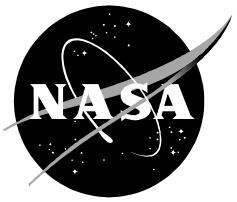
NASA/CR-2005-213690



NASA Summer Faculty Fellowship Program 2004

RESEARCH REPORTS

Volume 1

William A. Hyman, Donn G. Sickorez**, and Dawn M. Leveritt**,*

Editors

* *Texas A&M University*
NASA, Johnson Space Center
Houston, Texas

** *Lyndon B. Johnson Space Center*
Houston, Texas

August 2005

The NASA STI Program Office . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA's counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or cosponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and mission, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at ***http://www.sti.nasa.gov***

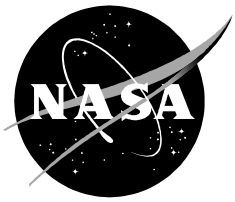
- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA Access Help Desk at (301) 621-0134

- Telephone the NASA Access Help Desk at (301) 621-0390

- Write to:
  NASA Access Help Desk
  NASA Center for AeroSpace Information
  7121 Standard
  Hanover, MD 21076-1320

NASA/CR-2005-213690



NASA Summer Faculty Fellowship Program 2004

RESEARCH REPORTS

Volume 1

William A. Hyman, Donn g. Sickorez**, and Dawn M. Leveritt**,*

Editors

* *Texas A&M UniversityDonn Sickorez, Ph.D.*
NASA, Johnson Space Center
Houston, Texas

** *Lyndon B. Johnson Space Center*
Houston, Texas

Grants NGT 9-1526 and NNJ04JF93A

National Aeronautics and
Space Administration

Johnson Space Center
Houston, Texas 77058-3696

August 2005

Available from:

NASA Center for AeroSpace Information
7121 Standard
Hanover, MD 21076-1320

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161

This report is also available in electronic form at http://techreports.larc.nasa.gov/cgi-bin/TRS

Preface

The 2004 Johnson Space Center (JSC) National Aeronautics and Space Administration Faculty Fellowship Program (NFFP) was conducted by Texas A&M University and JSC. The program was funded by the Office of Education, NASA Headquarters, Washington, D.C. and by JSC. Each faulty Fellow spent at least 10 weeks at JSC (or the White Sands Test Facility) engaged in a research project in collaboration with a NASA/JSC colleague.

This document is a compilation of the final reports on the research projects done by the Fellows during the summer of 2004. Volume 1 contains reports 1 through 12 and Volume 2 contains reports 13 through 22.

TABLE OF CONTENTS

Volume 1

Volume 2

Evaluation of Suited and Unsuited Human Functional Strength Using Multipurpose, Multiaxial Isokinetic Dynamometer

Final Report
NASA Faculty Fellowship Program – 2004
Johnson Space Center

Prepared By:	Fred Aghazadeh, Ph.D., P.E.
Academic Rank:	Associate Professor
University & Department:	Louisiana State University Industrial Engineering Department Baton Rouge, LA 70803
NASA/JSC	
Office:	Habitability and Environmental Factors Office (HEFO) Habitability & Human Factors Office Anthropometry and Biomechanics Facility (ABF)
JSC Colleague	Sudhakar L. Rajulu, Ph.D.
Date Submitted	August 10, 2004
Contract Number	NAG 9-1526 and NNJ04JF93A

ABSTRACT

The objective of the planned summer research was to develop a procedure to determine the isokinetic functional strength of suited and unsuited participants in order to estimate the coefficient of micro-gravity suit on human strength. To accomplish this objective, the Anthropometry and Biomechanics Facility's Multipurpose, Multiaxial Isokinetic dynamometer (MMID) was used. Development of procedure involved selection and testing of seven routines to be tested on MMID. We conducted the related experiemts and collected the data for 12 participants.

In addition to the above objective, we developed a procedure to assess the fatiguing characteristics of suited and unsuited participants using EMG technique. We collected EMG data on 10 participants while performing a programmed routing on MMID. EMG data along with information on the exerted forces, effector speed, number of repetitions, and duration of each routine were recorded for further analysis. Finally, gathering and tabulation of data for various human strengths for updating of MSIS (HSIS) strength requirement, which started in summer 2003, also continued.

INTRODUCTION

Extra-Vehicular Activities (EVA) are necessary part of being aboard International Space Station and the Space Shuttle. For EVA activities, astronauts must wear a pressurized suit called Extra-Vehicular Mobility Unit (EMU). Due to the pressurization and the bulkiness of the suit material, the EMU has been known to limit motion and impede generation of force. In order to determine the effect of suit on the strength and fatigue of the astronauts, several studies have been performed. The previous studies investigated such subjects as, effects of EVA gloves on performance, measurement of various strength variables whiled wearing EMU, determination of work and fatigue characteristics of suited and unsuited participants during isolated joint motions, etc.

However, the previous work on EMU focused solely on isolating individual joints and hence, lacks data on functional strength capabilities and limitations of a suited crewmember. The objectives of the planned summer research consisted of three parts;

Part I. Development of a procedure to determine isokinetic functional strength of suited and unsuited participants in order to estimate the coefficient of micro-gravity suit on human strength using the Anthropometry and Biomechanics Facility's Multipurpose, Multiaxial Isokinetic Dynamometer (MMID).

Part II. Develop a procedure to assess the fatiguing characteristics of suited and unsuited participants using the EMG methodology.

Part III. Continue the gathering of data for various human strengths for updating of MSIS strength requirement.

METHOD AND PROCEDURE

Apparatus

The initial stage of the project involved familiarization with the apparatus used in this project. This included obtaining the manual from the manufactures and writing of a JPA (Job Performance Aid) for the use of the equipment. The apparatus used in this project was Multipurpose, Multiaxial Isokinetic Dynamometer (MMID). The Multipurpose Multiaxial Isokinetic Dynamometer (MMID) is capable type of dynamometer used for measuring and stressing muscles in the arms, legs, and trunk. It monitors both strength and limb position in 3-D space. It is a new generation of physical fitness device capable of being used here on Earth as well as in the weightlessness of space. The key components of the MMID are the eight active modules. The modules reel in or spool out cable in unison to achieve a desired trajectory of the end effector. The MMID is capable of achieving complex, six degree-of-freedom motions by using all eight active modules. With all eight modules maintaining a given position, the end effector can be rigidly fixed

in space. Other advantage of the machine is the fact that each module is light (7 pounds) and requires almost minimum volume when stored (7 x7x7 inches).

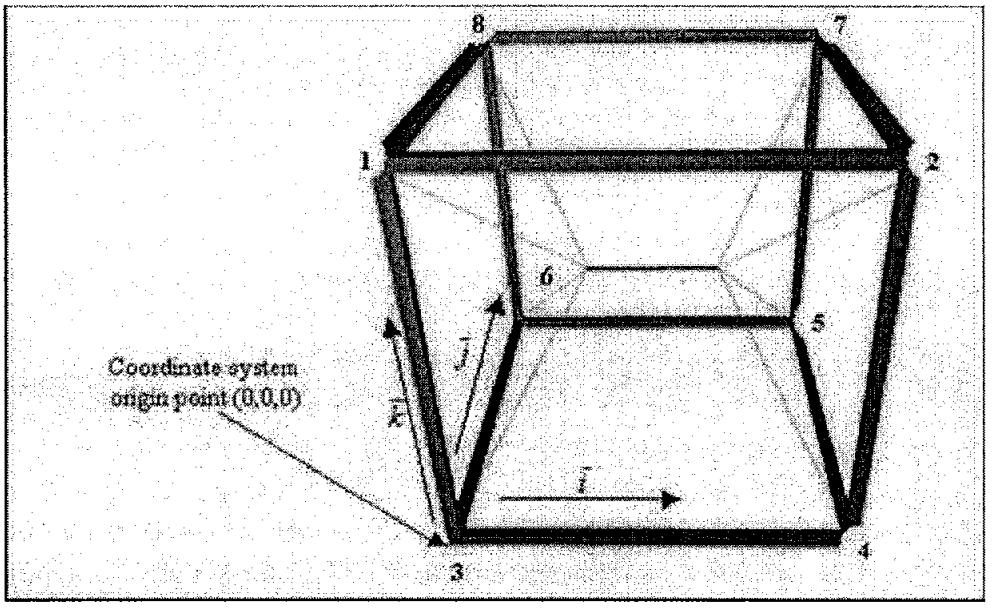


Figure 1: System Configuration and Coordinate Conventions

A diagram of the MMID system in its typical configuration is shown in Figure 1. This Figure shows the cubic configuration, the coordinate system origin point and orientation, and a typical end effector configuration (a bar). In this case, there are eight cables attached to eight points on the end effector, four points on each end. This configuration enables a comfortable balance between range of motion and force generating capability. Figure 2 illustrate one of the eight modules (pods). The module is small and may be mounted with ease.

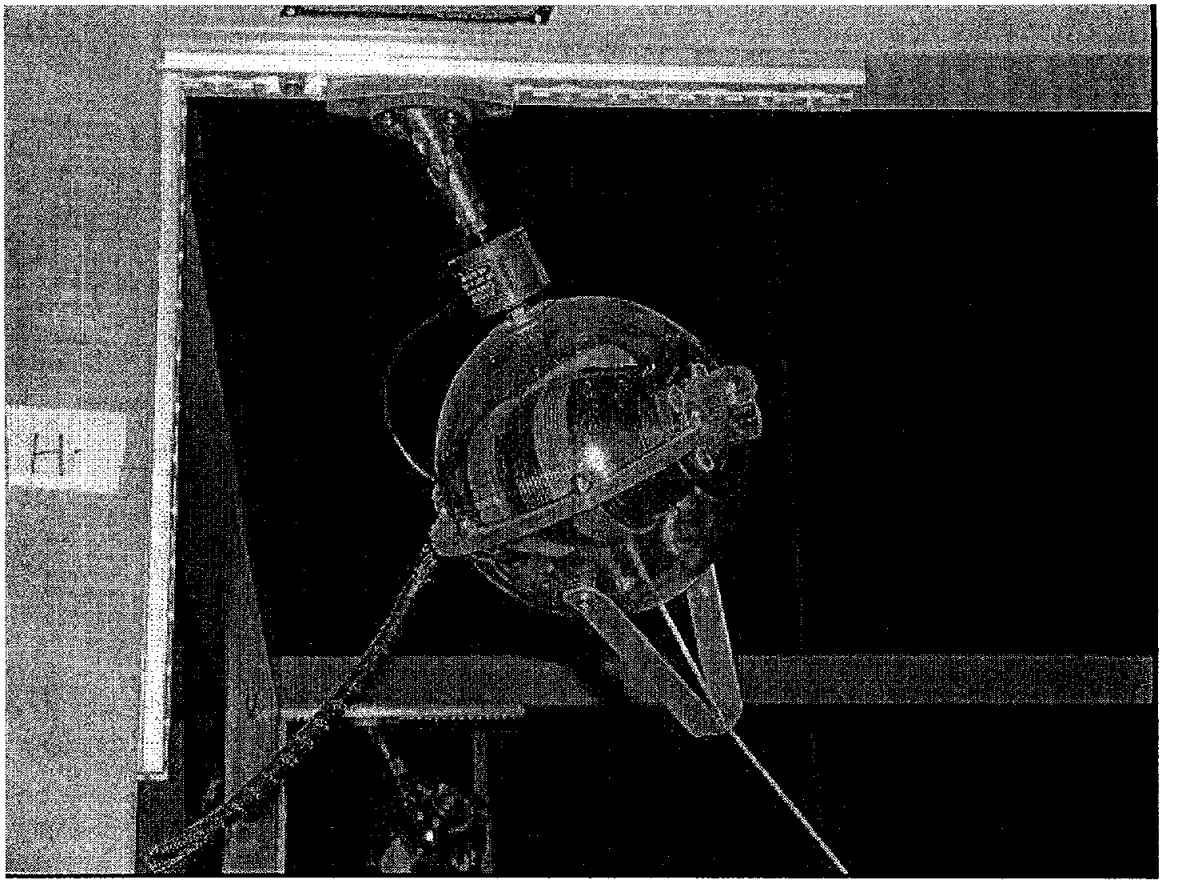


Figure 2. An Illustration of a Module (Pod)

Some measurement capability of MMID are listed below and shown in Figure 3 (Real-time Display Window) and Figure 4 (Real-time Information Display Window):

- Continuous recording of effector position in X,Y,Z axis
- Continuous recording of force (lbs), speed (in/s), acceleration (in/s^2) , deceleration,(in/s^2), and moment.

The apparatus is capable of measuring forces for the following routines:

Programmed Routines
Squat
Bench Press
Incline shoulder Press
Vertical Leg Press
Calf raises

Lat Pull-down (Latissimus dorsi)
Military Press
Bent-over Row
Butt Blaster
Inclined Leg Press
Triceps Press-down
Standing Curl
Cup Lift
Roto-swirl
Leg Curl
Seated Curl
Seated Curl
Inverted Push-up
Single Pod Pull
Bowl Move

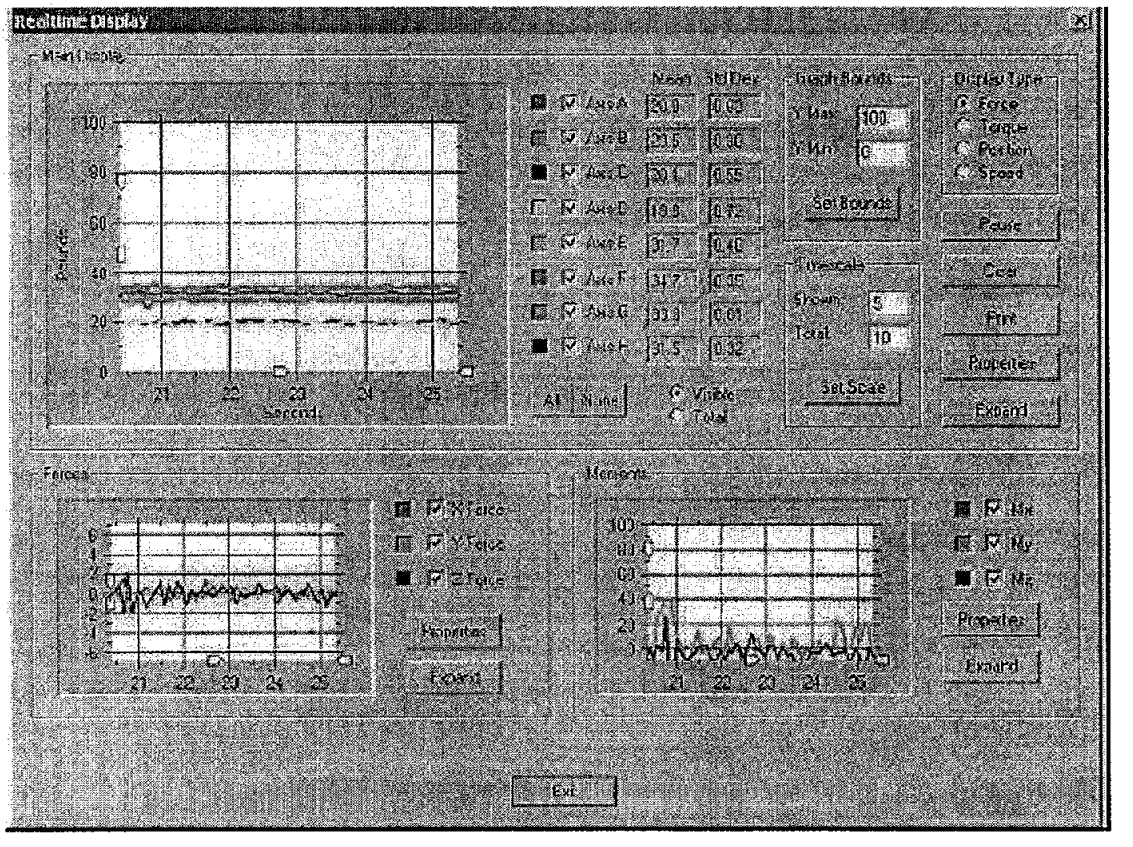


Figure 3. Real-time Display Window

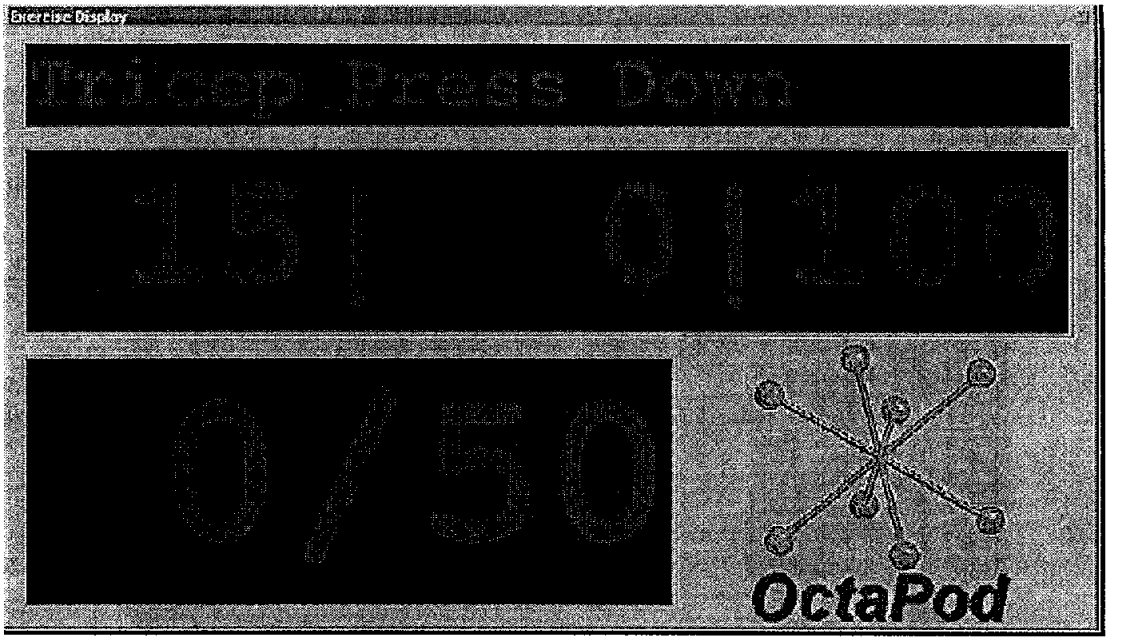


Figure 4: Real-time Information Display Window (Force in lbs.)

Procedure Development

Part I. In order to develop a procedure to determine isokinetic functional strength of suited and unsuited participants, we selected the routines, and specifications for each routine. This was done through trial and error, and was based on usability of the routine and capability of the measuring devise. The selected routines and their specifications are listed in Table 1.

Table 1. Specifications of Routines

Strength Test Positions & Specs	X-axis Position (inch)	Y-axis Position (inch)	Z-axis Position (inch)	Speed (in/s)	Acce-Leration (in/s^2)	Decele-Ration (in/s^2)	Min Move force (lbs)	Max registered Force (lbs)	Max move Force (lbs)
1. Sitting lat pull down	46	49	66-48	10	10	20	20	150	180
2. Sitting Mil Press	46	54	66-48	10	10	20	20	150	180
3. Sitting Push	46	36-63	46	10	10	10	20		
4. Sitting Pull	46	63-36	46	10	10	10	20		
5. Open Hatch	46	46	26-66	10	10	10	20		
6. Standing Curl	46	46	39-53	10	10	20	20	120	150
7. Standing Triceps Press	46	46	53-39	10	10	10	10	100	120
Fatigue Test									
1"/sec	46	36-63	46	1	1	1			
5"/ sec	46	36-63	46	5	5	5			
10"/sec	46	36-63	46	10	10	10			

Experiment Protocol for Parts 1 and 2

Once the routines were specified, we proceeded with the experimentation and data collection phase. The experiment involved 10 male and 2 female participants. The subjects were instructed to move the effector (bar) by pushing, pulling, raising, etc as fast and as hard they were able to do it without jerking the bar. They were told to build the speed and force gradually. The participants tried each routine a few times and became familiar with the routine. They were asked to perform the routine continuously till they were instructed to halt the routine.. The exerted force was monitored till values for three trials that were within 10 percent of each other were obtained. A rest period of 3 to 4 minutes between routines was used. The experimental order was randomized and is shown in Table 2.

The second part of this project was about determination of the fatiguing characteristics of participants using the EMG methodology. We used trial and error method to select a routine that was based on usability of the routine and capability of the measuring devise. The selected routine was the push-pull routine at speeds of 1, 5, and 10 inch/s. We selected the triceps muscle for electrode location. The EMG recording devise was the Bagnoli-2 EMG System (DelSys Inc., Boston, MA). This is a handheld, battery-operated 2-channel EMG system. Prior to the data collection, a surface electrode was placed on the right triceps muscle and a base EMG was recorded. The participants were asked to sit on the MMID bench and push the effector as hard and as fast they could without jerking the bar. The participants repeated the routine at the prescribed speed of 1, or 5, or 10 inch/s for 400 seconds. The order of trials was randomized. Table 2 shows the experiment order. It was observed that as the muscles fatigued, the exerted force did not change accordingly. The participants complained about fatigue in the hand and fingers. Data was collected and recorded for analysis in the future.

Table 2. Experiment Protocol

Subject Name and Number								
Gender								
Height								
Weight								
Age								
Strength Test Positions and order								
1. Sitting lat pull down	1	7	6	2	3	2	7	4
2. Sitting Mil Press	2	2	2	3	6	6	2	6
3. Sitting Push	3	3	3	4	1	5	5	1
4. Sitting Pull	4	4	7	6	5	3	3	3
5. Open Hatch	5	5	1	1	2	1	6	5
6. Standing Curl	6	6	4	5	4	4	1	7
7. Standing Triceps Press	7	1	5	7	7	7	4	2
Fatigue Test								
1"/sec	3	3	2	1	3	2	3	2
5"/ sec	2	1	1	2	2	3	1	1
10"/sec	1	2	3	3	1	1	2	3

SUMMARY AND FUTURE WORK

Part I included development of a procedure to determine isokinetic functional strength of suited and unsuited participants in order to estimate the coefficient of micro-gravity suit on human strength using the Anthropometry and Biomechanics Facility's Multipurpose, Multiaxial Isokinetic Dynamometer (MMID).

For this part of the project a detailed procedure was developed and tested. In addition, strength data for 10 male and 2 female unsuited participants were collected. The data includes strength data for seven routines at pre-determined X,Y, Z, coordinates, speeds, accelerations, decelerations, and resistive forces. The predetermined routines were "Lat Pull-down (Latissimus dorsi)," "Sitting Military Press," "programmed routines of Sitting Push, Sitting Pull, and Open Hatch," "Standing Curl," and "Standing Triceps Press."

Work to be completed includes:

A. Analysis of collected data to determine the functional strength capacity of participants.

B. Collection of the same data for suited participants.

C. Analysis of the suited data to determine the coefficient of micro-gravity suit on human strength.

Part II was about the development of a procedure to assess the fatiguing characteristics of suited and unsuited participants using the EMG methodology. After many trials and errors, we found that the push-pull routine at speeds of 1, 5, and 10 inch/s was the practical test for this part. EMG signal of the triceps muscle was recorded at 1000 hz. for 400 seconds for 10 male participants. Work to be completed includes:

A. Analysis of the collected data to determine the relationship among variables of generated force, effector speed, number of repetitions, and EMG signals.

B. Collection and analysis of the data for suited participants.

Part III was continuation of task of gathering of data for various human strengths for updating of NASA-STD-3000 Man-System Integration Standards (MSIS) strength requirement. At the present time about 190 references related to different human strengths have been located and documented. The following is a list of journals and the number of references collected from those Journals.

Journal Name	*Number of References*
Ergonomics	20
Research Quarterly	15
Applied Ergonomics	04
Journal of Biomechanics	06
International Journal of Industrial Ergonomics	03
Spine	01
Hand	01
Human Factors	04
Scandinavian Journal of Rehabilitation Medicine	02
Clinical Biomechanics	04
Clinical Orthopedics & Related Research	03
American Industrial Hygiene Association Journal	05
Proceedings of Human Factors	15
Journal of Applied Physiology	13
Archives of Physical Medicine & Rehabilitation	07
American Corrective Therapy Journal	05
Aviation Space and Environmental Medicine	03
Aerospace Medicine	02
Physical Therapy	07
Journal of Sports Medicine and Physical Fitness	03
Medicine and Science in Sports	05
Scandinavian Journal of Rheumatology	02
Journal of Hand Surgery	04
Journal of Bone and Joint Surgery	04
American Journal of Occupational Therapy	03
Engineering in Medicine	02
Journal of Orthopedics Research	03
Journal of Anatomy	04

Future work to be conducted includes:

A. Locating and documenting of additional sources

B. locating and collecting hard or electronic copies

C. Summarizing sources, extracting useful data, tables etc., and building an annotated bibliography

C. Tabulating sources into sections related to:

Strength in micro-gravity and hyper-gravity conditions
Strength of various limbs
Isokinetic strength data
Isometric strength data
Static strength data vs. functional strength
Strength in IVA (Intravehicular Activity) and EVA (Extravehicular Activity) suits

Design and Performance Evaluation of a UWB Communication and Tracking System for Mini-AERCam

Final Report
NASA Faculty Fellowship Program – 2004

Johnson Space Center

Prepared by:	Richard J. Barton, Ph.D.
Academic Rank:	Assistant Professor
University and Department:	University of Houston Dept. of Elec. and Comp. Eng. Houston, TX 77204

NASA/JSC

Directorate:	Engineering
Division:	Avionic Systems
Branch:	Electromagnetic Systems
JSC Colleague:	G. D. Arndt
Date: Submitted:	August 6, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

Abstract

NASA Johnson Space Center (JSC) is developing a low-volume, low-mass, robotic free-flying camera known as Mini-AERCam (Autonomous Extra-vehicular Robotic Camera) to assist the International Space Station (ISS) operations. Mini-AERCam is designed to provide astronauts and ground control real-time video for camera views of ISS. The system will assist ISS crewmembers and ground personnel to monitor ongoing operations and perform visual inspections of exterior ISS components without requiring extravehicular activity (EAV).

Mini-AERCam consists of a great number of subsystems. Many institutions and companies have been involved in the R&D for this project. A Mini-AERCam ground control system has been studied at Texas A&M University [3]. The path planning and control algorithms that direct the motions of Mini-AERCam have been developed through the joint effort of Carnegie Mellon University and the Texas Robotics and Automation Center [5]. NASA JSC has designed a layered control architecture that integrates all functions of Mini-AERCam [8]. The research described in this report is part of a larger effort focused on the communication and tracking subsystem that is designed to perform three major tasks:

1. To transmit commands from ISS to Mini-AERCam for control of robotic camera motions (downlink);
2. To transmit real-time video from Mini-AERCam to ISS for inspections (uplink);
3. To track the position of Mini-AERCam for precise motion control.

The ISS propagation environment is unique due to the nature of the ISS structure and multiple RF interference sources [9]. The ISS is composed of various truss segments, solar panels, thermal radiator panels, and modules for laboratories and crew accommodations. A tracking system supplemental to GPS is desirable both to improve accuracy and to eliminate the structural blockage due to the close proximity of the ISS which could at times limit the number of GPS satellites accessible to the Mini-AERCam. Ideally, the tracking system will be a passive component of the communication system which will need to operate in a time-varying multipath environment created as the robot camera moves over the ISS structure. In addition, due to many interference sources located on the ISS, SSO, LEO satellites and ground-based transmitters, selecting a frequency for the ISS and Mini-AERCam link which will coexist with all interferers poses a major design challenge. To meet all of these challenges, ultrawideband (UWB) radio technology is being studied for use in the Mini-AERCam communication and tracking subsystem. The research described in this report is focused on design and evaluation of passive tracking system algorithms based on UWB radio transmissions from mini-AERCam.

Introduction

UWB Technology

Ultrawideband radio, in particular impulse or carrier-free radio technology, is one promising new technology for low-power communications and tracking systems. It has been utilized for decades by the military and law enforcement agencies for fine-resolution ranging, covert communications and ground-penetrating radar applications. In February 2002, the Federal Communications Commission (FCC) approved the deployment of this technology in the commercial sector under Part 15 of its regulations [6]. UWB technology holds great potential to provide significant benefits in many applications such as precise positioning, short-range multimedia services and high-speed mobile wireless communications.

The DARPA study panel that coined the term *ultrawideband* in the 1990s defines it as a system with a fractional bandwidth greater than 25%. The basic concept of current UWB impulse radio technology is to transmit and receive an extremely short duration burst of RF energy – typically a few tens of picoseconds to a few nanoseconds in duration. Whereas conventional continuous-wave radio systems operate within a relatively narrow bandwidth, UWB operates across a wide range of frequency spectrum (a few GHz) by transmitting a series of low-power impulsive signals.

For the emerging technology of UWB radar and UWB wireless communications, the transmitted signal can be regarded as a uniform train of UWB pulses represented as

$$s(t) = \sum_{n=-\infty}^{+\infty} \omega(t - nT_r),$$

where T_r is the pulse repetition interval, and $\omega(t)$ is the pulse-shaping waveform which is often a Gaussian monocycle. In the time domain, the Gaussian monocycle is mathematically similar to the first derivative of Gaussian function. It has the form

$$\omega(t) = \frac{t}{\tau} e^{-(t/\tau)^2},$$

where τ is regarded as the duration of the monocycle. Figure 1 shows an ideal monocycle centered at 2 GHz in both the time and frequency domains [1].

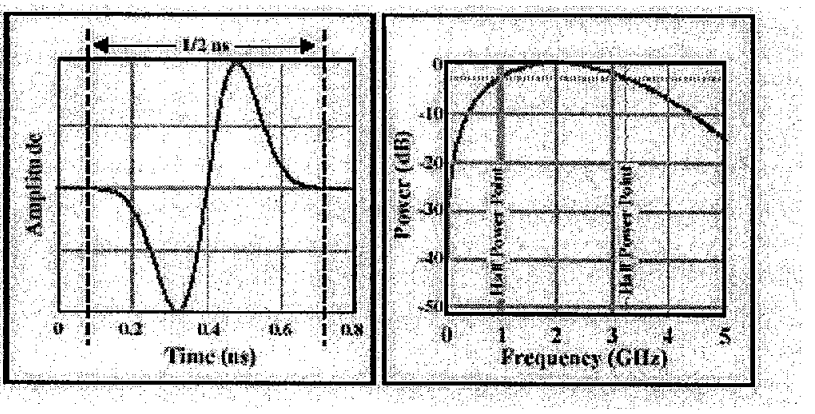


Figure 1. Gaussian monocycle in time and frequency domains

UWB impulse radio is characterized by several uniquely attractive features:

- Low-power, carrier-free, ultra-wide bandwidth signal transmissions
- Minimal interference on other RF systems due to extremely low power spectral density
- Immunity to interference from narrow-band RF systems due to extremely large bandwidth
- Immunity to multipath fading due to ample multipath diversity (RAKE receiver)
- Capable of precise positioning due to fine time resolution
- Capable of high data rate, multi-channel performance due to extremely large bandwidth
- Low-complexity, low-power baseband transceivers without intermediate frequency stage

Rapid technological advances have enabled the implementation of cost-effective UWB radar, communication, and tracking systems. Furthermore, antenna-array beamforming and space–time processing techniques promise further advancement in the capability of UWB technology to achieve long-range coverage, high capacity, and interference-free quality of reception [7].

Tracking Algorithm

To make Mini-AERCam coordinated maneuvers feasible, an accurate, robust, and self-contained tracking system that is small, low power, and low cost is required. Compared to GPS receivers, which can offer range resolution on the order of one meter, UWB radio can achieve sub-centimeter range resolution much faster and with less effort [1]. The experiment described in [2] demonstrates that UWB systems can provide range measurements accurate to the centimeter level over distances of kilometers, using only milliwatts of power from an omni-directional transceiver no bigger than a pager. In this research effort, the tracking subsystem will be designed to provide the precise positioning

required for Mini-AERCam motion control as a byproduct of the UWB video communication system.

Many technologies have been applied to locating the source of radio signals, such as angle of arrival (AOA), time difference of arrival (TDOA) and relative signal strength (RSS). The extremely high fidelity of the UWB timing circuitry permits very precise measurement of propagation times for transmitted signals. This fine time resolution feature of UWB motivates us to apply a TDOA approach for tracking system design. We will utilize the header of the video data packets as the pilot signal to implement a passive tracking system in which the tracking rides on top of video communications [10].

Since electromagnetic waves travel with constant velocity in free space, the distance between the transmitter and the receiver is directly proportional to the propagation time. The TDOA approach determines the possible position of the transmitter by examining the difference in time at which the same signal arrives at multiple receivers. Each TDOA measurement determines a hyperboloid corresponding to the surface of constant range difference between the two receivers. At least three receivers are needed for a 2-D location estimate and four receivers for a 3-D location estimation. The intersection of the hyperboloids corresponding to all the TDOA measurements provides the location of the transmitter.

Suppose N receivers measure the time of arrival (TOA) of pilot signals from the transmitter in a 2-D case. The TOA estimates of the signal from receiver i and j are denoted t_i and t_j respectively. A range difference measurement r_k can be calculated from the TDOA measurement as follows:

$$r_k = c(t_i - t_j) = d_i - d_j = f_k(x,y), \tag{1}$$

where d_i and d_j are the distances from the transmitter to receivers i and j, respectively, and c is the propagation velocity of the signals, which is generally taken to be the speed of light in free space. Since the positions of all the receivers are known, this equation is a function $f_k(x,y)$ only of the unknown coordinate position of the transmitter (x,y).

In many cases, the transmitter location is determined by finding a least-squares solution to a linearized version of Equation (1) . The linearization is given by

$$f_k(x,y) = f_k(x_0,y_0) + \frac{\partial f_k}{\partial x}(x - x_0) + \frac{\partial f_k}{\partial y}(y - y_0),$$

where the partial derivatives are evaluated at the *a priori* estimate for the transmitter position (x_0,y_0). This estimate is normally a previous solution for the transmitter position. The linearized version of Equation (1) can then be expressed as a matrix equation of the form $\mathbf{A}\mathbf{p}_T = \mathbf{b}$, where

$$\mathbf{A}=\begin{bmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \mathrm{M} & \mathrm{M} \\ \frac{\partial f_K}{\partial x} & \frac{\partial f_K}{\partial y} \end{bmatrix} \quad \mathbf{p}_T=\begin{bmatrix} (x-x_0) \\ (y-y_0) \end{bmatrix} \quad \mathbf{b}=\begin{bmatrix} r_1-f_1(x_0,y_0) \\ \mathrm{M} \\ r_K-f_K(x_0,y_0) \end{bmatrix}.$$

The least squares solution to this matrix equation is the estimated position of the transmitter, which is given by

$$\hat{\mathbf{p}}_T=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}.$$

The variance of the position estimate is related to the variance of the time estimate. Tracking requires that the direct path portion of the UWB signal be located and its arrival time inserted into the tracking algorithm. Hence, the accuracy of the TDOA estimates is very critical for position tracking. The conventional approach to estimating TDOA is to compute the cross-correlation of an identical transmitted signal arriving at different receivers. The TDOA estimate for each pair of receivers is given by the delay that maximizes the cross-correlation function. To complete the tracking algorithm, the sequence of position estimates is passed to a Kalman filter to update the current estimate of position.

Discussion of Results

The results of the research completed by the PI during the 2004 NASA Faculty Fellowship Program are discussed in the remainder of this report. The discussion is divided into two sections. The first presents a careful analysis of the statistical properties of a particular TDOA localization algorithm currently identified for use in the mini-AERCam communication and tracking subsystem. The second presents a proposed modification to this algorithm that should improve the localization accuracy.

Analysis of Current Algorithm

The TDOA localization algorithm currently proposed for use in the mini-AERCam communication and tracking subsystem was developed and analyzed by Chan and Ho in [4]. The statistical properties of the algorithm are analyzed to some extent in [4], and while the analysis presented there is essentially correct, it is also sketchy and incomplete, making a thorough performance evaluation of the algorithm problematic. To correct this deficiency, the PI performed a careful and complete analysis of the statistical properties of this algorithm in two dimensions. The results of that analysis are discussed in this section.

Assume that there is one transmitter located at an unknown location (x_0,y_0) in two-dimensional space and $M+1$ receivers located at positions $\{(0,0),(x_1,y_1),(x_2,y_2),\mathrm{K},(x_M,y_M)\}$, which are assumed to be known precisely.

Further, assume that measurements of the relative time delays $\{d_1, d_2, \mathsf{K}, d_M\}$ between the arrival of the transmitted signal at receiver $(0,0)$ and each of the other locations $(x_1, y_1), \mathsf{K}, (x_M, y_M)$ are available. If the propagation velocity of the signals is given by the constant c, then it can be shown that the following system of linear equations is satisfied:

$$\mathbf{G}_0 \mathbf{u}_0 = \mathbf{h}_0, \tag{2}$$

where,

$$\mathbf{u}_0 = \begin{bmatrix} x_0 \\ y_0 \\ r_0 \end{bmatrix}, \quad r_0 = \sqrt{x_0^2 + y_0^2}, \quad \mathbf{G}_0 = -2 \cdot \begin{bmatrix} x_1 & y_1 & cd_1 \\ x_2 & y_2 & cd_2 \\ \mathsf{M} & \mathsf{M} & \mathsf{M} \\ x_M & y_M & cd_M \end{bmatrix}, \quad \mathbf{h}_0 = \begin{bmatrix} c^2 d_1^2 - x_1^2 - y_1^2 \\ c^2 d_2^2 - x_2^2 - y_2^2 \\ \mathsf{M} \\ c^2 d_M^2 - x_M^2 - y_M^2 \end{bmatrix}.$$

If the time delay measurements are not precisely correct, we have instead the system

$$\mathbf{G}_1 \mathbf{u}_0 = \mathbf{h}_1 - (\Delta \mathbf{h}_1 - \Delta \mathbf{G}_1 \mathbf{u}_0), \tag{3}$$

where

$$\mathbf{G}_1 = \mathbf{G}_0 + \Delta \mathbf{G}_1, \quad \mathbf{h}_1 = \mathbf{h}_0 + \Delta \mathbf{h}_1, \quad \Delta \mathbf{G}_1 = -2 \cdot \begin{bmatrix} 0 & 0 & c\delta_1 \\ 0 & 0 & c\delta_2 \\ \mathsf{M} & \mathsf{M} & \mathsf{M} \\ 0 & 0 & c\delta_M \end{bmatrix}, \quad \Delta \mathbf{h}_1 = \begin{bmatrix} c^2\delta_1^2 + 2c^2 d_1 \delta_1 \\ c^2\delta_2^2 + 2c^2 d_2 \delta_2 \\ \mathsf{M} \\ c^2\delta_M^2 + 2c^2 d_M \delta_M \end{bmatrix},$$

and $\delta = \begin{bmatrix} \delta_1 & \delta_2 & \mathsf{L} & \delta_M \end{bmatrix}^T$ represents the vector of errors in the relative time delay measurements, which is assumed to be a zero-mean Gaussian random vector with covariance matrix $\mathbf{Q} = E\{\delta\delta^T\}$. For future reference, it is worthwhile to note that $\Delta \mathbf{h}_1$ can be rewritten as

$$\Delta \mathbf{h}_1 = 2c\mathbf{R}\delta + c^2 \delta \circ \delta,$$

where $\mathbf{R} = \mathrm{diag}(r_1, r_2, \mathsf{K}, r_M)$, $r_i = cd_i$, for $i = 1,2,\mathsf{K}, M$, and $\delta \circ \delta = \begin{bmatrix} \delta_1^2 & \delta_2^2 & \mathsf{L} & \delta_M^2 \end{bmatrix}^T$ represents the Hadamard product of the vector δ with itself.

The estimate $(\hat{x}, \hat{y})$ of (x_0, y_0) is computed recursively in three stages. In stage 1, a weighted least squares approach is used to find an estimate of $\mathbf{u}_0 = \begin{bmatrix} x_0 & y_0 & r_0 \end{bmatrix}^T$, which is given by

$$\hat{\mathbf{u}}_1 = \left(\mathbf{G}_1^T \mathbf{W}_1 \mathbf{G}_1\right)^{-1} \mathbf{G}_1^T \mathbf{W}_1 \mathbf{h}_1, \tag{4}$$

where the weighting matrix $\mathbf{W}_1$ is chosen as an approximation to the matrix $\mathbf{W}_0 = \left[E\left\{ (\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0)(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0)^T \right\} \right]^{-1}$. Since the statistical properties of the solution given by Equation (4) are relatively insensitive to small variations in the choice of the weighting matrix, we will make the simplifying assumption that $\mathbf{W}_1 = \mathbf{W}_0$ throughout the remainder of the statistical analysis. In practice, the matrix $\mathbf{W}_1$ is an estimate of $\mathbf{W}_0$ based on the observed data, but the error in the derived statistical properties introduced by this assumption is insignificant in comparison to the other simplifying assumptions made in the analysis.

Letting $\hat{\mathbf{u}}_1 = \mathbf{u}_0 + \Delta\mathbf{u}_1$, Equation (4) gives

$$\left(\mathbf{G}_0 + \Delta\mathbf{G}\right)^T \mathbf{W}_1 \left(\mathbf{G}_0 + \Delta\mathbf{G}\right)\left(\mathbf{u}_0 + \Delta\mathbf{u}_1\right) = \left(\mathbf{G}_0 + \Delta\mathbf{G}\right)^T \mathbf{W}_1 \left(\mathbf{h}_0 + \Delta\mathbf{h}\right).$$

This expression can be solved for an approximation to $\Delta\mathbf{u}_1$ by expanding, simplifying, and dropping terms involving powers of the vector δ greater than two. Proceeding in this manner and recalling that $\mathbf{G}_0\mathbf{u}_0 = \mathbf{h}_0$ and $\mathbf{W}_1 = \mathbf{W}_0$, we get

$$\begin{aligned}
\Delta\mathbf{u}_1 = &\left[\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0 + \Delta\mathbf{G}_1^T\mathbf{W}_0\mathbf{G}_0 + \mathbf{G}_0^T\mathbf{W}_0\Delta\mathbf{G}_1 + \Delta\mathbf{G}_1^T\mathbf{W}_0\Delta\mathbf{G}_1 \right]^{-1} \\
&\cdot \left(\mathbf{G}_0 + \Delta\mathbf{G}_1\right)^T \mathbf{W}_0 \left(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0\right) \\
\approx &\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} \left(\mathbf{G}_0 + \Delta\mathbf{G}_1\right)^T \mathbf{W}_0 \left(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0\right) \\
&- \left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} \left(\Delta\mathbf{G}_1^T\mathbf{W}_0\mathbf{G}_0 + \mathbf{G}_0^T\mathbf{W}_0\Delta\mathbf{G}_1\right)\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} \mathbf{G}_0^T\mathbf{W}_0 \left(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0\right).
\end{aligned}$$

Using this result, and assuming that $E\{\delta\} = 0$, we see that an approximation for the bias in Stage 1 of the algorithm is given by

$$\begin{aligned}
\mu_1 &= E\left\{\Delta\mathbf{u}_1\right\} \\
&\approx c^2 \left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} \mathbf{G}_0^T\mathbf{W}_0 \left(\begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \mathsf{M} \\ \sigma_M^2 \end{bmatrix} + 4\mathbf{Q}(\mathbf{R} + r_0\mathbf{I})\mathbf{W}_0\mathbf{G}_0 \left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right) \\
&\quad - 4c^2 \mathrm{Tr}\left(\mathbf{W}_0 \left[\mathbf{G}_0 \left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} \mathbf{G}_0^T\mathbf{W}_0 - \mathbf{I} \right](\mathbf{R} + r_0\mathbf{I})\mathbf{Q} \right) \left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},
\end{aligned} \tag{5}$$

where $\sigma_1^2 = E\left\{\delta_1^2\right\}$, $\sigma_2^2 = E\left\{\delta_2^2\right\}$, ..., $\sigma_M^2 = E\left\{\delta_M^2\right\}$. Proceeding in a similar fashion, and recalling that $\mathbf{W}_0 = \left[E\left\{(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0)(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0)^T\right\}\right]^{-1}$, we can derive an approximation for the autocorrelation matrix of $\Delta\mathbf{u}_1$. In particular, we get

$$\begin{aligned} E\left\{\Delta\mathbf{u}_1\Delta\mathbf{u}_1^T\right\} &\approx \left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1}\mathbf{G}_0^T\mathbf{W}_0 E\left\{(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0)(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0)^T\right\} \\ &\quad \cdot \mathbf{W}_0\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} \\ &= \left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1}. \end{aligned} \tag{6}$$

Note that this implies that the approximation for the covariance matrix of $\Delta\mathbf{u}_1$ takes the form

$$\Sigma_1 \approx \left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1} - \mu_1\mu_1^T,$$

which is slightly different than the result presented in [4], where the bias of the estimate was ignored. Finally, it is straightforward to show that

$$\mathbf{W}_0 = \left[E\left\{(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0)(\Delta\mathbf{h}_1 - \Delta\mathbf{G}_1\mathbf{u}_0)^T\right\}\right]^{-1} = \frac{1}{4c^2}\left(\mathbf{B}_1\mathbf{Q}\mathbf{B}_1\right)^{-1}, \tag{7}$$

where $\mathbf{B}_1 = \mathbf{R} + r_0\mathbf{I}$. Equations (5) and (6) with $\mathbf{W}_0$ given by Equation (7) constitute the desired statistical properties for Stage 1 of the algorithm.

For the second stage of the algorithm, the possible inconsistency between the estimated values for (x_0, y_0) and $r_0 = \sqrt{x_0^2 + y_0^2}$ obtained in the vector $\hat{\mathbf{u}}_1$ is resolved by computing a new estimate of $\left(x_0^2, y_0^2\right)$ based on $\hat{\mathbf{u}}_1$. The approach is again weighted least squares, and we begin by defining

$$\mathbf{G}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{h}_2 = \hat{\mathbf{u}}_1 \circ \hat{\mathbf{u}}_1 = \begin{bmatrix} (\hat{\mathbf{u}}_1(1))^2 \\ (\hat{\mathbf{u}}_1(2))^2 \\ (\hat{\mathbf{u}}_1(3))^2 \end{bmatrix}, \quad \Delta\mathbf{h}_2 = \mathbf{h}_2 - \mathbf{u}_0 \circ \mathbf{u}_0 = \mathbf{h}_2 - \mathbf{G}_2\begin{bmatrix} x_0^2 \\ y_0^2 \end{bmatrix}.$$

Clearly, the equation

$$\mathbf{G}_2\begin{bmatrix} x_0^2 \\ y_0^2 \end{bmatrix} = \mathbf{u}_0 \circ \mathbf{u}_0$$

is always satisfied, while the associated equation

$$\mathbf{G}_2\begin{bmatrix}(\hat{\mathbf{u}}_1(1))^2\\(\hat{\mathbf{u}}_1(2))^2\end{bmatrix}=\mathbf{h}_2 \tag{8}$$

will not generally be satisfied. The estimate $\hat{\mathbf{u}}_2$ of $\left(x_0^2,y_0^2\right)$ is computed as a weighted least squares solution to Equation (8) given by

$$\hat{\mathbf{u}}_2=\left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{W}_2\mathbf{h}_2,$$

where $\mathbf{W}_2$ is an estimate of the inverse of the autocorrelation matrix $E\left\{\Delta\mathbf{h}_2\Delta\mathbf{h}_2^T\right\}$. Assuming that the error $\Delta\mathbf{u}_1$ in the Stage 1 estimate $\hat{\mathbf{u}}_1$ is small, we have

$$\Delta\mathbf{h}_2\approx 2\mathbf{B}_2\Delta\mathbf{u}_1,$$

where $\mathbf{B}_2=\mathrm{diag}\left(x_0,y_0,r_0\right)$, and it follows from Equation (6) that

$$E\left\{\Delta\mathbf{h}_2\Delta\mathbf{h}_2^T\right\}\approx 4\mathbf{B}_2E\left\{\Delta\mathbf{u}_1\Delta\mathbf{u}_1^T\right\}\mathbf{B}_2\approx 4\mathbf{B}_2\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1}\mathbf{B}_2.$$

Again, the statistical properties of the estimate $\hat{\mathbf{u}}_2$ are insensitive to small variations in $\mathbf{W}_2$, so we simply assume for the remaining analysis that $\mathbf{W}_2^{-1}=4\mathbf{B}_2\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1}\mathbf{B}_2$. Finally, letting

$$\hat{\mathbf{u}}_2=\begin{bmatrix}x_0^2\\y_0^2\end{bmatrix}+\Delta\mathbf{u}_2,$$

it follows that

$$\begin{aligned}\Delta\mathbf{u}_2&=\hat{\mathbf{u}}_2-\begin{bmatrix}x_0^2\\y_0^2\end{bmatrix}=\left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{W}_2\mathbf{h}_2-\begin{bmatrix}x_0^2\\y_0^2\end{bmatrix}\\&=\left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{W}_2\mathbf{h}_2-\left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\begin{bmatrix}x_0^2\\y_0^2\end{bmatrix}\\&=\left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{W}_2\left(\mathbf{h}_2-\mathbf{G}_2\begin{bmatrix}x_0^2\\y_0^2\end{bmatrix}\right)=\left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{W}_2\Delta\mathbf{h}_2\\&\approx 2\left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{W}_2\mathbf{B}_2\Delta\mathbf{u}_1\\&\approx 2\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\Delta\mathbf{u}_1.\end{aligned}$$

Hence,

$$
\begin{aligned}
\boldsymbol{\mu}_2 &= E\left\{\Delta\mathbf{u}_2\right\} \\
&\approx 2\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0 E\left\{\Delta\mathbf{u}_1\right\} \\
&\approx 2c^2\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1} \\
&\quad\cdot\left[\begin{array}{l}
\mathbf{G}_0^T\mathbf{W}_0\left(\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\\ \mathsf{M}\\ \sigma_M^2\end{bmatrix}+4\mathbf{Q}\mathbf{B}_1\mathbf{W}_0\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1}\begin{bmatrix}0\\0\\1\end{bmatrix}\right) \\
\\
-4\mathrm{Tr}\left(\mathbf{W}_0\left[\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1}\mathbf{G}_0^T\mathbf{W}_0-\mathbf{I}\right]\mathbf{B}_1\mathbf{Q}\right)\begin{bmatrix}0\\0\\1\end{bmatrix}
\end{array}\right],
\end{aligned}
\tag{9}
$$

and

$$
\begin{aligned}
\boldsymbol{\Sigma}_2 &= E\left\{\Delta\mathbf{u}_2\Delta\mathbf{u}_2^T\right\} \\
&\approx \left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{W}_2 E\left\{\Delta\mathbf{h}_2\Delta\mathbf{h}_2^T\right\}\mathbf{W}_2\mathbf{G}_2\left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1} \\
&= \left(\mathbf{G}_2^T\mathbf{W}_2\mathbf{G}_2\right)^{-1} \\
&= 4\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}.
\end{aligned}
\tag{10}
$$

Equations (9) and (10) constitute the desired statistical properties for Stage 2 of the algorithm.

For the third and final stage of the algorithm, the estimated values of $\left(x_0^2, y_0^2\right)$ obtained in the vector $\hat{\mathbf{u}}_2$ are used to obtain a final estimate of $\left(x_0, y_0\right)$ based on $\hat{\mathbf{u}}_2$ and $\hat{\mathbf{u}}_1$ combined. The final estimate $\hat{\mathbf{u}}_3$ of $\left(x_0, y_0\right)$ is given by

$$
\hat{\mathbf{u}}_3 = \mathbf{P}\sqrt{\hat{\mathbf{u}}_3}\,,
$$

where, in this case, "$\sqrt{(\cdot)}$" indicates a component-wise square-root operation and $\mathbf{P} = \mathrm{diag}\left(\mathrm{sgn}\left[\hat{\mathbf{u}}_1(1)\right], \mathrm{sgn}\left[\hat{\mathbf{u}}_1(2)\right]\right)$. Assuming that the sign of each coordinate in $\hat{\mathbf{u}}_1$ is correct, it follows that

$$\Delta\mathbf{u}_3 = \hat{\mathbf{u}}_3 - \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} \approx \frac{1}{2}\mathbf{B}_3^{-1}\Delta\mathbf{u}_2$$

$$\approx \mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\Delta\mathbf{u}_1,$$

where $\mathbf{B}_3 = \text{diag}\{x_0, y_0\}$. Hence, the final approximations for the bias vector and autocorrelation matrix of the algorithm are given by

$$\mu_3 = E\{\Delta\mathbf{u}_3\}$$

$$\approx \frac{1}{2}\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0 E\{\Delta\mathbf{u}_1\}$$

$$\approx c^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}$$

$$\cdot\begin{bmatrix} \mathbf{G}_0^T\mathbf{W}_0\left(\begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \mathsf{M} \\ \sigma_M^2 \end{bmatrix} + 4\mathbf{Q}\mathbf{B}_1\mathbf{W}_0\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) \\ -4\text{Tr}\left(\mathbf{W}_0\left[\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\right)^{-1}\mathbf{G}_0^T\mathbf{W}_0 - \mathbf{I}\right]\mathbf{B}_1\mathbf{Q}\right)\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{bmatrix} \tag{11}$$

$$= c^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{B}_1^{-1}\mathbf{Q}^{-1}\mathbf{B}_1^{-1}\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}$$

$$\cdot\begin{bmatrix} \mathbf{G}_0^T\mathbf{B}_1^{-1}\mathbf{Q}^{-1}\mathbf{B}_1^{-1}\left(\begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \mathsf{M} \\ \sigma_M^2 \end{bmatrix} + 4\mathbf{B}_1^{-1}\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{B}_1^{-1}\mathbf{Q}^{-1}\mathbf{B}_1^{-1}\mathbf{G}_0\right)^{-1}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) \\ -4\text{Tr}\left(\mathbf{B}_1^{-1}\mathbf{Q}^{-1}\mathbf{B}_1^{-1}\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{B}_1^{-1}\mathbf{Q}^{-1}\mathbf{B}_1^{-1}\mathbf{G}_0\right)^{-1}\mathbf{G}_0^T\mathbf{B}_1^{-1} - \mathbf{B}_1^{-1}\right)\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \end{bmatrix},$$

and

$$
\begin{aligned}
\Sigma_3 &= E\left\{\Delta\mathbf{u}_3\Delta\mathbf{u}_3^T\right\} \\
&= \frac{1}{4}\mathbf{B}_3^{-1}E\left\{\Delta\mathbf{u}_2\Delta\mathbf{u}_2^T\right\}\mathbf{B}_3^{-1} \\
&\approx 4c^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{W}_0\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{B}_3^{-1} \\
&= 4c^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{B}_1^{-1}\mathbf{Q}^{-1}\mathbf{B}_1^{-1}\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{B}_3^{-1}.
\end{aligned}
\tag{12}
$$

The proposed modifications to this algorithm are discussed in the following and final section of this report.

Proposed Algorithm Modification

In an effort to improve the localization and tracking performance, we propose a simple modification to Stage 1 of the algorithm described above. We consider a Stage 1 estimate based upon a weighted *total-least-squares* solution instead of the weighted *least-squares* solution currently used.[1] In order to describe the proposed new Stage 1 algorithm, we let $\mathbf{G}_0$, $\mathbf{G}_1$, $\mathbf{h}_0$, and $\mathbf{h}_1$ be defined as above, but we change the definition of $\Delta\mathbf{G}_1$ just slightly. In particular, we let

$$
\Delta\mathbf{G}_1 = -2\cdot\begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & c\delta_1 \\ \varepsilon_{21} & \varepsilon_{22} & c\delta_2 \\ \mathrm{M} & \mathrm{M} & \mathrm{M} \\ \varepsilon_{M1} & \varepsilon_{M2} & c\delta_M \end{bmatrix},
$$

where $\left\{\varepsilon_{ij} : i,j = 1,2,\mathrm{K}\ M\right\}$ are independent, identically distributed random variables with mean zero and variance σ_ε^2, where $\sigma_\varepsilon^2 << \min\left\{\sigma_1^2,\sigma_2^2,\mathrm{K}\ ,\sigma_M^2\right\}$. This is merely a device for making $\Delta\mathbf{G}_1$ full rank with probability one while still reflecting the fact that the positions of the sensors $\left\{(x_1,y_1),(x_2,y_2),\mathrm{K}\ ,(x_M,y_M)\right\}$ are much more precisely known than the measured TDOA values.

With this change in mind, we now define the new matrices $\bar{\mathbf{G}}_0 = \left[\mathbf{G}_0 \mid \mathbf{h}_0\right]$, $\bar{\mathbf{G}}_1 = \left[\mathbf{G}_1 \mid \mathbf{h}_1\right]$ and $\bar{\Delta}_1 = \left[\Delta\mathbf{G}_1 \mid \Delta\mathbf{h}_1\right]$. Note that Equation (2) can now be rewritten as

$$
\bar{\mathbf{G}}_0\begin{bmatrix}\mathbf{u}_0 \\ \hline -1\end{bmatrix} = 0,
$$

and Equation (3) can be rewritten as

---

[1] For discussions of both total-least-squares solutions and least-squares solutions, see for example [11].

$$\left(\bar{\mathbf{G}}_1 - \bar{\Delta}_1\right)\begin{bmatrix}\mathbf{u}_0\\ \hline -1\end{bmatrix} = 0\,.$$

Now, let $\Sigma_L = E\left\{\Delta_1^T\Delta_1\right\}$ and $\Sigma_R = E\left\{\Delta_1\Delta_1^T\right\}$. Then it is straightforward to show that

$$\Sigma_R \approx 8\sigma_\varepsilon^2\mathbf{I} + 4c^2\left[\mathbf{Q} + \mathbf{RQR} - \mathbf{QR} - \mathbf{RQ}\right] + 2c^4\mathbf{Q}\,o\mathbf{Q} + c^4\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\\ \mathsf{M}\\ \sigma_M^2\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \sigma_2^2 & \mathsf{L} & \sigma_M^2\end{bmatrix},$$

and

$$\Sigma_L \approx \begin{bmatrix} 4M\sigma_\varepsilon^2 & 0 & 0 & 0\\ 0 & 4M\sigma_\varepsilon^2 & 0 & 0\\ 0 & 0 & 4c^2\mathrm{Tr}\mathbf{Q} & -4c^2\mathrm{Tr}(\mathbf{RQ})\\ 0 & 0 & -4c^2\mathrm{Tr}(\mathbf{RQ}) & c^2\mathrm{Tr}\left(2\mathbf{RQR} + 3c^2\mathbf{Q}\,o\mathbf{Q}\right)\end{bmatrix}.$$

We seek a matrix $\bar{\Delta}$ and an estimate $\hat{\mathbf{u}}_1$ of $\mathbf{u}_0$ such that the equation

$$\left(\bar{\mathbf{G}}_1 - \bar{\Delta}\right)\begin{bmatrix}\hat{\mathbf{u}}_1\\ \hline -1\end{bmatrix} = 0 \tag{13}$$

is satisfied, and the matrix $\bar{\Delta}$ has minimum possible norm of the form

$$\left\|\bar{\Delta}\right\| = \sqrt{\mathrm{Tr}\left(\bar{\Delta}^T\Sigma_R^{-1}\bar{\Delta}\Sigma_L^{-1}\right)}.$$

To find $\hat{\mathbf{u}}_1$ and $\bar{\Delta}$, we first factor Σ_L and Σ_R using the Cholesky decomposition to get $\Sigma_L^{-1} = \mathbf{W}_L\mathbf{W}_L^T$ and $\Sigma_R^{-1} = \mathbf{W}_R\mathbf{W}_R^T$ and note that Equation (13) can be rewritten as

$$\mathbf{W}_R^T\left(\bar{\mathbf{G}}_1 - \bar{\Delta}\right)\mathbf{W}_L\mathbf{W}_L^{-1}\begin{bmatrix}\hat{\mathbf{u}}_1\\ \hline -1\end{bmatrix} = 0\,,$$

or equivalently

$$\left(\tilde{\mathbf{G}}_1 - \tilde{\Delta}\right)\tilde{\mathbf{u}} = 0\,, \tag{14}$$

where

$$\tilde{\mathbf{G}} = \mathbf{W}_R^T \bar{\mathbf{G}}_1 \mathbf{W}_L,$$
$$\tilde{\Delta} = \mathbf{W}_R^T \bar{\Delta} \mathbf{W}_L,$$
$$\tilde{\mathbf{u}} = \mathbf{W}_L^{-1} \begin{bmatrix} \hat{\mathbf{u}}_1 \\ -1 \end{bmatrix}.$$

Now to solve Equation (13), we look first for $\tilde{\Delta}$ and $\tilde{\mathbf{u}}$ that satisfy Equation (14) such that $\tilde{\Delta}$ has minimum possible *Froebinius norm*, which is given by

$$\left\| \tilde{\Delta} \right\|_F = \sqrt{\mathrm{Tr}\left(\tilde{\Delta}^T \tilde{\Delta} \right)}.$$

The desired solution for Equation (14), which is well known [11], is derived as follows. Let the singular value decomposition of $\tilde{\mathbf{G}}$ be given by

$$\tilde{\mathbf{G}} = \mathbf{U}\mathbf{D}\mathbf{V}^T,$$

where the columns of $\mathbf{V}$ are the orthonormal eigenvectors of $\tilde{\mathbf{G}}^T \tilde{\mathbf{G}}$. Let $\mathbf{v}_{\min}$ be the column of $\mathbf{V}$ corresponding to the smallest eigenvalue $\lambda_{\min}$ of $\tilde{\mathbf{G}}^T \tilde{\mathbf{G}}$. The desired solution to Equation (14) is given by

$$\tilde{\Delta} = \sqrt{\lambda_{\min}} \mathbf{v}_{\min} \mathbf{v}_{\min}^T, \quad \tilde{\mathbf{u}} = \alpha \mathbf{v}_{\min},$$

where α is chosen such that

$$\tilde{\mathbf{u}} = \alpha \mathbf{v}_{\min} = \mathbf{W}_L^{-1} \begin{bmatrix} \hat{\mathbf{u}}_1 \\ -1 \end{bmatrix}$$

or equivalently,

$$\begin{bmatrix} \hat{\mathbf{u}}_1 \\ -1 \end{bmatrix} = \alpha \mathbf{W}_L \mathbf{v}_{\min}.$$

Hence, α is the negative of the inverse of the last component of the vector $\mathbf{W}_L \mathbf{v}_{\min}$, and the vector $\hat{\mathbf{u}}_1$ is the desired Stage 1 solution to Equation (13). Stage 2 and Stage 3 of the new algorithm are identical to the original Stage 2 and Stage 3.

References

[1] "PulsON Technology Overview," http://www.timedomain.com/Files/downloads/techpapers/PulsONOverview7_01.pdf.

[2] J. C. Adams, W. Gregorwich, L. Capots, and D. Liccardo, "Ultra-Wideband for Navigation and Communications," *Proceedings of IEEE Aerospace Conference*, 2001.

[3] C. W. Borst, "Telerobotic Ground Control of a Space Free Flyer," *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1998.

[4] Y. T. Chan and K. C. Ho, "An Efficient Closed-Form Localization Solution From Time Difference of Arrival Measurements," 1994.

[5] H. Choset, et al., "Path Planning and Control for AERCam, a Free-Flying Inspection Robot in Space," *Proceedings of IEEE International Conference on Robotics and Automation*, 1999.

[6] FCC, "First Notice and Order: Revision of Part 15 of the Commission's Rules Regarding Ultra-Wideband Transmission Systems," Feb. 2002.

[7] M. G. M. Hussain, "Principles of Space-Time Array Processing for Ultrawide Band Impulse Radar and Radio Communications," *IEEE Transactions on Vehicular Technology*, vol. 51, pp. 393-403, 2002.

[8] D. Kortenkamp, et al., "Applying a Layered Control Architecture to a Free-Flying Space Camera," *Proceedings of IEEE International Joint Symposium on Intelligence and Systems*, 1998.

[9] Y. C. Loh, et al., "Wireless Video System for Extra Vehicular Activity in the International Space Station and Space Shuttle Orbiter Environment," *Proceedings of IEEE 49th Vehicular Technology Conference*, 1999.

[10] K. Siwiak, et al., "Advances in Ultra-Wide Band Technology," *Proceedings of Radio Solutions 2001*, Commonwealth Conference & Events Centre, London, 2001.

[11] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

Appendix

In this Appendix, we derive the bias and mean-squared-error (MSE) for the current weighted-least-squares TDOA algorithm for a simple far-field example. In particular, we let $M = 4$, $\mathbf{Q} = \sigma^2\mathbf{I}$, and $r_0 >> \max_{1\le i\le 4}\left(\max\left\{|x_i|,|y_i|,cd_i\right\}\right)$. Then, from Equation (11), we get

$$\mu_3 = c^2\sigma^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{B}_1^{-2}\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}$$

$$\cdot\left[\begin{array}{l}\mathbf{G}_0^T\mathbf{B}_1^{-2}\left(\begin{bmatrix}1\\1\\1\\1\end{bmatrix}+4\mathbf{B}_1^{-1}\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{B}_1^{-2}\mathbf{G}_0\right)^{-1}\begin{bmatrix}0\\0\\1\end{bmatrix}\right)\\ \quad -4\mathrm{Tr}\left(\mathbf{B}_1^{-2}\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{B}_1^{-2}\mathbf{G}_0\right)^{-1}\mathbf{G}_0^T\mathbf{B}_1^{-1}-\mathbf{B}_1^{-1}\right)\begin{bmatrix}0\\0\\1\end{bmatrix}\end{array}\right]$$

$$\approx c^2\sigma^2r_0^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\left[\begin{array}{l}\frac{1}{r_0^2}\mathbf{G}_0^T\left(\begin{bmatrix}1\\1\\1\\1\end{bmatrix}+4r_0\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{G}_0\right)^{-1}\begin{bmatrix}0\\0\\1\end{bmatrix}\right)\\ \quad -\frac{4}{r_0}\mathrm{Tr}\left(\mathbf{G}_0\left(\mathbf{G}_0^T\mathbf{G}_0\right)^{-1}\mathbf{G}_0^T-\mathbf{I}\right)\begin{bmatrix}0\\0\\1\end{bmatrix}\end{array}\right]$$

$$= c^2\sigma^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\left(\mathbf{G}_0^T\begin{bmatrix}1\\1\\1\\1\end{bmatrix}+4r_0\begin{bmatrix}0\\0\\1\end{bmatrix}-4r_0(3-4)\begin{bmatrix}0\\0\\1\end{bmatrix}\right)$$

$$= 8c^2\sigma^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\left(\begin{bmatrix}\sum_{i=1}^4 x_i & \sum_{i=1}^4 y_i & \sum_{i=1}^4 cd_i\end{bmatrix}^T+8r_0\begin{bmatrix}0\\0\\1\end{bmatrix}\right)$$

$$\approx 8c^2\sigma^2r_0\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\begin{bmatrix}0\\0\\1\end{bmatrix}.$$

Continuing to simplify this, we get

$$\mathbf{B}_2^{-1} = \frac{1}{r_0 x_0 y_0}\begin{bmatrix} r_0 y_0 & 0 & 0 \\ 0 & r_0 x_0 & 0 \\ 0 & 0 & x_0 y_0 \end{bmatrix},$$

$$\mathbf{B}_2^{-1}\mathbf{G}_0^T = \frac{1}{r_0 x_0 y_0}\begin{bmatrix} r_0 y_0 x_1 & r_0 y_0 x_2 & r_0 y_0 x_3 & r_0 y_0 x_3 \\ r_0 x_0 y_1 & r_0 x_0 y_2 & r_0 x_0 y_3 & r_0 x_0 y_4 \\ x_0 y_0 cd_1 & x_0 y_0 cd_2 & x_0 y_0 cd_3 & x_0 y_0 cd_4 \end{bmatrix},$$

$$\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T = \frac{1}{r_0 x_0 y_0}$$
$$\cdot\begin{bmatrix} y_0(r_0 x_1 + x_0 cd_1) & y_0(r_0 x_2 + x_0 cd_2) & y_0(r_0 x_3 + x_0 cd_3) & y_0(r_0 x_4 + x_0 cd_4) \\ x_0(r_0 y_1 + y_0 cd_1) & x_0(r_0 y_2 + y_0 cd_2) & x_0(r_0 y_3 + y_0 cd_3) & x_0(r_0 y_4 + y_0 cd_4) \end{bmatrix},$$

$$\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2 = \frac{1}{r_0^2 x_0^2 y_0^2}$$
$$\cdot\begin{bmatrix} y_0^2\sum_{i=1}^{4}(r_0 x_i + x_0 cd_i)^2 & x_0 y_0\sum_{i=1}^{4}(r_0 x_i + x_0 cd_i)(r_0 y_i + y_0 cd_i) \\ x_0 y_0\sum_{i=1}^{4}(r_0 x_i + x_0 cd_i)(r_0 y_i + y_0 cd_i) & x_0^2\sum_{i=1}^{4}(r_0 y_i + y_0 cd_i)^2 \end{bmatrix},$$

$$\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}$$
$$= \frac{r_0^2}{\sum_{i=1}^{4}\left(x_i + \frac{x_0}{r_0}cd_i\right)^2\sum_{i=1}^{4}\left(y_i + \frac{y_0}{r_0}cd_i\right)^2 - \left[\sum_{i=1}^{4}\left(x_i + \frac{x_0}{r_0}cd_i\right)\left(y_i + \frac{y_0}{r_0}cd_i\right)\right]^2}$$
$$\cdot\begin{bmatrix} \frac{x_0^2}{r_0^2}\sum_{i=1}^{4}\left(y_i + \frac{y_0}{r_0^2}cd_i\right)^2 & -\frac{x_0 y_0}{r_0^2}\sum_{i=1}^{4}(r_0 x_i + x_0 cd_i)(r_0 y_i + y_0 cd_i) \\ -\frac{x_0 y_0}{r_0^2}\sum_{i=1}^{4}(r_0 x_i + x_0 cd_i)(r_0 y_i + y_0 cd_i) & \frac{y_0^2}{r_0^2}\sum_{i=1}^{4}\left(x_i + \frac{x_0}{r_0}cd_i\right)^2 \end{bmatrix},$$

$$\mathbf{G}_2^T\mathbf{B}_2^{-1}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \frac{1}{r_0}\begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{G}_2^T\mathbf{B}_2^{-1}\begin{bmatrix}0\\0\\1\end{bmatrix}$$

$$= \frac{1}{\sum_{i=1}^{4}\left(x_i+\frac{x_0}{r_0}cd_i\right)^2\sum_{i=1}^{4}\left(y_i+\frac{y_0}{r_0}cd_i\right)^2-\left[\sum_{i=1}^{4}\left(x_i+\frac{x_0}{r_0}cd_i\right)\left(y_i+\frac{y_0}{r_0}cd_i\right)\right]^2}$$

$$\cdot\begin{bmatrix}\frac{x_0}{r_0}\sum_{i=1}^{4}\left(y_i+\frac{y_0}{r_0^2}cd_i\right)^2-\frac{y_0}{r_0}\sum_{i=1}^{4}\left(r_0x_i+x_0cd_i\right)\left(r_0y_i+y_0cd_i\right)\\ \frac{y_0}{r_0}\sum_{i=1}^{4}\left(x_i+\frac{x_0}{r_0}cd_i\right)^2-\frac{x_0}{r_0}\sum_{i=1}^{4}\left(r_0x_i+x_0cd_i\right)\left(r_0y_i+y_0cd_i\right)\end{bmatrix}.$$

Now, let $x_0 = r_0\cos\theta$, $y_0 = r_0\sin\theta$,

$$\tilde{\mathbf{x}} = \begin{bmatrix}x_1+cd_1\cos\theta & x_2+cd_2\cos\theta & x_3+cd_3\cos\theta & x_4+cd_4\cos\theta\end{bmatrix}^T,$$

$$\tilde{\mathbf{y}} = \begin{bmatrix}y_1+cd_1\sin\theta & y_2+cd_2\sin\theta & y_3+cd_3\sin\theta & y_4+cd_4\sin\theta\end{bmatrix}^T,$$

$$\langle\tilde{\mathbf{x}}_0,\tilde{\mathbf{y}}_0\rangle = \|\tilde{\mathbf{x}}_0\|\|\tilde{\mathbf{y}}_0\|\cos\phi.$$

Then, putting this all together, it follows that the bias vector for this example is given by

$$\mu_3 \approx \frac{8c^2\sigma^2r_0}{\|\tilde{\mathbf{x}}_0\|^2\|\tilde{\mathbf{y}}_0\|^2-|\langle\tilde{\mathbf{x}}_0,\tilde{\mathbf{y}}_0\rangle|^2}\begin{bmatrix}\|\tilde{\mathbf{y}}_0\|^2\cos\theta-\langle\tilde{\mathbf{x}}_0,\tilde{\mathbf{y}}_0\rangle\sin\theta\\ \|\tilde{\mathbf{x}}_0\|^2\sin\theta-\langle\tilde{\mathbf{x}}_0,\tilde{\mathbf{y}}_0\rangle\cos\theta\end{bmatrix}$$

$$= \frac{8c^2\sigma^2r_0\left(\|\tilde{\mathbf{x}}_0\|^2+\|\tilde{\mathbf{y}}_0\|^2\right)}{\|\tilde{\mathbf{x}}_0\|^2\|\tilde{\mathbf{y}}_0\|^2\left(1-\cos^2\phi\right)}\begin{bmatrix}\frac{\|\tilde{\mathbf{y}}_0\|^2\cos\theta-\langle\tilde{\mathbf{x}}_0,\tilde{\mathbf{y}}_0\rangle\sin\theta}{\|\tilde{\mathbf{x}}_0\|^2+\|\tilde{\mathbf{y}}_0\|^2}\\ \frac{\|\tilde{\mathbf{x}}_0\|^2\sin\theta-\langle\tilde{\mathbf{x}}_0,\tilde{\mathbf{y}}_0\rangle\cos\theta}{\|\tilde{\mathbf{x}}_0\|^2+\|\tilde{\mathbf{y}}_0\|^2}\end{bmatrix}.$$

Similarly, Equation (12) gives

$$\Sigma_3 = 4c^2\sigma^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{B}_1^{-2}\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{B}_3^{-1}$$

$$\approx 4c^2\sigma^2 r_0^2\mathbf{B}_3^{-1}\left(\mathbf{G}_2^T\mathbf{B}_2^{-1}\mathbf{G}_0^T\mathbf{G}_0\mathbf{B}_2^{-1}\mathbf{G}_2\right)^{-1}\mathbf{B}_3^{-1}$$

$$= \frac{4c^2\sigma^2 r_0^2}{\sum_{i=1}^{4}\left(x_i+\frac{x_0}{r_0}cd_i\right)^2\sum_{i=1}^{4}\left(y_i+\frac{y_0}{r_0}cd_i\right)^2-\left[\sum_{i=1}^{4}\left(x_i+\frac{x_0}{r_0}cd_i\right)\left(y_i+\frac{y_0}{r_0}cd_i\right)\right]^2}$$

$$\cdot\begin{bmatrix}\sum_{i=1}^{4}\left(y_i+\frac{y_0}{r_0^2}cd_i\right)^2 & -\sum_{i=1}^{4}\left(r_0x_i+x_0cd_i\right)\left(r_0y_i+y_0cd_i\right)\\ -\sum_{i=1}^{4}\left(r_0x_i+x_0cd_i\right)\left(r_0y_i+y_0cd_i\right) & \sum_{i=1}^{4}\left(x_i+\frac{x_0}{r_0}cd_i\right)^2\end{bmatrix}$$

$$= \frac{4c^2\sigma^2 r_0^2\left(\|\mathbf{x}_\phi\|^2+\|\mathbf{y}_\phi\|^2\right)}{\|\mathbf{x}_\phi\|^2\|\mathbf{y}_\phi\|^2\left(1-\cos^2\phi\right)}\begin{bmatrix}\frac{\|\mathbf{y}_\phi\|^2}{\|\mathbf{x}_\phi\|^2+\|\mathbf{y}_\phi\|^2} & -\frac{\langle\mathbf{x}_\rho,\mathbf{y}_\rho\rangle\cos\phi}{\|\mathbf{x}_\phi\|^2+\|\mathbf{y}_\phi\|^2}\\ -\frac{\langle\mathbf{x}_\rho,\mathbf{y}_\rho\rangle\cos\phi}{\|\mathbf{x}_\phi\|^2+\|\mathbf{y}_\phi\|^2} & \frac{\|\mathbf{x}_\phi\|^2}{\|\mathbf{x}_\phi\|^2+\|\mathbf{y}_\phi\|^2}\end{bmatrix}.$$

Hence, the total MSE is given by

$$\text{MSE} = \mathbf{Tr}\left(\Sigma_3\right) = \frac{4c^2\sigma^2 r_0^2\left(\|\mathbf{x}_\phi\|^2+\|\mathbf{y}_\phi\|^2\right)}{\|\mathbf{x}_\phi\|^2\|\mathbf{y}_\phi\|^2\left(1-\cos^2\phi\right)},$$

which completes the example.

Studies of Carbon Nanotubes

Final Report
NASA Faculty Research Program - 2004
Johnson Space Center

Prepared by: Gerard T. Caneba, Ph.D.

Academic Rank:
Associate Professor

University & Department:
Michigan Technological University
Department of Chemical Engineering
1400 Townsend Drive
Houghton, MI 49931

NASA/JSC

Directorate:	Engineering
Division:	Structural Engineering
Branch:	Materials and Processes
JSC Colleague:	Brad S. Files
Date Submitted:	August, 2004
Contract Number:	NAG 9-1526 and NNJ04F93A

ABSTRACT

The fellowship experience for this summer for 2004 pertains to carbon nanotube coatings for various space-related applications. They involve the following projects: (a) EMI protection films from HiPco-polymers, and (b) Thermal protection nanosilica materials.

EMI protection films are targeted to be eventually applied onto casings of laptop computers. These coatings are composites of electrically-conductive SWNTs and compatible polymers. The substrate polymer will be polycarbonate, since computer housings are typically made of carbon composites of this type of polymer. A new experimental copolymer was used last year to generate electrically-conductive and thermal films with HiPco at 50/50 wt/wt composition. This will be one of the possible formulations. Reference films will be base polycarbonate and neat HiPco onto polycarbonate films. Other coating materials that will be tried will be based on HiPco composites with commercial enamels (polyurethane, acrylic, polyester), which could be compatible with the polycarbonate substrate.

Nanosilica fibers are planned for possible use as thermal protection tiles on the shuttle orbiter. Right now, microscale silica is used. Going to the nanoscale will increase the surface-volume-per-unit-area of radiative heat dissipation. Nanoscale carbon fibers/nanotubes can be used as templates for the generation of nanosilica. A sol-gel operation is employed for this purpose.

INTRODUCTION

Carbon nanotubes are of interest within NASA, as lightweight materials with enhanced mechanical, thermal, and electrical properties. For example, single carbon nanotube fibers have been shown to be stronger per weight compared to stainless steel (up to 100 times stronger) and Kevlar (14 times stronger). Also, they can posses electrical and thermal conductivities better than Copper. Within a polymeric matrix, thermal diffusivities can be at least 30 times that of the neat polymer.

EMI protection is normally associated with electrically conductive materials. Figure 1 below shows some of the previous results of comparisons of surface resistivities from various polymeric composites.[1]

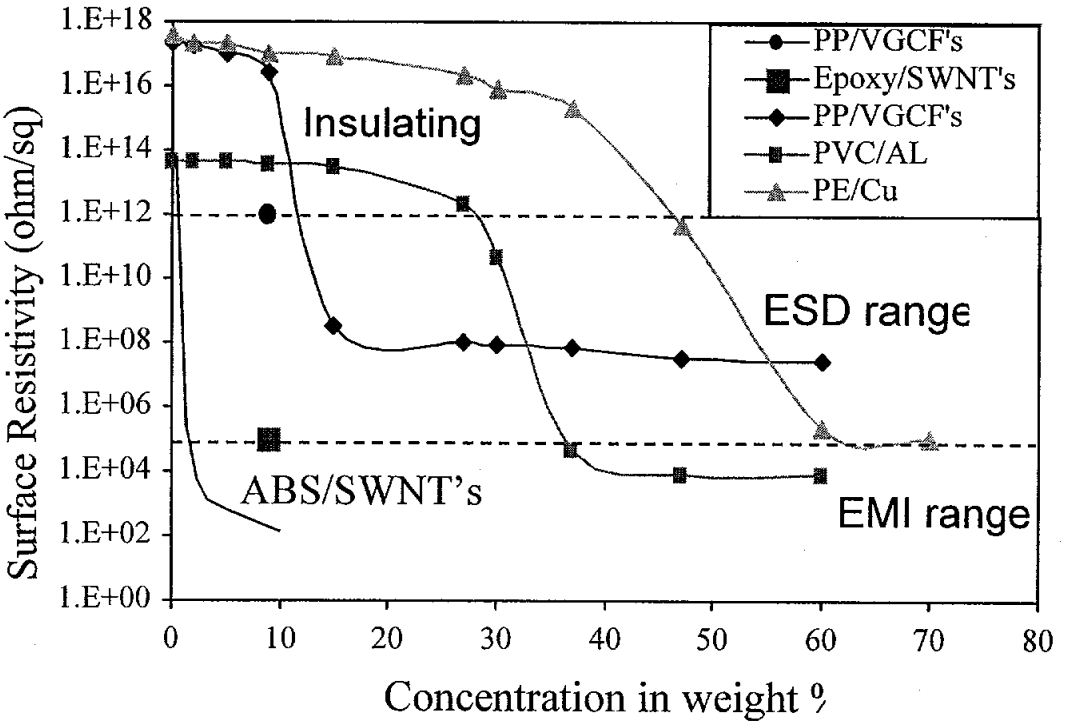


Figure 1: Surface resistivities of various polymeric composites.[1] Various nonconducting polymers are polypropylene (PP), epoxy, poly(vinyl chloride) (PVC), polyethylene (PE), and acrylonitrile-butadiene-styrene terpolymer (ABS). Conducting material fillers are vapor-grown carbon fibers (VGCF), single-walled carbon nanotubes (SWNT), Aluminum (AL), and Copper (Cu).

Based on Figure 1, EMI protection can be obtained if surface resistivities are below 10^5 Ohm/sq. This type of performance is obtained at much lower weight loadings for SWNT than for Copper and Aluminum fillers. This is not surprising, since single

SWNT fibers have similar electrical conductivities as metals at 1/5th the density. A possible stumbling block is that the Avionics has specified surface resistivities <50 mOhms/sq of 2.5-3 μm thick films. This is equivalent to a volume resistivities <0.125-0.15 μOhm-m or electrical conductivities >6.7-8x10^{6} (Ohm-m)$^{-1}$. Such electrical conductivity values are just an order of magnitude less than those of metals. For example, Silver, Copper, Gold, and Aluminum have electrical conductivities of 6.8, 6.0, 4.3, and 3.8x10^{7} (Ohm-m)$^{-1}$, respectively.[2] Even though a single crystal SWNT fiber could have a calculated electrical conductivity of 10^{8} (Ohm-m)$^{-1}$,[3] it will be a challenge to attain the minimum required value of 6.7-8x10^{6} (Ohm-m)$^{-1}$. The fellow believes that the required value was based on metallic systems, which are heavier than SWNT-based films. In the end, a balance between performance and weight would have to be determined.

Surface resistance (R_S) and surface resistivity (ρ_S) are obtained from the basic setup in Figure 2 below.[4]

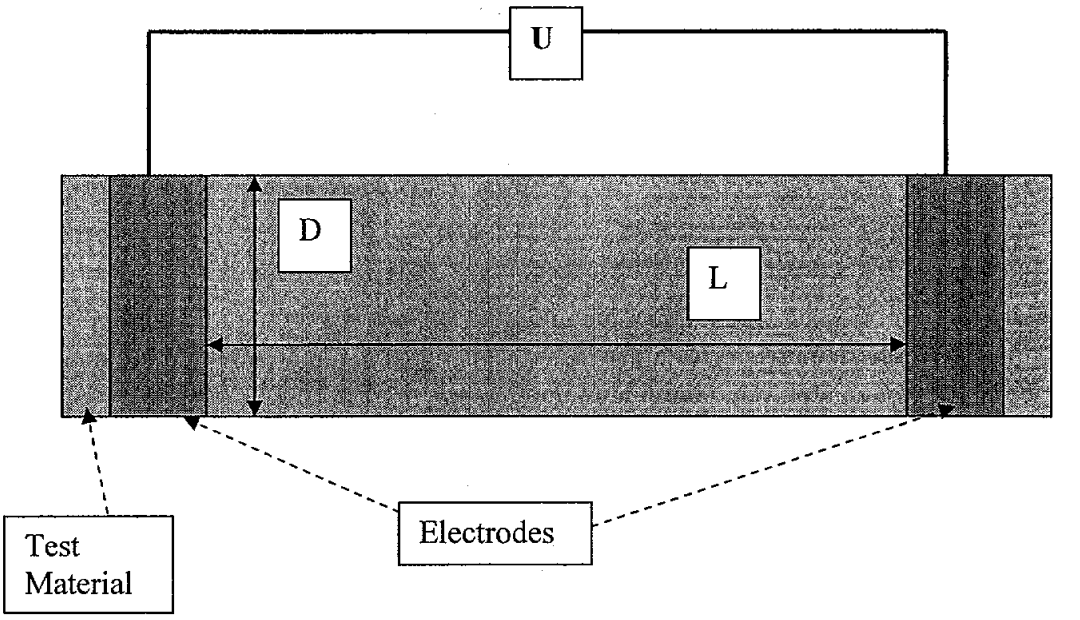


Figure 2: Top view of the basic setup for surface resistance (R_S) and surface resistivity (ρ_S) measurements.

The surface resistance is defined as the DC voltage (U) divided by the current (I_S) flowing between the two electrodes in contact with the surface of the test material (Figure 2).

$$R_S = \frac{U}{I_S} \qquad (1).$$

The surface resistivity, ρ_S, is determined by the ratio of the DC voltage drop (U) per unit length (L) to the surface current (IS) per unit width (D).

(2)

The surface resistivity is an inherent property of the material, and should remain the same regardless of the method and configuration of the electrodes. The surface resistance is specific to the setup and method of measurement. Based on Eqs. 1-2, both quantities should have dimensions of Ohms. To make the distinction, the surface resistivity is nominaly given the units of Ohms/sq instead.

To convert surface resistivity (ρ_S) to bulk resistivity (ρ)

$$\rho = (\delta_S)(\rho_S) \qquad (3)$$

where δ_S is the depth of the surface layer. Thus, the bulk resistivity would have units of Ohm-cm or Ohm-m. Finally, the conductivity (S) is defined as the reciprocal of the bulk resistivity, or

$$S = \frac{1}{\rho} \qquad (4).$$

The fellow will assist in the development of applications that involve polymeric coatings onto SWNTs. A particular focus area of application pertains to films for lightweight electrostatic and electromagnetic shielding. We will look into various organic polymer coatings that will be compatible with electrically conductive single-walled carbon nanotubes SWCNTs and various substrates within laptop computers used in the International Space Station (ISS). SWCNT/polymer films will be produced that could be applied onto an appropriate computer casing material. Working with engineers in the Avionics Division, electrical conductivity and static discharge properties will be obtained from these films in the future.

Another area of work is the use of carbon nanotubes for the development of thermal protection materials. This involves the application of silica coatings onto multi-walled CNTs. Silica-coated CNTs will be processed to remove the carbon core, in order to produce nanoscale silica tubes that could be used in the next generation of thermal protection tiles. Current thermal tiles are fused microscale silica rods. Since 90% of heat dissipation during the shuttle reentry is by radiative heat transfer, increasing the surface

area per volume of the tile material could be beneficial. Thus, there is interest in the development of silica nanotubes for this purpose.

EXPERIMENTAL

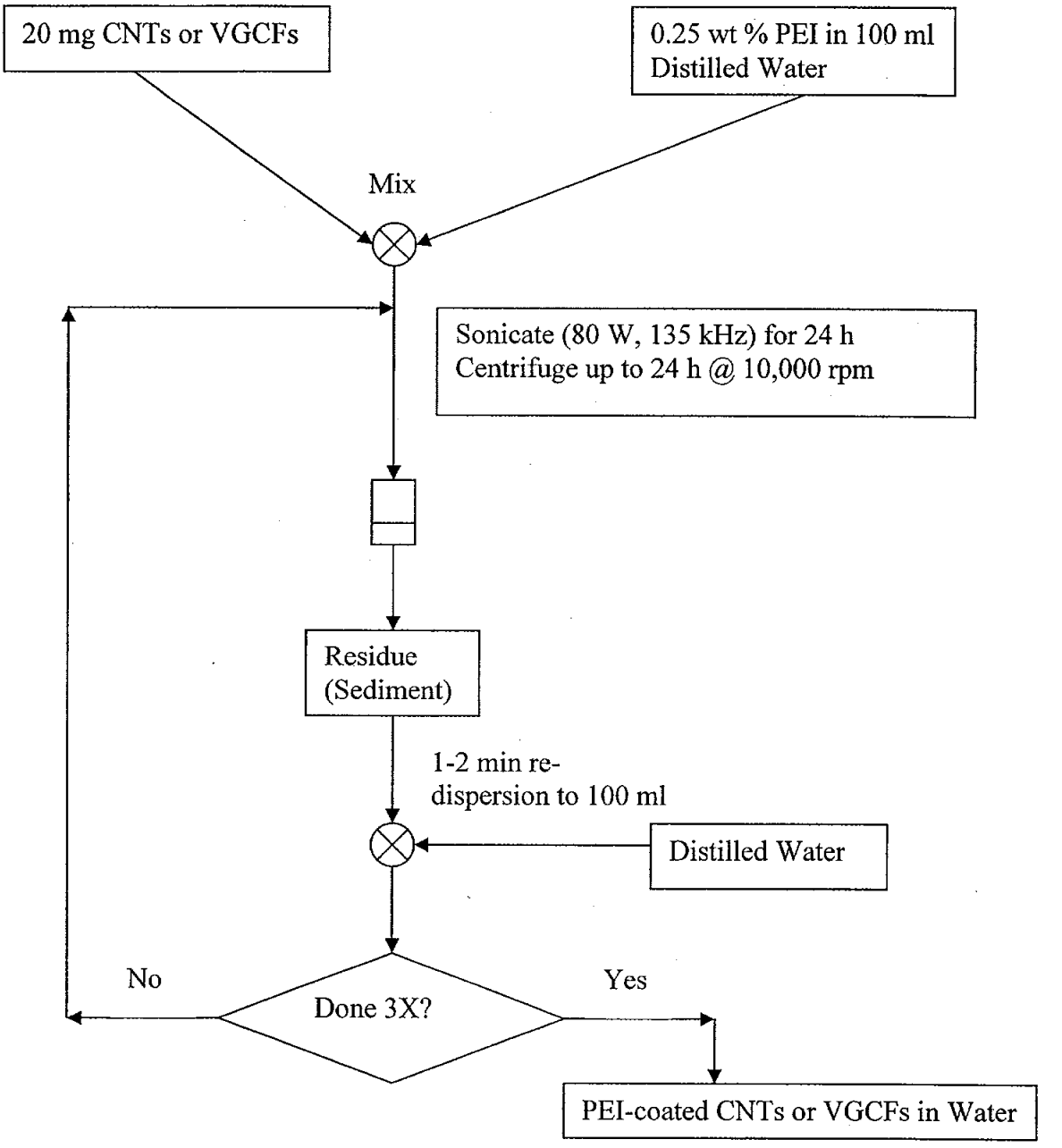
EMI Protection Films:

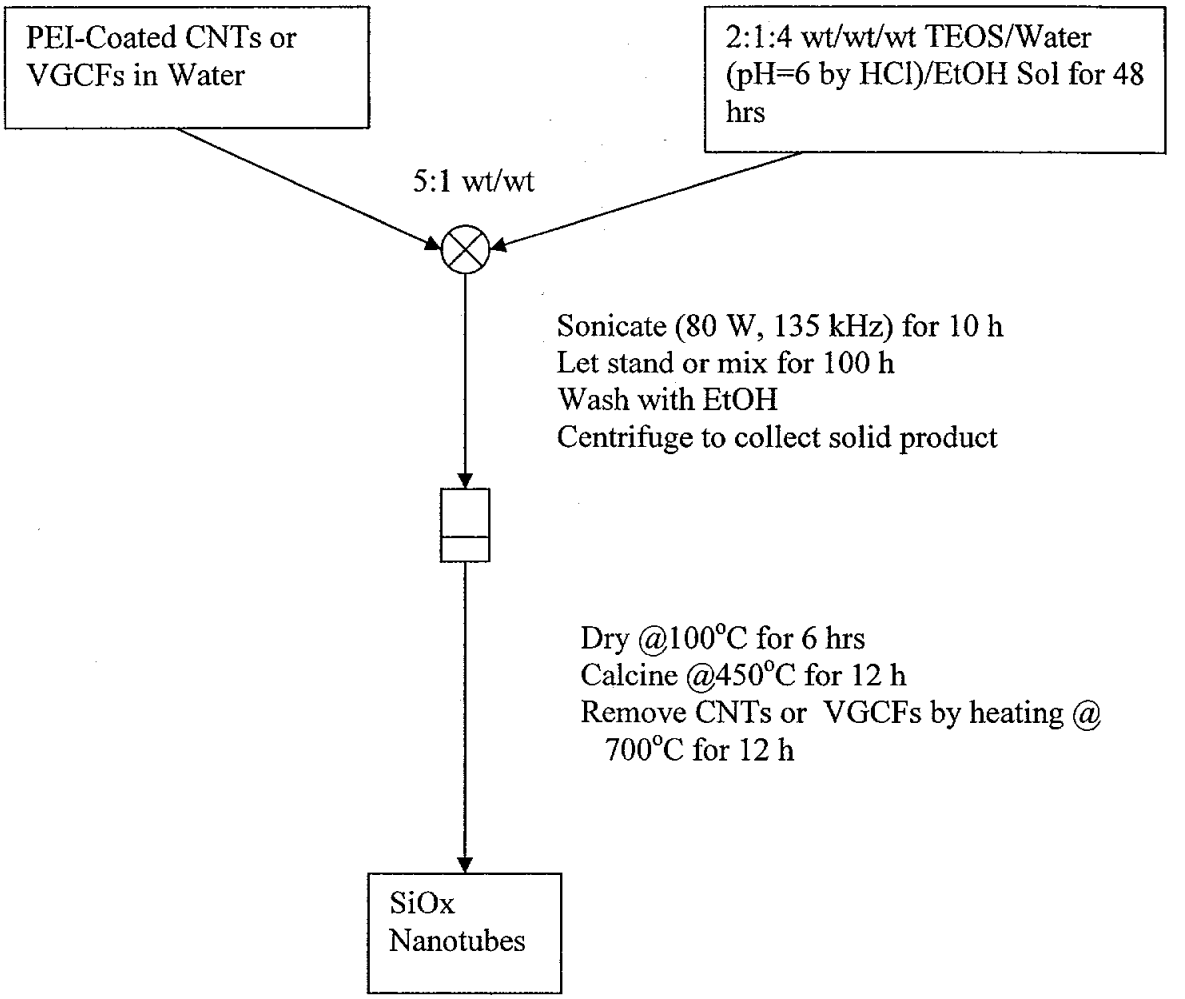
Since laptop computer cases have been found to comprise of polycarbonate composites, Lexan™ sheets have been envisioned as a good reference material. Base films will be purified HiPco and Laser nanotubes that will be cast onto Lexan™ sheets with DMF. The SWNT will be dispersed with the aid of an ultrasonic bath. Films will be cast using a handheld spray coating apparatus.

Other binders that will be investigated include: an experimental VA-t-AA copolymer with 6 wt % AA content[5] and commercial enamels, such as polyurethane, acrylics, polyester, and even epoxies. Film loading will be determined based on the amount of solid used per spray area. Finally, surface resistivity values will be obtained using the concentric cylinder apparatus at the RITF laboratory as well as a home-made version of Figure 2. Four measurements will be done at the sides of the four samples, and an average will be obtained.

Thermal Protection Nanosilica:

The procedure for the generation of silica nanofibers that can possibly be used as thermal protection materials is shown below.[6,7]





The mechanism of conventional sol-gel silica formation process is shown below. We note that there is a need for basic VGCF or CNT surface to form the silica via nucleation and growth.

- Hydrolysis

  $Si(OR)_4 + nH_2O = Si(OR)_m(OH)_n + nROH$

  Tetraethylorthosilicate or TEOS

  where n+m=4

  Also,

  $Si(OR)_m(OH)_n = Si(OR)_m(OH)_{n-k}O^{k-} + kH^+$

  where m=0 in a basic environment

  $[pK_k] = [9.8 \;\; 12.4 \;\; 15 \;\; 17.6]$

- **Condensation**
  (1) SiOH + SiOH --→ SiOSi + H_2O
  (2) SiOH + SiOR --→ SiOSi + ROH
  Condensation Rxn (1) is favored when $\alpha >> 2$
  Condensation Rxn (2) is favored when $\alpha << 2$
  α = H_2O/Si feed (mole/mole)

- **Thus, a basic substrate will favor silica adsorption when $\alpha >> 2$**

RESULTS AND DISCUSSION

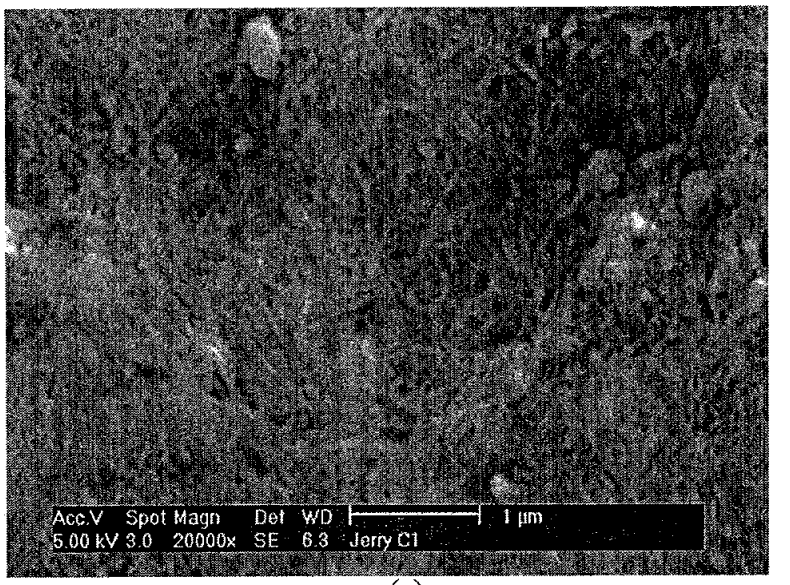
EMI Protections Films:

Since DMF is the solvent of choice for the dispersion of CNTs, its compatibility with proposed additives for coatings formation was tested first. This was done by attempting to dissolve a small amount of the additives in DMF. Complete dissolution is needed; otherwise downstream processing will not be successful.

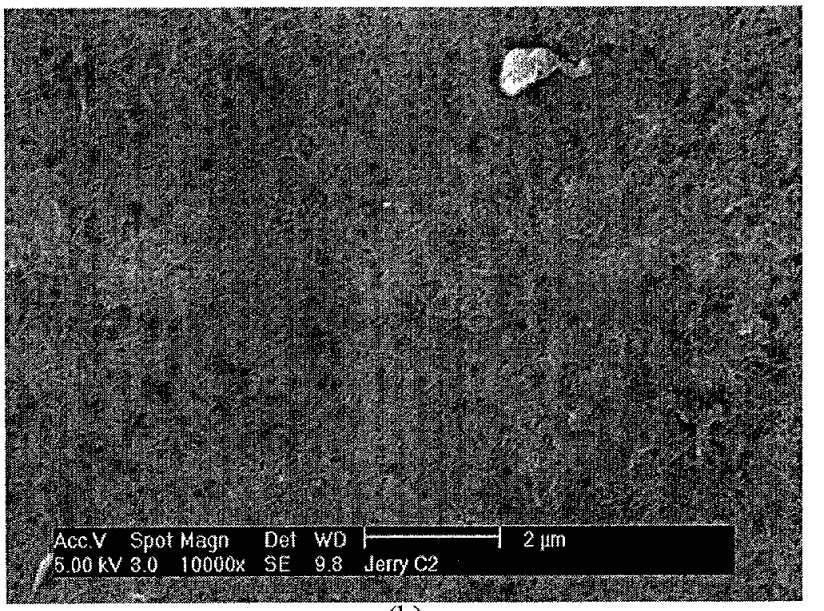
Results indicate that only the VA-t-AA copolymer would dissolve in DMF completely. The polyurethane enamel from Minwax did not dissolve at all. The acrylic material from Rust-Oleum did not completely dissolve either. Lastly, the polyester material formed a jelly fluid structure with DMF even without the presence of the catalyst. It is believed that the amine group in DMF acted as a catalyst. Therefore, these preliminary tests narrowed down our coating systems to SWNT/DMF and SWNT/VA-t-AA/DMF mixtures.

Another test was done on how DMF will dry onto a polycarbonate surface. This was done by applying a couple of drops of DMF onto a Lexan™ sheet and then leaving the DMF to dry in a fume hood. After a day of drying, area on the Lexan™ sheet that used to contain the DMF turned white, but the sheet did not deform. This means that the DMF partially swelled the Lexan™ surface and extracted low molecular weight moeties toward the surface. This might be a good thing, especially for the coating formulation that contains SWNT and DMF only. Here the SWNT would be deposited in the subsurface region of the Lexan™ sheet.

Mixtures containing 0.05-0.1 wt % SWNT (HiPco and Laser SWNT) were dispersed and sonicated; some of them contained various amounts of the VA-t-AA copolymer. The, the dispersed mixtures were sprayed onto the 4.5"x4.5"x0.093" Lexan™ sheets one layer at a time. It took at least one hour to dry a thin layer of the dispersion before the next spray. After drying the last spray, the coated Lexan™ coupon was vacuum dried at 60°C for 2 hrs. In Figure 3 below, SEM pictures of surfaces are shown.



(a)



(b)

Figure 3: SEM pictures of surfaces of HiPco onto Lexan™ at HiPco loadings of (a) 0.12 mg/cm^2 and (b) 0.020 mg/cm^2.

Based on Figures 3(a) and 3(b), SWNTs were clearly seen on the Lexan™ surfaces when they were spray-coated from SWNT/DMF dispersions. As the SWNT loading increased,

the more the SWNT bundles are exposed on the surface. It seems that the swelling of the Lexan™ can accommodate only a finite layer of the SWNT. With the addition of a polymeric additive (VA-t-AA Copolymer), the SWNT bundles seem to become more and more embedded in the polymer as the VA-t-AA Copolymer proportion increases relative to the SWNT (Figure 4).

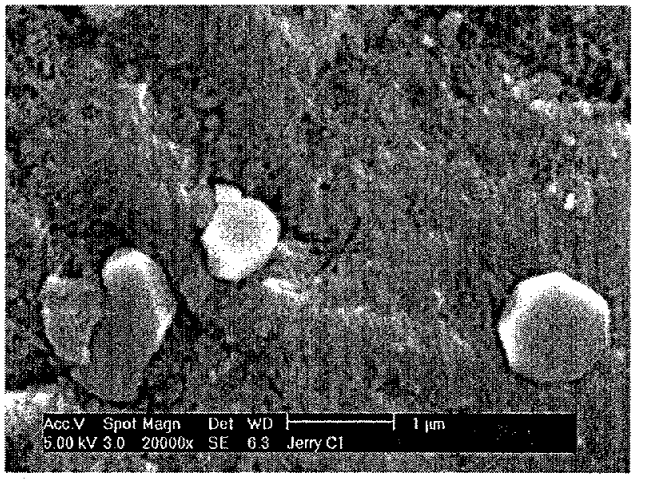


Figure 4: SEM of the surface of 91/9 VA-t-AA/HiPco on Lexan™ at a HiPco loading of 0.18 mg/cm^2.

Surface resistivities of four coated as well as the Lexan™ reference are shown in Table 1 below:

TABLE 1: Surface resistivities of costings onto Lexan™

Sample, Coating	Surface Resistivity based on Method in Figure 2, k-Ohms/sq		Surface Resistivity based on Concentric Circle Method, k-Ohms/sq	
	Readings	Average	Readings	Average
Reference - Lexan™	∞	∞	∞	∞
Pure HiPco @ 0.12 mg HiPco/cm^2	1.7,0.8,0.9,1.1	1.1	2.0,1.8,1.6,2.3	1.9
Pure HiPco @ 0.020 mg HiPco/cm^2	2.0,2.0,1.6,1.9	1.9	5.9,4.3,5.4,6.1	5.4
50/50 VA-t-AA/HiPco @ 0.10 mg HiPco/cm^2	3.3,2.3,1.8,2.8	2.6	3.5,5.6,4.5,2.8	4.1

91/9 VA-t-AA/HiPco @ 0.18 mg HiPco/cm^2	13,10,7,13	11	15,13,11,18	14

Readings based on the apparatus in Figure 2 are lower than those based on the Concentric Circle Method. The former is more reliable than the latter because test coupons are not completely flat. In general, more HiPco loading onto Lexan™ resulted in lower surface resistivities at the conductive range. For the VA-t-AA/HiPco composite onto Lexan™, surface resistivities depend on the HiPco loading and HiPco proportion in the surface composite layer. Even at a higher HiPco loading, surface resistivites can be relatively low at low HiPco proportion in the VA-t-AA/HiPco composite layer.

In order to determine the level of adhesion of the coatings onto the substrate, cross-cuts from a knife are made onto the coating surface in such a way that there are 25 1/8-inch-squares of coating material. Then, a Scotch™ tape is applied onto the cross-cuts with moderate pressure to ensure contact between the tape and the squares cut from the coating. Then, the tape is removed quickly. The equivalent number of squares removed determines the level of adhesion. In the samples indicated in Table 1, the pure HiPco coating @ 0.12 mg/cm^2 showed excellent adhesion at <1 Equivalent Squares removed. This is followed by the 91/9 VA-t-AA/HiPco coating with 16-20 Equivalent Squares removed. The rest of the coatings indicated 21-25 Equivalent Squares removed. Note that these adhesion measurements are done at the best regions of each of the coatings. With better spraying equipment, they would apply to the entire coatings.

Based on the above findings, it is better to use apply pure HiPco material onto Lexan™ using an appropriate solvent, such as DMF. In this case, DMF seems to be a slight swelling agent to Lexan™, which promotes some anchoring of SWNT onto the polymer surface. This work is preliminary, and more exhaustive studies are needed using better spraying equipment. Then the coatings have to be subjected to mechanical vibrational studies, to determine how performance will be affected.

In order to obtain better conductivity of SWNT-based coatings, the SWNT bundles and fibers can be aligned rheologically or by other means while the coating layers are drying. Also, metal nanoparticles can be added to the formulation, presumably to establish better connectivity between ends of SWNT fibers.

Thermal Protection Nanosilica:

The following carbon-containing starting mixtures were prepared for overnight sonication:
Mixture #1 – 20 mg VGCF, 250 mg 1.2 KDaltons PEI, 100 g DI Water
Mixture #2 – 20 mg VGCF, 250 mg 1.2 KDaltons PEI, 100 g DI Water

Mixture #3 – 20 mg MWNT#1, 250 mg 1.2 KDaltons PEI, 100 g DI Water
Mixture #4 – 20 mg MWNT#1, 250 mg 1.2 KDaltons PEI, 100 g DI Water
Mixture #5 – 20 mg VGCF, 250 mg 10 KDaltons PEI, 100 g DI Water
Mixture #6 – 20 mg VGCF, 250 mg 10 KDaltons PEI, 100 g DI Water
Mixture #7 – 20 mg MWNT#1, 250 mg 10 KDaltons PEI, 100 g DI Water
Mixture #8 – 20 mg MWNT#1, 250 mg 10 KDaltons PEI, 100 g DI Water

After overnight sonication, only Mixtures #7 and #8 were reasonably dispersed. The rest settled completely to the bottom of the containers when allowed to stand for at least a day. Dry residues of the bottom layers of Mixtures #7 and #8 are 27 and 22 mg, respectively. Since the residues contain both CNTs and PEI, relative amounts of the residues are 10 and 8 wt %, respectively. Thus, continuation of the above-mentioned procedure was done with Mixtures #7 and #8.

Mixtures #7 and #8 were centrifuged starting at 1,600 RPM but found it to work at 10,000 RPM for 1 hr. Supernatants were removed and their pH values were measured to be equal to 9. Then, 100 ml DI Water was added to each residue, and the mixtures were sonicated again overnight. The, it took 13 hrs of centrifugation at 10,000 RPM to reasonably settle the residue. Even with careful removal of supernatants, only 60-65 ml were removed. This means that the MWNTs were well dispersed. Both supernatants registered a pH of 6. Again, DI Water was added to approximate total volumes to 100 ml. After overnight sonication, Mixture #8 was centrifuged at 10,000 RPM for 16 hrs. Then, 70 ml of the supernatant was decanted off with a pH of 5.3. Finally, DI Water was added to both mixtures to bring them 10 100 ml each, and then sonicated overnight.

The following TEOS-containing mixture was prepared and stirred for 48 hrs: 5 g DI Water with pH adjusted to 5 using HCl, 10 g TEOS, and 20 g ethanol. For 75 g each of Mixtures #7 and #8, 15 g each of the TEOS-containing mixture. These two resulting mixtures were sonicated overnight and stirred for 1000 hrs. Gel formation was evident and the mixtures became a little grayish. Sol-gel reaction was stopped by addition of ethanol to 300 ml in each mixture. Grayish gel residues were obtained after the resulting mixtures were centrifuged at 5,000 RPM for 30 minutes. An attempt was made to disperse these wet gels by adding ethanol and sonicating overnight. In the end, only a small percentage (about 5 wt % at most) was dispersed. Thus, the gels and liquid with small dispersions were dried in air and then in vacuum at 100°C for 8 hrs. Dry materials became black and turned into bits of brittle material. Figure 5 below shows their basic morphologies.

If the objective is to produce silica nanotubes, the pre-cursor material shown in Figure 5 indicates that the silica-formation reaction onto the CNTs was allowed to proceed longer than needed. Structures in Figure 5 can be explained as that of the early stages of coarsening, whereby a slender network structure evolved into a thick-walled open cell structure with rounded nodes.

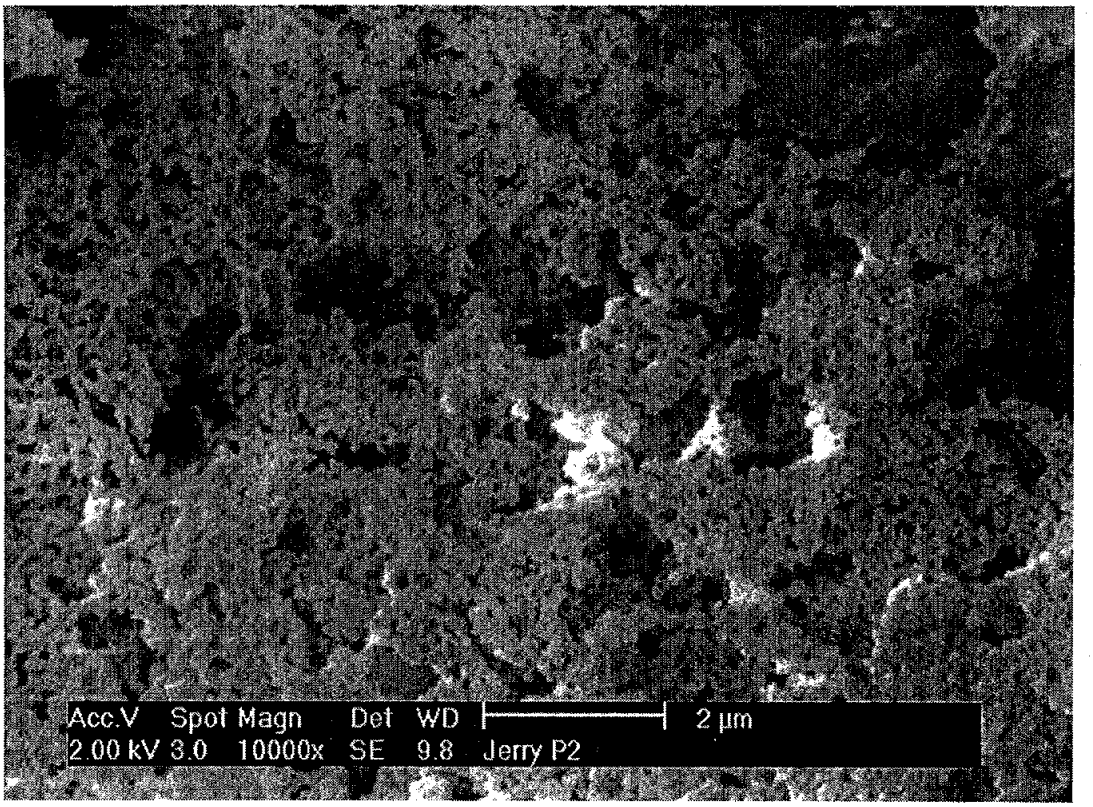


Figure 5: SEM of Silica templated from MWNT #8 before calcinations, showing an open cell network structure.

Calcination was done by heating the dried materials at 450°C for 12 hrs. We found that this resulted in some collapse of the open cell porous structure on Figure 5. After MWNT burn-off at 700°C for 12 hrs, surface area analysis was done with the products. Results still showed

In a new run, we took samples of various times during the MWNT-templated sol-gel process, and used a calcinations temperature of only 300°C for 12 hrs. Then, the MWNTs were burned off at 750°C for another 12 hrs. Resulting solid powdery materials are colored white, and Figure 6 below shows holes in the structure where the MWNTs have been burned off. These burned off regions appear to be look like eye sockets from spheroidal shells.

REFERENCES

1. Barrera, E. and Lozano, K., *J. of Materials*, 52, 32 (2000).

2. Callister, W.D., "Materials Science and Engineering: An Introduction", 6th Edition, John Wiley and Sons, New York, 2004, p. 620.

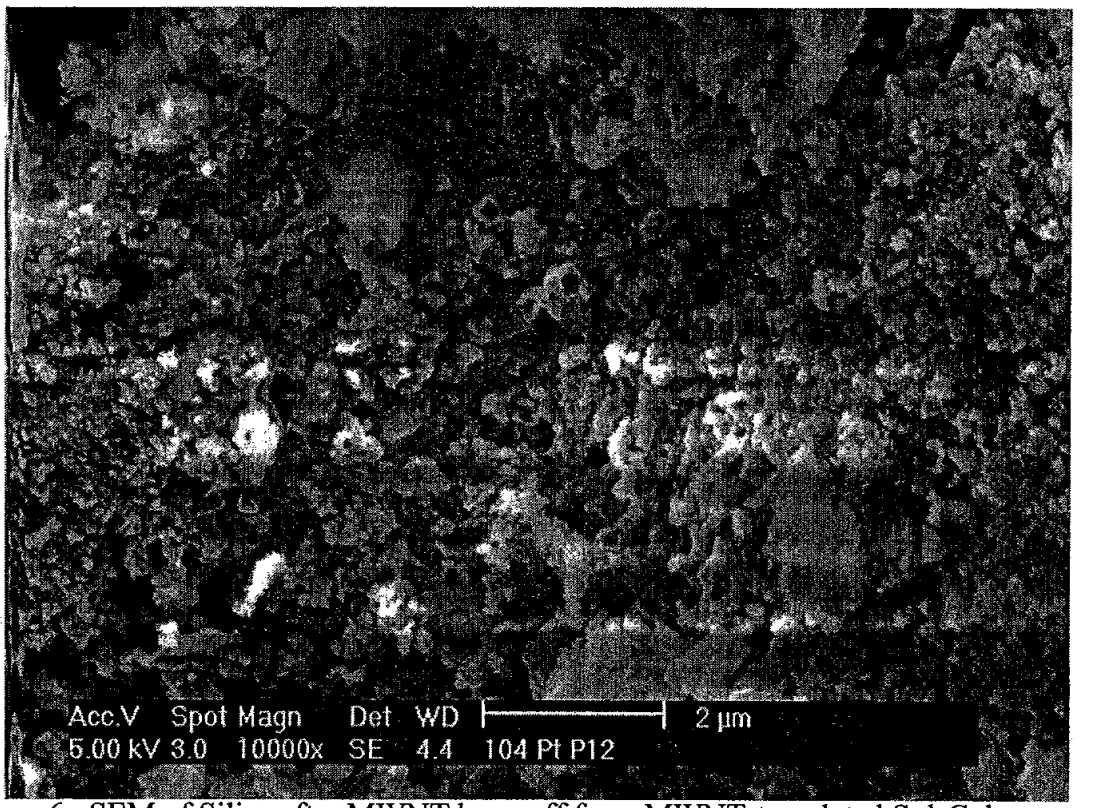


Figure 6: SEM of Silica after MWNT burn-off from MWNT-templated Sol-Gel process at 2 hrs of reaction. Regions where MWNT bundles (50-100 nm) have been burned off are seen in the form of "eye sockets" from spheroidal shells.

3. Dai, H., "Nanotube Growth and Characterization", in: Carbon Nanotubes: Synthesis, Properties, and Applications, M.S. Dresselhaus, G. Dresselhaus, and P. Avouris (Eds.), Springer, New York, 2001, pp. 29-53.
4. Heaney, M.B., "The Measurement, Instrumentation, and Sensors Handbook", Chapter on Electrical Conductivity and Resistivity, CRC Press, 1999.
5. Caneba, G.T. and Dar, Y.L., "FRRP Copolymers and Process for making the Same", submitted to *U.S. Patent and Trademark Office*, January, 2002, Publication 2003/0153708.
6. Ruhle, M., Seeger, T., Redlich, Ph., Grobert, N., Terrones, M., Walton, D.R.M., and Kroto, H.W., *J. Ceramics. Proc. Res.*, 3(1), 1 (2002)
7. Satishkumar, B.C., Govinddaraj, A., Nath, M., and Rao, C.N.R., *J. Mat. Chem.*, 10, 2115 (2000)

AC/DC Power Systems with Applications for future Lunar/Mars base and Crew Exploration Vehicle

Final Report
NASA Faculty Fellowship Program – 2004

Johnson Space Center

Prepared by:	Badrul H. Chowdhury, Ph.D.
Academic Rank:	Professor
University & Department	University of Missouri-Rolla Electrical & Computer Engineering Department Rolla, MO 65409-0040
NASA/JSC	
Directorate:	Engineering
Division:	Energy Systems Division
Branch:	Power Systems Branch
JSC Colleague:	Sabbir A. Hossain
Date Submitted:	August 6, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

The Power Systems branch at JSC faces a number of complex issues as it readies itself for the President's initiative on future space exploration beyond low earth orbit. Some of these preliminary issues – those dealing with electric power generation and distribution on board Mars-bound vehicle and that on Lunar and Martian surface may be summarized as follows:

- Type of prime mover – Because solar power may not be readily available on parts of the Lunar/Mars surface and also during the long duration flight to Mars, the primary source of power will most likely be nuclear power (Uranium fuel rods) with a secondary source of fuel cell (Hydrogen supply).
- The electric power generation source – With nuclear power being the main prime mover, the electric power generation source will most likely be an ac generator at a yet to be determined frequency. Thus, a critical issue is whether the generator should generate at constant or variable frequency. This will decide what type of generator to use – whether it is a synchronous machine, an asynchronous induction machine or a switched reluctance machine.
- The type of power distribution system – the distribution frequency, number of wires (3-wire, 4-wire or higher), and ac/dc hybridization.
- Building redundancy and fault tolerance in the generation and distribution sub-systems so that the system is safe; provides 100% availability to critical loads; continues to operate even with faulted sub-systems; and requires minimal maintenance.

This report describes results of a summer faculty fellowship spent in the Power Systems Branch with the specific aim of investigating some of the lessons learned in electric power generation and usage from the terrestrial power systems industry, the aerospace industry as well as NASA's on-going missions so as to recommend novel surface and vehicle-based power systems architectures in support of future space exploration initiatives. A hybrid ac/dc architecture with source side and load side redundancies and including emergency generators on both ac and dc sides is proposed. The generation frequency is 400 Hz mostly because of the technology maturity at this frequency in the aerospace industry. Power will be distributed to several ac load distribution buses through solid state variable speed, constant frequency converters on the ac side. A segmented dc ring bus supplied from ac/dc converters and with the capability of connecting/disconnecting the segments will supply power to multiple dc load distribution buses. The system will have the capability of reverse flow from dc to ac side in the case of an extreme emergency on the main ac generation side.

INTRODUCTION

On January 14, 2004, US President Bush announced a new vision for NASA [1]:

- Implement a sustained and affordable human and robotic program to explore the solar system and beyond;
- Extend human presence across the solar system, starting with a human return to the Moon by the year 2020, in preparation for human exploration of Mars and other destinations;
- Develop the innovative technologies, knowledge, and infrastructures both to explore and to support decisions about the destinations for human exploration.

The vision has two specific initiatives:

1. Lunar Exploration

- Begin robotic missions to the Moon by 2008, followed by a period of evaluating lunar resources and technologies for exploration
- Begin human expeditions to the Moon in the 2015 – 2020 timeframe

2. Mars Exploration

- Conduct robotic exploration of Mars to search for evidence of life, to understand the history of the solar system, and to prepare for future human exploration.
- Timing of human missions to Mars will be based on available budgetary resources, experience and knowledge gained from lunar exploration, discoveries by robotic spacecraft at Mars and other solar system locations, and development of required technologies and know-how.

Both initiatives will require NASA to develop and demonstrate power generation, propulsion, life support, and other key capabilities required to support more distant, more capable, and longer duration human and robotic exploration than ever done before. This report describes the requirements for power generation and distribution for human habitat on lunar/Martian surface and provides a preliminary system-level design. Preliminary discussions are also provided for a power system for the proposed VASIMR engine.

Power Systems for lunar/mars habitation

Fig. 1 shows the essential ingredients of a power system for future space applications.

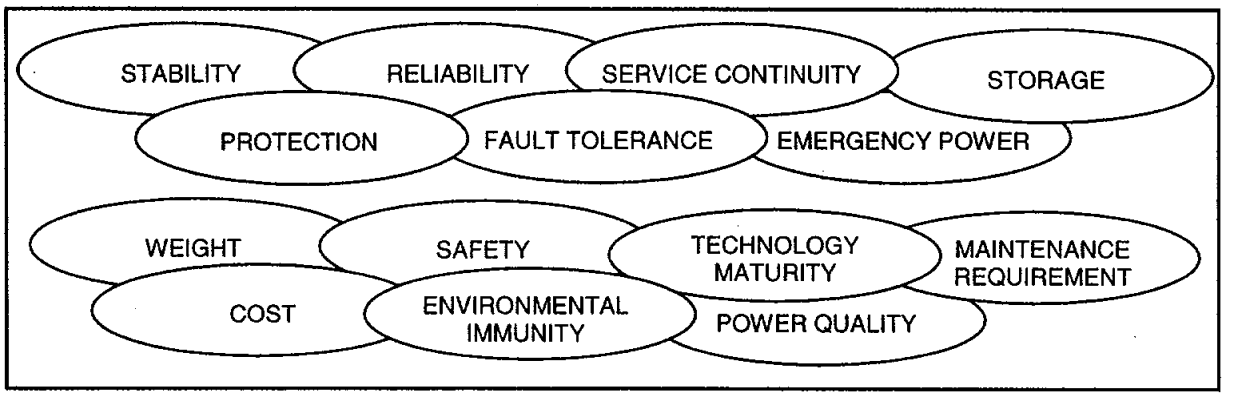


Fig. 1. Considerations for power systems for future space applications.

Undoubtedly, these characteristics are much more stringent than those expected of terrestrial power systems. The technological challenges in power systems for future space exploration are formidable and will require a careful study of what technology is available today and what needs to be developed before the vision can become a reality.

A.2 Power System for the VASIMR rocket engine

The Variable Specific Impulse Magnetoplasma Rocket (VASIMR) engine - a plasma-based propulsion system - will provide a new method of propulsion that can reduce interplanetary flight time [2]. High frequency radio waves are used to create electric fields to ionize gas (hydrogen, helium, or deuterium) particles creating plasma in a magnetic field. Low frequency radio waves are then used to add rotational energy to the ions in the cyclotron. A decrease in the magnetic field converts rotational energy of the ions into parallel energy. The ionized gas is then exhausted, thus providing thrust. The mass flow rate of the gas into the ionization chamber is controllable with a throttle, thereby changing the specific impulse of the engine. The VASIMR Engine has the capability to reduce travel time, making manned missions to Mars a distinct possibility in the future.

The engine can produce high power density with thrust velocities reaching 30,000 to 300,000 m/s. The engine is capable of varying the amount of thrust generated, allowing it to increase or decrease its acceleration. Electrical power sources for the VASIMR engine will most likely be a nuclear prime mover source running either a synchronous or an asynchronous machine with a rating of 5 to 10 megawatts of power.

AC AND DC SYSTEMS

Terrestrial power systems in the US and abroad consist of mostly ac generation with ac transmission serving mostly ac loads. More than 99% of the ac generation is accomplished by three phase ac synchronous generators. A very small percentage of generation in terrestrial power systems is done by solar photovoltaic systems (producing dc), and wind power generations systems (asynchronous generators generating variable frequency ac). Fig. 2 shows an example of a balanced 3-phase system represented by a one-line diagram. Generally, such a system is a 4-wire system.

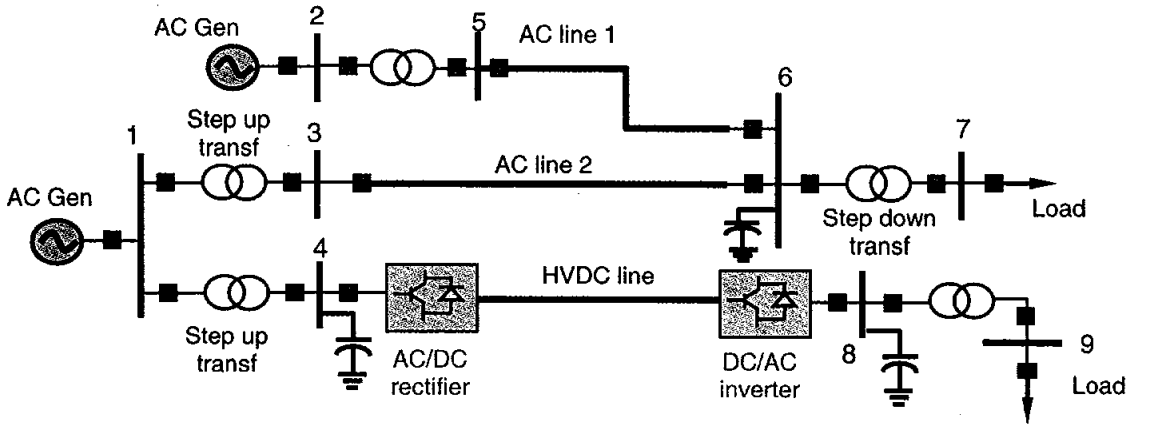


Fig. 2. A balanced 3-phase system represented by a one-line diagram showing 2 generators, 9 buses (nodes), 5 transformers, 2 ac lines, 1 HVDC line, 3 shunt capacitors, and 2 loads.

Three specific advantages of ac systems over dc systems should be noted. These are:

1. Synchronous generators will generate power in pure ac form. No filtering is required.
2. The synchronous generator source may be generally considered an infinite source, meaning that voltage (or frequency) will not drop excessively at the source with higher load currents. Dc sources, such as a battery, exhibits a drop in voltage as it discharges.

3. Voltage conversion is relatively easy by means of transformers with no additional filtering requirements.

Even a dc machine internally generates an ac voltage form which is mechanically converted to ac by means of commutators. Although the commutated voltage waveshape can be made to resemble a clean dc wave, the commutator requires periodic maintenance, has been known to fail more frequently than the rest of the dc machine, and is responsible for the higher cost of the dc machine. The concept of brushless dc machines was developed primarily to circumvent the disadvantages of the commutator. A brushless dc machine is simply a synchronous machine with a rectifier at the terminals to produce dc voltage. Hence, this machine suffers from the same disadvantages as an asynchronous machine with a power electronic interface. Incidentally, brushless dc motors are extensively used in hard disk drives and many industrial applications, and their market share is growing significantly in automotive, appliance and industrial applications.

Thus the technology exists to generate dc power. However, since voltages in dc systems cannot be transformed as easily as in ac systems, one has to resort to power electronics-based power conversion to bring about voltage step-up or step-down. Therefore, power electronic converters consisting of static semiconductor switches that are switched at fairly high frequencies, are used for this purpose. Unfortunately, converters create undesirable harmonics and electromagnetic interference (EMI) thus requiring use of filters. Additionally, the sensitive electronics inside the converter boxes have to be protected from faults.

Efficiency of AC systems

Transformers are about 90 – 95% efficient. Converters tend to be the same, perhaps a little less because of switching losses. DC systems are inherently less lossy because of the absence of reactive power. However, the resistive losses in dc lines are comparable to those in an ac line. For an understanding of how reactive power in ac systems leads to higher current magnitude, see Chapter 8 of IEEE Std 141-1993 - IEEE Red Book [3].

Line length and size of AC systems

Unlike a three phase ac system which requires at least three wires to carry current, dc power can be transmitted with only two wires (sometimes one). For a given ac voltage transmitted on a conductor, an HVDC system can carry 1.4 times that voltage on the same wire since an ac system's effective voltage is only 70.7% of the peak voltage. Therefore less insulation is needed for the wires.

Control and Stability of AC systems

DC power is inherently easier to control and re-route than ac power because of the absence of reactive power. In an ac system, active power can flow in one direction and reactive power in the other! Also, in the absence of line reactance in dc systems, power transfer is only limited by the thermal capacity of the line and not the steady state stability limit as is the case in ac lines.

Fault currents in AC systems

With high voltage DC, special attention should be given to fault propagation. Arc and corona suppression is of concern with higher voltage DC, especially at altitudes above 20,000

feet. Another consideration is the magnitude of the fault current with high voltage DC. Physical separation of high voltage and controls must be maintained. An advantage of DC is that simple make-before-break power transfers can be implemented to provide interrupt-free power transfers.

C. AC POWER CURCUITS

AC Power Generation

Typical options for bulk power generation, at the central station level based on fuel sources, include mostly coal-fired, natural gas-fired, water-powered, nuclear-powered, and petroleum-fired generations [4]. The common characteristics of terrestrial power generation are:

- A number of identical individual power generation units (hydro plants have larger units) make up a power plant usually in the hundreds of megawatts range.
- Many large power plants are located close to the fuel source. For example, the Jim Bridger power plant is located at Rock Springs, WY near a coal mine. The Grand Coulee power plant is located at the Grand Coulee dam on the upper Columbia River. Locating power plants close to the fuel source is still considered to be economically advantageous than transporting the fuel to a remote plant.
- More than 99% of bulk electric power is generated by 3-phase synchronous generators operating at 60 Hz frequency (50 Hz in most European and Asian countries).
- Maximum unit size in use today is 1300 MW.
- Most turbo-generator units (non-hydro) in the US have axial shafts and rotate at 1800 or 3600 rpm, while hydro-generators have vertical shafts and rotate at or below 1800 rpm.
- Power is generated at lower voltages. Typical generation voltage ratings (at the low side of the unit step up transformers) are: 240V, 480V, 600V, 2.4 kV, 4.16 kV, 6.9 kV, 13.8 kV [5]. Higher voltages (up to 23 kV) are possible with larger capacity generators. The limiting factor for high voltages are winding insulation and cooling requirements.
- The generated voltage is stepped up to much higher transmission voltages (138 kV to 765 kV. 1100 kV is also found in a few instances) by transformers in the switchyard.
- The generators are equipped with an exciter/AVR (for voltage and power factor control at the output) and a speed governor (for frequency control at the output of the gen).

AC Power Transmission

Because high voltage AC transmission lines have associated electric and magnetic fields, the mathematical model of a line includes both inductance and capacitance as shown in Fig. 3. Often, for short lines – lines shorter than 150 miles, the shunt capacitance of the line may be neglected. Additionally, line reactances for high voltage lines are much larger than line resistances. The electrical models of a synchronous generator and a transformer also include inductive reactances. The presence of these inductive reactances in an ac circuit can cause excessive voltage drops in the line unless somehow compensated for. Sometimes, mostly during high demands conditions, this situation leads to the voltage magnitude at the receiving end (load end) being much smaller than at the sending end.

Fig. 3a. One-line diagram of a transmission line

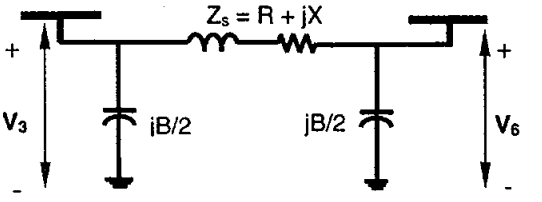


Fig. 3b. Electrical model of a transmission line

Mathematically, the voltage change between the sending and receiving ends of a line may be given by the following equation:

$$\Delta V = RI\cos\theta \pm XI\sin\theta \qquad (1)$$

where R = line resistance

X = line reactance

I = line current magnitude

θ = power factor angle

In (1), "plus" is used when the power factor is lagging and "minus" is used when the power factor is leading. Also, the $XI\sin\theta$ term is much larger than the $RI\cos\theta$ term and therefore, the reactive power flow has a much larger impact on the voltage change than the resistive power flow. In summary, ΔV is positive when the power factor is lagging and negative when the power factor is leading.

Three methods may be used to increase the receiving end voltage as listed below:

1. Increase voltage at the sending end. If using a synchronous generator, this simply amounts to increasing the set (reference) point of the voltage regulator associated with the excitation system which will then increase field excitation, thereby raising the reactive power output of the machine.
2. Compensate for the series reactance in the line by series capacitors.
3. Apply shunt compensation at the load. Use *CVT, Statcom* [6], etc.

Method 1 does not improve voltage regulation; it only helps to increase the voltage magnitude at the receiving end. In fact, the reactive power flow in the line increases in this method. Method 2 decreases the line reactance and is therefore capable of improving the load voltage as well as improving voltage regulation. However, there are resonance issues associated with this technique in addition to being a more expensive proposition. Method 3 is the most effective solution strategy from the perspectives of economy, voltage regulation and voltage stability. This method works well because it is able to decrease the reactive power flow in the line which also helps improve the power factor.

Stability of AC Circuits

Although the reactive power needs of the receiving end of an ac power system may be satisfied by means of reactive power generation at the source end, the reactive power flow increases in the transmission links leading to higher losses, which in turn leads to higher power generation at the source. In this situation, the maximum loadability point may be reached as shown in Fig. 4a. One may also observe the time evolution of the receiving end bus voltage which may continue to decrease even though shunt compensation is applied at the load, as shown in Fig. 4b. This condition is symptomatic of *voltage instability* and unless corrective actions are taken, the receiving end voltage may *collapse*.

Transmission Voltage And Frequency

Effect of voltage levels

Higher voltage levels are an advantage for power transmission because of lower transmission losses. However, higher voltages create higher electric field strength and are therefore conducive for the phenomenon of corona to occur. Thus line diameters have to be increased to compensate for this possibility, making lines heavier. Higher voltages also produce higher stress levels for insulation.

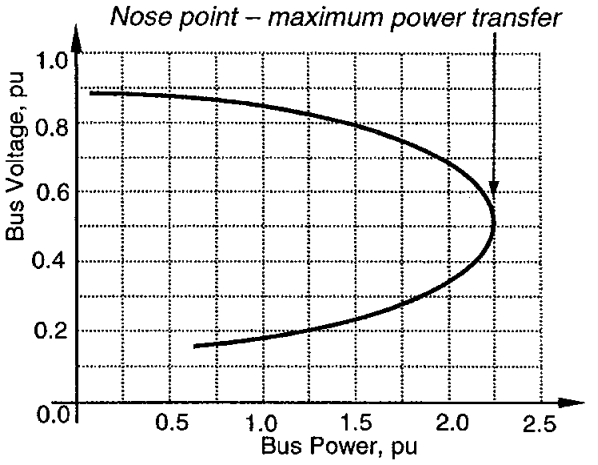


Fig. 4a. P-V curve showing bus voltage variation with real power

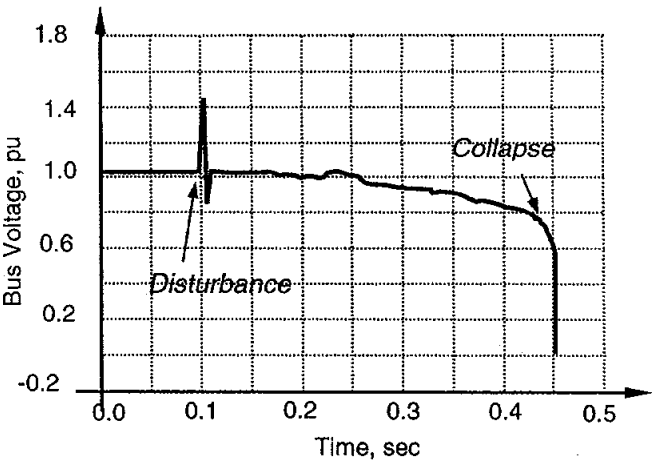


Fig. 4b. Voltage plotted as a function of time following a disturbance

Effect of frequency

The size and weight of transformers and energy storage elements decrease as the ac transmission frequency increases because of better utilization of the iron core. However, core losses in the magnetics (only 2-3% in 60 Hz transformers) and voltage regulation increase due to higher leakage reactance of transformers and line reactances of lines.

In dc transmission lines, the current is uniformly distributed through the conductor's cross section. However, in ac transmission lines, increasing frequency tends to crowd the current toward the outer perimeter of the conductor – *skin effect.* Since the line resistance is inversely proportional to the effective area of the conductor, it increases as frequency increases (proportional to square root of frequency) leading to higher losses. The resistive loss can be minimized and conductivity increased by plating the line with silver. Since silver is a better conductor than copper, most of the current will flow through the silver layer.

Higher frequency also has an impact on the short circuit capacities (SCC). It tends to lower the SCC as it is inversely proportional to the series reactance. Weaker systems have to be heavily capacitor compensated so as to maintain the required voltage at the load point. Additionally, at high frequencies, the X/R ratio of the wire tends to be large. This leads to lower amounts of damping during faults.

For buried cables, the shunt conductance is larger at higher frequencies leading to higher leakage currents. Thus ac frequency is an important parameter to consider for trade studies.

Building Fault Tolerance into Power Delivery

The rationale for building fault tolerance is to reduce the impact of faults by creating redundancies and providing looped sources with automatic transfer switches [7-11]. Fig. 5 shows a block diagram view of the various levels of redundancies in a simple power system. The configuration shown in Fig. 5(e) is preferred for the lunar/Mars habitat.

Voltage Transients – Switching Surges

Generally, any switching operation - fault initiation, interruption, operation of automatic transfer switches, etc. in a power system may be followed by a transient phenomenon in which transient overvoltages (TOV) can occur [12]. The sudden change in system condition can

generate damped oscillations with frequencies higher than the fundamental and determined by the resonant frequencies of the network. A capacitor switching voltage transient can be seen in Fig. 6. The magnitude of the switching overvoltages depends on:

- the type of circuit (RLC)
- the kind of switching operation (closing, opening, restriking)
- the type of loads
- the type of switching device or fuse

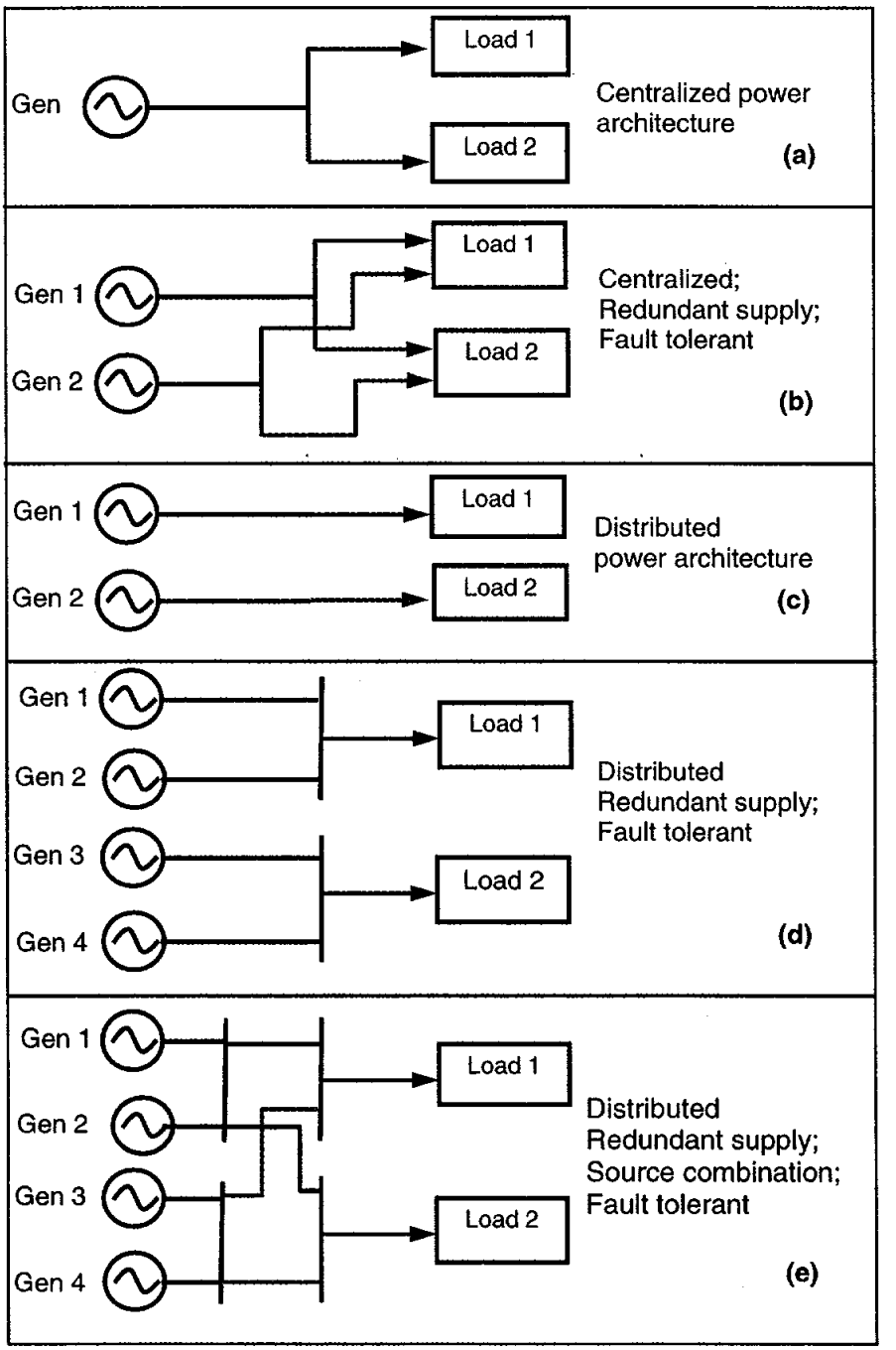


Fig. 5. Levels of redundancies provided by centralized and distributed power systems.

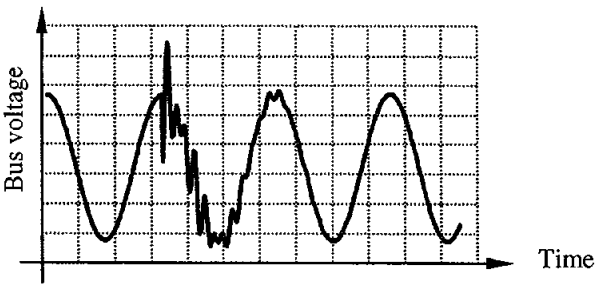


Fig. 6. Voltage transients due to capacitor switching.

The maximum transient voltage under any one or more of the following conditions:

- Instant at which the initiating event occurs – maximum TOV if it happens at the system voltage peak. (Note: the offset power-frequency current is greatest breaker closing at voltage zero).
- If a resonance condition exists in the system. System capacitor and inductance in resonance with load side capacitor.
- Whether ferroresonance condition exists
- Whether a capacitor restrikes.

TOV due to faults between line and ground

Depending on the configuration of the grounding, the fault current flows into one or more ground electrodes and generates ac overvoltages in the low-voltage system (load side) by ground coupling. The main parameter that influence the value and the duration of the TOV is the type of system grounding of the medium-voltage network

- Isolated (long time)
- Resonant-grounded (long time)
- Grounded through an impedance (longer time for high impedance and shorter for low-impedance grounded types)
- Solidly grounded (shorter time)

TOV due to a short circuit between line and neutral conductors

After the transient situation, the magnitude of the short-circuit current is limited only by the impedances of the supply and building wiring. The currents involved can be very high, ranging between one hundred and tens of thousands of amperes. A protective device operates to clear the fault. During this period of a few milliseconds to a few hundred milliseconds (but in all cases less than 5 s), a TOV can occur in the unfaulted lines of the affected power circuit.

THREE PHASE AC SYSTEMS

The advantages of three-phase AC systems are:

- Most AC generators are three-phase.
- Some power conditioning and electronic load equipment are operable only from a 3-phase power source.
- 3-phase systems may generally support larger loads with greater efficiency.
- The source impedance of three-phase systems is generally lower than 1-phase systems, which is important to minimize voltage waveform distortion due to nonlinear load currents.

Although three-phase voltage may be developed by different means, it is considered best practice to actually generate true three-phase power rather than convert single phase power (or dc power) to three-phase power. The methods to derive 3-phase power from 1-phase power ranges from using power electronics (1-ϕ ac to dc to 3-ϕ ac) to using single-phase motors turning a 3-phase generators, none of these is recommended by IEEE Std. 1100 [13].

Three Phase 3 or 4 Wire Systems

Three phase systems can be either 3-wire or 4-wire systems (Figs. 7a through 7d).

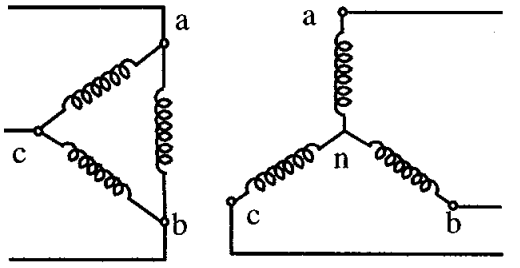


Fig. 7a. A 3-phase 3-wire ungrounded system.

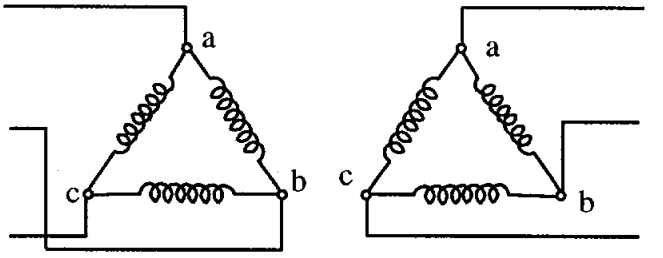


Fig. 7b. A 3-phase 3-wire ungrounded system

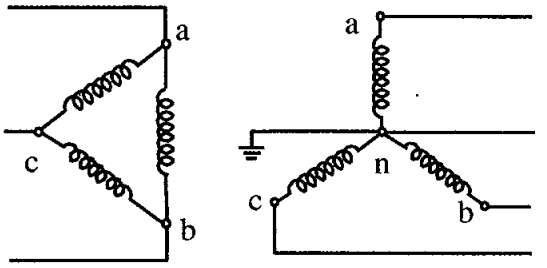


Fig. 7c. A 3-phase 4-wire system with neutral solidly grounded.

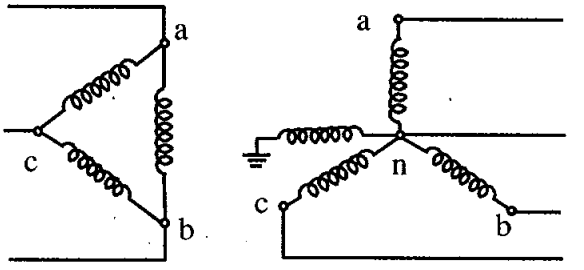


Fig. 7d. A 3-phase 4-wire system with neutral grounded through reactor.

Voltage Conversion (Transformer Connections)

Three phase transformers may be connected in any one of five configurations:

a. Y-Y (neutral grounded, ungrounded)
b. Y-Δ, Δ-Y (neutral grounded, ungrounded)
c. Δ-Δ (mid point grounded or ungrounded)
d. Special design (open delta)

Power systems do not generally use Δ-Δ configuration in low voltage systems because of its inability to provide lighting loads from a line to neutral voltage source. However, there are some advantages in using this configuration. If the Δ-Δ transformer is made of a bank of three single phase transformers, the system would continue to operate from just two transformers in an open delta configuration. However, the power capability is reduced and there is some amount of unbalance in the voltages.

Grounding

System grounding implies connecting the neutral point of a circuit element (rotating machine, transformer, etc.) on the system to ground either solidly or through a current limiting resistor or reactor. An ungrounded system has no intentional connection between a conductor

and ground. However, a capacitive coupling may exist between conductors and the adjacent grounded surfaces. Thus, an "ungrounded system" may be considered as a "capacitively grounded system" as shown in Fig. 8. Based on grounding, power systems may be classified as:

- Solidly grounded
- High resistance – used in low voltage (<600 V) systems
- Low resistance
- Ungrounded

Advantages of grounding

- Grounding offers a reference potential of zero volts.
- Allows ground faults to be detected thus allowing fast isolation of faulted part.
- The phase conductors are stressed at only line-to-neutral voltages above ground.
- Lower voltages in unfaulted phases during short circuit faults. See simulations.

The disadvantages include possibility of higher short circuit currents during short circuit faults.

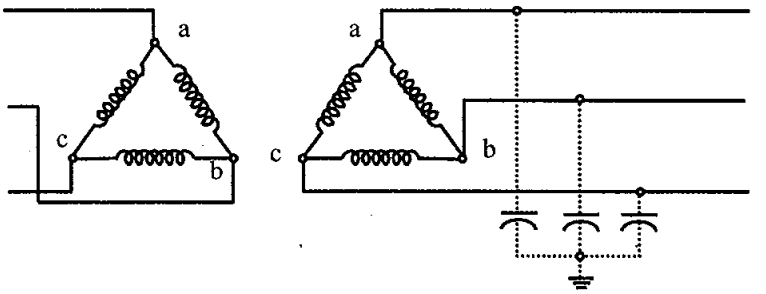


Fig. 8. A 3-phase 3-wire system with capacitive coupling to ground.

Ungrounded Systems

IEEE Standard 142-1991 [14] states that systems rated at 1000 V or less are suitable for ungrounded operation. The advantage of ungrounded systems is that no fault current will flow during a ground fault as there is no return path to the source. Therefore, the circuit will continue to operate safely after the first ground fault unless a second ground fault occurs. Such a situation is shown Case 3 of Appendix A.

A disadvantage of ungrounded systems is that in case of a ground fault, the other phase voltage will be subjected to full line-line voltages. Besides, the fault (or even normal switching activity) may result in high transient overvoltages (TOV) due to distributed capacitance of the line. Such high TOV may damage insulation or other pieces of sensitive equipment.

Another problem with ungrounded systems is the difficulty in locating or detecting faults. For these reasons, it may be advisable to apply high impedance grounding. Of course, in high impedance grounding, one must ensure that the impedance is low enough that the ground fault current is greater that the system's total capacitance-to-ground charging current. Otherwise, TOV may occur. IEEE Std. 141-1993 [3] states that high impedance grounding provides the same advantage as an ungrounded system, but additionally, limits the TOV.

400-Hz Power Generation

Some of US Navy's newest ships use 400 Hz, 3-phase nuclear power generation [15]. The current state-of-the-art in aerospace power is 400 Hz, 115 V (line-neutral), at either variable frequency or constant frequency [16-34]. Current aircraft electric power generation technology uses both the constant speed drive (CSD) [21-22] and the variable speed constant frequency

(VSCF) [23-27] technology. The CSD is an engine mounted generator system complete with an electrical generator and a variable displacement pump that constantly adjusts the output shaft rotation to maintain 24,000 RPM regardless of the throttle setting of the engine. The VSCF system is becoming the technology of choice lately mainly because of the advantages that power electronics provides from the perspective of weight and reliability. Fig. 9 shows a block diagram of such a system.

Electric power consumption on aircrafts will increase with the more electric aircraft and fly by wire paradigms now sweeping the industry [21, 22, 26]. Some of the concepts described here may be applied to future space exploration applications. However, the latter systems have to demonstrate a higher level of fault tolerance and reliability because of the long duration, relatively maintenance-free operational requirement in harsh environments.

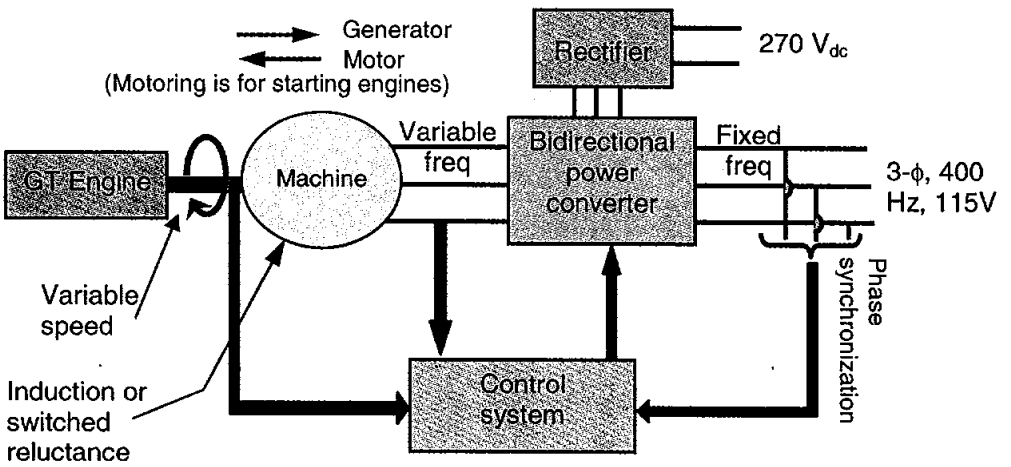


Fig. 9. Aircraft 400 Hz VSCF electric power generation system

SYSTEM LEVEL INTEGRATION

The proposed power system will consist of rotating and static, as well as ac and dc types of equipment both at the source and the load. For example, a fuel cell generating source, which has no rotating parts, will be interfaced with the ac side with a power electronic converter. On the other hand a synchronous machine is a complex rotating machine that depends on electromechanical principles to develop both active and reactive powers, but may or may not require a power electronic interface to the rest of the network depending on whether a voltage and/or frequency conversion is desired.

Fig. 10 shows the overall power system designed for lunar/Mars habitat. It consists of both an AC and a DC ring bus structures with generation sources located on both subsystems. The AC bus will have primary generation, while the dc bus hosts the backup/standby generation. Both ring bus structures allow isolation of specific segments for fault clearing while providing power to all load distribution centers from either the primary or the standby sources.

Generation

Fig. 11 shows the AC power generation and transmission at 400 Hz frequency. The solid state bus tie (SSBT) is normally open. It is closed only when any one of the sources is lost. Fault tolerance is added by providing a tie between the two sources by means of another SSBT. The solid state breaker (SSB) operates must faster than its mechanical counterpart. The assumption is

that power can be generated close to the habitat and thus long transmission line will not be needed. Power may be generated at a variable frequency at the source by means of a switched reluctance machine (SRM) or a double-fed induction machine (DFIM) while making use of the concepts of power electronic building blocks (PEBB) [35].

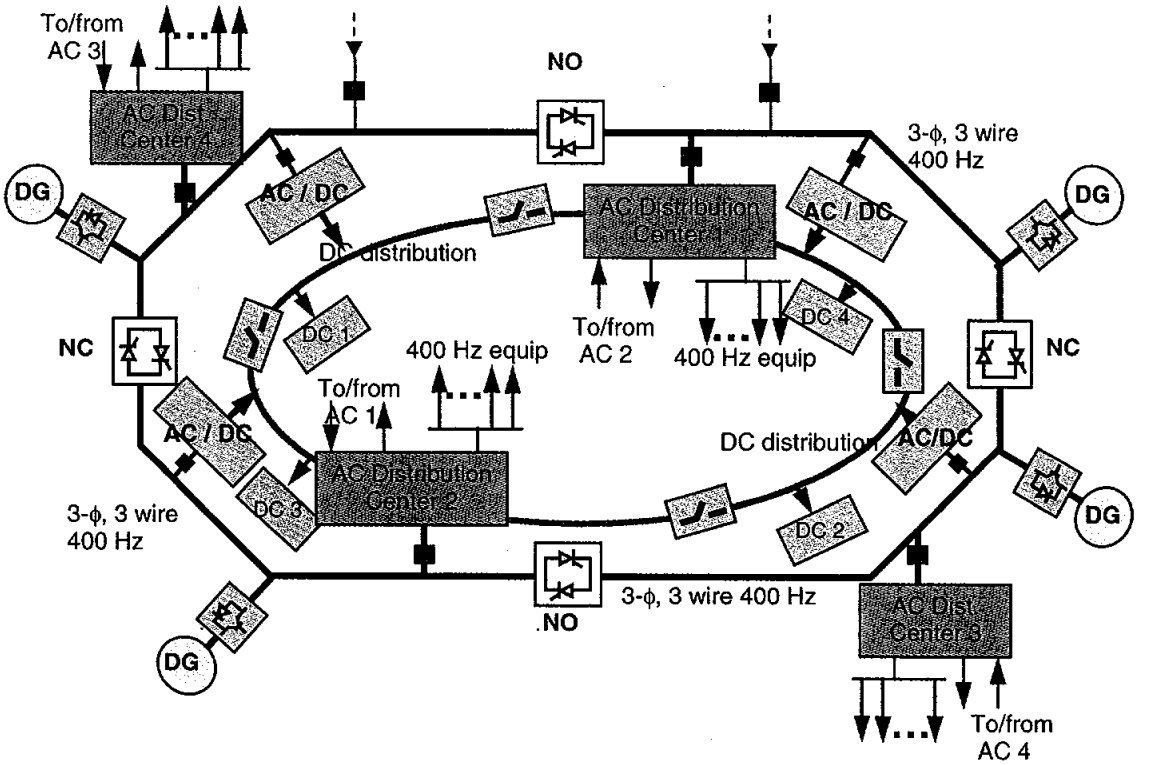


Fig. 10. The overall power system for lunar/Mars habitat

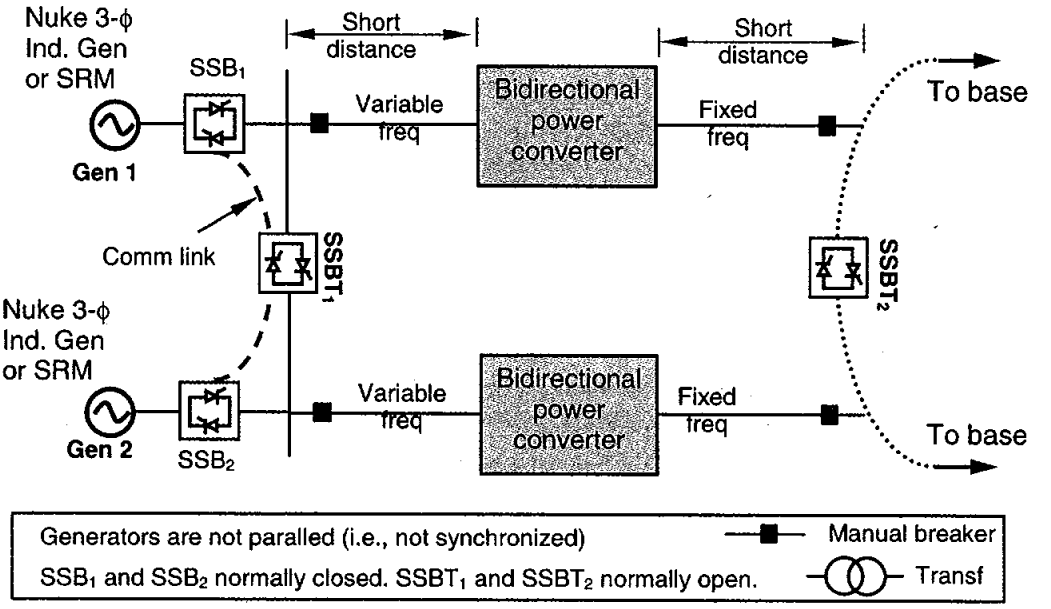


Fig. 11. Nuclear-based AC power generation and transmission using SRM and solid state converters.

CONCLUSIONS

Terrestrial and aircraft power system architectures are two of the most successful applications of engineering innovation and ingenuity that are at the core of the comforts and conveniences of modern living. They have both evolved over several years of design changes and, although neither is perfect, they operate very well under a great deal of uncertainty in operating conditions. On the other hand, the power systems on board the Shuttle Orbiter and the International Space Station offer the very best in power system technology, considering the enormous constraints of space travel and limitations in fuel supply. Incredibly, the technological challenges in building power systems for future space exploration and habitation on the Moon or Mars will be multiplied many fold. The requirements for applications in space beyond low earth orbit (LEO) as compared to terrestrial applications are listed below:

- Build N-redundancy for safety, fault tolerance and service continuity. N is a number much greater than that used on terrestrial system where N is usually 1 or at most 2.
- Build modularity for easy maintenance and repair. Terrestrial systems are inherently non-modular.
- Build distributed architecture for fault tolerance. This is in stark contrast to the centralized nature of terrestrial systems.
- Control and protection devices should be fast. Although protection devices for terrestrial applications are becoming faster now due to application of power electronics, the fastest device operates at about ¼ cycle of 60 Hz frequency. Space applications will most likely use 400 Hz frequency and thus will require faster protection devices.
- Avoid catastrophic wiring failure at all cost for safety. Conductor failures are common on terrestrial systems due to lightning, ice and wind.
- Total loading on the system should be monitored and kept within bounds. The allowable band is much more flexible on terrestrial systems because of a stiffer source.
- Overcurrents and overvoltages should be avoided. Again, the allowable band is much more flexible on terrestrial systems.
- Conductor wires should be as short as possible to avoid voltage drops and reduce weight. On terrestrial systems, high voltage transmission lines run for hundreds of miles.
- The system must have reserve and emergency backup for long durations. This is similar in nature to terrestrial systems.
- Generate power at higher voltages to reduce current and size/weight. Although the philosophy is similar for terrestrial systems, there is a limit to the highest voltage levels that may be used because of the prospect of corona occurring in outer space at lower levels.
- Use of power electronics for high density power systems. Generic terrestrial power systems do not generally use power electronics to generate or distribute power barring for a few exceptions.
- Use electromechanical systems utilizing permanent magnet (PM) machines to reduce weight. Generators on terrestrial systems use mostly synchronous machines where both the stator and rotor have windings.
- Use power electronic converters for voltage regulation. On terrestrial systems, automatic voltage regulation at the source is done by an exciter connected to the rotor circuit of a synchronous generator. Since a PM machine will be used for space application, power electronic conversion is required to bring about a similar effect.

REFERENCES

[1] NASA News Brief, "President Bush Offers New Vision For NASA – January 14, 2004," transcript available http://www.nasa.gov/missions/solarsystem/bush_vision.html

[2] ASPL, http://spaceflight.nasa.gov/shuttle/support/researching/aspl/index.html, JSC Advanced Space Propulsion Laboratory webpage.

[3] IEEE Std. 141-1993 (Red Book), *IEEE Recommended Practice for Electric Power Distribution for Industrial Plants*. Available from http://standards.nasa.gov/default.htm

[4] Energy Information Administration, http://www.eia.doe.gov/cneaf/electricity/ipp/ipp_sum.html

[5] ANSI C50.13-1977 – *ANSI National Standard Requirements for Cylindrical Rotor Synchronous Generators*, 1977.

[6] B.H. Chowdhury, *Power Quality Course Notes*, Available from Author.

[7] R.E. Johnson, "Practical considerations in the design of power system architectures for fault tolerant systems," *IEEE Digital Avionics Systems Conference*, 2001 DASC, vol. 1, pp 1A2/1 - 1A2/12, 14-18 Oct. 2001.

[8] M. Bailey, N. Hale, G. Ucerpi, J.-A. Hunt, S. Mollov, A. Forsyth, "Distributed Electrical Power Management Architecture," *IEE Colloquium on Electrical Machines and Systems for the More Electric Aircraft* (Ref. No. 1999/180), pp 7/1 - 7/4, 9 Nov. 1999.

[9] T.F. Glennon, "Fault tolerant generating and distribution system architecture," *IEE Colloquium on All Electric Aircraft* (Digest No. 1998/260), pp 4/1 - 4/4, 17 June 1998.

[10] M.W. Stavnes, A.N. Hammoud, "Assessment of safety in space power wiring systems," *IEEE Aerospace and Electronic Systems Magazine*, 9(1), pp 21 - 27, Jan. 1994.

[11] S. Green, D.J. Atkinson, B.C. Mecrow, A.G. Jack, B. Green, "Fault tolerant, variable frequency, unity power factor converters for safety critical PM drives," *IEE Proceedings-Electric Power Applications*, 150(6), pp 663 – 672, 7 Nov. 2003.

[12] IEEE C62.41.1 - *IEEE Guide on the Surge Environment in Low-Voltage (1000 V and Less) AC Power Circuits*, 2002.

[13] IEEE Industry Applications Society, *Recommended Practice for Powering and Grounding Electronic Equipment*, IEEE Std. 1100-1999, IEEE 1999.

[14] IEEE Std. 142-1991 (Green Book), *IEEE Recommended Practice for Grounding of Industrial and Commercial Power Systems*, Available from: http://standards.nasa.gov/default.htm

[15] Naval ships' Technical Manual - Chapter 320, *Electric power Distribution systems*, S9086-KY-STM-010/CH-320R2, 21 April, 1998.

[16] SAE, *Aerospace Recommended Practice for Electric Power Management*, Document SAE-ARP5584, 2003.

[17] International Organization for Standardization, *Aerospace Characteristics of aircraft electrical systems*, International Standard ISO 1540, Second edition - 1984-12-01

[18] IEEE Aerospace and Electronics Systems Society, *IEEE Guide for Aircraft Electric Systems, IEEE Std 128-1976*, New York, 1976.

[19] Department of Defense, *Aircraft Electric Power Characteristics*, MIL-Std-704f, March 2004.

[20] R. Schroer, Electric power. A century of powered flight: 1903-2003, *IEEE Aerospace and Electronic Systems Magazine*, 18(7), July 2003, 55 – 60.

[21] A.Emadi, M. Ehsani, "Aircraft power systems: technology, state of the art, and future trends, *IEEE Aerospace and Electronic Systems Magazine*, 15(1), pp 28 – 32, Jan. 2000.
[22] M.E. Elbuluk, M.D. Kankam, Potential starter/generator technologies for future aerospace applications, *IEEE Aerospace and Electronic Systems Magazine*, 12(5), pp 24 – 31, May 1997.
[23] M.H. Taha, D. Skinner, S. Gami, M. Holme, G. Raimondi, "Variable frequency to constant frequency converter (VFCFC) for aircraft applications," *IEEE International Conf. on Power Electronics, Machines and Drives*, (Conf. Publ. No. 487), pp 235 – 240, 4-7 June 2002.
[24] L.J. Feiner, "Power electronics transforms aircraft systems," *IEEE Conf. on 'Idea/Microelectronics'*, WESCON/94, pp 166 – 171, 27-29 Sept. 1994.
[25] L.J. Feiner, "Power electronics for transport aircraft applications," *Proc. of the Int. Conf. on Industrial Electronics, Control, and Instrumentation*, IECON '93, vol.2, pp 719 – 724, 15-19 Nov. 1993.
[26] L.J. Feiner, "Power-by-wire aircraft secondary power systems," *1993 AIAA/IEEE Digital Avionics Systems Conf.*, 12th DASC, pp 439 – 444, 25-28 Oct. 1993.
[27] T.M. Jahns, M.A. Maldonado, "A new resonant link aircraft power generating system," *IEEE Transactions on Aerospace and Electronic Systems*, 29(1), pp 206 – 214, Jan. 1993.
[28] Hamilton Sunstrand Corp., webpage: http://www.hamiltonsundstrandcorp.com/generic/0,3626,CLI1_DIV22_ETI2766,00.html
[29] L. Andrade, C. Tenning, "Design of Boeing 777 electric system," *IEEE Aerospace and Electronic Systems Magazine*, vol. 7(7), Pp 4 – 11, July 1992.
[30] R.L. Steigerwald, G.W. Ludwig, R. Kollman, "Investigation of power distribution architectures for distributed avionics loads," *26th Annual IEEE Power Electronics Specialists Conference*, PESC '95, vol. 1, pp 231 – 237, 18-22 June 1995.
[31] M.J.J. Cronin, "The all-electric aircraft," *IEE Review*, 36(8), pp 309 – 311, 13 Sept. 1990.
[32] R.E. Niggemann, S. Peecher, G. Rozman, "270-VDC/hybrid 115-VAC electric power generating system technology demonstrator," *IEEE Aerospace and Electronic Systems Magazine*, 6(8), pp 21 – 26, Aug. 1991.
[33] C. Anghel, "A novel start system for an aircraft auxiliary power unit," *35th Intersociety Energy Conversion Engineering Conference and Exhibit*, vol. 1, pp 7 – 11, 24-28 July 2000.
[34] D.M. Defreitas, "High performance electrical power systems for unmanned airborne vehicles," *Proc. of the 1988 IEEE Southern Tier Technical Conf.*, pp 40 – 51, 19 Oct. 1988.
[35] Virginia Power Electronics Center, "Power Electronics Building Blocks and System Integration," Final report 1999-2000, Office of Naval Research, N00014-98-1-0828, 2000.

Diagnostics of Carbon Nanotube Formation in a Laser Produced Plume: Spectroscopic *in situ* nanotube detection using spectral absorption and surface temperature measurements by black body emission.

Final Report
NASA/ASEE Summer Faculty Fellowship Program – 2004
Johnson Space Center

Prepared By:	Gary D. De Boer, Ph.D.
Academic Rank:	Associate Professor
University & Department	LeTourneau University Chemistry and Physics
NASA/JSC	
Directorate:	Engineering
Division:	Structural Engineering
Branch:	Materials and Processes
JSC Colleague	Carl Scott
Date Submitted	August 5, 2004
Contract Number	NAG-9-1526

ABSTRACT

Carbon nanotubes hold great promise for material advancements in the areas of composites and electronics. The advancement of research in these areas is dependent upon the availability of carbon nanotubes to a broad spectrum of academic and industrial researchers. Although there has been much progress made in reducing the costs of carbon nanotubes and increasing the quality and purity of the products, an increase in demand for still less expensive and specific nanotubes types has also grown.

This summer's work has involved two experiments that have been designed to further the understanding of the dynamics and chemical mechanisms of carbon nanotube formation. It is expected that a better understanding of the process of formation of nanotubes will aid current production designs and stimulate ideas for future production designs increasing the quantity, quality, and production control of carbon nanotubes.

The first experiment involved the measurement of surface temperature of the target as a function of time with respect to the ablation lasers. A peak surface temperature of 5000 K was determined from spectral analysis of black body emission from the target surface. The surface temperature as a function of various changes in operating parameters was also obtained. This data is expected to aid the modeling of ablation and plume dynamics.

The second experiment involved a time and spatial measurement of the spectrally resolved absorbance of the laser produced plume. This experiment explored the possibility of developing absorbance and fluorescence to detect carbon nanotubes during production. To attain control over the production of nanotubes with specific properties and reduce costs, a real time *in situ* diagnostics method would be very beneficial. Results from this summer's work indicate that detection of nanotubes during production may possibly be used for production feed back control.

INTRODUCTION

What is so important about carbon nanotubes?

Nanotechnology, the use of materials with dimensions of nanometers, represents engineering at the molecular scale, at a dimension beyond those typically used by chemists and much below those of the bulk dimensions used by engineers. It is within this interfacial domain of measure that materials of great promise for material and electronic advancement have been observed and proposed. Many of these promises have focused on the use of nanometer scaled tubes discovered in 1991 by Iijima. [1] These tubes, with dimension of nanometers in diameter and microns in length, can be described as the elongation of fullerenes into tubes. Fullerenes are spherical or elliptical in shape, the most well known being that composed of sixty carbon atoms and having the shape of a soccer ball, as proposed by Smalley in 1985. [2] Examples of a fullerene and two nanotubes can be seen in Figure 1. Carbon nanotubes are the building materials for many proposed nanostructures; therefore, an understanding of their properties and techniques for their utilization are essential to progress toward nanotube-based nanostructures.

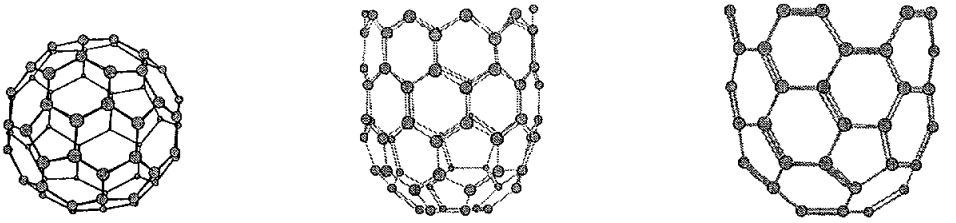


Figure 1: C_{60}, (9,0) metallic zigzag tube, and (5,4) semi-conducting armchair tube.

What are the problems?

Although much progress has been made in many areas of carbon nanotube production, characterization and applications, current production methods are still financially prohibitive for most commercial application and many academic research groups. Current production methods also result in tubes of various purity, diameter, length, and chirality. A more thorough understanding of the chemical mechanisms and better production feedback controls are essential to improve the production of carbon nanotubes and meet the demand for affordable quantities of nanotubes of selective properties.

What has been done to elucidate the chemical mechanisms?

Initial work on the elucidation of the chemical mechanisms has been done on the postproduction evaluation of the targets and products as a function of various production parameters. [3-6] Recent *in situ* work has followed various species during nanotube formation. Nickel atom, cobalt atom, C_2, and nonspecific larger carbonaceous materials

have been followed during nanotube formation in a laser-produced plume. [7-12] *In situ* work has been much more productive in explaining the chemistry involved in tube formation than the post analysis work. Scott *et al.* presents a summary of current thought with respect to the carbon nanotube formation mechanisms based on both the initial post production analysis and the recent *in situ* reports.[13] Questions remain about the role of the catalyst in its atomic and condensed particle form as well as the time and spatial variables involved in carbon nanotube formation.

What are the current methods of production feedback controls?

Currently, there is no production feedback control employed in the laser production methods at JSC. Production parameters such as gas flow, oven temperature, and laser output are monitored by the operators during production. The quality and quantity of tube production must be done post production.

The HiPco process of carbon nanotube production, at Rice University, does employ feedback through the monitoring of CO_2 produced during the disproportionation of CO on iron to form C_2 and CO_2. [14] Increased production of CO_2 is correlated to increased production of reduced carbon which will eventually lead to the formation of carbon nanotubes. By tuning parameters so that CO_2 output is optimized, the production of nanotubes is also optimized.

CO is not the feedstock in the laser method, and there is no CO_2 produced. The laser method would need a different species to provide feedback.

What else do we need to know?

Although there are many variables involved in carbon nanotube formation that can be explored, a method of detecting the presence of nanotubes *in situ* in real time during nanotube production would be very valuable in elucidating the chemical mechanisms and providing real time production feedback control.

EXPERIMENTAL

How can we do experiments that will give us the information we need?

At NASA-JSC any approach to studying surface temperature and the detection of nanotubes *in situ* during nanotube formation would have to be designed with respect to the current production configuration. The nanotube production configuration at JSC follows that developed at Rice University [15] and has been described previously by Arepalli, *et al.* [7, 8] Briefly, the setup includes a carbon target (19 mm diameter) which is doped with 1% nickel and 1% cobalt and is supported on a rod in an oven which is heated

to 1473 K during normal production. The target and rod are centered within a 50.8 mm quartz tube. A smaller 25.4 mm quartz tube is centered within the 50.8 mm tube and extends to within 6 mm of the target. Argon flows through the tubes toward the target at a pressure of 67 kPa and a flow rate of 100 sccm. Two Nd:YAG ablation lasers follow a path through the inner tube to strike the flat end of the target at normal incidence. The green (532 nm) Nd:YAG laser fires 50 ns prior to the IR (1064 nm) Nd:YAG laser.

The JSC nanotube production approach and facilities are very conducive to spectroscopic probing of intermediate species and products. We made use of spectroscopic techniques to measure the surface temperature of the target upon ablation and to measure the absorption of the laser produced plume during production. The former to provide empirical values for modeling projects and the latter to explore the possibility of developing production feedback controls.

Experiment 1: Target Surface Temperature Measurements

The surface temperature of the target was measured using existing fiber optics and optical dispersion techniques. A new optical collection configuration was introduced to collect blackbody emission directly from the target surface. The experimental setup is illustrated in Figure 2. This differs slightly from the nominal production configuration in that there is only a 25.4 mm tube rather than the 25.4 mm tube within the 50.8 mm tube as described above. Also there is a Y in the tube, at an angle of 45°, with the shorter leg being 19.0 mm in diameter. Due to the smaller diameter tube, a smaller diameter target was used, 12.0 mm rather than 19.0 mm.

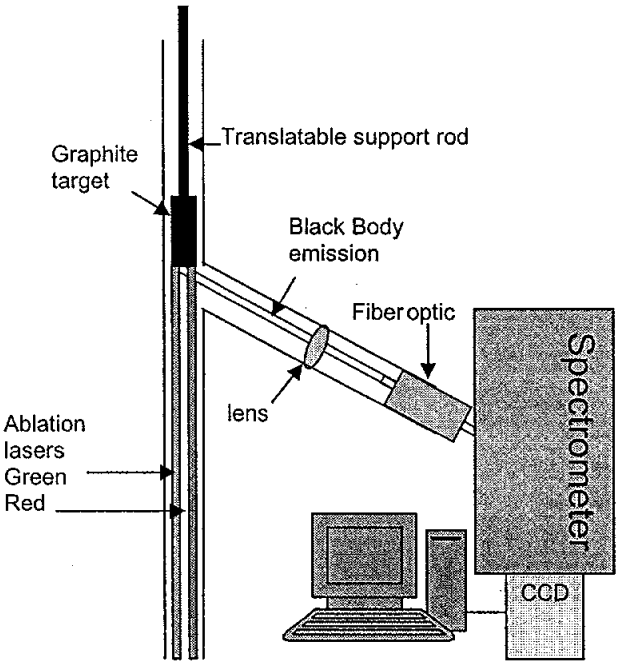


Figure 2: Surface temperature experimental setup

Studies involving the green laser also included a notch filter at 532 nm to avoid saturating the detector with scattered laser radiation. A temporal gate of 12 ns was used for collecting emission using various slit widths on the spectrometer depending on the amount of radiation emitted upon ablation.

Experiment 2. Absorbance measurements of the laser produced plume.

One of the attractive properties of the carbon nanotube is that its conductivity has been calculated to be a function of tube chirality and diameter. [16] An example of two different chiralities can be seen in Figure 1, the extremes of zigzag and armchair. It has been only recently that spectroscopic measurements of the band gaps associated with carbon nanotubes have been measured. [17-20] Absorption and fluorescence measurements have been well studied for nanotubes suspended in solution. Fluorescence of nanotubes requires very good solvation as it is thought that if any of the individual tubes within the ropes is a metallic conductor, fluorescence from excited electronic states will not be observable due to quenching by the metallic tubes which allow for a path of non-radiative electronic relaxation. Since there is a distribution of chirality and diameter in the production of nanotubes, the presence of metallic tubes in a rope is quite probable. Not until dispersion techniques had improved, was it possible to measure band gaps of isolated tubes by detection of fluorescence.

In the JSC nanotube production facility we hope to detect nanotubes *in situ* during carbon nanotube formation using recently reported absorption bands. Absorption was chosen rather than fluorescence because the JSC facility is equipped to measure light in the visible wavelength range of nanotube absorption but not in the infrared region of nanotube fluorescence. It was expected that tubes initially form individually in the gas phase before they flocculate into bundles later in time. Flocculation or the condensing of tubes into bundles would broaden the absorption bands and would likely quench fluorescence.

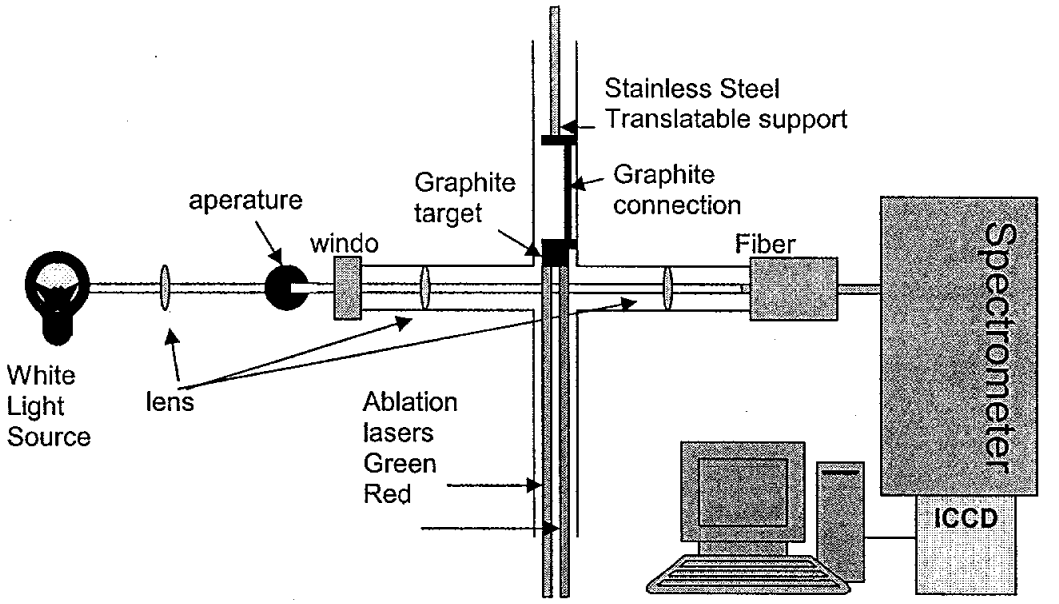


Figure 3: Experimental set up for absorbance measurements.

The experimental setup for absorption measurements differs from production in the following ways. An x-tube is used in place of the standard production tubes. The x-tube is 25.4 mm in diameter along the optical path of the ablation lasers and a 19 mm in diameter along the path perpendicular to path of the ablation lasers, rather than the 25.4 mm tube within the 50.8 mm tube as described above. The light transmitted through the optical path of the white light is collected by an optical fiber and dispersed with a spectrometer onto a CCD so that a wavelength resolved transmission spectrum is obtained. The CCD is gatable with respect to time of ablation and the graphite target is mounted on a translatable stage so that it is possible to probe for nanotubes in both temporal and spatial dimensions. A simplified experimental set up is illustrated in Figure 3.

RESULTS and ANALYSIS

Experiment 1. Surface Temperature

Emission from the target surface was collected using the y-tube design. The y-tube performed as designed, allowing for a consistent signal of much greater intensity than did previous diagnostic setups which collected emission transmitted through the standard quartz production tube. Although the y-tube design appears to be fairly robust, it was found that operating under lower pressures than 500 Torr at 1200 degrees Celsius caused the y-tube to begin a collapse that would slowly continue when operating at 1200 degrees Celsius even at the normal operating pressures of 500 Torr.

Emission from the surface of the target was collected under many different experimental parameters. In all the experiments the emission was resolved by wavelength. In experiments involving the ablation lasers, the emission was also resolved with respect to the time of ablation. This was done by collecting wavelength resolved emission at a variety of time delays from 200 nanoseconds prior to the laser pulse to 3 microseconds after the laser pulse with a time gate of 40 nanoseconds. Although greater time resolution is possible, 12 nanoseconds being the shortest time interval, shorter times result in poor statistics and poor signal quality. Experimental conditions included lasers operating in standard production parameters, operating singly, operating in reverse order, and operating with time delays of 0, 50, and 500 nanoseconds. Argon flow rates were also varied. Helium was used a substitute buffer gas for Argon. Oven temperature was operated at the standard 1200 and also 1000 degrees Celsius. Emission was collected from the center of the target to the edge of the target at 1 mm increments.

It would be difficult to report all the results of these experiments within the limits of a report of this nature, so only a few remarks will be made here in hopes of writing a more comprehensive report at a later date. Methodology of the analysis and then some of the general results are described below.

Although the y-tube is designed to collect emission from the target surface, emission from the laser-produced plume is also unavoidably present. Analysis requires a discernment to be made between the plume emission and the surface black body emission. Two methods were developed that would allow for this discernment, the first will be referred to as the ratio method and the second as the baseline curve fit method.

Method 1: The ratio method.

This method assumes an emission entirely from black body at a wavelength that was as far from the plume emission as possible while still being in a responsive region of the detector. This emission was then compared to the black body emission of the target under conditions in which no lasers were being used, assuming a black body temperature equal to the ambient temperature of the oven. The ratio of the emission intensity at a given wavelength to the intensity of emission under ambient oven conditions at the same wavelength can be used to determine the blackbody temperature of the emission of the former. Since the emission produced by laser ablation may include contributions from the plume in addition to emission from the surface the temperature obtained from the emission intensity ratios will give an upper limit temperature.

Method 2: The curve fit method.

The curve fit method involves correcting the raw data for the instrument response and then fitting the data to calculated black body curves. This method involves many data across the spectral range and therefore emission from C_2 and other sources is unavoidable. Therefore a subjectively determined baseline underneath any structured spectrum is interpreted as the blackbody emission. Spectra taken under ambient oven conditions and corrected for response fit very well to a black body curve of the ambient temperature, 1473 K.

Our results from method one and two indicate a peak surface temperature of 3000 K and 5000 K respectively. The ratio method was used in Figure 4 to calculate temperature at a number of different time delays with respect to the time of laser ablation. From this, it is noted that there is a steep temperature gradient across the target and that the target returns to near ambient temperatures within a few milliseconds. Figure 4 represents a temporal temperature profile at different positions on the target surface. The 'zero' position is taken to be the center of the target where the 4.8 mm diameter laser beam is also centered. As the target is moved one millimeter in, the 45 $^{\circ}$ angle of collection is such that emission is collected from a spot on the surface one millimeter outward from the zero position. Only three of these steps will move the emission spot from the laser spot. It should also be expected that the laser's energy profile across the spot is not flat and drops off from the center toward the edge. From Figure 4 it is clear that there is little change in temperature after moving 2 millimeters away from the center position.

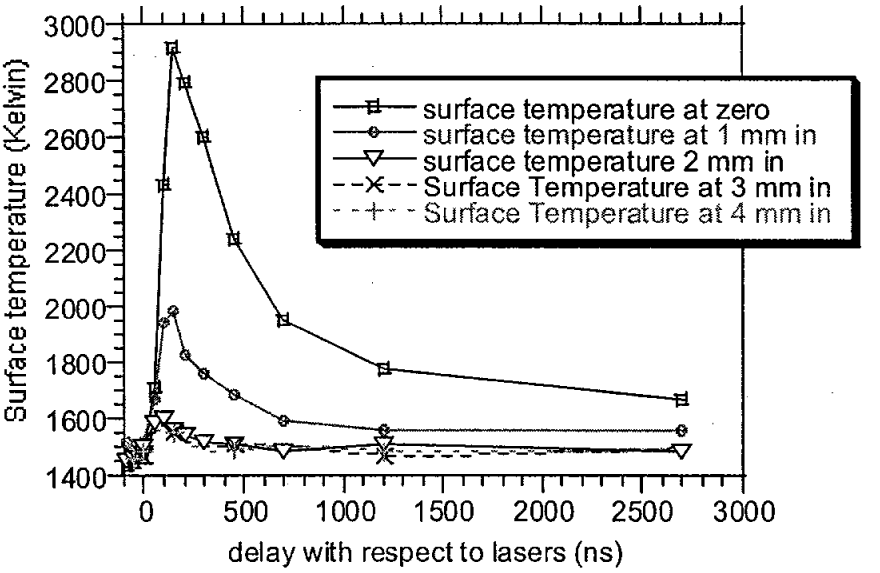


Figure 4: Surface temperature as a function of time for the standard laser combination. The different curves represent temporal profiles of temperature at different positions on the target surface.

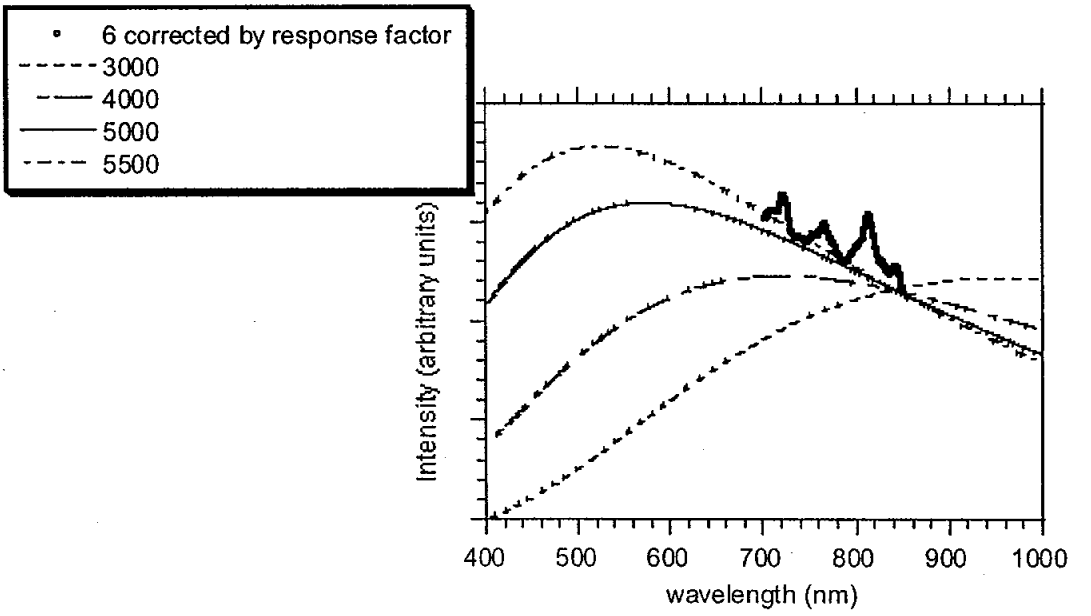


Figure 5: An illustration of the surface temperature obtained using a corrected spectrum fit to a calculated black body emission. The surface temperature obtained using this method, method two, appears to be between 5000 K and 5500 K rather than 3000 K as determined using the ratio method, method one.

Using the temporal temperature profile for the zero position, the spectrum collected at a time corresponding to the peak temperature was fit using the second method of analysis. An illustration of this method can be seen in Figure 5. Fitting to a background emission subjectively determined to fall in the valleys of what appears to be a C_2 emission spectrum, a black body curve fit of between 5000 and 5500 Kelvin seems reasonable.

Clearly, additional analysis needs to be done using both methods to determine there consistencies and inconsistencies in various parts of the temporal temperature profile. Such analysis will provide excellent opportunities for my undergraduate students to engage in this research. Results of this additional analysis will be reported to the nanotube team through student presentations and written reports.

Experiment 2. Absorption measurements.

Results of our absorption studies indicate a strong flat absorbance with a fairly linear bias toward shorter wavelengths. Although this absorption does not appear to be highly structured information on the absorbing material may be found by thorough analysis. However, even without a thorough analysis of the data a few preliminary results may be given.

It is clear that material begins to absorb within a few microseconds of ablation. The amount of absorption decreases after its initial peak to reach a fairly constant level after approximately one hundred microseconds. This level of absorption remains constant until the next laser pulse perturbs the system. All of our data was taken under sixty Hertz operation conditions. Figure 6 shows the relation of transmission with time of ablation.

There appears to be a great deal more absorbance in front of the target than behind the target. This may be because the ablated materials remain longer near their turning point. They are ablated from the target with the particles decreasing in velocity as they are slowed by the incoming buffer gas coming. The particles eventual stop and begin accelerating in the other direction to blow by the target. By time they have traveled past the target they will have gained a good deal of velocity, aggregated into clusters, or deposited out onto the walls of the quartz tubing. No absorption is observed at distances of 3 cm or more in front of the target surface and saw a much lesser amount of absorption behind the target.

A target without metal catalyst was also used under the same conditions as the standard production target. This 'blank' target also produced a fairly steady level of absorbance with a flat short wavelength biased spectrum. However, at early time delays, there is some difference in the spectrum obtained using the standard and the 'blank' target as can be seen in Figure 7.

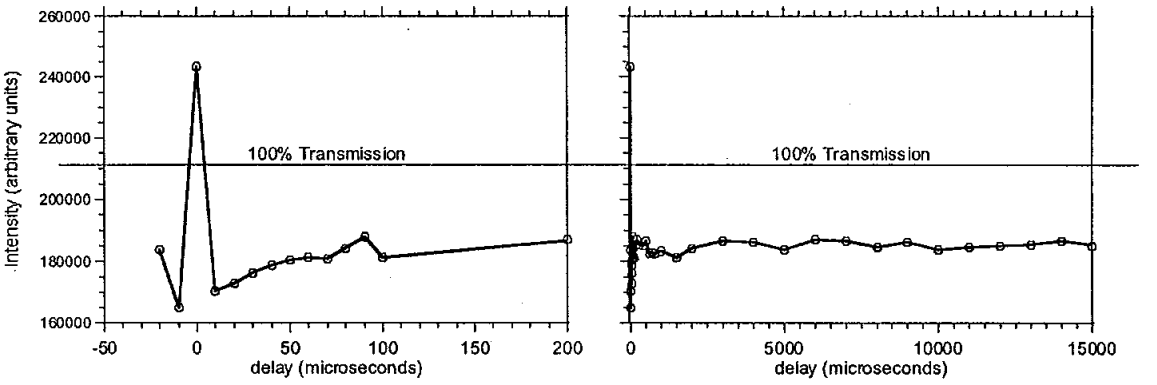


Figure 6: Temporal profile of transmission with respect to laser pulse. The profile shows a fairly constant transmission from some hundred microseconds after the laser pulse to 15 milliseconds after the laser pulse at a position 2 cm in front of the target. There are 16.7 milliseconds between laser pulses when operating at 60 Hertz.

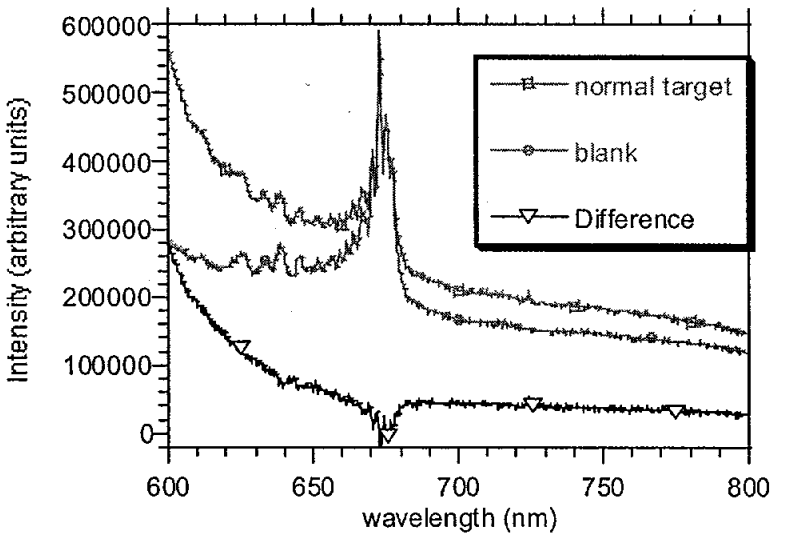


Figure7: Difference in emission during time immediately following ablation of target. The standard target exhibits additional emission at the shorter wavelengths.

The standard target appears to exhibit additional emissions at shorter wavelengths than does the 'blank' target. If the "blank" emission is assumed to be resulting from C_2 emissions, the emissions observed when using the standard target must be due to other species than C_2. It may be that the ablations lasers are acting as probes. It is known that nanotubes absorb the shorter 532 nm wavelengths of the 'green' Nd:YAG laser, although

one may expect that emission associated with that absorption should occur on a much faster time scale. The use of a 532 nm Nd:YAG laser as a possible probe laser in future experiments should be considered.

CONCLUSIONS

This summer's work has focused on a determination of target surface temperature as a function of ablation parameters and on the development of a method to measure absorption of species ablated from the target in spatial and temporal dimensions.

Emission from the target surfaces was measured and an initial analysis of that data appears to show a surface temperature in the range of 3000-5500 K under standard production conditions. Temporal temperature profiles under many different parametric conditions were taken. The data needs further analysis and once confidence in the results is obtained, they may be incorporated into other theoretical models of ablation and plume dynamics.

Absorption measurements were taken during carbon nanotube production that indicates a great deal of material is present at all times within the standard 16 msec window (Lasers run at 60 Hz). This absorption does not have any clear absorption features, but may have some wavelength dependence that may be useful when further analyzed. Also, there are indications that suggest other methods for probing carbon nanotubes during production.

There are some common difficulties with both of these experiments which should also be considered when planning for future studies. One of these difficulties is the changing of the target surface due to the 'pitting' of the target as material from the center of the target is ablated away while material outside the area of the laser spot remains. Spectra taken in the beginning of a run and hours later after significant pitting has occurred can be very different. A method to avoid pitting needs to be developed before reproducibility of spectra can be obtained over longer periods of time.

A second difficulty involves the coating of optical components with carbonaceous deposits. Deposits on the lenses decrease the transmission of the light source. This is a problem similarly encountered when using the production tube for diagnostics, but to a lesser degree. The optics of the y-tube are much less affected by these deposits than are the optics of the x-tube, probably because they are upstream and farther from the target. A method for introducing the buffer gas through the side arms of the x-tube or the design of longer arms on the x-tube which can incorporate longer lens may also be beneficial. A smaller hole than the current half inch hole in the 1 inch tube at the joint with the sidearm may also help prevent material from depositing on the lenses and would also help maintain a flow within the larger one inch tube.

Although progress has been made in developing methods of probing nanotubes during production, there are still other factors that also have remained elusive to the scientist's probing. Metals are thought to play a role as atoms in a 'scooter' mechanism but also as larger nanoparticles or clusters in a 'root' mechanism. Knowledge of the presence of the metal atoms and the metal clusters would help determine the plausibility of the two mechanisms. Although work has been done to follow the metal atoms, none has been done to detect *in situ* the metal nanoparticles. It would also be interesting to follow the progress of fullerenes with and without the presence of the metal catalysts during the formation of carbon nanotubes.

Work in carbon nanotubes has made great progress in the last few years. It is exciting to see that ideas that were only exploratory a few years ago have matured into rigorous scientific and engineering projects. Those working in the field today have a much firmer grasp of the issues, properties, challenges, and promise than they did just a few years ago. As the field of carbon nanotubes gathers momentum, it will continue to deliver new materials and applications beyond current imagination. NASA is well situated to take full advantage of these material advances. It has been great adventure for this author to be a small part of this project.

ACKNOWLEDGEMENTS

The faculty fellow would like to acknowledge the assistance of William Holmes, the JSC nanotube production laboratory supervisor, for his technical and creative assistance in performing these experiments. Sivaram Arepalli, Carl Scott, and Pasha Nikolaev must be recognized for their participation in the planning stages of these experiments. Leonard Yowell, as the project leader, should be acknowledged for his direction in managing the group and allocating the human resources, the equipment, and the time to make this work possible.

REFERENCES

1. Iijima, S., Nature, 1991. **354**: p. 56-58.
2. Smalley, R.E., et al., Nature, 1985. **318**: p. 162-163.
3. Yudasaka, M., T. Ichihashi, and S. Iijima, J. Phys. Chem. B, 1998. **102**: p. 10201-10207.
4. Yudasaka, M., et al., J. Phys. Chem., 1998. **102**: p. 4892.
5. Yudasaka, M., et al., Journal of Physical Chemistry, 1999. **103**: p. 3576-3581.
6. Asaka, S. and S. Bandow, Physical Review Letters, 1998. **80**(17): p. 3779-3782.
7. Arepalli, S. and C.D. Scott, Chemical Physics Letters, 1998. **302**: p. 139-145.
8. Arepalli, S., et al., Applied Physics A, 2000. **70**: p. 125-133.
9. Arepalli, S., et al., Appl. Phys. Lett., 2001. **78**: p. 1610-1612.
10. DeBoer, G., et al., J. Appl. Phys., 2001. **89**(10): p. 5760-5768.
11. Puretzky, A.A., et al., Appl. Phys. Letts., 2000. **76**(3): p. 182-184.
12. Puretzky, A.A., et al., Appl. Phys. A, 2000. **70**: p. 153-160.
13. Scott, C.D., et al., Appl. Phys. A., 2002. **74**: p. 11.
14. Bronikowski, M.J., et al., J. Vac. Sci. Technol. A, 2001. **19**(Jul/Aug): p. 1800-1805.
15. Smalley, R.E., et al., Science, 1996. **273**(July 26): p. 483-487.
16. Mintmire, J.W. and C.T. White, Carbon, 1995. **33**(7): p. 893-902.
17. O'Connel, M.J., et al., Science, 2002. **297**: p. 593-596.
18. Bachilo, S.M., et al., Science, 2002. **298**: p. 2361-2366.
19. Lefebvre, J., Y. Homma, and P. Finnie, J. Phys. Chem B., 2003. **107**: p. 1-4.
20. Lebedkin, S., et al., J. Phys. Chem B., 2003. **107**: p. 1949-1956.

Monte Carlo Simulation of Markov, Semi-Markov, and Generalized Semi-Markov Processes in Probabilistic Risk Assessment

Final Report
NASA Faculty Fellowship Program 2004

Johnson Space Center

Prepared by:	Thomas English
Academic Rank:	Professor
University & Department:	College of the Mainland Department of Mathematics Texas City, TX 77591
NASA/JSC	
Directorate:	Safety and Mission Assurance
Division:	Advanced Programs and Analysis Division
JSC Colleague:	Richard P. Heydorn, PhD
Date Submitted:	August 10, 2004
Contract Number:	NAG 9-1526

INTRODUCTION

A standard tool of reliability analysis used at NASA-JSC is the event tree. An event tree is simply a probability tree, with the probabilities determining the next step through the tree specified at each node. The nodal probabilities are determined by a reliability study of the physical system at work for a particular node. The reliability study performed at a node is typically referred to as a fault tree analysis, with the potential of a fault tree existing for each node on the event tree.

When examining an event tree it is obvious why the event tree/fault tree approach has been adopted. Typical event trees are quite complex in nature, and the event tree/fault tree approach provides a systematic and organized approach to reliability analysis.

The purpose of this study was two fold. Firstly, we wanted to explore the possibility that a semi-Markov process can create dependencies between sojourn times (the times it takes to transition from one state to the next) that can decrease the uncertainty when estimating time to failures. Using a generalized semi-Markov model, we studied a four element reliability model and were able to demonstrate such sojourn time dependencies. Secondly, we wanted to study the use of semi-Markov processes to introduce a time variable into the event tree diagrams that are commonly developed in PRA (Probabilistic Risk Assessment) analyses. Event tree end states which change with time are more representative of failure scenarios than are the usual static probability-derived end states.

BLOCK DIAGRAM ANALYSIS

Our study begins with a look at a four component reliability block diagram. Figure 1 shows the component block diagram.

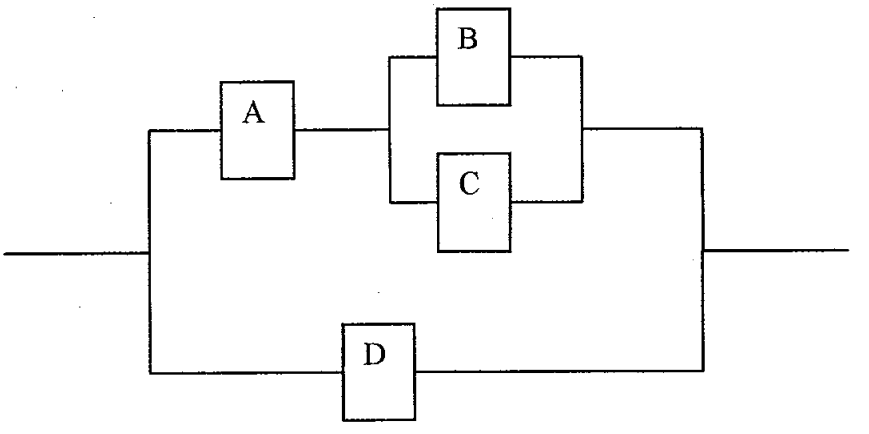


Figure 1: Four Component Block Diagram

This block diagram represents a system in which initially all four components are working independently of each other. Given that there are 4 components in the system, and two possible states for each component (working versus failed), there are a total of 16 possible states in which the system may reside at any given time. These 16 states are related to each other by the flow diagram illustrated in Figure 2. Figure 2 shows the possible paths to overall failure of the system, which is realized at nodes 8, 10, 12, 14, and 16. The 16 nodes of the flow diagram are defined in Table 1. The symbol "**O**" refers to an operating state, and the symbol "**F**" refers to a failed state.

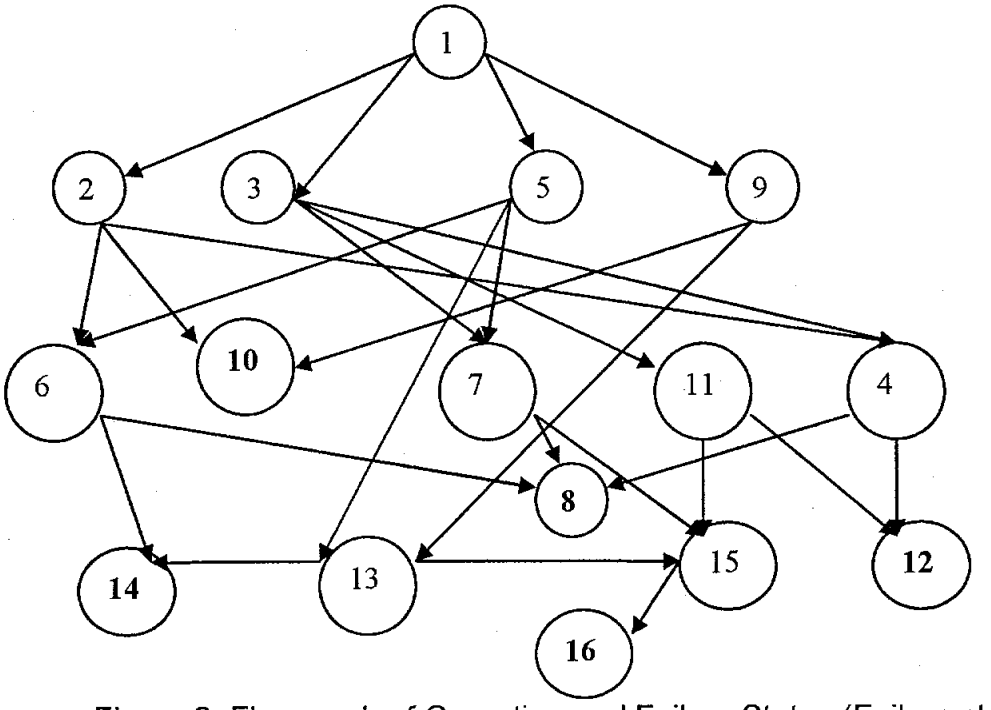


Figure 2: Flowgraph of Operating and Failure States (Failure states in boldface).

TABLE 1: OPERATING AND FAILURE STATES

Component	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
B	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
C	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
D	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
State	O	O	O	O	O	O	O	F	O	F	O	F	O	F	O	F

In effect, this four node diagram has incorporated into it the notion of a generalized semi-Markov process (GSMP), that is, the future node is predicted not only by the present node (as in a Markov Process), but also by a set of time

generators (think stopwatches) running in the present node and indicating the time to transition to a future node. Inherent in a GSMP is the notion of competition among the transition times, with the smallest transition time specifying the future node. A fundamental question is, can we use the node and time components to develop an improved reliability estimate? The first part of our investigation is to explore the concept of path dependence within the flow diagram.

Multiple runs through the flow diagram presented in Figure 2 are simulated using the S programming language with summary statistics presented in Table 2 and Table 3.

Table 2: PATH MEAN TIMES TO FAILURE

	The Weibull Model		
Path	**MTTF**	**S.E.**	**Prob**
1-2-10	0.99	0.01	0.11
1-9-10	1.00	0.01	0.00
1-2-6-14	1.00	0.00	0.12
1-2-6-8	1.00	0.01	0.00
1-2-4-8	0.99	0.01	0.10
1-2-4-12	0.99	0.01	0.10
1-3-4-8	0.99	0.01	0.10
1-3-4-12	0.99	0.01	0.10
1-3-7-8	1.00	0.01	0.00
1-3-11-12	1.00	0.01	0.00
1-5-6-8	1.00	0.00	0.00
1-5-6-14	1.00	0.00	0.00
1-5-7-8	1.00	0.00	0.00
1-5-13-14	1.14	0.11	0.03
1-9-11-12	0.99	0.01	0.00
1-9-13-14	1.13	0.11	0.03
1-3-7-15-16	1.23	0.18	0.11
1-3-11-15-16	1.23	0.18	0.11
1-9-11-15-16	1.22	0.17	0.00
1-5-7-15-16	1.24	0.17	0.00
1-5-13-15-16	1.33	0.18	0.03
1-9-13-15-16	1.33	0.18	0.03
Overall	1.08	0.16	1.00

Table 3: PARTIAL PATH MEAN TMES TO FAILURE

	Weibull Model	
Partial Path	**MTTF**	**S.E.**
1-2-4	0.99	0.01
1-2-6	1.00	0.00
1-2	1.00	0.01
1-9-11-15	1.22	0.17
1-9-13-15	1.33	0.18
1-9-11	1.21	0.18
1-9-13	1.23	0.18
1-9	1.23	0.18
1-3-7-15	1.23	0.18
1-3-11-15	1.23	0.18
1-3-4	0.99	0.01
1-3-7	1.23	0.18
1-3-11	1.23	0.18
1-3	1.12	0.18
1-5-7-15	1.24	0.17
1-5-13-15	1.33	0.18
1-5-6	1.00	0.00
1-5-7	1.23	0.17
1-5-13	1.23	0.18
1-5	1.14	0.18

In this simulation the time to failure of each component is assumed to have a Weibull distribution, with components A and B having their time to failure concentrated at a point by using a shape parameter of 200, while components C and D have their time to failure spread out by using a shape parameter of 3. The purpose for such a choice of shape parameters is to model a system where a specific path occurs with inter-arrival times that are almost fixed, while other

paths contain inter-arrival times with greater variation. This is an attempt to model real-world events where success at a node occurs at precise points in time, while failure can occur over a broader interval of time. The overall system time to failure is estimated by a weighted mixture of each path's time to failure.

Of interest is the apparent existence of path dependence within Table 2. What appears to be true from Table 2 is that the shorter paths have shorter mean time to failure. Table 3 is a look at the information gained by knowing the nodes passed through as one traverses the flow diagram. Note that passing through node 2 virtually guarantees a mean time to failure of the system of 1. Unfortunately, the standard error associated with the mean time to failure along any path is sufficiently large as to mask general distinctions among time to failure along arbitrary paths.

We examine the correlation among inter-arrival times T_1, T_2, T_3, T_4, and the time to failure TTF, by means of correlation matrices. Inter-arrival times (sojourn times) represent the times the process resides at a particular node before transitioning to the next node. In effect, the inter-arrival times for a particular path through the flow diagram are the differences between the failure times of components failing in sequence, with the first inter-arrival time being the time to failure of the first component. Hence, the time to failure (TTF) for a given path is the sum of the inter-arrival times along the path. Since all paths for the given reliability block diagram must have at least two component failures for the system to fail, there are at least two inter-arrival times on every path. A three by three matrix shows the correlations among the inter-arrival times along all paths (*i.e.* for paths of length 2, 3, or 4),, as well as the time to failure:

$$corr(T_1, T_2, TTF) = \begin{bmatrix} 1 & -.83 & .50 \\ & 1 & -.27 \\ & & 1 \end{bmatrix}.$$

A four by four matrix shows the correlations among the inter-arrival times along all paths containing at least three inter-arrival times (*i.e.* for paths of length 3 or 4),, as well as the time to failure:

$$corr(T_1, T_2, T_3, TTF) = \begin{bmatrix} 1 & -.81 & -.29 & .58 \\ & 1 & -.26 & -.23 \\ & & 1 & -.47 \\ & & & 1 \end{bmatrix}.$$

A five by five matrix shows the correlations among the inter-arrival times along all paths containing four inter-arrival times (*i.e.* for paths of length 4), as well as the time to failure:

$$corr(T_1, T_2, T_3, T_4, TTF) = \begin{bmatrix} 1 & -1 & .39 & -.04 & .78 \\ & 1 & -.39 & .04 & -.78 \\ & & 1 & -.10 & .64 \\ & & & 1 & .43 \\ & & & & 1 \end{bmatrix}.$$

Of interest in these correlation matrices is the exponential decay of correlation that occurs between inter-arrival times. In the last matrix we see that T_4 shares no correlation with the previous inter-arrival times. This loss of correlation will have an interesting consequence in the following section.

In an attempt to utilize the correlations present in the inter-arrival times and the time to failure (TTF) of the system, we construct a series of linear regression models and present the results in Table 4.

Table 4: PREDICTIVE MODELS

Model 1: TTF ~ T.1

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	2e+000	8e-003	2e+002	0e+000
T.1	2e+000	1e-002	2e+002	0e+000

Residual standard error: 0.8 on 99998 degrees of freedom
Multiple R-Squared: 0.3

Model 2: TTF ~ T.1 + T.2

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	0.12	0.02	7.40	0.00
T.1	3.35	0.02	186.96	0.00
T.2	1.94	0.02	96.04	0.00

Residual standard error: 0.7 on 99997 degrees of freedom
Multiple R-Squared: 0.3

Model 3: TTF ~ T.1 + T.2 + T.3

	Value		Std. Error	t value	Pr(>|t|)
(Intercept)	-2.04	0.03	-66.64	0.00	
T.1	5.62	0.03	189.95	0.00	
T.2	4.65	0.03	137.68	0.00	
T.3	1.70	0.03	55.98		0.00

Residual standard error: 0.5 on 88225 degrees of freedom
Multiple R-Squared: 0.5

Model 4: TTF ~ T.1 + T.2 + T.3 + T.4

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	0e+000	0e+000	4e+000	0e+000
T.1	4e+000	0e+000	4e+014	0e+000
T.2	3e+000	0e+000	3e+014	0e+000
T.3	2e+000	0e+000	2e+015	0e+000

ABSTRACT

Most probabilistic risk assessment (PRA) and reliability methods commonly used at Johnson Space Center (JSC) make the assumption that component failures in a system are independent random occurrences. There are some exceptions (e.g. modeling common cause events), but because of the mathematical complications that occur when full dependency is assumed, it is not done by the standard models.

This study investigates the use of models in which dependencies among the failure states has been considered via a variety of processes. Our study included: 1) analysis of a general block component diagram for path dependence and inter-arrival time correlations; 2) analysis of correlation among inter-arrival times on a small, generic event tree; 3) a semi-Markov approach designed to provide updated reliability predictions for general event trees.

T.4	1e+000	0e+000	3e+015	0e+000

Residual standard error: 1e-014 on 29902 degrees of freedom
Multiple R-Squared: 1

Model 5: T.2 ~ T.1

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	7e-001	1e-003	6e+002	0e+000
T.1	-7e-001	2e-003	-5e+002	0e+000

Residual standard error: 0.1 on 99998 degrees of freedom
Multiple R-Squared: 0.7

Model 6: T.3 ~ T.1 + T.2

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	9e-001	1e-003	7e+002	0e+000
T.1	-9e-001	2e-003	-6e+002	0e+000
T.2	-1e+000	2e-003	-6e+002	0e+000

Residual standard error: 0.06 on 88119 degrees of freedom
Multiple R-Squared: 0.8

Table 4 (CONTINUED): PREDICTIVE MODELS

Model 7: T.4 ~ T.1 + T.2 + T.3

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	6e-001	2e-001	4e+000	2e-004
T.1	-4e-001	2e-001	-2e+000	2e-002
T.2	-4e-001	2e-001	-2e+000	2e-002
T.3	1 -2e-001	2e-002	-1e+001	0e+000

Residual standard error: 0.2 on 30151 degrees of freedom
Multiple R-Squared: 0.009

We have considered models in which TTF is regressed upon the inter-arrival times as well as models in which inter-arrival times are regressed on prior inter-arrival times. In models designed to predict TTF, we see in general that the multiple R-squared values are small and hence we gain poor predictive value from the model. The only model which predicts TTF well is model 4, which says knowing all the inter-arrival times allows one to predict TTF. Since TTF is the sum of all inter-arrival times, one hardly finds this regression model useful. In models 5 and 6, we see more promise in gaining predictive ability, with multiple R-squared values improving. In model 7 we see the artifact of the loss of correlation previously mentioned. The "return to randomness" of the inter-arrival times masks our ability to gain predictive power.

We consider another method of extracting predictive power from the simulation in terms of a simulated reliability curve. Figure 3 presents an empirical reliability curve for the data used in the simulation of the block diagram and the empirical mixture pdf of the distributions for the four components.

In Figure 3 we have plotted a 95% confidence interval on the empirical reliability curve, which appears as a very narrow band about the mean reliability curve. This simulation ran 100,000 tests replicated 50 times (hence 5,000,000 path simulations) to generate the lower and upper confidence intervals. The empirical mixture pdf for the block diagram is calculated using data from one of the 50 replications. We see the reliability curve generated from the simulation is quite accurate in terms of the confidence intervals, and can alleviate the difficulty of analytic calculations when the pdf for the block diagram is a mixture (Figure 3) and hence analytically more difficult to work with.

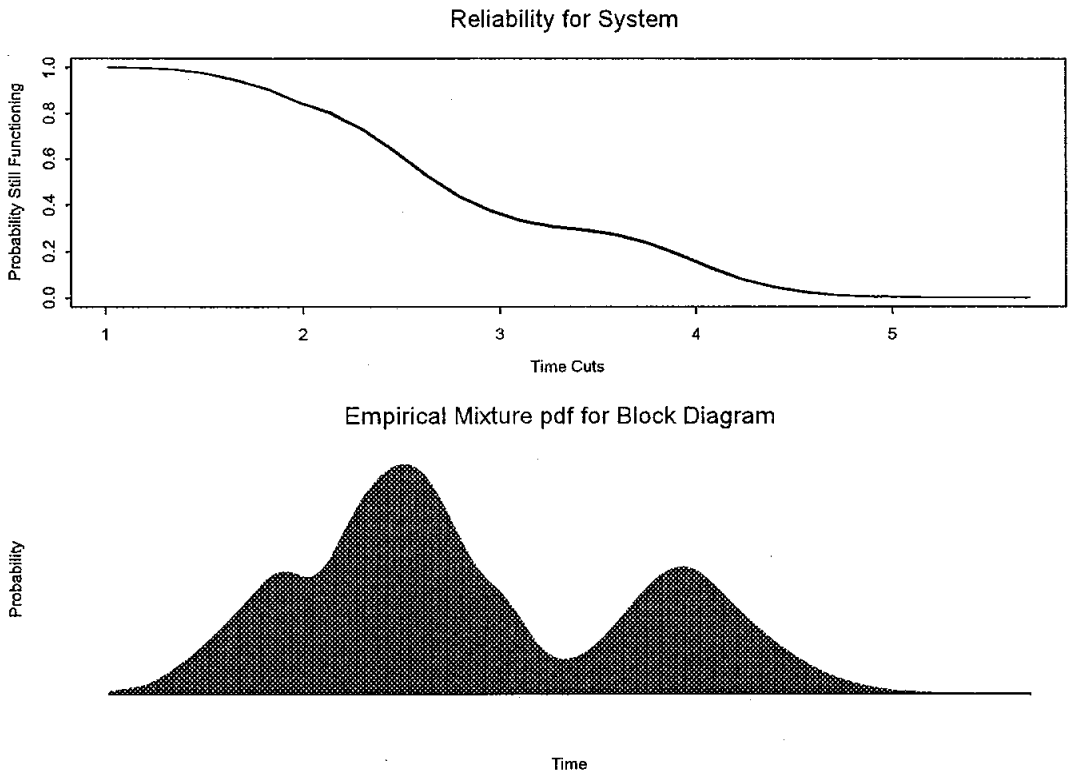


Figure 3: Reliability Curve for Block Diagram; pdf for Block Diagram

SIMPLE EVENT TREES

We began the study of the application of semi-Markov processes to event tees with the simple flow diagram shown in Figure 4.

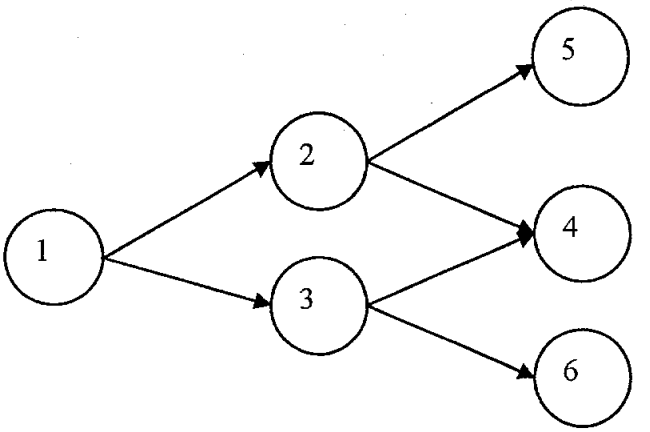


Figure 4: The simple event tree flowdiagram.

Event tree models are such that there are two primary paths through the flow diagram: a success path and a failure path. The transition times for the success path are not random, whereas the transition times for the failure paths are random. The study of correlation among inter-arrival times for the flow diagram shown in Figure 4 is carried out both as a semi-Markov process (SMP) in which the path is chosen by binomial distributions located at each node, while the inter-arrival times are Weibull in nature, and a generalized semi-Markov Process (GSMP) in which case the inter-arrival times compete to transition at each node. In both cases the inter-arrival times for these models do not show dependencies between inter-arrival times. This result indicates that there is a fundamental difference between the block diagram and the event tree, and motivates our desire to try a different approach to analyzing event trees.

A SIMPLE NASA EVENT TREE

We turn our focus to the analysis of a simple NASA-JSC event tree representing a lunar mission (Figure 5, Table 5), and show how Monte-Carlo simulation techniques can provide a more dynamic view of the probability associated with each path, as well as capture underlying information associated with node and inter-arrival time values.

TABLE 5: EVENT TREE TIMES

Node	Mission Events	Time (min)	Explanation of Event Times
1	Booster Launch With Payload	0.167	From engine ignition to clearing the tower
2	Booster Ascent With Payload	8.5	Tower clear to engine cutoff
3	Launch Abort	10	From abort declaration to descent (abort declaration could be anywhere during ascent)
4	Payload Orbit Insertion	2	Orbital engine burn time
5	Mission In Orbit	7200	

6	Mission Abort And Return	90	Maximum time from declaration of abort to orbital engine burn
7	Deorbit Burn	2	Orbital engine burn time (deorbit)
8	Vehicle Entry	60	Vehicle entry from engine burn to below Mach 1
9	Vehicle Descent	10	Mach 1 to final approach
10	Vehicle Landing	2	Final approach, landing, and rollout

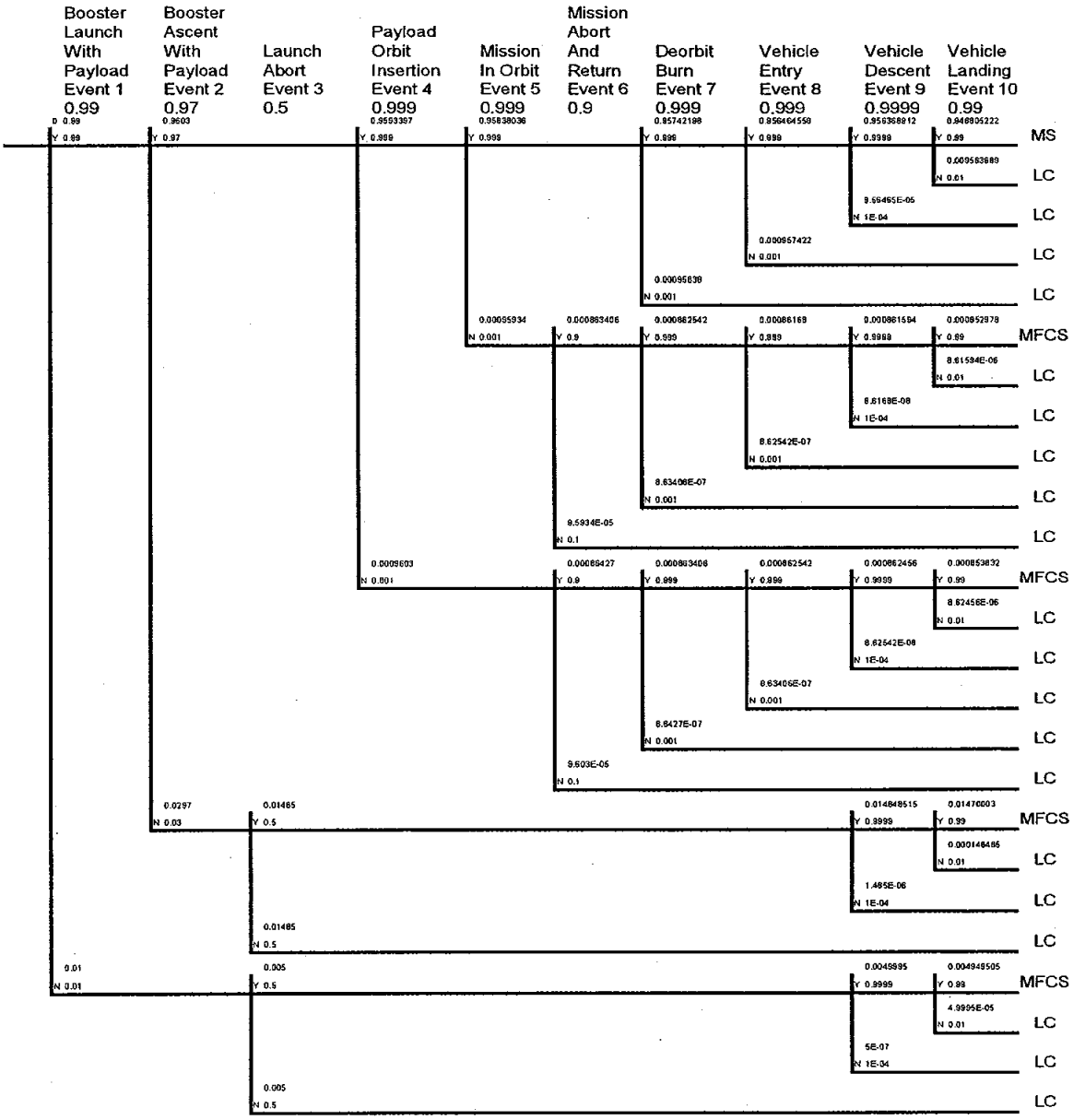


Figure 5: A simple NASA event tree; MS = mission success, LC = loss of crew, MFCS = mission fails, crew survives.

The interpretation of the event tree as a flowgraph results in figure 6.

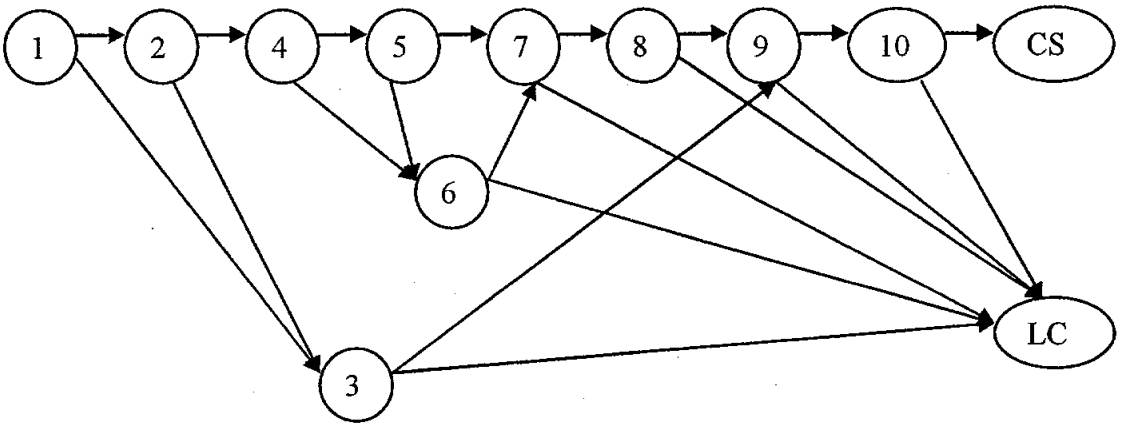


Figure 6: Flowgraph of Event Tree

The 25 paths (Table 6) associated with the flowgraph of the event tree are simulated in a process by which the probabilities provided on the tree diagram are used to calculate the number of simulations needed for each path. Note that 12 millions simulations are required to ensure that the relatively unlikely paths (path 8 and path 14) will be represented at least once in the simulation.

TABLE 6: PATHS THROUGH THE FLOW DIAGRAM

Path	Nodes										Number Simulations
1	1	2	4	5		7	8	9	10	CS	11,361,663
2	1	2	4	5		7	8	9	10	LC	114,764
3	1	2	4	5		7	8	9		LC	1,148
4	1	2	4	5		7	8			LC	11,489
5	1	2	4	5		7				LC	11,501
6	1	2	4	5	6	7	8	9	10	CS	10,236
7	1	2	4	5	6	7	8	9	10	LC	103
8	1	2	4	5	6	7	8	9		LC	1
9	1	2	4	5	6	7	8			LC	10
10	1	2	4	5	6	7				LC	10
11	1	2	4	5	6					LC	1,151
12	1	2	4		6	7	8	9	10	CS	10,246
13	1	2	4		6	7	8	9	10	LC	103
14	1	2	4		6	7	8	9		LC	1
15	1	2	4		6	7	8			LC	10
16	1	2	4		6	7				LC	10
17	1	2	4		6					LC	1,152

18	1	2	3	9	10	CS	176,400
19	1	2	3	9	10	LC	1,782
20	1	2	3	9		LC	18
21	1	2	3			LC	178,200
22	1		3	9	10	CS	59,394
23	1		3	9	10	LC	600
24	1		3	9		LC	6
25	1		3			LC	60,000

Furthermore, the event tree times provided in Table 5 are used to construct shape and scale parameters for the inter-arrival times along each path. The shape parameters are chosen to either be 2 along an event to failure (such as a mission abort) or 200 along an event to success. The scale parameter for inter-arrival times associated with success were taken to be the time to events in Table 5, or ½ the time to events in Table 5 for failure events. This was done in order to concentrate the mass of the success distribution more or less at the point in time at which success is to occur, while failure is more spread out over the entire time interval.

The simulation program uses the inter-arrival times, in conjunction with the present node location to construct the probability mass function (pmf) for the distribution of end states, MS = Mission Success, LC = Loss of Crew, MFCS = Mission Fails, Crew Survives, along 10 equally spaced time slices of the possible inter-arrival times for a given node. Given that there are 10 time slices, we generate 10 pmf's for each of the 10 nodes and display the pmf's as a continuum for each node. For sake of brevity in this report, we only show the results for a particular node – node 1. Figure 6 shows the change in the pmf for the 10 time slices along the inter-arrival times for node 1. We see a re-apportionment of the total probability among the three end states as time evolved. Practically speaking, this means that if we can accurately construct the distributions of failure times along a tree diagram, then this dynamic approach to risk assessment will allow us to update our probability of success based on two observations – the present node, and the time elapsed since entering the present node.

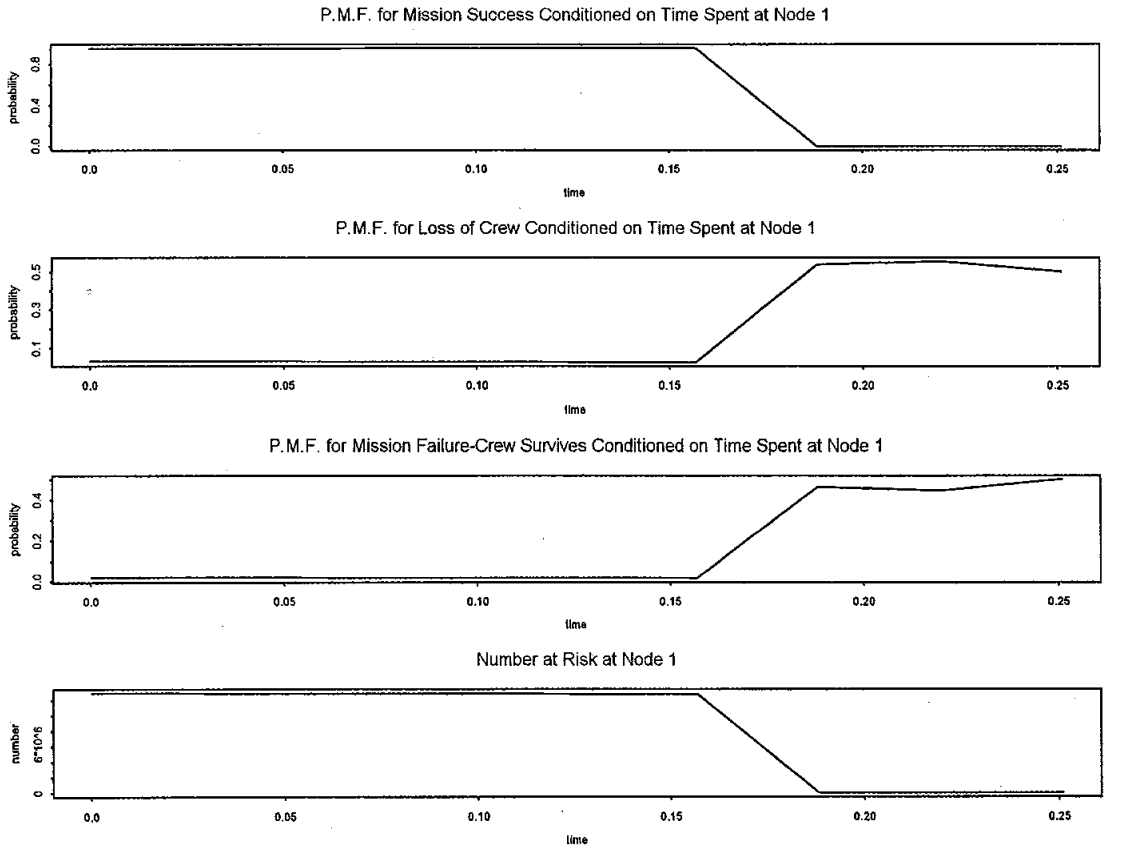


Figure 6: Simulation of pmf for node 1 as time evolves.

The simulation that generated Figure 6 is one in which the probability of ending at a particular node is empirically computed for 10 discrete time intervals, t_i, i = 1, ..., 10, on the range of inter-arrival times for node 1 by considering only those simulations that have not failed at t_i. In effect,

$\Pr(MS \mid t_{i+1} > t > t_i, node\ \ 1) = n(MS \mid t_i)\ \ /\ \ [n(MS \mid t_i) + n(MFCS \mid t_i) + n(LC \mid t_i)],\ \ i = 1,...,9.$

Graphics for the other 9 nodes are constructed in a similar manner.

The simulation performed is somewhat contrived in that it requires the distribution of inter-arrival times be known, and if these inter-arrival times are known, then the pmf can be computed analytically as

$\Pr(MS \mid t_{1,2} > t_i) = \Pr(T > t_{1,2} > t_i) P_{2,4} P_{4,5} P_{5,7} P_{7,8} P_{8,9} P_{9,10} P_{10,MS}$,

where $t_{1,2}$ = the fixed time for the node 2 to be achieved along the path to mission success,

T = the Weibull distributed random variable for the time to failure at node 1,

$P_{i,j}$ = the specified probability of transfer between nodes i and j.

Although contrived, the simulation does carry out two important tasks in that it lays the foundation for a simulation of a GSMP, and it shows the effect of time in the analysis of reliability of the event tree.

The improvement one gets by using the semi-Markov model for an event tree diagram is that the end states now depend upon time. This means that the end state probabilities fluctuate with time and so, for example, if one end state is "loss of crew", then it can happen that the probability that the crew is lost can be high early in the mission, but small late in the mission.

CONCLUSIONS

The study has shown several results of interest to reliability studies at NASA-JSC:

1) Evidence is found to show that correlation between inter-arrival times exists for the general block diagram. This correlation is shown to be difficult to exploit using classical predictive models, and therefore it is suggested that simulation will provide a superior estimate of the distribution of time to failure for the system. It is suggested that each fault tree component of an event tree may be analyzed by the simulation technique, which would provide an empirical pdf for each node of the event tree.

2) Attempts should be made to incorporate the GSMP approach to modeling the event tree. Unless all success events within an event tree are precisely timed and executed such that they effectively have no variance in time, then the GSMP is a more natural modeling assumption.

3) The event tree simulation should be replicated, say 50 times, in order to allow confidence intervals to be placed on the time dependent pmf presented in the study.

REFERENCES

Nilsen, Frode B, 1998. GMSim: A tool for compositional GSMP modeling. Proceedings of the 1998 Winter Simulation Conference. P555-562.

Haas, Peter J. Simulation (Class notes). Retrieved July 9, 2004, from Stanford University website:
http://www.stanford.edu/class/msande223/

Computer Simulation of the VASIMR Engine

Final Report

NASA Faculty Fellowship Program – 2004

Johnson Space Center

Prepared by:	David Garrison
Academic Rank:	Assistant Professor
University & Department	University of Houston-Clear Lake
	Physics Department
	Houston, TX 77058
NASA/JSC	
Directorate:	Space and Life Sciences
Division:	Advanced Space Propulsion Laboratory
Branch:	N/A
JSC Colleague:	John Shebalin
Date Submitted:	August 13, 2004
Contract Number:	NAG 9-1526

ABSTRACT

The goal of this project is to develop a magneto-hydrodynamic (MHD) computer code for simulation of the VASIMR engine. This code is designed be easy to modify and use. We achieve this using the Cactus framework, a system originally developed for research in numerical relativity. Since its release, Cactus has become an extremely powerful and flexible open source framework. The development of the code will be done in stages, starting with a basic fluid dynamic simulation and working towards a more complex MHD code. Once developed, this code can be used by students and researchers in order to further test and improve the VASIMR engine.

INTRODUCTION

Variable Specific Impulse Magneto-plasma Rocket

The Variable Specific Impulse Magneto-plasma Rocket (VASIMR) is a project at the Advanced Space Propulsion Laboratory (ASPL) at JSC [2]. The project is led from NASA JSC, and has contracts with several government research centers, industrial companies and universities. In addition, researchers from universities and institutes all around the world collaborate with ASPL.

The Magneto-plasma rocket engine provides propulsion by ionizing and heating neutral gases to high temperatures and then guiding them out of a magnetic nozzle in order to produce thrust, much like a chemical rocket engine. However, the essential difference between VASIMR and a chemical rocket engine is that VASIMR will produce very high specific impulse at relatively low thrust (*i.e.*, a low density, high velocity exhaust), while a chemical rocket engine produces high thrust at relatively low specific impulse (*i.e.*, a high density, low velocity exhaust).

The particular niche filled by VASIMR in the electric propulsion community is that of a relatively high-power plasma propulsion system that is focused on human space flight, rather than on less massive unmanned, robotic space flight missions. The efficiency of the engine permits a favorable ratio of payload mass to spacecraft mass, one that allows long-duration space exploration missions to be realistically contemplated.

In its research configuration, VASIMR utilizes four co-axial magnetic coils and two co-axial antennas to achieve its purpose. The first antenna is a so-called helicon antenna, which serves as a plasma generator in that it ionizes an injected neutral gas (typically hydrogen, deuterium or helium). The second antenna is known as the ion cyclotron resonance heating (ICRH) antenna and it boosts the energy of the plasma by feeding electromagnetic energy preferentially into the ions.

While the helicon antenna is primarily responsible for creating the plasma, the second antenna is used to increase the ion energy and exhaust velocity, and thus the specific impulse of the rocket engine. The magnetic coils work in concert to shape the strong axial magnetic field that guides the strongly magnetized plasma (*i.e.*, magneto-plasma). The final magnetic coil (or a smaller auxiliary coil) serves as a magnetic nozzle, by which the specific impulse and thrust of the plasma exhaust may be varied. When the components are operating together, the result is the Variable Specific Impulse Magneto-plasma Rocket, or VASIMR.

The magnetic nozzle gives VASIMR the unique ability to modulate the plasma exhaust so as to maintain maximum power and efficiency. This technique is termed "Constant Power Throttling" and is similar to adjusting the transmission on an automobile. The VASIMR engine (specific impulse, $I_{sp} \sim 15{,}000$ sec), is designed to run continuously, so that, although it has low thrust, any interplanetary transit time is considerably reduced. In contrast, a chemical rocket, such as the space shuttle main

engine (I_{sp} ~ 450 sec), is designed to provide very high thrust, but only for about eight minutes. A traditional chemical rocket lifts a space ship off of a planet and gives it an initial velocity, after which it is in free flight towards its objective.

The role of VASIMR is to provide thrust during what would have been unpowered free flight, thereby shortening travel time. For example, using only a chemical rocket would give a transit time of about 300 days to reach Mars. Adding VASIMR for the interplanetary section of the journey (equipped with a nuclear power generation system) would reduce the trip to as little as 39 days carrying 20 tons of cargo, or 115 days for a larger 61-ton cargo load. Also, by minimizing transit time, physical stress and risk to the crew is also minimized.

Our goal in this project is to create a computerized model of the VASIMR system in order to understand the fluid dynamics and thermodynamics of plasma flow in the engine and in its exhaust [4,5]. This model will incorporate variations of such system parameters as magnetic coil current values and magnetic field structure. We found Cactus to be the best framework for developing these models.

Cactus

Cactus [1] is an open source problem-solving environment designed for scientists and engineers. The Cactus framework, which was originally developed for numerical relativity research, has become an extremely powerful and flexible tool. Cactus originated in the academic research community, where it was developed and used over many years by a large international collaboration of physicists and computational scientists. Its modular structure easily enables parallel computation across different architectures and collaborative code development between different groups.

The name Cactus comes from the design of a central core (or "flesh") that connects to application modules (or "thorns") through an extensible interface[1]. Thorns can implement custom developed scientific or engineering applications, such as computational fluid dynamics. Other thorns from a standard computational toolkit provide a range of computational capabilities, such as parallel I/O, data distribution, or checkpointing.

Cactus runs on many architectures. Virtually all Unix based systems as well as Windows NT are supported. Applications, developed on standard workstations or laptops, can be seamlessly run on clusters or supercomputers. Cactus provides easy access to many cutting edge software technologies being developed in the academic research community, including the Globus Metacomputing Toolkit, HDF5 parallel file I/O, the PETSc scientific library, adaptive mesh refinement, web interfaces, and advanced visualization tools.

[1] See Appendix A

We chose to use Cactus because it is flexible, modular and well documented. The Cactus development groups are quick to respond to questions and communication within the development community is freely available. Also, efforts by the Cactus organization as well as third party developers ensure that new features and bug fixes are constantly being developed [3].

GOALS

The goal of this Faculty Fellowship Program (FFP) project was to test the feasibility of using the Cactus framework to develop a magneto-hydrodynamic (MHD) code for use with the VASIMR project. There are many differences between the existing Cactus codes used in numerical relativity and the MHD codes used within the VASIMR project. These differences had to be addressed in order to develop VASIMR simulations within the Cactus framework.

An alternative to using Cactus would be to either develop a new MHD code from scratch or to modify existing codes. However, the main motivation for switching to the Cactus framework is to gain the support of existing documentation and a large development community. Unlike existing software, a program developed with Cactus will be relatively easy for short-term workers (such as students) to modify and use because of the well-designed structure of Cactus and its extensive support network and documentation.

The Physics Program at the University of Houston – Clear Lake (UHCL) focuses on a Masters degree in Physics. Our graduate students are required to complete a research project or thesis but typically only have about a year to work on such a project. Existing codes usually require several years to learn enough about the software to modify and use on original research projects and are therefore not useful for short-term student projects. This research program will provide a suitable vehicle for student theses because original work can be completed in just a few months. This will also provide a framework for the controlled evolution of software suitable for ASPL.

Development will be done in stages, starting with a basic fluid dynamic simulation and working towards a more complex MHD code. The fluid code is designed primarily to test the feasibility of installing and running Cactus on ASPL and UHCL machines. The fluid code will eventually evolve into a full MHD code but before that can happen several technological steps must be taken. These steps are outlined in the section titled "Development Thorn". Eventually, this code will then be used by students and researchers to further design and improve the VASIMR engine.

INSTALLATION

The first step of this project involved installing Cactus on each of the development machines and testing them using several existing sample thorns. The three development machines were a dual processor Macintosh G4 machine at UHCL, a Linux Beowulf cluster at ASPL and my personal Macintosh Powerbook G4. Each computer

was already equipped with both Fortran and C compilers but I also added additional visualization tools (xgraph, ygraph and gnuplot) to the Macintosh machines.

The biggest challenge during the installation process was finding the correct configuration for Cactus for each different hardware/software setup. The only way to find the correct configuration for each operating system, compiler and software package was to review the documentation and search through the computer's directory structure for the right parameters. This involved some trial and error and in a few cases, we had to correct a few Unix login files. After a couple weeks of searching for the right configurations, all three machines were compiling and running the example codes well.

The test examples ranged for a simple "Hello World" screen printout to a scalar wave simulation that used Cactus' ability to steer computer simulations through a web browser. These tests proved that all the compilers and tools were working correctly so we could move on to the next step, developing an original thorn.

THE COMFLUID THORN

Instead of jumping right into the development of a full MHD thorn, we thought it would be a good idea to first develop a compressible fluid simulation code which has a similar geometry to the VASIMR engine. This involves using a cylindrical coordinate grid and a set of coupled differential equations representing the number density and velocity of particles in the fluid. This is effectively the same problem as in MHD except that the fluid is not charged and there are no magnetic fields. The comfluid thorn was then developed to further test to concept of using Cactus for fluid simulations. The equations, which it evolved, are given below:

$$\frac{\partial n}{\partial t} + \nabla \cdot \left(n\vec{V} \right)$$

$$\frac{\partial \vec{V}}{\partial t} + \vec{V} \cdot \nabla \vec{V} = -\frac{k_B T}{m} \frac{\nabla n}{n}$$

where all units are MKS, $\vec{V}$ is the particle flow velocity, n is the number density (particle/m^3), m is the atomic mass for the fluid under consideration, mass density $\rho = mn$, T is the fluid temperature (assumed to be constant here) and k_B is the Boltzman constant. In addition the energy and momentum is calculated at each grid point for use in our analysis of the code's performance. It should be noted that the above equations are relatively simple, but are suitable to start with.

The equations were evolved in two dimensions in cylindrical coordinates ignoring the angular direction. Because Cactus is based on a Cartesian grid, we had to write subroutines to calculate gradients and divergences in cylindrical coordinates. We also used periodic boundary conditions to "roll" the Cartesian grid into a cylindrical one. As soon as a cylindrical grid thorn becomes available for Cactus, we plan to implement it into our program.

The code compiled and ran on all three development machines without any platform specific modifications. Slices taken in the radial and axial coordinates where then used for data analysis[2]. The initial data for the system modeled a Gaussian distribution of particles with velocities pointing out towards the radial and axial directions. As time evolved the particle distribution dispersed and the particles disappeared out the edges of the simulation domain. Towards the end of the simulation, boundary value errors begin to appear.

This test revealed two problems with the way the simulation was designed. 1) Further work is needed to increase the stability of the code so that it can run longer before significant errors occur. 2) Customized boundary conditions need to be implemented so that we can make some boundaries reflective (example when the radial direction rho = 0) while others are absorbing. Also, the stability of a finite differenced numerical code such as this depends on several factors such as boundary conditions, grid spacing, time step size and other parameters choices. Future work will involve increasing the stability of the code as well as adding new features to make the simulation more realistic.

In order to coordinate the development of improvements to the code while not destroying the progress that we have already made, we split the code and began work on an advanced "development" version. The "stable" version was saved for later study while the development version is continuously changed to improve stability and experiment with new features.

DEVELOPMENT THORN

Time Integration

The first technique adopted in the development code is the Iterated Crank-Nicholson time integrator. By using a second or higher order time integration technique such as Iterated Crank-Nicholson or Runge-Kutta, we can further increase the stability of the code. These techniques work well in numerical relativity and should work well for our program. These systems work by correcting for small errors, which occur as we evolve the equations from the solution at one time to the next. Instead of the growth in errors depending directly on the time step, they depend on the time step squared. This can decrease error growth by several orders of magnitude without a significant decrease in computational speed.

Boundary Conditions

The stable version of our code currently depends on Cactus' built in boundary conditions. By developing our own boundary condition subroutines, we can reduce errors at the boundary by "tuning" the boundaries to our system. Eventually we can introduce absorbing boundary conditions, which eliminate computational artifacts such as unwanted reflections and further reduce boundary errors. Most importantly, we can choose where to apply reflective and absorbing boundary conditions in order to make our

[2] See Appendix B

simulation more realistic. If we are working in cylindrical coordinates, no information should leave the grid when it passes through rho equals zero.

Spectral methods

Cactus is currently designed to use finite differencing as a method of numerically calculating the derivatives of functions. Spectral methods have been shown to be much more accurate and stable than finite difference methods but more difficult to implement. There is currently an effort to develop a general spectral methods thorn for Cactus. Once it has been released, we can begin testing it and eventually add it to our code.

Adaptive mesh refinement

Adaptive mesh refinement (AMR) is a technique where the grid spacing can change depending on the dynamics of the code. This leads to greater accuracy in parts of the grid where it is needed and less accuracy where it is not. This increases both accuracy and computational efficiency. There is currently a third party Cactus Thorn that adds AMR to Cactus.

Other improvements

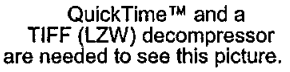
There are several other improvements that can be made to the code including improved initial conditions, the addition of dissipative terms, viscosity, temperature variations in the fluid and much more. These improvements can be added as needed, however, the focus of the code will be to test the concept of using Cactus for VASIMR research and then to develop a MHD code.

Add MHD equations

The long-term goal of this project is to add the MHD equations and turn this fluid dynamic code into a full MHD code [4,5]. This will involve adding a several more evolution equations to the list of coupled differential equations. These include equations for charge density, magnetic field, and the energies and momentum carried by both. There is an additional difficulty at this point in understanding the dynamics of how these equations are evolved and making them as stable as possible. Because of this it is to our advantage to develop a modular, well documented and easy to understand code so that future students can add equations with minimal intimidation.

APPENDIX A

Cactus program structure



APPENDIX B

Preliminary Numerical Results

The energy of the compressible fluid flows to the boundary and disappears, boundary errors develop. For both plots: Blue = early times, Red = late times, x = radial, z = axial. Both plot where produced with ygraph.

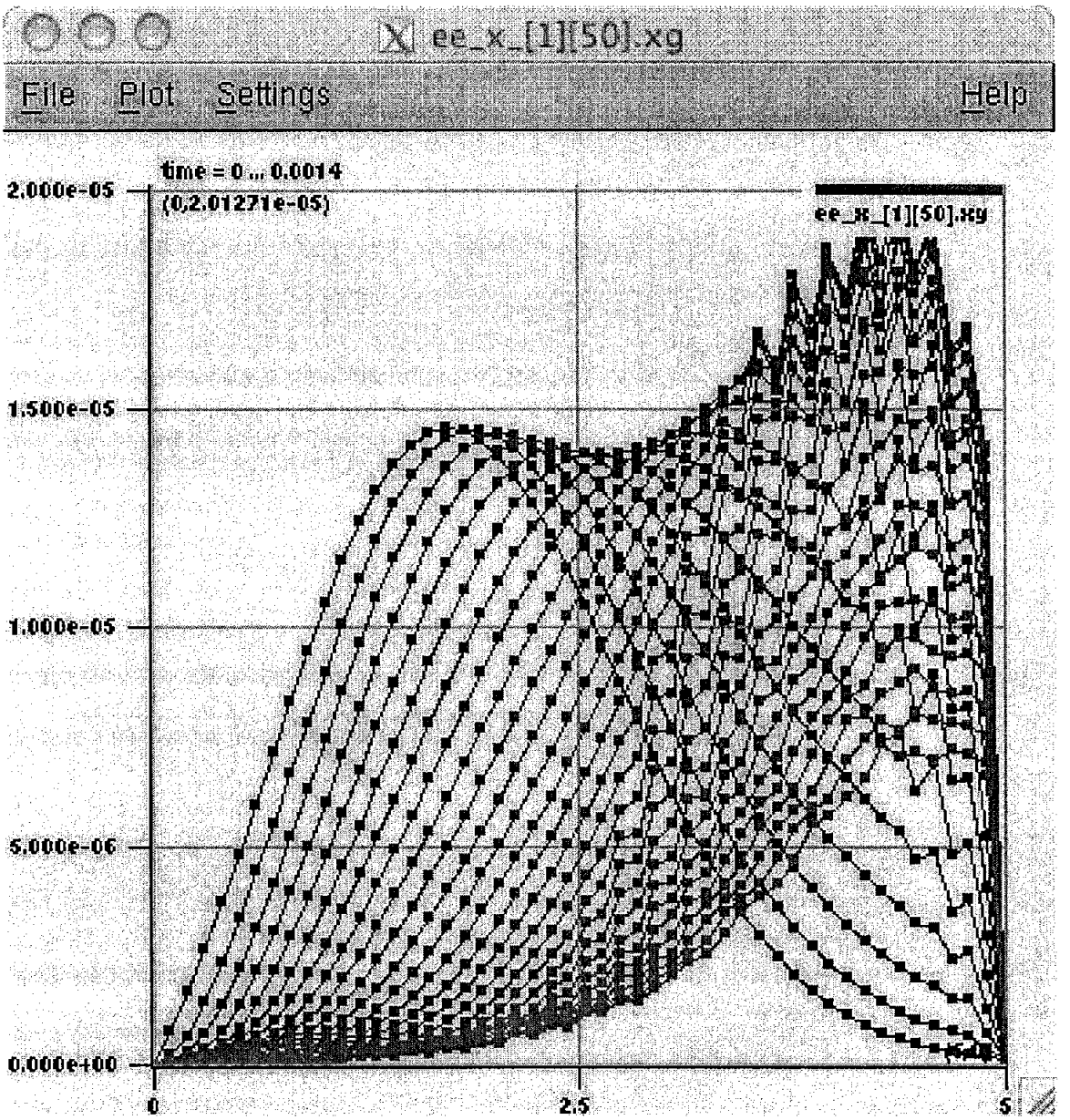


Figure 1 : Above is a plot of energy vs. radial position taken at several times. The y axis is energy amplitude while the x axis shows radial position.

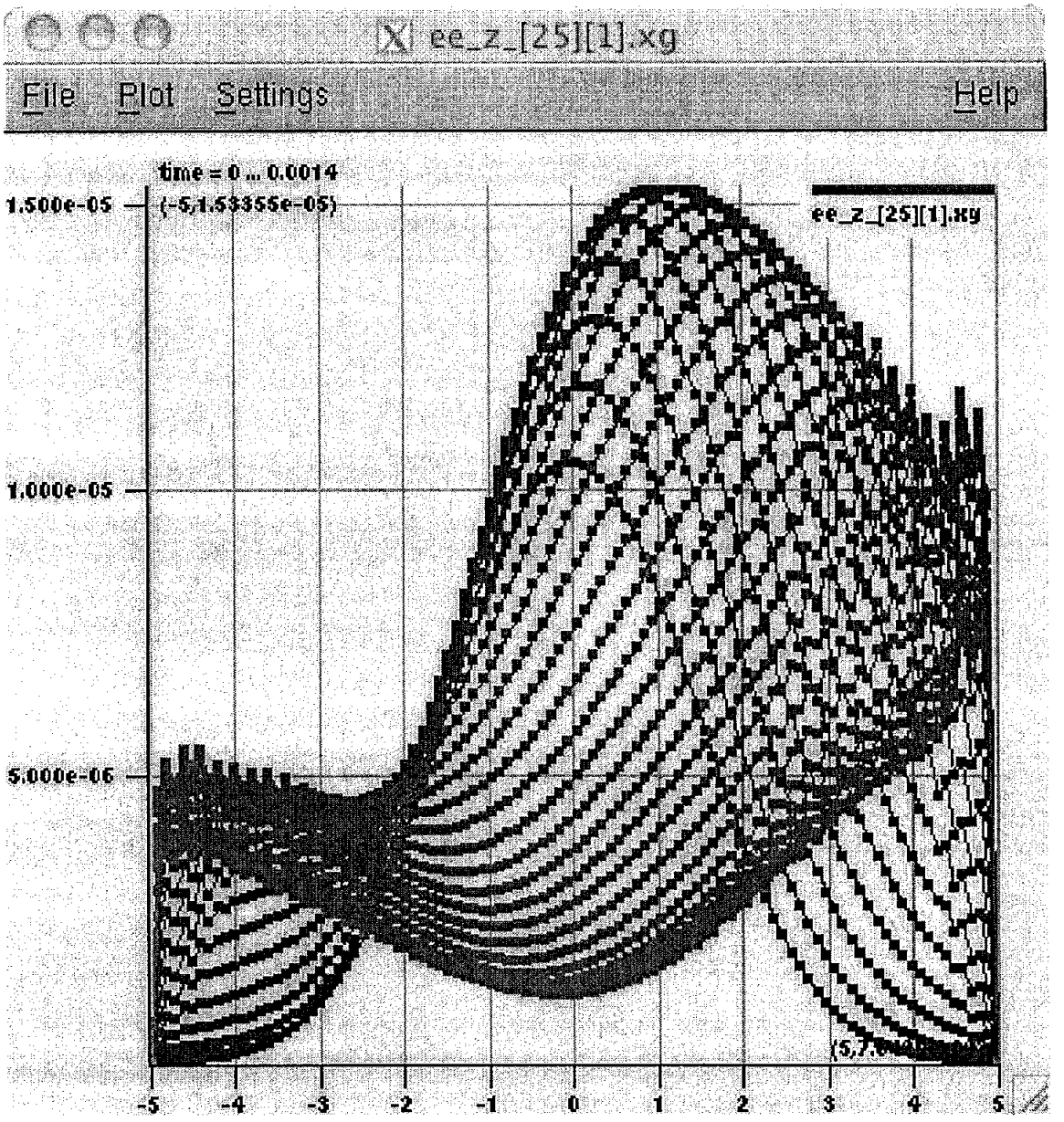


Figure 2 : Above is a plot of energy vs. axial position taken at several times. The y axis is energy amplitude while the x axis shows axial position.

REFERENCES

1. Cactus, **http://www.cactuscode.org**

2. Chang Diaz F. R. (2000) ***The VASIMR Rocket***, Scientific American, 283, (5), 90-97.

3. Goodale, Tom, ***Cactus 4.0: An Introduction and Perspectives On Future Plans***, PowerPoint presentation.

4. Tarditi, A. G. and J. V. Shebalin, ***MHD simulation of flow through the VASIMR magnetic nozzle***, APS Division of Plasma Physics Meeting, Oct. 2003.

5. Tarditi, A. G., J. V. Shebalin, and E. A. Bering, ***MHD simulation of the exhaust plume in the VASIMR advanced propulsion concept***, XXIII General Assembly of the International Union of Geodesy and Geophysics, IUGG2003, Sapporo, Japan, June 2003

Real-Time Analysis of Electrocardiographic Data for Heart Rate Turbulence

Final Report

NASA Faculty Fellowship Program – 2004

Johnson Space Center

Prepared By:	E. Carl Greco, Jr., Ph.D.
Academic Rank:	Associate Professor
University and Department:	Arkansas Tech University Department of Electrical Engineering Russellville, AR 72801
NASA/JSC	
Directorate:	Space and Life Science
Office:	Human Adaptation and Countermeasures
Mail Code:	SK3
JSC Colleague:	Todd T. Schlegel, M.D.
Date Submitted:	August 13, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

Episodes of ventricular ectopy (premature ventricular contractions, PVCs) have been reported in several astronauts and cosmonauts during space flight. Indeed, the "Occurrence of Serious Cardiac Dysrhythmias" is now NASA's #1 priority critical path risk factor in the cardiovascular area that could jeopardize a mission as well as the health and welfare of the astronaut. Epidemiological, experimental and clinical observations suggest that severe autonomic dysfunction and/or transient cardiac ischemia can initiate potentially lethal ventricular arrhythmias. On earth, Heart Rate Turbulence (HRT) in response to PVCs has been shown to provide not only an index of baroreflex sensitivity (BRS), but also more importantly, an index of the propensity for lethal ventricular arrhythmia. An HRT procedure integrated into the existing advanced electrocardiographic system under development in JSC's Human Adaptation and Countermeasures Office was developed to provide a system for assessment of PVCs in a real-time monitoring or offline (play-back) scenario.

The offline heart rate turbulence software program that was designed in the summer of 2003 was refined and modified for "close to" real-time results. In addition, assistance was provided with the continued development of the real-time heart rate variability software program. These programs should prove useful in evaluating the risk for arrhythmias in astronauts who do and who do not have premature ventricular contractions, respectively.

The software developed for these projects has not been included in this report. Please contact Dr. Todd Schlegel for information on acquiring a specific program.

INTRODUCTION

Heart Rate Turbulence, HRT, is the sinus nodal response following an isolated premature ventricular contraction, PVC. Typically a short initial acceleration in heart rate immediately follows the PVC's compensatory pause. This initial acceleration is then later followed by a deceleration of the heart rate. Figure 1 depicts a typical electrocardiogram and arterial blood pressure response for a normal, healthy individual just prior to and following an isolated PVC.

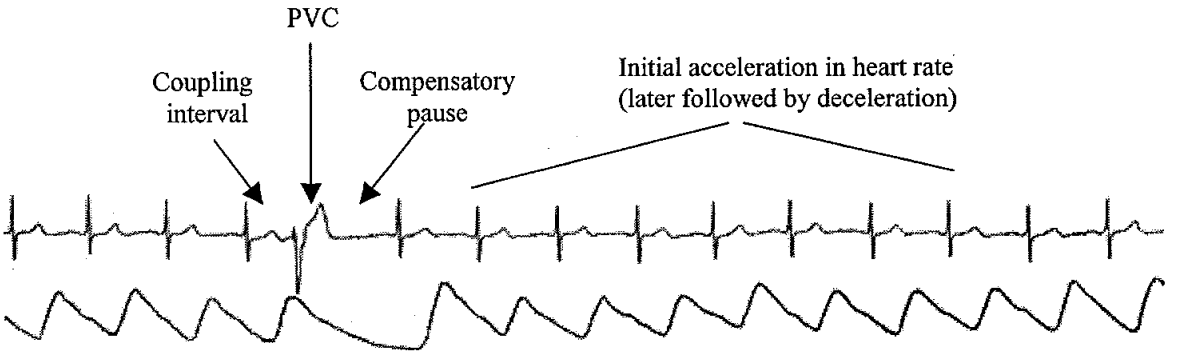


Figure 1: Electrocardiogram (upper trace) and arterial blood pressure preceding and following a single premature ventricular contraction, PVC[2]

Although the underlying mechanisms of HRT have not been fully identified, HRT likely represents a baroreflex response. The premature ventricular contraction causes a brief decrease in the mean arterial blood pressure. When the autonomic control system is intact, the change in arterial blood pressure elicits an instantaneous response in the normally conducted heartbeats that follow the PVC which result in HRT. If the autonomic control system is impaired, this reaction is either weakened or entirely missing. Two parameters have been used to quantify HRT: Turbulence Onset and maximum Turbulence Slope[1]. Turbulence Onset is a measure of the sinus acceleration following single PVC whereas the maximum Turbulence Slope is an indicator of the deceleration phase.

The HRT for a normal healthy individual is shown in Figure 2 and is represented by the RR intervals just preceding and following qualified PVCs. Beats –2 and –1 are the two normal sinus rhythm (NSR) RR intervals just prior to the PVC, beat 0 is the coupling RR interval between the last NSR beat and the PVC, and beat 1 is the compensatory pause RR interval between the PVC and the NSR beat immediately following the PVC. Beats 2 – 16 are all NSR RR intervals. Figure 2 depicts the average of five acceptable separate HRT responses from a total of seven PVC's. An HRT response is deemed to be acceptable if a predetermined number of normal sinus beats preceded and followed the single PVC. In this example, the acceptance criteria were 10 normal beat RR intervals

proceeding and 15 following the PVC. Two of the seven detected PVCs were excluded based on these acceptance criteria. The total number of detected and acceptable PVCs was reported on the graphical display. Straight lines connect the average RR intervals between each consecutive beat relative to the PVCs' average coupling interval. The vertical bold line at each RR interval location spans the standard deviation of that PVC relative RR interval. The horizontal dashed line represents the mean of the RR intervals for the two pre-PVC NSR beats preceding all of the included PVCs.

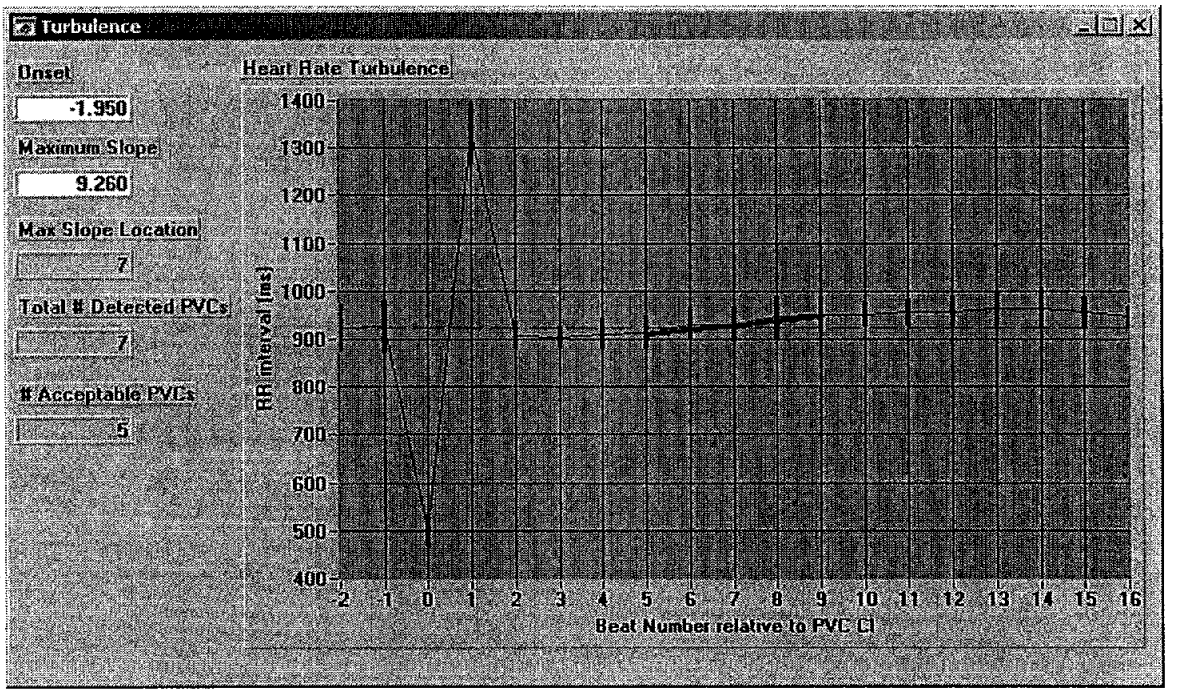


Figure 2: A typical HRT response

Traditional HRT analysis utilized long term data recording for example from 24-hour Holter recordings. The real-time online analysis approach described here evaluated the turbulence response for each identifiable PVC as it occurred.

HRT Parameters

HRT has been characterized by two parameters, Turbulence Onset (TO) and Turbulence Slope (TS) defined as follows:

Turbulence Onset:

$$TO = \frac{(RR_2 + RR_3) - (RR_{-2} + RR_{-1})}{(RR_{-2} + RR_{-1})} * 100\%$$

where RR_{-2} and RR_{-1} represent the two NSR beats intervals immediately preceding the single PVC's coupling interval, RR_0. The RR_2 and RR_3 intervals are the first two NSR beat intervals immediately following the PVC's compensatory pause interval, i.e., RR_1. Turbulence Onset represents the fractional (ms/ms) differential change (expressed in percent) for the two-beat NSR average prior to and following the PVC. TO was calculated for each individual PVC and averaged over all acceptable PVCs[2].

Turbulence Slope:

Turbulence slope was determined from the averaged RR interval response for all PVCs within a patient's record as the maximum slope of a five beat NSR sequence within a 15-beat interval following the PVC. The required normal sinus beat intervals window bracketing each PVC was designated by the operator prior to analysis. Linear regression was applied to each overlapping 5-beat sequence. For example, for the data shown in Figure 2 the slopes were found for the following eleven 5-beat number sequences:

[(2, 3, 4, 5, 6); (3, 4, 5, 6, 7); ... (12, 13, 14, 15, 16)]

and recorded at the center beat number for each sequence, i.e., [4, 5, 6, ..., 14]. The maximum slope for each of these five beat intervals was reported in ms/beat units and represented a measure of the deceleration in heart rate following the single PVC. A straight line drawn through the center point of the sequence with the maximum slope is shown in Figure 2. The maximum slope and its location were reported on the HRT display panel. In the online evaluation of HRT, the HRT display was updated following the prerequisite contiguous interval of normal sinus beats trailing each PVC.

Clinical studies utilizing 24 hr. Holter recordings have found significant predictive risks associated with abnormal turbulence parameters. Schmidt et al[1] used TO > 0% and TS < 2.5 ms/beat to stratify patients into a high risk group. The turbulence slope was found to be more significant of the two parameters for risk assessment. Taken together TO and TS were found to be the best predictors of mortality in their post myocardial infarction patient populations with reduced left ventricular ejection fraction. On the other hand, normal healthy individuals who have PVCs but who do not have structural heart disease have turbulence parameters in the normal range. For example, Diaz et al[3] found TO < 0 for all participants in a study of healthy subjects with PVCs. Values of TO ranged from -1.1% to –11.2%, with a mean of –4.9%. The TS values were not reported in this study.

METHODS

A real-time online HRT parameter estimation routine was developed in the JSC Neurosciences Laboratory to analyze an electrocardiogram obtained from the CARDIAX PC-based computerized ECG system. The CARDIAX system was developed by

International Medical Equipment Developing Co. Ltd. (IMED), Budapest, Hungary, and distributed by CardioSoft, Houston, Texas. The HRT program was written in the C programming language using the CVI system software integration and development environment from National Instruments. Data exchange between the CardioSoft/CARDIAX PC based system and the HRT/HRV application occurred via a named pipe shared memory communication channel as described in the MSDN Library. The HRT code was integrated into a system containing additional heart rate variability analysis routines[4].

The first stage of the HRT application was beat classification based on interval analysis to identify dysrhythmias resulting in significantly altered beat-to-beat intervals. With rare exception, single beat ectopic foci of ventricular origin produce an earlier than a normal conducted beat and result in a delayed compensatory pause recovery beat. The structural evaluation of the P and QRS complexes was not deemed necessary for beat classification requirements for HRT and was therefore not preformed[+]. Analysis of the normal sinus heart rate response that followed an isolated PVC was then performed.

HRT Beat Classification

The first step in the HRT process was the classification of individual beats. Two separate beat classification algorithms were investigated. The first was a modification of the technique developed last summer and was based on the algorithm described on the HRT website[2]. The second procedure was a subset of the one developed by Hamilton and Tompkins[5] and provided on their website[6]. Both beats classification algorithms were based on interval analysis. The HRT algorithm classified each interval as one of the following: normal; PVC coupling interval; PVC compensatory pause; artifact; and unknown. The i-th RR interval, RR_i, was designated a normal RR interval if all of the following conditions were met:

Normal beat requirements:

$$300ms \le RR_i \le 2000ms$$
$$|RR_i - RR_{i-1}| < 200ms$$
$$|RR_i - RA| < 0.20 * RA$$

where RA is the running average of the last 5 normal beats. Traditional classification of PVCs, as defined on the h-r-t website[2], was based on the following procedure: A PVC was defined as a sequence of two consecutive RR intervals wherein the coupling interval is 20% less than the running average and the compensatory pause is 20% greater, or

[+] Beat classification for HRT analysis was found to be less stringent than that required for ectopic beat exclusion in HRV parameter estimation. A small number of false negatives were acceptable in HRT but not in HRV analysis.

$RR_{i-1} < 0.80 * RA$ for a coupling interval

And

$RR_i > 1.20 * RA$ for a compensatory pause

RA was the running average or filtered average as define below. The user can choose the **HRT.org** analysis option using the switch in the upper right portion of the program initiation panel, as shown in Figure 3 below:

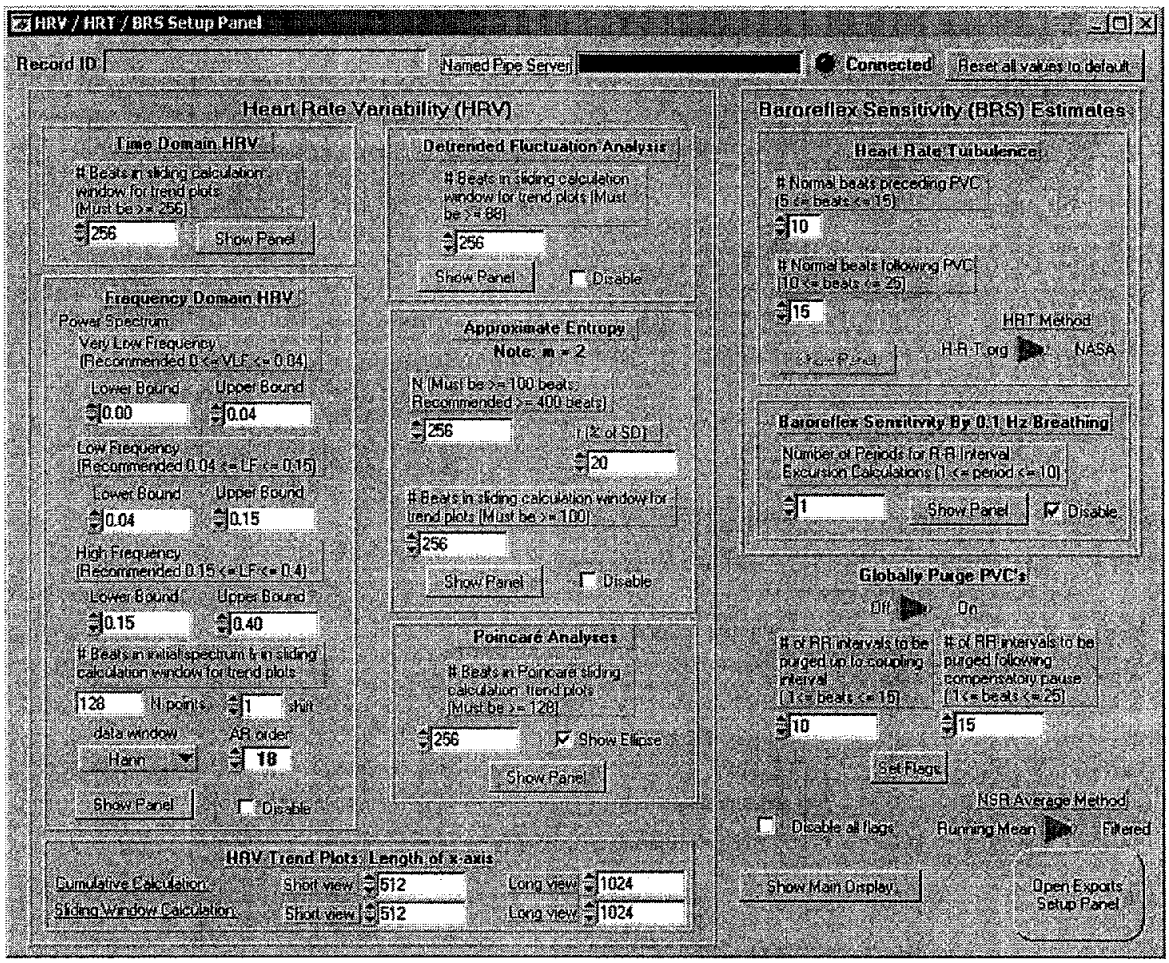


Figure 3: HRV Setup Panel showing the setting for HRT beat classification

A filtered running average for RA weighted the most recent NSR higher with the sequence [5^2, 4^2, 3^2, 2^2, 1] scaled to the sum of the coefficients. The switch in the lower right corner allows the user to select the standard or filtered option.

A slightly modified approach was developed to handle records where the coupling intervals were larger and compensatory pauses were smaller than the traditional criteria. This is designated as the **NASA** criteria on the initiation panel and is defined by the following:

$$RR_{i-1} < RA - RSD \qquad \text{for a coupling interval}$$

And

$$RR_i > \max[1.17 * RA, RA + RSD] \qquad \text{for compensatory pause}$$

where RSD is the running standard deviation.

An interval that does not qualify as a normal interval, coupling interval or compensatory pause is designated as an artifact. An "unknown" classification is necessary to label the first beat in the record since it does not have a preceding beat and cannot be classified as normal, PVC or artifact without it. The first beat was typically the only beat so designated.

Rhythm Check Beat Classification

The second algorithm investigated for beat classification was based on a subset of the algorithm development by Hamilton and distributed from his website as GPL code. RhythmChk, or rhythm check, compared the current interval to past seven intervals and their respective classification for a beat classification of the current beat. Intervals were classified as either normal, PVC or unknown. RhythmChk required a minimum of four intervals to begin. The first three intervals were designated as unknown. Subsequent unknown classifications were converted to artifact for integration into the HRV program.

PVC Qualification and Inclusion in HRT

RR intervals classified as a coupling interval followed by a compensatory pause were identified as individual PVCs. For inclusion in the HRT analysis, each PVC must also be evaluated to determine if it had the prerequisite number of normal sinus beats preceding and following it. The default values were 10 consecutive before and 15 consecutive NSR beats following each PVC. Otherwise the PVC was excluded from the HRT analysis. The program was written to allow the user to adjust these numbers within a specified range.

HRT Analysis

The TO and TS parameters were determined for all acceptable PVCs. If none were detected, a popup message was printed to the screen. Refer to Figure 2 for the format of the TO and TS parameters along with the total number of detected PVCs, acceptable

PVCs for analysis and a plot of the averaged HRT response. The maximum slope location is displayed and a line with this slope is drawn at that location (beat 7 in the above plot). If the TO and/or TS values are out of the normal range, they were highlighted in red on the display. The maximum slope for this record was 9.26 ms/beat and was in the normal range (> 2.5) and the TO value –1.95% was also in the normal range (<0%).

CVS

A version control system was installed for code maintenance and documentation. CVS required a CVS server installation to retain the code database and communicate with the user clients. The WinCvs[7] client and CVSNT[8] server package were installed. Both were licensed under GNU General Public License (GPL).

WinCVS

The WinCvs client interface provided the following features[7]:

- Sophisticated graphical user interface helps to utilize full power of CVS for experts and quickly learn basics for beginners.
- Native look-and-feel on Windows, Mac and Unix/Linux thanks to the use of popular GUI frameworks like MFC, Metrowerks PowerPlant and gtk+.
- Scripting support allows to easily automate, extend and customize common tasks.
- Realtime sandbox view with visual indication of the local state of files.
- Various filters to monitor any folder or all its subfolders in a flat view.
- Command line support makes any CVS commands or command options not directly handled by GUI possible.
- Repository tags, modules and files browser allows to easily enter command parameters.
- Changes in the files can be verified using diff command or external diff application.
- File revisions history can be displayed as a graph.
- Supports text, binary and Unicode file types.
- The type of the files is automatically detected upon import and add command.
- Reserved edits help to organize team work.
- Close cooperation with CVSNT project resulting in very dynamic and effective development of new features.

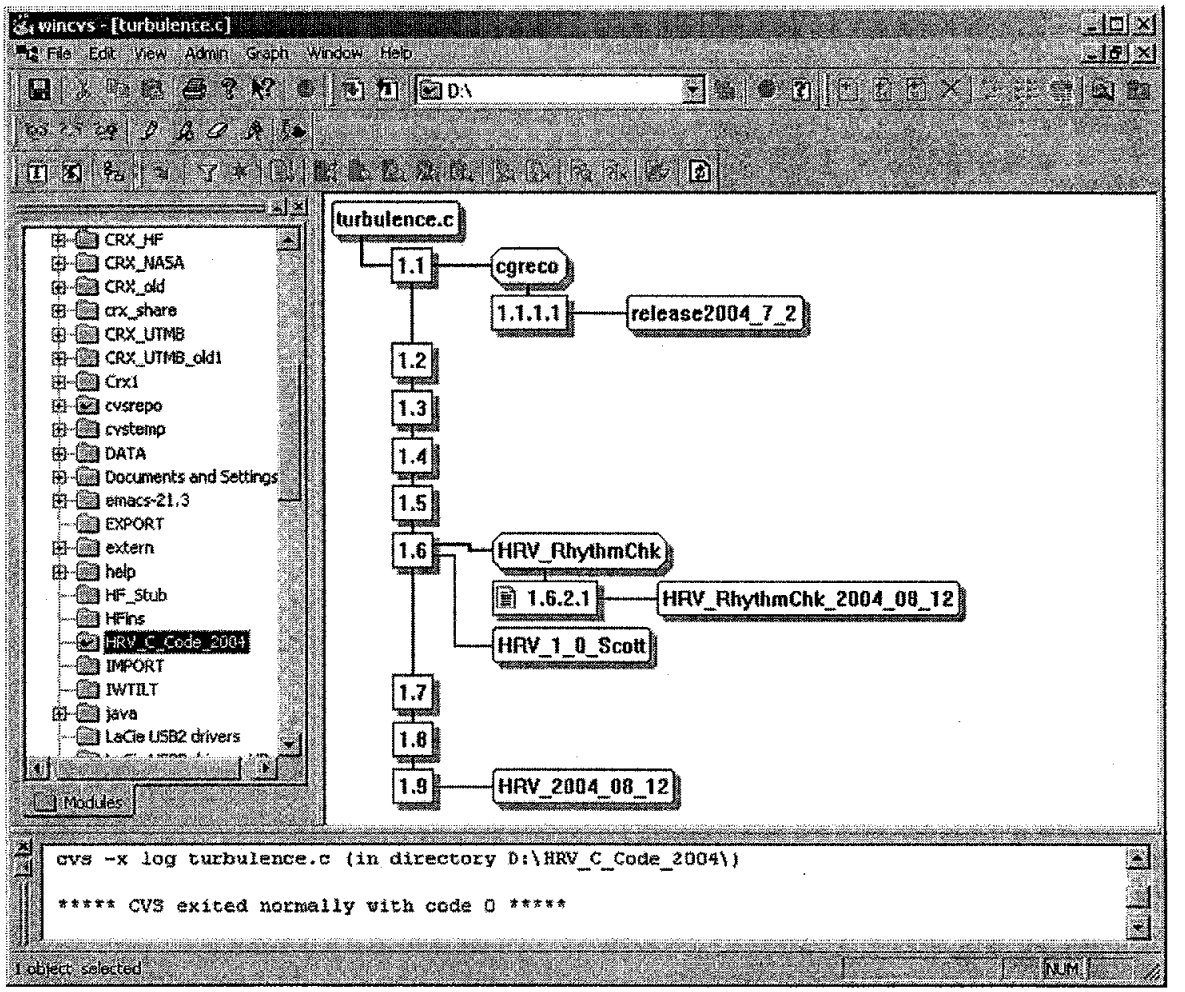


Figure 4: WinCvs client interface with graphical view of the revision history for turbulence.c. The HRV_2004_08_12 and HRV_RhythmChk_2004_08_12 tags correlate this module with the others at this revision level

The WinCvs provided the client side access to the code database required for daily updates (commits), updates and checkouts and yet was easy and straight forward to use.

The CVSNT server maintained the code database and provided network access for designated users. CVSNT supports several communication protocols. The Microsoft's Security Support Provide Interface (sspi) was selected for its support of domain name access and its encryption features. Two Groups were setup on the CVSNT server: CVSAdmins and CVSUsers, Figure 5 and 6. Users added to the CVSUsers group had access to the CVS code database to checkout, update and commit changes. The CVSAdmins group included users given additional administrative access not required by the general user. The CVS database was then assigned the necessary permissions for the

CVSUsers and CVSAdmins groups as stipulated in the installation CVSNT installation guide[8]. A command script, SetACL, was obtained from the CVSNT website and modified to facilitate this process.

The CVS database, or repository, contains the following packages: HRV_C_Code_2004, Beat_Class_Test, CBuf, dsp, and NamedPipeClient. Each package contains several modules, or files. Figure 4 displays the version history of the turbulence.c module from HRV_C_Code_2004 in graphical format. This turbulence.c module was initially checked in (committed) as version 1.1. It was revised five times, versions 1.1 – 1.6. Version 1.6 was tagged HRV_1_0_Scott along with all other modules in this package at the same level of development. A member of the CVSUsers group may then check out all modules associated with the HRV_1_0_Scott tag to obtain the package at this level of development. Also, at version 1.6 a branch was started and tagged HRV_RhythmChk. All additional modules included in this branch were similarly tagged. Again the user can checkout all modules associated with this branch by specifying the HRV_RhythmChk tag on checkout. The main branch was tagged HRV_2004_08_12 and the branch was tagged HRV_RhythmChk_2004_08_12 to identify and synchronize the modules at that development level.

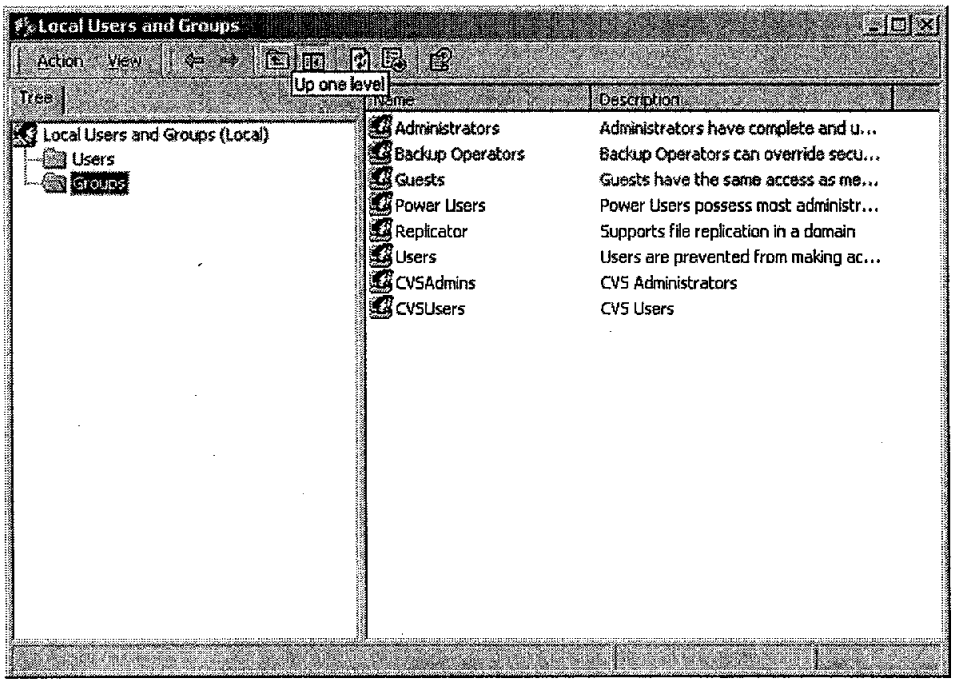


Figure 5: CVSNT Groups CVSAdmins and CVSUsers added to the CVSNT server

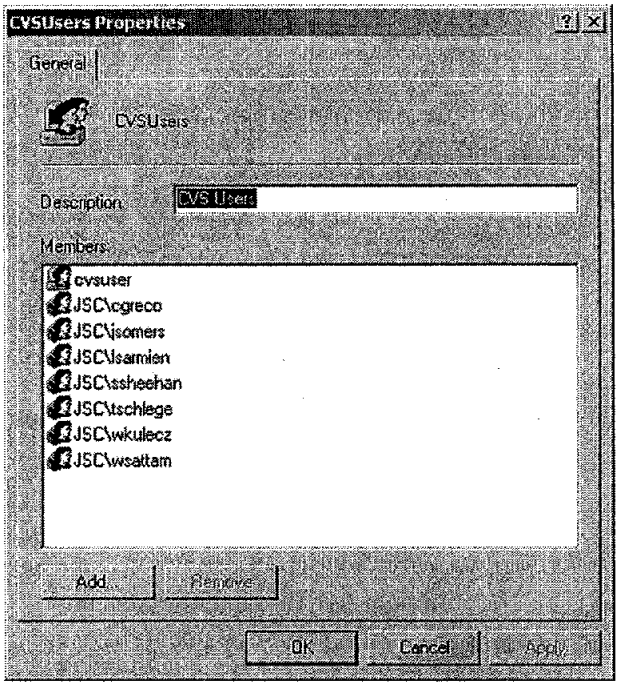


Figure 6: CVSUsers identified by their domain usernames

WORK COMPLETED THIS SUMMER

The work completed this summer included the development of the computer software to classify RR intervals based on the interval analyses noted above. A program was then written to identify PVCs to be included in the real-time HRT analysis as described in detail above. The code was integrated into a package of advanced electrocardiography software applications for heart rate variability analysis used in Dr. Schlegel's laboratory. The HRT software application described herein was written in the C computer language and was constructed in a LabWindows CVI (National Instruments, Austin, TX) programming environment, compatible with Microsoft Windows. Interested parties can view a demonstration of the HRT program by contacting Todd T. Schlegel, M.D. in the JSC Neurosciences Laboratory.

I worked with Scott Sheehan, a NSBRI student, to purge ectopic beats from several HRV parameters.

CONCLUSIONS

A series of tools were developed to facilitate the analysis of the electrocardiogram online in real time, and assist NASA flight surgeons and other physicians with cardiovascular

diagnoses. The evaluation of HRT onset and slope following premature ventricular contractions might eventually allow for a non-invasive and unobtrusive way to assess both susceptibility to arrhythmia and changes in baroreflex sensitivity in astronauts during and after space flight. Both cardiac arrhythmias and reduced baroreflex responsiveness are known to occur during space flight. The HRT software developed this summer has therefore been designed to ultimately allow NASA flight surgeons to follow trends in baroreflex sensitivity and arrhythmic risk in astronauts who have premature ventricular contractions. Similarly, the software should eventually allow other physicians to monitor at-risk cardiac patients on the ground, particularly those who may have a propensity for cardiac arrhythmias.

REFERENCES

1. Schmidt G, Malik M, Barthel P, Schneider R, Ulm K, Rolnitzky L, Camm AJ, Bigger JT, Jr., Schomig A: Heart-rate turbulence after ventricular premature beats as a predictor of mortality after acute myocardial infarction. Lancet 353:1390-6, 1999
2. Heart Rate Turbulence, Technische Universitat Munchen, 2004, pp http://www.h-r-t.org
3. Diaz JO, Castellanos A, Moleiro F, Interian A, Myerburg RJ: Relation between sinus rates preceding and following ectopic beats occurring in isolation and as episodes of bigeminy in young healthy subjects. *In* Am J Cardiol. Vol 90, 2002, pp 332-5
4. Pino F, Schlegel T: HRV Online Analysis Program. In Preparation, 2004
5. Hamilton PS, Tompkins WJ: Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. IEEE Trans Biomed Eng 33:1157-65, 1986
6. Hamilton PS, Curley MG: OSEA: RhythmChk. Vol 2004, 2.1 edition, EP Limited, 2004, pp http://www.eplimited.com/software.htm
7. WinCvs, 1.3.17.2 edition, CvsGui, 2004, pp http://www.wincvs.org/
8. CVSNT Wiki. Vol 2004, 2.0.51c edition, March Hare Pty Ltd, 2004, pp http://www.cvsnt.org/wiki/FrontPage

Effective Crew Operations: An Analysis of Technologies for Improving Crew Activities and Medical Procedures

Final Report
NASA Faculty Fellowship Program – 2004
Johnson Space Center

Prepared By:	Craig Harvey, Ph.D., P.E.
Academic Rank:	Assistant Professor
University & Department:	Louisiana State University Industrial Engineering Department Baton Rouge, LA 70803
NASA/JSC	
Office:	Habitability and Environmental Factors Office (HEFO) Habitability & Human Factors Office Usability and Testing Facility(UTAF)
JSC Colleague	Mihriban Whitmore, Ph.D.
Date Submitted	August 10, 2004
Contract Number	NAG 9-1526 and NNJ04JF93A

ABSTRACT

NASA's vision for space exploration (February 2004) calls for development of a new crew exploration vehicle, sustained lunar operations, and human exploration of Mars. To meet the challenges of planned sustained operations as well as the limited communications between Earth and the crew (e.g., Mars exploration), many systems will require crews to operate in an autonomous environment. It has been estimated that once every 2.4 years a major medical issue will occur while in space. NASA's future travels, especially to Mars, will begin to push this timeframe. Therefore, now is the time for investigating technologies and systems that will support crews in these environments. Therefore, this summer two studies were conducted to evaluate the technology and systems that may be used by crews in future missions.

The first study evaluated three commercial Indoor Positioning Systems (IPS) (Versus, Ekahau, and Radianse) that can track equipment and people within a facility. While similar to Global Positioning Systems (GPS), the specific technology used is different. Several conclusions can be drawn from the evaluation conducted, but in summary it is clear that none of the systems provides a complete solution in meeting the tracking and technology integration requirements of NASA. From a functional performance (e.g., system meets user needs) evaluation perspective, Versus performed fairly well on all performance measures as compared to Ekahau and Radianse. However, the system only provides tracking at the room level. Thus, Versus does not provide the level of fidelity required for tracking assets or people for NASA requirements. From an engineering implementation perspective, Ekahau is far simpler to implement that the other two systems because of its wi-fi design (e.g., no required runs of cable). By looking at these two perspectives, one finds there was no clear system that met NASA requirements. Thus it would be premature to suggest that any of these systems are ready for implementation and further study is required.

The second study evaluated current medical packs, used on-board the International Space Station (ISS), in the execution of an emergency medical procedure as compared to a modified design. An experiment using 13 participants found no difference in performance time between the two packs; however, it did find a marginally significant difference ($p = 0.08$) in the number of errors with the modified design resulting in less errors. Using the experimental data collected, a computer model was developed that allowed for running larger sample sizes. Results from this model found a statistically significant differences for time and errors ($p<0.05$). Further modeling evaluated the effect of errors on performance time. Results once again found a statistically significant difference for time and found that the current pack design's performance in time was 4 times greater when errors were considered as compared to the design when errors were ignored. However, the modified pack only saw a 2 times increase when errors were considered. Given that NASA typically is dealing with small samples and limited resources to test participants, modeling should be considered to evaluate designs prior to experimentation. Future work will investigate the value in developing a desktop application for modeling medical procedures independent of experimentation. This modeling does not preclude experimental efforts, but it does provide guidance in conducting experiments.

Both studies highlight issues that require further investigation. These studies are just one step needed to prepare systems and technologies for future planned human exploration.

INTRODUCTION

Two studies were conducted as a part of this report. The first study evaluated Indoor Positioning System (IPS) technology for use on-board the International Space Station (ISS) as well as any future NASA exploration vehicles. The second study evaluated current medical packs used on board the ISS and explored the use of modeling techniques for simulating human experiments.

STUDY 1: LOCATION TRACKING STUDY

IPSs are being used to track equipment and people in many different settings commercially including hospitals, university libraries, and museums. While similar to Global Positioning System (GPS) tracking devices, the specific technology used is different. IPSs are functionally in-door equivalents to GPS tracking systems.

The Biomedical Systems Division in the JSC Engineering Directorate was tasked to evaluate the usability of three commercial IPSs that are currently being used by several industries. Biomedical Systems Division personnel requested the Usability Testing and Analysis Facility (UTAF) to provide a human factors assessment of the three systems, Ekahau, Versus, and Radiance. All three systems had been procured as evaluation systems. This study report summarizes the UTAF human factors assessment findings, including: a general description of each system, the human factors assessment approach employed, identified issues and recommendations, and a plan for possible future work.

Location Tracking System Descriptions

Three commercial systems were evaluated as a part of this assessment. A brief overview of each of the systems evaluated is provided.

Ekahau: The Ekahau Positioning system is developed by Ekahau, Inc. of Saratoga, CA. The Ekahau Site Calibration™ method is used for collecting radio network sample points from different site locations. Each sample point contains received signal strength intensity (RSSI) and the related map coordinates, stored in an area-specific positioning model for accurate tracking.

Ekahau advertises up to a 1 meter (3½ ft) average positioning accuracy based on the positioning model technology. Ekahau allows an administrator to define "logical" areas that can be used to identify an individual's location. These logical locations could correspond to a room or a portion of a room. Ekahau's positioning and site survey technologies work with all industry-standard wi-fi (IEEE 802.11a/b/g) access points and most network cards without proprietary hardware. Therefore, Ekahau when implemented can take advantage of existing wi-fi networks.

Versus: Versus Information System (VIS) is developed by Versus Technologies, Inc. of Traverse City, MI. Versus is a combination radio frequency (RF) and infrared (IR) system that can identify at "room-level" the location of an individual or piece of equipment. Versus has two types of badges: those whose purpose is solely to monitor an individual's location, and those that allow an individual to send an alert to an operator console.

Radianse: Radianse, Inc., formerly Sentinel Wireless, Inc., is headquartered in Lawrence, Massachusetts. Radianse has developed an active-Radio Frequency Identification (RFID) technology that provides identification. Radianse receivers use standard Ethernet wiring and connect directly to a network where they require miniscule bandwidth. Internet Protocol (IP) addresses can be either Dynamic Host Configuration Protocol (DHCP) or static IP. Each receiver covers up to a 60-foot diameter. Location data from Radianse badges are collected by Internet

protocol-based Radianse receivers, and transmitted over any existing local area networks (LAN) to Radianse Location Software. Radianse location data can be shared with other systems and databases using standards such as Open Database Connectivity (ODBC), Extensible Markup Language (XML), Short Messaging System (SMS), Java, and JavaScript. Radianse badges contain two programmable buttons that can be set to send different messages to the operator console.

HUMAN FACTORS ASSESSMENT APPROACH

Participants

The UTAF requested 10 (4 male, 6 female) different individuals from the Habitability and Human Factors Office (JSC-SF3) and the Biomedical Systems Division (JSC-EB) to serve as evaluators of the system. A total of five participants were used for the individual tests as described below in the procedure section and six pairs of individuals were used for the team evaluation described in the procedure. Individuals participated in one or more of the tests. Since the systems and not the participants were being tested, participants served only as a means of moving the transmitters through the facility. No crew members were used for this evaluation.

Test Facility and Materials

This study was conducted in the Advanced Integration Matrix (AIM) habitat located at NASA's Johnson Space Center, building 29. Three areas in the module were used including an area with two floors. Transmitters were carried or worn by participants in this study depending on the location system. Participants carried a HP iPaq Pocket PC Personal Digital Assistant (PDA) with a compact flash wi-fi card for the Ekahau system. Participants wore a battery operated RF badge for the Radianse system. Participants wore the Versus Personnel Alert badge, a battery operated RF/IR badge, which allows for monitoring and sending alerts. We only tested the monitoring function.

A dedicated laptop computer was used to monitor the personnel traversing the AIM facility and also used to capture the location data. This laptop was configured with each system's software as provided by three vendors in their evaluation packages. Each system was configured by an in-house engineer trained on the systems from JSC-EB. Prior to any testing, all targets were evaluated by placing the particular transmitter on predefined targets in the same orientation to ensure the system would recognize the transmitter independent of people transporting the transmitter.

Existing video cameras located throughout the AIM facility were used to capture the participants traversing the facility on VHS tapes. In addition, the existing intercom system was used to tell participants when to travel to the next target in the path.

Procedure

Participants were given an overview of the testing being conducted, signed a consent form approving video recording of the session, completed a demographic survey, and received a safety briefing so as to avoid the potential tripping hazards within the AIM facility. In addition, participants completed a walkthrough of path they were to follow so that they were aware of the targets in the facility. Once the system was enabled, the participant was either given the PDA to carry or had the receiver attached to their clothing. In either case, all participants carried the PDA or wore the receiver in the same position. These individuals traversed the AIM facility while wearing/holding the transmitters of each of the systems.

Five conditions were tested for this study and the exact procedure followed for each condition is described below.

Condition 1: Individual walking (Individual)

Participant Procedure: A single participant, holding/wearing a transmitter, walked to each of the sequentially numbered targets of the red path (refer to Figure 1) when they were instructed to do so (i.e., test conductor announced "Next" over the intercom system). Participants were told when to begin walking to the first target (target 2) based on the computer clock used by the systems to time stamp their location. The participant would stop at each target for a period of approximately 15 seconds, and advance to the next target based on a verbal command given through the intercom system. The participant continued the walk through all 15 targets. A total of five different participants were used for this condition.

Condition 2: Two-person side-by-side walking (Pairs)

Participant Procedure: Two participants walked the identified red path together. Participants advanced through the targets as in condition 1. A total of six participants (3 teams) were used for this condition.

Condition 3: Two-person opposite walking (Opposite)

Participant Procedure: Two participants walked the identified paths (blue and green, refer to Figure 1). One participant started on the blue path and one participant started on the green path. The participants passed each other on the paths. Participants advanced through the targets as in condition 1. In the module with two floors, participants first walked in parallel along the module (e.g., one on the 1st floor and one on the 2nd floor) in the same direction. On a second pass through the module, participants began at different ends on different levels and crossed over one another in the middle. This required participants to pass one another, but on different levels. A total of six participants (3 teams) were used for this condition.

Condition 4: Hidden transmitter detection (Individual)

Participant Procedure: A participant was asked to walk to predefined locations for each test location. In this evaluation, transmitter (PDA or badge) was hidden in different clothing and a computer bag to determine if it could still be detected by the system. One participant at several randomly selected locations was used.

Condition 5: Obstacle transmitter detection (Individual)

Participant Procedure: A participant was asked to walk to predefined locations for each test location. For this condition, several obstacles were selected for testing and these obstacles were placed in path of receivers. Locations were randomly chosen and obstacles included metal, plastic, wood, Plexiglas, and cardboard. One participant at several randomly selected obstructed locations.

Route Maps

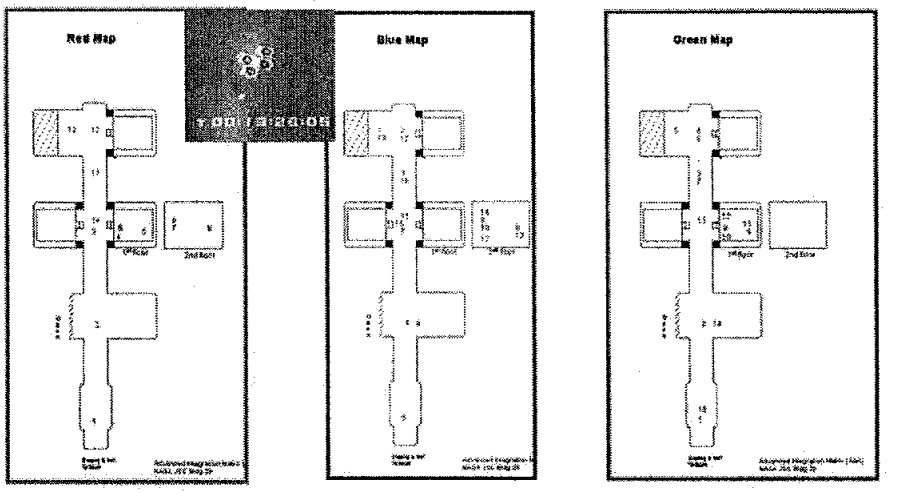


FIGURE 1: MAPS OF PATHS WALKED AND EXAMPLE NUMBERING (PHOTO)

Measures of Performance

Three different measures of performance were identified for evaluation of the three systems: errors, time to detect, and percentage time correct. Each measure is described in Table 1.

TABLE 1: MEASURES OF PERFORMANCE

Measure	Description of Measure
Number of Errors	An error could be of two types: fail to detect (Miss) or detection of the participant in one location when they were actually in another location (False Alarms). An error occurred whenever the system failed to detect the person correctly while standing at a target for a period of 15 seconds or failed to detect them upon arrival at a new target. This method of counting error follows the Signal Detection Theory (SDT) method as developed by Green and Swets (1966).
Time to Detect	This measure evaluates the time it took for the system to identify the individual at a target. Four methods of time to detect were defined: (1) early (-) and late detection (+) time were counted, these times could potentially cancel each other out; (2) absolute value time to detect (e.g., early or late detection was counted); (3) all times late or early, within +/- 2 seconds were dropped and average time calculated; (4) late detection only time was counted.
Percentage Time Correct (used as replacement for time to detect measure as will be discussed in report)	This measure is the total number of seconds the system properly detected the participant compared to the total number of seconds for which the participant is standing at the path targets. Time was only counted for the time the participant was at an actual target. No attempts were made to assess the accuracy of the system while the participant was traversing the path. Using the SDT model above, this method assessed the percentage time of hits compared to total time available for a hit.

Limitations of the Study

Before discussing the results of the study, it is important to understand the limitations of this study up front.

- The capabilities of the systems were not equivalent. For example, Ekahau provided a display showing a map of the AIM facility and the current location of the individual; however, the other two evaluation systems tested did not have this capability.
- The systems were set up and calibrated by a trained individual on the EB team, rather than the respective vendors. It is possible that with vendor support of the set up, each system might have performed better.
- The four wi-fi routers used by Ekahau in this study included NetGear (1), Linksys (1), and D-Link (2). There were some indications that accuracy may have differed for the different vendor routers. This study did not evaluate this possibility.
- The Ekahau system periodically indicates fluctuations (1 to 2 second) that will place an individual in the wrong location. As a result, our analysis looks at the impact of these fluctuations on system performance. There was some discussion that these anomalies might be due to the wi-fi points, settings within the software, or versions of the software. This study did not evaluate all of these factors, and thus they should be considered in future evaluations.

RESULTS

The results will be discussed by looking at each of the measures discussed in Table 1.

Number of Errors

Figure 2 displays the errors per system. Ekahau was prone to minor fluctuations (1 -2 seconds) in its detection of a participant. The system would sometimes flip between areas when the participant was actually standing still on a target. Because of these fluctuations, the initial analysis of Ekahau showed an error rate of over 100% when each fluctuation was counted as an error since the number of errors exceeded the number of targets. Consulting with JSC-EB and review of the fluctuations found that the fluctuations were minor. Therefore, it was decided that Ekahau would be evaluated by eliminating any of the minor fluctuations (e.g., 1-2 seconds). Thus one will notice in Figure 2 that the errors displayed for Ekahau display a NF or no fluctuations indicator. In addition, since Ekahau can detect down to a logical area and the other two systems, as tested, could only detect to room level, Figure 2 displays the error rate for Ekahau at both the logical and room level. Figure 2 also shows the error rate for conditions 1 (Individual), 2 (Pairs), and 3 (Opposite) for all systems.

As one can see in Figure 2, Versus had the lowest error rate among the systems tested and Ekahau by logical area had the highest error rate. When Ekahau is evaluated at room level in comparison to the other two systems, its error rate decreases substantially; however it is still quite a bit larger than the other two systems. Even with Versus having the lowest error rate, its rate is still approximately 14%.

To assess the severity of the errors received, each of the errors was classified into three categories: (1) Adjacent - identified the individual in an adjacent room to their actual location; (2) Nonadjacent – identified the individual in an area non adjacent to their actual location; and, (3) Different floor – identified the individual on a floor different that where they were located. Because of the uniqueness of the AIM facility compared to facilities in which these systems are typically implemented - for example, the open second floor and total metal structure, we felt it was important to understand the types of errors occurring. This may be useful to vendors of the systems as well as to JSC. Table 2 provides a percentage breakdown of the types of errors encountered.

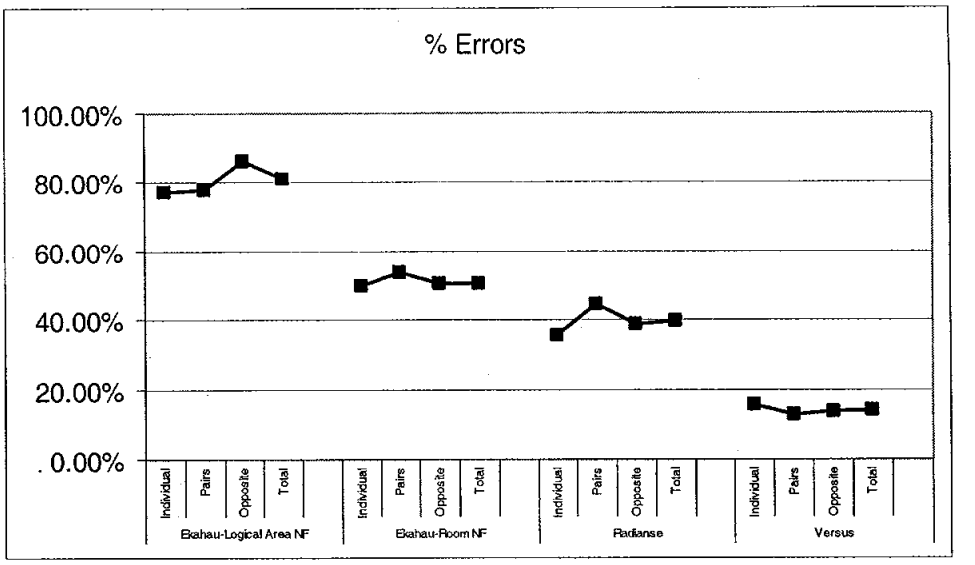


Figure 2. Percentage of error by system and by condition

TABLE 2: PERCENTAGE OF TYPES OF ERRORS

System	Adjacent	Nonadjacent, Same Floor	Different Floor	Nonadjacent, Different Floor
Ekahau	42.70%	0.00%	53.93%	3.37%
Radianse	54.17%	9.72%	19.44%	16.67%
Versus	63.64%	0.00%	13.64%	22.73%

Time to Detect

Initially the time to detect measure was identified as a system measure of performance. However, because of the fluctuations in the Ekahau system, it was felt that this measure would be difficult to assess fairly. Therefore, this measure was replaced by the percentage time correct measure.

Percentage Time Correct

Figure 3 displays the percentage time correct on target for each system. Once again, Ekahau was evaluated several different ways because of the differences between it and the other systems. Ekahau was first evaluated including any and all fluctuations in a location and then excluding all small fluctuations (1-2 sec). Once again, it was also evaluated at the logical and room level.

Versus once again had the largest percentage correct, 96.5%, as compared to the other systems. Ekahau including fluctuations at the logical area had the lowest percentage time correct, 41.45%. The percentage time correct and number of errors in detection follow similar patterns of performance except in the opposite direction. Thus, the higher the numbers of errors, the smaller the time percentage correct as confirmed by a Pearson's correlation of 0.98.

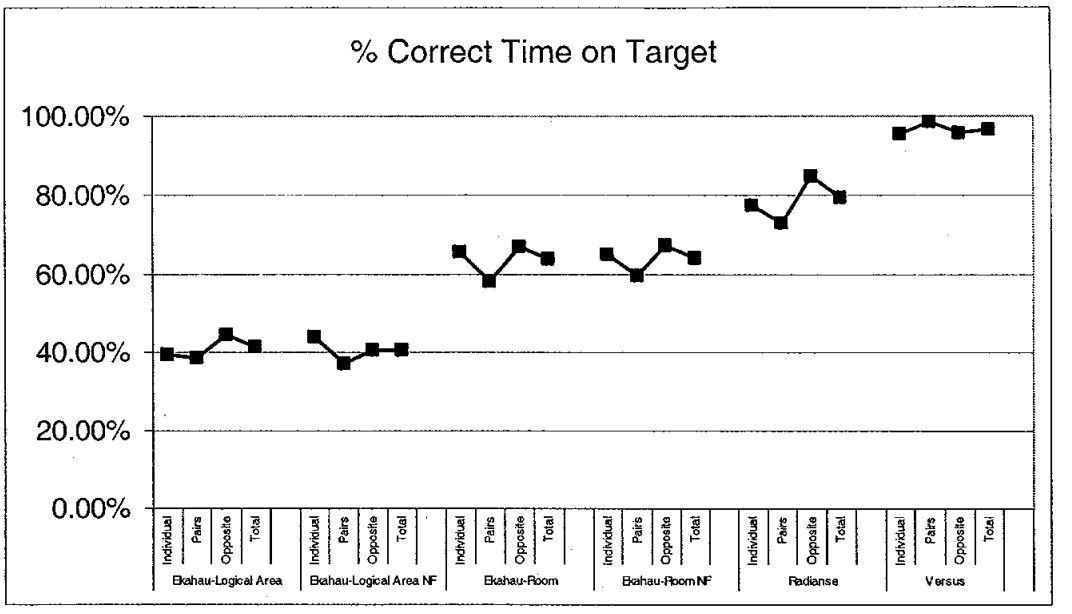


Figure 3: Percentage of time a system detected correctly by system and by condition

Hidden Transmitter Detection

For this evaluation, a single subject was used for the evaluation. The transmitter for each system was placed in different potential items that could hide the signal. The five tested scenarios were: (1) in the pocket of a coat (light colored); (2) under a light colored shirt; (3) in a computer bag; (4) in pant (jeans) pocket; and (5) under a dark colored shirt. Ekahau and Radianse were found in all hidden locations. Versus was found in all locations with the exception of three locations: (1) in a computer bag; (2) in pant (jeans) pocket; and (3) under a dark colored shirt. This finding was expected since the Versus system requires a good IR signal to detect the badge.

Obstacle Transmitter Detection

Several different obstacles were evaluated to determine if the transmitters would be impacted including Styrofoam, Plexiglas partitions, metal panels, and a large piece of equipment. In addition, we asked the participant to enter areas with the transmitter away from the receiver. Groups of people also were used for evaluation. Each of the systems were discovered in all locations with some minor exceptions. Versus could be found in all locations except when the participant entered an area with their back turned or when they stood behind a large piece of equipment. Ekahau experienced fluctuations is signal when a group of four people were in front of the receiver. Ekahau would locate the individual in the wrong location.

DISCUSSION

As this was a preliminary study, only a small sample of participants was used for the study and no statistical analysis was conducted. However, this study did allow for testing of the study protocol and the collection of some initial data for assessment of the three systems. The small sample size of participants is a limitation; however, it did provide a very good assessment of the capability of the systems as they stand today in the AIM facility.

When all three systems are compared on their ability to identify an individual's location within a small area, Ekahau was the only system tested that provided this capability. The other

A Fundamental Mathematical Model of a Microbial Predenitrification System

Final Report

NASA Faculty Fellowship Program – 2004

Johnson Space Center

Prepared By:	Karlene A. Hoo, Ph. D.
Academic Rank:	Professor
University and Department:	Texas Tech University Chemical Engineering Department Lubbock, TX 79409
NASA/JSC	
Directorate:	Engineering
Division:	Automation, Robotics and Simulation
Branch:	Intelligent Systems
JSC Colleague:	David Overland
Date Submitted:	August 13, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

Space flight beyond Low Earth Orbit requires sophisticated systems to support all aspects of the mission (life support, real-time communications, etc.). A common concern that cuts across all these systems is the selection of information technology (IT) methodology, software and hardware architectures to provide robust monitoring, diagnosis, and control support. Another dimension of the problem space is that different systems must be integrated seamlessly so that communication speed and data handling appear as a continuum (un-interrupted). One such team investigating this problem is the Advanced Integration Matrix (AIM) team whose role is to define the critical requirements expected of software and hardware to support an integrated approach to the command and control of Advanced Life Support (ALS) for future long-duration human space missions, including permanent human presence on the Moon and Mars. A goal of the AIM team is to set the foundation for testing criteria that will assist in specifying tasks, control schemes and test scenarios to validate and verify systems capabilities.

This project is to contribute to the goals of the AIM team by assisting with controls planning for ALS. Control for ALS is an enormous problem it involves air revitalization, water recovery, food production, solids processing and crew. In more general terms, these systems can be characterized as involving both continuous and discrete processes, dynamic interactions among the sub-systems, nonlinear behavior due to the complex operations, and a large number of multivariable interactions due to the dimension of the state space. It is imperative that a baseline approach from which to measure performance is established especially when the expectation for the control system is complete autonomous control.

ACKNOWLEDGEMENTS

This work was made possible not only by the financial support of the NASA's NFFP program, but also by the active participation of NASA civil servant and my collaborator David Overland (Engineering, ER2). Technical assistance and many useful discussions were provided by Marvin Ciskowski (Lockheed) and David Overland. Program outreach support was provided by Gretchen Thomas (NASA), Heather Paul (NASA), Laura Labuda (Lockheed Martin), and Debbie Berdich (Lockheed Martin).

1. INTRODUCTION

The objective of the project is to treat grey water using a microbial based wastewater treatment system. Grey water is any water that has been used in the home, except water from toilets. This may include dish, shower, sink, and laundry water. This type of "water" may be reused for other purposes, especially landscape irrigation. In space studies: gray water will contain, 10% urine, hygiene (hand, shower, and oral waters), laundry water (high surfactant concentrations).

True nitrifying bacteria are considered to be those belonging to the family *nitro-bacteraceae*. These bacteria are strictly aerobic, gram-negative, chemolithic autotrophs. They require oxygen, utilize mostly inorganic (without carbon) compounds as their energy source, and require carbon dioxide (CO_2) for their source of carbon. In the case of the *nitrobacteraceae* these energy sources are derived from the chemical conversion of ammonia (NH_4^+-N) to nitrite (NO_2^--N) or, nitrite (NO_2^--N) to nitrate (NO_3^--N). *Nitrosomonas* is the most common ammonia-oxidizer while *nitrobacter* is the most common nitrite-oxidizer.

The nitrifiying reactor studied contains three types of microorganisms: aerobic heterotrophs, *nitrosomonas* and *nitrobacters*. Assume that the cell can be represented generically by the formula $C_5H_7O_2N$ (mw=113g/gmol) and that the organic carbon donor has the formula $C_{10}H_{19}O_3N$ (mw= 201g/gmol).

Nitrosomonas oxidize the ammonium-ion to nitrite and *nitrobacters* oxidize the nitrite to nitrate. Overall

$$\tfrac{1}{8}NH_4^+ + \tfrac{3}{8}H_2O \mapsto \tfrac{1}{8}NO_3^- + \tfrac{5}{4}H^+ + e^- \qquad (1)$$

These two species work together to achieve the overall oxidation of ammonium to nitrate. The nitrate is recycled to the denitrifying reactor where it is used by the microorganisms (*pseudomonas*) under anoxic conditions.

Dentrifiers are chemotrophs that use organic and inorganic electron donors. Common gram-negative dentrifiers are *Proteobacteria* such as *pseudomonas*. All dentrifiers are facultative microbes, which means they shift to either nitrate or nitrite ion respiration when oxygen is limited. Organic carbon users are heterotrophs. Denitrification proceeds in a stepwise manner:

$$\tfrac{1}{2}NO_3^- + e^- + H^+ \mapsto \tfrac{1}{2}NO_2^- + \tfrac{1}{2}H_2O : \text{nitrate reduction}$$
$$NO_2^- + e^- + 2H^+ \mapsto NO + H_2O : \text{nitrite reduction}$$
$$NO + e^- + H^+ \mapsto \tfrac{1}{2}N_2O + \tfrac{1}{2}H_2O : \text{nitric oxide reduction}$$
$$\tfrac{1}{2}N_2O + e^- + H^+ \mapsto \tfrac{1}{2}N_2(g) + \tfrac{1}{2}H_2O : \text{nitrous oxide reduction}$$

Electrons provided by the donor are used for energy (R_e) and cell synthesis (R_s),

$$R = f_e R_e + f_s R_s$$
$$R_e = R_a - R_d \qquad (2)$$
$$R_s = R_c - R_d$$

where R_d is the donor reaction, R_a is the electron acceptor reaction, and R_c is the cell synthesis reaction. The relationship between f_e, fraction used for energy, and the fraction used for cell synthesis is f_s=1-f_e.

2. PREDENITRIFICATION SYSTEM

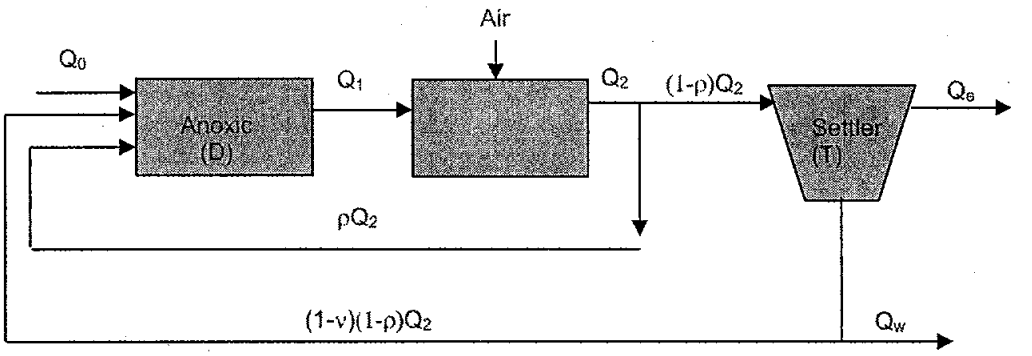


Figure 1: Denitrifier and nitrifier in series with a settler that return active sludge. The nitrifier returns a high concentration of nitrate-ion to the denitrifier.

Consider the following two reactors in series and a settler, shown in Figure 1. The first reactor is a denitrifier that converts organic carbon compounds under anoxic conditions (absence of oxygen) for energy and nitrate-ion (electron acceptor) for cellular respiration. Carbon dioxide and a source of nitrogen (ammonium-ion) are used for cell synthesis. Other nutrients such as phosphorous should also be supplied. The second reactor is the nitrifier that operates aerobically and consists of *nitrosomonas*, *nitrobacters*, and aerobic heterotrophs. Oxygen is used for cellular respiration and is supplied in the form of air (79% N_2, 21% O_2). Any organic carbons unused by the *pseudomonas* are the source of energy for the aerobic heterotrophs. Ammonium-ion: ($NH_4^+ - N$)) is used by *nitrosomonas* for energy (produce nitrite-ion ($NO_2^- - N$)) and both *nitrosomonas* and *nitrobacters* also use the ammonium-ion for cell synthesis. The nitrite-ion is used (reduced) by the *nitrobacters* for energy. The product, the nitrate-ion ($NO_3^- - N$)) is recycled to the denitrifier for cellular respiration. The rate of cell synthesis in the denitrifier is limited by the supply of nitrate-ion. The recommended recycle rate is 4 to 6 times the influent stream to the denitrifier.

$$f_d = 1 - \frac{1}{1+b\theta_x}\left(\frac{f_s}{f_s^0}(1+b\theta_x) - 1\right)$$
$$f_s = \frac{f_s^0\left(1+(1-f_d)b\theta_x\right)}{1+b\theta_x} \qquad (3)$$

2.1 STOICHIOMETRY

2.1.1 DENITRIFIER (ANOXIC REACTOR)

Parameter	*Pseudomonas*
f_s^0	0.52
Y mgVSSa/mgBOD$_L$	0.26
$\hat{q}$ mgBOD$_L$/mgVSSa-d	12
$\hat{q}$ mgNO$_3$/mgVSSa-d	16
K mgBOD$_L$/L	1
K_{no} mgNO3/L	10
θ_x d	5
b d	0.052
$\left[\theta_x^{min}\right]_{lim}$ d	0.33
S_{min} mgNO$_3$/L	0.017
f_d	0.8
f_s	0.4342

Table 1: Typical parameters for a Denitrifier: T=20^{0}C (Rittman and McCarty)

Pseudomonas

$$R_a : \frac{1}{5}NO_3^- + \frac{6}{5}H^+ \mapsto \frac{1}{10}N_2 + \frac{3}{5}H_2O$$

$$R_c : \frac{1}{5}CO_2 + \frac{1}{20}HCO_3^- + \frac{1}{20}NH_4^+ + H^+ \mapsto \frac{1}{20}C_5H_7O_2N + \frac{9}{20}H_2O$$

$$R_d : \frac{1}{50}C_{10}H_{19}O_3N + \frac{9}{25}H_2O \mapsto \frac{9}{50}CO_2 + \frac{1}{50}HCO_3 + \frac{1}{50}NH_4^+ + H^+$$

$$\frac{1}{50}C_{10}H_{19}O_3N + \frac{fe}{5}NO_3^- + (\frac{1}{5} - \frac{fs}{5})H^+ + (\frac{fs}{5} - \frac{9}{50})CO_2 + (\frac{fs}{20} - \frac{1}{50})HCO_3^- + (\frac{fs}{20} - \frac{1}{50})NH_4^+ \mapsto \frac{fe}{10}N_2 + (\frac{4}{25} - \frac{3fs}{20})H_2O + \frac{fs}{20}C_5H_7C$$

2.2 NITRIFIER (AEROBIC REACTOR):

Parameter	*Nitrosomonas*	*Nitrobacters*	Aerobic Heterotrophs
f_s^0	0.14	0.10	0.7
Y	0.33 mgVSSa/mgNH$_4$	0.083mgVSSa/mgNO2	0.45mgVSSa/mgBOD$_L$
$\hat{q}$	1.7mgNH$_4$/mgVSSa-d	7.3 mgNO$_2$/mgVSSa-d	10mgBOD$_L$/mgVSSa-d
$\hat{q}$ mgO$_2$/mgVSSa-d	5.1	7.5	
K	0.57mgNH$_4$/L	0.62 mgNO$_2$/L	10 mgBOD$_L$/L
K_{Ox} mgO$_2$/L	0.5	0.68	
b d	0.082	0.082	0.1
$\left[\theta_x^{min}\right]_{lim}$ d	2.1	1.9	
$S_{min,Ox}$ mgO$_2$/L	0.084	0.12	
$S_{min,N}$	0.094 mgNH$_4$/L	0.1 mgNO$_2$/L	
f_d	0.067	0.067	0.8

Table 2: Typical parameters for a Nitrifier: T=15^{0}C (Table 9.1, p 472, Rittman and McCarty)

Nitrosomonas

$$R_a : \frac{1}{4}O_2 + H^+ + e^- \mapsto \frac{1}{2}H_2O$$

$$R_c : \frac{1}{5}CO_2 + \frac{1}{20}HCO_3^- + \frac{1}{20}NH_4^+ + H^+ + e^- \mapsto \frac{1}{20}C_5H_7O_2N + \frac{9}{20}H_2O$$

$$R_d : \frac{1}{6}NH_4^+ + \frac{1}{3}H_2O + e^- \mapsto \frac{1}{6}NO_2^- + \frac{4}{3}H^+$$

$$R : \frac{(1-fs)}{4}O_2 + \frac{fs}{5}CO_2 + \frac{fs}{20}HCO_3^- + (\frac{1}{6} + \frac{fs}{20})NH_4^+ \mapsto \frac{fs}{20}C_5H_7O_2N + \frac{1}{6}NO_2^- + (\frac{1}{6} - \frac{fs}{20})H_2O + \frac{1}{3}H^+$$

Nitrobacter:

$$R_a : \tfrac{1}{4}O_2 + H^+ \mapsto \tfrac{1}{2}H_2O$$

$$R_c : \tfrac{1}{5}CO_2 + \tfrac{1}{20}HCO_3^- + \tfrac{1}{20}NH_4^+ + H^+ \mapsto \tfrac{1}{20}C_5H_7O_2N + \tfrac{9}{20}H_2O$$

$$R_d : \tfrac{1}{2}NO_2^- + \tfrac{1}{2}H_2O \mapsto \tfrac{1}{2}NO_3^- + H^+$$

$$R : \tfrac{1-fs}{4}O_2 + H^+ + \tfrac{fs}{5}CO_2 + \tfrac{fs}{20}HCO_3^- + \tfrac{fs}{20}NH_4^+ + \tfrac{1}{2}NO_2^- \mapsto \tfrac{1}{2}NO_3^- \tfrac{fs}{20}C_5H_7O_2N + \tfrac{fs}{20}H_2O$$

Aerobic Heterotrophs:

$$R_a : \tfrac{1}{4}O_2 + H^+ \mapsto \tfrac{1}{2}H_2O$$

$$R_c : \tfrac{1}{5}CO_2 + \tfrac{1}{20}HCO_3^- + \tfrac{1}{20}NH_4^+ + H^+ \mapsto \tfrac{1}{20}C_5H_7O_2N + \tfrac{9}{20}H_2O$$

$$R_d : \tfrac{1}{50}C_{10}H_{19}O_3N + \tfrac{9}{25}H_2O \mapsto \tfrac{9}{50}CO_2 + \tfrac{1}{50}HCO_3^- + \tfrac{1}{50}NH_4^+ + H^+$$

$$R : \tfrac{1-fs}{4}O_2 + \tfrac{1}{50}C_{10}H_{19}O_3N + (\tfrac{fs}{5} - \tfrac{9}{50})CO_2 + (\tfrac{fs}{20} - \tfrac{1}{50})HCO_3^- + (\tfrac{fs}{20} - \tfrac{1}{50})NH_4^+ \mapsto \tfrac{fs}{20}C_5H_7O_2N + (\tfrac{7}{50} - \tfrac{fs}{20})H_2O$$

3. MODELING

Streams appear as subscripts; components appear as superscripts. Properties of a unit operation appear as subscripts (N: nitrifier, D: denitrifier, T: settler). For example V_N: volume of the nitrifier. The following balances are with reference to Figure 1.

3.1 DENITRIFIER

Input streams are: Q^0 (L/d), ρQ_2 (L/d), ν (1-ρ) Q2 (L/d). Input biochemical oxygen demands (BOD_L) are the carbon source, S^0 (mg BOD_L/L), the nitrogen (total Kjedahl nitrogen or TKN) source, Sn^0 (mg N/L), the nitrate-ion, *Sno* (mg NO3-N/L), and the initial amount of inerts, x_i^0 (mgVSS/L). *Pseudomonas* (x_P: mgVSSa/L) biomass is the *active* volatile suspended solids (VSSa) and x_v (mgVSS/L) = $x_P + x_i$ is the total volatile suspended solids in the effluent.

3.1.1 COMPONENT BALANCES

$$
\begin{aligned}
&\text{biomass}: V_D \frac{dx_{P1}}{dt} = Q^0 x_P^0 + Q_2 f_d(\rho x_{v2} + \nu(1-\rho))x_w - Q_1 x_{p1} + \left(Y_P r_P - b_P x_{P1}\right)V_D \\
&\text{orgC}: V_D \frac{dS_1}{dt} = Q^0 S^0 + \rho Q_2 S_2 + \nu(1-\rho)Q_2 S_2 - Q_1 S_1 - r_P V_D \\
&\text{NO}_3^- : V_D \frac{dSno_1}{dt} = \rho Q_2 Sno_2 + \nu(1-\rho)Sno_2 - Q_1 Sno_1 - r_{no} V_D \\
&\text{TKN}: V_D \frac{dSnh_1}{dt} = Q^0 Sn^0 + \rho Q_2 Snh_2 + \nu(1-\rho)Q_2 Snh_2 - Q_1 Snh_1 - r_{nh} V_D \\
&\text{inerts}: V_D \frac{dx_{i1}}{dt} = Q^0 x_i^0 - Q_1 x_{i1} + (1-f_d)b(x_{v2} + x_w + x_{P1})V_D \\
&\text{utilRate}: r_P = \frac{\hat{q}_P S_1}{K_P + S_1} x_{P1} (\text{mgBOD}_L/\text{L - d})
\end{aligned}
\tag{4}
$$

where Y_P is the true yield for synthesis; f_d is the fraction of the active biomass (x_P) and recycled volatile suspended solids that is biodegradable, x_i^0 is the influent inert concentration, K_P is that concentration that gives one-half the maximum growth rate, $\hat{q}_P$ is the maximum specific rate of substrate utilization, b_P: is the endogenous decay coefficient, ρ is that fraction of the effluent stream (Q_2) of the nitrifier that is recycled to the denitrifier, Snh is the ammonium-ion,$(1-\nu)$ is that fraction of the influent stream ($(1-\rho)Q_2$) to the settler that is recycled to the denitrifier, and x_{v2}, x_w are the mixed liquor compositions that are recycled from the nitrifier and waste stream of the settler, respectively.

Define the hydraulic detention time (HDT, units of days), θ_D, in the denitrifier, and the nitrifier, θ_N, respectively by

$$
\theta_D = \frac{V_D}{Q^0} \text{ and } \theta_N = \frac{V_N}{Q_1} \tag{5}
$$

The mean cell retention time, θ_{xD}, (MCRT same as solids retention time or the sludge age, units of days) is defined by

$$
\theta_{xD} = \frac{V \cdot \text{active biomass}}{Q_1 \cdot \text{produced biomass}} \tag{6}
$$

A useful relationship between the MCRT and the HRT is given by

$$
\theta = \frac{\theta_x}{w_v}\left(\frac{\Delta(V x_v)}{\Delta t}\right) = \lim_{t \to 0} \frac{\theta_x}{w_v}\left(\frac{d(V x_v)}{dt}\right) = \frac{\theta_x}{w_v}(Q_1 x_v) \tag{7}
$$

where w_v (mgVSSa/L) is the mixed-liquor volatile suspended solids (MLVSS or holdup) and $Q_1 x_v$ is the mass production rate of the total VSS (active, inerts, soluble microbial products (SMP)).

3.1.2 ASSUMPTIONS

1. There is no active biomass in the input stream, thus $x_p^0 = 0$.
2. There are no TKN and organic carbons present in the effluent stream of the nitrifier. Thus, Snh_2=0, S_2=0.
3. There is no nitrate-ion in the effluent stream of the settler, thus Sno_e=0 but Sno_w =Sno_2.
4. Only f_d of the mixed liquor (x_{v2}, x_w) returned to the denitrifier in either recycle streams is active biomass.

Applying the assumptions and simplifying the system of equations in (4) give,

$$\begin{aligned}
\frac{dx_{P1}}{dt} &= \frac{Q_2}{V_D} f_d\left(\rho x_{v2} + \nu(1-\rho)x_w\right) - \frac{Q_1}{V_D} x_{P1} + \left(Y_P r_P - b_P x_{P1}\right) \\
\frac{dS_1}{dt} &= \frac{S^0}{\theta_D} - \frac{Q_1}{V_D} S_1 - r_P \\
\frac{dSno_1}{dt} &= \frac{Q_2}{V_D}\left(\rho + \nu(1-\rho)\right)Sno_2 - \frac{Q_1}{V_D} Sno_1 - r_{no} \qquad (8) \\
\frac{dSnh_1}{dt} &= \frac{Sn^0}{\theta_D} - \frac{Q_1}{V_D} Snh_1 - r_{nh} \\
\frac{dx_{i1}}{dt} &= \frac{x_i^0}{\theta_D} - \frac{Q_1}{V_D} x_{i1} + (1-f_d)b(x_{v2} + x_w + x_{P1})
\end{aligned}$$

Design variables are recycle fractions (ρ, ν), mean cell retention time, θ_{xD}, MLVSS, w_{vD}, Q_1, Q^0. Reasonable choices of ρ,$(1-\nu)(1-\rho)$ are 6 and 0.25, respectively. In the case of HDT and MCRT, good choices are 0.75 and 15 days, respectively.

REMARKS

1. Nitrogen balance on the nitrifier will indicate the amount of nitrate-ion (Sno_2) in stream 2 (effluent from nitrifier) that is an input stream to the denitrifier. Define

$$\gamma = \frac{14\,\text{mgN}}{113\,\text{mgVSS}} \qquad (9)$$

that represents the mass of nitrogen present in the biomass. A steady-state component balance on the nitrate-ion in the effluent of the reactor is given by,

$$\begin{aligned}
Q_2 Sno_2\,(\text{mgNO3/d}) &= Q^0 Sn^0 - \left(Q_1 x_{v1}^a + Q_2 x_{v2}^a\right)\gamma \\
x_{v1}^a &= x_{P1} + x_{i1} - x_i^0 \qquad (10) \\
x_{v2}^a &= x_{H2} + x_{N2} + x_{i2} - x_i^0
\end{aligned}$$

The dynamic balance is given by,

$$V_N \frac{dSno_2}{dt} = Q^0 Sn^0 - Q_1 x_{vD} \gamma - Q_2 Sno_2 - (r_N - r_H) V_N \qquad (11)$$

In the dynamic balance equation, r_N and r_H are the utilization rates of the nitrate-ion by the nitrifiers and the aerobic heterotrophs.

2. The steady-state value of the ammonium-ion is a function of amount of organic substrate consumed. By stoichiometry,

$$\bar{S}nh_1 = \frac{(\frac{f_s}{20} - \frac{1}{50})}{50} (\frac{14\,\text{mgN}}{201\,\text{mgBOD}_L}) \text{ (substrate consumption rate)} \qquad (12)$$

3. The utilization, r_H, of the organic substrate, $C_{10}H_{19}O_3N$, by the aerobic heterotroph is based on the amount of nitrate-ion present. Thus, the nitrate-ion is the limiting component.

The total sludge leaving the denitrifier is given by:

$$\frac{\Delta(V_D x_{vD})}{\Delta t} (\text{mgVSS/d}) = Q_1 x_v = \frac{w_v Q^0 \theta_D}{\theta_{xD}} = \frac{w_v V_D}{\theta_{xD}} \qquad (13)$$

3.1.3 STEADY STATE

The steady state of the denitrifier (variables with an overbar) is given by,

$$\bar{x}_{P1} = \left(\frac{Q_2}{V_D} f_d (\rho \bar{x}_{v2} + (1-\nu)(1-\rho)) \bar{x}_w - Y_P (\frac{S^0}{\theta_D} - \frac{Q_1 \bar{S}_1}{V_D}) \right) \frac{V_D}{Q_1 + b_P V_D}$$

$$\bar{S}_1 = \frac{-(\frac{Q^0 S^0}{Q_1} - K_P - \bar{q}_P \bar{x}_{P1} \frac{V_D}{Q_1}) \pm \sqrt{(\frac{Q^0 S^0}{Q_1} - K_P - \bar{q}_P \bar{x}_{P1} \frac{V_D}{Q_1})^2 - 4 \frac{Q^0 S^0 K_P}{Q_1}}}{2}$$

$$\bar{S}no_1 = \frac{-(\frac{Q_2}{Q_1} \bar{S}no_2 - K_{no} - \bar{q}_{no} \bar{x}_{P1} \frac{V_D}{Q_1}) \pm \sqrt{(\frac{Q_2}{Q_1} \alpha \bar{S}no_2 - K_{no} - \bar{q}_{no} \bar{x}_{P1} \frac{V_D}{Q_1})^2 - 4\alpha \frac{Q_2 \bar{S}no_2 K_{no}}{Q_1}}}{2} \qquad (14)$$

$$\alpha = (\rho + \nu(1-\rho))$$

$$\bar{S}nh_1 = \frac{-(\frac{Q^0}{Q_1} Sn^0 - K_{nh} - \bar{q}_{nh} \bar{x}_{P1} \frac{V_D}{Q_1}) \pm \sqrt{(\frac{Q^0}{Q_1} Sn^0 - K_{nh} + \bar{q}_{nh} \bar{x}_{P1} \frac{V_D}{Q_1})^2 - 4 \frac{Q^0 Sn^0 K_{nh}}{Q_1}}}{2}$$

$$\bar{x}_{i1} = \frac{V_D}{Q_1} \left(\frac{x_i^0}{\theta_D} + (1 - f_d) b_P (\bar{x}_{v2} + \bar{x}_{vw} + \bar{x}_{P1}) \right)$$

3.2 NITRIFIER

Nitrogen supply to the nitrifier starts as effluent TKN from the denitrifier (amount not used by the *pseudomonas*). TKN hydrolyzes to ammonium, $NH_4^+ - N$, and nitrite, $NO_2^- - N$. Ammonium is used in the nitrifier by all microorganisms, heterotrophs, *nitrosomonas* and *nitrobacters* for cell synthesis; but *nitrosomonas* also uses ammonium for energy while *nitrobacters* uses nitrite for energy. $NH_4^+ - N$ (S_{nh}) is the limiting substrate for *nitrosomonas*, $NO_2^- - N$ (S_{no}) is the limiting substrate for *nitrobacters*. The amount of TKN nitrogen available for *nitrobacters* is a function of the amount that is unused by the heterotrophs, *nitrosomonas* and any inerts (or SMP). Heterotrophs: x_H, *Nitrosomonas*: x_s, *Nitrobacters*: x_B, organic carbon: S

3.2.1 COMPONENT BALANCES

$$
\begin{aligned}
V_N \frac{dx_{H2}}{dt} &= Q_1 x_H^0 - Q_2 x_{H2} + (Y_H r_H - b_H x_{H2}) V_N \\
V_N \frac{dS_2}{dt} &= Q_1 S_1 - Q_2 S_2 - r_H V_N \\
V_N \frac{dx_{S2}}{dt} &= Q_1 x_{S1} - Q_2 x_{S2} + (Y_S r_S - b_S x_S) V_N \\
V_N \frac{dSnh_2}{dt} &= Q_1 Snh_1 - Q_2 Snh_2 - V_N r_S \\
V_N \frac{dx_{B2}}{dt} &= Q_1 x_{B1} - Q_2 x_{B2} + (Y_B r_B - b_B x_{B2}) V_N \\
V_N \frac{dSno2_2}{dt} &= Q_1 Sn_1 - Q_2 Sno2_2 - V_N r_B \\
V_N \frac{dx_{i2}}{dt} &= Q_0 x_i^0 - Q_2 x_{i2} + (1 - f_d)(b_H x_{H2} + b_N x_{N2} + b_B x_{B2}) V_N
\end{aligned}
\tag{15}
$$

The utilization terms r_H, r_S, and r_B are given by:

$$
\begin{aligned}
r_H &= \frac{\hat{q}_H S_2}{K_H + S_2} x_{H2} \\
r_S &= \frac{\hat{q}_S Snh_2}{K_S + Snh_2} x_S \\
r_B &= \frac{\hat{q}_B Sno2_2}{K_B + Sno2_2} x_{B2}
\end{aligned}
\tag{16}
$$

The total sludge leaving the denitrifier is given by:

$$
\frac{\Delta(V_N x_{vN})}{\Delta t}(\text{mgVSS/d}) = Q_2 x_v = \frac{w_{vN} Q_1 \theta_N}{\theta_{xN}} = \frac{w_v V_N}{\theta_{xN}} \tag{17}
$$

The difference between the influent nitrogen and the amount of biomass (heterotrophs and *nitrosomonas* and the biodegradable inerts) produced indicates how much nitrate-ion remains for cell maintenance of the *nitrobacters*.

$$V_N \frac{dSno2_2}{dt} = Q_1\left(Sn_1 - \frac{\theta_N}{\theta_{xN}}(\bar{x}_H + \bar{x}_S + \bar{x}_i)\right)\gamma - Q_2 Sno2_2 - V_N r_B \qquad (18)$$

The amount of unused nitrogen is given by

$$\bar{S}n_2 = \bar{S}n_1 - \frac{\theta_N}{\theta_{xN}}(\bar{x}_a + \bar{x}_S + \bar{x}_B + \bar{x}_i)\gamma - \bar{S}nh_1 - \bar{S}no2_2 \qquad (19)$$

3.2.2 STEADY STATE

At steady state, the above equations can be solved to give,

$$\begin{aligned}
\bar{x}_H &= \frac{\theta_{xN}}{\theta_N}\left(\frac{\bar{S}_1 - \bar{S}_2}{1 + b_H\theta_{xN}}\right)Y_H \\
\bar{S}_2 &= \frac{K_H(1 + b_H\theta_{xN})}{\theta_{xN}(Y_H\hat{q}_H - b_H) - 1} \\
\bar{x}_S &= \frac{\theta_{xN}}{\theta_N}\left(\frac{Snh_1 - \bar{S}nh_2}{1 + b_S\theta_{xN}}\right)Y_S \\
\bar{S}nh_2 &= K_S\frac{(1 + b_S\theta_{xN})}{\theta_{xN}(Y_S\hat{q}_S - b_S) - 1} \\
\bar{x}_B &= \frac{\theta_{xN}}{\theta_N}\left(\frac{\bar{S}no2_1 - \bar{S}no2_2}{1 + b_B\theta_{xN}}\right)Y_B \\
\bar{S}no2_2 &= K_B\frac{(1 + b_B\theta_{xN})}{\theta_{xN}(Y_B\hat{q}_B - b_B) - 1} \\
\bar{x}_{i2} &= \frac{\theta_{xN}}{\theta_N}\left(x_{i1}^0 + \theta_N(1 - f_d)(b_H\bar{x}_H + b_S\bar{x}_S + b_B\bar{x}_B\right)
\end{aligned} \qquad (20)$$

3.3 SETTLER

Because a net growth of microorganisms is obtained, the net growth called excess sludge or waste sludge is removed from the system for subsequent sludge treatment and disposal. It is crucial that the quantity of waste sludge (bio-solids) produced must be removed continually in order to maintain the steady-state conditions. The rate of sludge wasting is essential for operating the treatment system and for determining the total cost of construction and operation of the system.

3.3.1 MATERIAL AND COMPONENT BALANCES

$$\frac{dV_T}{dt} = (1-\rho)Q_2 - (Q_e + Q_w) - (1-\nu)(1-\rho)Q_2$$

$$V_T \frac{dSno2_e}{dt} = (1-\rho)Q_2 Sno2_2 - (Q_e + Q_w + (1-\nu)(1-\rho)Q_2)Sno2_e \qquad (21)$$

$$V_T \frac{dx_{vw}}{dt} = (1-\rho)Q_2 x_{v2} - (Q_e + (1-\nu)(1-\rho)Q_2)x_{vw}$$

3.3.2 ASSUMPTIONS

1. No unused organic carbon substrate or nitrogen (in the form of ammonium-ion or nitrite-ion).
2. The concentration of nitrate-ion is the same in the waste, recycle, and effluent streams.
3. The concentration of biomass is the same in the recycle and waste streams with no biomass losses in the effluent stream.

3.3.3 STEADY STATE

$$\begin{aligned} Q_e + Q_w &= \nu(1-\rho)Q_2 \\ Sno2_e &= Sno2_2 \qquad (22) \\ x_{vw} &= \frac{(1-\rho)}{(1-\nu)(1-\rho)Q_2 + Q_w} Q_2 \end{aligned}$$

4. SUMMARY

Models of the prednitrifying system were developed in the Mathworks (Natick, MA) Matlab® and simulated to study parameter sensitivities and design considerations.

NOMENCLATURE

BAP	Biomass associated products
BOD	Biochemical oxygen demand
COD	Chemical oxygen demand
$C_{10}H_{19}O_3N$	Organic carbon substrate
$C_5H_7O_2N$	Cell or biomass empirical formula
HDT	Hydraulic detention time
MCRT	Mean cell retention time
MLVSS	Mixed liquor volatile suspended solids, mgVSS/L
OD	Oxygen demand (oxygen equivalents)
SMP	Soluble microbial products
SRT	Solids retention time
TKN	Total Kjeldahl nitrogen
UAP	Utilization associated products
VSS,VSSa	Volatile suspended solids, active volatile suspended solids
b	Endogenous rate of decay coefficient, 1/d
d	Day
f_e	Fraction of electrons used for energy
f_d	Fraction of active biomass that is biodegradable
f_s	Fraction of electrons used for synthesis
f_s^0	Total number of electrons transferred to the electron acceptor
r_k	Rate of utilization by biomass component K (K=B,H,P,S)
K	Substrate concentration that give ½ the maximum rate, mgBOD$_L$/L
$\hat{q}$	Maximum specific rate of substrate utilization, mgBOD$_L$/mgVSS-d
Q^0	Inlet volumetric feed rate, L/d
Q_j	Volumetric rate of stream j, L/d
r	Rate of substrate utilization, mgBOD$_L$/L-d
R_e	Energy reaction
R_d	Donor reaction
R_s	Cell synthesis reaction
S^0	Organic substrate influent concentration, mgBOD$_L$/L
S_j	Substrate concentration in stream j, mgBOD$_L$/L
Sno	Nitrate-ion concentration, mgNO3-N/L
$Sno2$	Nitrite-ion concentration, mgNO2-N/L
Snh	Ammonium concentration, mgNH4-N/L
Sn^0	Influent nitrogen concentration mgN/L
x_{kj}	kth biomass component in stream j, mgVSSa/L
x_i^0	Influent inert concentration, mgVSS/L
x_{vj}^a	Active volatile suspended solids in stream j, mgVSS/L
x_i	Non-biodegradable portion of the biomass, mgVSS/L
x_B	Concentration of biomass associated products, mgBOD/L
x_U	Concentration of utilized associated products, mgBOD/L
V_D, V_N, V_T	Denitrifier volume, nitrifier volume, settler volume, L/d

Y	True cell synthesis yield, mgVSS/mgBOD_L
Y_n	Net yield, mgVSS/mgBOD_L

GREEK LETTERS

ρ	Fraction of effluent of nitrifier that is recycled to the denitrifier
$(1-\nu)$	Fraction of influent to the settler that is recycled to the denitrifier
γ	Ratio of ½ molecular weight of nitrogen to that of biomass
θ_D, θ_N	Hydraulic detention time (HDT) in denitrifier and nitrifier, d
θ_x	Mean cell retention time (MCRT) or solids retention time (SRT), d
$\theta_{x,min}$	Min MCRT) or min SRT, d
$[\theta_{x,min}]_{lim}$	Lower bound on the min MCRT) or min SRT, d

SUBSCRIPTS

e	Effluent of the settler
v	Volatile suspended solids
w	Waste stream
B	*Nitrobacters,* biomass associated SMP
D	Denitrifier
H	Aerobic heterotrophs (nitrifier)
N	Nitrifier
P	*Pseudomonas*
S	*Nitrosomonas*
T	Settler
U	Utilization associated SMP

REFERENCES

Brogan, W. (1991) Modern Control Theory, 3rd ed, Prentice-Hall, Engelwood Cliff, NJ.

Campbell, M., L. Vega, E. Ungar, and K. Pickering (2003) "Development of a Gravity Independent Nitrification Biological Water Processor," 2003-01-2560.

Campbell, M., B. Finger, C. Verostko, K. Wines, G. Pariani, and K. Pickering (2003) "Integrated Water Recovery System Test," SAE International, 2003-01-2577.

Crew and Thermal Systems Div (EC3) (2000) "Advanced Water Recovery System (WRS) Integrated Test Plan," CTSD-ADV-2163, JSC 39893, Rev. A.

Dold, P.L., G. A. Ekama, and G. vR. Marais (1980) "A General Model for Activated Sludge Process," Prog. Wat. Tech., 12, pp 47-77.

Drysdale, G., H. Kasan, and F. Bux (1999) "Denitrification by Heterotrophic Bacteria During Activated Sludge Treatment," Water SA, 25(3), pp 357-362.

Gamble, T., M. Betlach, and J. Tiedje (1977) "Numerically Dominant Denitrifying Bacteria from World Soils," Appl. & Env. Microbio., 33(4), pp 926-939.

Hart, S. , P. Currier, and D. Thomas (2000) "Denitrification by *Pseudomonas aeruginosa* Under Simulated Engineering Martian Conditions," ISIS conference.

Kissel, J. (1986) "Modeling Mass Transfer in Biological Wastewater Treatment Processes," Wat. Sci. Tech., 18, pp 35-45.

Muirhead, D., T. Rector, W. A. Jackson, H. Keister, A. Morse, K. Rainwater, and K. Pickering (2003) "Performance of a Small Scale Biological Water Recovery System," SAE International, 2003-01-2557.

Overland, D., M. Ciskowski, I. Cavanaugh, L. Labuda, F. Mount, M. Whitmore, R. Gomez, H. Slade, and J. Park (2004) "Advanced Integration Matrix Team Bravo White Paper: Investigation of Advanced Control Architecture Issues and Technology Gaps," June 3.

Pickering, K., K. Wines, G. Pariani, L. Franks, J. Yeh, M. Campbell, "Early Results of an Integrated Water Recovery System Test," SAE International, 2001-01-2210.

Rittman, B. E. and P. L. McCarty (2001) Environmental Biotechnology, McGraw-Hill, New York, NY.

Rector, T., W. Jackson, and K. Rainwater (2003) "Determination of the Fate and Behavior of a Commerical Surfactant in a Water Recycle System (WRS)," SAE International, 2003-01-2558.

Sakano, Y., K. Pickering, P. Strom, and L. Kerkhof (2002) "Spatial Distribution of Total Ammonia-Oxidizing and Denitrifying Bacteria in Biological Wastewater Treatment Reactors for Bioregenerative Life Support," Appl. & Env. Microbio., 68(5), pp 2285-2293.

Varma, A. and M. Morbidelli (1997) Mathematical Methods in Chemical Engineering, Oxford Press., UK.

Vega, L. (2004) "Hydraulic Retention Time (HRT) as a Factor in Total Organic Carbon (TOC) Degradation in a Small-scale Anaerobic Bioreactor," *private communication*.

Wentzel, M., M. Ubisi, M. Lakay, and G. Ekama (2002) "Incorporation of Inorganic Material in Anoxic/Aerobic Activated Sludge System Mixed Liquor," Wat. Res., 36, pp 5074-5082.

ADAPTABLE CONSTRAINED GENETIC PROGRAMMING: EXTENSIONS AND APPLICATIONS

Final Report

NASA Faculty Fellowship Program – 2004

Johnson Space Center

Prepared By:	Cezary Z. Janikow, Ph. D.
Academic Rank:	Associate Professor
University and Department:	University of Missouri - St. Louis Department of Mathematics and Computer Science St. Louis, Missouri 63121
NASA/JSC	
Directorate:	Engineering
Division:	Automation, Robotics and Simulation
Branch:	Intelligent Systems
JSC Colleague:	Dennis Lawler
Date Submitted:	July 22, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

An evolutionary algorithm applies evolution-based principles to problem solving. To solve a problem, the user defines the space of potential solutions, the representation space. Sample solutions are encoded in a chromosome-like structure. The algorithm maintains a population of such samples, which undergo simulated evolution by means of mutation, crossover, and survival of the fittest principles. Genetic Programming (GP) uses tree-like chromosomes, providing very rich representation suitable for many problems of interest. GP has been successfully applied to a number of practical problems such as learning Boolean functions and designing hardware circuits.

To apply GP to a problem, the user needs to define the actual representation space, by defining the atomic functions and terminals labeling the actual trees. The sufficiency principle requires that the label set be sufficient to build the desired solution trees. The closure principle allows the labels to mix in any arity-consistent manner. To satisfy both principles, the user is often forced to provide a large label set, with *ad h*oc interpretations or penalties to deal with undesired local contexts. This unfortunately enlarges the actual representation space, and thus usually slows down the search. In the past few years, three different methodologies have been proposed to allow the user to alleviate the closure principle by providing means to define, and to process, constraints on mixing the labels in the trees. Last summer we proposed a new methodology to further alleviate the problem by discovering local heuristics for building quality solution trees. A pilot system was implemented last summer and tested throughout the year.

This summer we have implemented a new revision, and produced a User's Manual so that the pilot system can be made available to other practitioners and researchers. We have also designed, and partly implemented, a larger system capable of dealing with much more powerful heuristics.

INTRODUCTION

Genetic programming (GP), proposed by Koza [1], is an evolutionary algorithm, and thus it solves a problem by utilizing a population of solutions evolving under limited resources. The solutions, called chromosomes, are evaluated by a problem-specific user-defined evaluation method. They compete for survival based on this evaluation, and they undergo simulated evolution by means of simulated crossover and mutation operators.

GP differs from other evolutionary methods by using trees to represent potential problem solutions. Trees provide a rich representation that is sufficient to represent computer programs, analytical functions, and variable length structures, even hardware circuits. The user defines the representation space by defining the set of functions and terminals labeling the nodes of the trees (for the internal and the external nodes, respectively). One of the foremost principles is that of *sufficiency* [1], which states that the function and terminal sets must be sufficient to solve the problem. The reasoning is obvious: every solution will be in the form of a tree, labeled only with the user-defined elements. In the absence of specific knowledge and heuristics, sufficiency will usually force the user to artificially enlarge the sets to avoid missing some important elements. This unfortunately dramatically increases the search space.

Even disregarding sufficiency, GP practitioners still face another problem. Consider a 3-argument *if* function (corresponding to the *if-else* conditional statement in any programming language). This function should have a test argument, and then two action arguments. But GP has no way of defining or processing this knowledge. To allow GP to operate nevertheless, Koza has proposed the principle of *closure* [1], which requires very elaborate semantic interpretations to ensure the validity of any arity-consistent label in any context. Structure-preserving crossover was introduced as the first attempt to handle such specific local constraints [1].

Structure-preserving crossover wasn't a generic method. In the nineties, three independent generic methodologies were developed to allow problem-independent constraints on tree construction. Montana proposed STGP [5], which used types to control the way functions and terminals can label local subtrees. For example, the function *if* can be required to use Boolean-producing subtrees on its first argument.

We proposed CGP, developed at NASA/JSC during summer research, which originally required the user to explicitly specify allowed and/or disallowed local tree structures [2]. These local constraints could be based on types, but also on some problem specific heuristics. In a follow-up version, we also added explicit type-processing capabilities, with polymorphic functions. For example, the + function could be overloaded so that it produces an integer from integers but it produces an angle from angles.

Finally, those interested in program induction following specific syntax structure have used similar ideas to propose CFG-based GP [6]. However, those systems still require the user to enter all possible constraints that can be processed within the methodology. What happens if the user is not aware of any, not aware of the best constraints to solve a particular problem, or aware of the constraints but not of some rule-of-a-thumb heuristics? To deal with this case, last summer we introduced Adaptable Constrained

Genetic Programming v1.1 (ACGP1.1), which is a methodology (and a pilot implementation) to automatically adapt the user-specified constraints and heuristics to improve GP's performance. This summer, we have improved that implementation (ACGP1.1.1), and provided a User's Manual, so that the system can be distributed to interested practitioners and researchers.

ACGP1.1 discovers limited local heuristics and only on labels. This summer, we have designed a new methodology, ACGP2.1, which is capable of discovering much richer heuristics on labels, types, and combinations of these. We have implemented the system, capable of discovering seven different heuristics. In the next few months, we plan to build tools to analyze the heuristics, and then to use them to improve GP's searching capabilities.

ACGP HEURISTICS

ACGP is a methodology to discover useful heuristics on GP solution trees. Such heuristics, if available, have been shown to greatly enhance GP problem solving capabilities [2]. ACGP discovers the heuristics by observing the distribution of labels (and types if applicable) in those better-off solutions, assuming that those solutions are better because they were generated with better distribution of local heuristics on average. ACGP1.1.1 deals with limited heuristics on labels, while the newer ACGP2.1 handles heuristics on both labels and types.

ACGP1.1.1 and First-Order Label-Based Heuristics

Last year we have developed ACGP1.1, which processes heuristics on labels only, while limited to parent-child relationship only. That is, ACGP1.1 doesn't process type information, and it cannot discover any heuristics taking a node's siblings into account. We call this kind of limited heuristics the *first-order* heuristics, as illustrated in Figure 1 center. *Zero-order* heuristics use one node at a time, while *second-order* heuristics take all siblings into account.

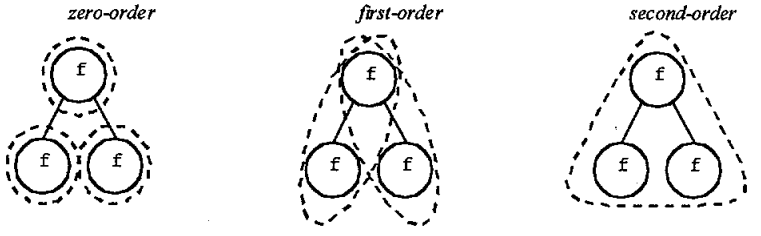


Figure 1: The three different levels of heuristics. Note that zero-order heuristics are meaningless if the only node information is the node's label.

ACGP1.1.1 does not process type information, and thus it only uses a function/terminal label in a node. This is illustrated in Figure 2 (left). Apparently, this limited data provides no information for zero-order heuristics. ACGP1.1.1 does not employ any second-order heuristics either, as it was used as a proof of concept. Even the limited first-order

heuristics have been demonstrated to both allow the user to discover very useful knowledge and to improve problem-solving capabilities [3][4].

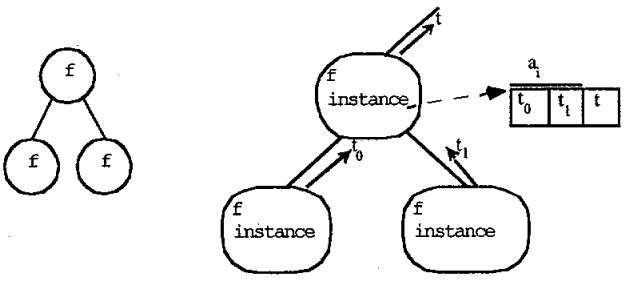


Figure 2: Information available in a tree node: left in ACGP1.1.1 (label only), right in ACGP2.1 (label, type generated, and the polymorphic function instance used).

ACGP2.1

ACGP2.1 uses both labels and types, with polymorphic functions. Therefore, the information retained in a tree node is much richer, as illustrated in Figure 2 (right): each node retains the label (function for an internal node and terminal for an external node) and the specific polymorphic instance (for functions only – terminals are not overloaded). For example, in Figure 2 right, the top node retains the information that the function used in the node is **f**, and that the function generates type **t**, using its overloaded instance which requires types $\mathbf{t_0}$ and $\mathbf{t_1}$ on the two arguments, left to right respectively.

The richness of this information allows useful zero-order heuristics. For example, the same node in Figure 2, if expressed disproportionally in the better solutions, would suggest that **f** should use this polymorphic instance whenever it generates type **t**. ACGP2.1 uses a number of heuristics of all three kinds: zero, first, and second-order. Moreover, it uses heuristics on labels only (for typeless applications), on types only, and on combinations of these. Example heuristics are illustrated in Figures 3-6.

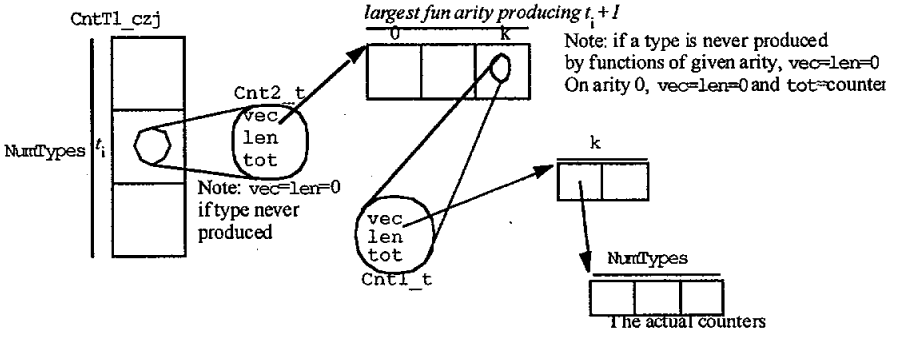


Figure 3: First-Order Type-Based Heuristics CntT1.

Figure 3 illustrates a very specific first-order heuristic on types only. This particular heuristic is capable of discovering, for example, that if a given type needs to be generated

(t_i), what arity function would be most beneficial to generate it, and for that arity, what should be the types of all of the arguments.

Figure 4 illustrates a specific first-order heuristic on labels only. This heuristic is capable of discovering what should be the labels of all the children, independently, of a function such as f_i. This heuristic alone is in fact equivalent to all the capabilities available in ACGP1.1.1.

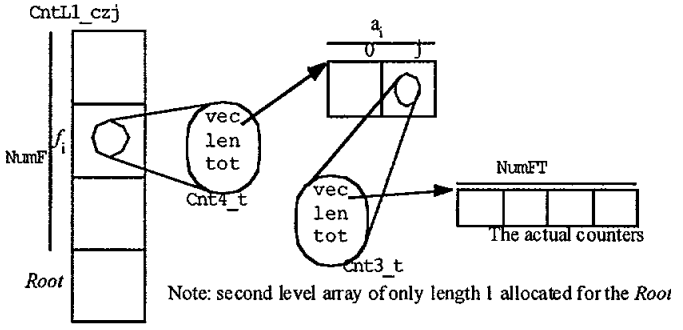


Figure 4: First-Order Label-Based Heuristics CntL1.

When we combine labels and types, even one node provides some meaningful information, resulting in combined zero-order heuristics. One such heuristic is illustrated in Figure 5. It is capable of discovering that if a given type needs to be generated from a node (such as t_i), what functions/terminals are most likely to generate it successfully.

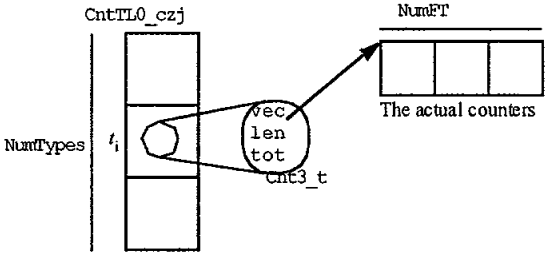


Figure 5: Zero-order Combined Heuristics CntTL0.

Figure 6 illustrates another combined zero-order heuristic. This one is capable of discovering that if a given node needs to be labeled with a function such as f_i, which polymorphic type it should use to be most successful (terminals are not polymorphic).

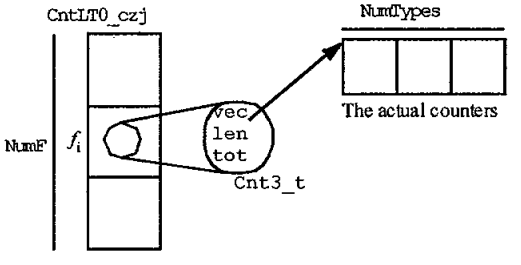


Figure 6: Zero-order Combined Heuristics CntLT0.

Currently, ACGP2.1 discovers seven such heuristics, printing them into 28 files (using different scenarios). In the near future, tools to analyze, and to use the heuristics in improving the search, are needed.

REFERENCES

[1] Koza, J. R. 1994, *Genetic Programming. On the Programming of Computers by Means of Natural Selection*, Massachusetts Institute of Technology.

[2] Janikow, Cezary Z. "A Methodology for Processing Problem Constraints in Genetic Programming". *Computers and Mathematics with Applications*, Vol. 32, No. 8, pp. 97-113, 1996.

[3] Janikow, Cezary Z. "*ACGP: Adaptable Constrained Genetic Programming*". *Proceedings of GPTP04* TBP.

[4] Janikow, Cezary Z. "*Adapting Representation in Genetic Programming*". *Proceedings of GECCO 2004.*

[5] Montana, D. J 1995, "*Strongly Typed Genetic Programming*", *Evolutionary Computation*, Vol. 3, No. 2.

[6] Whigham, P.A.. "*Grammatically-based genetic programming*". In J. P. Rosca, editor, *Proceedings of the Workshop on Genetic Programming: From Theory to Real-World Applications*, pages 33–41, Tahoe City, California, USA, 9 1995.

Systems Engineering and Integration for Advanced Life Support System and HST

Final Report
NASA Faculty Fellowship Program – 2004

Johnson Space Center

Prepared by:	Ali K. Kamrani, Ph.D.
Academic Rank:	Associate Professor
University & Department	University of Houston Industrial Engineering Houston, TX 77204
NASA/JSC	
Directorate: Engineering	
Division: Crew & Thermal Systems	
Branch: N/A	
JSC Colleague:	Donald L. Henninger, Ph.D.
Date Submitted:	August 31, 2004
Contract Number:	NAG 9-1526&NNJ04JF93A

ABSTRACT

Systems engineering (SE) discipline has revolutionized the way engineers and managers think about solving issues related to design of complex systems. With continued development of state-of-the-art technologies, systems are becoming more complex and therefore, a systematic approach is essential to control and manage their integrated design and development. This complexity is driven from integration issues. In this case, sub-systems must interact with one another in order to achieve integration objectives, and also achieve the overall system's required performance. Systems engineering process addresses these issues at multiple levels. It is a technology and management process dedicated to controlling all aspects of system life cycle to assure integration at all levels.

The Advanced Integration Matrix (AIM) project serves as the systems engineering and integration function for the Human Support Technology (HST) program. AIM provides means for integrated test facilities and personnel for performance trade studies, analyses, integrated models, test results, and validated requirements of the integration of HST. The goal of AIM is to address systems-level integration issues for exploration missions. It will use an incremental systems integration approach to yield technologies, baselines for further development, and possible breakthrough concepts in the areas of technological and organizational interfaces, total information flow, system wide controls, technical synergism, mission operations protocols and procedures, and human-machine interfaces.

This report provides the summary of results based on the proposed SOW during the 2004 fellowship at NASA's Johnson Space center for NFFP. These tasks were:

1. Benchmarking and the evaluation of systems engineering processes in order to identify best practices and lesson learned.
2. Propose a SE process template for the identification of functional requirements and its decomposition for human life support systems.

INTRODUCTION - ADVANCED INTEGRATION MATRIX (AIM)

The Advanced Integrated Matrix (AIM) project attempts to provide SE&I services to highly advanced and complex projects (e.g. missions beyond Lower Earth Orbit or LEO) through ground-based testing and integration. The goal of the AIM is to develop the enabling environment and tool for gap analysis and commonality. The roadmap to AIM is through incremental implementation. The incremental approach evolving from a single-enterprise into a multi-enterprise, multi-center concern focused on developing and testing integrated mission concepts. AIM will initiate with projects with low level of complexity for integration and testing and then gradually evolve into a full mission scenario through "***fly the mission on the ground***" concept. Figure 1 conceptualizes this concept.

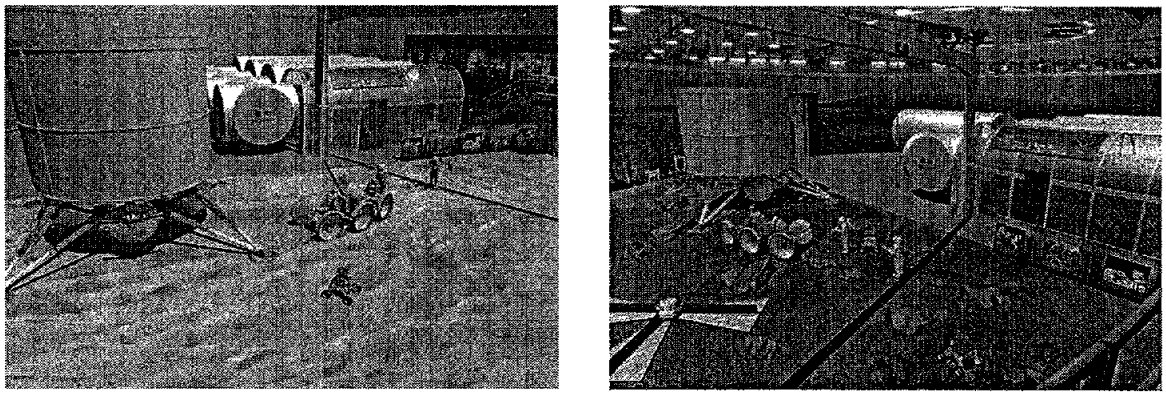


Figure 1. Moon and Mars *"fly the mission on the ground"* Configurations[1]

This incremental approach will generate the lessons learned and baselines for further designs of more complex systems. The possibility of identifying new breakthrough managements concepts and technology to support interfaces, information flow and sharing, managements, controls, operations and procedures, and man-machine interfaces are the goal of the this approach. This would require participations from different programs in the agency. AIM will[2]:

- Address system-level integration and interface issues to support agency commitment to an exploration mission
- Investigate issues common to multiple vehicles, architectures, and mission scenarios, and develop solutions in a scalable format
- Aggressively pursue participation from academia and other NASA Centers to address Agency's Strategic Plan.

AIM will collect the scientific and technological knowledge of individual projects into an integrated ground-based test environment where system-level interactions are studied and optimized for commonality, performance efficiency, cost and mass savings. The Office

[1] D. Henninger, *Integrity: A Program Concept*, Johnson Space Center, NASA, 2002.
[2] G. Thomas, *AIM Project Quarterly Report*, AHST Program, Johnson Space Center, NASA, 2003.

of Biological and Physical Research (OBPR) has authorized initiation of the formulation phase. The purpose is to identify and solve system-level integration issues for exploration missions beyond Low Earth Orbit through design and development of a ground-based facility for developing an integrated system for joint human-robotic missions.

PART I – SYSTEMS ENGINEERING PROCESS AND GENERIC MODELS

Past experience has shown that lack of planning and clear identification of objectives has been the major problem associated with the design and development of any complex system. This approach has resulted in a system's lack of performance and finally design failure. Traditionally, systems have been developed based on "***Deliver now and Fix Later.*** [3]" This process has suffered from lack of clear planning which resulted in failure and high cost of design modifications. In this scenario, requirements at the systems level were kept general in order to reduce complexity to allow for new technology integration. This has routinely evolved into last minute modifications that impacted the schedule and cost. Decisions made at the early stages of development life cycle will significantly impact the overall life-cycle including cost and system's effectiveness. Therefore, there is a need for a disciplined approach for integrated design and the development of new systems. In this case, all aspects of the development are considered early in the process and used for continuous improvement. Systems Engineering is "*The effective application of scientific and engineering efforts to 1) transform an operational need into a defined system configuration through the top-down iterative process of requirement analysis, functional analysis and allocation, synthesis, design optimization, test, evaluation and validation, 2) integrate related technical parameters and ensure the compatibility of all physical, functional, and program interfaces in a manner that optimizes the total definition and design, and 3) integrate reliability, maintainability, usability, safety, serviceability, disposability to meet cost, schedule, and technical performance objectives*[4]. Systems engineering is also considered a process for Managing Technology. System engineering process has evolved in seven different paradigms[5]. A summary of these processes are discussed below.

1. **Build–Test–Fix:** This method consists of three basic steps, *fabricate a design, test the system, and then operate.* This method is typically used for in-house software development where the customer is also the developer. It is considered to be a simple but effective approach. Although, it lacks the requirements analysis phase that makes it not suitable for any complex systems design.
2. **Staircase:** The Staircase method allows for better management and control of system development. This method is considered to be a systematic flow through the SE process. It is well suited for the developments of existing systems variants. In this

[3] Benjamin S. Blanchard, *Systems Engineering Management*, Wiely Interscience, 1998.
[4] Systems Engineering Fundamentals, Defense Acquisition University Press, 2001.
[5] Norman B. Reilly, *Successful Systems Engineering for Engineers and Managers*, Kluwer Academic Publishers, 1993.

case, already developed requirements are modified. The export version of a military aircraft is an example this concept. **R**equirements-**S**pecification-**D**esign-**F**abrication-**T**esting-**A**cceptance-**O**perate area steps in the staircase SE cycle.

3. **Waterfall:** This method improves the staircase method by adding feedback loops between successive stages as shown in Figure 2. Through these feedbacks, each stage is capable of gaining knowledge from the subsequent stages. The success of this model is dependant on understanding and processing revisions through feedback analysis.

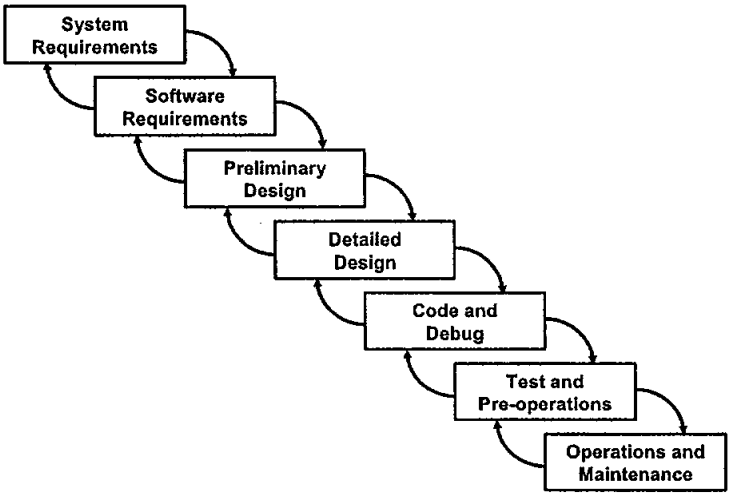


Figure 2. Waterfall Systems Engineering Process Model

4. **Early Prototype:** The early prototype process is an extension to the staircase with feed back cycle. This method is used when it is difficult for customer to identify requirements, but could recognize them through some model or prototype representation. The advantage of this method is due to direct interaction between stakeholders. Some of the difficulties with this approach are 1) the initial prototype could discourage the customer 2) it suggests an unattainable goals and 3) prototype design becomes the main objective rather than the actual customer's need.
5. **Spiral:** The Spiral method, as shown in Figure 3 is an extension to the early prototype concept. The primary advantage of the spiral method is the detailed development of requirements, specifications, and designs. The significant challenge for the spiral method is managing the prototype transitions. Some of the advantages are:
   - Risk-driven sequential phases with user involvement.
   - Considers highest risks issue first (Requirements understanding, Technical feasibility and System operations).
   - Cycles of risk-driven phases, spiral around and end with a final waterfall wrap.
6. **Rapid Development:** The success of this process dependent on fast-paced innovation (**RSDFTAO--- RSDFTAO...**) while completing multiple small starts to the final system. This process requires cross-functional teams with the ability to work across the functional boundaries.

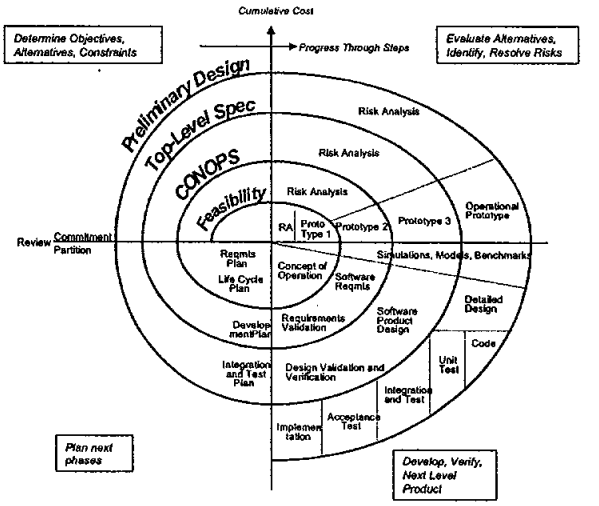


Figure 3. Systems Engineering Spiral model[6]

7. **Integrated:** Integrated or concurrent development method consists of cross-functional teams with members from all of the functional areas working closely together, sharing details of their portion of the design as it progresses, and developing all aspects of the system simultaneously. The result is overlapping and managing the overall life cycle. Concurrent Engineering (CE) is defined as the systematic and integrated approach to systems life cycle design. CE is also known as the *Design for Life Cycle* model. Concurrent engineering is the implementation of parallel designs by cross-functional teams including suppliers. Without empowered team members and the free flow of communications, this method will not function. Figure 4 illustrates an overview of the integrated SE infrastructure. Table1 lists partial comparison among these models.

A model that is typically used to define the critical elements of SE process is the Systems Engineering Capability Maturity Model (SE-CMM)[7]. These elements consist of activities which define '*what*' has to be done. Table 2 lists these tasks. Tasks for design and development are listed in engineering group, project elements provide the management infrastructure and finally the organization elements provide the business infrastructure to support systems engineering effort.

[6] B. W. Boehm, *Spiral Model of Software Development*, Tutorial Software Project Management edited by R. H. Thayer and M. Dorfman, IEEE Press, 1988.

[7] Systems Engineering Capability Maturity Model[SM], Version 1.1, Software Engineering Institute, Carnegie Mellon University, 1995.

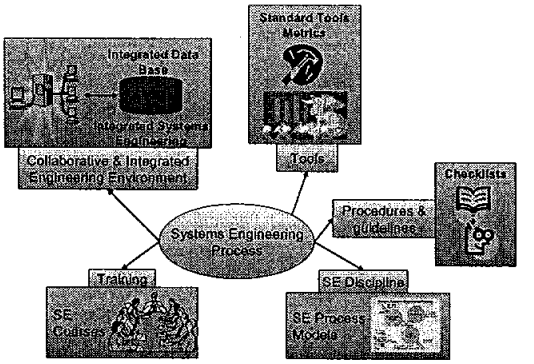


Figure 4. Integrated Systems Engineering Infrastructure

SE Paradigm	Easy to Implement	Supports Prototyping	Early Customer Involvements	Ease of Engineering Changes	Meets Customer's Schedule	Requirement Driven
Build-Test-Fix	++	+	-- --	--	+	--
Staircase	+	--	--	++	0	-+
Waterfall	-	--	+	+	--	+
Early Prototype	--	+	+	+	--	+
Spiral	-- --	++	+	+	+	+
Rapid Development	-- --	++	-- --	+	--	+
Integrated	-- --	++	++	-- --	+	++

Table 1. SEP sample comparison, strong (++), average (+), none (0), low (-) very low(--)

Engineering	Project	Organization
✓ Understand Customer Needs	✓ Ensure Quality	✓ Coordinate with Suppliers
✓ Derive and Allocate Requirements	✓ Manage & Control Configurations	✓ Define SE Process
✓ Analyze Alternative Solutions	✓ Manage Risk	✓ Manage System Evolution
✓ Evolve System Architecture	✓ Monitor & Control Technical Effort	✓ Manage Systems Engineering Support Environment
✓ Integrate System	✓ Plan Technical Effort	✓ Continuous Improvement
✓ Integrate Disciplines	✓ Integrate Technical Efforts	
✓ Testing and Acceptance		

Table 2. SE-CMM Model

SYSTEMS ENGINEERING SAMPLE PROCESS MODLES

"Vee" Process is widely used by many industries (Figure 5).

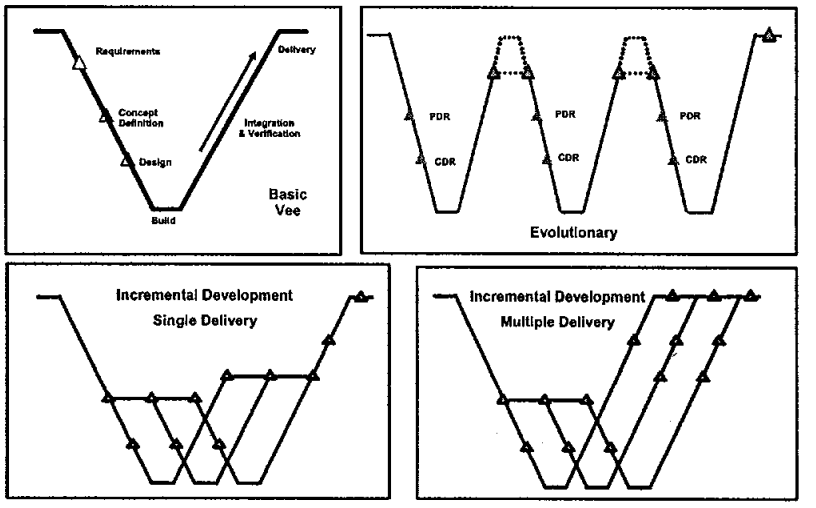


Figure 5. "Vee" Process as applied to NASA Projects[8]

This concept is based on the iterative and parallel processes on the left hand side that will manage the verification functions on the right hand side. Verification is completed in a serial fashion, resulting in minimal rework. This method is cost effective and improves the success of the project. It also provides the necessary infrastructure for alternative design analysis and selection. A system that fits the requirements with best performance, voiced by the stakeholders. It is believed that by using this approach the probability of design of a reliable and satiable system is high[9]. Within the "Vee" process, the ***3-Bubble Method*** (Figure 6) assures that the design performance and feasibility (schedule) are continuously compared with the requirements. This allows for analysis of alternative designs against the verified requirements.

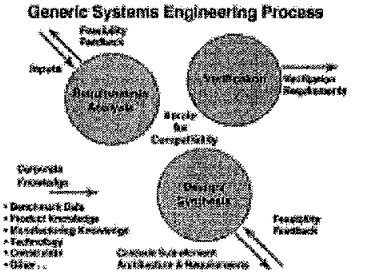


Figure 6. 3-bubbles method[10]

[8] K Forsberg, H. Mooz and H. Cotterman, *Visualizing Project Management: A Model for Business and Technical Success*, 2nd Edition, Wiley and Sons, 2000.
[9] Ford Design Institute, *Systems Engineering Fundamentals Course*,, Ford Motor Company, 2000.
[10] Ibid.

The International Council in Systems Engineering, ***INCOSE***, defines the Systems Engineering Process as *"... an iterative process of technical management, acquisition and supply, system design, product realization, and technical evaluation at each level of the system, beginning at the top (the system level) and propagating those processes through a series of steps which eventually lead to a preferred system solution. At each successive level there are supporting, lower-level design iterations which are necessary to gain confidence for the decisions taken. During each iteration, many concept alternatives are postulated, analyzed, and evaluated in trade-off studies. There are many cross-coupling factors, where decisions on one subsystem effect other subsystems*[11]." INCOSE model is illustrated in Figure 7.

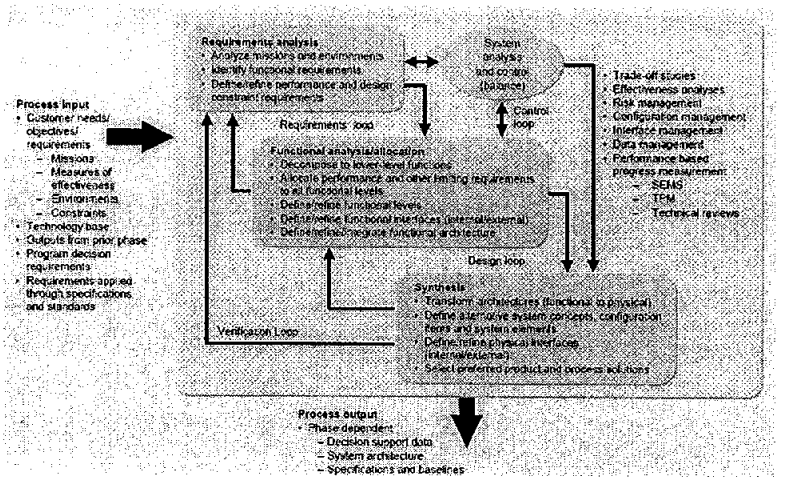


Figure 7. INCOSE Systems Engineering Model[12]

The Department of Defense (DoD) defines the systems engineering process as the transformation of the operational needs and requirements into an integrated system design solution through concurrent consideration of all life-cycle needs. This process will ensure the compatibility and integration of all physical interfaces and system definition and design reflect the requirements for all system elements (hardware, software, data, people, etc.). Finally, the SE process will identify and manage technical risks associated with the system development. Figure 8 illustrates the DoD SE process. ***Cost as an Independent Variable Concept (CAIV)*** is defined in Section 3.3.4 of DoD 5000.2-R, as:"*. . . a process that helps arrive at cost objectives (including life-cycle costs) and helps the requirements community set performance objectives. The CAIV process shall be used to develop an acquisition strategy for acquiring and operating affordable DoD systems by setting aggressive, achievable cost objectives and managing achievement of these objectives. Cost objectives shall also be set to balance mission needs with projected out-year resources, taking into account anticipated process improvements in both DoD and defense industries.*" Cost in this process is a constraint. Identification and use of viable range of alternatives with knowledge of real and potential impacts, is essential for making

[11] International Council on Systems Engineering (INCOSE) SE Handbook Working Group, 2000.
[12] Ibid.

the right decisions to meet stakeholders VOC while minimizing the Total Ownership Cost (TOC).

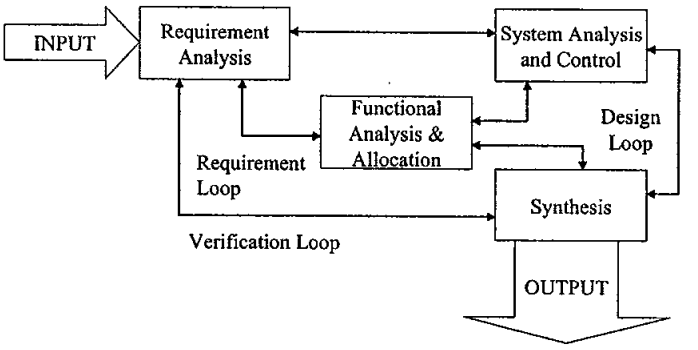


Figure 8. DoD SE Process Cycle[13]

CAIV principle, as proposed by USAF, is further decomposed into five pillars[14]. Figure 9 illustrates these five pillars.

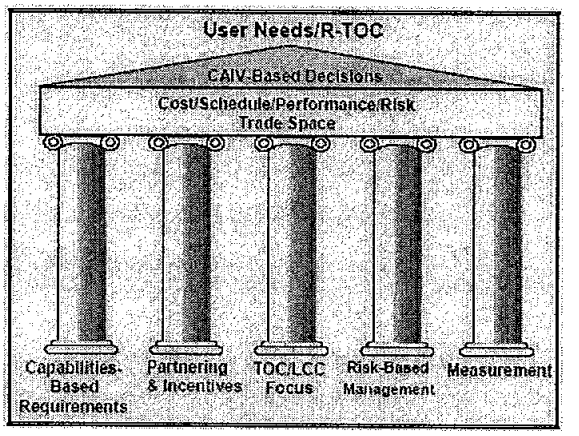


Figure 9. CAIV Process Pillars[15]

CAIV is based on capability-based requirements. In this case, user must first define "what" the system needs to do and how subsystems are allocated. Cost Performance Integrated Product Team (CPIPT) is a major component of the CAIV process. This team performs cost-performance-schedule tradeoffs leading to CAIV-based cost, performance, and schedule objectives. The CPIPT and stakeholder work closely together to resolve issues and decide on the final range and objective values for schedule cost and

[13] Systems Engineering Fundamentals, Defense Acquisition University Press, 2001.
[14] Col. M. A. Kaye (USAF), Lt. Col. M. S. Sobota (USAF), D. R. Graham, A. L. Gotwald, "*Cost as an Independent Variable (CAIV) Principles and Implementation*," American Institute of Aeronautics and Astronautics, 1999.
[15] Ibid.

performance. The total life cycle cost is closely monitored. TOC in CAIV is LCC. Risk management is an important part of the CAIV process. Program management and means for measure of progress assures the success of CAIV process. **Simulation-Based Acquisition Concept (SBA)** is the integrated and collaborative approach to systems design and through computer–based modeling and simulation. SBA concept is based on the DoD acquisition reform initiative. The new process as defined by DoD 5000.1, 23 Oct 2000 consists of "*an acquisition process in which DoD and Industry are enabled by robust, collaborative use of simulation technology that is integrated across acquisition phases and programs.*" SBA concept is based on collaborative engineering concept and environment[16]. Industrial partners, academia experts and government agencies will closely collaborate using COTS, technologies, developed methodologies and resources. This will reduce the development time and cost associated increased performance and functionality. Five principal architectural concepts used for SBA implementation are:

1. Collaborative Environment
2. Collaborative Environments Reference Systems Architecture
3. Distributed Product Descriptions
4. DoD/Industry Resource Repository
5. Data Exchange Format

SBA is not the replacement for systems engineering process. It is a distributed and integrated approach to design using the systems engineering principles. It is a modeling and simulation (M&S) technique used to support managers during their decision making process. It must maintain the integrity and security of all shared data including responsibility and accountability at all levels of proprietary and security.

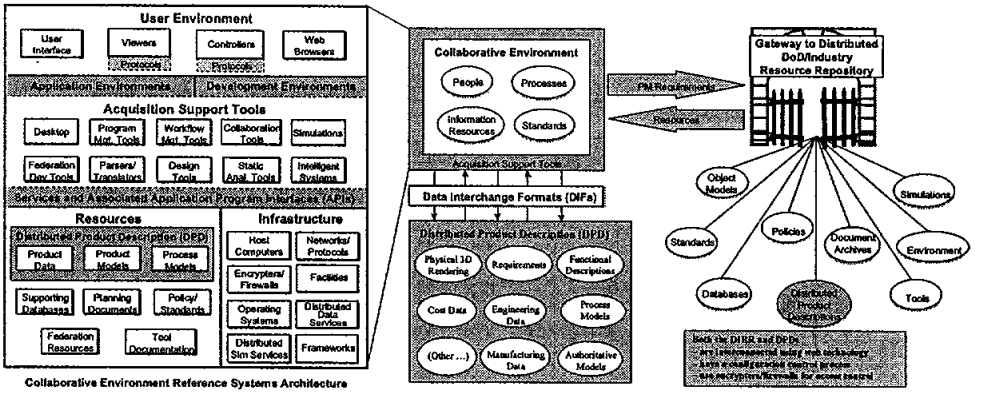


Figure 10. SBA infrastructure[17]

[16] Simulation Based Acquisition; A New Approach, Defense Systems Management College Press, 1998.
[17] J. E. Coolahan, F. T. Case, Lt. Col. R. J. Hartnett, Jr., (USAF), "*The Joint Strike Fighter (JSF): Strike Warfare Collaborative Environment (SWCE),*" Proceedings of Fall Simulation Interoperability Workshop, 2000.

THE JOINT STRIKE FIGHTER PROGRAM

JSF program goal is to develop an affordable family of multi-role attack aircraft with high commonality. JSF has been described as a supersonic, single-engine, single-seat airplane with F-16/F/A-18 class performance and stealthy. The three JSF variants are a conventional takeoff and landing (CTOL); a carrier-based attack (CV) variant and a STOVL replacement. The Air Force may also buy some STOVL variants to replace its A-10s. (Figure 11).

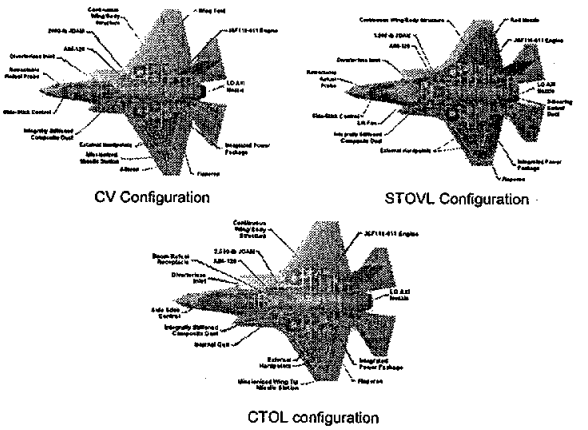


Figure 11. Lockheed Martin JSF Family for Next Generation of Military Fighters

Modeling and simulation (M&S) concept has been significantly used for the JSF program through Program Definition and Risk Reduction (PDRR) to integrate warfighters and maintainers into the development process early in the design life cycle. The objective is to reduce requirements changes and development costs. JSF program charter focused on six key initiatives:

- Establishing an integrated team of users and developers
- Facilitating evolution of fully developed and validated operational requirements
- Evolving requirements over time
- Reducing development, procurement, and support costs
- Performing tradeoff analyses of critical user defined performance parameters
- Conducting unprecedented levels of simulation

Within the four major JSF requirement (lethality, survivability, supportability, and affordability) 34 attributes are identified. These are, acquire the target, generate sorties, situational awareness, countermeasures, ID the target, weapon/sensor integration, accurate navigation, mission level intelligence, pass/receive timely information, assess battle damage, route planning, emissions control, maintainability, carrier suitability, inter operability, basing flexibility, all weather/night capability, multi-role capability, weapon carriage versatility, mission flexibility, accuracy, radar cross section, range, logistic

footprint, payload, infrared signature, speed, maneuverability, electro-optical signature, redundancy, hardening, acoustic signature, reliability, and ferry range. The JSF attributes are the combination of functions, operational flexibility, operational constraints, and parameters. Simulation modeling and virtual prototyping has allowed the Joint Strike Fighter (JSF) concept demonstration for assembly to be accomplished with fifty percent reduction in staffing and time compared to actual planned levels[18]. *"For JSF developments, simulations have improved the mechanical tolerances where the originally projected shim stock weight of 40 lbs per aircraft, as in the F-16, was reduced to less than 1 pound.*[19] " Cost was considered to be a major criteria during JSF requirement analysis phase. It was used as criteria for trade-studies. JSF is developed based on the simulation and acquisition program concept and collaborative engineering. Simulation was extensively used during the requirements development process, and will be used throughout the program. Virtual prototyping and collaborative engineering was used to integrate all stakeholders into the systems engineering process. Analysis provides the incremental approach for complete system definition, design and integration.

PART II: ALS Functional Analysis and Decompositions within AIM

The life support system provides for crew the necessary resources for activities such as food, water and waste management. Figures 12 and13 illustrate the level of interface between the crew, life support sub-systems and the four main sub-systems interaction of ALS.

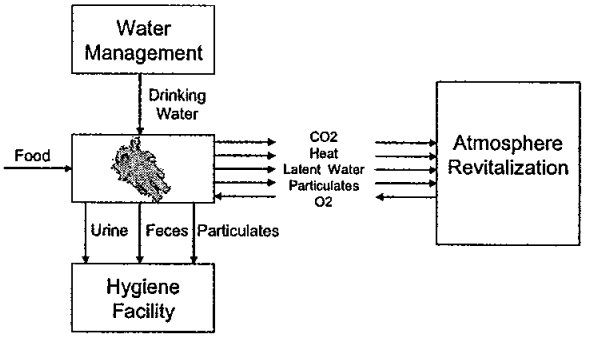


Figure 12. Crew and the Life Support System Interface[20]

A three phase methodology was proposed to further identify and decompose the life support functions. The steps within the methodology are:

1. Understand the System
    - Collect Information about the problem

[18] Boeing.com/defense-space/military/jsf/lean_mfg.html
[19] Building A Business Case for M&S, Acquisition Review Quarterly—Fall 2000
[20] D. Henninger, *"Lunar Base Life Support and Crew Health,"* Lunar Base Handbook , McGrawHill, 1999.

- Organize the Requirements
- Develop the Theme

2. Define Function
   - Select requirement for functional analysis
   - Define Functions
   - Define function criteria
3. Decompose and Systematize Functions
   - Identify main functions
   - Relate functions
   - Check functions series
   - Establish criteria

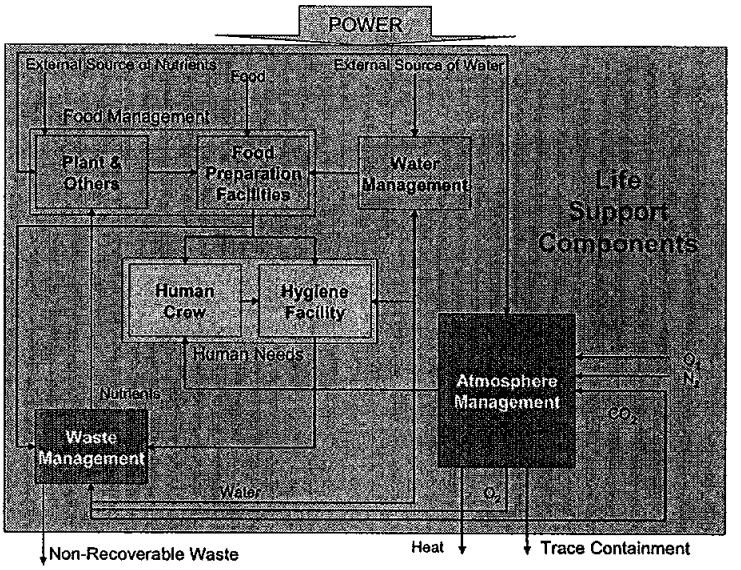


Figure 13. Life Support Sub-Systems[21]

To define the scope of the life support system, the affinity diagram was used to collected data from the team members. Affinity analysis is a process used to gather large amounts of data based on opinions, concepts and issues and then organizes them into sets of groups using their natural relationship. The Affinity process uses brainstorming sessions to generate and collect group ideas. Steps for developing the affinity diagram are:

1) Identify the problem
2) Generate ideas
3) Display ideas
4) Sort ideas into groups
5) Create header cards

[21] S. Doll, *"Life Support Functions and Technology Analysis for Future Missions,"* Proceeding of 20th Intersociety Conference on Environmental Systems, SAE Technical Paper901216, 1990.

6) Draw clustered groups and finished diagram

The issue of ***"How to Support Crew Life Beyond LEO?"*** was used to collect data from the team. The team included engineers, systems engineers and managers from both NASA and Lockheed. Figure 14 illustrate the results in the affinity diagram format.

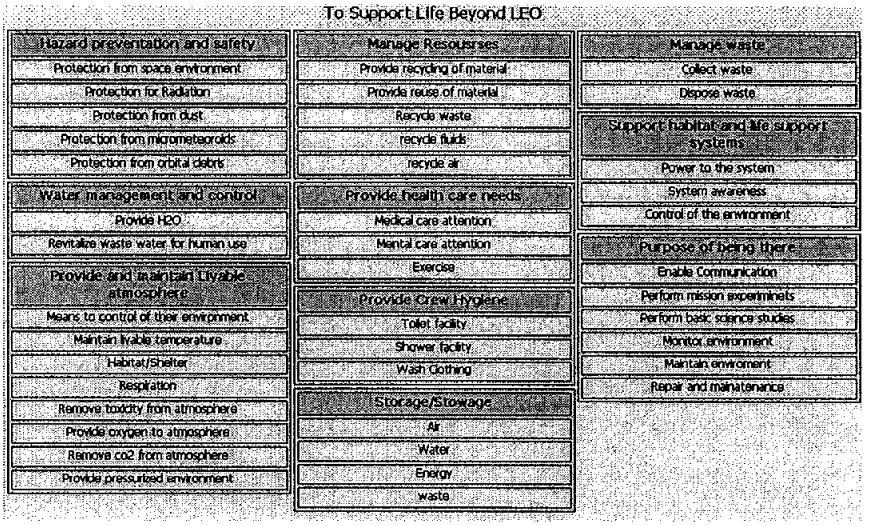


Figure 14. Result of the Affinity Analysis

Defining the functions of a given system involves asking the question of "**What is it action?**" Functions are typically expressed using the verbal model that combines a verb and a noun. Typically function is characterized by the degree to which it performance is required and fulfilled under certain condition. These are called criteria for functions; (e.g. mile/gallon). The criteria for functions are determined using the so-called 5W1H questions; what, who, when, where, why and how much. After further analysis of the affinity data and its interpretation to the requirements, the following partial functional requirements for ALS were identified:

- Maintain a safe, habitable and operational environment.
- Provide resources for atmosphere, maintenance, crew consumption, and crew hygiene.
- Manage wastes for resource recovery.

Axiomatic design was used for further decomposition, systemization and mapping of the technologies. Axiomatic design provides the scientific basis and structure design approach to design identification, decomposition and mapping[22]. AD is a structured approach that associates the needs, requirements and solutions for system design problem. In Axiomatic Design approach, the customer wants is processes in such a way

[22] Num Suh,, "*The Principles of Design*," New York: Oxford University Press, 1990.

that lead to the definition of the functional requirements (FRs) which then identifies the design parameters (DPs). The FRs identify *what* needs to be done and DPs identify *how* each FR is implemented. Each DP is decomposed into lower level of FRs through zigzagging for further identifications of DPs and FRs[23]. This decomposition process continues until the DPs explain the design; design is complete. Axiomatic design was used functional requirements analysis of ALS. Sample results for decomposition and traceability chart is listed in Figures 15 and 16.

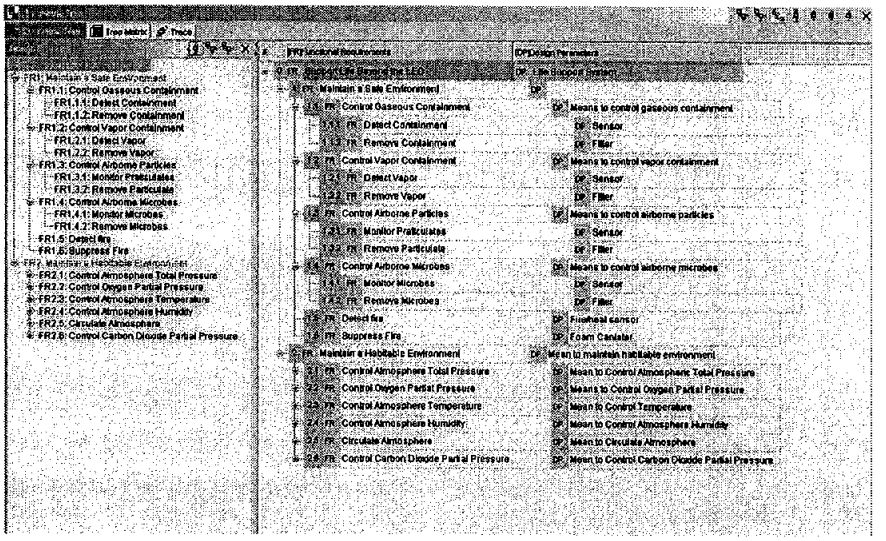


Figure 15. Partial listings of the ALS Decomposed functions

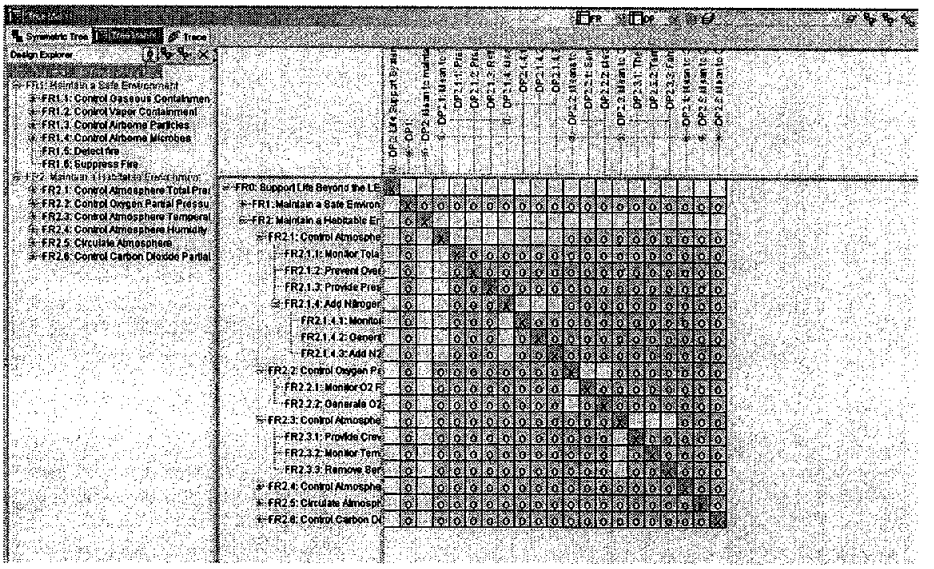


Figure 16. Partial traceability matrix

[23] D. S. Cochran and A.K. Chu, "*Measuring Manufacturing System Design Effectiveness Based on the Manufacturing System Design Decomposition*", 3rd World Congress on Intelligent Manufacturing Processes & Systems Cambridge, MA – June 28-30, 2000

CONCLUSION AND FUTURE PLANS

The mission of AIM is also to consider the HST technologies as an integral part of Advanced Life Support system development. The process focusing on the embodiment of technologies as part of the system's design is known as *technology-push*. This approach requires a new engineering paradigm that considers technology feasibility analysis and its integration into the customer-pull traditional SE process in order to validate the performance(s) of the technology within the overall system using parallel structure. Since no structured approach is available for the technology push method of design, potential risks of missing the mission needs and requirements are high. Despite these risks ***a successful process has significant benefits***. A new engineering paradigm is proposed[24] to consider perform this task. It will perform technology feasibility analysis and integrate it into the ALS SE development process in order to validate the performance(s) of the technology within the overall system using parallel structure. A step-by-step process is used to guide the systems engineer through testing and integration of the ALS technologies and then identifying corresponding HST design parameters. The three stages proposed for technology capability and feasibility analysis are ***1) Technology Evaluations, 2) Technology Opportunity Identification, and 3) Technology Mapping*** as shown in Figure 17.

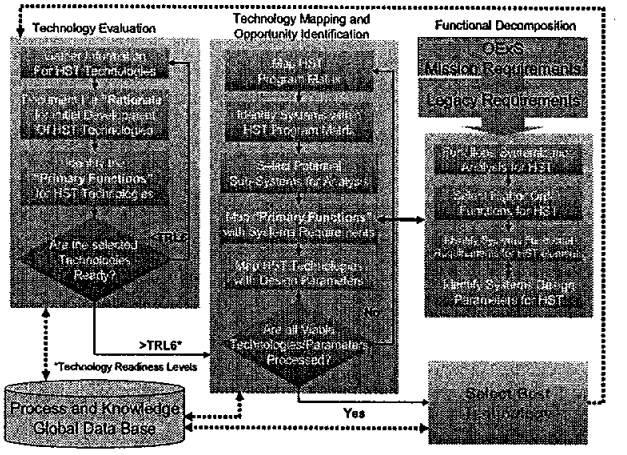


Figure 17. Integrated TP and SE Process for ALS and HST Mapping Methodology

Further research is necessary for the development and implementation of the proposed method. Potential sources of funding are being considered for the continuation.

[24] A. Kamrani, ALS Sub-Systems Design Integration & Testing within AIM, NFFP Directorate Proposal, (not funded), 2004.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 2005	NASA Contractor Report

4. TITLE AND SUBTITLE
NASA Summer Faculty Fellowship Program 2004, Volumes 1 & 2

5. FUNDING NUMBERS

6. AUTHOR(S)
Edited by William A. Hyman, Donn G. Sicorez , and Dawn M. Leveritt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lyndon B. Johnson Space Center
Houston, Texas 77058

8. PERFORMING ORGANIZATION REPORT NUMBERS
S-961

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
CR-2005-213690

11. SUPPLEMENTARY NOTES
2 Volumes. Volume 1 186 pages, voilue 2 162 pages.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Available from the NASA Center for AeroSpace Information (CASI)
7121 Standard
Hanover, MD 21076-1320 Category: 99

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

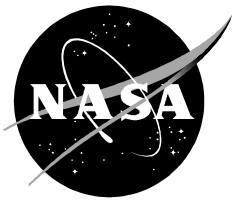
The 2004 Johnson Space Center (JSC) National Aeronautics and Space Administration Faculty Fellowship Program (NFFP) was conducted by Texas A&M University and JSC. The program was funded by the Office of Education, NASA Headquarters, Washington, D.C. and by JSC. Each faculty Fellow spent at least 10 weeks at JSC (or the White Sands Test Facility) engaged in a research project in collaboration with a NASA/JSC colleague.

This document is a compilation of the final reports on the research projects done by the Fellows during the summer of 2004. Volume 1 contains reports 1 through 12 and Volume 2 contains reports 13 through 22.

14. SUBJECT TERMS	15. NUMBER OF PAGES	16. PRICE CODE
Human performance, abilities; life support systems; systems engineering; Medical science, cardiology; aerospace medicine; communication; exploration	338	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NASA/CR-2005-213690



NASA Summer Faculty Fellowship Program 2004

RESEARCH REPORTS

Volume 2

William A. Hyman, Donn g. Sickorez**, and Dawn M. Leveritt**,*
Editors

* *Texas A&M UniversityDonn Sickorez, Ph.D.*
NASA, Johnson Space Center
Houston, Texas

** *Lyndon B. Johnson Space Center*
Houston, Texas

August 2005

The NASA STI Program Office . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA's counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or cosponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and mission, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at ***http://www.sti.nasa.gov***

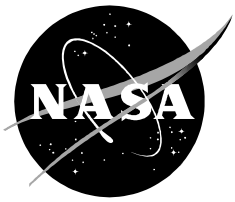
- E-mail your question via the Internet to help@sti.nasa.gov

- Fax your question to the NASA Access Help Desk at (301) 621-0134

- Telephone the NASA Access Help Desk at (301) 621-0390

- Write to:
  NASA Access Help Desk
  NASA Center for AeroSpace Information
  7121 Standard
  Hanover, MD 21076-1320

NASA/CR-2005-213690



NASA Summer Faculty Fellowship Program 2004

RESEARCH REPORTS

Volume 2

William A. Hyman, Donn G. Sickorez**, and Dawn M. Leveritt**,*

Editors

* *Texas A&M University*
NASA, Johnson Space Center
Houston, Texas

** *Lyndon B. Johnson Space Center*
Houston, Texas

Grants NGT 9-1526 and NNJ04JF93A

National Aeronautics and
Space Administration

Johnson Space Center
Houston, Texas 77058-3696

August 2005

Available from:

NASA Center for AeroSpace Information
7121 Standard
Hanover, MD 21076-1320

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161

This report is also available in electronic form at http://techreports.larc.nasa.gov/cgi-bin/TRS

Preface

The 2004 Johnson Space Center (JSC) National Aeronautics and Space Administration Faculty Fellowship Program (NFFP) was conducted by Texas A&M University and JSC. The program was funded by the Office of Education, NASA Headquarters, Washington, D.C. and by JSC. Each faulty Fellow spent at least 10 weeks at JSC (or the White Sands Test Facility) engaged in a research project in collaboration with a NASA/JSC colleague.

This document is a compilation of the final reports on the research projects done by the Fellows during the summer of 2004. Volume 1 contains reports 1 through 12 and Volume 2 contains reports 13 through 22.

TABLE OF CONTENTS

Volume 2

Physics-based Simulation of Human Posture Using 3D Whole Body Scanning Technology for Astronaut Space Suit Evaluation

Final Report
NASA Faculty Fellowship Program – 2004
Johnson Space Center

Prepared By:	Kyu-Jung Kim, Ph.D.
Academic Rank:	Assistant Professor
University & Department:	Mechanical Engineering Department Univ. of Wisconsin-Milwaukee Milwaukee, Wisconsin 53201
NASA/JSC	
Directorate:	Space and Life Sciences
Division:	Habitability and Environmental Factors
Branch:	Habitability and Human Factors
JSC Colleague:	Sudhakar Rajulu, Ph.D.
Date Submitted:	August 6, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

Over the past few years high precision three-dimensional (3D) full body laser scanners have been developed to be used as a powerful anthropometry tool for quantification of the morphology of the human body. The full body scanner can quickly extract body characteristics in non-contact fashion. It is required for the Anthropometry and Biomechanics Facility (ABF) to have capabilities for kinematics simulation of a digital human at various postures whereas the laser scanner only allows capturing a single static posture at each time.

During this summer fellowship period a theoretical study has been conducted to estimate an arbitrary posture with a series of example postures through finite element (FE) approximation and found that four-point isoparametric FE approximation would result in reasonable maximum position errors less than 5%. Subsequent pilot scan experiments demonstrated that a bead marker with a nominal size of 6 mm could be used as a marker for digitizing 3-D coordinates of anatomical landmarks for further kinematic analysis. Two sessions of human subject testing were conducted for reconstruction of an arbitrary postures from a set of example postures for each joint motion for the forearm/hand complex and the whole upper extremity.

INTRODUCTION

Three-dimensional whole body laser scanning can extract body surface and volume characteristics in seconds to construct a 'digital human' and is more repeatable than traditional anthropometry. 3D scan data from a standard pose of an astronaut can be processed to create a 3D whole body CAD model for further computer simulation. However, this digital human scanned at a single static posture may not be useful for assessment of suit accommodation and further evaluation the biomechanical requirements of a particular suited/unsuited pose or task. Thus, it requires for the ABF to have capabilities for kinematics simulation of a digital human at various postures without losing the integrity of physical parameters (such as segment lengths, widths, depths, soft tissue deformation, etc).

Current ergonomic tools only allow kinematics simulations using a standard model with anthropometric scaling, thereby lacking simulations of an individual with diverse anthropometric variations and physical features. Furthermore, the ABF has been using a stick figure model (ERGO model) written in MATLAB. The ERGO model is based on traditional anthropometric measurements and has the ability to be repositioned but lacks surface definitions (Figure 1).

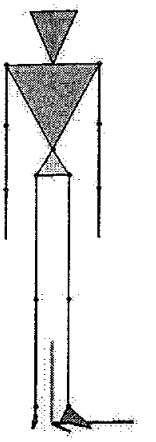


Figure 1. ERGO model in the anatomic position

WORK ACCOMPLISHED

During this summer a theoretical study has been conducted to estimate an arbitrary posture with a series of example postures through finite element (FE) approximation. Subsequent pilot experiments were conducted using a 3D whole body scanner (Human Solutions, Inc) to find optimum makers for digitizing 3D coordinates of anatomical landmarks for kinematic analysis. Two human subject testing was conducted for reconstruction of an arbitrary postures from a set of example postures for each joint motion.

FE Approximation of Arbitrary Postures

Human joint motion can be idealized as a circular arc motion with one degree of freedom (DOF) for the elbow joint or two DOF's for the wrist joint and three DOF's for the shoulder joint. First, a single DOF joint with unit distance and 90 degree range of motion (ROM) was used to approximate the arbitrary positions at an interval of 3 degrees along the ROM with three example angular positions of 0, 45, and 90 degrees (Figure 2 left). The 3-point FE approximation was conducted using an isoparametric formulation, resulting in quadratic Legendre interpolation (Logan 2002). The maximum position error was estimated to be 3%. However, the error would be substantially increased to 22% when the same 3-point FE approximation using three example angular positions of 0, 90, and 180 degrees was used to interpolate for a joint with 180 degree ROM (Figure 2 right). Higher order FE approximation could substantially reduce the maximum position error. With four example angular positions of 0, 30, 60, and 90 degrees it became 0.3% and 4.5% for 90 and 180 degree ROM's, respectively (Figure 3).

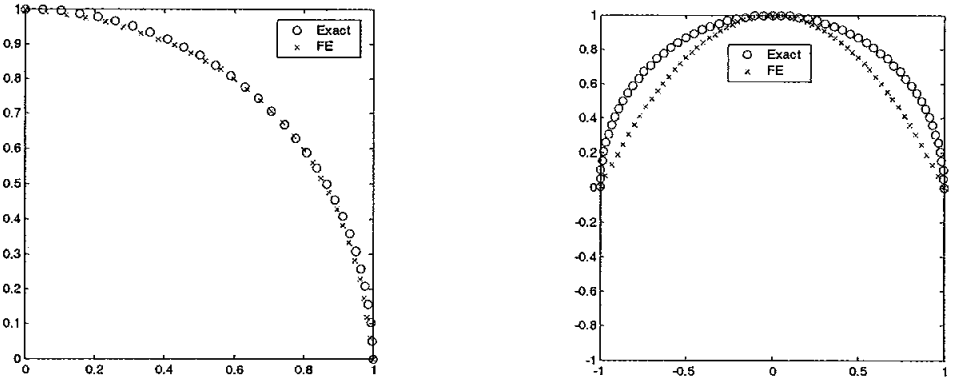


Figure 2. 3-point FE approximation of a single DOF joint with 90 and 180 degree ROM.

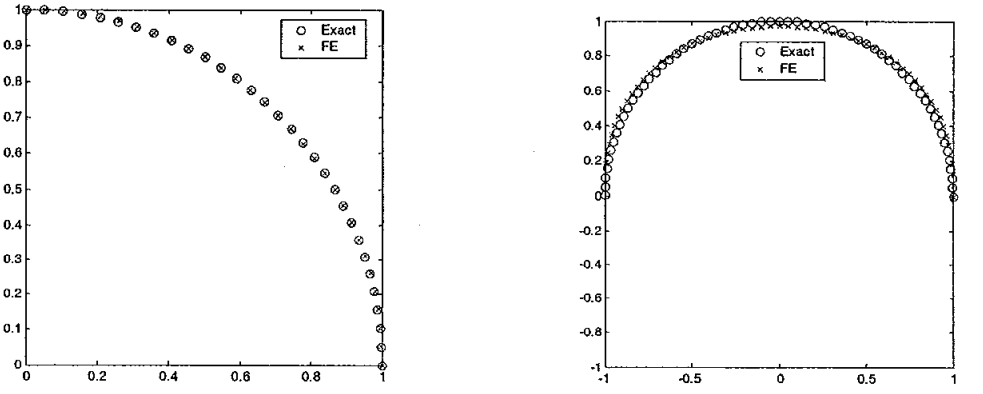


Figure 3. 4-point FE approximation of a single DOF joint with 90 and 180 degree ROM.

Similar theoretical study was conducted for a two DOF joint. The joint had unit distance and 90 and 180 degree ROM's. Both 3- and 4-point FE approximations were used for each DOF so that a total of 9 and 16 example positions were used to interpolate arbitrary positions in the joint space. The 3-point FE approximation resulted in maximum position error of 4.5% and 22.0% for 90 and 180 degree ROM, respectively (Figure 4). On the other hand, the 4-point FE approximation resulted in maximum position error of 0.5% and 4.6% for 90 and 180 degree ROM, respectively (Figure 5).

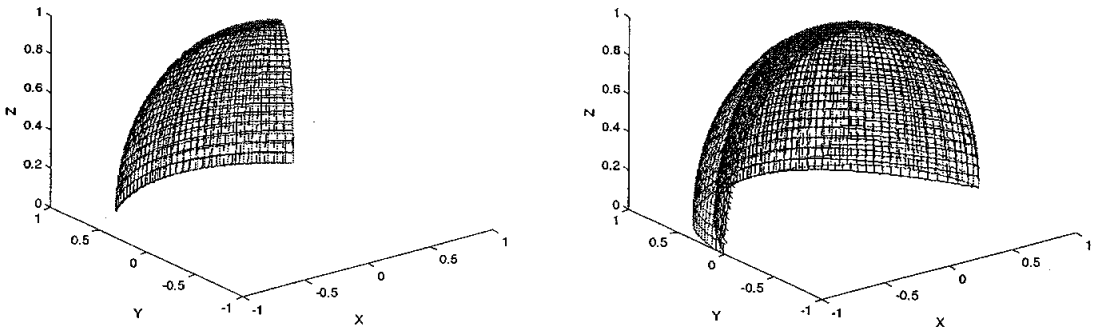


Figure 4. 3-point FE approximation of a two DOF joint with 90 and 180 degree ROM.

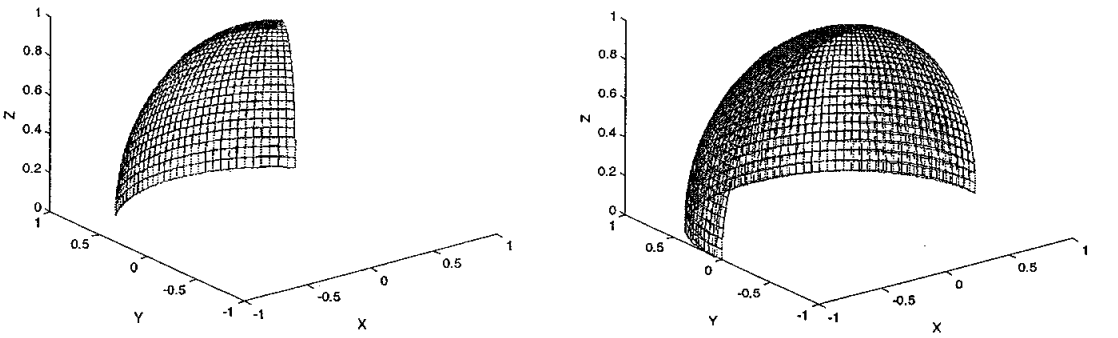


Figure 5. 4-point FE approximation of a two DOF joint with 90 and 180 degree ROM.

From these studies it was concluded that at least four example postures are needed to interpolate an arbitrary posture using FE approximation for each joint DOF. Thus subsequent experimental studies were conducted based on these results.

Scanning of the Forearm/Hand Complex and the Upper Extremity at Example Postures

After a few trial scans with various markers it was concluded that a plastic bead with a diameter of 6 mm could be captured and digitized in ScanWorX V2.8 software (Human Solutions, Inc). To test the results of the theoretical study, a 3D laser scanning study was conducted with a human subject for the motion of the forearm/hand complex (FHC) since

it has both one and two DOF joints. A total of 27 bead markers were attached on each anatomical landmark for 3D kinematic analysis of joint motion (Figure 6).

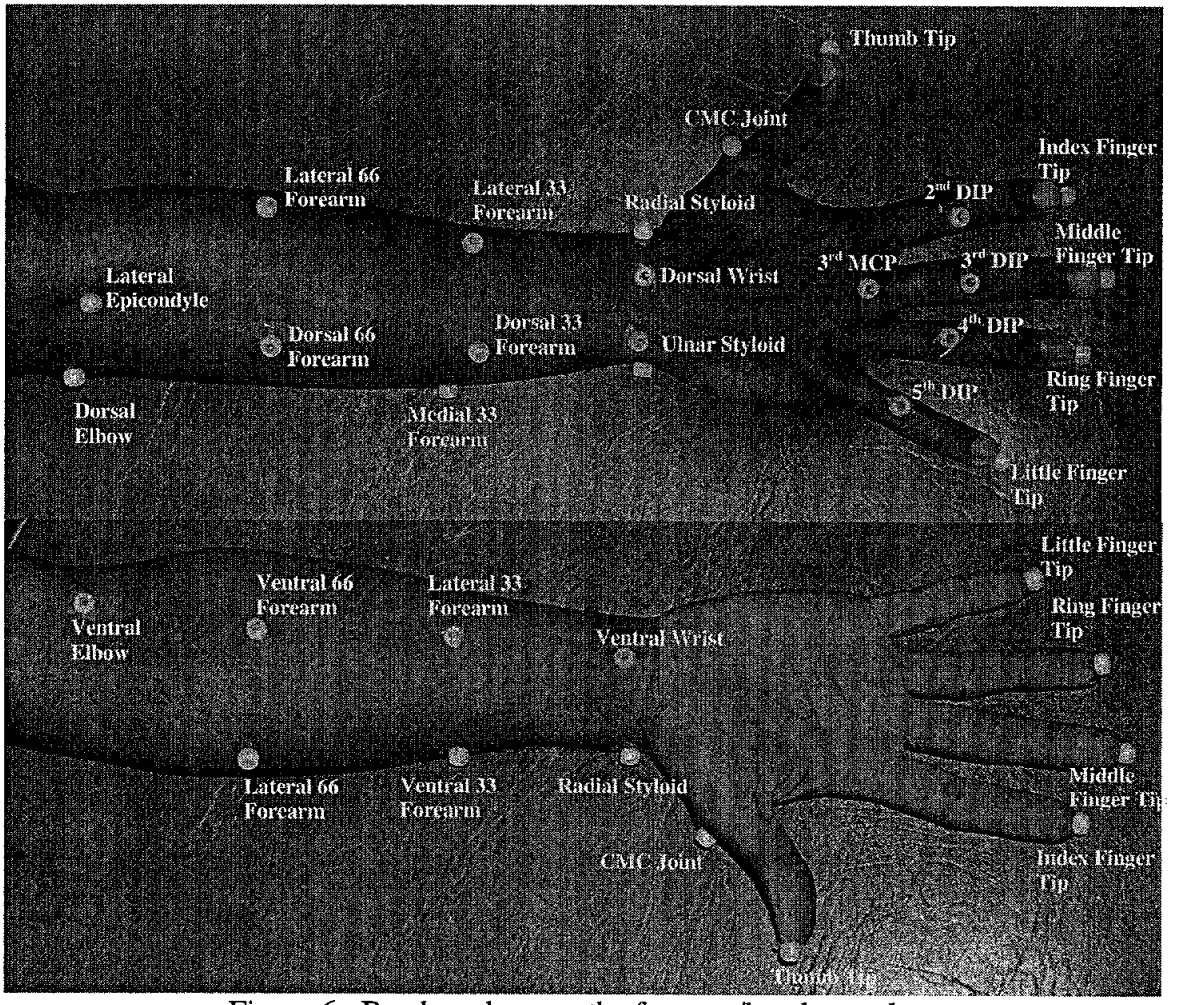


Figure 6. Bead markers on the forearm/hand complex.

The FHC with bead markers were scanned and processed using ScanX V2.8 software. The subject was in supine position with the FHC in full pronation. Then, for each joint DOF a series of four scans were made. The subject voluntarily moved each joint at zero, one-, two-thirds, and full ROM. Wrist flexion/extension, radial/ulnar deviation, forearm pronation/supination, and elbow flexion/extension pairs were scanned so that a total of 16 scans plus one anatomical neutral scan were made as example scans for FE approximation. To minimize horizontal gaps in the scan the upper extremity was maintained in vertical position as much as possible. In the viewer window with "Standard Projection", the region of interest was selected from the whole-body scan image by removing the non-interest areas with Shift-Control and polygonal selection. Then, each of the bead markers were manually located using "persistent" markers in the Measure3D window (Figure 7).

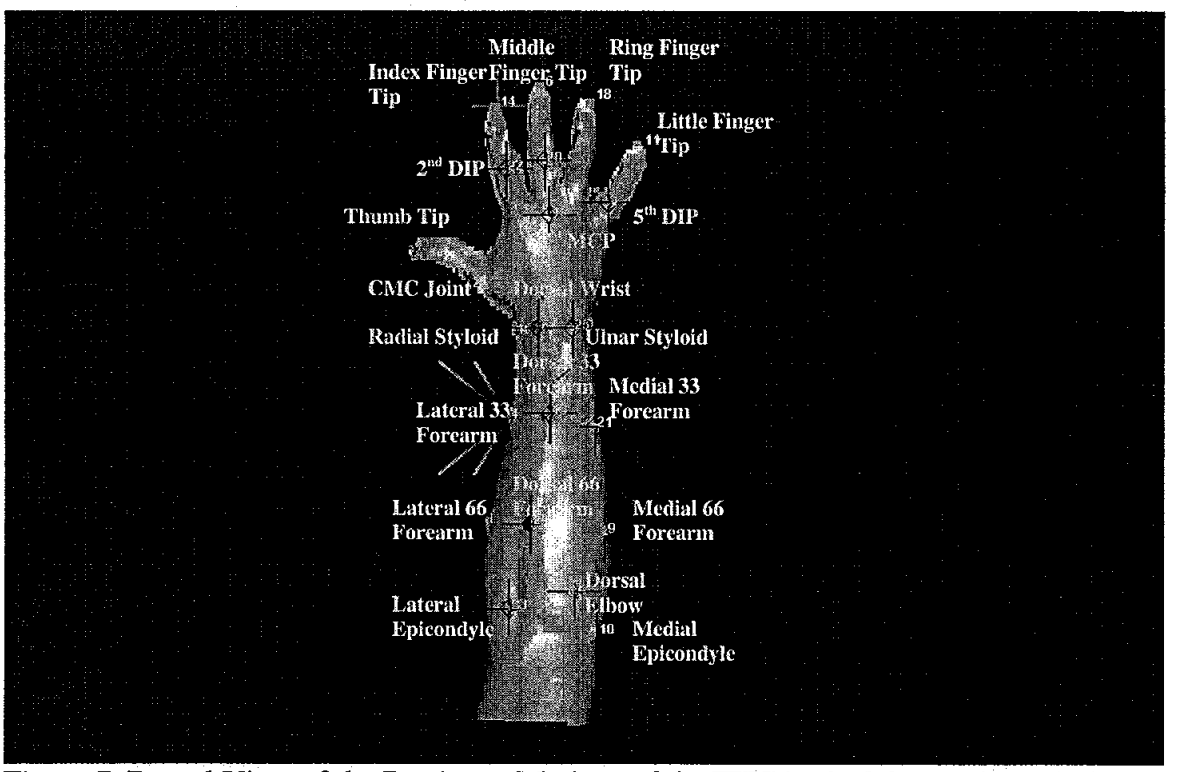


Figure 7. Dorsal View of the Persistent Markers of the FHC in the Measure3D Window.

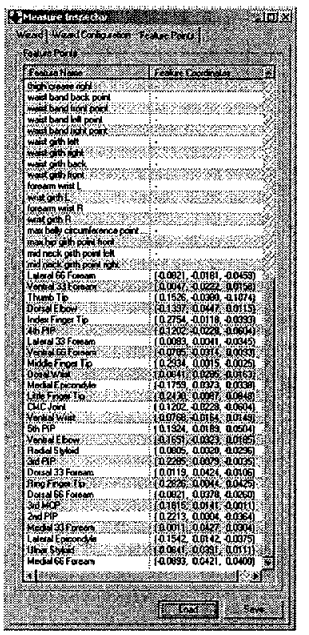


Figure 8. 3D Coordinates of the Persistent Markers from the Measure Inspector.

The 3D coordinates of the persistent markers were retrieved from the list of feature points by selecting Measure>Inspect Measures in the Measure3D menu (Figure 8). The feature files were saved into a text file and imported into MATLAB.

Similar procedures were taken to scan the upper extremity motion at various example postures. A total of 51 bead markers are used (Figure 9).

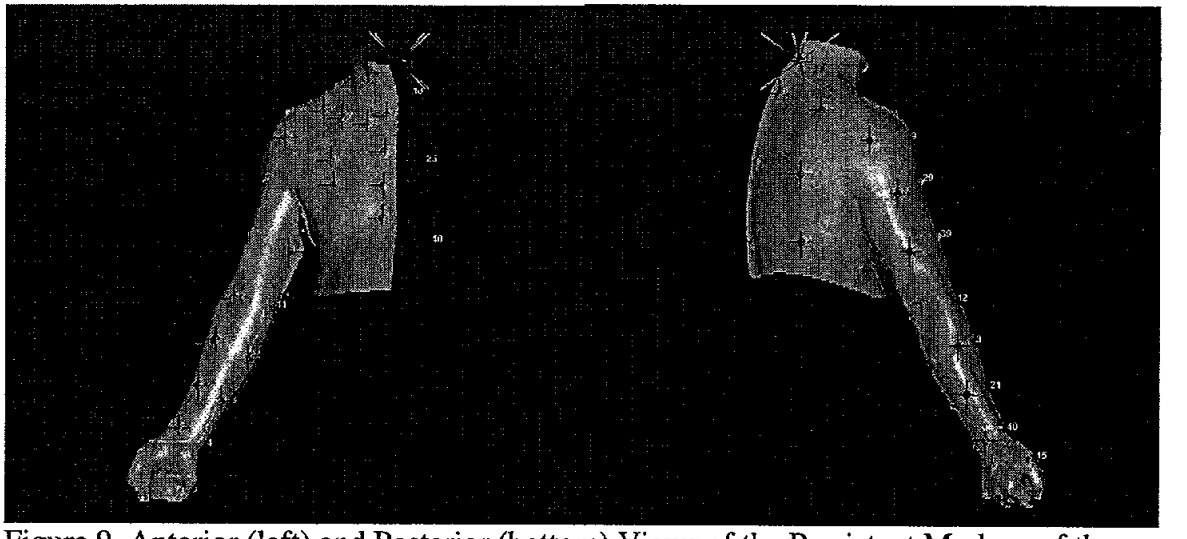


Figure 9. Anterior (left) and Posterior (bottom) Views of the Persistent Markers of the upper extremity in the Measure3D Window.

FUTURE WORKS

- Conducting 3D kinematic analysis using the persistent marker data to estimate the joint angles for each different example postures
- Filling holes and gaps in the scans for surface parameterization and correspondence
- Building posable 3D forearm/hand complex and upper extremity models
- Validating the models with the scans at arbitrary postures to estimate the maximum position errors
- Documenting the results for submitting conference abstracts and/or journal papers

REFERENCES

Logan DL (2002) "A First Course in Finite Element Methods", 3rd Ed, Brooks/Cole, Pacific Grove, CA.

Artificial Neural Network Test Support Development for the Space Shuttle PRCS Thrusters

Final Report
NASA/Faculty Fellowship Program – 2004

Johnson Space Center

Prepared by:	Mark E. Lehr, Ph.D.
Academic Rank:	Assistant Professor
University & Department	Riverside College Computer Systems Riverside, Ca 92506
NASA/JSC	
Directorate:	White Sands Test Facility
Division:	Laboratories
Branch:	Science and Engineering
JSC Colleague:	Regor Saulsberry
Date Submitted:	September 15, 2004
Contract Number:	NAG 9 -1526 and NNJ04JF93A

ABSTRACT

A significant anomaly, Fuel Valve Pilot Seal Extrusion, is affecting the Shuttle Primary Reaction Control System (PRCS) Thrusters, and has caused 79 to fail. To help address this problem, a Shuttle PRCS Thruster Process Evaluation Team (TPET) was formed. The White Sands Test Facility (WSTF) and Boeing members of the TPET have identified many discrete valve current trace characteristics that are predictive of the problem. However, these are difficult and time consuming to identify and trend by manual analysis. Based on this exhaustive analysis over months, 22 thrusters previously delivered by the Depot were identified as high risk for flight failures. Although these had only recently been installed, they had to be removed from Shuttles OV103 and OV104 for reprocessing, by directive of the Shuttle Project Office. The resulting impact of the thruster removal, replacement, and valve replacement was significant (months of work and hundreds of thousands of dollars). Much of this could have been saved had the proposed Neural Network (NN) tool described in this paper been in place.

In addition to the significant benefits to the Shuttle indicated above, the development and implementation of this type of testing will be the genesis for potential Quality improvements across many areas of WSTF test data analysis and will be shared with other NASA centers. Future tests can be designed to incorporate engineering experience via Artificial Neural Nets (ANN) into depot level acceptance of hardware. Additionally, results were shared with a NASA Engineering and Safety Center (NESC) Super Problem Response Team (SPRT). There was extensive interest voiced among many different personnel from several centers. There are potential spin-offs of this effort that can be directly applied to other data acquisition systems as well as vehicle health management for current and future flight vehicles.

The preliminary ANN tool developed during this fellowship for the Component Test Facility (CTF) program was designed with the following concepts in mind:

1) Engineering expertise can be incorporated into software and utilized as a consistent diagnostic tool.
2) The tool will run in parallel with existing test equipment.
3) The tool will help technicians in the CTF evaluate and categorize hardware.
4) Data augmentation, storage, and ease of access will enhance the effectiveness of the tools diagnostic capability.

Further development of this tool is warranted based upon the results to date. Risks have been minimized with a proof of concept approach and the cost will be less than one refurbished valve. Plans have been developed with budgets and schedules to study, configure, test, document, and train operators for the final configuration in the CTF.

INTRODUCTION

Solenoids are used to control fluid flow electronically. A solenoid valve is a control unit which, when electrically energized or de-energized, either shuts off or allows fluid to flow. The actuator takes the form of an electromagnet. When energized, a magnetic field builds up which pulls a plunger or pivoted armature against the action of a spring. When de-energized, the plunger or pivoted armature is returned to its original position by the spring action.

With a direct-acting solenoid valve, the seat seal is attached to the solenoid core. In the de-energized condition, a seat orifice is closed, which opens when the valve is energized. With direct-acting valves, the static pressure forces increase with increasing orifice diameter which means that the magnetic forces required to overcome the pressure forces, become correspondingly larger. Internally piloted solenoid valves are therefore employed for switching higher pressures in conjunction with larger orifice sizes; in this case, the differential fluid pressure performs the main work in opening and closing the valve. Figure 1 illustrates the RCS pilot actuated valve which is the point of focus for this report.

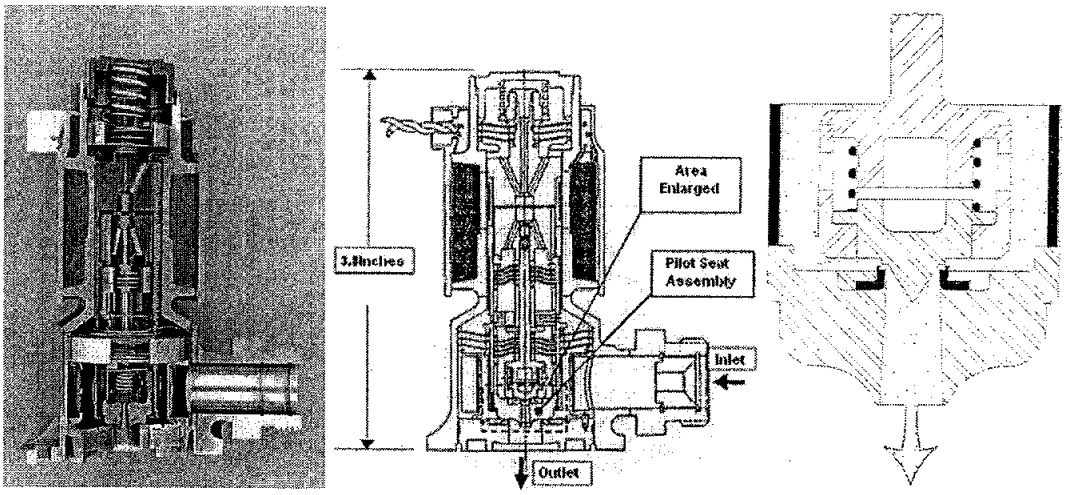


Figure 1: (Left) A cutaway internal view of the pilot-actuated value, (Middle) a diagram of the valve indicating the area of interest, (Right) an enlarged view of the pilot valve and extruded seal which hinders flow and reduces the pressure differential.

The deformation and extrusion of the pilot seal will cause an obstruction of fluid flow from the upper chamber during operation. Many papers have been written as to the cause of this extrusion and a definitive paper is being published. A full explanation as to the cause of pilot seal extrusion may be obtained by contacting the NASA colleague of this paper. However, if the seal obstructs the flow, then a sufficient pressure differential will not be created thereby causing main stage failure. This problem can be characterized

by data acquisition of the current trace in the CTF which would then identify the valve for refurbishment.

ARTIFICIAL NEURAL NETWORKS

An Artificial Neural Network is modeled upon the human brain's interconnected system of neurons. Neural networks imitate the brain's ability to sort out patterns and learn from trial and error, discerning and extracting the relationships that underlie the data with which it is presented. Most neural networks are software simulations run on conventional computers. Each neuron in the network has one or more inputs and produces an output; each input has a weighting factor, which modifies the value entering the neuron. The neuron mathematically manipulates the inputs, and outputs the result. The neural network is simply neurons joined together, with the output from one neuron becoming input to others until the final output is reached. The network learns when examples (with known results) are presented to it; the weighting factors are adjusted by training algorithms which bring the final output closer to the known result. Neural networks are good at providing very fast, very close approximations of the correct answer. Although they are not as well suited as conventional computers for performing mathematical calculations, neural networks excel at recognizing shapes or patterns, learning from experience, or sorting relevant data from irrelevant. Their applications[1] can be categorized into classification, recognition and identification, assessment, monitoring and control, forecasting and prediction.

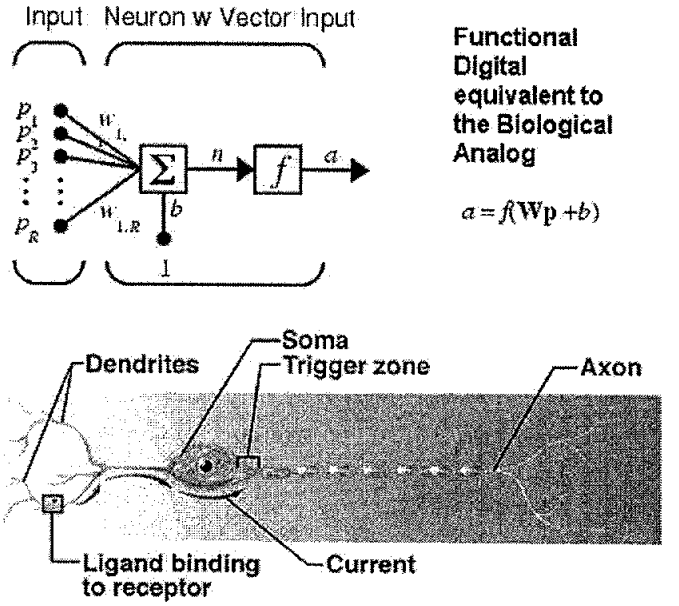


Figure 2: Depicting the Digital analogy of a single Neuron with that of its Biological equivalent.

DATA ANALYSIS

The first data set analyzed consisted of 20 current traces pulled from nominal and failed valves. These were typed utilizing a TPET presentation guideline followed in the CTF. A new category was developed for those square traces that were very close to nominal. Due to limited data, the NN was trained using a statistical approach. Boundaries were developed for the data using 2 standard deviations from the mean which limited the training set to 15 derived current traces given 5 categories. A back propagation neural network was built and trained with this initial data set. A cost function was developed to flatten the results outside of the normal NN training process.

The second data set consisted of 17 current traces. The NN found anomalies in several of the data traces. It identified 1 of the traces as unknown. On closer inspection this trace was taken using gas instead of water. Water traces are more consistent and they are the only data to be analyzed with this technique. Another anomaly occurred with a valve that went from unknown to round to nominal all within the same test day. This valve was thrown out since the data was extremely inconsistent. These 4 pieces of data were rejected and not included in this data set augmentation. Another new category was developed for valves that were rounded but very close to nominal. Figure 3 shows a sample analysis of the nominal data traces. The training data for the NN is located in the bottom panel. This represents the mean with 2 standard deviation limits.

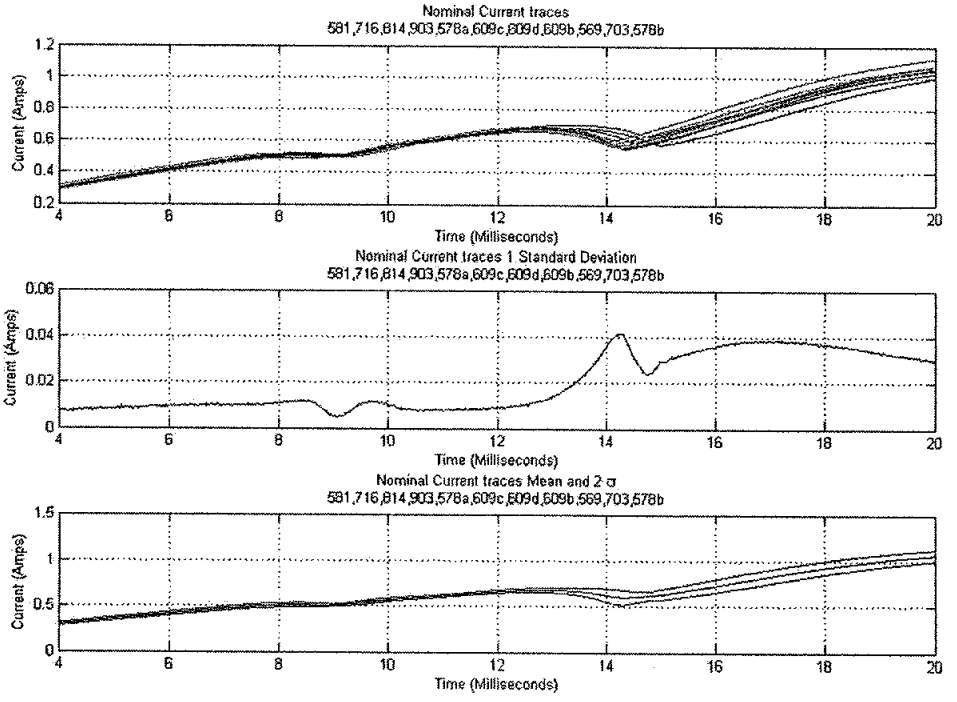


Figure 3: (Upper) Nominal current trace, (Middle) Standard deviation, (Bottom) Mean and 2 standard deviation boundaries.

Figure 4 depicts all categorical bounds and the finalized 18 training sets used in the NN analysis which became the network implemented in the Optics Lab and the CTF. There is substantial overlap in the boundaries which indicates that manual analysis would tend to be error prone for valves that fell between categories.

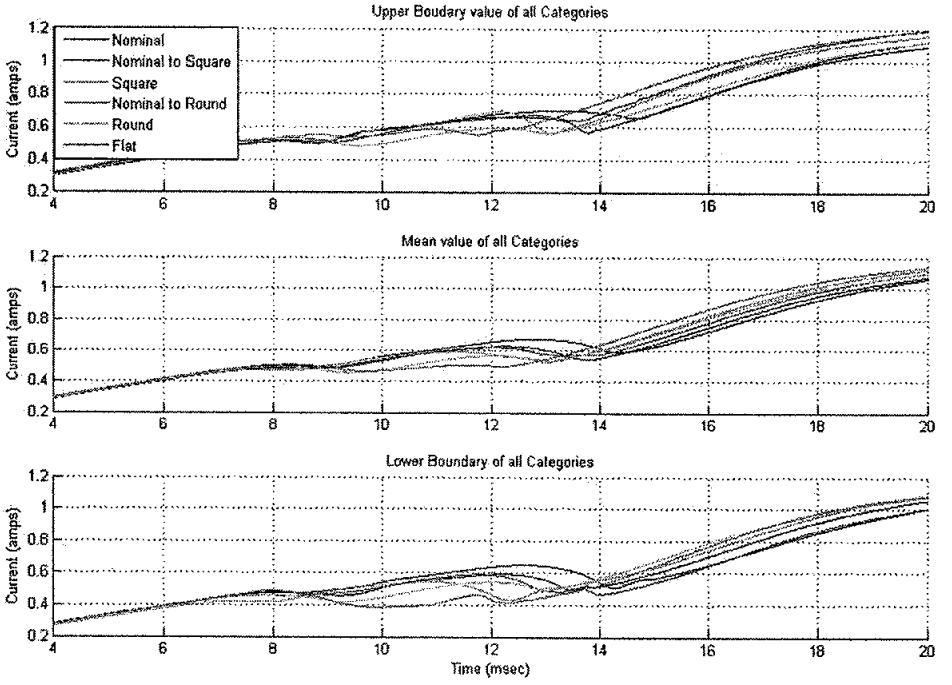


Figure 4: (Upper) Upper Boundary current trace, (Middle) Mean, (Bottom) Lower Boundary

STUDIES CONDUCTED

Once the NN was trained[2], then it was necessary to determine its predictive capabilities in the presence of noise, bias, and scale factor errors. A noise profile was developed to include all three of these characteristics. Though not depicted, the noise was normally distributed and increased with respect to time to its stated value in the figures to follow. Similarly, the biases where increased as a function of time to mimic bias conditions in the presence of scale factor errors.

Figure 5 shows that without any error, the training set and the extended data set obtained zero categorical errors when noise was absent. As the noise increased then so did the errors. The magnitude of the noise experienced in the CTF is 1/40 of that which would cause any significant degradation in performance. However, small negative biases or scale factor problems would tend to create unwanted errors. The NN could have been

trained to account for these errors but the training time proved to be unnecessary. This result justified the preconditioning of the data by removal of biases and the careful determination of scale factors. All of which are easily accomplished when the interfacing hardware is built and the software are initialized.

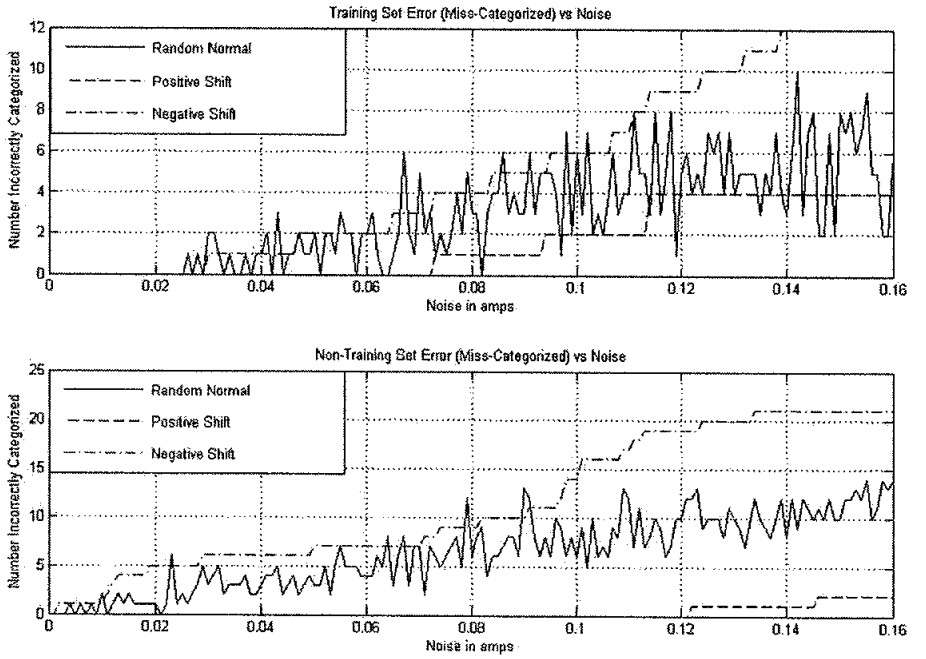


Figure 5: (Top) Categorization error for the training set as a function of Noise, Bias, and Scaling Errors, (Bottom) Categorization error for the data set as a function of noise.

The importance of passing good hardware vs. refurbishing bad hardware prior to an In-Flight Failure (IFF) or In-Flight Anomaly (IFA) should be explored in a little more detail than indicated above. Let us presume that instead of looking at numbers of errors we determine what types of errors occur given noise. Figure 6 shows in the top panel that as noise increases nominal hardware starts to fail. However, the bottom panel indicates by comparing the top that no failed hardware ever becomes nominal with an increase in noise. Also, the other categories only indicate the degree of failure so misdiagnosing a valve due to noise would still require action that would prevent an IFF/IFA. Therefore, the worst response that could happen with the present design configuration and an inordinate amount of noise would be refurbishment of good valves. The primary failure mode is not a safety concern but a cost and schedule consideration. This is an acceptable mode of failure which can be caught by monitoring failure rates and data traces which are already done in the facility as a matter of course. In addition, software can calculate the noise levels during any given run and indicate if this has impacted the ability of the NN to categorize the valve properly.

The studies above indicate that the present design concept would work very well in the CTF facility. All data traces were properly characterized given the statistical approach to train the network and correctly functioned under abnormal conditions. These promising results prompted the decision to attempt an actual test under simulated as well as actual CTF conditions.

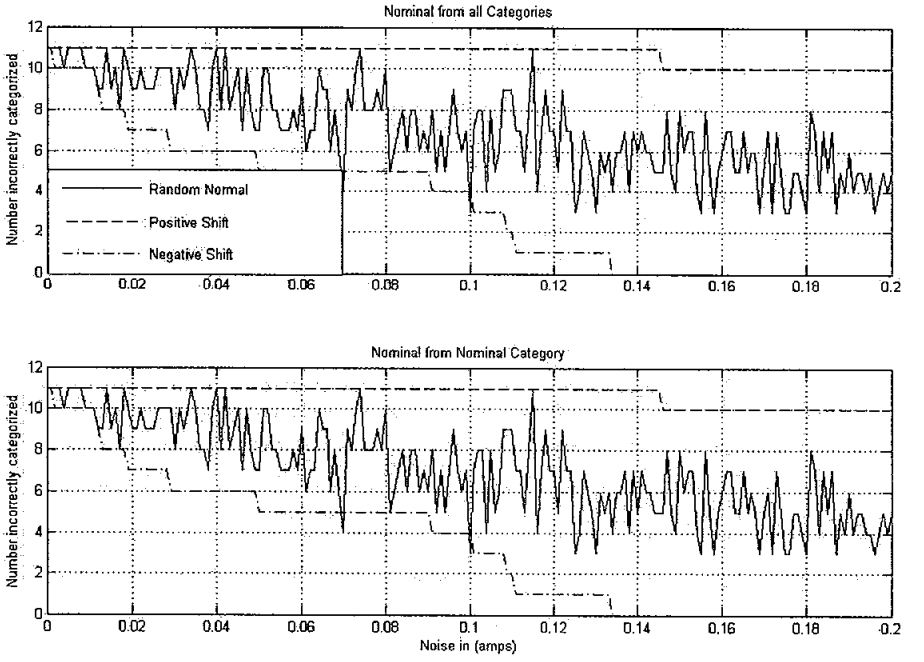


Figure 6: (Top) Nominal categorization error for the full data set as a function of noise, (Bottom) Nominal categorization error for the nominal data set as a function of noise.

PROGRAM IMPLEMENTATION AND DEVELOPMENT

The Artificial Neural Network (ANN) Algorithm was developed using MATLAB and the data recorded from the CTF which was reduced using the Znet tool. To develop software in the shortest time frame feasible it would be necessary to use as many existing components as possible. Therefore, the NN coefficients were transferred from MATLAB to a specially designed C++ program to duplicate the MATLAB results. The Znet program was written in C and its developer made minor modification to write the real time data to a file wherein the program would call the C++ NN program which would write the results to another file. Though dissimilar programs, the interface was reduced to reading and writing files.

Finally, the operator interface had to be designed and developed to convey all important information that the technician and engineers currently have to analyze by hand. The important information is plotted, compared as well as written. To categorize the valve as to type, the nominal limits are plotted in the top panel of Figure 7. The

actual trace is simultaneously plotted for visual verification of the type which is printed in the title. (The bottom panel duplicates the trace as currently output in the strip charts.) The title contains the pilot opening time, bias, and noise characteristics to monitor data integrity.

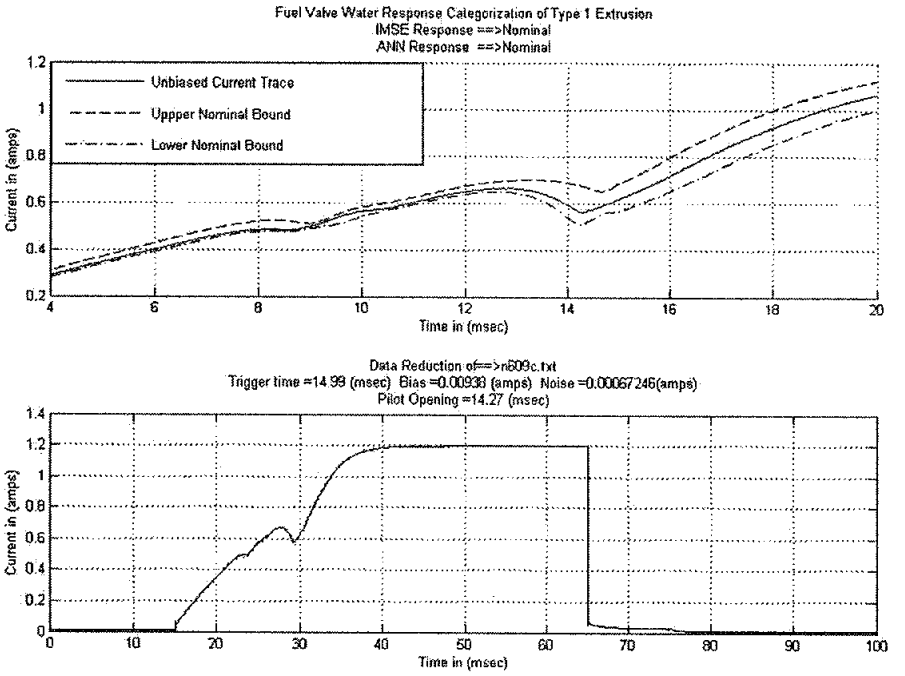


Figure 7: (Top) Operator interface showing the nominal boundary, a nominal data run and indicators for the category, (Bottom) Strip Chart analog of the current data trace with bias and pilot opening time.

OPTICS LAB TEST

The decision was made to proceed with actual hardware testing based upon encouraging research results and available time to structure a test of a mule valve in the Optics Lab. Hardware needed to be designed and built while software was simultaneously written and developed to accomplish a test which would put the design concept through its paces.

Figure 8 depicts the recorded results that would have been obtained using the ANN tool with valve SN 754. The data indicates a failed valve with a flat response. This represents the same operator interface which depicts the valve category, pilot opening time, bias, and noise characteristics. The purpose of presenting this valve characterization is simply to juxtapose with a like categorization of the mule valve that was available for our trial test.

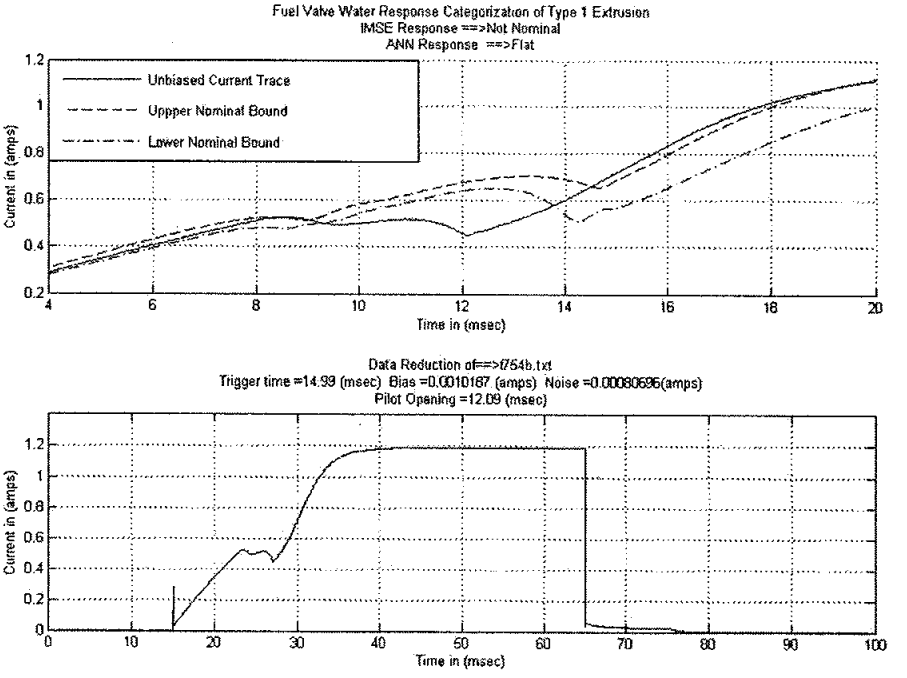


Figure 8: (Top) Operator interface showing the nominal boundary, a nominal data run and indicators for the category, (Bottom) Strip Chart analog of the current data trace with bias and pilot opening time.

The test setup in the Optics Lab consisted of a Laptop computer with the ANN software installed, a National Instruments DAQ pad, an oscilloscope, amplifier, FET switch and mule valve. The DAQ pad had been utilized in previous tests on a different project which resulted in several damaged I/O ports. This coupled with the power conditions in the lab probably were the cause of larger than expected noise levels. The mule valve was never an operational piece of equipment but did serve to simulate what could be expected in the CTF facility with actual hardware.

Therefore, when reviewing the mule valve SN 009 results (Figure 9) we see very similar tendencies as the previous failed flat response. The differences are due to the test setup and equipment utilized to obtain and analyze the data. The system experienced an increase of noise by a factor of 25 and still successfully categorized the valve. This is not surprising considering the levels that would be required to create categorical errors indicated in the Study section. The data in the CTF is much cleaner then our jury rigged testing done in the Optics Lab. The trigger pulse was initiated by hand and turned off by hand which is evident by the valve on time and the spike which initiates the current trace.

Characteristics that caused problems were not directly related to the NN software but had more to do with how to identify the trigger pulse in noise and where the pilot opening time was located in the noise. To compensate for noise effects, a variable bandwidth filter was developed and added to the software which feeds data to the ANN

algorithm. With this small modification, any noise problem carried into the CTF facility with our test equipment would be neutralized.

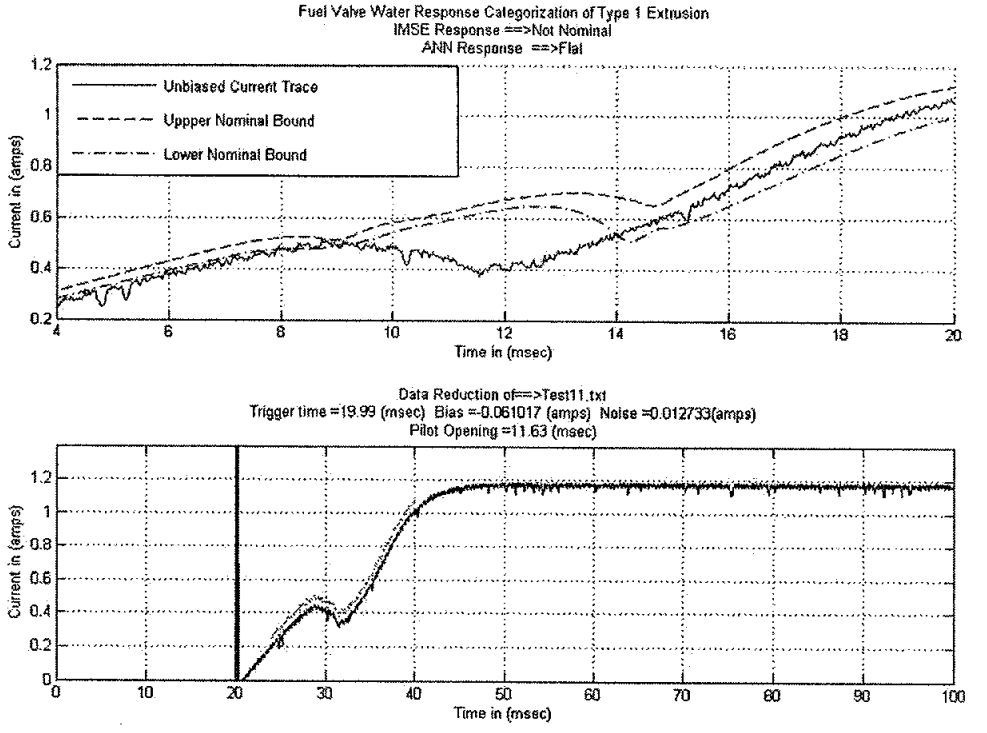


Figure 9: (Top) Operator interface showing a flat response from the Mule Valve in the Optics Lab, (Bottom) Strip Chart analog of the current data trace with bias and pilot opening time.

COMPONENT TEST FACILITY (CTF)

There was sufficient time during the development of the ANN tool to devise a demonstration of its utility with hardware and software in the CTF using actual RCS hardware. Cables were modified to run this Tool in parallel with the existing depot level test equipment. Valves that had been previously failed were identified for inclusion in this test process as well as the mule valve used in the Optics Lab Test.

The first process involved running the mule valve to verify that the cable modification did not make any difference to the results of the data. The first three runs were taken without any modification to the test equipment. The next 3 runs were taken with cable modifications to the test equipment as well as the ANN tool running in parallel. Figure 10 shows the results of all 6 runs. The variation from run to run was well within normal run to run variations associated with repeating the tests on the same valve. Valve opening times varied by less than .25 milliseconds

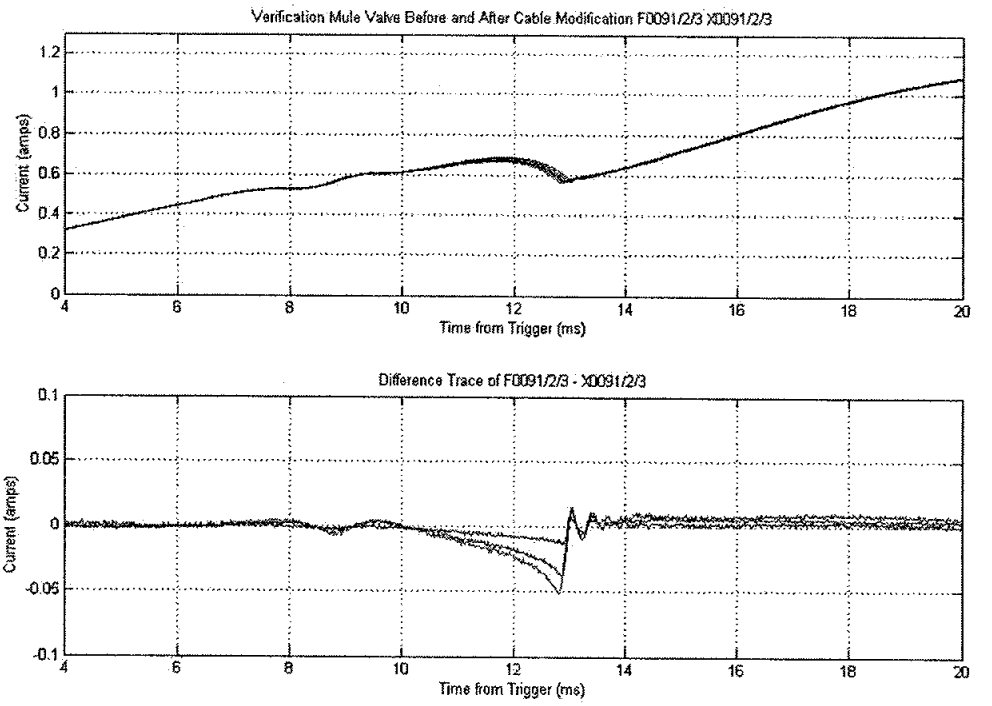


Figure 10: (Top) CTF verification run of mule valve, (Bottom) Difference trace of valve current comparing the modified cables.

The second process involved running the failed valves (SN 256 and 534) previously categorized as rounded by manual analysis. Figure 11 shows the results of this comparison. The test stand digital recording and the ANN tool recording are plotted and then differenced. The noise levels are about 10 times the magnitude as seen in the original data. This is due to the DAQ pad, as first seen in the Optics Lab test. The neural net was built to handle more noise than this. There is a .02 amp bias indicated as well from the DAQ pad but the ANN tool compensates by deleting the bias. Very little time was allocated for setting the scaling of the DAQ pad so initial results used a value with 2% error. This is shown by the difference slope increase. When corrected the scaling error was essentially zeroed, however the ANN tool was built to handle larger scaling errors so there was no need to compensate for this small error.

Figure 12 shows the final result for valve 256. Just as it had initially failed in the CTF facility prior to running the ANN Tool, it is evident that the valve is not nominal. It agrees with previous manual assessments that the valve has rounded characteristics which identifies the valve as unacceptable and should be scheduled for refurbishment.

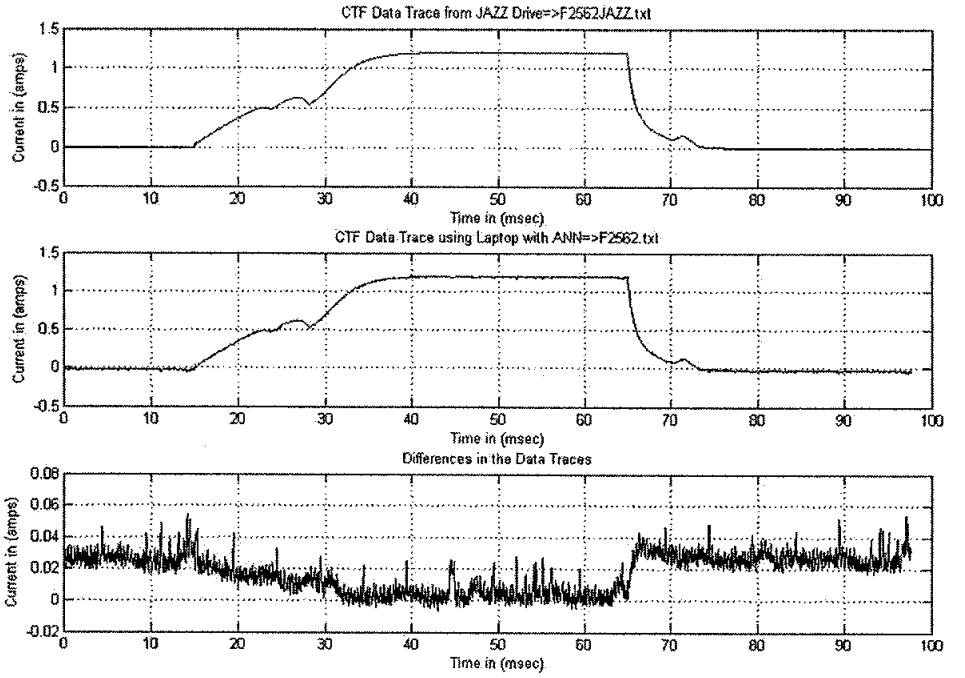


Figure 11: (Top) CTF 256 Valve trace using test stand hardware, (Middle) 256 Valve trace using parallel ANN Tool, (Bottom) Difference trace of current.

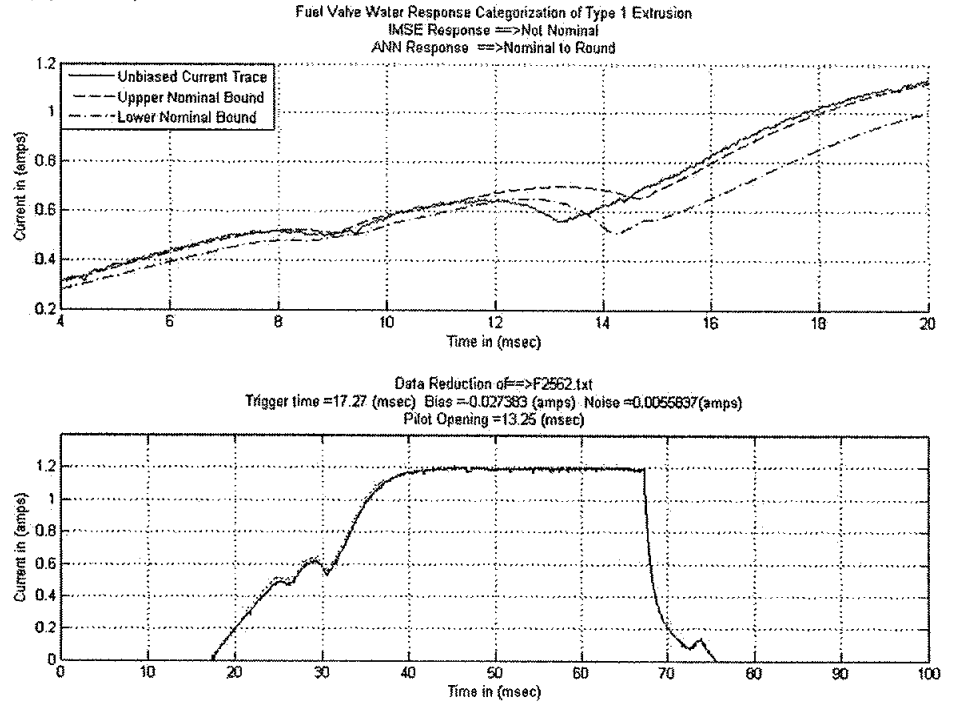


Figure 12: (Top) Operator interface showing a rounded response for the 256 Valve, (Bottom) Strip Chart analog of the current data trace with bias and pilot opening time.

CONCLUSIONS

It comes as no surprise that engineering expertise can be incorporated into software. Neural networks are suitable for pattern recognition which is the task currently required for categorization of the PRCS valves. The difference is that algorithmic implementations are consistent and non-subjective tools when compared to manual analysis.

The ANN tool is a software algorithm that will run in parallel with existing test equipment. It has been shown that there is no loss of data integrity and with minor modifications will represent a diagnostic tool that enhances the CTF's capability of typing and storing data. Further development of this tool is warranted based upon the results to date. Implementation risks have been minimized, success has been maximized, and the cost will be less than one refurbished valve.

PROPOSED PROJECT CONTINUATION

A single data trace was used to characterize and categorize the PRCS valves as a suitability test. Additional information in the form of pressure, accelerometer, and Hall-Effect traces should be incorporated to completely characterize the status of the thruster valve. Therefore, existing and historical data traces should be analyzed to enhance the accuracy of the results while augmenting the output of the current tool with the main valve opening times. Hardware and software should be configured and incorporated into the CTF as an assessment tool. Plans to this effect have been developed and presented during the final presentation of results to the WSTF management and staff.

Future studies would be important to optimize the performance of the neural network in the proposed ANN tool. These studies would determine the optimal neural network type, number of layers and neurons, training algorithm, and category refinement. In addition, the RCS valve simulation should be further development, understood and documented. This would help in fully understanding the properties of the valve and might be included in the ANN tool which could then predict seal extrusion and useful valve life before refurbishment.

REFERENCES

1) Corder, Mike. "Crippled but Not Crashed", Scientific American, August 2004, vol. 291, no. 2, pages 94-95

2) Hagen, Martin, et.al., Neural Network Design, University of Colorado, 1996.

Urban forms, physical activity and body mass index: a cross-city examination using ISS Earth Observation photographs

Prepared by :	Ge Lin, Ph.D
Academic Rank	Assistant professor
University & Department	West Virginia University Department of Geography and Geology Morgantown, WV 26506
NASA/JAC	
Directorate	Space and Life Sciences
Division	Earth Observations
Branch	
JSC Colleague:	Kam Lulla, Ph.D
Date Submitted	September 12, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

Johnson Space Center has archived thousands of astronauts acquired Earth images. Some spectacular images have been widely used in news media and in k-12 class room, but their potential utilizations in health promotion and disease prevention have relatively untapped. The project uses daytime ISS photographs to define city forms and links them to city or metropolitan level health data in a multicity context. Road connectivity, landuse mix and Shannon's information indices were used in the classification of photographs. In contrast to previous remote-sensing studies, which tend to focus on a single city or a portion of a city, this project utilized photographs of 39 U.S. cities. And in contrast to previous health-promotion studies on the built environment, which tend to rely on survey respondents' responses to evaluate road connectivity or mixed land use for a single study site, the project examined the built environments of multiple cities based on ISS photos.

It was found that road connectivity and landuse mix were not statistically significant by themselves, but the composite measure of the Shannon index was significantly associated with physical activity, but not BMI. Consequently, leisure-time physical activity seems to be positively associated with the urban complexity scale. It was also concluded that unless they are planned or designed in advance, photographs taken by astronauts generally are not appropriate for a study of a single-site built environment nor are they appropriate for a study of infectious diseases at a local scale. To link urban built environment with city-wide health indicators, both the traditional nadir view and oblique views should be emphasized in future astronauts' earth observation photographs.

INTRODUCTION

In the last three decades, the number of overweight and obese individuals has increased at an alarming rate in the U.S. Substantially reducing overweight and obesity nationwide (i.e., by one-third) has become a top public health objective (Healthy People 2010). Numerous articles and special issues of leading medical and public health journals, such as the September 2003 issues of the *American Journal of Public Health* and the *American Journal of Health Promotion* (Giles-Corti, 2003, Ewing 2003), and numerous issues of the *American Journal of Preventive Medicine* have been devoted to environmental risk factors and determinants that predispose people to being overweight or obese. The February 2003 issue of *Science* also devoted a special section to obesity and environmental factors (Hill 2003). The consensus is that a decrease in physical activity, together with increased energy intake at the societal and population level, largely contributes to the obesity epidemic.

Many investigators have related neighborhood characteristics with physical activity and body mass index (BMI; Egger and Swinburn 1997). A neighborhood, which is usually defined as several city blocks or sometimes by census tract, is conducive to outdoor activities if it can provide various incidental opportunities for walking and biking to shops and parks and to conduct daily business (e.g., banks, postal service). Numerous studies have associated a greater level and intensity of physical activity with mixed land use, neighborhood trails, hilly landscapes, and proximity to parks and recreational facilities (Giles and Donovan, 2003; Leyden, 2003, Lindström 2003). Most studies that have examined environmental determinants of physical activity, however, used respondents' perceived environmental and neighborhood factors (Brownson et al 2001). Objectively measured neighborhood environments are found only in very localized studies that involve very few neighborhoods (Saelens et al 2003), and there is a great demand in the field of health promotion to provide and to link objectively measured neighborhood characteristics to community health indicators such as physical activity.

There are three elements of the built urban environment: urban configuration (i.e., arrangement of physical elements), land use (e.g., the location and density of residential, commercial, and other spaces), and transportation network (e.g., roads, railroad tracks, bridges). Previous remote-sensing studies of urban environments have covered a wide range of topics (e.g., urban sprawl, change in land use, estimates of housing and population density, urban morphology) about built environments (Lo and Yang 2003; Civco et al. 2003) Most of these studies, however, have addressed only one element of the urban built environment or a single study site (i.e., a city). To date, no multicity study has been conducted.

There are many sources of digital data on urban built environments, but most of them are inaccessible at a neighborhood scale. Aerial photographs and high-resolution satellite images are available for all cities in the United States, but they cost tens of thousand dollars for a multicity study. The National Aeronautics and Space Administration (NASA) has more than 500,000 Earth photographs housed at the Johnson Space Center that provide a unique data source for investigating urban environments around the world (Lulla and Dessinov, 2000). Astronauts took the photographs, of varying quality, during numerous shuttle flights and while on board the Mir Space Station

or the International Space Station (ISS). The recent ISS photographs are more consistent in quality than the others and include images of major U.S. cities.

ISS photographs have several advantages for exploring built environments. First, they are in the public domain and are downloadable free of charge. Second, the quality of the photographs is acceptable for testing a research hypothesis or for making a quick assessment of a study area. ISS photos taken from a 400-mm or 800-mm lens provide a ground resolution between 20 and 6 meters, depending partially on the lens used and partially on the distance to the Earth's surface. The photographs taken at a point directly below the ISS (i.e., the nadir view) have the best resolution because of the shorter distance to the Earth. Third, the photographs taken by the astronauts are prepared for the general public, and they require much less knowledge to process or to read than satellite images, such as TM, yet they have spatial resolution that is comparable to TM or to SPOT. Fourth, astronauts are trained observers and can capture images of cities from various angles with variable scales, which may reveal more information about urban morphology than images taken from the nadir view only.

Herein, I describe how I used ISS photographs in an investigation of urban built environments in an attempt to link individual physical activity data for selected U.S. metropolitan areas with characteristics that are likely to be amenable to physical activity. I first identify the national survey data that we linked to the ISS images and then describe how I developed and tested a set of operational measures of the amenability of urban environments to outdoor activities based on selected ISS images.

METHODS

(1) The Behavioral Risk Factor Surveillance System Survey.

I evaluated several national-level data sources that included information about physical activity and BMI at the individual level—the National Health Interview Survey (NHIS), the National Health and Nutrition Examination Survey (NHANES), and the Behavioral Risk Factor Surveillance System (BRFSS) survey. I chose to use the data from the BRFSS survey because it was the only study that had geographic information at the county or metropolitan–area level. The survey was conducted by telephone in 2003 and sampled more than 250,000 respondents from among the noninstitutionalized adult (≥18 years of age) population in the U.S. The 2003 BRFSS survey had several sample-weight variables to correct for variations in sampling schemes (Holtzman 2002). The *final person weight* variable was designed to correct biases so that the final BRFSS survey sample for analysis could be nationally representative. Each respondent was asked about basic demographic and socioeconomic variables and about body weight, height, and the frequency and intensity of leisure-time physical activity. They were also asked about such chronic conditions as hypertension, diabetes, and heart disease and about such health-related behaviors as smoking, diet, and alcohol consumption.

The 2003 BRFSS survey data provide the total number of minutes per week in which the subjects engaged in 1) moderate physical activity or 2) vigorous physical activity during their leisure time. Both moderate and vigorous physical activities were

combined into a single measure of time (in minutes) spent engaged in physical activity outside of the workplace.

(2) Measuring outdoor amenability to physical activity.

There are many ways to measure the outdoor amenability of an urban environment. The new urbanism movement, which emerged in the late 1980s, embraces urban design and planning principles that both create great public places and reduce automobile use. The *Good City Form* (Lynch, 1981) provided a language and conceptual framework for describing and evaluating the built environment and defined physical characteristics. In his early (1961) and more recent work (1981), Lynch suggested that the more physically complex a city is, the greater number of incentives there are for residents to walk. According to Lynch, a city can be portrayed vertically, horizontally, and architecturally. These dimensions can be defined abstractly by using the Shannon index (Haken and Portugali 2002), which is an information index that portrays the complexity of three combined dimensions. For the purposes of the current projeect, I only considered the horizontal arrangements of a city that are amenable to physical activity.

Appropriately measuring the link between the built environment and outdoor activities is not a trivial task, even when dealing with a single dimension. The measures most commonly used by researchers reflect the availability of data, as well as the traditional concerns of transportation planning, and are not necessarily suited to the study of the link between the built environment and physical activity. When examining interactions between the built environment and travel behavior, various elements of the environment are more appropriately measured at various geographic scales. Past research has typically focused either on the neighborhood, an area that is often conceptualized as encompassing several city blocks, or on broader regional scales, such as several square miles within a large city or metropolitan area or even an entire metropolitan area. I chose to test measures of the built environment by dividing those measures into local (i.e., neighborhood) and regional characteristics.

Based on Lynch (1981), there are at least three interrelated and often correlated elements of the built environment at the neighborhood scale:

1. Density and intensity of development. Density is a measure of the amount of activity found in a geographic area. It is usually defined as population, employment, or building square footage per area unit (e.g., people per acre, jobs per square mile). The floor-area ratio, defined as the ratio between the floor space in a building and the size of the parcel on which that building sits, is another popular measure of density. Density is perhaps the easiest characteristic of the built environment to measure and is thus widely used. Although the ISS photographs can be used to derive the average building setback, the problem with varying geographic scales makes it difficult to uniformly assess the building setback in a multicity context, so I did not evaluate it. I relied instead on population-density statistics from the 2000 U.S. population census to control and evaluate the density effect.

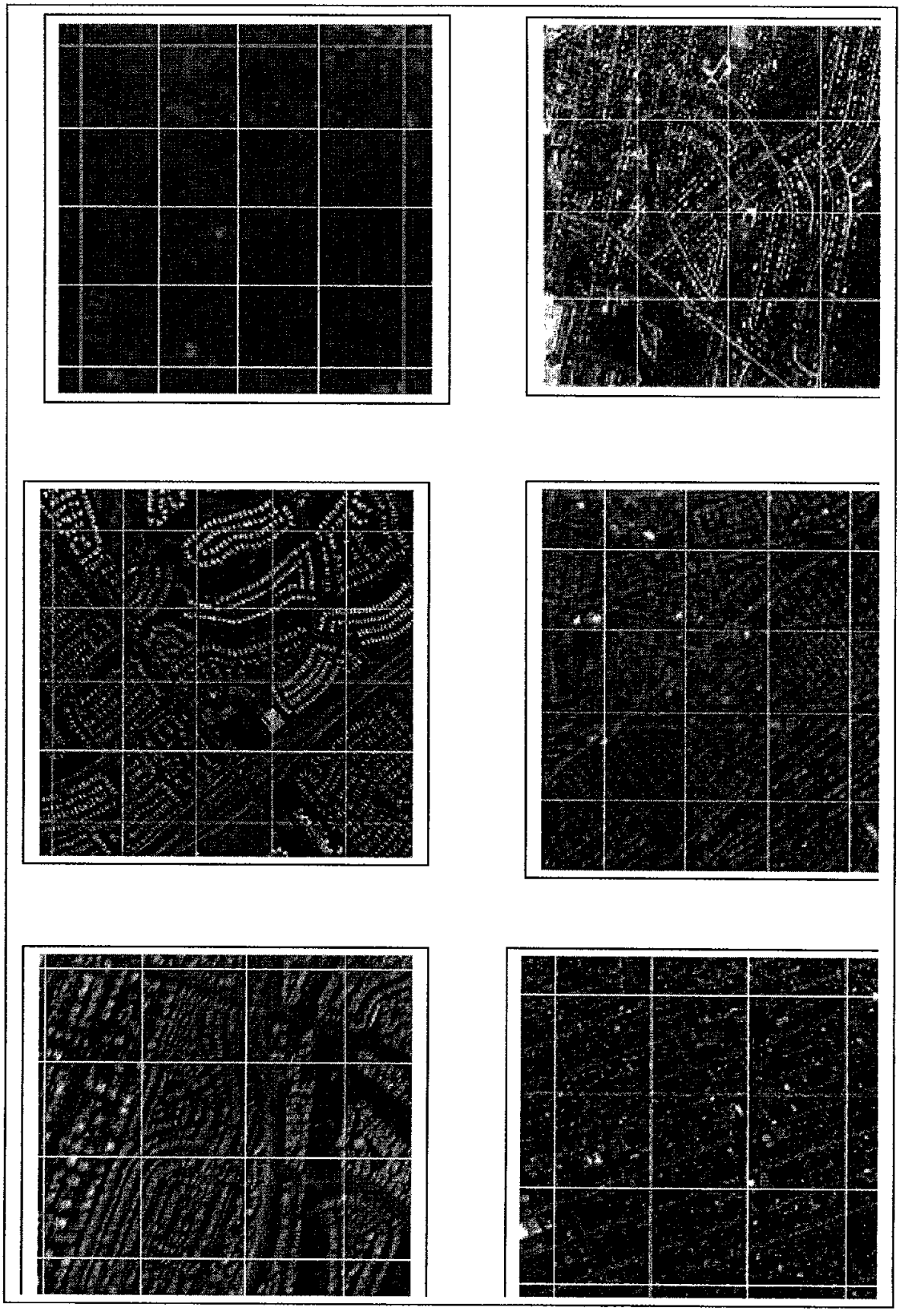
2. Land use mix. Land use mix is defined as the relative proximity of different land uses within a given geographic area. A mixed-use neighborhood would not only include homes but also stores, offices, parks, and perhaps other land uses. I devised a random grid with a center cell and 8 adjacent cells (see Figure 1 for an example). For each grid cell, I counted the number of neighboring cells occupied by different land uses.(Cervero and Kockelman, 1997).
3. Neighborhood connectivity. Connectivity is defined as the directness and availability of alternative routes from one point to another within a number of neighborhoods (Hess, 1997). There are several measurements, such as the number of intersections per square mile, the node-to-line segment ratio, and average block length. The node-to-line segment ratio is hard to manipulate in a global information system (GIS) or remote-sensing environment, because a node at the dead end of a street is still recognized to be a node by the database. To avoid this ambiguity in the GIS database, we used a simple index—the intersection-to-line segment ratio. The greater the node-to-line segment ratio is believed to be associated with greater connectivity of alternative routes (Greenwald and Boarnet 2002). I again used the same random grid from the land-use measure in 2) and counted the number of intersections and road segments within a grid. This process was repeated 20 times for each ISS photograph of a city, and the total numbers of intersections and road segments were used to derive the intersection-to-line segment ratio for each city.

A neighborhood was considered to be pedestrian friendly if it was densely developed with a mixed land-use pattern and a highly interconnected street network. Although different types of land use may be attractive, more road connectivity in an urban landscape also means simple layout, such as the Manhattan road network, which, according to Lynch (1981), could make walking during leisure time less attractive (more incidental physical activity from necessary walks). So one way to deal with this problem is to use a modified Shannon index (Haken and Portugali 2002). Shannon indices have several forms of operational equations, and almost all of them are taken as a ratio of possibility. If a road network is simple, the possibility of a road being extended to other possible routes is greater. In contrast, a road network that is more complex and, therefore, more interconnected, is less likely to be reconfigured, which means that the amount of possible information is less.

The $\log_2$ of the inverse of the road connectivity index is a proxy of the Shannon index. Similarly, the greater the number of land-use types, the greater the possibility of spatial configuration and the greater the amount of information. For any land-use pattern that has more than 1 category, the simplest land use for a grid cell would be one type of use only, and the most complex landuse would be 3 categories. Hence, when each neighboring cell has all three types, the total would by 3 times 8 (cells), or 24, the $\log_2$ (total neighboring land-use types/8) is another proxy of the Shannon index. By combining the calculations of both road and landuse complexity indices, I derived information about the complexity index of road and landuse as

$$\text{Shannon Index} = \log_2(1/(\text{road index}) + (\text{land-use index})/8) \qquad (1)$$

Figure 1. Random grid-samples of selected IIS photo (from left to right Boulder, CO; Bangor, ME; Phoenix, AZ; Boston, MA; Las Vegas, NV; Louisville, KY)



Anticipating some problems with the road-connectivity index based on the intersection-to-line segment ratio, I also tried to calculate the number of circular networks within a mile grid by using the U.S. Census TIGER file and comparing the results for the Houston metropolitan area. The TIGER road network, however, proved to be too approximative. Many streets in the Houston metropolitan area, for instance, would have a dead end, but the TIGER file does not contain this information (not shown, available upon request). For this reason, the TIGER file data would result in more error than those obtained from the ISS photos if they can be carefully selected.

Since I want to be able to distinguish between three types of land use—residential, recreational (green and water), and other built-up lands—I set the ground resolution to be ± 10 meters. This requirement limited the camera lens to > 400 mm. In addition, the ISS photographs must cover a part of metropolitan area where the BRFSS survey had some respondents so that survey data could be linked to the ISS photos. Combining these two requirements, and some photo-quality requirements (e.g., clear sky), I found about 70 ISS photos that cover 39 U.S. cities. For a complete list of the ISS photos used in this study, see Appendix 1.

(3) Deriving city measures.

I used a center grid-cell about 100 meter wide to form a 3 × 3 grid to move around for evaluating land use based on the method in 2); I then this 3 × 3 grid to evaluate the road network based on the method in 3). For each photo, I randomly move this center grid 20 times to select sample grid. Since each selected grid location is randomly generated, I could not control its location. Nevertheless, I made sure that there was no overlap between two center cells in any randomly selected location. In addition, if the center cell did not cover a residential area, I regenerated the grid until it covered a residential neighborhood. An analytical algorithm for each city is:

Step 1. Generate a 100-meter box, together with 8 neighboring boxes, by using the queen's rule.
Step 2. If the center box touches a residential neighborhood, then selected the grid, otherwise, repeat step 1 until the condition is met.
Step 3. For each center box, count the number of land uses (3 types; bodies of water are excluded) in an adjacent cell and record it.
Step 4. Count the number of intersections and road segments within the 9-cell grid and store them in the memory (if a street extends outside the grid, do not count it).
Step 5. Repeat steps 1 through 4 a total of 20 times. For each repetition, the center box cannot overlap any other center box.
Step 6. Summarize the measures made in step 3 and divide them by 20; the result becomes the average land-use mix index. Add the total number of intersections and road segments in step 4 and the ratio of the two becomes the intersection-to-road segment ratio.

When there was more than one photo for each city, the two with the sharpest images were used, and each was sampled 10 times. The sampling was done in ArcView

3.2. For several cities (e.g., Green Bay, Wisconsin; Kansas City, Missouri; Indianapolis, Indiana), the ISS photographs had insufficient information, so I used some aerial photographs to supplement the evaluation process. I first tried to automate the process. However, the photo image has limited spectrum information, both supervised and unsupervised classification resulted in more time spending on visual correction of the resulted classification. I eventually did visual classification. The whole process took about 20 working days.

After deriving each measure for all 39 cities, I used equation 1 to calculate a Shannon index. I downloaded the population density information from the U.S. Census Bureau and added it to the three measures. I then linked the three measures of physical friendliness (i.e., population density, mixed land use, and road connectivity) along with the Shannon index to individual data in the BRFSS survey for further analyses.

(4) Multivariate analysis.

I then conducted a multivariate analysis using a multilevel model with the number of minutes spent per week engaging in physical activity being the dependent variable. Assuming that individuals from each city or metropolitan area would respond similarly to the three measures of outdoor amenability to physical activity, a multilevel analysis is appropriate because individual variables are nested within the city variables (Duncan, Jones and Moon 1998). In other words, the city variables will not change among individuals living in the same city. I fit a fixed-effect multilevel model with an intercept that accounts for this hierarchical structure. In the model, I also controlled for individual risk factors such as age, sex, race, vegetable intake, and smoking status. Suppose, for example, that the dependent variable is physical activity y, a simple fixed effect model is

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + ,...,+ \beta_k x_{kij} + \phi_1 road_j + \phi_2 landuse_j + \phi_3 popd_j + \mu_{0j} + \varepsilon_{ij} \qquad (2)$$

wherein the βs are parameter estimates for individual-level variables (x) that are indexed by individual (i) and county (j) and the ϕs are parameter estimates for county-level variables that are indexed only by j. Both β and ϕ are the fixed effects, leaving μ_{0j} as the only county-specific random effect (Goldstein et al. 2002). Preliminary results showed that BMI had no relationship with any of the measures, so I will not report them here.

RESULTS

Descriptive analysis. The complete list of the three measures, together with physical activity, is given in Table 1. Overall, the mixed-land-use scores were high for the larger cities and low for the smaller cities. Phoenix, Arizona, and Boston, Massachusetts, had the lowest mixed–land-use score at 11, and Boulder, Colorado, had the highest score at 18. Recall that the score equals 8, but it means that each neighboring cell has one type of land use. If the score is more than 16, it means that each neighboring cell on average has two types of land use (i.e., other built environments versus residential, or recreational versus residential, or recreational versus other built environments. Ironically, Boulder also had a lower score in street connectivity, as measured by the intersection-to-road segment ratio. So superficially, at least, one amenity measure does not predispose the

other. Interestingly, the average BMI in Boulder was second lowest among the 39 cities, and its average physical activity score was 93.47, which is much higher than the mean (85.8) for all the cities.

Table 1. Road connectivity, mixed landuse and Shannon indices among 39 cities

Metropolitan Area	Road index	Land Use index	Shannon Index	BMI	Physical activity (mim)	# Obs Matched
Arlington, VA	0.56	16.37	1.82	24.14	82.66	133
Atlanta, GA	0.55	13.20	1.67	26.96	78.41	460
Austin, TX	0.43	17.90	2.06	26.04	83.76	256
Baltimore	0.54	14.73	1.76	27.58	71.97	293
Bangor, ME	0.35	15.19	2.10	27.13	85.95	247
Boston, MA	0.71	11.52	1.39	25.88	79.53	935
Boulder, CO	0.38	18.77	2.19	24.77	93.47	246
Buffalo, NY	0.52	16.23	1.86	26.51	86.95	284
Chicago, IL	0.56	15.23	1.76	26.59	80.98	2041
Cincinnati, OH	0.65	14.26	1.61	27.28	95.93	546
Dallas, TX	0.72	12.21	1.42	26.51	88.89	501
Denver, CO	0.56	14.92	1.75	25.67	89.92	429
Des Moines, IA	0.43	15.43	1.94	26.99	78.28	591
Detroit, MI	0.45	13.92	1.84	27.84	79.65	501
District of Columbia	0.69	15.39	1.65	26.27	81.57	1951
Duluth, MN	0.43	17.18	2.03	26.29	90.20	163
Fort Worth, TX	0.67	14.47	1.61	26.63	78.54	378
Gary, IN	0.49	12.54	1.71	27.76	94.02	402
Grand Junction, CO	0.55	16.15	1.82	25.79	114.59	125
Green Bay, WI	0.35	13.77	2.04	26.56	76.38	157
Houston, TX	0.51	14.42	1.78	27.31	78.27	728
Indianapolis, IN	0.40	15.65	2.02	27.04	92.79	731
Kansas City, MO-KS	0.51	14.33	1.78	26.12	71.47	701
Las Vegas, NV	0.28	11.53	2.13	26.60	103.80	880
Los Angeles-Long Beach, CA	0.48	16.94	1.94	26.87	95.79	958
Louisville, KY	0.65	14.32	1.61	27.05	52.30	368
NYC, NY	0.76	13.70	1.49	25.75	76.79	486
Nashville, TN	0.44	14.56	1.89	26.49	67.61	230
New Orleans, LA	0.58	13.21	1.63	26.83	65.59	389
Oakland, CA	0.45	15.14	1.91	26.08	102.34	199
Oklahoma City, OK	0.37	14.73	2.03	26.90	74.14	1278
Omaha, NE	0.52	15.11	1.81	26.63	72.73	1121
Philadelphia, PA	0.62	14.21	1.64	27.55	86.31	310
Phoenix, AZ	0.34	11.04	1.94	26.21	82.04	805
Providence, RI	0.63	16.79	1.78	26.32	82.53	2319
San Diego, CA	0.28	15.14	2.28	26.53	106.08	367
San Francisco, CA	0.71	14.22	1.56	24.75	93.23	99
Virginia Beach, VA	0.38	17.35	2.13	27.37	108.09	262
Waterloo-Cedar, IA	0.34	15.50	2.14	27.10	71.83	129

Both San Diego, California, and Las Vegas, Nevada, had low road connectivity scores, but both cities had higher average physical activity scores. A connectivity index close to 0.3 suggests a triangle-like road system, where T-shaped intersections and dead-end streets are fairly common. Even though neighborhood road systems in San Diego and Las Vegas are not well connected, residents there might be more likely to either work out in a gym (Las Vegas) or walk in neighborhood (San Diego). New York City and Boston had higher road connectivity scores. A connectivity index close to 0.7 suggests a Manhattan network system. Since the dimensions of a single grid is about 300 × 300 meters (or 9 100-meter grids), an index close to 7 is very high. If an even larger grid were used, the index could be higher. Regardless of the scale used, a higher road connectivity score apparently is associated with less physical activity, because both New Yorkers and Bostonians tend to engage in less physical activity than the average. One problem with categorizing physical activity simply as moderate or vigorous is that no distinction is made between outdoor and indoor physical activities. In addition, such categorization seems to reflect only purposeful leisure-time physical activity, and it does pick up many incidental daily activities, such as running errands and shopping at a local store.

It seems counterintuitive that Boston, Dallas, and New York had greater outdoor amenability than San Diego or Virginia Beach, but the Shannon indices suggest that this is the case. If the vertical dimension of a city could be added to our analysis, the Shannon index would have a different meaning. In the context of this study, however, there appears to be no relationship between the Shannon index and physical activity scores without controlling other individual characteristics.

Multivariate analysis. I first ran a model that included the road and land-use indices—(model 1); I then dropped the road and land-use variables and added the Shannon index as the key area-level explanatory variable (model 2). The results for both models are shown in Table 2. In both models, the results for the demographic and behavioral risk factors were consistent with the literature. The age, sex, and race/ethnicity of the respondents to the 2003 BRFSS survey were also important demographic factors. The physical activity decreased with age, and males engaged in more physical activity than did females. Most minority groups (i.e., Asian, Black, and Hispanic) tended to engage in less physical activity, and physical activity increased with educational level. As expected, smokers had a higher level of physical activity, as did persons who ate more servings of fruits and vegetables per day.

None of the three variables that represented outdoor amenability to physical activity—road connectivity, mixed land use, and population density—were significant in model 1. Several explanations could be offered for the non-significant results for the road-connectivity and land-use variables. It is possible that there was some spatial mismatch in a particular area. The ISS photographs typically cover a part of a county, but the survey data may cover the entire area or a different part of the county. It is possible that there is no direct relationship between these indices and physical activity. It is noteworthy that the Shannon index was significant in model 2. Presumably, as

suggested by Lynch (1981.), the more complex the combined land use and road system, the greater the physical activity.

Table 2. Multilevel regression on BMI and physical activity

	Model I			Model II		
Individual variables	Estimate	Error	t Value	Estimate	Error	t Value
Intercept	66.9769	19.3112	3.47*	27.6349	17.5787	1.57
Age (18-29)**						
Age 30-44	-8.7808	1.8505	-4.75*	-8.7883	1.8505	-4.75*
Age 45-64	-16.1527	1.8663	-8.65*	-16.1598	1.8663	-8.66*
age 65-74	-25.6303	2.9451	-8.7*	-25.6418	2.945	-8.71*
age 75 or older	-50.8325	3.2427	-15.68*	-50.8386	3.2426	-15.68*
Sex (Male)	22.5497	1.4085	16.01*	22.5466	1.4085	16.01*
Race/ethnicity (White)						
Black	-4.8239	2.2042	-2.19*	-4.7633	2.2036	-2.16*
Asian	-22.2135	3.3292	-6.67*	-22.1685	3.3283	-6.66*
Hispan	-4.6968	2.1312	-2.2*	-4.6832	2.1306	-2.2*
Other races	17.6372	3.1253	5.64*	17.6376	3.1251	5.64*
Education (< high school)						
High School	4.9743	1.283	3.88*	4.9701	1.2829	3.87*
Associate Degree	5.4863	2.6272	2.09*	5.4739	2.6271	2.08*
College or higher	0.3139	2.6164	0.12*	0.3231	2.6162	0.12*
Fruit/veget Servings/day	3.4979	0.321	10.9*	3.5006	0.321	10.91*
Current smokers	6.0563	1.7742	3.41*	6.0536	1.7742	3.41*
City-wide variables						
intersection/road-segment ratio	-27.5512	16.3993	-1.68			
mixed landuse index	1.0315	1.1104	0.93			
Shannon index				22.1429	9.261	2.39*
Population density	0.06391	0.1812	0.35	0.0635	0.1698	0.37

* significant at $P<0.05$

**category in the parentheses is the referent

CONCLUDING REMARKS

In this project, I explore the feasibility of using ISS photographs to assess the built environment of urban areas. In contrast to previous remote-sensing studies, which tend to focus on a single city or a portion of a city, I used images of 39 U.S. cities. And in contrast to previous health-promotion studies on the built environment, which tend to rely on survey respondents' responses to evaluate road connectivity or mixed land use for a

single study site, I examined the built environments of multiple cities based on ISS photos. Although none of the citywide indicators was statistically significant by itself, the composite measure of the Shannon index was significant. Consequently, leisure-time physical activity seems to be positively associated with the urban complexity scale.

A truly representative Shannon index for a city also requires vertical measures, which I could not evaluate because of the lack of oblique-view data. The oblique view is important, especially for its ability to capture a vertical dimension. Oblique view is more intuitive to visualization. For cities located near a mountain or a hill, the oblique view can be a better source of information source than a three-dimensional model. It is suggested implicitly in the literature on remote sensing that the oblique view is inferior because people tend to correct the distortion caused by the view. Three-dimensional views also are distorted; using an oblique photograph sidesteps artificial distortion by a computer algorithm.

Unless they are planned or designed in advance, photographs taken by astronauts generally are not appropriate for a study of a single-site built environment nor are they appropriate for a study of infectious diseases, such as West Nile virus (Rogers et al., 2002). The ground resolution is usually sufficient for an infectious disease study, but an astronaut usually does not systematically take photographs of a city. This practice often leaves various holes in the landscape that are critical for studying the infectious environment. For a large-area study that requires a coarse ground resolution, an ISS photograph is a fine choice, because only one photograph is needed for each time period. Finer ground resolution requires that several photographs be taken, some from an azimuthal viewpoint from the ISS window. The latter requirement is problematic in the current ISS observations of the Earth.

Reference:

Alexander C. 1997 A pattern language: towns, buildings, construction. New York: Oxford University Press.

Beck, LR Lobitz BM and Wood BL 2000. Remote sensing and human health: new sensors and new opportunities.. Emerging Infectious Diseases 6:217-226.

Brownson RC, Baker EA, Housemann RA, Brennan LK, Bacak, SJ. 2001. Environmental and Policy Determinants of Physical Activity in the United States. Am J of Public Health ;91:1995–2003.

Calthorpe P. The next American metropolis: ecology, community and the American dream. New York: Princeton Architectural Press, 1993.

Cervero R, Kockelman K. 1997 Travel demand and the 3 Ds: density, diversity, and design. Transportation Res Part D 1997;3:199 –219.

Civco DL Hurd JD Wilson EH Arnold CL and Pristoe MP 2002. Quantifying and describing urban size in the Northeast Ustated States. PE & RS 69 1083-90

Egger G, Swinburn B. 1997. An ecological approach to the obesity pandemic. *BMJ.* 315:477–480.

Ewing R, Schmid T, Killingsworth R, Zlot A, Raudenbush S. 2003. Relationship between urban sprawl and physical activity, obesity, and morbidity. *Am J Health Promotion.* 18:47–57.

Giles-Corti B, Donovan RJ. 2003. Relative influences of individual, social, environmental, and physical environmental correlates of walking. *Am J Public Health.*93:1583–1589.

Goldstein H, Browne W and Rasbash J. 2002. TUTORIAL IN BIOSTATISTICS: Multilevel modeling of medical data. *Statist. Med.* 21:3291–3315

Greenwald M, Boarnet MG.2002. The built environment as a determinant of walking behavior: analyzing non-work pedestrian travel in Portland, Oregon. Transportation Res Record,1780:33–42.

Haken H and Portugali J. 2002. The face of the city is its information Journal of environmental psychology 23: 382-405

Hess PM. 1997. Measures of connectivity. Places 1997;11:58 –65.

Hill JO, Wyatt HR, Reed GW, Peters JC. 2003. Obesity and the environment: where do I go from Here? *Science.* 299:853–855.

Holtzman, D. 2003. The Behavioral Risk Factor Surveillance System. In: Blumenthal D, DiClemente R, ed. Community-based Health Research: Issues and Methods. New York: Springer Publishers:115-131

Katz P. The new urbanism: toward an architecture of community. New York: McGraw-Hill, 1994.

Leyden KM. 2003. Social Capital and the Built Environment: The Importance of Walkable Neighborhoods. *Am J of Public Health.* 93:1546-50.

Lo CP and Yang X 2003. Drivers of landuse/land-cover changes and dynamic modeling for Atlanta, Georgia Metropolitan Area. PE & RS 68: 1073-82.

Lulla KP and Dessinov LV (2000) *Dynamic Earth Environments: remote sensing observations from shuttle-Mir missions.* Ed. John Wiley & Son, NewYork.

Lynch K. 1981. *Good city form.* Cambridge, MA: MIT Press.

Lynch K., 1960 *The image of the city.* Cambridge, MA MIT Press

Rogers DJ Myers MF Tucker CJ Smith PF White DJ Backenson B Eidson M Kramer LD Bakker B and Hay SI 2002. Predicting the distribution of West Niles Fever in North America using satellite sensor data.

Saelens BE, Sallis JF, Black JB, and Chen D. 2003. Neighborhood-Based Differences in Physical Activity: An Environment Scale Evaluation. *Am J of Public Health.* 93: 1552-57.

Appendix I. The final list of ISS photos used in the report

WATERLOO_IA_ISS002-E-8856
Virginia_Beach_ISS001-E-6812
San_Fran_ISS002-E-9248
San_Fran_ISS002-E-6047
San_Diago_ISS002-E-7445
SAN DIEGO _ISS002-E-7443
SAN DIEGO _ISS002-E-7442
SAN DIEGO _ISS001-E-6361
San DeigoISS002-E-7441
San Deigo_upperISS002-E-7444
PHOENIX_ISS001-E-6350
PHILADELPHIA_ISS002-E-8024
OMAHA_ISS002-E-6957
Oklohoma_cityISS002-E-6358
OKLAHOMA CITY_ISS007-E-17125
OAKLAND, BRIDGES_CA_ISS002-E-9247
NORFOLK _VA_ISS006-E-52127_2
NIAGARA FALLS_ISS003-E-5109
NIAGARA FALLS _ISS002-E-6180
NewOrleans_ISS002-E-6936
NewOrleans_ISS002-E-6935
NEW ORLEANS_ISS002-E-7092
MANHATTEN ISLAND ISS002-E-6333a
MANHATTEN ISLAND ISS002-E-6333
MANHATTAN_ISS001-E-6630
MANHATTAN, _NY_ISS006-E-46068
LousvilleISS002-E-5323
LOUISVILLE_KY_ISS002-E-5323
LOS ANGELES, MARINA DEL REY ISS007-E-11930
LOS ANGELES, ISS007-E-11931
Las_VegasISS001-E-6659
Las_Vegas_ISS002-E-8486
Las_Vegas_ISS002-E-6229
LA_port_CA_ISS002-E-9253
LA_CA_ISS002-E-9252
KANSAS SPEEDWAY, KANSAS CITY _ISS006-E
Indiapolis_ISS004-E-7390
Houston_ISS001-E-6283
HOUSTON, DOWNTOWN_ISS007-E-15620
HOUSTON, DOWNTOWN_ISS007-E-15617
HOUSTON, DOWNTOWN_ISS007-E-15613
Greenbay_WI_NM21-763-37
GRAND JUNCTION_CO_ISS002-E-7452
Gary_IN_ISS006-E-49805
FORT WORTH _ISS001-E-6696
Duluth_MN_ISS002-E-7893
Detroit_ISS004-E-13710
Des Moines_IA_NM21-763-35
DC_ISS006-E-50925
DC_ISS002-E-8127
DC_ISS002-E-8125
DALLAS_ISS001-E-6699
Dallas_ISS001-E-6697
Cincinnati_ISS004-E-10467
CHICAGO_ISS006-E-49802
Chicago_ISS002-E-9840
Chicago_ISS002-E-8801
CHICAGO, HARBOR, PARK _ISS002-E-8798
CHICAGO O'HARE AIRPORT ISS003-E-5071
CHARLOTTE_SC_ISS003-E-6957
BOULDER_CO_ISS007-E-17065
BOSTON_ISS007-E-17770_2
BOSTON_ISS007-E-17770
Boston_ISS002-E-5553
Bangor_ME_ISS002-E-6171
Baltimore_ISS006-E-50932
Baltimore_ISS005-E-17522
Baltimore_ISS004-E-10673
AUSTIN_ISS007-E-11256_2
Arlington_VI_ISS002-E-8126

Advanced Water Recovery Technologies for Long Duration Space Exploration Missions

Final Report
NASA Faculty Fellowship Program – 2004

Johnson Space Center

Prepared by:	Sean X. Liu, Ph.D.
Academic Rank:	Assistant Professor
University & Department	Rutgers University Department of Food Science 65 Dudley Road New Brunswick, NJ 08901
NASA/JSC	
Directorate:	Engineering
Division:	Crew and Thermal System Division
Branch:	Advanced Life Support Office
JSC Colleague:	Daniel J. Barta, Ph.D.
Date Submitted:	August 9, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

Extended-duration space travel and habitation require recovering water from wastewater generated in spacecrafts and extraterrestrial outposts since the largest consumable for human life support is water. Many wastewater treatment technologies used for terrestrial applications are adoptable to extraterrestrial situations but challenges remain as constraints of space flights and habitation impose severe limitations of these technologies. Membrane-based technologies, particularly membrane filtration, have been widely studied by NASA and NASA-funded research groups for possible applications in space wastewater treatment. The advantages of membrane filtration are apparent: it is energy-efficient and compact, needs little consumable other than replacement membranes and cleaning agents, and doesn't involve multiphase flow, which is big plus for operations under microgravity environment. However, membrane lifespan and performance are affected by the phenomena of concentration polarization and membrane fouling. This article attempts to survey current status of membrane technologies related to wastewater treatment and desalination in the context of space exploration and quantify them in terms of readiness level for space exploration. This paper also makes specific recommendations and predictions on how scientist and engineers involving designing, testing, and developing space-certified membrane-based advanced water recovery technologies can improve the likelihood of successful development of an effective regenerative human life support system for long-duration space missions.

INTRODUCTION

Water is essential to all lives. Currently, potable water has been provided fully for the entire duration of the mission to all manned low orbit space missions including Space Shuttle and International Space Station (ISS) missions. The wastewaters generated in current space-related missions, including urine, hygiene water, and condensate water, are either discharged into space or brought back to the earth (although there is limited water recovery of certain streams of space wastewater in ISS, the recovered water is not used as potable water). This arrangement of water supply for long-duration manned space missions would be, of course, unattainable. And wastewater recycling becomes an essential part of Environmental Control and Life Support Systems (ECLSS) for all long-duration space exploration missions. Many terrestrial wastewater treatment technologies can be potentially adopted for extraterrestrial applications. However, many challenges need to be overcome in order to converting wastewaters to potable water as required by extended duration space exploration.

The main challenge of human presence in space is to duplicate the critical functions of intricate, interdependent processes that occur and sustain lives on earth. The water aspect of this challenge is to completely recover potable water from wastewater generated in the outer space with no need of water replenishment and substantial amount of consumables in a restricted and confined microgravity environment. Outer space is an unforgiving and daunting place, far away from any help from the earth, and this puts a huge premium on reliability and easiness of maintenance of water recovery systems. Microgravity introduces another dimension to an already-difficult problem in wastewater treatment in space exploration. Forces that are small in terrestrial flow situations such as surface tension become dominant while buoyancy is absent in a weightless environment. Flotation and sedimentation, for example, two common and inexpensive wastewater treatment processes, have no use in a microgravity environment. Some of other common wastewater treatment unit operations have to be modified to address solid/liquid, gas/liquid, and gas/solid separations in the form of additional equipment or/and processes. The difficulty of wastewater treatment associated with microgravity is also extended to the issues of process scale-up and modeling/simulation. All process models, past or current, theoretical or empirical, are subject to validation in space because of microgravity factor. The cost of doing that is so prohibitive that there have been few attempts made to field-test the equipment or design.

The labeling of "microgravity compatible" technologies for water treatment are often based on whether the technology in question is mono-phased and/or whether gravity plays any significant role in driving process performance. This approach is imprecise and sometimes questionable since it ignores the forces such as surface tension that are insignificant under normal gravity but are important in the microgravity environment. The uniqueness of microgravity environment has unsettling implications on various water treatment processes where the bulk fluid is mono-phased but involved solid-fluid interfaces between the fluid and the material of the equipment. One such a

sample is membrane filtration that is being used in ISS for water recovery from hygiene water and humidity condensate. There is no evidence yet to suspect that microgravity has adverse effect on operability of the reverse osmosis unit, however, no one can rule out the possibility of adverse effect of surface tension on membrane filtration at the membrane surface and/or in the boundary layer since surface tension is manifested at interfaces, where concentration polarization and membrane fouling occur.

Recognizing the gap between a basic process of a particular water processing technology and an on-board water treatment module for extended-duration space missions, NASA has devised a systematic assessment scheme, called Technology Readiness Level (TRL), to assess the maturity of a technology for all space-bound technologies, making comparison of sophistication levels among different technologies designed for a particular application of space exploration. For each technology, the larger the number of TRL, the closer the technology is eligible for being used for space missions. The definitions of TRL are can be found in many NASA documents (White, 1995).

MEMBRANE SEPARATION TECHNOLOGIES FOR WATER RECOVERY

Membrane Filtration

Membrane filtration is a technology that utilizes semi-permeable materials in a specific arrangement (configuration) to exclude most organic or inorganic maters in wastewaters based on size or molecular weight while allowing water and, for some variations of membrane filtration systems, small molecules to permeate through. The most common variations of membrane filtration are based on the ability of a membrane to reject materials of certain range of size and/or molecular weight (Liu, 2003). Membrane filtration technology for water treatment has advanced rapidly as demand for potable water worldwide increases. The last two decades have witnessed new reverse osmosis membrane materials that can be operated at ever lower pressure and with increasing salt rejection. Current commercial membranes for reverse osmosis have been claimed to have 97% - 99.5% salt rejection rate (usually obtained from lab-scale membrane units with NaCl solution) and 7 bar operating net driving pressure (Nicolaisen, 2002). Realistically, many reverse osmosis water treatment plants operate at much lower salt rejection rate (about 50 % - 75%) as concentration polarization and fouling take their tolls. The effect of the "evil twin," concentration polarization and fouling, on potable water production from brackish water and seawater is significant and has limited the wide acceptance in the U.S. as a main water treatment technology because of high energy cost and disposal issue related to the concentrated brine from reverse osmosis plants. In long-duration space missions, however, the requirement for water recovery from space wastewater is ideally 100% (not accounting for additional water from foods). This would require either the development of low-pressure and less prone to fouling membranes situated in a membrane module that has minimal concentration polarization in operation or incorporation and optimization of several membrane processes or/and other separation

technologies into water recovery systems. Various membrane separation types in common uses have different possible TRL rating with microfiltration (MF) at the high end, and ultrafiltration (UF) and reverse osmosis (RO) in the middle range of the TRL spectrum.

MF is a pressure-driven membrane filtration process that has a membrane with a pore size typically of 0.01-2 μm and able to retain particles with molecular weights equal or larger than 200 kDa and is used in a number of applications, as either a pre-filtration step or as a process to separate a fluid from a process stream. MF membranes are symmetric with characteristic sponge-like network of interconnecting pores. Cartridge filters are typically composed of microfiltration media. Multi-units of MF have been used in spacecrafts and habitats including MIR and ISS as a pretreatment unit for subsequent water processors such as vapor-compression distillation. MF as pretreatment process could be considered as TRL 8 or 9 technologies.

UF involves the use membrane with a pore size less than 0.1 μm (500 – 100 kDa). Ultrafiltration is not as fine a filtration process as reverse osmosis, but it also does not require the same energy to perform the separation. Applications of ultrafiltration in water recovery for space adventures can mostly likely be found in situations where pretreatment is needed for reducing or removing certain compounds from the feed stream of a reverse osmosis unit in order to alleviate the energy demand and fouling. In UF, the chemical nature of membrane materials has only little effect upon the separation (but not fouling) since ultrafiltration separation like microfiltration is based upon sieving mechanisms thus ultrafiltration is only somewhat dependent upon the charge of the particle and is much more concerned with the size of the particle.

The presence of large quantity of mixed surfactants in space wastewater poses a unique problem for ultrafiltration. On the one hand, micellar-enhanced ultrafitration is widely credited for removing certain particulates and solutes that would be impossible to be removed without the assistance of surfactant aggregates, micelles; on the other hand, surfactant monomers are believed to be responsible for membrane fouling by adhering to the membrane surface. The susceptibility of UF membranes to fouling by proteins has generated interests in fundamental studies in membrane fouling. It is no surprising to see prevalence of fouling in these applications since polymeric UF membranes (polysulfone, for instance) are more or less hydrophobic and proteins have tendency to adhere their hydrophobic cores to the membrane surface thus forming a strong bond – irreversible fouling. In space wastewater treatment, however, membrane fouling is mainly caused by deposition of minerals on the surface and blockage of the pores in addition to adsorption of surfactant monomers, and biofouling. The extent of mineral fouling in relation to surfactant fouling is yet to be determined. Biofouling of UF and other membranes is another important subject that is not adequately studied. In addition to composition of wastewater feed stream, the membrane surface characteristics are the most important factors that determine the extent of biofouling. One recent paper (Vrijenhoek et al., 2001) suggested that the smoothness of the RO membrane surface had a lot to do with whether

biofilms would form on the membrane because, they argued, without crevices or holes or folds, it is difficult to for microorganisms to establish their colonies. This conclusion obviously needs to be further studied. But even the above argument is accurate; one has to wonder if it is also applicable to ultrafiltration since UF membranes contain relatively large-sized pores.

RO, also known as hyperfiltration, is the finest filtration known. This process will allow the removal of particles as small as ions from a solution. Reverse osmosis is used to purify water and remove salts and other impurities in order to improve the color, taste or properties of the fluid. Most reverse osmosis technology uses a process known as cross-flow to allow the membrane to continually clean itself. As some of the fluid passes through the membrane the rest continues downstream, sweeping the rejected species away from the membrane. The process of reverse osmosis requires a driving force to push the fluid through the membrane, and the most common force is pressure from a pump. A reverse osmosis process involves pressures 5-10 times higher than those used in ultrafiltration. As the concentration of the fluid being rejected increases, the driving force required continuing concentrating the fluid increases. Reverse osmosis is capable of rejecting bacteria, salts, sugars, proteins, particles, fats, and other constituents that have a molecular weight of greater than 0.15-0.25 kDa. The separation of ions with reverse osmosis is aided by charged particles. This means that dissolved ions that carry a charge, such as salts, are more likely to be rejected by the membrane than those that are not charged, such as organics. The larger the charge and the larger the particle, the more likely it will be rejected. The transport mechanism of RO is now believed to be the solution diffusion mechanism.

Other Membrane Processes

There are several other membrane processes that involve separate dissolved species from water. Among them are pervaporation and membrane distillation. Pervaporation is defined as a separation process in which a liquid feed mixture is separated by means of partial diffusion-vaporization through a non-porous polymeric membrane while vacuum or a sweep gas is applied to the downstream side of the membrane. Membrane pervaporation has been used in removal of VOC from groundwater and wastewater (for example, Peng et al., 2003; Peng and Liu, 2003ab) and in removal of water from highly-concentrated alcohol (for example, Verkerk et al., 2001). The strength of pervaporation technology lies in its ability to separate trace amount of component(s) from the remaining components in the bulk liquid with less energy requirement and high recovery rate than other separation technologies including other membrane processes. The potential application of pervaporation and its cousin processes such as temperature swing adsorption and thick film absorption in space wastewater treatment is limited to dehydration of the high concentrated brine discharged from an RO unit. It should be noted that the issues such as concentration polarization and membrane fouling also affect pervaporation. Scaling of minerals is a potentially worrisome problem since many pervaporation units operate at 30 – 50 °C to be most effective.

Membrane distillation is another membrane technology that can be used as a part of water recovery system. Membrane distillation (MD) is a type of low temperature, reduced pressure distillation using porous hydrophobic polymer materials. It is a process that separates two aqueous solutions at different temperatures and has been developed for the production of high-purity water, and for the separation of volatile solvents such as acetone and ethanol. MD can achieve higher concentration than RO. In MD, the membrane must be hydrophobic and microporous. The hydrophobic nature of the material prevents the membrane from being wetted by the liquid feed and hence liquid penetration and transport across the membrane is avoided, provided the feed side pressure does not exceed the minimum entry pressure for the pore size distribution of the membrane. The driving force of MD is temperature gradient and the two different temperatures produce two different partial vapor pressures at the solution-membrane interface, which propels consequent penetration of the vapor through the pores of the membrane. The vapor is condensed on the chilled wall by cooling water, producing a distillate. This process usually takes place at atmospheric pressure and temperature that may be much lower than the boiling point of water. Membrane could be used to compliment a hybrid membrane process such as UF-RO unit in space missions. The effect of microgravity on MD operations needs further research.

Membrane Materials

A membrane is undoubted the center of membrane technology. It is no surprise there are many efforts devoted to this area. Many companies have developed and manufactured a variety of membrane materials and configurations for water purification. Current commercial membranes for membrane filtration are mainly made from synthetic polymers and inorganic materials with varied durability under harsh and prolonged operating conditions. Table 2 lists several typical membrane materials and their respective properties (Cheryan, 1998; Peng and Liu, 2003; Cortalezzi et al., 2003):

Table 2. Properties of selective membrane filtration materials

Materials	Maximum temperature (C)	pH range	Solvent resistance
Cellulose acetate	30/65	2-2.75	Low
Fluoropolymer	60	1.5-12	High
Polyamide	60	2-10	High*
Polyethersulfone	80	1.5-9.5	Medium
Polysulfone	80	1.5-12	Medium
PVDF	80	1.5-12	Medium
Polyacrylonitrile	80	1.5-12	High
Alumina oxide	300	0-14	High†
Zirconia oxide	300	0.5-13.5	High†
Iron oxide	300	0-14	High

* susceptible to chlorine attack.

† not recommended for phosphorus.

The first-generation RO membrane materials such as cellulose acetate, though less prone to fouling, has seen its market share declining in desalination and wastewater treatment operations due to newly arrived composite RO membranes. The fragility of this type of membranes has ruled itself out in applications in space missions. Currently, the second-generation RO membranes such as composite membranes made from thin polyamide active layer on top of UF or MF substrates made from polysulfone has been adopted for seawater desalination. However, owing to different components in space wastewater, the applicability of this type of materials remains unclear and needs further long-term studies. Inorganic membranes represented by alumina oxide and zirconia oxide (third-generation) are very resistant to high temperature, organic solvents and acids. However, the processability and cost issues related to inorganic membranes are main road blocks to successful commercial applications. The potential applications of inorganic membranes for space missions are not very encouraging now due to the processability issue. This situation could change with the improvements in processing techniques. A current trend in membrane development is modification of membrane surface characteristics to achieve certain operational goals. Improvement in hydrophilicity by copolymerizing another monomer or functional group is a common technique to reduce membrane fouling by proteins or organic colloids. Another emerging area of membrane materials is nanocomposite membrane. This type of membrane materials involve the use of polymeric materials as substrate embedded with nano-sized property-enhancers such as carbon nanotubes.

Membrane Modules

Spiral wound

In spiral wound modules, a flat membrane envelope or set of envelopes is rolled into a cylinder as shown in Figure 1. The envelope is constructed from two sheets of membrane, sealed on three edges and each sheet is sandwiched between two turbulent-promoting spacers. The open end of envelope is sealed to a perforated tube (the permeate tube) with a proper glue so that the permeate can pass through the perforations. Another spacer is laid on top of the envelope before it is rolled, creating the flow path for the feed liquid. This feed spacer generates turbulence, thereby enhancing the feed side mass transfer rate. The spiral wrapped envelopes and spacers are then wrapped again with tape or glass or net-like sieve before fitting into a pressure vessel. In this way, a reasonable membrane area can be housed in a convenient module, resulting in a very high surface area to volume ratio. One noticeable drawback lies in the permeate path length. A permeating component that enters the permeate envelope farthest from the permeate tube must spiral inward several feet. Depending upon the path length, permeate spacer design, gel layer, and permeate flux, significant permeate side pressure drops can be encountered. The other disadvantage of this module is that it is a poor choice for treating fluids containing particulate matters. This configuration is widely used in desalination plants with RO and generally is well-suited for space wastewater treatment.

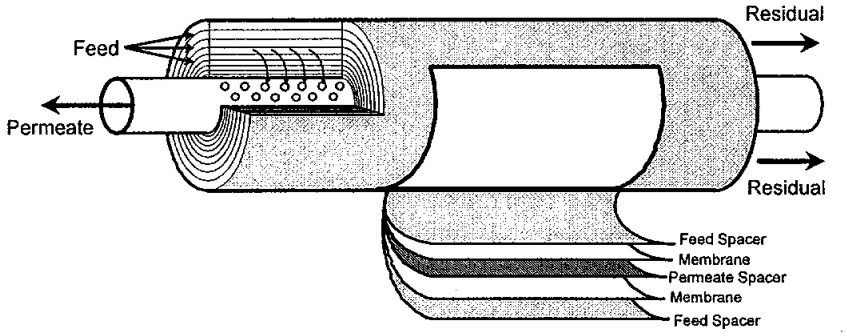


Figure 1. A schematic illustration of a spiral wound module
(Liu, 2003)

Hollow fiber

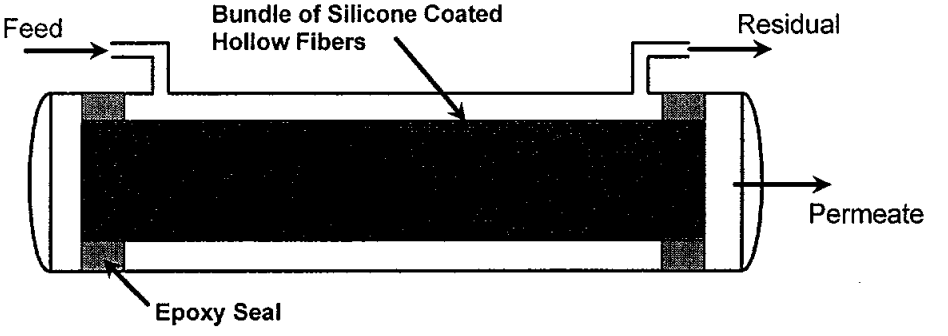


Figure 2. A schematic illustration of a hollow fiber module
(Liu, 2003)

In a hollow fiber configuration, small diameter polymer tubes are bundled together to form a hollow fiber module like a shell and tube heat exchanger (Figure 2). These modules can be configured for liquid flow on the tube side, or lumen side. These

tubes have diameters on the order of 100 microns. As a result, they have a very high surface area to module volume ratio. The drawback is that the liquid flow inside the hollow fibers is normally within the range of laminar flow regime due to its low hydraulic diameter. The consequence of prevalent laminar flows is high mass transfer resistance on the liquid feed side. However, because of laminar flow regime, the modeling of mass transfer in a hollow fiber module is relatively easy and the scale-up behavior is more predictable than that in other modules. One noticeable problem with a hollow fiber module is that a whole unit has to be replaced if failure occurs.

Plate & frame

Plate-and-frame configuration is a migration from filtration technology, and is formed by the layering of flat sheets of membrane between spacers. The feed and permeate channels are isolated from one another using flat membranes and rigid frames (Figure 3). A single plate and frame unit can be used to test different membranes by swapping out the flat sheets of membrane. Further it allows for the use for membrane materials (e.g., inorganic membranes) that cannot be conveniently produced as hollow fibers or spiral wound elements. The disadvantages are that the ratio of membrane area to module volume is low compared to spiral wound or hollow fiber modules, dismounting is time-consuming and labor-intensive, and higher capital costs associated with the frame structures. Although a lot of tests related to membrane characterization or optimization in NASA or elsewhere use variations of this type of membrane modules, it is highly unlikely that any of this type of membrane configuration would end up in a spacecraft or space habitat.

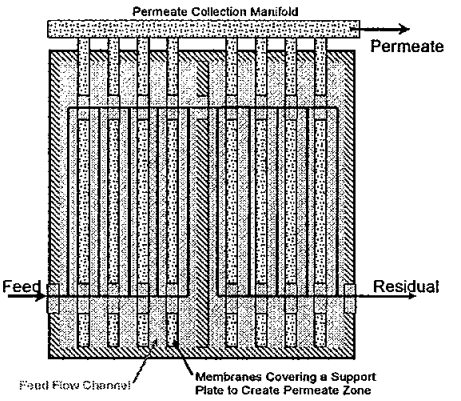


Figure 3. A schematic illustration of a plate & frame module (Liu, 2003)

Tubular

Polymeric tubular membranes are usually made by casting a membrane onto the inside of a pre-formed tube, which is referred to as the substrate tube. The tube is generally made from one or two piles of non-woven fabric such as polyester or polypropylene. The diameters of tubes range from 5-25 mm (Figure 4). The advantage of the tubular membrane is its mechanical strength if the membrane is supported by porous stainless steel or plastic tubes. Tubular arrangements often provide good control of flow to the operators and are easy to clean. Additionally it is the only membrane format for inorganic membranes, particularly ceramics. The disadvantage of this type of modules is mainly higher costs in investment and operation. The arrangement of tubular membranes in a housing vessel is similar to that of hollow fiber element. Tubular membranes sometimes are arranged helically to enhance mass transfer by creating a second flow (Dean vortex) inside the substrate tube (Moulin et al., 1999).

Other Configurations

Several membrane configurations were developed in response to concentration polarization issue in water treatment. The main thrust of these membrane unit designs is to induce high shear on the membrane surfaces (Murase et al., 1991; Engler and Wiesner, 2000; Al-Akoum et al., 2002; Lee and Lueptow, 2002). Vane and Alvarez (2002) used a VSEP (Figure 5) to improve mass transfer at the interface for pervaporation of VOCs.

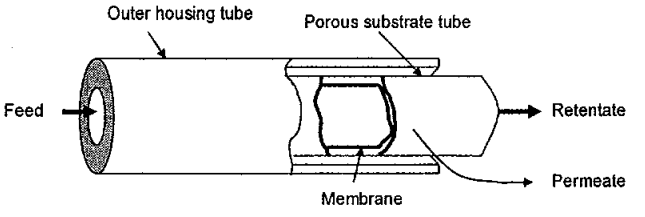


Figure 4. A schematic illustration of a tubular module (Liu, 2003)

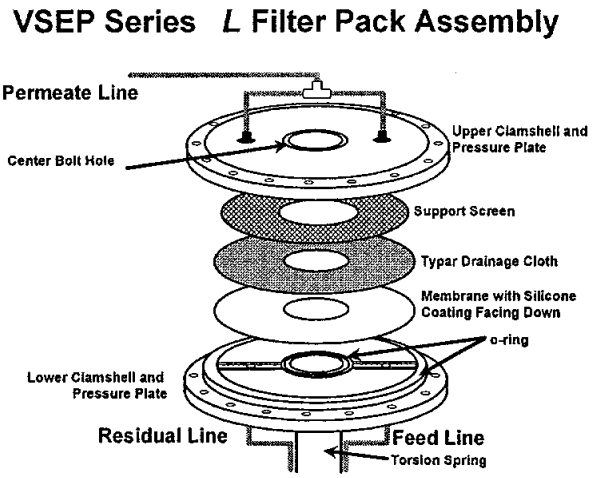


Figure 5. A Vibrating Membrane Module (Courtesy of New Logic, Inc.)

Concentration Polarization

Concentration polarization is an adverse process phenomenon that affects almost all membrane systems and types. In membrane filtration, it builds up retained components to such an extent that the retained components begin to back-diffuse to the bulk. Concentration polarization is also partially responsible for causing membrane fouling because of certain sparsely soluble minerals often reach saturation concentration at the membrane surface and precipitate on the membrane forming a layer of irreversible bonded minerals. Concentration polarization, however, is considered reversible and can often be alleviated by introduce mixing-promoting spacers and increasing flow rate in a cross-flow membrane configuration. Some innovative design of membrane systems such as rotating RO (Lee and Lueptow, 2000), vibrating membrane module (Vane and Alvarez, 2002; Al-Akoum et al., 2002) have been demonstrated to be effective in the systems involved. Beyond increasing shear to counter concentration polarization, there are several other possibilities utilizing other external forces/fields to reduce concentration gradients and enhance trans-membrane mass transfer. One of such forces utilized is electrical force. Huotari and others were able to increase the limiting flux of a cross-flow ultrafiltration unit dealing with oily wastewater by applying electric field (Huotari et al., 1999). However, direct application of the set-up from the above-mentioned authors to space wastewater is very difficult since there are too many components including surfactants, ions, microorganisms, and urea with diverse electrophoretic motilities. The other possibilities of using non-shear forces are acoustic separation or ultrasound in the boundary layer (Athaide and Govind, 1987) and the use of magnetic force. These new areas are promising and need to be further studied.

Membrane Fouling

Fouling is a phenomenon of irreversible loss of membrane permeability leading to reduction in permeation flux. Fouling is caused by adsorption of feed components, clogging of the pores (UF and MF), chemical bonding reaction between the solutes and the membrane, gel formation, and microbial growth and biofilm formation (Koltuniewicz and Noworyta, 1994). The major factors that influence membrane fouling are the hydrodynamics of the process, and the physicochemical properties of the membrane and the feed solution (Huisman et al., 2000). Membrane fouling is a direct result of interaction between solutes in the feed stream and the membrane. As such, the properties of the membrane and solutes in the feed stream as well as operating parameters have strong bearing on fouling. For a UF/MF membrane, the hydrophilicity, surface topography, charge on the membrane, and pore size contribute individually or in several of combinations, to the fouling while organic colloids, pH, soluble minerals, and surfactants appear to be the contributing factors from solutes in feed streams (Cheryan, 1998). As alluded previously, proper selection or modification of membrane surface, pretreatment of membranes with certain surfactants and enzymes, and use of biocides can reduce fouling. The measures used to fight concentration polarization can also mitigate fouling since concentration polarization is partially responsible for fouling. Temperature also affects the extent of fouling (Goosen et al., 2002).

In addition to hydrophilicity, membrane surface topography and pore size also affect the interaction between foulant molecules and the membrane thus membrane fouling. Membrane surface morphology can influence the membrane fouling in two ways: the rough surface tends to trap macromolecules and the surface area of a rough membrane is larger than that of a smooth membrane, which increases likelihood and number of protein adsorption sites. Additionally, in a cross-flow mode operation, a foulant molecule that deposits on a rough membrane surface is less likely to tear off from the surface. Pore size role in membrane fouling seems to be obvious. However, large pore size only gives initial high flux. Once foulants deposit onto the surface of the pore and aggregates are formed in the pore, the pore becomes constricted and lower flux ensues. If pore size is in the same magnitude as size of the molecules, the chance of the molecules clogging some of pores increases. Cheryan (1998) suggests a ratio of pore size to particle size of 1:10.

CONCLUDING REMARKS

Membrane separation technologies are the logic choice for space water recovery. Membrane filtration is a physical process that requires no additional chemicals and less energy than a typical thermal process, and is compact, modular, and perceivably insensitive to microgravity. Great leap has been made in many areas of membrane filtration technology ranging from materials to new module/unit designs. A lot of this advancement will ultimately be migrated to space wastewater treatment, resulting in better and reliable space water recovery systems. The most challenging task that NASA

scientists and engineers face is the difficulty of quickly bringing the existing technology to TRL 7 or higher. The lack of experimental data regarding long-term membrane performance under microgravity environment is the major obstacle for this implementation. Additional critical areas that need further studies include biofouling mechanism and removal strategies, fouling by mixed surfactants, novel fouling resistant membranes and innovative countermeasures to concentration polarization.

The future of membrane technologies for space missions will be no doubt very bright and it is highly likely there will be a membrane subsystem in the ECLSS of a spacecraft or space habitat. The water recovery systems for various mission scenarios need to be tailored and fully integrated into the ECLSS of the space living environment. The decision of which water treatment component should be included in a water recovery system ought to be based on a variety of important factors including energy consumption and energy sources, equivalent system mass, reliability, and simplicity in operation and maintenance.

REFERENCES

Al-akum, O., Ding, L.H., and Jaffrin, M.Y. (2002). Microfiltration and ultrafiltration of UHT skim milk with a vibrating membrane module. Separ. Purifi. Technol. 28(3), 219-234.

Althaide, A. and Govind, R. (1987). The effect of fouling on the stability of membrane bioreactors. Chem. Engr. Sci. 42(1), 172-175.

Cheryan, M. (1998). Ultrafiltration and Microfiltration Handbook. Technomic Publishing Company, Inc., Lancaster, Pennsylvania, USA.

Cortalezzi, M.M., Rose, J., Wells, G.F., Bottero, J.-Y., Brron, A. R., and Wiesner, M. R. (2003). Ceramic membranes derived from ferroxane nanoparticles: a new route for the fabrication of iron oxide ultrafiltration membranes. J. Membr. Sci. 227, 207-217.

Engler, J. and Wiesner, M.R. (2000). Particle fouling of a rotating membrane disk. Water Res. 34(2), 557-565.

Goosen, M.F.A., Sablani, S.S., Al-Maskari, S.S., Al-Belushi, R.H., and Wilf, M. (2002). Effect of feed temperature on permeate flux and mass transfer coefficient in spiral wound reverse osmosis systems. Desalination 144, 367-372.

Huotari, H., Huisman, I.H., and G. Trägårdh (1999). Electrically enhanced cross-flow membrane filtration of oily wastewater using the membrane as a cathode. J.Membr. Sci. 156, 49-60.

Koltuniewicz, A. and Noworyta, A. (1994). Dynamic properties of ultrafiltration systems in light of the surface renewal theory. Ind. Engr.Chem. Res. 33, 1771-1779.

Lee, S. and Lueptow, R. M. (2000). Toward a reverse osmosis membrane system for recycling space mission wastewater. Life Supp. & Biosph. Sci. 7, 251-261.

Lee, S. and Lueptow, R.M. (2002). Experimental verification of a model for rotating reverse osmosis. Desalination 146, 353-359.

Liu, S. X. (2003). Design of membrane systems. In Encyclopedia of Agricultural, Food and Biological Engineering, ed., D. R. Heldman, pp. 614-620, Mercel Dekker, New York.

Moulin, P.; Manno, P.; Rouch, J.C.; Serra, C.; Clifton, M.J.; Aptel, P. (1999). Flux improvement by dean vortices: ultrafiltration of colloidal suspensions and macromolecular solutions. J. Membr. Sci. 156, 109-130.

Murase, T., Iritani, E., Chidphong, P., Kano, K., Atsumi, K., and Shirato, M. (1991). High-speed microfiltration using a rotating cylindrical ceramic membrane. Ind. Chem. Engr. 31(2), 370-378.

Nicolaisen, B. (2002). Developments in membrane technology for water treatment. Desalination 153, 355-360.

Peng, M., L.M. Vane, and S.X. Liu, (2003). Recent advances in voc removal from water by pervaporation. J. Hazard. Mater. 98 (1-3), 69-90.

Peng, M. and S.X. Liu (2003a).VOC Removal from contaminated groundwater through membrane pervaporation. part I: water-1,1,1- trichloroethane system," J. Environ. Sci. 15(6), 815-820.

Peng, M. and S.X. Liu(2003b). VOC removal from contaminated groundwater through membrane pervaporation. part II: 1,1,1- trichloroethane – surfactant solution system," J. Environ. Sci. 15(6), 821-927.

Vane, L.M. and Alvarez, F.R. (2002). Full-scale vibrating pervaporation membrane unit: VOC removal from water and surfactant solutions. J. Membr. Sci. 202, 177-193.

Verkerk, A.W., van Male, P., Vorstman, M.A.G. and Keurentjes, J.T. F. (2001). Properties of high flux ceramic pervaporation membranes for dehydration of alcohol/water mixtures. Separa. Purifi. Technol. 22-23, 689-695.

http://atdo.jsc.nasa.gov/services/library/documents/TRLs.pdf

Solar Modulation of Inner Trapped Belt Radiation Flux as a Function of Atmospheric Density

Final Report
NASA / ASEE Summer NASA Faculty Fellowship Program – 2004
Johnson Space Center

Faculty Fellow:	M. A. K. Lodhi
Academic Rank:	Professor of Physics
University & Department:	Texas Tech University Department of Physics Lubbock, TX 79409
NASA / JSC Directorate:	Space and Life Science
Division or Office:	Astrmaterials Research and Exploration Science (ARES)
Branch Office:	Astromaterials Research
JSC Colleague:	Thomas L. Wilson
Date Submitted:	July 29, 2004
Contact Number:	NAG9-1526 and NNJ04JF93A

Simplified Solar Modulation Model of Inner Trapped Belt Proton Flux as a Function of Atmospheric Density

ABSTRACT

No simple algorithm seems to exist for calculating proton fluxes and lifetimes in the Earth's inner, trapped radiation belt throughout the solar cycle. Most models of the inner trapped belt in use depend upon AP8 which only describes the radiation environment *at* solar maximum and solar minimum in Cycle 20. One exception is NOAAPRO which incorporates flight data from the TIROS/NOAA polar orbiting spacecraft. The present study discloses yet another, simple formulation for approximating proton fluxes at any time in a given solar cycle, in particular *between* solar maximum and solar minimum. It is derived from AP8 using a regression algorithm technique from nuclear physics. From flux and its time integral fluence, one can then approximate dose rate and its time integral dose. It has already been published in this journal that the absorbed dose rate, D, in the trapped belts exhibits a power law relationship, $D = A\rho^{-n}$, where A is a constant, ρ is the atmospheric density, and the index n is weakly dependent upon shielding. However, that method does not work for flux and fluence. Instead, we extend this idea by showing that the power law approximation for flux J is actually bivariant in energy E as well as density ρ. The resulting relation is $J(E,\rho) \sim \sum A(E^n)\rho^{-n}$, with A itself a power law in E. This provides another method for calculating approximate proton flux and lifetime at any time in the solar cycle. These in turn can be used to predict the associated dose and dose rate.

1. Introduction

Studies of space radiation and its effects are concerned with the impact of charged species on the functionality and lifetime of human beings as well as scientific instrumentation and advanced electronic systems in space. Two aspects of the near-Earth space environment are very relevant, particularly in the thermosphere (85km<h<500km) where Shuttle and International Space Station (ISS) orbits occur. One is the existence of energetic proton and electron populations trapped by the Earth's magnetic field in "Van Allen" belts (E.g., Schulz and Lanzerotti, 1974; Spejeldvik and Rothwell, 1985). The other is the realization (Jacchia, 1960, 1961) that the properties of the upper atmosphere of the Earth are strongly coupled to solar activity, in particular atmospheric density and temperature. Throughout the course of the solar cycle, the Earth's atmospheric neutrals expand and contract the thermosphere in response to the behavior of the Sun. Clearly, the density in Jacchia's concept of a dynamic atmosphere couples to the charged-belt species as these undergo multiple scattering off the neutrals. That in turn reduces their lifetime in the belts (Blanchard and Hess, 1964; Cornwell et al., 1965; Dragt, 1966; Ray, 1966; Kern, 1994; Pfitzer, 1989; Watts et al., 1989).

Therefore, it becomes necessary to understand how atmospheric density *per se* couples to charged-belt population levels as a function of solar activity. This is the simplified goal of the present investigation.

Pfitzer (1989; 1990) has succeeded in developing a reasonable parametric method for estimating dose in the thermosphere from atmospheric density. However, the method does not work for flux. Inspired by that preliminary investigation, Badhwar and his colleagues (1999, 1997, 1996a,b; Golightly et al., 1996) have examined flight data for a correlation between dose and atmospheric density. They have extensively studied and analyzed the low-Earth radiation and time lag of the twenty-two year solar modulation of the trapped proton radiation exposure inside the Space Shuttle. They have shown that the daily trapped-particle dose rate is an approximate power law function of daily atmospheric density, thus supporting the Pfitzer model and method. Their further analysis of the trapped absorbed dose rate, D, at six fixed locations in the habitable volume of the Shuttle exhibits a power law relationship, $D = A\rho^{-n}$, where ρ is the atmospheric density. The index, n, is weakly dependent on the shielding, decreasing as the average shielding increases (Badhwar, 1999).

This present study further examines the AP8 proton flux question and its relationship to atmospheric density. It enhances the previous Pfitzer and Badhwar density analyses by developing a dynamic trapped-belt proton radiation algorithm that is applicable to the ISS and other space flights in the Earth's thermosphere throughout the solar cycle. Although only a very limited range of energies is considered, the method addresses several of the shortcomings and over-simplifications in that earlier work.

2. Analysis

The limitations with the original NASA trapped-belt models (Sawyer and Vette, 1976; Bilitza, 1987) known as AP8 and AE8 have been thoroughly discussed (Watts et al., 1989; Pfitzer, 1989, 1990; Badhwar, 1999). The AP8 model was constructed from satellite data in solar cycle 20, a small one compared to more recent events. AP8MIN derives from the epoch of 1964, and AP8MAX from that of 1970. The solar radio flux at 10.7 cm, $F_{10.7}$, is 150 for AP8MAX and 70 for AP8MIN. These baseline values will be adopted here.

One other promising approach to overcoming the AP8 model limitations has already been produced. It involved the development of a new computer technique known as NOAAPRO (Huston and Pfitzer, 1998a,b). This method has since been adapted by Singleterry *et al.* (2004) to enhance the out-of-date AP8 and AE8 models at Shuttle and ISS altitudes using the computer program SIREST.

Since the original AP8 model is readily available elsewhere (Heynderickx et al., 2004), it will be used to modify the atmospheric density method of Pfitzer and Badhwar by producing a bivariant energy-density algorithm and then compare the result with the NOAAPRO-enhanced AP8 model of Singleterry *et al.* At the outset, AP8 is adopted here primarily in order to be consistent with the Pfitzer method. The analysis can be applied to other simulation methods such as NOAAPRO and SIREST. Only the omnidirectional fluxes are studied in this analysis, noting that the anisotropic nature of these has been discussed by Watts et al. (1989).

Upon examining the proton flux data from the AP8 model program, and in view of the overall problem as studied for more than 40 years (Pfitzer, 1989, 1990), several observations can be summarized.

1. The atmospheric density decreases rapidly as the altitude increases.
2. Both integrated and differential flux for protons increases as the altitude increases (or density decreases) for both solar maximum and minimum cycles.
3. The fluxes in general are smaller at solar max than at solar min, at least for low altitudes and low energies.
4. However, the difference in the proton flux is wider at low energy than for the higher-energy protons at the same altitude.
5. The flux decreases as the energy increases.

These observations prompt the idea that the proton flux J can be expressed, empirically at least, as a function of two variables, density (altitude) and energy, namely $J(E,\rho)$ or $J(E,h)$.

Solar Modulation of Atmospheric Density

The altitude and density relationship has a long history (Jacchia, 1960, 1961; Harris and Priester, 1962, 1963). The form to be used here has been described recently

by several authors (Pfitzer, 1989, 1990; Watts et al., 1989). A simple parameterization of the US Air Force Model made by Pfitzer (1989, 1990) is

$$\rho = \rho_o \exp\{-(h-120) / [M(h-103)^{1/2}]\}, \quad (1)$$

where the solar-cycle modulation term M is expressed as

$$M = 0.99 + 0.518 \{ (F_{10.7} + F_{bar}) / 110 \}^{1/2} , \quad (2a)$$

$$F = F_{10.7} + F_{bar} . \quad (2b)$$

In (1) ρ denotes the atmospheric density, ρ_o is assumed to be $\rho_o = 2.7 \times 10^{-11}$ g/cm^3, h is the altitude in km, $F_{10.7}$ is the instantaneous value of the solar radio flux at 10.7 cm, and F_{bar} is the weighted value of $F_{10.7}$ for averaging, such as three prior solar rotations.

The density in (1) is a multi-variant function of h and $F_{10.7}$. Similarly, the AP8 proton flux J is a multi-variant function of h and energy E. The problem at hand, then, is to generate the multi-dimensional surface of J as a function of E and h or ρ. By selecting an altitude h and emulating the solar cycle with a solar radio flux $F_{10.7}$, the modulated proton flux J follows as a function of energy E. Dynamically speaking, furthermore, all of these variables are functions of time t.

A "carpet" plot (in the sense of Pfitzer, 1989, 1990) is merely a projection of these surfaces onto a two-dimensional graph. This can be obtained for the solar-cycle terms altitude $h = f(\rho)$ and flux $F_{10.7} = g(\rho)$ as a function of atmospheric density ρ in (1) and (2), by taking the inverse of (1) for constant surfaces of $F_{10.7}$ and h respectively. The result is provided in Fig. 1. As pointed out by Pfitzer, the trapped protons have a very slow response time to dynamic changes in atmospheric density $\rho(t)$. Therefore, the values of $F_{10.7}$ and F_{bar} are assumed identical, whereby $F = F_{10.7} + F_{bar} = 2 F_{10.7}$ in (2b). As stated earlier, the values of $F_{10.7}$ used for solar max and solar min in the following regression analysis become 150 and 70 respectively.

Regression Algorithm

Several methods and approaches are available to generate a semi-empirical formula for proton flux J as a bivariant function of density ρ (altitude h) and energy E. The method adopted here is taken from theoretical nuclear physics (Lodhi and Waak, 1975, 1976) based upon a polynomial regression analysis. It is used to determine the functional relationship between fluxes at solar maximum and minimum. Since the proton lifetime τ is determined by energy losses primarily due to multiple Coulomb scattering from charged constituents in the upper thermosphere, as well as some neutral scattering, it is a function of time t, proton energy E, and atmospheric density ρ or altitude h. That is, $\tau = \tau(E, \rho^{-1}, t)$ or $\tau = \tau(E, h, t)$ since the atmosphere expands and contracts at different times during the solar cycle.

Utilizing the regression analysis technique, one keeps the regression coefficient greater than 90%. A ratio J_{max}/J_{min} or J_{min}/J_{max} is generated for differential proton fluxes at solar max and solar min as a function of density (and 1/density) for energies between 30 and 300 MeV from AP8 model data within 300 to 600 km. Next this ratio is fitted to some polynomial ranging from linear to fifth-order in ρ and $1/\rho$ for proton energies of 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 125, 150, 175, 200, and 300 MeV. Note that this ratio is non-linear in ρ or $1/\rho$. One also finds that polynomials of higher-order than second result in a better fit for given energy than the second-order. However, a common expression for the entire energy range gets worse than that of the second-order. This observation limits the method to be confined to a second-order expression for the flux ratio as a function of ρ_{max} or $\rho^{-1}{}_{max}$ for all energies. The regression analysis works well within the energy and altitude range adopted. For other ranges of approximation one has to be careful and do the analysis again, piecewise fitting the entire dynamical range.

Density Dependence

First the flux ratio is generated as a function of ρ in the quadratic form. The coefficients of ρ vary while progressing from one energy to the next. The flux ratio can then be written in the following fashion:

$$J_{min}/J_{max} = a\rho^2 + b\rho + c \quad , \tag{3}$$

where a, b, and c are proton energy coefficients. Naively, one might hope that these are constant coefficients. However, one discovers that a, b, and c are functions of energy E for the 15 different energy values chosen.

The next step is to find a common expression for this ratio for all energy values. That is achieved by obtaining a functional relation for the coefficients in (3) as a function of energy, again by the method of regression. One obtains four relations for ρ ranging from linear to fourth-order in energy. One also finds that expressions for coefficients a, b, and c cannot be of the same polynomial order in reproducing the flux ratio. The best fits found for two different energy ranges are as follows:

$$J_{min}/J_{max} = (a_2E^2 + a_1E + a_0)_{le}\rho^2 + (b_4E^4 + b_3E^3 + b_2E^2 + b_1E + b_0)_{le}\rho$$

$$+ (c_4E^4 + c_3E^3 + c_2E^2 + c_1E + c_0)_{le} \quad , \quad 30 < E \leq 60\ MeV \tag{4}$$

$$J_{min}/J_{max} = (a_1E + a_0)_{he}\,\rho^2 + (b_1E + b_0)_{he}\,\rho$$

$$+ (c_2E^2 + c_1E + c_0)_{he} \quad , \quad 70 \leq E < 300\ MeV \tag{5}$$

The coefficients within parentheses are different for *le* (low energy) and *he* (high energy). The actual coefficients and a check for the accuracy are given below under *Results*.

Inverse Density Dependence

In contrast to (3), the inverse algorithm can be derived. It is known that the atmospheric densities decrease as the altitude increases or the reciprocal of the density increases as the altitude increases. The flux variation with respect to the inverse of the density must convey a direct relation to the variation of the altitude. One must therefore search for a similar expression giving the flux ratio as a function of the inverse density. After several trials the best-fitted function thus obtained is given in the form:

$$J_{max}/J_{min} = (a_4 E^4 + a_3 E^3 + a_2 E^2 + a_1 E + a_0)\rho^{-2} + (b_3 E^3 + b_2 E^2 + b_1 E + b_0)\rho^{-1} + (c_2 E^2 + c_1 E + c_0) \quad (6)$$

for all energies E under consideration. This expression is further approximated by writing the coefficients in the exponential form, thus yielding:

$$J_{max}/J_{min} = Ae^{\alpha E}\rho^{-2} + Be^{\beta E}\rho^{-1} + Ce^{\gamma E} \quad (7)$$

for all proton energy ranges from 30 to 300 MeV.

Results at Solar Max

The two algorithms (3) and (7) are now compared at solar maximum. The resultant semi-empirical formula for the flux ratio in (3) as a function of ρ (in units of 10^{-15} g/cm^3) is given by:

$$(J_{min}/J_{max})_{le} = (-3x10^{-8}E^2 + 1x10^{-5}E - 8.0x10^{-4})\rho^2 + (7x10^{-10}E^4 - 5x10^{-7}E^3 + 1x10^{-4}E^2 - 1.4x10^{-2}E + 0.695)\rho + (-2x10^{-11}E^4 - 5x10^{-8}E^3 + 4x10^{-5}E^2 - 8.6x10^{-3}E + 1.897)\ , \quad 30 \le E \le 60\ MeV \quad (8)$$

$$(J_{min}/J_{max})_{he} = (2x10^{-6}E + 1x10^{-4})\rho^2 + (-7x10^{-4}E + 0.181)\rho + (6x10^{-6}E^2 - 3.7x10^{-3}E + 1.66)\ . \quad 70 \le E \le 300\ MeV \quad (9)$$

To check formulas (8) and (9) for J_{min}, an example at 400 km and proton energy of 100 MeV is taken, and the results summarized in Table 1. Let us define

$$f(\rho) = a\rho^2 + b\rho + c \tag{10}$$

on the right-hand-side of (3), (8), and (9).

At altitude 400 km (Heynderickx et al., 2004), the AP8 model in SPENVIS gives

$$\rho_{max} = 3.8x10^{-15} g/cm^3$$

$$\rho_{min} = 9.57x10^{-16} g/cm^3$$

$$J_{max}^{(100MeV)} = 2.79x10^{-2}\ cm^{-2}s^{-1}MeV^{-1}\ \ (SPENVIS) \quad .$$

For this density at solar max one obtains $f(\rho_{max}) = f(3.8)$ in (10) and $J_{min}/J_{max} = 1.78$ in (3). It then follows from (3) that

$$J_{min}^{(Algorithm)} = J_{max}^{(SPENVIS)} f(\rho_{max}) = (2.79x10^{-2}\ cm^{-2}s^{-1}MeV^{-1})(1.78)$$

$$= 4.9554x10^{-2}\ cm^{-2}s^{-1}MeV^{-1} \ . \tag{11}$$

On the other hand, AP8 (Heynderickx et al., 2004) gives

$$J_{min}^{(SPENVIS)} = 5.151x10^{-2}\ cm^{-2}s^{-1}MeV^{-1} \ . \tag{12}$$

Comparison of (11) and (12) yields a difference of $0.183x10^{-2}\ cm^{-2}s^{-1}MeV^{-1}$ with an error of *3.5%*. These are summarized in Table 1.

Next, the resultant semi-empirical formula for the flux ratio J_{max}/J_{min} in (7) as a function of $1/\rho$ (in units of $10^{+15}\ cm^3/g$) is determined by regression analysis to have the form:

$$J_{max}/J_{min} = -0.0241e^{0.0007E}\rho^{-2} + 0.1966e^{-0.0007E}\rho^{-1}$$

$$+ \ 0.3208e^{+0.0032E} \tag{13}$$

for all energies between 30 and 300 MeV.

Following the same procedure used in (10) and (11), we can define

$$g(\rho^{-1}) = Ae^{\alpha E}\rho^{-2} + Be^{\beta E}\rho^{-1} + Ce^{\gamma E} \tag{14}$$

for the right-hand-side of (7), (8), (9), and (13). One determines that $g(\rho^{-1}_{max}) = 0.4912$ in (14). The resultant J_{min}, the difference from SPENVIS, and the error are summarized in Table 1.

For further comparison, from expression (13) for $J_{max}(cm^{-1}s^{-1}MeV^{-1})$ a differential flux is calculated and contrasted with the AP8 data in Fig. 2, derived from SPENVIS (Heynderickx et al., 2004) for an ISS orbit of 350-478 km altitude and inclination 51.65°. Fig. 2 is a two-dimensional projection of the three-dimensional surface $J(h,E)$ at various selected altitude h. The NOAAPRO results in SIREST are shown in Fig. 3 and Fig. 4 along with the algorithm at 400km and 500km altitudes, for solar max with $F_{10.7} = 150$ in the algorithm.

Results Midway between Solar Max and Solar Min

Method 1. By varying the solar-cycle modulation parameter M in (2), one obtains a different atmospheric density model in (1). This can be accomplished by changing $F_{10.7}$ and F_{bar} whereby a different value of density ρ is obtained, either from (1) or from the carpet plot in Fig. 1. Upon insertion of the new value of density ρ, a proton differential flux follows from (3) and (7). The baseline regression algorithms (3) and (7) assume $F_{10.7} = 70$ and $F_{10.7} = 150$ in Fig. 3 and Fig. 4 for solar min and max respectively. In order to determine the proton differential flux mid-way through this same adopted cycle, the solar flux becomes $F_{10.7} = 110$ whereby $F = 220$ in (2b). The resulting proton differential spectrum is shown in Fig. 5 for 400km in Fig. 6 for 500km.

Method 2. Given a proton flux J at either solar maximum or minimum, such as the algorithm in (3) and (7), then an interim flux is determinable as a linear time-interpolation,

$$J(E,h,\tau) \sim J_{max}(E,h,\tau)(1-\Gamma) \;+\; \Gamma\, J_{min}(E,h,\tau)\;, \tag{15}$$

or alternatively,

$$J(E,h,\tau) \sim J_{min}(E,h,\tau)(1-\Gamma) \;+\; \Gamma\, J_{max}(E,h,\tau)\;, \tag{16}$$

where $\Gamma(E,\rho,\tau)$ is the dimensionless fraction

$$\Gamma(E,\rho,\tau) = \frac{\tau - \tau_{max}(E)}{\tau_{min}(E) - \tau_{max}(E)}\quad . \tag{17}$$

The lifetime τ is assumed to be limited to one solar cycle or 11 years.

To calculate the desired intermediate proton flux $J(E,h,\tau)$ at a time *between* solar maximum and solar minimum using Method 2, the right-hand-side of (17) represents the interpolation fraction Γ of the solar cycle since last solar minimum. Then either of (15) and (16) gives the interim flux in this approximation. Substituting (3) and (7) into (15) and (16) respectively, one has

$$J(E,h,\tau) \sim J_{max}(E,h,\tau)\left[(1-\Gamma) + \Gamma(a\rho^2 + b\rho + c)\right], \quad (18)$$

$$J(E,h,\tau) \sim J_{min}(E,h,\tau)\left[(1-\Gamma) + \Gamma(Ae^{\alpha E}\rho^{-2} + Be^{\beta E}\rho^{-1} + Ce^{\gamma E})\right]. \quad (19)$$

The various coefficients in (18) and (19) are given in (4)-(5) and (8)-(9). Further study is planned to conduct an error analysis between Method 1 and Method 2.

3. Conclusions

The proton differential flux has been expressed, empirically, as a bivariant function of density (altitude) and energy, broken into two ranges of proton energy, *viz.*, 30 to 60 MeV and 70 to 300 MeV. The corresponding expression in terms of inverse density is relatively compact and works for the entire range of proton energy, 30 to 300 MeV. From this analysis it is observed that the proton differential flux has a nonlinear dependence on energy and density (altitude). The flux ratio has been expressed in a semi-empirical form given by (3) and (7). It has been compared with AP8 model data as shown in Fig. 2 for Shuttle and ISS altitudes of current interest. An additional comparison with NOAAPRO is given in Fig. 3 and Fig. 4. An illustration of the algorithm mid-way through the adopted solar cycle is produced in Fig. 5 and Fig. 6. Finally, the algorithm provides a simple means for calculating the trapped-belt proton flux when the $F_{10.7}$ solar modulation flux is 200 or greater. The analysis thus avails itself to other more prominent solar sycles where AP8 is not valid. However, a thorough error analysis will be required in the future in order to determine the general limitations of this regression-analysis algorithm. As a concluding remark, the selection of solar flux $F_{10.7}$ is a matter of convention due to its known correlation with sunspot number. The physics of the thermosphere is not completely understood and there is current interest in the extreme ultraviolet parameter $E_{10.7}$. Should another modulation factor be found, the regression analysis presented here can be modified to accommodate it.

Acknowledgement

During the NASA Faculty Fellowship Program (NFFP) at the Johnson Space Center, Houston, Texas, in 2001 the initial phase of this investigation began with Dr. Gauttam D. Badhwar (deceased). We thank Dr. Stuart Huston for valuable comments regarding NOAAPRO. We are also grateful to Dr. Karl Pfitzer for providing details of his model.

References

Badhwar, G.D., 1999. Radiation dose rates in Space Shuttle as a function of

atmospheric density, Rad. Meas. 30, 401-414.
Badhwar, G.D., Shurshakov, V.A., and V.V. Tsetlin, V.V., 1997. Solar Modulation of Dose Rate Onboard the Mir Station, IEEE Trans. on Nuclear Science 44, no. 6, 2529-2541.
Badhwar, G.D., Konradi, A., Atwell, W., Golightly, M.J., Cucinotta, F.A., Wilson, J.W., Petrov, V.M., Tchernykh, I.V., Shurshakov, V.A., and Labokov, A.P., 1996a. Measurements of the linear energy transfer spectra on the Mir orbital station and comparison with radiation transport models, Rad. Meas. 26, 147-158.
Badhwar, G.D., Golightly, M.J., Konradi, A., Atwell, W., Kern, J.W., Cash, B., Benton, E.V., Frank, A.L., Sanner, D., Keegan, R.P., Frigo, L.A., Petrov, V.M., Tchernykh, I.V., Akatov, Y.A., Shurshakov, V.A., Arkhangelsky, V.V., Kushin, V.V., Klychin, N.A., Vana, N., and Schoner, W., 1996b. In-flight radiation measurements on STS-60, Rad. Meas. 26, 17-34.
Bilitza, D., 1987. Models of the trapped particle fluxes AE-8 (electrons) and AP8 (protons) in inner and outer radiation belts, Nat. Spa. Sci. Data Center, Goddard Space Flight Center, NSSDC, October.
Blanchard, R.C., and Hess, W.N., 1964. Solar cycle changes in inner-zone protons, J. Geophys. Res. 69, 3927-3938.
Cornwell, J.M., Simms, A.R., and White, R.S., 1965. Atmospheric density experienced by radiation belt protons, J. Geophys. Res. 70, 3099-3111.
Dragt, A.J., 1966. Solar cycle modulation of the radiation belt proton flux, J. Geophys. Res. 76, 2313-2344.
Golightly, M.J., Badhwar, G.D., Dunlap, M.J., Patel, S.H., 1996. Solar-Cycle modulation of the trapped proton radiation exposure inside the Space Shuttle. In: Balasubramaniam, K.S., Keil, S.L., and R.N. Smartt, R.N. (Eds.), Solar Drivers of the Interplanetary and Terrestrial Disturbances, Vol. 95, Astronomical Society of the Pacific Conference Series, 505-517.
Harris, I., and Priester, W., 1962. Theoretical models for the solar-cycle variation of the upper atmosphere, J. Geophys. Res. 67, 4585-4591.
Harris, I., and Priester, W., 1963. Relation between theoretical and observational models of the upper atmosphere, J. Geophys. Res. 68, 5891.
Heynderickx, D., *et al.*, 2004. SPace ENVironment Information System (SPENVIS), European Space Agency (ESA), http://www.spenvis.oma.be/spenvis/ .
Huston, S.L., and Pfitzer, K.A., 1998a. A new model for the low altitude trapped proton environment, IEEE Trans. On Nuclear Science 45, no. 6, 2972-2978.
Huston, S.L., and Pfitzer, K.A., 1998b. Space environment effects: Low-altitude trapped radiation models, Marshall Space Flight Center, NASA/CR-1998-208593, 63 pp., available at http://see.msfc.nasa.gov/ire/irepub.html.
Jacchia, L.G., 1960. A variable atmospheric-density model from satellite accelerations, J. Geophys. Res. 65, 2775-2782.
Jacchia, L.G., 1961. A working model for the upper atmosphere, Nature 192, 1147-1148.
Kern, J.W., 1994. A note on vector flux models for radiation dose calculations, Rad. Meas. 23, 43-48.

Lodhi, M.A.K., and Waak, B.T., 1975. Solution of bound state problems in nuclear shell models with momentum-dependent potentials, Comm. Phys. Comm. 10, 182-193.

Lodhi, M.A.K., and Waak, B.T., 1976. A Momentum-Dependent Potential Approximated by the Morse Function for Studying Nuclear Systematics, Ann. Phys. 101, 1-21.

Pfitzer, K.A., 1989. Space station radiation dose as a function of atmospheric density, McDonnell Douglas Space Syst. Co., Advanced Technology Center, Huntington Beach, CA 92647, MDSSC Rep. no. H5387.

Ray, E.C., 1966. On the theory of protons trapped in the Earth's magnetic field, J. Geophys. Res. 65, 1125-1134.

Sawyer, D.M., and Vette, J.I., 1976. AP8 trapped proton environment for solar maximum and solar minimum, Nat. Spa. Sci. Data Center, Goddard Space Flight Center, NSSDC/WDC-A-R&S 76-06.

Schulz, M. and Lanzerotti, L.J., 1974. Particle Diffusion in the Radiation Belts, Springer-Verlag, New York.

Singleterry, R., *et al.*, 2004. Space Ionizing Radiation Environments and Shielding Tools (SIREST), NASA, http://sirest.larc.nasa.gov/.

Spejeldvik, W.N. and P.L. Rothwell, 1985. The Radiation Belts. In: Handbook of Geophysics and the Space Environment, Jursa, A.S. (Ed.), AFGL, USAF, DA167000, 5-1.

Watts, J.W., Parnell, T.A., and Heckman, H.H., 1989. Approximate angular distribution and spectra for geomagnetically trapped protons in low-Earth orbit. In: High-Energy Radiation Background in Space, Rester, A.C., Jr., and Trombka, J.I. (Eds.), AIP Conf. Proc. 186, New York, 75-85.

Figure 1 Caption

Carpet plot of the solar-cycle terms altitude h and modulation flux $F_{10.7}$, as a function of atmospheric density in Equations (1) and (2).

Figure 2 Caption

Graph of proton differential flux versus energy at various Shuttle and International Space Station altitudes, comparing the present algorithm with predictions of AP8 at solar maximum.

Figure 3 Caption

Graph of proton differential flux versus energy at 400 km altitude. Proton flux models AP8MAX, SIREST/NOAAPRO, and the algorithm presented here are compared at solar maximum with $F_{10.7} = 150$. AP8MIN is also given.

Figure 4 Caption

Graph of proton differential flux versus energy, like Figure 3, except at 500 km altitude. Proton flux models SIREST/NOAAPRO, and the algorithm presented here are compared at solar maximum with $F_{10.7} = 150$. AP8MAX is not illustrated since it is essentially identical to SIREST at this altitude. AP8MIN is also given.

Figure 5 Caption

The same graph as in Figure 3, except with the algorithm initialized for half-way through the assumed solar cycle assuming $F_{10.7} = F_{bar} = 110$ in Eq. (1) and (2).

Figure 6 Caption

The same graph as in Figure 4, except with the algorithm initialized for half-way through the assumed solar cycle assuming $F_{10.7} = F_{bar} = 110$ in Eq. (1) and (2).

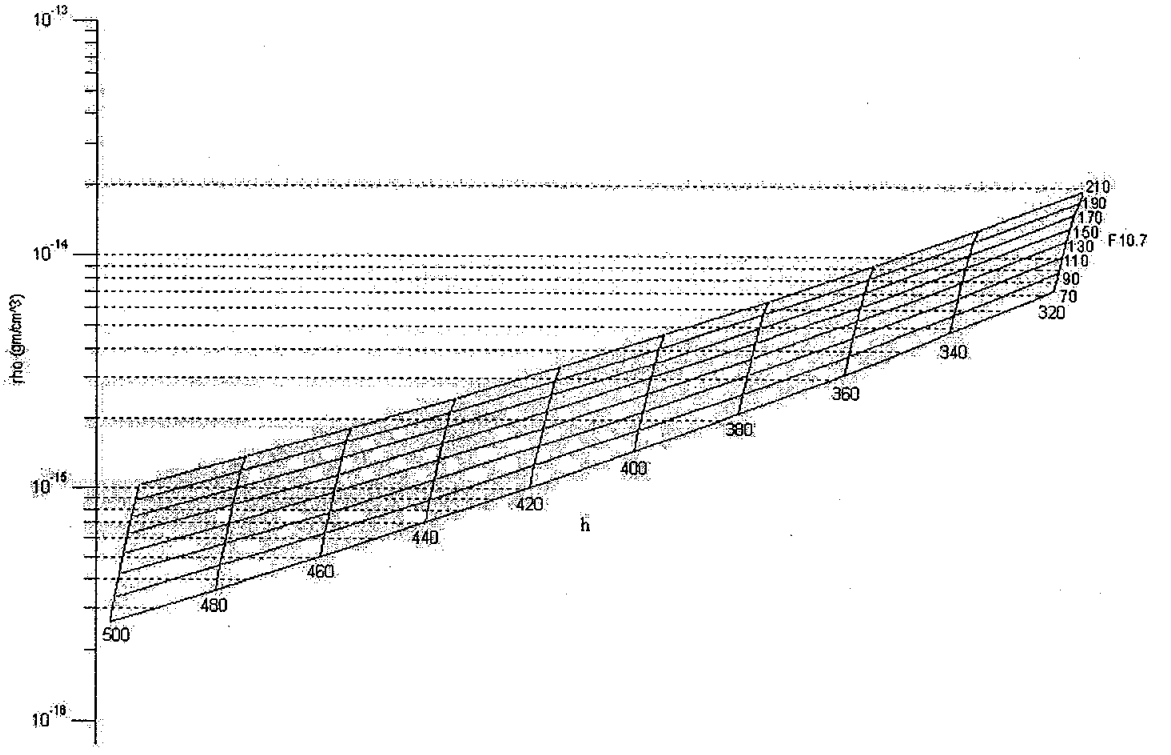

10^-13
10^-14
10^-15
10^-16
rho (gm/cm^3)
210
190
170
150
130
110
90
70
F10.7
320
340
360
380
400
420
440
460
480
500
h

Figure 1

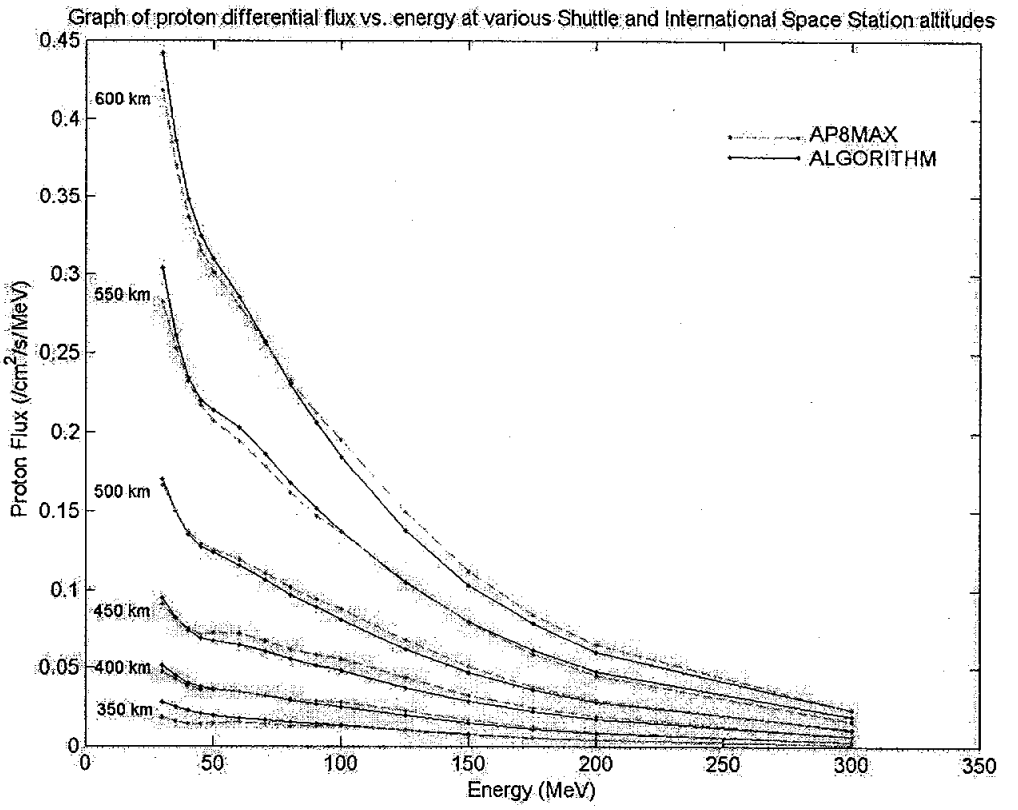

Graph of proton differential flux vs. energy at various Shuttle and International Space Station altitudes
Proton Flux (/cm²/s/MeV)
Energy (MeV)
600 km
550 km
500 km
450 km
400 km
350 km
AP8MAX
ALGORITHM
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0
50
100
150
200
250
300
350

Figure 2

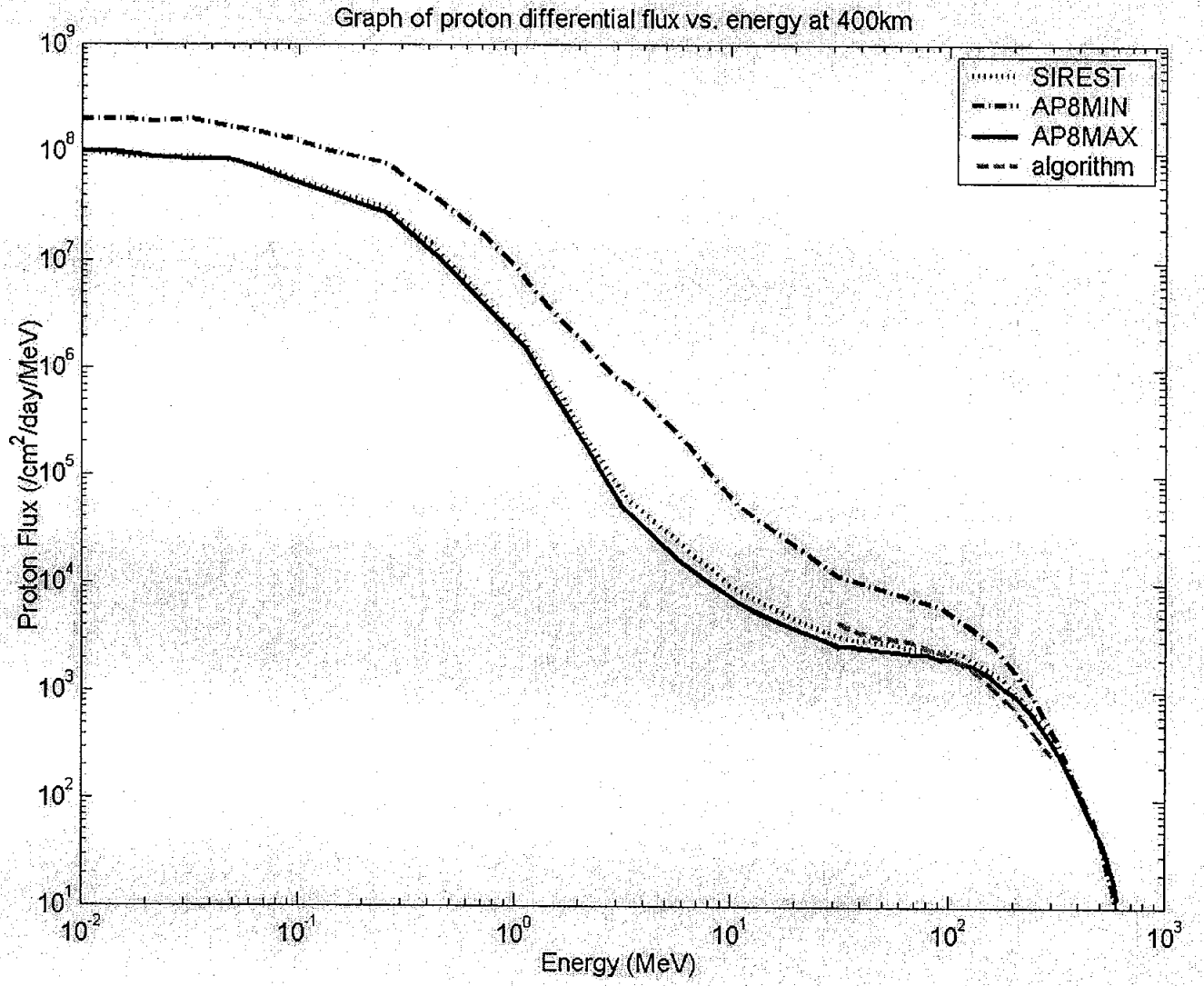

Graph of proton differential flux vs. energy at 400km
SIREST
AP8MIN
AP8MAX
algorithm
Proton Flux (/cm2/day/MeV)
Energy (MeV)
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^-2
10^-1
10^0
10^1
10^2
10^3

Figure 3

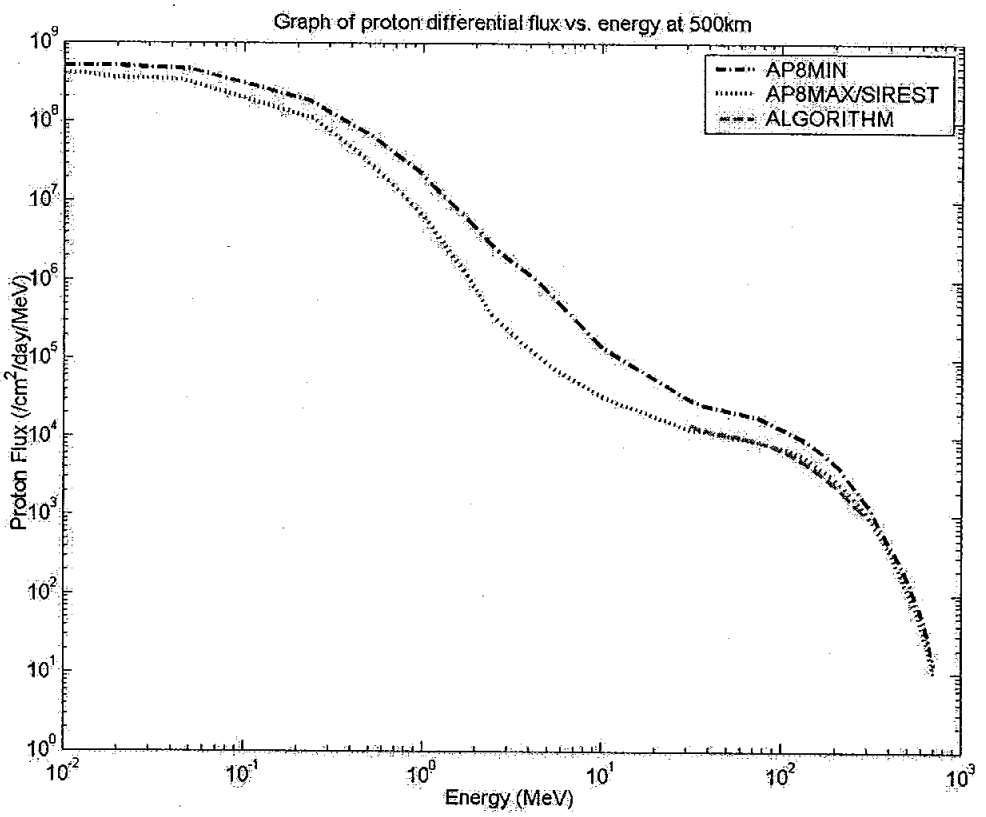

Graph of proton differential flux vs. energy at 500km
AP8MIN
AP8MAX/SIREST
ALGORITHM
Proton Flux (/cm²/day/MeV)
Energy (MeV)
10⁹
10⁸
10⁷
10⁶
10⁵
10⁴
10³
10²
10¹
10⁰
10⁻²
10⁻¹
10⁰
10¹
10²
10³

Figure 4

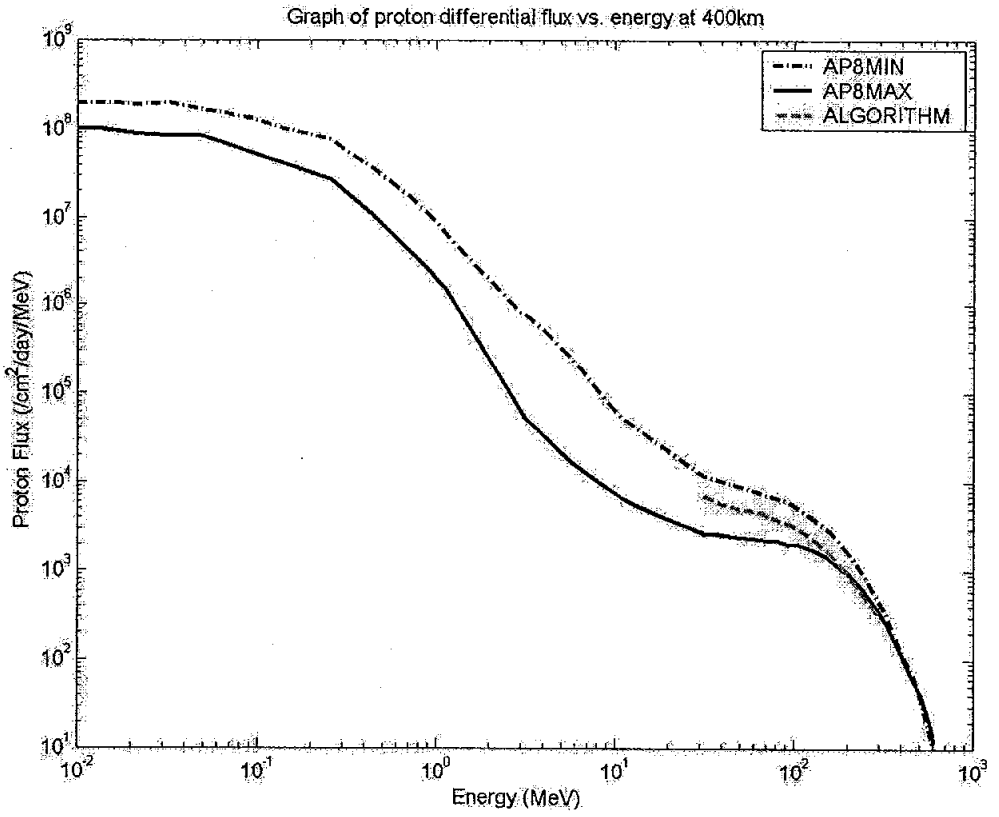

Graph of proton differential flux vs. energy at 400km
AP8MIN
AP8MAX
ALGORITHM
Proton Flux (/cm²/day/MeV)
Energy (MeV)

Figure 5

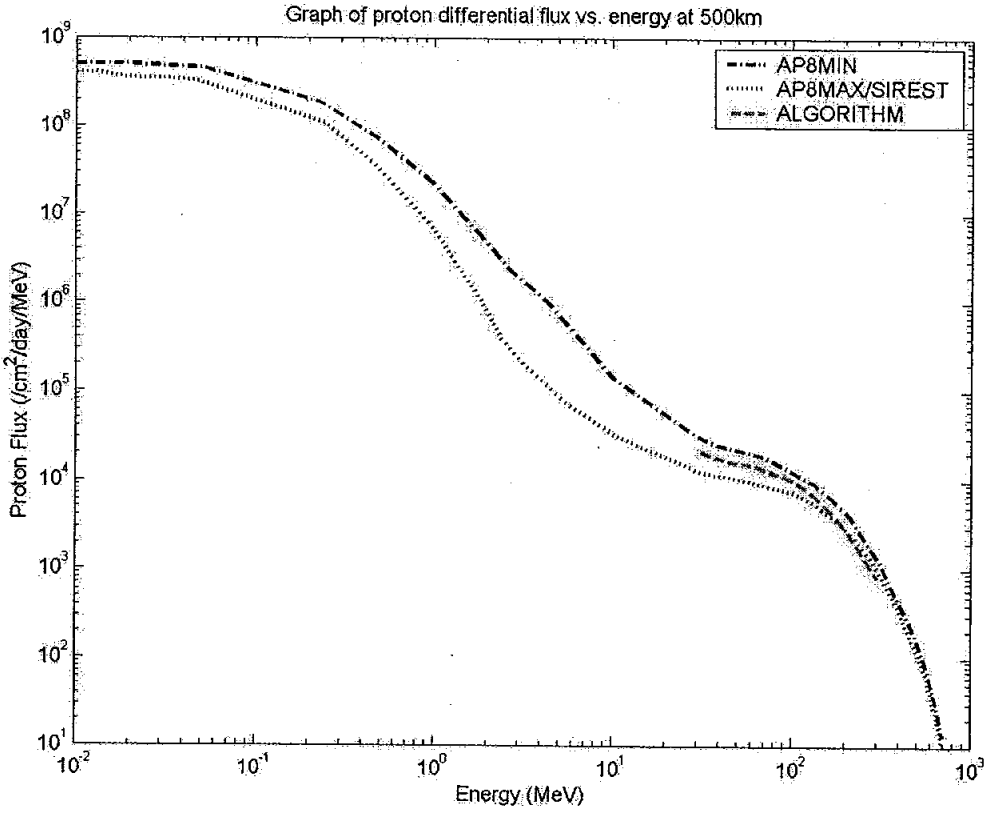

Graph of proton differential flux vs. energy at 500km
AP8MIN
AP8MAX/SIREST
ALGORITHM
Proton Flux (/cm²/day/MeV)
Energy (MeV)
10⁹
10⁸
10⁷
10⁶
10⁵
10⁴
10³
10²
10¹
10⁻²
10⁻¹
10⁰
10¹
10²
10³

Figure 6

Table and Table Caption

Table 1

Comparison of $J_{min}^{(Algorithm)}$ with $J_{min}^{(SPENVIS)}$ for proton energy of 100 MeV at 400 km.

SPENVIS Parameter	Value	Value
ρ_{max}	$3.8x10^{-15}$	
ρ_{min}	$9.57x10^{-16}$	
$J_{max}^{(SPENVIS)}$	$2.79x10^{-2}$	
$J_{min}^{(SPENVIS)}$	$5.15x10^{-2}$	
$J_{min}^{(SPENVIS)} / J_{max}^{(SPENVIS)}$	1.78	
Comparison with Algorithm	Algorithm (3)	Algorithm (7)
$J_{min}^{(Algorithm)}$	4.955	5.680
Difference from $J_{min}^{(SPENVIS)}$	0.183	0.53
% Error	3.5	10.3

Units of ρ are in g/cm^3 and those of J are $cm^{-2}s^{-1}MeV^{-1}$.

An XML Representation for Crew Procedures

Final Report
NASA Faculty Fellowship Program - 2004

Johnson Space Center

Prepared by:	Richard C. Simpson, PhD, ATP
Academic Rank:	Assistant Professor
University & Department:	University of Pittsburgh Department of Rehabilitation Science and Technology Pittsburgh, PA 15260
NASA/JSC	
Directorate:	Engineering
Division:	Automation Robotics and Simulation
Branch:	Intelligent Systems
Directorate:	Space Operations
Division:	Automation and Robotics
Branch:	Engineering
JSC Colleague:	Cliff Farmer
Date Submitted:	August 6, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

NASA ensures safe operation of complex systems through the use of formally-documented procedures, which encode the operational knowledge of the system as derived from system experts. Crew members use procedure documentation on the ground for training purposes and on-board space shuttle and space station to guide their activities. Investigators at JSC are developing a new representation for procedures that is *content-based* (as opposed to *display-based*). Instead of specifying how a procedure should look on the printed page, the content-based representation will identify the components of a procedure and (more importantly) how the components are related (e.g., how the activities within a procedure are sequenced; what resources need to be available for each activity). This approach will allow different sets of rules to be created for displaying procedures on a computer screen, on a hand-held personal digital assistant (PDA), verbally, or on a printed page, and will also allow intelligent reasoning processes to automatically interpret and use procedure definitions.

During his NASA fellowship, Dr. Simpson examined how various industries represent procedures (also called *business processes* or *workflows*), in areas such as manufacturing, accounting, shipping, or customer service. A useful method for designing and evaluating workflow representation languages is by determining their ability to encode various *workflow patterns*, which depict abstract relationships between the components of a procedure removed from the context of a specific procedure or industry. Investigators have used this type of analysis to evaluate how well-suited existing workflow representation languages are for various industries based on the workflow patterns that commonly arise across industry-specific procedures. Based on this type of analysis, it is already clear that existing workflow representations capture discrete flow of *control* (i.e., when one activity should start and stop based on when other activities start and stop), but do not capture the flow of data, materials, resources or priorities. Existing workflow representation languages are also limited to representing sequences of discrete activities, and cannot encode procedures involving continuous flow of information or materials between activities.

INTRODUCTION

NASA ensures safe operation of complex systems through the use of formally-documented procedures, which encode the operational knowledge of the system as derived from system experts. Crew members use procedure documentation on the ground for training purposes and on-board space shuttle and space station to guide their activities. NASA is currently moving from a print-oriented PDF representation to an XML representation for procedures, but the XML representation seeks simply to mimic the PDF look and feel without including any semantic or syntactic information. In other words, there is no explicit identification of procedure *components* (e.g., resources, activities, warnings, pre-conditions, post-conditions) or rules about how components interact. This makes it impossible for intelligent reasoning processes to use this representation for tasks like validation and verification, execution tracking and procedure assistance.

As an alternative, investigators at JSC are developing a new representation for procedures that is *content-based* (as opposed to *display-based*). Instead of specifying how a procedure should look on the printed page, the content-based representation will identify the components of a procedure and (more importantly) how the components are related (e.g., how the activities within a procedure are sequenced; what resources need to be available for each activity). This approach will allow different sets of rules to be created for displaying procedures on a computer screen, on a hand-held personal digital assistant (PDA), verbally, or on a printed page, and will also allow intelligent reasoning processes to automatically interpret and use procedure definitions. The initial goal of the project is to develop a content-based representation for procedures that can be used in place of the existing display-based representation. Once the representation has been developed, editing tools will be developed and tested using actual NASA systems, procedures and system experts. Ultimately, the representation will be used by intelligent systems to provide adaptive training, assistance and monitoring.

During his NASA fellowship, Dr. Simpson examined how various industries represent procedures (also called *business processes* or *workflows*), in areas such as manufacturing, accounting, shipping, or customer service. Content-based workflow representations can be displayed graphically or textually, and there is often a direct mapping between a graphical and textual representation of a workflow. Graphical workflow representations (e.g., UML Activity Diagrams, Petri-Nets, Gantt Charts, BPMN [1]) are typically easier for humans to understand and manipulate, but textual representation languages (e.g., XPDL [2], BPML [3]) can be interpreted by *workflow engines* to automatically manage and monitor execution of business processes.

WORKFLOW MANAGEMENT

Workflow management technology is used to automate business processes in which data and tasks are passed between (human and machine) participants according to a defined set of rules to achieve an overall business goal. Workflow management technology is most frequently used in office environments in applications such as

accounting, shipping and general administration, but it is also applicable to design, engineering and manufacturing [4]. An emerging use of workflow management technology is within web sites [5], to automate interactions between a user, the website, and the website's underlying business infrastructure.

A workflow management system automates a business process by managing the sequence of work activities and invoking the appropriate human and/or information technology (IT) resources associated with each activity as specified in a process definition [4]. A process definition consists of a network of activities and their relationships, criteria to indicate the start and termination of the process, and information about each activity within the process such as participants, associated IT applications and data [[6]].

The process definition is expressed in a textual or graphical form or in a formal language notation, which we refer to as a workflow representation language [4]. Textual formats work well for linear tasks, but not for tasks with lots of branching. Difficult to get an "overview" of the task. Difficult to express dependencies within task. Graphical formats provide a good overview of a process but the symbols don't provide room for much detail. Graphical formats often don't have a natural way to represent groupings or hierarchies among steps [7].

Each activity within a process is a single logical step in the process (e.g., making a payment, filing an invoice). It is sometimes not practical to automate all activities within a process, but the process definition will still describe all activities whether they are performed automatically or manually. For example, if a document must be signed in front of a witness, then this might be the one manual activity within an otherwise automated process [6].

WORKFLOW REPRESENTATION LANGUAGES

Unified Modeling Language (UML)

UML provides a visual, object oriented (OO) modeling notation that is valuable for designing and understanding complex systems. UML is the most widely known modeling notation, has a graphical notation which is readily understood, and a rich set of semantics for capturing key features of OO systems [8]. Unfortunately, no single type of UML diagram captures all of the information needed to describe a process. UML activity diagrams can represent complicated sequences and parallelism, but are not the best choice for representing the relationships between activities and objects. UML interaction diagrams do a much better job describing how actions and objects collaborate [9].

Petri-Nets

A Petri Net is a particular kind of directed graph with an initial state called initial marking. The underlying graph of a Petri Net is a directed, bipartite graph consisting of two kinds of nodes, called places and transitions. Arcs represent connections between nodes. An arc can only connect from a place to a transition or from a transition to a

place. Connections between two nodes that are of the same kind are not allowed. In graphical representation, places are drawn as circles and transitions as bars or boxes. A marking (state) is an assignment of tokens to the places of the Net. A transition is enabled if each place connected to the transition input arc (input place), contains at least one token. The firing of an enabled transition removes a token from each input place and deposits a token on each place connected with its output arcs (output place). At any given time instance, the distribution of tokens on places defined the current state of the Petri Net; thus, the modeled system. Petri Nets also allow the determination of reachability (if a reachable/obtainable from a given state) and deadlocking (if a state could be reached where the process can not proceed) [10].

Business Process Modeling Notation (BPMN) [1]

The Business Process Modeling Notation (BPMN) specification provides a graphical notation for expressing business processes in a Business Process Diagram (BPD). The objective of BPMN is to support business process management by both technical users and business users by providing a notation that is intuitive to business users yet able to represent complex process semantics. The BPMN specification also provides a mapping between the graphics of the notation to the underlying constructs of execution languages, particularly BPEL4WS [1].

Process Specification Language (PSL) [9]

PSL allows for the possibility of multiple syntaxes, with the choice of syntax depending on factors such as the nature of the process being described and the data source and destination. Key to PSL are the formal definitions (ontology) that underlie the language. Because of these explicit and unambiguous definitions, information exchange can be achieved without relying on hidden assumptions or subjective mappings. PSL semantics are represented using a formal language developed for the exchange of knowledge among disparate computer programs. Thus concepts can within a process be defined unambiguously, a necessary characteristic to exchange process information using the PSL ontology [9].

Business Process Modeling Language (BPML) [3]

BPML provides an abstract model for expressing business processes and supporting entities. BPML defines a formal model for expressing abstract and executable processes that address all aspects of enterprise business processes, including activities of varying complexity, transactions and their compensation, data management, concurrency, exception handling and operational semantics. BPML also provides a grammar in the form of an XML Schema for enabling the persistence and interchange of definitions across heterogeneous systems and modeling tools [3].

Business Process Execution Language for Web Services (BPEL4WS) [5]

BPEL provides an XML notation and semantics for specifying business process behavior based on Web Services. A BPEL4WS process is defined in terms of its interactions with partners. A partner may provide services to the process, require services from the process, or participate in a two-way interaction with the process. Thus BPEL orchestrates Web Services by specifying the order in which it is meaningful to call a collection of services, and assigns responsibilities for each of the services to partners [8].

WORKFLOW PATTERNS

A useful method for designing and evaluating workflow representation languages is by determining their ability to encode various *workflow patterns* [11], which depict abstract relationships between the components of a procedure removed from the context of a specific procedure or industry. Investigators have used this type of analysis [10-12] to evaluate how well-suited existing workflow representation languages are for various industries based on the workflow patterns that commonly arise across industry-specific procedures.

Based on this type of analysis, it is already clear that existing workflow representations capture discrete flow of *control* (i.e., when one activity should start and stop based on when other activities start and stop), but do not capture the flow of data, materials, resources or priorities. Existing workflow representation languages are also limited to representing sequences of discrete activities, and cannot encode procedures involving continuous flow of information or materials between activities.

USING XML AS THE BASIS FOR A WORKFLOW REPRESENTATION LANGUAGE

An XML markup scheme for process data should take advantage of what XML does best, while minimizing the impact of where XML falls short. XML's "tag-centric" syntax makes it a natural fit for representing ordered sequences and hierarchies. Thus it is well suited for ordering time points and occurrences of activities. It is also good at representing sub-activities and sub-occurrences. Another capability of XML, useful for process representation, is XML's modularity. For example, using XML namespaces I can embed an arbitrary object description into a process specification and leave it up to a software tool, separate from the process specification interpreter, to parse the object description. I can also employ namespaces to modularize our process markup language itself (perhaps mirroring PSL's modularization) [9].

Although XML has many advantages for representing processes, it has a major disadvantage. While XML excels as a serialization syntax for exchanging data structures between applications, XML is not very good at expressing the kinds of complex constraints needed for process descriptions. For example, it might be difficult for an

XML schema for a process description language to enforce scheduling constraints involving shared resources. Such constraints could be more easily expressed in a rich language for knowledge representation such as KIF [9].

Because XML is deficient when it comes to representing complex constraints on populations of data elements, its process representation capabilities are limited. However, this does not mean that I cannot use XML to exchange process descriptions. Rather, it means that I probably would not want to exchange all of a process description's underlying ontology in XML, and I cannot count on an XML language to enforce all constraints on process data. It also means that XML would be a poor authoring environment for all but the most simple process descriptions [9].

REFERENCES

[1] S. A. White, "Business Process Modeling Notation (BPMN)," Business Process Management Initiative (BPMI) 3 May 2004.

[2] WfMC, "Workflow Process Definition Interface -- XML Process Definition Language," Workflow Management Coalition, Lighthouse Point, FL, Workflow Standard WFMC-TC-1025, 25 October 2002.

[3] A. Arkin, "Business Process Modeling Language," Business Process Management Initiative, Specification 13 November 2002.

[4] "Workflow Management Facility Specification, V1.2," Object Management Group, Needham, MA April 2000.

[5] T. Andrews, F. Curbera, H. Dholakia, Y. Goland, J. Klein, F. Leymann, K. Liu, D. Roller, D. Smith, S. Thatte, I. Trickovic, and S. Weerawarana, "Business Process Execution Language for Web Services, v1.1," 5 May 2003.

[6] R. Allen, "Workflow: An introduction," in *The Workflow Handbook 2001*, L. Fischer, Ed. Lighthouse Point, FL: Future Strategies, Inc., 2001, pp. 15-38.

[7] D. R. Wieringa and D. K. Farkas, "Procedure writing across domains: Nuclear power plant procedures and computer documentation," presented at International Conference on Systems Documentation, Chicago, IL, 1991.

[8] K. Mantell, "From UML to BPEL," vol. 2004: IBM developerWorks, 2003.

[9] D. Dodds, A. Watt, M. Birbeck, J. Cousins, D. Rivers-Moore, R. Worden, M. Nic, D. Ayers, K. Ahmed, A. Wrightson, and J. Lubell, *Professional XML Metadata*. Hoboken, NJ: Wrox Press, 2001.

[10] A. Knutilla, C. Schlenofff, S. Ray, S. T. Polyak, A. Tate, S. C. Cheah, and R. C. Anderson, "Process Specification Language: Analysis of Existing Representations," National Institute of Standards and Technology, Gaithersburg, MD NISTIR 6133, 1998.

[11] W. M. P. van der Aalst, A. H. M. ter Hofstede, B. Kiepuszewski, and A. P. Barros, "Workflow Patterns," *Distributed and Parallel Databases*, vol. 14, pp. 5-51, 2003.

[12] C. Schlenoff, A. Knutilla, and S. Ray, "Unified Process Specification Language: Requirements for Modeling Process," National Institute of Standards and Technology, Gaithersburg, MD NSTIR 5910, September 1996.

Experimental Reproduction of Olivine rich Type-I Chondrules

Final Report
NASA Faculty Fellowship Program-2004

Johnson Space Center

Prepared by:	Robert K. Smith, Ph.D.
Academic Rank:	Professor
University & Department:	The University of Texas at San Antonio Dept. of Earth and Environmental Science San Antonio, Texas 78249-0663
NASA/JSC	
Directorate:	Space and Life Sciences
Division:	Astromaterials Research & Exploration Science (ARES)
Branch:	Astromaterials Acquisition & Curation
JSC Colleague:	Gary E. Lofgren, Ph.D.

Date Submitted:	August 6, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

Ordinary chondritic meteorites are an abundant type of stony meteorite characterized by the presence of chondrules. Chondrules are small spheres consisting of silicate, metal, and sulfide minerals that experienced melting in the nebula before incorporation into chondritic meteorite parent bodies. Therefore, chondrules record a variety of processes that occurred in the early solar nebula. Two common types of unequilibrated chondrules with porphyritic textures include FeO-poor (type I) and FeO-rich (type II) each subdivided into an A (SiO_2-poor) and B (SiO_2-rich) series. Type IA chondrules include those with high proportions of olivine phenocrysts (>80% olivine) and type IB chondrules include those with high proportions of pyroxene phenocrysts (<20% olivine). An intermediate composition, type IAB chondrules include those chondrules in which the proportion of olivine phenocrysts is between 20-80%. We conducted high-temperature laboratory experiments (melting at 1550° C) to produce type I chondrules from average unequilibrated ordinary chondrite (UOC) material mixed with small amounts of additional olivine. The experiments were conducted by adding forsteritic rich olivine (San Carlos olivine, Fo 91) to UOC material (GRO 95544) in a 30/70 ratio, respectively. Results of these high temperature experiments suggest that we have replicated type IA chondrule textures and compositions with dynamic crystallization experiments in which a heterogeneous mixture of UOC (GRO 95544) and olivine (San Carlos olivine) were melted at 1550°C for 1 hr. and cooled at 5-1000°C/hr using graphite crucibles in evacuated silica tubes to provide a reducing environment.

INTRODUCTION

Chondritic meteorites are the oldest and most primitive rocks in the solar system (Brearley and Jones, 1998). Additionally, chondrites are the hosts for interstellar grains that predate solar system formation and survived processing in the protoplanetary disk (solar nebula) environment (Brearley and Jones, 1998). The abundance of chondrules in chondrites implies that melting of small particles was a common phenomenon in the early solar system (Hewins, 1997). Therefore, chondrites offer planetary scientists the opportunity to study the earliest history of formation of our solar system.

Ordinary chondritic meteorites are an abundant type of stony meteorite characterized by the presence of chondrules. Chondrules are small spheres consisting of silicate, metal, and sulfide minerals that experienced melting before incorporation into chondritic meteorite parent bodies. Therefore, chondrules record a variety of processes that occurred in the early solar nebula. Two common types of unequilibrated chondrules with porphyritic textures include FeO-poor (type I) and FeO-rich (type II) each subdivided into an A (SiO_2-poor) and B (SiO_2-rich) series. Type IA chondrules include those with high proportions of olivine phenocrysts (>80% olivine) and type IB chondrules include those with high proportions of pyroxene phenocrysts (<20% olivine). An intermediate composition, type IAB chondrules include those chondrules in which the proportion of olivine phenocrysts is between 20-80%.

Type I chondrules have a distinctive chemical composition (Lofgren and Le, 2002) that is largely devoid of oxidized iron (Jones and Scott, 1989; and Jones, 1994). The silicate phases usually contain less than 5 wt. % FeO and metallic Fe is usually abundant (Lofgren and Le, 2002). Type I chondrules display a wide range of textures from barred to porphyritic to partially melted aggregates (Jones and Scott, 1989; Jones, 1994; and Lofgren and Le, 2000). Because of the reduced nature of the chondrules, experimental duplication of their crystallization histories is difficult (Lofgren and Le, 2002). Experiments conducted in silica tubes with the sample in graphite crucibles provides for a reducing environment, but also sets an upper limit on the temperature of the experiments. This limit of 1550°C is well below the melting temperatures required, but does allow testing of the hypothesis that many type I chondrules experienced a partial melting history that does not involve large amounts of melting (Lofgren and Le, 2002). The starting materials used in these experiments are unequilibrated ordinary chondrites (UOC) of the L petrologic type (i.e., low total Fe content). The bulk composition of these chondrites when reduced approaches the composition of the type IAB and IA chondrules. The experiments suggest that such a partial melting history is consistent with the observed textures and mineral chemistries in many type IAB and IA chondrules.

EXPERIMENTAL TECHNIQUES

Experiments were conducted in evacuated silica tubes with the sample in a graphite crucible using techniques similar to McCoy et al. (1999). The experimental configuration

produces a reducing environment in which the oxygen partial pressure (f_{O2} or oxygen fugacity) is 3 to 5 orders of magnitude below the Iron-Wüstite (IW) buffer in the temperature range 800 to 1550°C, respectively. This f_{O2} is most likely lower than for type I chondrules, but does provide a lower limit. The starting meteorite material, collected from the Grosvenor Mountains, Antarctica, was comprised of fragments of the unequilibrated ordinary chondrite (UOC) GRO 95544 (L3.1) that were ground to an average grain size of approximately 50μm (<10 to 100 μm) and then mixed with San Carlos olivine (Fo91) in the ratio of 70/30, respectively. Starting materials differed principally in the grain size of the olivine. Heterogeneous aliquots (70/30 ratio, i.e., UOC/SC olivine) ranging between 145 and 155 mg were then pressed into pellets that were then placed in a graphite crucible. Each graphite crucible was sealed in an evacuated silica tube and placed in a furnace. Dynamic crystallization experiments (i.e., controlled cooling conditions) were brought to1550°C for 1 hour and then cooled at rates from 5-1000°C/hr. Each sample was quenched by removing the silica tube from the furnace and placing it in a stream of compressed air.

Polished microprobe mounts were prepared for textural and chemical analyses. Backscatter (BSE) images and mineral analyses were collected on the JEOL JSM-5910 LV SEM and the Cameca SX-100 microprobe, respectively at the NASA-JSC. BSE images and microprobe analyses of minerals were collected using an accelerating potential of 15 Kv, and 15 Kv and a beam current of 20 nA, respectively. Natural minerals were used as standards for the microprobe analyses.

RESULTS

All the experimental charges show evidence of partial melting in a reducing environment. The degree of silicate partial melting ranges from approximately 10-15 % to near 70 %. All of the silica phases crystallized from the melt plus the glassy mesostasis contain less than 1 wt. % FeO and metallic Fe occurs as large rounded blebs and as abundant, small blebs (<1μm) of exsolved Fe-metal in the olivine. Some zoned "relict" olivine however, show Fe contents that range from 8.3 to 1.4 wt. %. Olivine and pyroxene grains crystallized from the melt have uniform compositions with no obvious chemical zoning. Additionally, the metal blebs show a tendency to migrate to the outer edge of the experimental charges. All dynamic cooling experiments have elliptical to rounded shapes with large to small vesicles.

The experiment with the best developed type I characteristics is GRO-295. It was melted at 1550°C for 1 hour and cooled at 5°C/hr to 800°C. Figure 1 shows a type IA chondrule from QUE 97008 and compared with experiment GRO-295. GRO-295 has a well developed type IA texture (Figure 1D) in which the olivine is in contact with a clinopyroxene (cpx) set in a glassy mesostasis. Typical compositions of the phases analyzed in GRO-295 are given in Table 1. The olivine and orthopyroxene are relatively homogeneous and consistent in composition. Clinopyroxene has a variable composition relative to Ca and Al. The glassy mesostasis shows the greatest compositional variation,

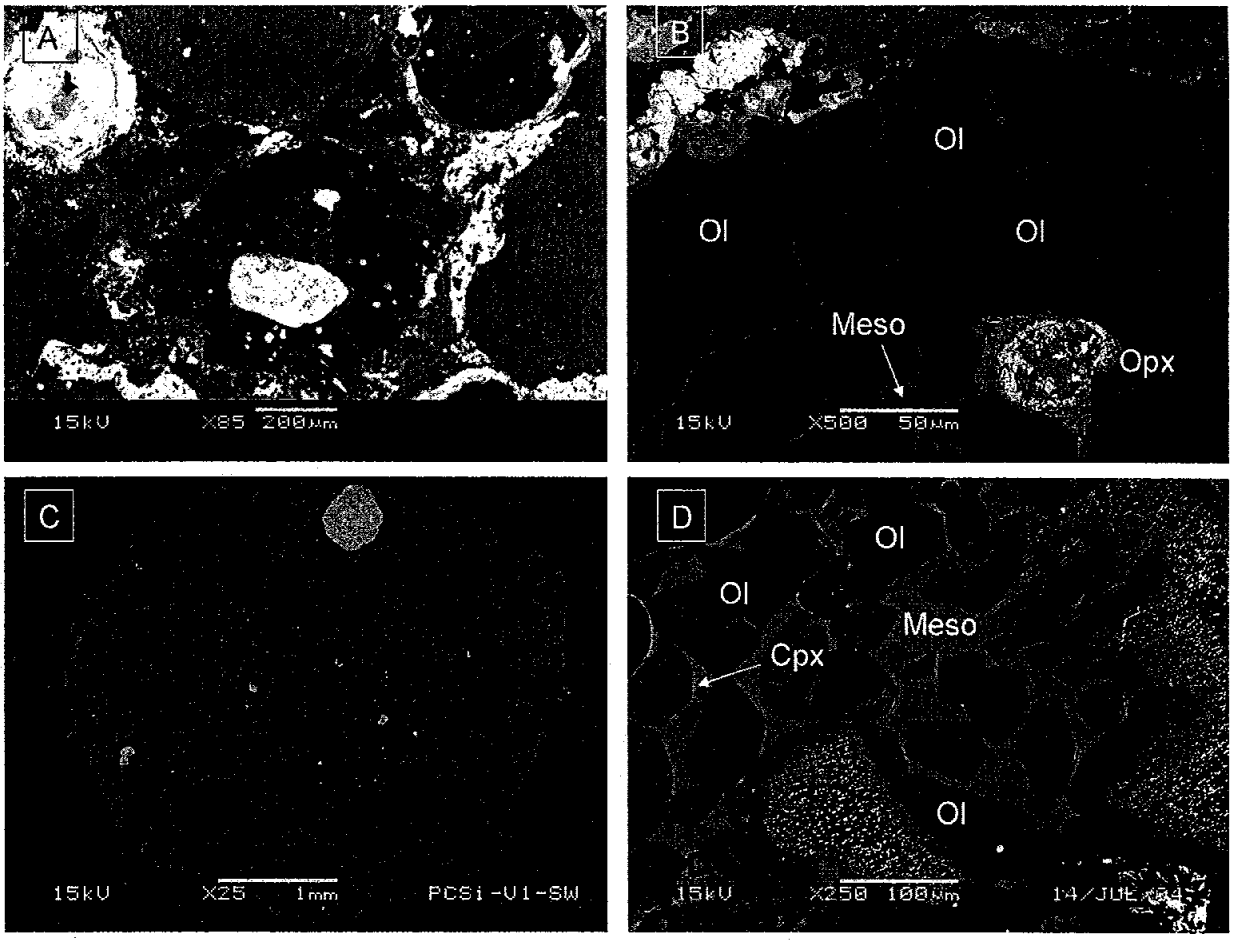


Figure 1. SEM backscatter images showing a comparison of an experimentally produced type IA chondrule (photo C) with a natural type IA chondrule from QUE 97008, lower center of photo A. A) Type IA chondrule. Olivine-orthopyroxene (dark grains) and metal (white). B) BSE image of QUE 97008-6, 500x. Ol = olivine (dark grains), Opx = orthopyroxene (medium gray grains), and Meso = glassy mesostasis (light gray). C) Experimental charge GRO-295 melted at 1550°C for 1 hour and cooled at 5°C/hr. Chondrule is circular in shape with numerous vesicles. Olivine (dark grains), orthopyroxene (medium gray grains), and glassy mesostasis (light gray). D) BSE image of experimental charge GRO-295 shown in photo C at 250x. Ol = olivine (dark grains), Cpx = clinopyroxene (light gray), and Meso = glassy mesostasis (medium gray). "Relict", more Fe-rich olivine shows partial melting with exsolving Fe metal blebs and heterogeneous nucleation and growth of more Mg-rich olivine rims. Smaller, equant, euhedral to anhedral crystals of Mg-rich olivine (dark) that nucleated and grew from the melt during cooling are present. Orthopyroxene has then nucleated and rims some olivine crystals (not seen in this BSE image). The texture and mineral proportions between photos B and C are similar, but more glassy mesostasis exists in the experimental charge relative to the natural chondrule.

Table 1.Typical compositions of olivine, orthopyroxene, clinopyroxene, and glassy mesostasis in experiment GRO-295, all in oxide wt. %.

Oxide	Olivine	Orthopyroxene	Clinopyroxene	Mesostasis
SiO_2	42.95	59.12	50.35	54.69
TiO_2	0.01	0.22	1.09	0.23
Al_2O_3	0.03	1.44	9.54	24.32
Cr_2O_3	0.06	0.11	0.17	0.03
FeO	0.69	0.37	0.33	0.47
MnO	0.05	0.02	0.07	0.06
MgO	56.61	38.90	18.08	5.97
CaO	0.13	0.46	21.01	14.72
Na_2O	0.00	0.00	0.00	0.03
K_2O	0.00	0.00	0.00	0.00
P_2O_5	0.00	0.00	0.15	0.10
Total	100.54	100.65	100.78	100.61

but is dependent on the amount of crystallization of the experimental charge. Sodium (Na) in experiment GRO-295 is very low, reflecting the long duration of the cooling event.

DISCUSSION

Type IA chondrules are characterized by highly forsteritic olivine and En (enstatite) rich pyroxenes (Jones and Scott, 1989). Additionally, type IA chondrule mesostasis varies from 100% glass to partly microcrystalline, occupying 5-15 volume % of the chondrule (Jones and Scott, 1989). The mineralogy in GRO-295 experiment shows all silicate phases (olivine and pyroxene) to meet the definition of a type IA chondrule, i.e., olivines are forsteritic and the pyroxenes are magnesium-rich. However, because the experimental starting material is heterogeneous modal percentages of the silicate minerals and mesostasis vary from one area to another within the experimental charge. Therefore, experimental preliminary results suggest that the formation of type I chondrules by reduction of ordinary chondrite debris in the solar nebula is a viable mechanism. The consensus is that most chondrules are crystallized melt droplets from the near total melting of crystalline precursor material (Lofgren ,1996; Lofgren and Le, 2002). The process proposed here, however is one of partial melting (<70% melting). The reduction and partial melting would take place at temperatures equal to or less than 1550°C

CONCLUSIONS

In this study we have reproduced type I chondrules by the reduction of UOC material (GRO 95544) mixed with olivine (San Carlos olivine, Fo 91) in the ratio 70/30, during

dynamic crystallization experiments. The heterogeneous powdered UOC-olivine mixture used as starting material in the experiments simulates crystalline chondrule precursors in the solar nebula. These chondrules vary in silicate mineral modes and texture depending on the degree of melting of the precursor UOC-olivine mixture and the ultimate cooling rate after the melting-crystallization event.

REFERENCES CITED

Brearley, A.J., and Jones, R., 1998, Chondritic meteorites. Reviews in Mineralogy, vol. 36, Planetary Materials; Mineralogical Society of America, p. 3-1 to 3-398.

Hewins, R.H., 1997, Chondrules. Annual Review of Earth and Planetary Sciences, vol. 25, p. 61-83.

Jones, R. H., 1994, Petrology of FeO-poor, porphyritic pyroxene chondrules in the Semarkona chondrite. Geochimica Cosmochimica Acta, vol. 58, p. 5325-5353.

Jones, R. H., and Scott, E. R. D., 1989, Petrology and thermal history of type IA chondrules in the Semarkona (LL3.0) chondrite. Proc. 19th Lunar and Planetary Science Conference, p. 523-536.

Lofgren, G.E., 1996, A dynamic crystallization model for chodrule melts. In Chondrules and the Protoplanetary Disk, editors, R.H. Hewins, R.H. Jones, and E.R.D. Scott, p. 187-196.

Lofgren. G. E., and Le, L., 2000, Experimental evidence for a partial melting origin for most porphyritic chondrules. In Lunar and Planetary Science XXXI, Abstract #1809, Lunar and Planetary Institute, Houston, Texas (CD-ROM).

Lofgren, G. E., and Le, L., 2002, Experimental reproduction of Type-1B chondrules. In Lunar and Planetary Science XXXIII, Abstract #1612, Lunar and Planetary Institute, Houston, Texas (CD-ROM).

McCoy, T.J., Dickinson, T.L., and Lofgren, G.E., 1999, Partial melting of the Indarch (EH4) meteorite: A textural, chemical, and phase relations view of melting and melt migration. Meteoritics & Planetary Science, vol. 34, p. 735-746.

Packaging Materials for Thermally Processed Foods in Future Space Missions

Final Report
NASA Faculty Fellowship Program – 2004
Johnson Space Center

Prepared by:	Juming Tang, Ph.D.
Academic Rank:	Professor
University and Department:	Washington State University Department of Biological Systems Engineering Pullman, WA 99164-6120
NASA/JSC	
Office:	Habitability and Environmental Factor Office (HEFO) Habitability&Human Factors Office Space Food Systems Laboratory (SFSL)
JSC Colleague	Vickie Kloeris
Date Submitted:	August 12, 2004
Contract Number:	NAG 20-1526 and NNJ04JF93A

ABSTRACT

Thermally processed shelf-stable foods are important in International Space Station (ISS) programs and essential to the success of future long-duration manned space missions. NASA uses military MRE pouch material to package thermally processed foods for ISS. But the packaging material for MRE pouches contains aluminum (Al) foil as moisture and oxygen barrier. Al foils create potential problem for solid waste disposal in long duration missions, adds much weight, and are not compatible with some of the emerging processing technologies. This is a need to explore the use of non-foil materials that can provide designed shelf-life of 3-5 years for future space missions. This report presents a review on the current status of package options for thermally processes shelf-stable foods and provides an assessment on the potential of using commercially available O_2 and moisture barrier films as a part of package materials for thermally processed foods in future long-duration manned space missions. Based on several criteria, including potential problem in solid waste disposal, weight, mechanical and barrier properties, as well as commercial readiness, laminated EVOH films and SiOx coated films hold most promise as the future package materials for long-term manned space missions. But as of today, none of the commercial pouch films can provide the required O_2 barrier for 3-5 year shelf-life at ambient temperature. Research is needed to investigate the synergistic effects of better engineered laminated structures, shorter processing times at elevated temperature, and controlled storage conditions to meet the requirements of long-duration space missions, especially missions to Mars.

1. INTRODUCTION

Thermally processed shelf-stable foods are a major component of food supply for International Space Station (ISS) and future long-duration manned space missions. Retorting is the only thermal processing method currently used to produce shelf-stable low acid (pH>4.5) foods in North America. It is also used by NASA in producing shelf-stable moist foods. Retorting systems rely on convectional surface heating and internal heat conduction to kill anaerobic spores in packaged foods to make them free from pathogenic and most of spoilage bacteria. The high processing temperatures (120-130°C) and relative long processing times (30-60 min) used in those processes cause severe degradation in the processed foods. Emerging food processing technologies that use volumetric heating through microwaves or adiabatic heating though high pressure hold promise to produce high quality shelf-stable foods.

NASA currently uses military MRE pouches for thermally processed foods. But the packaging material for MRE pouches contains aluminum (Al) foil as moisture and oxygen barrier. Al foils create potential problem for solid waste disposal in long-duration manned missions, adds weight, and are not compatible with some of the emerging processing technologies (e.g., microwave sterilization). This is a need to explore the use of non-foil materials to provide 3-5 year shelf-life for future space missions. The objectives of this project were to study the current status of package options for thermally processes shelf-stable foods through a survey of the literature and food package suppliers and provide an assessment on the potential of using commercially available O_2 and moisture barrier films as a part of package materials for thermally processed foods in future long-duration manned space mission.

2. CRITERIA IN SELECTING PACKAGE MATERIALS

To meet the requirements for long-duration space missions, package materials for thermally processed foods need to satisfy the following criteria:

1). can be heat sealed and withstand retort temperatures (120-130°C),
2). retain adequate mechanical and barrier properties to provide a shelf-life of 3-5 years,
3). do not cause problems in solid waste disposal.
4). have relatively light weights.

3. PRODUCT SHELF-LIFE

Shelf-life refers to the period of time beyond which the product is no longer acceptable to consumers (Perchonok, 2002). Factors that determine the shelf-life of a food product include microbial growth, chemical reactions (e.g., lipid oxidation, Maillard browning) and physical changes, which are, in turn, influenced by storage conditions, such as temperature, relative humidity and other compositions of the ambient air/gases. The amount of O_2 in packaged foods and/or rate of O_2 ingress into a package are among the most important factors controlling microbial growth and chemical reactions. Different foods have different sensitive to O_2 (see Table 1) with dairy, meat, fish, poultry among the most O_2 sensitive groups Controlling the O_2 permeability is one of the most important considerations in selecting and designing package materials, especially for shelf-stable foods. Information in Table 1 serves as a general guide for selecting package materials. But in food companies, direct sensory evaluation is often used to determine the shelf-life of shelf-stable products. For example, in Hormel, processed foods are stored at

a pre-determined temperature, and sensory attributes and microbial accounts were evaluated every month. Color, odor, texture, general taste, and/or microbial counts were used as criteria to determine the shelf-life of the products. Hormel's shelf-stable products processed in trays indicate between 12 and 18 month shelf-life.

Table 1. Maximum allowable ingress of O_2 or loss or gain of moisture in shelf-stable products (Armstrong, 2002)

Foods	Max O_2 ingress, ppm	Max H_2O gain (+) or loss (-) %
Canned milk, meats, fish, poultry, vegetables, soups, spaghetti, catsup, sauces	1-5	- 3%
Beer, wine	1-5	- 3%, - 20% CO_2 or SO_2
Canned fruit	5-15	- 3%
Dried foods	5-15	+1%
Carbonated soft drinks, fruit juices	10-40	- 3%
Oils, shortenings, salad dressings, peanut butter	50-200	+10%
Jams, jellies, syrups, pickles, olives vinegar	50-200	- 3%

4. OXYGEN BARRIERS

Numerous barrier materials are used in the industry and many are being developed. This section presents a review of several most relevant barrier materials that are or will potentially be used for retortable package materials.

4.1. Polyvinylidene chloride (PVDC)

The copolymer of vinylindene chloride with vinyl chloride was first called Saran by the Dow Chemical Company (Hanlon, 1992). The name, Saran, is now used for any form of polyvinylidene chloride film or coating. PVDC film (normally co-polymerized with 30-50% of vinyl chloride) is soft and transparent. Saran has an excellent barrier properties compared to other films (Table 2), cost about 12 cents per 1000 in^2. Saran films can be stretched extensive, making them ideal as household wrapping films (thus came the name of Saran Wrap).

Although FDA approves the use of most Saran films as food packaging materials, Japan recently banned PVDC coated films because of suspected negative effects on human health and environment (Jahromi and Moosheimer, 2000). In particular, chlorine presented in PVDC and polyvinyl chloride (PVC) may lead to formation of toxic dioxins on combustion in solid waste disposal (Lange and Wyser, 2003). The solid waste treatment experts at NASA (John Fisher, Lead for Solid Waste Processing Element in Advanced Life Support Division) suggests to avoid chlorinated package materials, and advocate the use of package materials made from high value regenerative elements such as carbon, hydrogen and oxygen in future long-duration manned missions (Email communications with Michele Perchonok, 2004).

4.2. EVOH

Ethylene vinyl alcohol copolymers (EVOH) is a copolymer of ethylene and vinyl alcohol. The structure of EVOH comprises of highly ordered crystalline regions, which provide the high barrier property, and glassy amorphous regions, giving EVOH films flexibility. In a dry state, (EVOH) with 25-45 mole % ethylene are excellent barriers to O_2 (Zhang *et al.*, 2001). A major drawback of EVOH films in food applications is their hydrophilic nature. EVOH absorbs moisture at high relative humidity and in retorting conditions. It was believed that the absorbed water molecules in EVOH interact with the OH groups of the polymer matrix and weakens the hydrogen bonds between polymer changes. This enhances the motion of polymer segments, thus changing the polymer mechanical and barrier properties (Zhang *et al.*, 1999 and 2001).

4.2.1. Glass transition temperature of EVOH

A polymer in a glassy state has better barrier properties than in an amorphous state. When the glass transition temperature T_g of a polymer drops below the storage temperature, the permeability of the package material will increase, thus shortening the shelf-life of stored products. Fully exposed commercial EVOH films with 32-44 mole% ethylene (EF-F15, EF-XL15 and EF-E15 from EVAL) absorb about 3% H_2O (dry basis) at relative humidity (RH), and 6–8% of H_2O in a 100% environment (Zhang *et al.*, 1999). The glass transition temperature of the films drops from 55-62°C to 20°C when the RH increases from 0 to 75%.

4.2.2. O_2 and H_2O permeability of EVOH

Zhang *et al.* (2001) reported that under a given test condition, 15 μm EVOH films containing 32 mol % ethylene (EF-F15 and EF-XL15) have an O_2 transmission rate (0.2-1 cc/m^2-day-atm, depending upon temperature) about 1/7 that of 15 μm EVOH films with 44 mol% ethylene (EF-E15) in a RH range between 0-60%. The O_2 transmission rate (O_2TR) decrease with RH from 0 to 35% at all three tested temperatures (15, 25 and 35°C), and then slightly increase with RH from 35 to 60%. The O_2TR increases sharply after RH increases beyond 75% which corresponding to a T_g of room temperature for EVOH.

The O_2TR of the biaxially oriented film (EF-XL15) is not different from that of non oriented film (EF-F15) in the RH range between 0 and 60%, while the oriented film has slightly better O_2 barrier properties at higher RH. 15 μm EVOH films reduced their water barrier property (e.g., 0.3 – 2 g/m^2 –day-atm, depending on temperature, to 1-10 g/m^2 – day-atm) as RH increases from 0 to 90% (Zhang *et al.*, 2001).

In a recent report (Kucukpinar and Doruker, 2004) suggests that moisture molecules in EVOH films form hydrogen bonds with the side groups of the EVOH chains and disrupt the hydrogen bonding among the polymers. The increased moisture in EVOH film also enlarges voids among the polymer chains in the amorphous zones. The micro-structure changes lead to increased permeability to water vapor and O_2.

The degree of crystallinity in an EVOH film of a given % mol ethylene can be changed during re-crystallization through orientation and/or heat treatment. For example, for an EVOH film of 32 mol% ethylene, the crystallinity may vary from 27% for a none-orientation film with no heat treatment to 58% with 140°C heat treatment, and to 70% for a biaxial orientation film with 140°C heat treatment. The degree of crystallinity does not

influence O_2 transmission rate (OTR) at 0%RH (0.01 cc-mil/in^2–day–atm). But at 100% RH, the OTR decreases from ~3 cc-mil/in^2–day–atm to 0.2 cc-mil/in^2–day–atm as the crystallinity increases from 25% to 70% (Armstrong, 2000).

4.2.3. EVOH in package materials.

As stated in the above sections, properties of EVOH films are influenced by moisture. In addition to their sensitivity to moisture uptake, EVOH copolymers do not have good compatibility (adhesion and miscibility) with other polar or non-polar polymers (Lagaron *et al.,* 2003). As a result, EVOH films are usually sandwiched by coextrusion in multilayer structures in which the inner and outer layers are hydrophobic polymers, such as polypropylene and PET, for retortable pouches or trays (see table 3). EVOH films may be difficult to laminate and may delaminate during a retort process (personal conversation with Kapak Marketing Director). Special surface treatments were developed to help the adhesion and maintain integrity of the films during retorting processes. Co-extruding EVOH films with other materials may be a less expensive method to produce large quantity of package films (personal conversation with Curwood, WI).

Attempts have also been made to blend EVOH with other polymers. In a recent study, Lagaraon *et al.* (2003) reported on the effect of blending amorphous polyamide with EVOH. The blending did not improve moisture nor O_2 barrier properties, compared with pure EVOH films.

A systematic study on the influence of thermal processing on O_2 ingress in EVOH laminated trays was reported by Zhang *et al.*, (1998). The trays used in the experiments consisted of PP/PE scrap/tie/EVOH (25 μm) /tie/scrap/PP (total 367 μm or 14.5 mil) (200 g capacity) or PP/PE/scrap/tie/EVOH (36 μm/tie/scrap/PP (total 387 μm or 15.2 mil) (256 g capacity). The major findings are: 1) O_2 ingress increased linearly with retorting time – the rate of increase was constant; 2) higher process temperature resulted in high rate of O_2 ingress in package; 3) processing with N_2/62% steam at 121°C resulted in 1/10 of O_2 ingress compared to air/62% steam (~0.001 cc O_2/ min-package vs 0.01cc O_2/min-package); and 4) storage temperature greatly influenced O_2 ingress (e.g., ~0.002cc O_2/day-package at 21.1°C and 60% RH vs ~0.018 cc O_2/day-package at 32.2°C and 75% RH for trays processed at 126°C for 120 min).

EVOH is commonly used in rigid polymeric containers for thermally processed foods. The thickness of retortable pouches (~4 mil) is, however, much smaller than that of rigid containers. The thin PP, PET or nylon films protecting EVOH in those pouch materials would allow large amount water (20-30% of the solid weight of package materials) to be absorbed in EVOH during long retort times. The high moisture content would reduce the melting temperature of the EVOH film, causing it to melt at retorting temperatures. After retorting and upon cooling, EVOH would re-crystallize, expelling the absorbed water. Too tight a material over the EVOH film would cause blisters to form and make the film opaque and degrade the barrier properties. There is a delicate balance in designing retortable EVOH pouch films. The success of using EVOH in pouch materials to a large extend depends upon the thermal processes. EVOH/PP laminated pouch materials are used in commercial applications in Japan, but only used for 3 oz retorted pet foods in the USA (Curwood, MN). According to EVALCAL (Rober Armstrong, July, 2004) the retort times used in the USA is generally much longer than in Japan, making it difficult to meet the requirements. But the parent company of

EVALCAL, Kuraray, has recently developed a retort films (XEF-630) (Table 2) that can meet the shelf-life requirements (1-1.5 year) for most retail products. Kuarary is working on extraction tests before seeking FDA approval of the package materials. New processing technologies such as high pressure and microwave sterilization use much shorter process times. With modification to the current structure of XEF-630 and using the new processing technologies with much shorter process times, it may be possible to produce high quality products with 3-5 year shelf-life.

4.3. Liquid Crystal Polymers (LCP)

Liquid crystal polymers (LCP) are a class of polymers with a high level of crystallity formed by linear polymer chains (Lusignea, *et al.*, 1999). Polyester is the most used backbone for LCPs. LCPs have unique electric properties: the dielectric properties are constant over a wide range of frequencies and are not sensitive to moisture. This has made LCPs very attractive materials as high frequency substrates in circuit applications. LCPs also have excellent gas and moisture barrier properties. Flodberg *et al.* (2001) listed Vectra A950, a LCP film produced by Ticona, as one of the best barrier polymers at the time of their report. It is a co-polyester consisting of 73 mol% p-hydroxybenoic acid and 27 mol% 2-hydroxy-6-naphthoic acid, with a density of 1400 kg/m^3 and a melting point of about 280°C. Schut (2001) reported on four Ticona Vextran LCP films. From the both reports, it appears that the permeability of LCP films to O_2 and water vapor is comparable to that of dry EVOH film. Schut (2001) also reported that Ticona had developed PP/LCP laminated film for retortable food pouches. It was a single co-extruded PP-tie-LCP (5-10 μm)-tie-PP stiff film, and FDA has given a green light on Feb 21, 2001 for use in USA.

LCPs are expensive. In pure form, the cost of LCP ranged between $26 and $33/kg (Lusignea *et al.*, 1999) which is significantly higher than all other polymers, e.g., (~$1.0/kg, Hanlon, 1992), PET ($1.3-1.6/kg), PVDC ($2.8-3.3/kg) and EVOH (~ $4.8/kg) (Osborn and Jenkins, 1992). The high cost of LCP discourages packaging companies from using it for retortable package materials. In addition, Dr. Shepherd from Ticona (personal conversion, 2004) stated that thin LCP films are difficult to handle and tear easily. The laminated structure can help to overcome this problem. But the adhesion for the tie in the laminated structure is relatively weak. Non-traditional adhesions can be used, but obtaining FDA approval of these adhesions can be a daunting task. Because of many technical challenges and shifting of company priorities, some of the earlier active companies in this field, such as Superex Polymer Inc. (Waltham, MA) and Ticona, have decided to stop R&D activities related to LCP in food applications (personal conversation with Jim Shepherd, Ticona, July 23, 2004). The best that Ticona was able to provide WSU was a10 mil thick film, consisting of HDPE/tie/LCP Vectran V300P (5 μm)/tie/HDPE, which is not suited for retort applications.

4.4. Aluminum foil

Aluminum foil of >0.7 mil is almost impermeable to moisture and gases. Thinner foil has pinholes that make it slightly permeable. The chance of finding a pinhole in size ranging from 0.0000001 to 0.00003 in^2 in a one ft^2 foil is about 15% for 0.7 mil and 8% in 1 mil thick foil (Hanlon, 1992), which leads to a moisture vapor transmission rate of 0.03 g/100 in^2 –day (at 100% RH and 100 F) for the 0.7 mil foil and close to 0 for the 1 mil foil.

Military ready-to-eat meal (MRE) uses aluminum foil of about 0.35 or 0.7 mil thick in a laminated structure (with ~0.5 mil in polyester and 3 mil polypropylene) to provide O_2 and H_2O barrier (see Table 2). The pouches can stand retort temperatures of up to 135C. The specific weight of MRE foil films is 133 g/m^2 for the 4 mil thickness film (NASA silver pouch bag) and 163 g/m^2 for the 6 mil thickness film (NASA brown pouches), which is significantly heavies than 112 g/m^2 for the 4.1 mil SiO_x coated film (Alcan 17000) and 117 g/m^2 for the 4.7 mil B-Pack EVOH film (100 μm PP/EVOH 20 μm). But most importantly, aluminum foils present difficulty in solid waste disposal for future long-duration manned space missions. There is also a general desire, especially in Europe, to avoid commercial use of aluminum foils because of the difficulty in waste disposal and consumers' perception of high energy use in making aluminum foil (Lange and Wyser, 2003).

Packaging films containing aluminum foil or a metalized layer of polymer film are not suited for microwave sterilization because the metal sheet prevents electromagnetic energy from penetrating into the packaged foods. High pressure processes can compromise the integrity of those materials. For example, visible signs of delamination were observed between the polypropylene (PP) and aluminum (Al) layers in MRE pouches processed at >200 MPa at 90°C (Schauwecker *et al.,* 2002). High pressure processes (600-800 MPa for 5-10 min at 40-60°C) also caused structure damages to metalized PET films (Caner *et al.*, 2003) and significantly reduce barrier properties of those package materials (Caner *et al.*, 2000).

4.5. Silicon Oxide (SiO_x) Coating

Early development of high barrier coated polymeric films started in 1959 with aluminum metallization techniques. In 1980's, transparent SiO_x coated films were developed as alternatives to metalized plastics (Leterrier, 2003). Similar to aluminum metallization, deposition of a thin layer of low permeability SiO_x on a thermoplastic substrate sharply improves barrier properties. Many deposition methods have been developed over the years, including sputtering, electron-beam deposition and plasma-enhanced chemical vapor deposition (PECVD) (Erlat *et al.,* 2000). PECVD was developed in early 1990s and used extensively in the microelectronics industry to deposit silicon dioxide on thin films as electric insulation. It can deposit SiO_x below the glass transition temperature of the substrate polymers and has the advantages in adhesion and step coverage of silica coating. This makes the silica coating more flexible and resistant to cracking during the converting processes, namely lamination printing and making pouches for food applications (Teshima *et al.*, 2003).

Typical thickness of the SiOx is between 10-50 nm, and the polymer substrate between 12 and 25 μm. Thicker layer becomes brittle (Hedenqvist and Johansson, 2003). The permeability of these laminated films for O_2 is about 0.3-0.5 cc/m^2 –day –atm, which is very low compared to that for most polymer films, but still several order of magnitude higher than that of silica glass (Roberts *et al.*, 2002). Many papers suggest that defects exist in the coating, creating pathways for O_2. A recent study has found that that gas permeates directly through oxide matrix, suggesting a quite different lattice structure of SiO_x than annealed silica glass, with the former being a more open structure (Erlat *et al.*, 2000). Roberts *et al.* (2002) used mathematical models to study the influence of three different possible types of defects, namely macro (>1nm), nano (0.1-1 nm) and lattice

(0.2-0.3 nm), on permeability of O_2. Their study indicates the presence of macro-defects, but those defects are small and rare, and could not be detected by atomic force microscopy (AFM). ATM images show clearly irregular surface array of columnar grain-like structures likely caused by growth around isolated nucleation centers during the coating in a vacuum deposit coater.

The S_iO_x coat is, however, fragile. Hedenqvist and Hohansson (2003) studied the effects of 90 degree folding of S_iO_x coated films and found that 1-2 times folding initiated cracks in the coating and increased O_2 permeability by 13-74 %. Preliminary tests in my laboratory show that the O_2 transmission rate of the films increased four times after retorting, from 1.2 cc/m^2-day-atm to 5.0 cc/m^2-day-atm (Table 2). Packaging companies have also expressed concern with regard to the converting processes (making laminated package materials from the films) in that cracks may develop. For example, the economic losses might be very large, if 0.5 million pouches are produced with defects that are not detected (personal conversation with CURWOOD representatives).

4.6. AlOx and AlO_xN_y Coating

Thin transparent Aluminum oxide coatings are also used to improve barrier properties of package films. Recent studies (Erlat *et al.*, 2004) have shown that inclusion of nitrogen in AlO_x coating offers a significant improvement of O_2 and H_2O properties in those films. However, the plasma enhanced chemical vapor deposition (PECVD) coating method commonly used for AlO_x and SiO_x requires the use of silane as a precursor, which makes it unattractive in commercial production of the films. Erlat's group at Oxford University developed a reactive magnetic sputtering method to overcome the difficulty. The O_2 transmission rate for the PET (50 μm)/AlO_xN_y films decreased from ~50 cc/m^2-day to 1 cc/m^2 –day as the coating thickness increased from 0 to 80 nm (Erlat *et al.*, 2004). The still relatively high O_2 transmission rates were attributed to a significant numbers of micron-scale defects that that provide permeation pathways through the films. Water vapor transmission rate decreased from 1 g/m^2-day to 0.1 g/m^2-day, as the coating thickness increased from 20 to 80 nm.

AlOx coated films suffer from the same fragility problem as SiO_x coated films. A recent study at Illinois Institute of Technology shows that the O_2 transmission rate increased from 0.54 cc/m^2-day-atm to 25 cc/m^2-day-atm after a high pressure process (preheated to 90°C, and processed at 688 MPa), a 46 fold increase (Table 2).

4.7. Nanocomposites

Nanocomposite films incorporate nano particles, typically 100-1000 nm, in a polymer matrix to increase tortuosity of diffusion pathway and, thus, increase barrier properties (Lange and Wyser, 2003). Those particles can be inorganic clays or LCP. The difficulty of making those films lies in dispersion of those particles in the matrix and difference in their melting points. To the best of the author's knowledge, noncomposite films not are used in commercial food applications.

Table 2. Properties of commercial package films/trays

Barrier	Film description	O_2 Transmission Rate (cc/m^2-day) Before Retort	O_2 Transmission Rate (cc/m^2-day) Post Retort	H_2O Transmission Rate (gm/m^2-day)	Description of the films outside -- inside	Source of information
EVOH	Combitherm	0.51, measured at 24 C and 50% rh	Process condition: 121C and 30 min	3.1	Nylon/EVOH/nylon/LF adhesive/HV PE/LLDPE	NASA Food Lab
	B-Pack Film	≤ 0.7, measured at 23 C and 0% rh		<15, measured at 38C and 90% R.H.	PP/PA/EVOH retortable 10 micron/PA/PP	B-Pack, S.p.A.-Italy
	Eval XEF-630 -- new product (May, 2004)	0.3 at 20C and 85%	0.4 -- after retort at 120C and 30 min with water		Barrier/ Ony (0.6mil)/CPP (2.0 mil)	Kuraray, Con., LTD, EVAL Company
SiOx	Alcan 17000	<1.55		<2.33		Alcan specification
	Alcan 17000	1.2	5.0		Polyester (0.48 mil)-SiOx (0.6) – PP (3)	Juming Tang, WSU
AlOx	Pyramid 4381	0.54	25 (post high pressure process)		N/A	Tiariana N. Koutchma, IIT
LCP	Vectra A950 Time in compression moulding					Film from Ticona, data by Flodberg *et al.*, 2001
	8 min	0.45 (6.5 mil thick)				
	30 min	0.2 (5.4 mil)				
Foil Pouches	MRE	0.06		0.01	N/A	Natick

5. EMERGING PROCESSING TECHNOLOGIES

Over the past five years, major processing technology developments have been supported by the US Army Natick Soldier Center and US Department Defense to provide high quality shelf-stable foods for the military. The required shelf-life for shelf-stable military rations is three years at 80°F (26.6°C). This is not much different from the requirement for NASA long duration mission programs. The two most promising technologies for shelf-stable food products for long- duration manned space missions are: 1) 915 MHz Single Mode Microwave Sterilization Technology; 2) High Pressure/Thermal Processing Technology. The following sections provide brief introductions to these two technologies and their affect package barrier properties.

5.1. WSU 915 MHz Microwave Sterilization Technology

Microwave heating is a result of the polarization effect of electromagnetic radiation on foods at frequencies between 300 MHz and 300 GHz. Microwaves can interact directly with foods to generate heat in hermetically sealed polymeric containers. Microwave sterilization has an advantage over retorting of canned foods because of the short heating time and potential for more uniform heating. Commercial microwave sterilization

processes are now used in Belgium (Tops Foods, Belgium) and Japan (Otsuka Chemical Co., Osaka, Japan). Products from Tops Foods and Otsuka Chemical Co. demonstrate that microwave sterilized products containing pasta, rice, and meats have better organoleptic quality and appearance than frozen products.

No commercial microwave sterilization systems are used in North America. The 2,450 MHz microwave sterilization technology used in commercial applications in Europe and Japan could not be adopted in the USA due to more stringent FDA requirements. A major technical problem with the 2,450 MHz microwave systems is the unpredictable cold spot in the food packages. As a result, the processed foods require 100% incubation before being released to market. To overcome this problem, the Advanced Thermal Processing Technology Team at Washington State University (WSU) used 915 MHz microwave and developed signal-mode sterilization concept (Pathak *et al.*, 2003). The WSU design combines microwave heating with circulating water at 121°C to shorten process time (Guan *et al.*, 2002). A Microwave Sterilization Consortium was formed in 2001 with support from the US Department of Defense Dual Use Scientific and Technology (DUST) Program, Washington State University and it's industrial partners. The aim of this consortium is to develop and scale-up the 915 MHz single-mode microwave sterilization technology for industrial processing of shelf-stable package foods for military and civilian uses. WSU Microwave Sterilization Consortium consists of US Army Natick Soldier Center and eight companies, including Kraft, Masterfoods, Hormel, Ocean Beauty Seafoods, Truitt Brothers, Rexam Container, Graphic Packaging, and Ferrite Component. The Technical Service Center of National Food Processors Association (Dublin, CA) serves as a technical advisor on microbial safety and FDA approval. Over the past three years, the consortium has addressed several major technical issues, including developing a pilot-system for demonstration and engineering studies, developing techniques for locating cold spots in packaged foods (Lau *et al.*, 2003), monitoring temperature during MW processes and microbial challenge studies (Guan et *al.*, 2003). Our newly developed pilot-scale system can complete a thermal process for 7 oz trays ($F_o = 6$) in 5-8 min, a significant reduction in processing times compared to the conventional processes of approximately 30 min. With the system, the WSU team and industrial partners tested on some heat sensitive products that could suffer significant quality losses in conventional retort systems (e.g., Guan *et al.*, 2002). The new technology shows promises for producing very high quality fish, diary, pasta, and meat products. During the June 22-23 2004 consortium meeting at WSU, the consortium decided that we are ready to contact FDA in developing approval documents and that we should move ahead to scale-up the processes for industrial applications. It may still take approximately 2-4 years before commercial application of FDA approved industrial systems. But the WSU pilot-scale system, designed to mimic industrial operations, is available for product and packaging feasibility studies.

MRE foil pouches shields electromagnetic fields from reaching food in packages and, therefore, are not suited for microwave sterilization processes. At Washington State University, we use pouches made of SiO_x coated films or polymeric trays with EVOH barrier and Seran based lid stock. The O_2 transmission rate for both the Seran based lid stock and the SiO_x coated films after the post process is approximately 2 cc/m^2-day (Table 2). Rexam Containers and EVALCAL are very excited about the short process times (5-8 min) when using our microwave sterilization system in connection with

EVOH package films. Two unique design features of the WSU Microwave Sterilization Technology offer opportunities to maintain the integrity of EVOH based films: 1) water immersion during the MW heating reduces the partial pressure of O_2 and thus reduces the gas ingress; 2) the short exposure times of food packages reduces moisture migration into package materials and, thus, eliminates the problem associated with melting down of soaked EVOH films.

5.2 High Pressure Process (HPP)

High pressure processes (HHP) refer to a novel food preservation technology that uses high pressures >300 MPa to inactive pathogenic and spoilage micro-organisms in foods. HHP is effective in inactivating most vegetative pathogens, and are now used worldwide in commercial operations to extend the shelf-life of many commodities, including orange juice, avocado pulp, sliced ham, meats, oyster, salsas, and guacamole. A major advantage of HHP is that it does not rely on thermal energy or use little thermal energy to inactive most of food pathogens, thus help to retain most quality attributes of the processed foods. HHP sterilization for low-acid (pH>4.5) foods is still under development (Sizer *et al.*, 2002), because HHP alone is not adequate in inactivating spores that pose great risk in low-acid shelf-stable foods. Current research and development activities related to HHP sterilization processes relies on the adiabatic heating of foods via high pressure from a relatively high initial product temperature (e.g., 90°C) to a final temperature of 121°C (Morris, 2001). This process is commonly referred to as a thermally-assisted HHP process. Several technique issues remain to be addressed before HHP sterilization processes can be approved by FDA and implemented by the food industry: 1) selection of the most resistant food pathogen spore as the target bacterium to design a HHP process; 2) selection of a surrogate that is non pathogenic micro-organism and more resistant than the target bacterium for process development and validation; and 3) identification of the least treated location in foods during HPP processes for process development.

In order to speed up the technology development of HHP sterilization technology for low acid foods, a consortium led by US Army Natick Soldier Center was formed in 2000 with support from the DoD Dual-Use Scientific and Technology Program. Two active consortium members are the National Center for Food Safety and Technology at Illinois Institute of Technology (Summit-Argo, ILL) and Avure, a subsidiary of Flow International (Kent, WA). In addition, the National Food Laboratory of National Food Processors Association serves as a subcontractor for the microbial studies and is conducting a research to study the kinetics of C. *botulinum* spores under HHP conditions. It is anticipated that the HHP Consortium will be ready by the end of 2004 to approach FDA in preparation for petition document (personal email communication with Dr. Pat Dunne, US Army Natick Soldier Center, MA). HHP is suited for food products that have little air voids and do not undergo undesirable texture changes under high pressure.

Delamination is often observed in MRE foil pouch materials after HHP, while transparent SiO_x coating fractures and becomes translucent under pressure. Laminated EVOH films are not affected by HHP in pasteurization processes and are currently used for many HHP pasteurized foods, including guacamole (personal communication with Rober Armstrong, EVALA, TX).

Table 3 summarizes advantages and limitation of the films discussed in this section.

6. SUMMARY AND DISCUSSIONS

The technologies for LCPs and nano-clay/polymer films are still immature to allow the use of those films as food packages in long-duration manned missions. There are significant technical hurdles to be overcome, and it is difficulty to predict when some of those films may be commercially available for retort applications. There appears to be large momentum in developing transparent films coated with SiOx, $AliO_x$ or TiO to provide unique barrier properties. Films of this type are commercially available, but can still not meet the NASA requirements in terms of post-process barrier properties. EVOH films in laminated structures appear to be most promising in providing the required shelf-life for packaged products. In addition, EVOH films contain basic elements of C, O and H, all of high regenerative in space missions, and can serve as a source for recovery of O_2 and water.

MRE pouches currently used in NASA and the US Army can hold 8 oz (226 g) moist products. An 8 oz capacity pouch has approximately 0.052 m^2 surface area. The upper limit for the maximum allowable O_2 in most sensitive food category (e.g., fish, meat, dairy, and poultry) is 5 ppm which corresponds to about 0.8 cc of O_2 at standard conditions (Table 1). For a 3 yr (1092 days) self-life, the maximum rate of O_2 ingress into the package should be less than:

$$O_{2,ir} = \frac{0.8cc}{1092day * 0.052m^2} = 0.014cc/m^2 - day$$

The transmission rate for package materials should be less than:

$$OTr = \frac{O_{2,ir}}{\Delta P_{O2}}$$

where, ΔP_{O2} (in atm) represents the difference in partial pressure of O_2 across the package material. In dry atmospheric air, ΔP_{O2} is 0.2095 atm. This value is reduced in humid air. So for a conservative estimation, we would require that the package material provides an OTr of 0.067 cc/m^2-day-atm (0.0043 cc/100 in^2-day-atm).

It is clear from the above calculation and Table 2 that none of the current non-foil package materials meet the requirements. It may, however, be possible to use the synergistic effects of good processing technologies (shorter processing times) and right storage conditions (reduced partial pressure of O_2 and controlled temperature) to enhance the performance of EVOH laminated package materials (and thicker barrier films) to provide 3-5 year shelf-life for thermally processed products.

Table 3. Summary of options of package materials in future space missions

Package material	Mechanical properties	Environmental	Barrier property	Commercial readiness	Compatible with MW and HHP
Foil/metalized	Good	Poor	Good	Good	Poor
PVDC	Good	Poor	Fair	Good	Good
EVOH	Good	Good	Fair	Fair	Good
SiOx coated	Poor-fair	Good	Fair	Fair	Poor-fair
LCP	Poor	Good	Good	Poor	N/A
Nano	N/A	Fair	N/A	Poor	N/A

REFERENCES

Armstrong, R.A. 2002. Effects of polymer structure on gas barrier of ethylene vinyl alcohol (EVOH) and considerations for package development. *TAPPI 2002 PLACE Conference.*

Brody, A.L. 2003. Predicting packaged food shelf-life. Food Technology 57(4):100-102.

Caner, C., Hernandez, R.J., Pascall, M.A., Riemer, J. 2003. The use of mechanical analyses, scanning electron microscopy and ultrasonic imaging to study the effects of high temperature processing on multilayer films. *J. of the Sciences of Food and Agriculture* 83:1095-1103.

Caner, C., Hernandez, R.J. and Pascall, M.A. 2000. Effect of high-pressure processing on the permeance of selected high-barrier laminated films. *Packaging Technology and Science* 13:183-195

Erlat, A.G., Henry, B.M., Grovenor, C.R.M., Briggs, A.G.D. 2004. Mechanism of water vapor transport through PET/AlO_xN_y gas barrier films. *J. Phys. Chem.* 108:883-890.

Erlat, A.G., Wang, B.C., Spontak, R.J., Tropsha, Y., Mar, K.D., Montgomery, D.B., Vogler, E.A., 2000. Morphology and gas barrier properties of thin SiO_x coating on polycarbonate: correlations with plasma enhanced chemical vapor deposition conditions, *J. Mater. Res.*, 15:704.

Guan, D., Gray, P., Kang, DH, Tang, J., Shafer, B., Ito, K., Younce, F., and Yang, C.S. 2003. Microbiological validation of microwave-circulated water combination heating technology by inoculated pack studies, *J. Food Sci.* 68(4):1428-1432.

Guan, D., Plotka, V. C. F., Clark, S., and Tang J. 2002. Sensory evaluation of microwave treated macaroni and cheese. *J. Food Processing and Preservation*, 26:307-322.

Flodberg, G., Axelson-Larsson, L., Hedenqvist, M.S., Gedde, U.W. 2001. Liquid crystalline polymer pouches for local anaesthetic emulsion. *Packaging Technology and Science* 14:159-170.

Hanlon, J. F. 1992. *Handbook of Package Engineering* (2nd Ed.). Technomic Publishing Company, Inc., Lancaster, PA.

Hedenqvist, M.S., Johansson, K.S., 2003. Barrier properties of SiO_x –coated polymers:multi-layer modeling and effects of mechanical folding, *Surface and Coatings Technology*, 172:7-12.

Jahromi, S. and Moosheimer, U. 2000. Oxygen barrier coatings based on supramolecular assembly of melamine. *Macromolecules* 33:7582-7587.

Kucukpinar, E. and Doruker, P. 2004. Effect of absorbed water on oxygen transport in EVOH matrices: a molecular dynamics study. Polymer 45:3555-3564.

Lagaron, J.M., Gimenez, E., Altava, B., Del-Valle, V. and Gavara, R. 2003. Characterization of extruded ethylene-vinyl alcohol copolymer based barrier blends with interest in food packaging applications. *Macromol. Symp.* 198:473-482.

Lang, J. and Wyser, Y. 2003. Recent innovations in barrier technologies for plastic packaging – a review. *Package Technology and Science* 16:149-158.

Lau, H., Tang, J., Taub, I.A., Yang, T.C.S., Edwards, C.G. and Mao, R. 2003. Kinetics of chemical marker formation in whey protein gels for studying high temperature short time microwave sterilization. *J. Food Engineering* 60:397-405.

Leterrier, Y. 2003. Durability of nanosized oxygen-barrier coatings on polymers. *Progress in Materials Science* 48:1-55.

Lopez, A. 1987. *A Complete Course In Canning and Related Processes.* Vol. 2. Packaging, Aseptic Processing, Ingredients, 12th edition, The Canning Trade Inc., Baltimore, MD.

Lusignea, R.G. 1999. Orientation of LCP blown film with rotating dies. Polymer Engineering and Science, 39 (2):2326-2334.

Massey L.K. 2003. Permeability properties of plastics and elasttomers.

Osborn, K.R. and Jenkins, W.A. 1992. *Plastic Films: Technology and Packaging Applications.* Technomic Publishing Company, Inc., Lancaster, PA.

Morris, C.E. 2001. Thermally-assisted high-pressures lifts quality of shelf-stable foods. *Food Engineering, Sept issue* - 2001.

Pathak, S, Fen, L., and Tang, J. 2003. Finite difference time domain simulation of single-mode 915 MHz cavities in processing pre-packaged foods. *J. Microwave Powers and Electromagnetic Energy* 38(1): 37-48.

Perchonok, M. 2002. Shelf-life considerations and techniques. Food Product Development Based on Experience. Catherine Side (Ed.), Iowa State Press, Ames, Iowa.

Roberts, A.P., Henry, B.M., Sutton, B.M., Grovenor, C.R.M., Briggs, G.A.D., Miyamoto, T., Kano, M., Tsukahara, Y., and Yanaka, M., 2002. Gas permeation in silicon-oxide/polymer (SiO_x/PET) barrier films: role of the oxide lattice, nano-defects and macro-defects. *J. Membrane Science,* 208:75-88.

Schauwecker, A., Balasubramaniam, V.M., Sadler, G., Pascall, M.A., Adhikan, C. 2002. Influence of high-pressure processing on selected polymer materials and on the migration of a pressure-transmitting fluid. *Packaging Technology and Science* 15:255-262.

Schut, J.H. 2001. Materials close-up: LCPs break new ground in film co-extrusion and thermoforming. Plastics Technology On-line Article, www.plasticstechnology.com/articles/200104cu2.html

Sizer, C.E., Balasubramaniam, V.M., and Ting Ed. 2992. Validating high-pressure processes for low-acid foods. Food Technology 56(2):36-42.

Teshima, K., Sugimura, H., Inoue, Y., Takai, O. 2003. Gas barrier performance of surface-modified silica films with grafted organosilane molecules. *Langmuir* 19:8331-8334.

Zhang, Z.B., Britt, I.J., Tung, M.A. 1998. Oxygen ingress in plastic retortable packages during thermal processing and storage. *J. Plastic Film and Sheeting* 14:287-307.

Zhang, Z.B., Britt, I.J., Tung, M.A. 1999. Water absorption in EVOH films and its influence on glass transition temperature. J. Applied Polymer Science: Part B: Polymer Physics 37: 691-699.

Zhang, Z.B., Britt, I.J., Tung, M.A. 2001. Permeability of O_2 and H_2O vapor through EVOH films as influenced by relative humidity. J. Applied Polymer Science 82:1866-1872.

ATTPENDIX

Definitions/Equations

Transfer of mass through a package film can be described by the following equation (Brody, 2003):

$$\frac{dM}{dt} = \frac{k}{l} A\Delta P \qquad \ldots. (1)$$

where, M is the mass or volume of a gas transmitted through the film (g or cc), t is time (day), k is the **permeability** of the film material for the gas under consideration (cc-mil/m^2-day-atm), A is the area of the film (m^2), l represents the thickness of the film (mil), and ΔP is the difference in partial pressure of the gas across the film (atm). In our ambient environment, ΔP is 0.209 for O_2 and 0.0003 for CO_2 (Brody, 2003).

Transmission rate, Tr (cc/m^2–day-atm), is defined as:

$$Tr = \frac{k}{l} \qquad \ldots (2)$$

From Eq. (2):

$$k = Tr \times l \qquad \ldots (3)$$

Permeability reflects the intrinsic barrier property of one material, while transmission rate represents the barrier property of the film. The latter is inversely proportional to the thickness of the film (see Eq. 2).

Calculation of O_2 mass from volume:

Molar mass of O_2 is 32, 1 mole of an ideal gas at 1 atm and 20 °C is ~22,400 cc. So 32g O_2 occupies 22,400 cc volume.

The maximum allowable O_2 in most sensitive foods is 5 ppm (Table 1). The maximum allowable volume of O_2 in a 200 g food package is thus:

$$V_{O2} = \frac{5x10^{-6} * 226g * 22{,}400cc}{32g} = 0.79cc$$

D-Side: A Facility and Workforce Planning Group Multi-criteria Decision Support System for Johnson Space Center

Final Report

Faculty Fellowship Program - 2004

Johnson Space Center

Prepared by:	Madjid Tavana, Ph.D.
Academic Rank:	Professor
University & Department:	La Salle University Management Department Philadelphia, PA 19141
NASA/JSC	
Directorate:	Mission Operations Directorate
Division:	Advanced Operations Development Division
Branch:	Operations, Research, & Strategic Development Branch
JSC Colleague:	Anthony C. Bruins
Date Submitted:	July 28, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

"To understand and protect our home planet, to explore the universe and search for life, and to inspire the next generation of explorers" is NASA's mission. The Systems Management Office at Johnson Space Center (JSC) is searching for methods to effectively manage the Center's resources to meet NASA's mission. *D-Side* is a group multi-criteria decision support system (GMDSS) developed to support facility decisions at JSC. *D-Side* uses a series of sequential and structured processes to plot facilities in a three-dimensional (3-D) graph on the basis of each facility's alignment with NASA's mission and goals, the extent to which other facilities are dependent on the facility, and the dollar value of capital investments that have been postponed at the facility relative to the facility's replacement value. A similarity factor rank orders facilities based on their Euclidean distance from Ideal and Nadir points. These similarity factors are then used to allocate capital improvement resources across facilities. We also present a parallel model that can be used to support decisions concerning allocation of human resources investments across workforce units. Finally, we present results from a pilot study where 12 experienced facility managers from NASA used *D-Side* and the organization's current approach to rank order and allocate funds for capital improvement across 20 facilities. Users evaluated *D-Side* favorably in terms of ease of use, the quality of the decision-making process, decision quality, and overall value-added. Their evaluations of *D-Side* were significantly more favorable than their evaluations of the current approach.

Keywords: NASA, Multi-Criteria Decision Making, Decision Support System, AHP, Euclidean Distance, 3-D Modeling, Facility Planning, Workforce Planning.

1. INTRODUCTION

Over the past decade, NASA's budget has been the target of economizing. Like many other organizations throughout the public and private sectors, one of the most pressing problems facing the Johnson Space Center (JSC) is deciding how to allocate its resources in the face of fiscal constraints. The Systems Management Office at JSC is responsible for developing methods to optimize JSC's facilities and workforce. To do so, the Systems Management Office seeks to allocate increasingly limited resources in a way that is aligned with NASA's mission and strategic goals while also ensuring that operations at each facility are not adversely affected by operational interruptions at other facilities. Finally, the Systems Management Office seeks to develop a resource allocation process that is perceived by stakeholders as fair and free from undesirable personal or political biases.

At many organizations, decisions about allocating resources across facilities follow an annual cycle that is aligned with the organization's fiscal budgeting cycle (Gregory and Pearce 1999). Resource allocation typically begins by considering the investments that each facility would like to make in order to meet its needs and objectives. Because resources are usually limited, these potential investments are prioritized by evaluating them against a set of criteria. Texts on finance and accounting generally recommend that such decisions be evaluated using financial metrics such as payback period, net present value, or profitability index (e.g., Garrison and Noreen 2002). However, for mission-driven agencies in the public sector (such as NASA), these financial metrics are often less helpful than they are in other (e.g., private sector) settings. Whatever the criteria used to rank order facilities, resources are then allocated over the period covered by the budget cycle on the basis of each facility's relative priority. At the end of the cycle, any new or unmet needs and objectives serve as inputs in the next annual cycle.

Currently, the Systems Management Office uses a multi-criteria decision making (MCDM) model to allocate capital improvement among its facilities. A facility review team selected by JSC leadership uses the following five criteria to assess each facility with a weighted sum method:

- Mission - How much does this asset support NASA?
- Availability - How available does this asset need to be?
- Exclusivity - Can this asset be found elsewhere?
- Potential Future Need - Can there be a critical need for this asset in the future?
- Advanced Technology Development - Does this asset contribute to cutting edge research?

The weighting for the criteria are captured with the Analytic Hierarchy Process (AHP) and Expert Choice software (Expert Choice 2004). The facility review team rates each facility on the above five criteria using a 0 (asset doest not support the criterion) to 5 (asset fully supports the criterion) rating scale. Each facility's score is the weighted sum of its asset ratings across the criteria. Facilities with higher scores are considered critical and receive more funding.

As part of the workforce planning conducted annually, a staffing review team is formed by JSC leadership to determine the impact of potential changes in work based on changes in strategy and identified center goals. Budget constraints, programmatic changes, probable attritions, and development needs/interests of current staff are used to develop full-time equivalent work demand for each directorate for the fiscal year. The staffing review team then compares the work demand with actual headcount and recommends full-time equivalent numbers for each directorate. The current approach reports on current workforce supply, forecast Center

workforce requirements (demand), and assess future gaps so the Center can allocate appropriate funds to close the gap.

D-Side is a group multi-criteria decision support system (GMDSS) developed for JSC to guide funding decisions for its facilities and workforce. Group decision support systems (GDSS) are interactive, computer-based systems that help a group of decision-makers solve problems and make choices. A group decision support system should support not only group interactions but also structure decision making processes with modeling tools (DeSanctis and Gallupe 1987). Group decision making is a complex human activity and involves assessments of multiple conflicting criteria and various decision alternatives. *D-side* embeds a multi-criteria decision making model (MCDM) into its group decision support system.

Over the last several decades, a philosophy and a body of intuitive and analytical MCDM models have been developed. Schoemaker and Russo (1993) describe four general approaches to MCDM ranging from intuitive to highly analytical. These methods include intuitive judgments, rules and shortcuts, importance weighting, and value analysis. They argue that analytical methods such as importance weighting and value analysis are more complex but also more accurate than the intuitive approaches (Schoemaker and Russo 1993).

Embedding a MCDM into a GDSS requires aggregation of individual preferences into a group decision. In a loosely coupled procedure (*Parallel Coupling*), group members individually assess all alternatives based on their own preferences and reach an individual decision. At the end, individual decisions are synthesized to arrive at a group decision. In contrast, in a tightly coupled procedure (*Sequential Coupling*), group members collectively assess all alternatives based on group preferences and arrive at a group decision. According to Arrow's Impossibility Theorem (Arrow 1963), no group decision making method is perfect. This theorem is regarded as a very important work in modern social choice theory and shows that a group decision outcome can never satisfy every group member. Cao et al. (2003) show that parallel coupling tends to make group members more satisfied with process outcome while the sequential coupling tends to produce better results with respect to the decision quality and decision confidence. *D-Side* supports both parallel and sequential coupling. However, the facility and workforce planning model described in this paper will focus on sequential coupling for simplicity. Results from previous research show that groups using a GDSS outperform groups not using a GDSS or using a GDSS without a problem-modeling tool (Barkhi 2002, Fjermestad and Hiltz 2001, Lam 1997).

D-Side, a group multi-criteria decision support system (GMDSS), offers several advantages. First, *D-side* has a strong strategic focus; resource allocation decisions are guided by explicit consideration of each facility's contribution to the attainment of the organization's mission and strategic goals. Second, *D-side* employs the Analytic Hierarchy Process and Expert Choice software to determine the relative importance (i.e., weight) of each of the organization's strategic goals. Facilities that are closely aligned with the most important strategic goals are given higher priority than facilities that are closely aligned with less important strategic goals. Third, *D-side* also considers interdependencies among facilities; resource allocation decisions are influenced by the extent to which operational interruptions at each facility might adversely affect operations at other facilities. This is accomplished through a modified version of the Mission Dependency Index (developed by the Naval Facilities Engineering Service Center). Fourth, a core metric in the facility management industry, the dollar value of capital investments that have been postponed at the facility relative to the facility's replacement value, is used to compare facilities. Fifth, the *D-side* process produces a plot of the facilities in three-dimensional space,

thereby allowing decision makers to view simultaneously the rank ordering of the facilities, the relative standing of each facility on each of the underlying criteria, and facilities that are similar (i.e., that cluster together) in terms of the underlying criteria. That is, *D-side* incorporates recommendations concerning the visual display of quantitative data by enabling users to view simultaneously a large amount of multivariate and comparative information from many different perspectives (Tufte 1990, 2001). Sixth, *D-side* incorporates several tools from the industrial and organizational psychology research literature to enhance the accuracy of judgments and minimize the likelihood that judgments will be distorted by self-serving or other political biases. Seventh, on the basis of the criteria described above (i.e., alignment with strategic goals, interdependence of facilities, and relative value of postponed capital investments) *D-side* provides clear guidance about the proportion of available resources that should be allocated to each facility. Eighth, *D-Side* employs a shared methodology for both facility and workforce planning models, thereby enabling information sharing between the two models. Ninth, *D-Side* offers a set of integrated decision-support features including visualization and animation, consolidation and aggregation, modeling, what-if, and goal-seek analysis.

Section 2 of this paper presents a detailed explanation of the facilities model. Section 3 provides an overview of the workforce model. Section 4 presents guidelines for the rating process. Section 5 illustrates a hypothetical example. Section 6 describes the results of a pilot study of *D-Side* and presents users' evaluations of *D-Side* versus the current approach in terms of ease of use, the decision-making process, decision quality, and overall value-added. Section 7 presents some concluding comments and managerial implications.

2. THE FACILITIES MODEL

Two panels of expert judges are selected by JSC leadership to participate in this process: the directorate representative panel (*D-Panel*) and the facility management panel (*F-Panel*). The *D-Panel* members are Directorate representatives selected from Center Operations, Engineering, Life Sciences, Mission Operations, Space Shuttle, International Space Station, and Systems Management Office. The *F-Panel* is comprised of the facility mangers selected from various Directorates by JSC leadership. *D-Side* uses a series of intuitive and analytical methods to plot each facility in a three-dimensional (3-D) graph based on its Euclidean distance from the Ideal and Nadir points. The overall methodology is depicted in Figure 1.

Insert Figure 1 Here

Conceptually, the methodology is identical for facilities and workforce units, working through three distinct but related phases. The eight-step procedure described below is utilized to systematically evaluate each facility:

1.a. Center Priority Weights: Center Priority Weights are used to determine the Facility Priority Index (P_i) for each facility under review at JSC. P_i attempts to quantitatively score facilities based on how well they support NASA's mission and goals. To determine facility alignment with NASA's mission, each existing facility is scored against a set of strategic goals. This requires the scoring of over 250 separate facilities against 10 goals each planning period. These goals along with their respective missions identified in the 2003 NASA Strategic Plan (National Aeronautics and Space Administration 2003) are presented in Figure 2.

Insert Figure 2 Here

The *D-Panel* initially identifies the relative importance of each of the four missions (W_j; $j=1,...,4$). Then, for each mission, the *D-Panel* identifies the relative importance of each goal associated with the mission (W_{jk}; $j=1,...,4$ and $k=1,...,\ l_j$ where l_j is the number of goals associated with mission j). Next, the *F-Panel* collectively scores each facility against each goal in Phase 2 (Step 2.a).

The relative importance (i.e., weighting) for the missions and goals is captured through a series of pairwise comparisons using the AHP and Expert Choice software in a GMDSS. The *D-Panel* members are asked to provide their subjective assessment of each pairwise comparison. Saaty's AHP (Saaty and Vargas 1998, Forman and Gass 2001) uses these pairwise comparisons to derive a weight for each mission and goal.

Initially, the *D-Panel* members are asked to compare each possible pair of missions by providing their judgments about which mission is more important and by how much. Because there are four missions, this requires each *D-Panel* member to make six pairwise comparisons [$n(n-1)/2$]. Then, within each mission, the *D-Panel* members are asked to compare each possible pair of strategic goals (g_i and g_j) and provide judgments about which goal is more important and by how much. For example, three strategic goals are associated with the mission 'Understand and protect our home planet'; weighting these three goals requires three pairwise comparisons. Assuming that the *D-Panel* member is evaluating q goals, AHP quantifies these judgments and represents them in a $q \times q$ matrix:

$M = (m_{ij}) \qquad (i, j = 1, 2,....,q)$

If g_i is judged to be of equal importance as g_j, then $m_{ij} = 1$

If g_i is judged to be more important than g_j, then $m_{ij} > 1$

If g_i is judged to be less important than g_j, then $m_{ij} < 1$

$m_{ij} = 1/m_{ji} \qquad m_{ij} \neq 0$

Because the entry m_{ij} is the inverse of the entry m_{ji}, the matrix M is a reciprocal matrix. m_{ij} reflects the relative importance of goal g_i compared with goal g_j. For example, $m_{12}=1.25$ indicates that g_1 is 1.25 times as important as g_2.

Then, the vector v representing the relative weights of each of the q goals can be found by computing the normalized eigenvector corresponding to the maximum eigenvalue of matrix M. An eigenvalue of M is defined as λ which satisfies the following matrix equation:

$$M v = \lambda v$$

where λ is a constant, called the eigenvalue, associated with the given eigenvector v. Saaty (1977, 1983, 1989, 1990, 1994) has shown that the best estimate of v is the one associated with the maximum eigenvalue (λ_{max}) of the matrix M. Because the sum of the weights should be equal to 1.00, the normalized eigenvector is used. Saaty's algorithm for obtaining this v is incorporated in Expert Choice software used in this study.

The relative importance of each mission is found through a series of pairwise comparisons among missions. The relative weights of the missions (W_j; $j=1,...,4$) are the v values computed using the Matrix M, where m_{ij} is the relative importance of mission i compared with mission j. Similarly, the relative importance of each strategic goal for each mission (W_{jk}) is computed.

One of the advantages of AHP is that it encourages panel members to be consistent in their pairwise comparisons. Saaty suggests a measure of consistency for the pairwise comparisons. When the judgments are perfectly consistent, the maximum eigenvalue, λ_{max}, should equal q, the number of goals that are compared. In general, the responses are not

perfectly consistent, and λ_{max} is greater than q. The larger the λ_{max}, the greater is the degree of inconsistency. Saaty defines the consistency index (*CI*) as (λ_{max} - q) / (q - *1*), and provides a random index (*RI*) for matrices of order *3* to *10* based on a simulation of a large number of randomly generated weights. Saaty recommends the calculation of a consistency ratio (*CR*), which is the ratio of *CI* to the *RI* for the same order matrix. A *CR* of *0.10* or less is considered acceptable. When the *CR* is unacceptable, the panel member is made aware that his or her pairwise comparisons are logically inconsistent and is encouraged to revise them.

The responses are processed with Expert Choice and panel members with inconsistency ratios greater than 0.10 are asked to reconsider their judgments as suggested by Saaty. The mean importance weights are calculated after the necessary adjustments are made to inconsistent responses. Each panel member is presented with his or her individual score along with the *D-Panel* group mean weights. Panel members are given the opportunity to revisit their judgments and make revisions to their pairwise comparison scores based on this feedback.

There has been some criticism of AHP in the operations research literature. Harker and Vargas (1987) show that AHP does have an axiomatic foundation, the cardinal measurement of preferences is fully represented by the eigenvector method, and the principles of hierarchical composition and rank reversal are valid. On the other hand, Dyer (1990a) has questioned the theoretical basis underlying AHP and argues that it can lead to preference reversals based on the alternative set being analyzed. In response, Saaty (1990) explains how rank reversal is a positive feature when new reference points are introduced. We use the geometric aggregation rule to avoid the controversies associated with rank reversal (Dyer 1990a, Saaty 1990, Harker and Vargas 1990, and Dyer 1990b).

The result of this step is a set of weights representing the relative importance of the four missions and 10 strategic goals.

1.b. Facility Backlogs and Replacement Values: The Facilities Engineering Division at JSC regularly monitors its facilities and keeps data concerning deferred maintenance and current replacement value of its facilities. An official *Deferred Maintenance Parametric Estimating Guide* is used to perform this continuous assessment. The data collected by the Facilities Engineering Division is used to determine the Facility Condition Index (C_i) of each facility in Phase 2 (Step 2.c).

2.a. Facility Priority Index (P_i): Facility Priority Index attempts to quantitatively score facilities based on how well they support NASA's mission and strategic goals. *F-Panel* members rate each facility against each of the strategic goals identified in the 2003 NASA Strategic Plan (NASA 2003) using a rating scale between 0 and 100 (S_{ijk}). A simple additive weighting model is used to compute a P_i for facility i by multiplying the relative importance weight of each mission (W_j) by the relative importance weight of each goal (W_{jk}) and the facility score (S_{ijk}):

$$P_i = \sum_{j=1}^{4} \sum_{k=1}^{l_j} W_j W_{jk} S_{ijk} \qquad \text{(where } l_j \text{ is the number of goals associated with mission } j\text{)} \qquad (1)$$

The P_i calculation yields a score between 0 and 100. Facilities with large P_i are highly aligned with the mission and goals while facilities with small P_i are less closely aligned with NASA's mission and goals.

2.b. Facility Dependency Index (D_i): The purpose of the Facility Dependency Index is to reduce the risk of consequences associated with not funding or under-funding facilities. D_i is a modified version of the Mission Dependency Index developed by the Naval Facilities Engineering Service Center (NFESC) to describe the relative importance of infrastructure in

terms of mission criticality (Antelman and Miller 2002). The Facility Dependency Index is an operational risk management metric that reflects the extent to which other facilities are dependent on the focal facility. It does this by evaluating responses to two questions that together yield a score from 0 to 100, with 100 being the highest score. The first question focuses on the impact that an interruption of operations at one facility would have on other facilities. The second question focuses on the difficulty of relocating the function or service provided by the facility. *F-Panel* members respond to the following two questions for each facility:

Question 1: How long could the operations at this facility be stopped without adverse impact on other facilities? The possible responses are:

- None : Any interruption will immediately have an adverse impact on other facilities
- Very Brief: A few hours
- Brief : One day
- Very Short: Two days
- Short : Three to seven days
- Long: One to two weeks
- Very Long : More than two weeks

Question 2: If the facility was not functional, could its operations be relocated to another facility or a temporary facility? The possible responses are:

- Impossible: Alternate facility is not available.
- Very Difficult: Alternate facilities exist, but utilization would require an extraordinary effort with respect to person-hours and money and the job may be compromised.
- Difficult: Alternate facilities exist, but utilization would require a moderate effort with respect to person-hours and money, however, the job would not be compromised.
- Possible: Alternate facilities are readily available with little effort.
- Very Possible: Alternate facilities are readily available with no effort.

The responses to Questions 1 and 2 are referred to Table 1 to determine the Facility Dependency Index of the facility. Facilities with high D_i scores should be viewed as 'critical.' Other facilities are adversely affected by work interruptions at such facilities and the operations of such facilities cannot be easily relocated.

Insert Table 1 Here

2.c. Facility Condition Index (C_i): The Facility Condition Index, first cited in 1991, has long been a core metric in the facility management industry to provide a comparison between different facilities (Rush 1991). C_i measures the relative condition of a single facility and is defined as the facility backlog (B_i) divided by the facility replacement value (V_i) times 100 ($C_i = B_i/V_i \times 100$). Facility backlog is the dollar value of capital investments that have been postponed due to lack of funding or other reasons. The facility replacement value shows how much it would cost to replace a facility if it had to be built from scratch. The C_i calculation yields a score between 0 and 100. Facilities with lower C_i's are in better condition while facilities with higher C_i's are in worse condition.

While C_i is a useful index, relying only on C_i to allocate resources can be misleading because what is broken should not always be considered a top priority for funding (e.g., when a facility is not closely linked to missions/goals and when the facility's operations have little impact on other facilities). Our model therefore combines three indices to address this concern. Each facility is represented as a bubble in the 3-D model. The diameter of the bubbles represents the amount of investment needed to bring the facility to the desired condition (Facility Backlog).

3.a. Similarity Factors (Z_i): The Ideal point represents a hypothetical facility that would have received the maximum score (i.e., 100) on each of the evaluation criteria (Facility Priority, Facility Dependency, and Facility Condition). In contrast, the Nadir point represents a hypothetical facility that would have received the minimum score (i.e., 0) on each of the evaluation criteria. Our model relies on the premise that the facility that is closest to the Ideal and furthest from the Nadir is preferred to other facilities and is a better candidate for funding. (Note that the facility that is closest to the Ideal would be, by definition, furthest from the Nadir.) That is, the rank ordering of facilities from the Ideal will be inversely (and highly) correlated with the rank ordering of facilities from the Nadir. A graphical view of this model is presented in Figure 3.

Insert Figure 3 Here

Assuming:

n	=	Number of facilities under consideration.
P_i	=	Facility Priority Index of the i-th Facility ($i=1,...,n;\ 0 \le P_i \le 100$)
D_i	=	Facility Dependency Index of the i-th Facility ($i=1,...,n;\ 0 \le D_i \le 100$)
C_i	=	Facility Condition Index of the i-th Facility ($i=1,...,n;\ 0 \le C_i \le 100$)
d_i^*	=	Euclidean Distance of the i-th Facility from the Ideal Point ($i=1,...,n$)
d_i^-	=	Euclidean Distance of the i-th Facility from the Nadir Point ($i=1,...,n$)
Z_i	=	Similarity Factor of the i-th Facility to Ideal ($i=1,...,n;\ 0 \le Z_i \le 1$)

Our objective is to *Maximize* $Z_i = \dfrac{d_i^-}{d_i^* + d_i^-}$ (2)

where:

$$d_i^* = \sqrt{(P_i - 100)^2 + (D_i - 100)^2 + (C_i - 100)^2} \quad (3)$$

$$d_i^- = \sqrt{P_i^2 + D_i^2 + C_i^2} \quad (4)$$

3.b. Sensitivity Analysis: The process as proposed here presumes equal weights for P_i, D_i, and C_i. However, the process could be easily modified by attaching different weights to P_i, D_i, and C_i. A sensitivity analysis could be performed to evaluate the sensitivity of the results to the weights associated with P_i, D_i, and C_i. Assuming that $\hat{w}_P$, $\hat{w}_D$, and $\hat{w}_C$ are the relative importance weights associated with P_i, D_i, and C_i, the Euclidean distance of the i-th Facility from Ideal ($\hat{d}_i^*$) and the Euclidean distance of the i-th Facility from Nadir ($\hat{d}_i^-$) can be found as:

$$\hat{d}_i^* = \sqrt{\hat{w}_P (P_i - 100)^2 + \hat{w}_D (D_i - 100)^2 + \hat{w}_C (C_i - 100)^2} \quad (5)$$

$$\hat{d}_i^- = \sqrt{\hat{w}_P P_i^2 + \hat{w}_D D_i^2 + \hat{w}_C C_i^2} \quad (6)$$

Consider two facilities, A (P_A, D_A, C_A) and B (P_B, D_B, C_B). Equations (5) and (6) can be used to calculate $\hat{d}_A^*$, $\hat{d}_A^-$, $\hat{d}_B^*$, and $\hat{d}_B^-$. We can find the critical value of one weight assuming that another one is fixed. For example, assuming $\hat{w}_D$ is fixed, since $\hat{w}_P + \hat{w}_D + \hat{w}_C = 1$, both $\hat{w}_P$ and $\hat{w}_C$ will change with sensitivity analysis. We can find the critical weight associated with the P_i for the Nadir point ($\hat{w}_P^-$). The critical weight is the weight that causes rank reversal when $\hat{d}_A^- = \hat{d}_B^-$:

$$\hat{w}_P^- = \frac{\hat{w}_D\left(D_B^2 - D_A^2 + C_A^2 - C_B^2\right) + C_B^2 - C_A^2}{C_B^2 - C_A^2 + P_A^2 - P_B^2} \tag{7}$$

If A is preferred to B when $\hat{w}_P < \hat{w}_P^-$ then B is preferred to A when $\hat{w}_P > \hat{w}_P^-$

If A is preferred to B when $\hat{w}_P > \hat{w}_P^-$ then B is preferred to A when $\hat{w}_P < \hat{w}_P^-$

A and B are equally preferred if $\hat{w}_P = \hat{w}_P^-$

We can also find the critical weight associated with the P_i for the Ideal point ($\hat{w}_P^*$) assuming $\hat{d}_A^- = \hat{d}_B^-$:

$$\hat{w}_P^* = \frac{\hat{w}_D\left[(D_B - 100)^2 - (D_A - 100)^2 + (C_A - 100)^2 - (C_B - 100)^2\right] + (C_B - 100)^2 + (C_A - 100)^2}{(C_B - 100)^2 - (C_A - 100)^2 + (P_A - 100)^2 + (P_B - 100)^2} \tag{8}$$

If A is preferred to B when $\hat{w}_P < \hat{w}_P^*$ then B is preferred to A when $\hat{w}_P > \hat{w}_P^*$

If A is preferred to B when $\hat{w}_P > \hat{w}_P^*$ then B is preferred to A when $\hat{w}_P < \hat{w}_P^*$

A and B are equally preferred if $\hat{w}_P = \hat{w}_P^*$

There are several other approaches to sensitivity analysis. For example, one might consider performing sensitivity analysis on different weights such as the Center Priority Weights or different scores such as the facility scores.

3.c. Resource Allocation: In the final step of the process, Similarity Factors (Z_i) are used to allocate capital improvement funds to each facility. Facilities with higher P_i, C_i, and D_i are more closely aligned with goals, more critical, and more in need of improvement. Therefore, a larger percentage of their backlog should be funded. Initially, we normalize Z_i's to determine an allocation factor for each facility ($\overline{Z}_i = Z_i \Big/ \sum_{i=1}^{n} Z_i$). Next, $\overline{Z}_i$'s are multiplied by the backlogs to calculate the tentative capital improvement budget of each facility ($A_i = \overline{Z}_i \,.\, B_i$). A_i's are then summed to calculate a total tentative capital improvement for the Center ($\overline{A} = \sum_{i=1}^{n} A_i$). Systems Management Office uses $\overline{A}$ as a benchmark for capital improvement. Once the actual capital improvement budget for the Center ($\overline{\overline{A}}$) is determined, all A_i's are adjusted downward (if $\overline{A} > \overline{\overline{A}}$) or upward (if $\overline{A} < \overline{\overline{A}}$) by $(\overline{A} - \overline{\overline{A}})/\overline{A}$ percent.

3. THE WORKFORCE MODEL

The workforce model is conceptually identical to the facilities model and uses three indices to plot each workforce unit in a 3-D graph. A workforce management panel (*W-Panel*), which is selected from Center Operations, Engineering, Life Sciences, Mission Operations, Orbital Space Plane, Space Shuttle, International Space Station, and the Systems Management Office, is formed in addition to the *D-Panel* to participate in this process. Similarity factors are used to rank order workforce units based on their Euclidean distance from the Ideal and Nadir points. This model provides a systematic and structured approach to measure workforce needs and priorities in the short and long-term. 'Workforce units' or functionally-oriented job types are used as the unit of measure for this model. The workforce units are the civil service employees and support service contractors who share work with civil servants, but not the product-oriented

contractor workforce. The workforce units for three JSC Directorates: Mission Operations, Engineering, and Space and Life Sciences are presented in Table 2 for demonstration purposes.

Insert Table 2 Here

Directorates provide workforce backlogs and workforce value data. Workforce backlog refers to the gap between anticipated and optimal human resources investments (including human resources investments that have been postponed due to lack of funding). In order to determine this figure, the Directorate must first determine the optimal civil service level during the upcoming budget period. The optimal civil service level refers to the salary, benefit, and training costs associated with the number of civil service employees (and support service contractors who share work with civil servants) that will be required to meet forecasted workload expectations during the upcoming budget period. Each Directorate must also determine the anticipated civil service level during the upcoming budget period. When identifying the anticipated civil service level, attention should be paid to attrition (retirements and other turnover) levels, the quantity of hires currently allowed into the Directorate, and the typical movement rates among workforce units. If the optimal level of human resources investments is greater than the anticipated level, then additional investments in human resources are required during the upcoming budget period. If the optimal level of human resources investments is less than the anticipated level (e.g., because forecasts indicate that workload expectations are decreasing for the workforce unit), then no additional investments in human resources are required during the upcoming budget period (and the level of human resources investments in the workforce unit might even need to be reduced).

Alternative definitions of workload backlog could also be considered. For example, workforce backlog could also refer to the extent to which the workforce unit lags competitors in compensation (and hence experiences higher than expected rates of turnover and related human resources problems) or insufficient investment in training and development for current employees (e.g., in-house training, external education).

The most straightforward definition of workforce value is the current value of the workforce unit's compensation costs (including direct compensation and benefits). However, if we want to parallel the 'replacement value' concept from the facilities model, workforce value should be defined as the cost to replace the workforce unit if it had to be hired from scratch. In addition to compensation costs, this would include an estimate of the recruiting and selection costs that would be involved in replacement (likely to be much higher for some workforce units than others due to talent scarcity in some disciplines) and the orientation and training costs that would be required to make the new employees fully productive (also likely to be much higher for some workforce units than others). The Workforce Condition Index is workforce backlog divided by workforce value times 100, yielding a score between 0 and 100.

Next, the Workforce Priority Index is calculated. This index quantitatively scores workforce units based on how well they support NASA's mission and goals. The *W-Panel* collectively rates the workforce units against each of the goals identified in the 2003 NASA Strategic Plan (NASA 2003). An additive weighting model is used to compute this index for each workforce unit by multiplying the Center Priority Weights developed earlier by the *D-Panel* and the workforce unit scores developed by the *W-Panel*. This calculation yields a score between 0 and 100 for the Workforce Priority Index.

Next, the Workforce Dependency Index is calculated for each workforce unit. Workforce Dependency Index represents the consequences associated with not funding or under-funding workforce units. Similar to the facility model, this index determines workforce dependency

based on responses from two questions. The *W-Panel* members are asked to respond to the following two questions for each workforce unit:

Question 1: How long could the operations supported by this workforce unit be stopped without adverse impact on other workforce units?

Question 2: If the workforce unit was not functional, could its operations be reassigned to another workforce unit or a temporary workforce unit?

The responses to both questions are referred to Table 2 to determine the Workforce Dependency Index of the workforce unit (similar to the facility model). Workforce units with high dependency should be viewed as 'critical' units.

Finally, the three indices (Workforce Priority, Workforce Dependency, and Workforce Condition) are used to calculate the similarity factors and plot each workforce unit as a bubble in the 3-D graph. The workforce unit that is closest to the Ideal and at the same time furthest from its Nadir would receive higher priority in allocating human resources investments than other workforce units. The similarity factors are used to allocate funds for human resources for each workforce unit.

4. GUIDELINES FOR THE RATING PROCESS

Because the final ranking of facilities (and workforce units) will depend heavily on the ratings provided by *F-Panel* (and *W-Panel)* members, it is important that ratings be perceived as reasonably accurate and fair. If the rating process is viewed by stakeholders as biased, inaccurate, or contaminated by self-serving motives, then resource allocation decisions will be viewed as unfair.

Research in the field of industrial and organizational psychology has distinguished between the fairness of outcomes (e.g., the resources allocated to each facility) and procedural fairness (Gilliland and Langdon 1998). Procedural fairness refers to perceptions about the fairness of the process used to make the decision and is strongly influence by factors such as the relevance of criteria, opportunities for input to the decision making process, and the consistency with which the process and standards are applied (e.g., across facilities). In research on performance ratings, procedural fairness has been shown to be positively related to acceptance of performance evaluations, trust in the supervisor (rater), satisfaction with the evaluation process, motivation to improve performance, and organizational commitment. Also, supervisors appear to be less likely to distort or manipulate their ratings after steps are taken to build fairness into the process (Taylor et al. 1995). One consistent finding has been that fair procedures can make up for negative outcomes. That is, when the process is perceived to be fair and the basis for decisions is thoroughly and adequately communicated, individuals who receive negative outcomes are much more likely to perceive those outcomes as fair (Gilliland and Langdon 1998). Thus, procedural fairness is likely to be especially important to those facilities receiving negative outcomes (i.e., less than desired resource allocations).

Gilliland and Langdon (1998) suggest a number of steps that can be taken to enhance the perceived fairness of employee evaluation systems; these steps have direct parallels for evaluating facilities. For example, it is important to communicate consistently with all stakeholders about the development of the resource allocation process (including the relevance of the criteria that will be used to rank order facilities) and to alleviate any concerns about the new process through two-way communication forums (e.g., meetings, conferences). It is also important that facility managers are given opportunities to provide input, the evaluation process is standardized (i.e., the same standards are applied consistently to all facilities), multiple raters

are used to evaluate each facility (to minimize biases of individual raters), raters are knowledgeable about the operations at facilities they are asked to evaluate, and administrative decisions (i.e., resource allocations) are closely linked to the rating process. It would also be desirable to allow facilities an opportunity to rebut and request a review of ratings that they perceive to be unfair (thereby creating a sense of due process).

In developing the rating processes described here, we note two types of rating errors that can occur. Some rating errors are unintentional. For example, two raters who agree about a facility's contribution to a specific strategic goal might assign different ratings to the facility because they interpret the rating scale differently. However, some rating errors are intentional and reflect self-serving or political motives. In this case, raters may have the ability to make accurate ratings but they are unwilling to do so. In sum, raters can play organizational games and distort their ratings to achieve organizational or personal goals (Kozlowski et al. 1998). Kozlowski et al. (1998) have noted that politics and associated rating distortions are more likely when (a) there is a direct link between the ratings and desired organizational rewards (as is the case in resource allocation decisions), (b) there is a lack of surveillance of rater behavior, and (c) there is a widespread perception that others will distort their ratings (e.g., that others will provide inflated ratings concerning their own facilities). Kozlowski et al. (1998) describe several actions that organizations can take to minimize the role of politics in ratings; each has a clear implication for evaluating facilities. These recommendations include having top management serve as role models by providing fair evaluations and discouraging political game playing, allowing stakeholders to suggest potential improvements to the system itself, ensuring that evaluation criteria are widely viewed as relevant, training raters, using multiple raters for each facility, and making raters accountable for their evaluations (e.g., having to explain the reasons for their evaluations). It is noteworthy that the recommendations to reduce politics in ratings closely parallel the recommendations for enhancing procedural fairness.

When raters are motivated to provide accurate ratings, rater training can enhance the accuracy of their ratings (Hauenstein 1998). Hauenstein (1998) reviewed empirical research in this area and described key elements in successful rater accuracy training. First raters are familiarized with the rating criteria. Second, raters are given examples of facility characteristics associated with different points (e.g., 10, 30, 50, 70, 90) on the rating scale for each criterion (e.g., characteristics of facilities that might serve as indicators concerning their degree of alignment with specific strategic objectives). Third, raters complete practice ratings. For example, they can be provided with relevant information about a real or hypothetical facility and asked to rate the facility on one or more criteria (e.g., alignment with strategic objectives or questions 1 and 2 of the Facility Dependency Index). Fourth, the distribution of ratings from all raters is then displayed, often accompanied by 'true' ratings (developed earlier based on consensus discussions among subject matter experts who are highly knowledgeable about the facility being rated and the rating criteria). Raters then discuss reasons for differences in their ratings (e.g., "What information led you to give this facility a rating of 70 in terms of its alignment with strategic goal *i* whereas others gave the facility a rating of 40?"). For example, raters might discuss what information about a facility seems most relevant to them in arriving at a rating on a specific criterion. Discussion might also focus on whether inaccurate raters were simply too lenient (or harsh) or whether they did not recognize what information about the facility was most important for evaluating the specific criterion. Raters might also be cautioned to avoid halo errors (the tendency to form an overall impression about a facility and incorrectly rate the facility as high or low on all criteria on the basis of that overall impression). For

example, a facility whose operations would be difficult to relocate (or where interruption of operations would adversely affect other facilities) is not necessarily closely aligned with many strategic goals. Fifth, the process of completing practice ratings and subsequent discussions is repeated concerning several real or hypothetical facilities.

Based on rating research and recommendations described above, we offer the following guidelines for the rating processes used to generate P_i and D_i.

1. Raters should receive rater accuracy training prior to completing ratings that will influence resource allocation decisions.
2. Rating scales should include definitions that describe the meaning of several points on the scale (e.g., 90 = This facility exists primarily to support attainment of this strategic objective).
3. Multiple raters (e.g., 5 or more) should be used. Raters should not be asked to rate facilities with which they have little or no familiarity.
4. Facility managers should always participate in discussing and rating their own facility.
5. The reliability (i.e., consistency) of raters should be tracked. If a set of i raters evaluates j facilities, the reliability of their ratings can be indexed using an intraclass coefficient (Shrout and Fleiss 1979). Using an i raters x j facilities analysis of variance framework, $ICC=(MS_{between} - MS_{within})/MS_{between}$. Note that ICC yields the reliability of the average of the ratings made by i raters. ICC values greater than 0.70 are considered acceptable. To estimate the reliability that could be expected from using more or fewer raters, the Spearman-Brown prophecy formula can be used: $r_{nn} = nr_{xx}/[1+(n-1)r_{xx}]$ Where r_{nn} is the estimated reliability based on n times as many raters as those at hand, and r_{xx} is the reliability based on the current number of raters. For example, if r_{xx} was 0.80 using 4 raters, then the estimated reliability for 2 raters (i.e., $n = 0.50$) would be $0.67 = (0.50)(0.80)/[1+(0.50-1)0.80]$.
6. Raters whose ratings are consistently outliers should receive additional training or be removed from the process. When rating a single facility, outliers can be detected by displaying the distribution of ratings. Across j facilities, a rater would be considered an outlier if ICC increases when the rater's evaluations are removed from the dataset.
7. Substantial disagreements among raters should be discussed thoroughly and the facility rating on the criterion should be reached through consensus. Where only minor disagreements occur among raters, the average rating can serve as the facility rating.
8. A facilitator should guide rating sessions to ensure that the same process is applied systematically to all facilities. The facilitator can also provide training to new raters and facilitate discussions to reach consensus when there are substantial disagreements among raters. A few studies have focused on the method participants used to interact with GDSS, with emphasis on the use of human facilitators. In general, it has been shown that facilitation enhances the effectiveness of groups using GDSS (Khalifa et al. 2002).

5. HYPOTHETICAL EXAMPLE

This section illustrates the concepts of Ideal and Nadir using four hypothetical facilities at JSC. First, the Facility Priority Index (P_i) is computed for each of the four facilities. P_i Shows how closely aligned each facility is with NASA's mission and strategic goals. The first step in determining P_i is to have the *D-Panel* at JSC identify the relative importance of each mission and strategic goal by using AHP and Expert Choice software. Each panel member is presented with his or her individual weights along with the *D-Panel* group mean weights. The panel members

are given the opportunity to revise their judgments according to this feedback. Table 3 presents the *D-Panel* mean weights after two rounds of judgment.

Insert Table 3 Here

The next step in determining P_i is having the *F-Panel* rate each facility against each strategic goal using a rating scale between 0 and 100. Higher scores are given to facilities that are highly aligned with a goal while lower scores are given to facilities that are less closely aligned with the goal. Table 3 shows the scoring of the facilities along with the P_i for each facility.

Following the calculation of P_i's, we calculate the Facility Dependency Index (D_i) of each facility. The *F-Panel* is asked to respond to the two questions used to determine the Facility Dependency Index. Table 4 shows the responses to the two questions along with the corresponding D_i for each facility. The results show that facility 3 is the most critical facility while facility 2 is the least critical facility.

Insert Table 4 Here

Next, the Facilities Engineering Division provides pertinent information concerning capital investment backlog and current replacement value of each facility. These figures are used to determine a Facility Condition Index (C_i) for each facility. As shown in Table 5, facility 3 has the largest C_i while facility 2 has the smallest C_i. Furthermore, facility 1 has the largest capital investment backlog while facility 2 has the smallest backlog.

Insert Table 5 Here

Using the mission priority, facility dependency, and facility condition indices, we calculate the similarity factor (Z_i) of each facility and rank order them in descending order. As show in Table 6, facility 3 is most similar to the Ideal with a similarity factor of *0.64* while facility 2 is least similar to the Ideal with a similarity factor of *0.17*. The proximity of facility 3 to the Ideal and facility 2 to the Nadir can also be viewed in the 3-D model presented in Figure 4(a). All four facilities are represented as bubbles in these figures. The size of the bubbles represents the backlog. Bubble 1 has the largest diameter (B_1=*$2,800,000*) while bubble 2 has the smallest backlog (B_2=*$1,100,000*). The user interface in *D-Side* supports visualization and animation by allowing the decision maker to rotate the 3-D model and view the facilities from different angles in the cube as it is shown in Figure 4(b). Additional 2-D views of the 3-D model given in Figure 4(c, d and e) can also provide valuable information.

Insert Table 6 and Figure 4 Here

D-Side empowers the decision makers with tools to perform what-if analysis, goal-seeking, and sensitivity analysis. For example, to study the sensitivity of weight to rank reversal, we can find the critical weight associated with the P_i for the Nadir point ($\hat{w}_P^-$) and the Ideal point ($\hat{w}_P^*$). Note that the default relative importance weight associated with P_i, D_i, and C_i, is 0.33 and $\hat{w}_P = \hat{w}_D = \hat{w}_C$. Given the following P_i, D_i, and C_i:

Facility 1 (*40, 40, 25*)
Facility 2 (*20, 20, 10*)
Facility 3 (*80, 80, 40*)
Facility 4 (*30, 60, 20*)

facility 3 totally dominates others while facility 2 is totally dominated by the others. Therefore, we focus our sensitivity analysis on facilities 1 and 4. Using equation (7), we find $\hat{w}_P^-$, the critical value of P_i as a function of D_i:

$$\hat{w}_P^- = \frac{1775\ \hat{w}_D - 225}{475}$$

Considering different values for $\hat{w}_D$, we calculate $\hat{w}_P^-$. As shown in Table 7, when $\hat{w}_D = 0.20$, we find $\hat{w}_P^- = 0.27$, meaning that 0.27 is the reversal weight for P_i (compared with the default weight of 0.33 for P_i). In other words, the rank order changes when moving the weight from below 0.27 to above 0.27. $\hat{w}_P^- > 1$ or $\hat{w}_P^- < 0$ corresponds to the infeasible cases where there is no reversal and the order does not change. Similar calculations are presented in Table 7 for $\hat{w}_P^*$ where $\hat{w}_P^* > 1$ or $\hat{w}_P^* < 0$ corresponds to infeasible cases:

$$\hat{w}_P^* = \frac{-2775\ \hat{w}_D + 775}{-525}$$

Insert Table 7 Here

Normalized Similarity Factors ($\overline{Z}_i$) are used to determine the required funding needed for each facility (A_i). The tentative capital improvement amount suggested by *D-Side* for all 4 facilities equals $1,769,000 (see Table 8). This amount is used by the Systems Management Office as a benchmark for capital improvement distribution. Depending on the available budget, A_i's are adjusted downward or upward. For example, if JSC leadership approves $690,000 (10% of total backlogs) for facility improvement, all A_i's will be adjusted downward by (1,769,000-690,000)/1,769,000 or 61.0% (see table 8).

Insert Table 8 Here

6. PILOT STUDY

Twelve facility managers volunteered to participate in the pilot study. Their mean number of years at NASA was 18.58 (SD = 7.15). They had on average 8.83 years (SD = 3.83) of facility management experience at NASA and 6.58 years (SD = 1.78) of experience with the current facility evaluation process. Seventeen percent had a bachelor's degree, 66% had a master's degree, and 17% had a doctoral degree.

Participants were randomly assigned to one of two sessions. Both groups were asked to evaluate 20 randomly selected facilities at JSC with a total backlog of $19,100,000. Group I participants first evaluated the facilities using the current approach (115 minutes) and then evaluated the facilities using *D-Side* (200 minutes). Group II participants first evaluated the facilities using *D-Side* (225 minutes) and then evaluated the facilities using the current approach (80 minutes). In both sessions, the survey containing the dependent variables (described below) was completed at the end of the session (i.e., after all facilities had been evaluated using both approaches).

Table 9 shows the combined results for the two groups using the current approach. The similarity factors presented in Table 10 provide an overall ranking of the facilities using *D-Side* (sorted in a descending order of Z_i). Since we had two groups working on the same problem, the average priority (P_i) and dependency (D_i) indices for the two groups were used to arrive at the final figures. Note that the C_i's remain unchanged for both groups. The correlation between the

Overall Score from the current approach (Table 9) and Z_i from *D-Side* (Table 10) was .88. That is, the two approaches yielded similar rank orderings of the facilities. The 3-D and 2-D views of the *D-Side* results are presented in Figure 5. Using the normalized similarity factors, $969,510 was considered as the tentative capital improvement amount for the 20 facilities. However, it was expected that JSC leadership approves 10% or $1,910,000 towards capital improvement for these 20 facilities. Therefore, all figures were adjusted upward by a factor of (969,510-1,910,000)/969,510 or 97.0%.

Insert Figure 5 and Tables 9 and 10 Here

Dependent Variables

Our choice of dependent variables was influenced primarily by Bharati and Chaudhury's (2004) decision satisfaction model, Benbasat and Lim's (1993) meta-analysis, DeLone and McLean's (1992) usability model, Gallupe and DeSanctis's (1988) effectiveness model, and Limayem and DeSanctis's (2000) decisional guidance model. Collectively, these (and many other) studies point to the importance of ease of use, the quality of the decision process, and decision quality as key antecedents of users' satisfaction with and use of decision support systems.

We therefore created four scales: Ease of Use (7 items), Decision Process (8 items), Decision Quality (6 items), and Overall Value-Added (4 items). Where feasible, items were adapted from items used in previous studies (e.g., Aldag & Power, 1986; Bharati and Chaudhury, 2004; Davis, 1989; DeLone and McLean, 1992; Gallupe, DeSanctis & Dickson, 1988; Niederman & DeSanctis, 1995; Watson, DeSanctis, & Poole, 1988; Srinivasan, 1985). All items are presented in the Appendix. Each item was answered using a 7-point rating scale where 1 = strongly disagree, 4 = neutral, and 7 = strongly agree.

The reliability (Cronbach's alpha) of each of the scales was good. For the seven Ease of Use items, *alpha* = .74 and .85 for the current approach and *D-Side*, respectively. For the eight Decision Process items, *alpha* = .77 and .67 for the current approach and *D-Side*, respectively. For the six Decision Quality items, *alpha* = .88 and .84 for the current approach and *D-Side*, respectively. For the four Overall Value-Added items, *alpha* = .92 and .91 for the current approach and *D-Side*, respectively.

Following Campbell and Fiske (1959), Davis (1989) pointed out that items intended to measure the same construct for the same approach (convergent validity) should be more highly correlated than items intended to measure different constructs for the same approach (discriminant validity) or different constructs for different approaches. We found that the average correlation among items measuring the same construct for the same approach (r = .39) was higher than the average correlation among items measuring different constructs for the same approach (r = .32) or different constructs for different approaches (r = -.26).

We created a score for each participant on each scale (for each approach) by computing the mean responses to the Ease of Use, Decision Process, Decision Quality, and Overall Value-Added items.

Users' Evaluations

Descriptive statistics concerning the dependent variables are presented in Tables 11 and 12. Inspection of the means in Table 11 indicates that users' reactions to the current approach were slightly below or near the center of the 7-point rating scale whereas their reactions to *D-Side* were more favorable. Inspection of individual-level ratings indicted that 11 of the 12 facility managers evaluated *D-Side* more favorably than the current approach on all four dependent variables. Independent groups t-tests indicated that there were no significant differences on the dependent variables for users in the two sessions. The correlations in Table 12

indicate that, for each approach, correlations among the dependent variables were moderate to high. Also, users' opinions about the current approach were negatively related to their opinions about *D-Side*. That is, positive opinions about *D-Side* were associated with negative opinions about the current approach.

Insert Tables 11 and 12 Here

We used multiple regression to examine the extent to which Overall Value-Added was predicted by Ease of Use, Decision Process, and Decision Quality. For the current approach, $R^2 = .74$, $p = .01$. For *D-Side*, $R^2 = .80$, $p < .01$. Inspection of the standardized beta coefficients (and significance tests associated with them) indicated that Decision Quality was the strongest predictor of Overall Value-Added for both the current approach and *D-Side* (see also the correlations in Table 12).

We examined the relationship between users' background characteristics and the dependent variables. 'Years at NASA' was positively related to opinions about the current approach (for Ease of Use, $r = .57$, $p = .05$; for Decision Process, $r = .69$, $p = .01$, for Decision Quality, $r = .82$, $p < .01$, for Overall Value-Added, $r = .65$, $p = .02$). However, 'years at NASA' was unrelated to opinions about *D-Side*. 'Years of facility management experience at NASA' was positively related to opinions about the current approach (for Decision Process, $r = .58$, $p = .05$; for Decision Quality, $r = .71$, $p < .01$). 'Years of facility management experience at NASA' was unrelated to opinions about *D-Side*. 'Years of experience with the current facility evaluation process' was unrelated to opinions about the current process or to opinions about *D-Side*. However, education level was positively related to opinions about *D-Side* (for Ease of Use, $r = .68$, $p = .02$; for Decision Quality, $r = .59$, $p < .04$, for Overall Value-Added, $r = .71$, $p = .01$) but was unrelated to opinions about the current approach. In sum, users with many years of facility management experience at NASA viewed the current approach more favorably than those with less facility management experience. (Although, as noted above, 11 of 12 users viewed *D-Side* more favorably than the current approach.) And users with higher levels of education viewed *D-Side* more favorably than those with less education.

We used paired t-tests to compare users' opinions concerning the current approach and *D-Side* on each of the four dependent variables. Significance tests indicated that users had more favorable opinions concerning *D-Side* than the current approach on each of the four dependent variables (for Ease of Use, $t = 7.75$, $df = 11$, $p < .001$; for Decision Process, $t = 8.12$, $df = 11$, $p < .001$; for Decision Quality, $t = 6.99$, $df = 11$, $p < .001$; for Overall value-Added, $t = 7.31$, $df = 11$, $p < .001$).

For each dependent variable, we calculated Cohen's d (i.e., the mean difference divided by the standard deviation of the difference scores) as a measure of the effect size. According to Cohen (1988), small, medium, and large values of d are about .20, .50, and .80 respectively. All effect sizes were very large (for Ease of Use, $d = 2.23$; for Decision Process, $d = 2.35$; for Decision Quality, $d = 2.01$; for Overall value-Added, $d = 2.12$), thereby indicating that the mean differences between the current approach and *D-Side* were both statistically and practically significant.

7. CONCLUSION AND MANAGERIAL IMPLICATIONS

The pilot study illustrated that, despite the longer amount of time required to use *D-Side* relative to the current approach (and the similar rank ordering of facilities using the two approaches), facility managers viewed *D-Side* much more favorably in terms of ease of use, decision process, decision quality, and overall value-added. *D-Side* is quite flexible in that it can easily be adapted

by other organizations, for example, by substituting other criteria for those listed here or adding additional criteria. Although this paper uses three criteria (e.g., facility priority index, facility dependency index, and facility condition index), the mathematics of *D-Side* can accommodate any number of relevant criteria and determine the distance of each facility or workforce unit from the Ideal and Nadir (although the 3-D views are naturally limited to only three criteria at a time). As illustrated by the workforce model presented here, the basic framework of *D-Side* can be readily adapted to other decision-making contexts (e.g., evaluating proposals within facilities, evaluating new business opportunities, and so on). Another advantage of *D-Side* is the ease with which it places both inherently subjective criteria (e.g., the alignment of facilities with strategic goals) and more objective criteria (e.g., facility condition index) on a common measuring scale (0 to 100). Finally, sensitivity analyses enable decision makers to understand the conditions (criterion weights) that would cause reversals in the rank ordering of facilities.

We believe that one of strength of *D-Side* is its integration of research from multiple disciplines. That is, it combines research from literature on decision-making (e.g., AHP, MCDM) and research from the industrial and organizational psychology literature (e.g. procedural justice, rater accuracy training). The use of tools from either one of these disciplines without the other is likely to be inadequate. For example, if facility priority index ratings were derived primarily or solely on the basis of ratings made by facility managers concerning their own facilities, self-serving and political biases would likely contaminate ratings and create widespread perceptions that the process was unfair despite its appearance of objectivity in terms of the underlying mathematics. At the same time, the industrial and organizational psychology literature does not offer useful guidance about (a) optimal approaches to weight criteria (e.g., AHP), (b) how changing criterion weights will affect resource allocation decisions (e.g., sensitivity analysis), or (c) how to combine subjective (e.g., ratings) and objective (e.g., financial) data in a way that offers an easily understood rationale to support resource allocation decisions.

While the focus on three criteria might be viewed by some as limiting, we believe it is actually a strength. *D-Side* forces decision-makers to carefully consider and select the most relevant criteria rather than simply brainstorming and including a potentially lengthy list of relevant criteria. *D-Side* also emphasizes the importance of combining strategic (and hence long-term) criteria along with more tactical and short-term criteria (e.g., facility backlog and the extent to which operational interruptions at each facility might adversely affect operations at other facilities).

As with any MCDM, selecting the 'right' criteria is critical. In each application of D-Side, one empirical question of applied interest is whether criteria are highly correlated with each other. For example, if workforce unit scores on one criterion (e.g., alignment with strategic goals) were found to be highly correlated (e.g., $r > .80$) with scores on another criterion (e.g., workforce condition index), then it could be argued that the two criteria are redundant. In this situation, one of the two criteria might be replaced with another relevant yet conceptually independent criterion. Moreover, we expect that when decision makers are allowed to include a lengthy list of criteria (an outcome deliberately constrained by *D-Side*), it is highly likely that some criteria will be highly correlated (and therefore redundant) with other criteria. When this occurs, the net effect is that whatever is being measured by the highly correlated (redundant) criteria is given additional mathematical weight in final decisions. This points to the importance of (a) initially selecting a small number of conceptually independent criteria that address both

strategic and tactical perspectives, and (b) periodically examining the correlations among criteria.

Cluster analysis can be used to identify groups of facilities that are similar on the underlying criteria. Note that these groups of facilities would be clustered together in the 3-D model. Identifying relatively homogeneous clusters of facilities can be helpful because it enables senior managers to identify resource allocation and related management actions that are likely to address the needs of all the facilities in a cluster (rather than having to make separate decisions about each facility on its own).

Using a structured, step-by-step approach like *D-Side* is not intended to imply a deterministic approach to facility and workforce planning. While *D-Side* enables decision makers to crystallize their thoughts and organize data by placing both inherently subjective criteria and more objective criteria on a common measuring scale, it should be used very carefully. As with any decision analysis model, the researchers and practicing managers must be aware of the limitations of subjective estimates. *D-Side* should not be used blindly to plug-in numbers and crank-out solutions. The effectiveness of the model relies heavily on the ability and willingness of decision makers to provide sound judgments. Potentially, decision makers could make poor judgments as they do with any approach. Such judgments can generate misleading results and ultimately poor decisions.

REFERENCES AND TABLES ARE AVAILABLE UPON REQUEST

Influence of Hydroponically Grown Hoyt Soybeans and Radiation Encountered on Mars Missions on the Yield and Quality of Soymilk and Tofu

Final Report
NASA Faculty Fellowship Program 2004
Johnson Space Center

Prepared by:	Lester A. Wilson, Ph.D.
Academic Rank:	Professor
University and Department:	Iowa State University Food Science and Human Nutrition Ames, IA 50011
NASA/JSC	
Directorate:	Space and Life Sciences
Division: Factors Office (HEFO)	Habitability and Environmental
Branch:	Habitability and Human Factors Office
JSC Colleague:	Michele Perchonok,
Date Submitted:	August 17, 2004
Contract #	NAG 9-1526 and NNJ04JF93A

ABSTRACT

Soybeans were chosen for lunar and planetary missions due to their nutritive value and ability to produce oil and protein for further food applications. However, soybeans must be processed into foods prior to crew consumption. Wilson et al. (2003) raised questions about (1) the influence of radiation (on germination and functional properties) that the soybeans would be exposed to during bulk storage for a Mars mission, and (2) the impact of using hydroponically grown versus field grown soybeans on the yield and quality of soyfoods. The influence of radiation can be broken down into two components: (A) affect of surface pasteurization to ensure the astronauts safety from food-borne illnesses (a Hazard Analysis Critical Control Point), and (B) affect of the amount of radiation the soybeans receive during a Mars mission. Decreases in the amount of natural antioxidants and free radical formation and oxidation induced changes in the soybean (lipid, protein, etc.) will influence the nutritional value, texture, quality, and safety of soyfoods made from them. The objectives of this project are to (1) evaluate the influence of gamma and electron beam radiation on bulk soybeans (HACCP, CCP) on the microbial load, germination, ease of processing, and quality of soymilk and tofu; (2) provide scale up and mass balance data for Advanced Life Support subsystems including Biomass, Solid Waste Processing, and Water Recovery Systems; and (3) to compare Hoyt field grown to hydroponically grown Hoyt soybeans for soymilk and tofu production. The soybean cultivar Hoyt, a small standing, high protein cultivar that could grow hydroponically in the AIMS facility on Mars) was evaluated for the production of soymilk and tofu. The quality and yield of the soymilk and tofu from hydroponic Hoyt, was compared to Vinton 81 (a soyfood industry standard), field Hoyt, IA 2032LS (lipoxygenase-free), and Proto (high protein and antioxidant potential). Soymilk and tofu were produced using the Japanese method. The soymilk was coagulated with calcium sulfate dihydrate. Soybeans and tofu were evaluated using chemical, microbial, and instrumental sensory methods. The surface radiation of whole dry soybeans using electron beam or gamma rays at 10 or 30 kGy did provide microbial safety for the astronauts. However, these doses caused oxidative changes that resulted in tofu with rancid aroma, darkening of the tofu, lower tofu yields, more solid waste, and loss of the ability of the seeds to germinate. While lower doses may reduce these problems, we lose the ability to insure microbial safety (cross-contamination) of bulk soybeans for the astronauts. Counter measures could include vacuum packaging, radiating under freezing conditions. A No Effect Dose for food quality, below 10 kGy needs to be determined. Better estimates of the radiation that the food will be exposed to need to determined and shared. Appropriate shielding for the food as well as the astronauts needs to be developed. The Hoyt soybean did not provide a high yielding, high quality tofu. A new small scale system for evaluating soybeans was developed using 50 g quantities of soybeans.

INTRODUCTION

Soybeans were chosen for lunar and planetary missions due to their nutritive value and ability to produce oil and protein for further food applications. However, soybeans must be processed into foods prior to crew consumption. Wilson et al. (2003) raised questions about (1) the influence of radiation (on germination and functional properties) that the soybeans would be exposed to during bulk storage for a Mars mission, and (2) the impact of using hydroponically grown versus field grown soybeans on the yield and quality of soyfoods. The influence of radiation can be broken down into two components: (A) affect of surface pasteurization to ensure the astronauts safety from food-borne illnesses (a Hazard Analysis Critical Control Point), and (B) affect of the amount of radiation the soybeans receive during a Mars mission. Decreases in the amount of natural antioxidants and free radical formation and oxidation induced changes in the soybean (lipid, protein, etc.) will influence the nutritional value, texture, quality, and safety of soyfoods made from them. The NASA Advanced Food Technology team needs to evaluate small quantities of hydroponically grown soybeans to determine their acceptability for Lunar and Mars missions. In addition, due to limited quantities of soybeans, it is necessary to develop a methodology for evaluation of smaller quantities of soybeans. Due to the fact that the exact dose the astronauts and foods would be exposed to during a Mars mission was not available from NASA, we concentrated on the use of radiation as a HACCP step to insure that the bulk soybeans would not be vectors for food-borne illness microorganisms. Both e-bean and gamma radiation were used as a CCP to insure the safety of the astronauts. While radiation can be used to pasteurize and sterilize foods, the radiation can influence the sensory and nutritive value of the food.

Radiation and Food Systems

Irradiation is one of the approved methods to extend the shelf life, and preserve foods. Less than one kGy can be used to inhibit sprouting of potatoes, control insects in fruits and grains, and delay ripening. 1-10 kGy treatments can kill pathogenic microorganisms, whereas 30kGy to 67 kGy can be used to sterilize foods (meats, dried spices, etc.) (Bennion and Scheule, 2004; Potter and Hotchkiss, 1995).
Ionizing radiation is a primary concern not only for human health but also for food quality and functionality. The foods shipped to Mars must be free from food-borne illness microorganisms and pathogens. An extended inter-planetary mission to Mars, as proposed by NASA, will require a 5-year shelf life for prepackaged foods for the return flight. In addition, any ingredients or food items used on the planetary surface will require a shelf life of close to 5 years. Understanding the effects of safety measures taken prior to transit, as well as adverse conditions to which these products will be exposed, is necessary to enable development of a high quality nutritious food system.

Lipid Oxidation: The basic components of food consist of water, carbohydrates, proteins, lipids, vitamins, and minerals. Lipids may be the most affected by radiation because free radicals will participate in the initiation step of the lipid oxidation reaction. This oxidative process consists of three steps (initiation, propagation, and termination) and leads to the development of short chain acids, aldehydes, ketones, carbonyls, and

peroxides that may impart off-odors and off-flavors in lipids and foods containing lipids. Even small amounts of these compounds may leave foods, oils or fats unfit for consumption. Unsaturated fatty acids (containing double bonds) and other unsaturated compounds in foods are easily oxidized, which may make soybeans (high in unsaturated fatty acids (oleic, linoleic and linolenic acids) very susceptable to this type of preservation technique. High Temperatures, presence of oxygen, iron, and UV light are all known to catalyze the formation of free radicals and the ensuing autoxidation. The task of keeping foods from becoming oxidized is not an easy task. High barrier opaque packaging, use of vacuum packaging, anf the addition or chelating agents (EDTA) and antioxidants (natural Vitamin E and C; synthetic BHA, BHT, TBHQ) are currently used to slow these oxidation reactions. While soybeans contain Vitamin E, it will be 'used up' protecting the unsaturated fatty acids, thus allowing oxidation to start after the initial lag phase in this reaction.

Radiation effects on soybeans: Radiation has been used to create mutagens to develop new soybean oil cultivars. Hammond and Fehr (1975) used X-rays and ethyl methylsulfonate to create cultivars with low linolenic acid content. Seeds from parent strains, F2 seeds, and crosses were irradiated (X-ray) at levels of 10, 15, 20, and 25 Kr. While seeds treated with 20 and 25 Kr had a poor germination rate, the 10 Kr gave near normal germination.

Wilson during his NASA Faculty Fellow Program (NFFP) in 2003 at Johnson Space Center (Houston, TX) raised questions about the influence of radiation (on germination and functional properties) that the soybeans would be exposed to during bulk storage prior to and during a Mars mission. The influence of radiation can be broken down into two components: the affect of surface pasteurization to ensure the astronauts safety from food-borne illnesses (HACCP, CCP), and the affect of the amount of radiation the soybeans receive during a Mars mission.

Hydroponic Soybeans on Mars

It has been proposed that astronauts on extended duration missions might hydroponically grow their own soybeans for production. Specifically, NASA has chosen the Hoyt variety of soybean for its short full-grown stature and relatively high protein count. Research has shown (Watanabe, T., et al. 1964; Wilson, 83, 85, 86, 04) that differences in soybeans will have an effect on the sensory characteristics of soy products. As such, there was a need to evaluate the composition of the HOYT soybeans, their ability to imbibe water, and their protein extractability for tofu manufacture, in order for the astronauts to produce food from the soybeans. All tofu-making processes have a set of steps necessary to produce acceptable products. Although minor changes can be made depending on the technique used and the desired type of tofu (firm or silken style), all processes involve: soaking of dry beans, grinding the hydrated beans, pasteurization, filtering solids out of the soymilk, coagulation, and pressing. Soybeans must be evaluated for their ability to produce acceptable products with maximum yields and minimum waste streams. During Wilson's 2003 NFFP at JSC, no hydroponically grown soybeans were available. However, the field grown Hoyt soybeans were found to be of

poor quality and that they produced inferior soymilk and tofu (compared to Vinton 81 and IA 2032LS cultivars). The Hoyt cultivar had less protein in the bean and resulting tofu than all of the other cultivars. The tofu and okara were an unacceptable gray/black mottled color compared to the cream colored Vinton 81 standard. The IA 2032 LS cultivar is a lipoxygenase- free (lacks the enzymes that catalyze lipid oxidation) that had a much milder aroma and flavor, than the other cultivars evaluated in this study.

Therefore, the objectives of this research was to (1) determine the influence of radiation (pasteurization and sterilization), as a HACCP, CCP step, on the germination rate and the quality of tofu; (2) to compare Hoyt field grown to hydroponically grown Hoyt soybeans for soymilk and tofu production, (3) provide scale up/down procedure to evaluate small quantities of soybeans for food use, and (4) supply additional mass balance data for Advanced Life Support subsystems including Biomass, Solid Waste Processing, and Water Recovery Systems.

Research Approach: Soybean cultivars were selected based upon the results of my 2003 NFFP (Vinton 81 and IA 2032LS), the availability of hydroponic and field grown Hoyt soybeans, and the availability of a high antioxidant capacity Proto soybean. All of the cultivars selected were non-GMO, which were grown at known locations. Vinton 81 is a high protein, large seeded cultivar that is considered the gold standard by the Soyfoods Industry around the world. It is used for soymilk and tofu production. IA 2032LS is a large seeded, high protein cultivar that is lacking all three lipoxygenase isoenzymes. Lipoxygenase enzymes catalyze the formation of hydroperoxides that break down unsaturated fatty acids into low molecular weight flavor compounds, described as green, grassy, beany, painty, oxidized odors and flavors. All soybeans were stored in the dark at 20 C prior to and after irradiation

One pound of each cultivar was put into a large ziploc bag, the air squeezed out, sealed, and labeled prior to being treated. The amount in the bag allowed a single layer of seeds to be exposed to the radiation treatment. The bagged soybeans were irradiated at 5 doses (0, 10, 20 kGrays) at the Iowa State University Linear Accelator Facility, Texas A&M Electron Beam Facility for e-beam and the University of Illinois for gamma irradiation (Irradiation costs were covered by grants from NASA FT CSC and USDA, CSREES Regional Research Project NC-136) . One set of each cultivar was shipped to each location, but not irradiated to serve as a control. After each treatment, the soybeans were divided into four batches: (1) for chemical analyses [proximate analyses, peroxide value, thiobarbartic acid], using standard AOCS procedures, and antioxidant potential (PhotoChem), aroma by GC (Wilson, 1998, 2004]; (2) microbial analyses (standard plate count, coliforms, Salmonella, yeasts and molds) using standard methods in the NASA Food Microbiology Lab (JSC, Houston, TX),; (3) germination test (Bugbee, 2004) and (4) soymilk and tofu production.

The functionality of the soybeans was evaluated by manufacturing soymilk and tofu. The standardized methods of Johnson and Wilson (1984), Moizuddin, et al. (1999a), Moizuddin, Johnson, and Wilson (1999b); Wilson, 2003 were used. The Japanese method of soymilk production (Wilson (1995) from whole soybeans was utilized (soak beans 8-12 hours, grind beans, cook at 95 °C for 7 minutes, filter out okara, coagulate the

soymilk, cut the curds to release the whey, press in tofu press, refrigerate overnight prior to chemical and instrumental tests). Three different tofu presses (Fig.1) were manufactured at ISU (funded by NASA FT CSC, 2004) to allow scale-down procedures to be developed (based upon 300 g, 100 g, and 50 g batches).In addition, two NASA large tofu presses were used (Wilson, 2003; Fig.1)An 8% soluble solids soymilk was produced and coagulated using calcium sulfate dihydrate at 85 °C. The amount of coagulant needed was determined by the method of Moizuddin, Johnson, and Wilson (1999b). Yields of soymilk, tofu, okara, and whey along with the color, texture, and aroma of the soymilk and tofu were determined utilizing instrumental methods. Color was measured by using a Hunter Color Difference Meter Model XE under D65 light with a 10-degree standard observer. Texture was determined by using a Texture Profile Analysis (TPA) procedure (Bourne, 1978) to determine hardness, brittleness, adhesiveness, cohesiveness, and elasticity of each sample. A 1 cm-cube of tofu was compressed (80%) using a compression head in a Texture Technology TA XT2ci instrument. pH and conductance measurement of the whey were used to determine the optimum coagulation of the milk (Moizuddin, Johnson, and Wilson, 1999b; Wilson, 2003). All procedures were replicated in triplicate. The results were analyzed statistically for treatment affects, and correlations.

FIGURE 1. TOFU PRESS BOXES. THE TWO PRESSES ON THE RIGHT WERE USED IN NFFP 2003. THE FOUR PRESSES ON THE LEFT WERE USED FOR THIS STUDY.

Results and Discussion

While awaiting the arrival of the tofu presses and the irradiated soybeans, two major activities were initiated. Dr. Juming Tang and I prepared a White paper on the retortable pouch thermal processing system used by the NASA ISS product development team. Action was taken based upon the recommendations given in this report. The second activity used Vinton 81 and Hoyt soybeans from my NFFP 2003 to standardize new calcium sulfate dihydrate and to teach an intern how to manufacture soymilk and tofu.

To optimize the processing of soybeans using the new tofu presses and the evaluation of the irradiated soybeans, a preliminary characterization of the chosen soybean varieties was necessary. The preliminary data using Hoyt, Vinton 81, and IA 2032LS soybeans from NFFP03 were used to evaluate the new soymilk and tofu manufacturing system. Likewise, these soybeans were used to scale down the amounts of soybeans needed per test from 3,000 g to 50 g batches. After the arrival of the irradiated soybeans, control and treated samples were run in order to get an estimate of their

behavior and the amount of coagulate needed per treatment. This preliminary data yielded information about the characteristics of the beans themselves as well as how they performed in soymilk, tofu, okara, and whey processing. The initial processing was done in small batches on the stove in the Space Food Systems Laboratory to allow for manipulation of a number of variables at once, and to maximize the used of the limited supply of hydroponic Hoyt soybeans (300 g).

Upon the completion of the preliminary experiments, an optimal processing technique was developed for the Hoyt beans and the irradiated soybeans using 50 g batches. Additional tests were performed using a jacketed kettle in the food lab to verify conditions on a larger scale. Results were compared to NFFP 2003 and to commercial ISU Pilot Plant data.

With scale-ups and base-line data obtained, the main experiment with replication was initiated.

Preliminary Hoyt soybean characterization and process optimization

Hoyt soybeans were selected by NASA because they are believed to be high in protein, low growing, and can be grown hydroponically. Hoyt beans contain black hilum and seed coat staining in NFFP03 and both field and hydroponically grown crops this year had the same characteristics. In both years, this stain was carried into the soymilk and tofu (Wilson, 2003). Only a limited amount of the Hoyt soybeans grown under hydroponic conditions was available (300 g), so 50 g batches were used in the scale-up and in final evaluations of all treatments. Addition hydroponicall grown soybeans should be grown and evaluated in the future due to the small sample size available this year, Compositional data for all the cultivars are given in Table 1.

Table 1. Soybean Composition

			13% moisture basis		
	Sample	Moisture	Protein	Oil	Fiber
2002	Vinton 81	8.9	38.4	17.4	4.6
	IA 2032 LS	10.8	39.0	18.2	4.6
	Proto	9.4	38.2	16.9	4.7
2003	Vinton 81	7.9	39.3	18.2	4.6
	IA 2032 LS	7.8	37.6	20.2	4.5
	Proto	9.0	39.6	16.0	4.7
JSC	Hoyt-F	9.6	36.2	18.4	5.1
Utah	Hoyt-F	10.8	37.5	17.2	4.8
ISU	Hoyt-F	9.0	36.3	19.1	4.6
ISU	Hoyt-H	10.0	32.0	na	na

The protein content of traditional soybean cultivars ranged from 37.6 to 39.6% at 13% moisture. The field grown Hoyt ranged from 36.2-37.5%, whereas the hydroponically grown Hoyt was only 3 (Table 1). In general the Hoyt cultivar was lower in protein, slightly higher in fiber, and similar in oil content. (Table 1). This compositional data was compared to three other cultivars: Vinton 81, the soyfoods industry 'gold standard' high protein and seed size; IA 2032LS, a high protein, large

seeded lipoxygenase free (low beany flavor, non-GMO) cultivar; and Proto, a high protein cultivar from North Dakota that is proported to have a higher anti-oxidant level. All of the food-grade cultivars were higher in protein than the Hoyt cultivars. Figure 2 demonstrates the higher antioxidant capacity of the Proto cultivar based upon PhotoChem (chemluninance) data from this study.

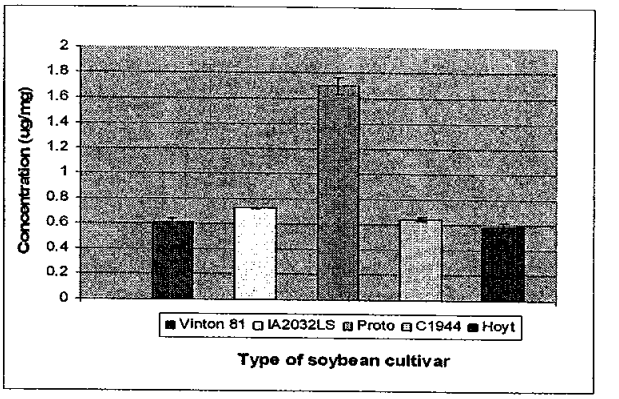


Figure 2. Antioxidant potential of five soybean cultivars.

The new calcium sulfate dehydrate was ordered by NASA. However, it was in a small granular form rather than a powder. We ordered a new food-grade powdered from from Custom Gypsum (OK). In our evaluations it took almost twice the amount of granular form than the powdered form to get the same degree of coagulation. Initial runs of all of the control and treated soybeans were made to determine tha amount of coagulant needed, as was done in NFFP 2003. The Hoyt soybeans from NFFP 2003 still needed 0.052 N calcium sulfate dihydrate to form curds, compared to 0.023-0.025 N for the new Hoyt (Field grown and hydroponic), the cultivars from NFFP 2003, and the new crop control cultivars. The requirement for more coagulant by Hoyt from this years study, confirms our data from last year (NFFP 2003). Likewise, pH and conductivity data matched the data from last year (NFFP 2003). Amounts of needed resuppliable inputs are of concern to NASA because they would contribute to the overall mass of a mission.

Fifty, 100, and 300g batches of each cultivar were soaked and processed to evaluate the new pressing boxes for their ability to scale up or scale down a process. The 300g press worked well and scaled up to the kettle and pilot plant systems. The 100 and 50g were harder to use due to the thinner diameter and higher height. The 50 g tofu box gave consistantly higher yields, due to less efficient pressing, even with the same force/area^2. Due to the small quality of hydroponic beans, the 50 g press box was used. All color and two cubes of tofu/press box were available from each 50 g batch. The hydroponically grown Hoyt soybeans absorbed less water than the field grown Hoyts during the soaking step (Table 2). This was due to the presence (14%) of "stone" or "hardshell" beans. These beans did not rehydrate, even after 24 hours of soaking. There were no differences in the amount of okara or yield of tofu between field and hydroponically grown beans.

Previous studies found that Hoyt beans soak up 2.3 times their weight compared to 2.38 this year for field grown and 2.22 for hydroponically grown (Tables 2 and 3). The Vinton cultivar (commonly used in the soyfoods

Table 2. Comparison of Hydroponically and Field Grown Hoyt Soybeans

	FIELD	HYDROPONIC
Water Uptake	2.38*	2.22
Okara Yield	83.70	1.20
Tofu Yield	81.90	2.27

industry) is much more efficient in making tofu than the other beans. The Vinton cultivar produced tofu that was 2.63 times the dry weight of bean used.

As reported in 2003, the color of seed coat and cotyledon were carried into the soymilk and the tofu. Using the HunterLab Model XE, significant differences in the color of the tofu, okara, whey, and soymilk were found. The Hoyt soybeans from all locations and crop years produced what could be described as an unappealing grayish color soymilk, tofu and okara. The Proto cultivar produced a more tan in color tofu, due its brown hilum. The industry standard Vinton 81 and the IA 2032 LS cultivars produced a more usual creamy light yellow color soymilk, tofu and okara.

Irradiation Study

TABLE 3. SELECT PROCESSING DATA FOR THE PROTO CULTIVAR

Dose	Water Uptake	Okara	Tofu
0	2.37	1.38	2.20
10E	2.44	1.30	2.38
10G	2.46	1.48	2.10
030E	2.40	1.16	1.92

Microbial Load of the Soybean Cultivars

- All soybeans in this study would meet Shuttle Food Microbiological Requirements (They can fly!!)
- No coliforms or Salmonella were found.
- Average total aerobic counts ranged from 0 to 250 CFU/g.
- Yeasts and Molds ranged from 0 to18 CFU/g. Aspergillus flavus was found on one sample, which may be due to contamination after irradiation (sampling).

TBA, A Measure of Oxidation: E-Beam and Gamma

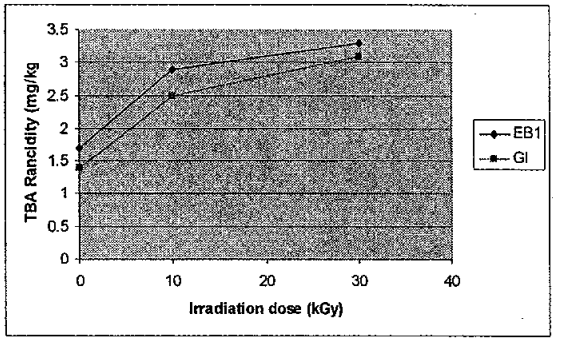


FIGURE 3. TBA RANCIDITY FOR VINTON 81 AT IRRADIATION DOSES OF 0KGY, 10KGY AND 30KGY

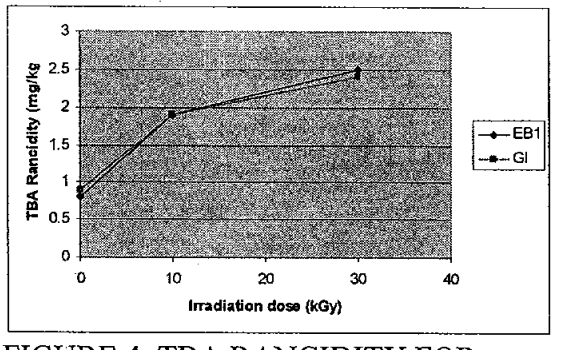


FIGURE 4. TBA RANCIDITY FOR IA2032LS AT IRRADIATION DOSES OF 0KGY, 10KGY AND 30KGY.

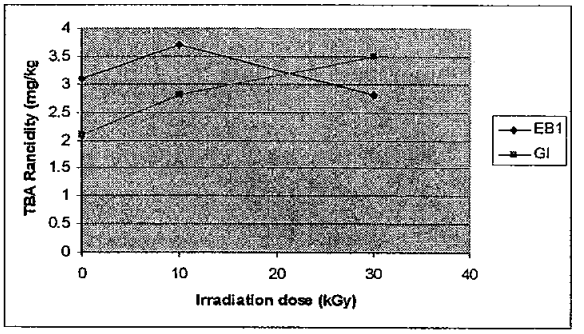


FIGURE 5. TBA RANCIDITY FOR PROTO AT IRRADIATION DOSES OF 0KGY, 10KGY AND 30KGY.

While TBA values increased with irradiation, the TBA levels were lower for the IA 2032 LS. It was also noted that this cultivar had a less rancid/oxisdized odor.

Free Fatty Acids

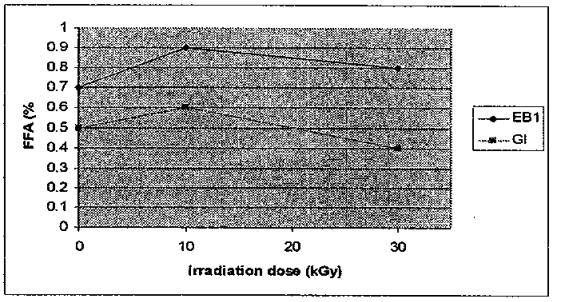


FIGURE 6. PERCENT FREE FATTY ACID (FFA) FOR VINTON 81 AT IRRADIATION DOSES OF 0KGY, 10KGY AND 30KGY

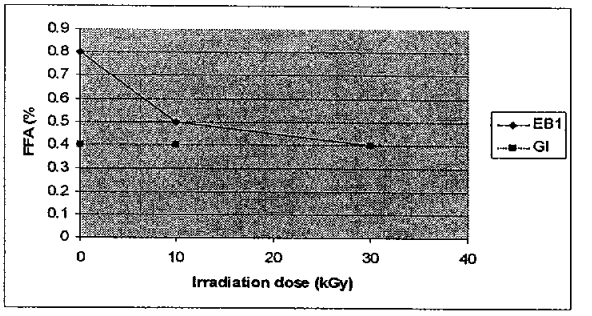


FIGURE 7. PERCENT FREE FATTY ACID (FFA) FOR IA2032LS AT IRRADIATION DOSES OF 0KGY, 10KGY AND 30KG

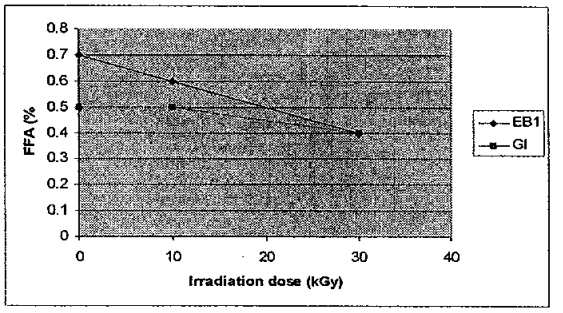


FIGURE 8. PERCENT FREE FATTY ACID (FFA) FOR PROTO AT IRRADIATION DOSES OF 0KGY, 10KGY AND 30KGY.

In general, FFA levels decreased with increasing dosage. However, Vinton 81 FFA decreased at a lower rate than the other two cultivars. Gamma radiation FFA values were lower than the E-bean at 10 kGy. The decrease in FFA may be due to the destruction of the free polyunsaturated FFA, as noted by increasing TBA and oxidized/rancid aroma values.

In addition to the changes in appearance, noted above, the soybeans

irradiated at 30 kGy had less okara (small particles passed the filtering system) and lower tofu yields (Table 3). The tofu also had a softer texture, more pasty.

Appearance and Aroma of the Soaked Soybeans

- Irradiated raw soaked soybeans:
  - Were visually damaged at the ends by the treatment (electron and Gamma)
- 30 kGy caused more damage.
  - Both 10 and 30 kGy soaked beans had a rancid aroma
  - IA2032LS had less of this aroma
- The 30 kGy soybeans were softer than the other treatments.
- After grinding and during cooking, the IA2032LS beans had less rancid-oxidized aroma than the other irradiated cultivars.
- The hydroponically grown Hoyt soybeans contained 'hardshell' or 'stone' soybeans "Like a Rock" to quote an Insurance company.
- The hydroponically grown beans absorbed less water than the field grown Hoyts.
- The hardshell beans were not altered by irradiation treatments.
- The irradiated beans lost more solids into the soak water than the control beans (concern for waste water treatment)
- Control < 10 kGy < 30 kGy
- 0.2 to 4% solids lost into the soak water

Color of the Tofu

TABLE 4. VINTON 81 AND PROTO TOFU COLOR

Cultivar	L*	a	b
Vinton 81			
0E	83.70	0.9	14.9
10E	81.90	1.2	13.7
30E	80.95	1.7	12.9
0G	85.91	0.8	17.6
10G	84.96	1.4	16.5
30G	82.92	1.9	16.2
Proto			
0*	87.82	0.8	14.9
10E	86.65	1.2	14.6
10G	85.76	1.3	15.1
30E	84.57	1.7	15.2

E= Electron G=Gamma * L= 0 (black) /100 (white); a-green/ + red; b-blue/ + yellow

Color Summary

- Color from the seed coat and cotyledon were carried into the soymilk and the tofu, as reported by Wilson (NFFP 03).
- Visual and instrumental methods detected these color changes.
- Electron and gamma irradiation to obtain pasteurization or sterility reduced the lightness and yellowness, white increasing the redness (more tan color) of the soymilk and tofu (Table 4).
- These color changes occurred for all cultivars.

Texture of the Tofu

TABLE 5. VINTON 81 TOFU TEXTURE AS INFLUENCED BY IRRADIATION OF THE SOYBEANS

Dose (kGy)	Hardness (kg)	Adhesiveness Kg-sec	Springiness	Cohesiveness	Resilience
E-Beam					
0	1.454	-0.00751	0.875	0.712	0.380
10	1.247	-0.01087	0.870	.0674	0.317
30	1.006	-0.03474	0.865	0.520	0.220
Gamma					
0	1.393	-0.02546	0.874	0.684	0.348
10	1.287	-0.03056	0.872	0.593	0.288
30	.0522	-0.05216	0.748	.0386	0.124

Texture Summary

- **Irradiation of the soybeans resulted in (Table 5):**
  - Softer tofus
  - Less adhesive curds
  - Less springiness
  - Less cohesiveness
  - Less resilience
  - Lower yields
  - More okara leaking into the tofu

The comparative studies showed that the Vinton 81 soybeans were more efficient in tofu making than the Hoyt. The Vinton soybeans yielded 263% of their dry weight in tofu while the Hoyt yielded only 120% of their dry weight in tofu.

Finally, the Vinton 81 soybeans (which are standard in industry) produced better sensory qualities than the Hoyt beans. The color was more similar to commercially available tofu for the Vinton 81. Although no taste panels were conducted, it was noted that the aroma during the making of the two tofus was much more pleasant for the Vinton beans.

For these reasons, we feel that NASA should consider replacing the hydroponic Hoyt beans with Vinton 81. As an alternative to growing their own soybeans, the Vinton beans would be sent with crew as a dry ingredient. Although this would take up more space, the benefits of more abundant, high-quality finished soy product using less of a chemical input make this scenario appropriate for future research.

CONCLUSIONS

- While the use of irradiation as a HACCP CCP will help prevent food-borne illness hazards to the crew, 10-30 KGy causes undesirable sensory, yield, and physical changes in the soymilk and tofu. This will most likely translate into altered functional properties when used as ingredients in other foods.
- Aroma, color and textural changes made unacceptable soymilk and tofu.
- The natural antioxidant level in the soybean cultivars was not sufficient to protect the soybeans from these dose levels.
- The lipoxygenase-free soybean cultivar was less oxidized, due to (Hypothesis) the lack of enzyme-substrate interaction after the radiation has, essentially, punched holes in the membranes and structural material.
- Based upon a small sample size, the hydroponically grown Hoyt behaved similarly to field grown Hoyt cultivars, with the exception of 'stone' beans.
- As noted in my NFFP2003 report, the black hilum and seed coat staining detracts from the quality of the tofu and it's okara. A clear hilum cultivar should be used.
- Counter measures could include vacuum packaging, radiating under freezing conditions.
- A No Effect Dose for food quality, below 10 kGy needs to be determined.
- Better estimates of the radiation that the food will be exposed to need to determined and shared.
- Appropriate shielding for the food as well as the astronauts needs to be developed.
- The actual doses that the crew and food will experience during normal and solar flair transit to Mars needs to be determined (or made known).
- More hydroponically grown soybeans are needed in order to verify the finding of this study.

Deliverables

A new small scale system for evaluating soybeans was developed using 50 g quantities of soybeans. In a separate study with Dr. Juming Tang, the retotable pouch procedures and processing system at Texas A&M were evaluated and a White paper was produced. Action was taken, based upon the recommendations from this report.

References

1. Bennion, M. and B. Scheude (2004). "Food Preservation and Packaging." Introductory Foods, Chapter 28: 667-669. Prentice Hall, N.J.
2. Hammond, E.G. and W.R.Fehr. (1975). Oil Quality Improvement in Soybeans-Glycine max (L.) Merr. Sonderdruck aus fette seifen anstrichmittel. 77: 97-101.
3. Johnson, L.D. and L.A. Wilson, 1984. Influence of soybean variety and method of processing in tofu manufacturing: comparison of methods for measuring soluble solids in soymilk. J. Food Sci. 49 1:202-204.

4. Liu, K. 1997. Soybeans: Chemistry, Technology, and Utilization.ITP International Thomson Publishing. New York. p114-217.
5. Moizuddin, S., G. Buseman, A.M. Fenton, and L.A. Wilson. 1999. Tofu production from soybeans or full-fat soyflakes using direct and indirect heating processes. J. Food Sci. 64:145-148.
6. Moizuddin, S., L.D. Johnson, and L.A. Wilson. 1999. Rapid method for determining optimum coagulant concentration in tofu manufacture. J. Food Sci. 64:684-687.
7. Murphy, P.A., H.P. Chen, C.C. Hauck, and L.A. Wilson. 1997. Soybean Storage Protein Composition and Tofu Quality. Food Technology 51(3) 86-88, 110.
8. Potter, N.N. and Hotchkiss, J.H. (1995). Food Science, 5th edition. Gaithursburg, MD, Chapman and Hall.
9. Torres-Penaranda, A.V., C. Reitmeier, L. A. Wilson, W. Fehr, and J.M. Narvel. 1998. Sensory characteristics of soymilk and tofu made from lipoxygenase-free and normal soybeans. J. Food Sci. 63:1084-1087.
10. Watanabe, T., et al. 1964. Research into the standardization of the tofu making process. National Food Research Institute Report (apan). Parts 1-3.
11. Wilson, L.A., N.P. Senechal, and M.P. Widrlechner. (1992). Headspace analysis of the volatile oils of Agastache. J Agric. Food Chem. 40:1362-1366.
12. Wilson, L.A., 2003 NFFP Report, Chapter 20.
13. Wilson, L.A., P.A. Murphy, and P. Gallagher. 1992. Soyfood. Product markets in Japan: U.S. export opportunities. MATRIC, Ames, IA. Library of Congress #92-60367. 64 p.
14. Wilson, L.A. 1995. "Soyfoods" in Practical Handbook of Soybean Processing and Utilization. Chapter 22. D.R. Erickson, Editor. AOCS Press and USB Press, St. Louis, MO. pp 428-459.
15. Wilson, L.A. 1996. "Comparison of Lipoxygenase-null and Lipoxygenase-containing soybeans in foods." Chapter 12. In: Lipoxygenase and Lipoxygenase Pathway Enzymes. G.J. Piazza, Editor. AOCS Press, St. Louis, MO. pp. 209-225.

DEVELOPING A FRAMEWORK FOR EFFECTIVE NETWORK CAPACITY PLANNING

Final Report
NASA Faculty Fellowship Program – 2004
Johnson Space Center

Prepared By:	Ece Yaprak, Ph.D.
Academic Rank:	Associate Professor
University & Department:	Wayne State University College of Engineering Division of Engineering Technology Detroit, MI 48202
NASA/JSC	
Directorate:	Information Resources Directorate
Division:	Information Technology Division
Branch:	Communications Branch
JSC Colleague:	Jose Nunez
Date Submitted:	July 9, 2004
Contract Number:	NAG 9-1526 and NNJ04JF93A

ABSTRACT

As Internet traffic continues to grow exponentially, developing a clearer understanding of, and appropriately measuring, network's performance is becoming ever more critical.

An important challenge faced by the Information Resources Directorate (IRD) at the Johnson Space Center in this context remains not only monitoring and maintaining a secure network, but also better understanding the capacity and future growth potential boundaries of its network. This requires capacity planning which involves modeling and simulating different network alternatives, and incorporating changes in design as technologies, components, configurations, and applications change, to determine optimal solutions in light of IRD's goals, objectives and strategies.

My primary task this summer was to address this need. I evaluated network-modeling tools from OPNET Technologies Inc. and Compuware Corporation. I generated a baseline model for Building 45 using both tools by importing "real" topology/traffic information using IRD's various network management tools. I compared each tool against the other in terms of the advantages and disadvantages of both tools to accomplish IRD's goals. I also prepared step-by-step "how to design a baseline model" tutorial for both OPNET and Compuware products.

INTRODUCTION

As networks became more and more involved, deploying different and sometimes incompatible technologies and trying to monitor the behavior of these networks become increasingly difficult. Round-the-clock network availability along with optimized network and application performance is mission-critical to IRD. For this reason, products like OPNET by OPNET Technologies Inc. and Vantage suite by Compuware Corporation enable users to be more effective in understanding their IT infrastructure, applications, and its anticipated network growth.

Capacity planning in this regard helps provide a clearer understanding of the limits to, and capacity of, IRD's network infrastructure before changes are implemented in light of evaluation of various scenarios. For instance, the individual merits of new technology (such as VoIP) can be assessed and planned before deployment. These tools are able to model the network, and measure the network and application performance. Understanding the detailed traffic volumes, flows, and network architectures is the first step in identifying and correcting problems before any new deployment [1,2].

In order to establish and evaluate various design alternatives however, a baseline model of the IT infrastructure must be generated first. Importing "real" network traffic and topology information from various network management tools can create baseline model. Using this baseline model, "what-if" (sensitivity) analyses can be conducted by altering specific attributes of the baseline model to determine when the infrastructure under analysis might exceed capacity, how re-routing might affect network performance, and how much load a link is able to handle before it begins to degrade. This process can help identify appropriate network design options based on different utilization characteristics and selected topology changes. Both OPNET and Compuware products are evaluated for network capacity planning and baseline models were developed using both packages. Below both products are summarized and then the developed model is explained.

PRODUCT REVIEW

As networks get more complex, with diverse set of technologies and more stringent requirements for application performance, problem solving becomes more difficult. When slowdowns of mission-critical applications occur, IT staff needs tools to track the source of the bottlenecks. In order to troubleshoot performance problems, IT staff needs a complete picture of the IT infrastructure first. There are a variety of tools to help IT staff more effectively design and deploy networks, diagnose network and application performance problems, and predict the impact of network changes. Of those, OPNET and Compuware products are summarized below.

OPNET Technologies, Inc. is a leading provider of management software for networks and applications. **OPNET's IT Guru** enables users to model the entire network, including its routers, switches, protocols, servers, and the individual applications they support. The function of the various modules is summarized below. Of those, Multi-Vendor Import and Flow Analysis modules were used extensively along with other network management tools in our study [3]:

- **Multi-Vendor Import (MVI) Module** is able to import both the topology as well as traffic information in order to create an accurate baseline model.
- **Flow Analysis Module** provides the capability to visualize traffic flows and analyze the impact of failures.
- **Net Doctor Module** identifies potential or existing trouble spots in the network, as well as helps determine whether or not the network is optimally configured.
- **Expert Service Prediction (ESP) Module** helps determine the topology and traffic service projections, perform iterative simulations to automatically analyze the impact of increased load over time.
- **Application Troubleshooting and Deployment (ACE) Module** helps to capture, filter, and synchronize applications from multiple network segments.
- **Application Decode (ADM) Module** enhances the visualizations and diagnoses offered by ACE module, using the application and protocol decode engine from Sniffer Technologies.
- **Virtual Network Environment (VNE) Server** module provides an on-line, integrated view of the network by collecting data and creating a unified network view for planning, engineering, and operations.

Compuware Corporation has products for every aspect of the application life cycle. **Compuware's Vantage** suite is an application service management solution to manage application performance from the end-user perspective. It is able to troubleshoot application performance problems of Web-services-based applications in production. This helps users to understand why transactions are not meeting their expected service level agreements (SLAs) by giving insight into the transaction's programming components. Response time metrics, integrated with end-to-end performance analysis proactively identify and solve performance problems. It also has a number of modules and the function of various modules is summarized below. Application Expert and Predictor are used in our study.

- **ClientVantage** ensures that applications are available and perform at acceptable levels by tracking response times, resource usage, application faults and availability.
- **ServerVantage** monitors servers, applications and databases and produces web-based network management reports.

- **NetworkVantage** handles network performance management with an application perspective by uncovering, who the infrastructure is serving, what applications demand the most resources and how to troubleshoot performance problems.
- **Application Expert** is able to demonstrate how changes in network bandwidth, latency, load and TCP window size affect each end user's response time.
- **Predictor** is able to perform simple growth planning and WAN provisioning based on the key performance metric, the application response time experienced by the end users.

The following section summarizes the baseline models that were developed using both products.

MODELING

The baseline network infrastructure model is developed using the OPNET's MVI module for Building 45. MVI module lets IT people to import both the topology as well as traffic information in order to create an accurate model of the current network. This module is able to obtain network topology, configuration, and utilization data from a variety of sources, leveraging existing information to enable advanced network and application troubleshooting and planning [4].

Network topology is created first using this module. There are a number of different options to create this topology such as creating an empty scenario and building the topology manually using the icons provided, or using the importing option of the MVI module. We have chosen the "importing" function because it gives us the exact network topology. The MVI module is able to pull data from sources such as CiscoWorks, HP OpenView, and various other management tools. We have utilized CiscoWorks for the generation of the necessary device configuration files (config, cdpneighbors, vlan and version) for each switch and router [5].

The MVI module's "Import Device Configurations" function lets you to either import a completely new topology or lets you to import only some of the modified devices (using the files generated). Initially, an entire new topology for Building 45 is generated as can be seen in Fig. 1. This topology also includes Building 46 and 32 core switches. Fig. 2 shows the multilayer topology. If one of the icons is double-clicked, you can see the underlying topology. For example, by double-clicking the B45 icon, you can get the topology seen in Fig. 1.

Once the topology is created, then the traffic information needs to be added. There are different ways of adding this traffic information. The network model can be loaded with "traffic matrix" information, which represents end-to-end traffic flows. Alternatively, a background traffic load of the baseline network can be loaded onto the model. Traffic and/or utilization data can be loaded from various sources, such as Cisco NetFlow

Collector, HP OpenView, Sniffer Pro, MRTG, etc. First, we have used MRTG traffic data and created the model. Later, the traffic flow was collected from Sniffer and used for importing. (Switch and router names on Fig. 1 are deliberately cleared).

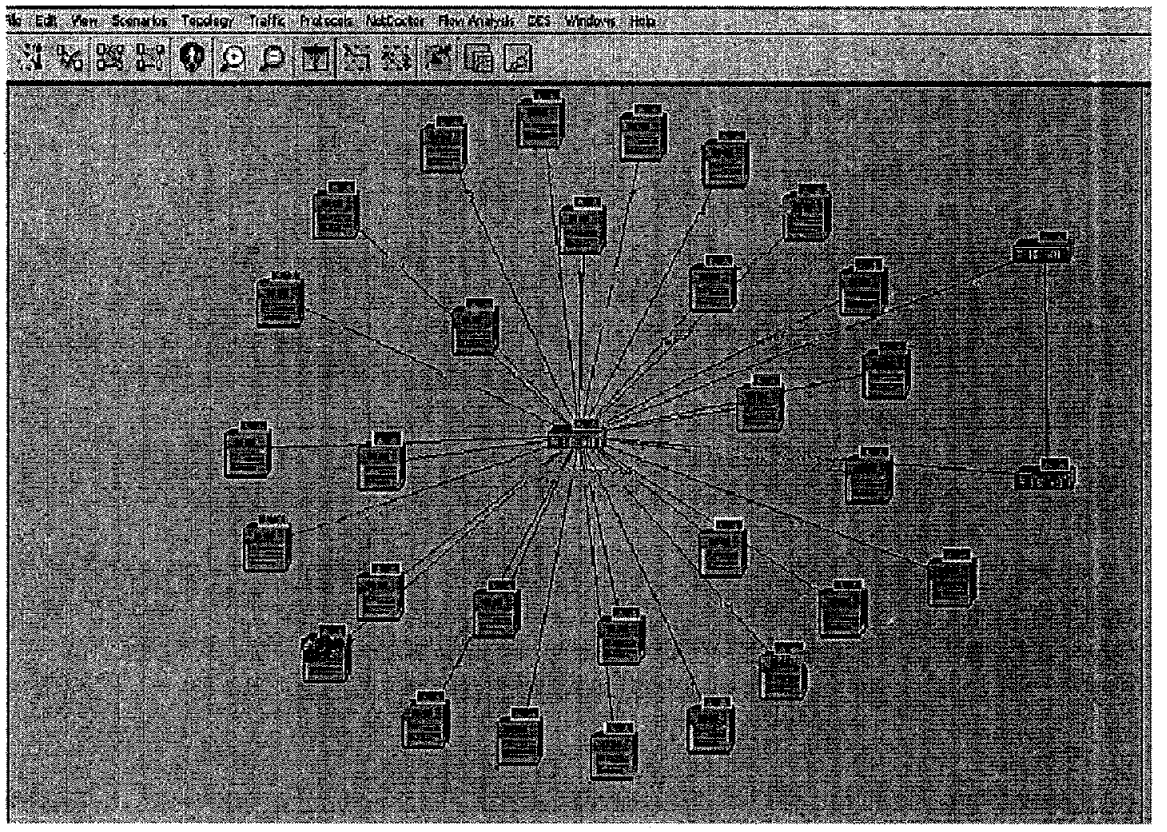


Fig. 1. Building 45 Network Topology

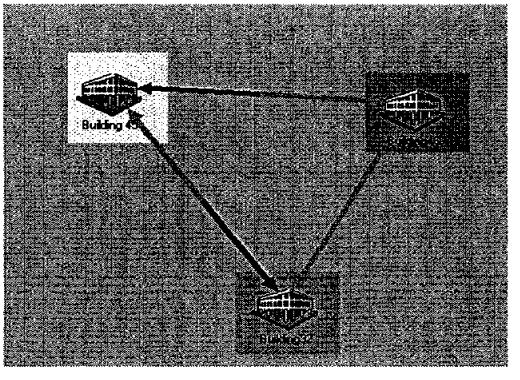


Fig. 2. Multilayer Topology

Once the baseline model is created, the Flow Analysis Module of OPNET is used in order to provide us the capability to visualize traffic flows, analyze the impact of failures, and design fault tolerant networks by pinpointing the actual and potential bottlenecks of the network. This module is also capable of predicting the impact of router failures that uses multiple IP routing protocols and showing how utilizations change in the network over a specific time period. Various global, node and link statistics, such as link utilization, throughput in bit/sec and traffic sent and received were also collected.

Baseline model was also constructed using the **Compuware** tools. For that purpose, a Sniffer file was used to create new tasks in Application Expert. Both filtered and unfiltered trace files were used for the creation of various tasks. We have used Response Time Analysis chart to show the impact that various nodes have on the overall response time of a task. This chart shows the response time and percentage of total response time attribute to each node in the task. We also have used Thread Analysis to view the files or commands such as SQL statement or HTTP commands an application is sending over the network to easily understand the time and duration of a thread and the relationship between threads [6].

Once the tasks were created using Application Expert, then the topology was created. We have created the topology for Building 45 manually using Predictor. The user profiles, which were created with Application Expert, were used for the traffic generation. Traffic import in this case did not work because the Compuware software was not able to detect any of the Sniffer capture cards. Fig. 3 shows the model, which was developed manually using Predictor.

CONCLUSION

Both OPNET and Compuware products were evaluated for network capacity planning. Baseline model for Building 45, which includes the topology and traffic information, was built using both products. Compuware Vantage suite is easier to model. It quickly provides application performance from the end user perspective. Modeling capabilities of OPNET on the other hand, was superior. It is difficult to model; however, detailed network and application performance analysis can be done using this product.

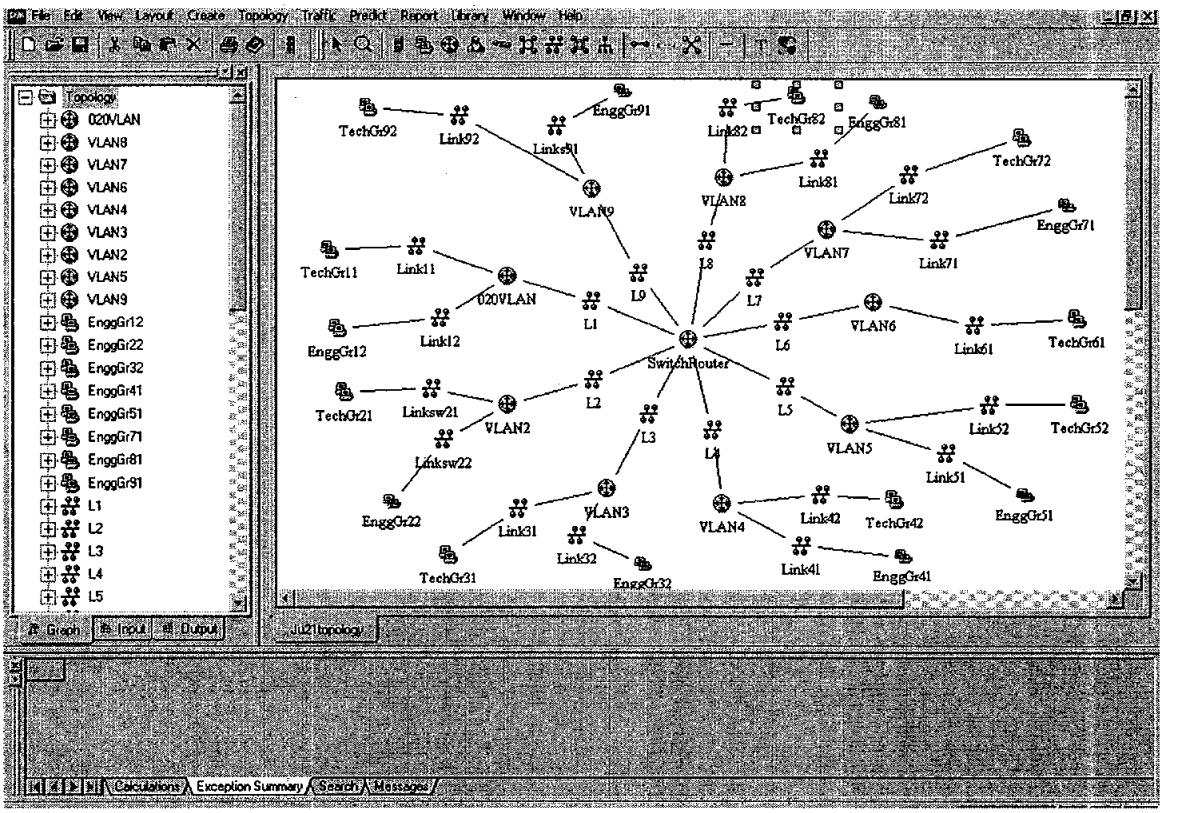


Fig. 3. Building 45 Network Model using Compuware

REFERENCES

1. Behrouz Forouzan, "Data Communications and Networking," McGraw Hill, 2001.

2. Mani Subramanian, "Network Management: Principles and Practice", Addison-Wesley, 2000.

3. Network Monitoring Tools, http://www.slac.stanford.edu/xorg/nmtf/nmtf-tools.html

4. Opnet ITGuru, http://www.opnet.com/products/itguru/home.html

5. CiscoWorks, http://www.cisco.com/warp/public/44/jump/ciscoworks.shtml

6. Compuware's Predictor, http://www.compuware.com/products/vantage/1020_ENG_HTML.htm

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 2005	NASA Contractor Report

4. TITLE AND SUBTITLE
NASA Summer Faculty Fellowship Program 2004, Volumes 1 & 2

5. FUNDING NUMBERS

6. AUTHOR(S)
Edited by William A. Hyman, Donn G. Sicorez , and Dawn M. Leveritt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lyndon B. Johnson Space Center
Houston, Texas 77058

8. PERFORMING ORGANIZATION REPORT NUMBERS
S-961

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
CR-2005-213690

11. SUPPLEMENTARY NOTES
2 Volumes. Volume 1 186 pages, voilue 2 162 pages.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Available from the NASA Center for AeroSpace Information (CASI)
7121 Standard
Hanover, MD 21076-1320 Category: 99

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The 2004 Johnson Space Center (JSC) National Aeronautics and Space Administration Faculty Fellowship Program (NFFP) was conducted by Texas A&M University and JSC. The program was funded by the Office of Education, NASA Headquarters, Washington, D.C. and by JSC. Each faculty Fellow spent at least 10 weeks at JSC (or the White Sands Test Facility) engaged in a research project in collaboration with a NASA/JSC colleague.

This document is a compilation of the final reports on the research projects done by the Fellows during the summer of 2004. Volume 1 contains reports 1 through 12 and Volume 2 contains reports 13 through 22.

14. SUBJECT TERMS	15. NUMBER OF PAGES	16. PRICE CODE
Human performance, abilities; life support systems; systems engineering; Medical science, cardiology; aerospace medicine; communication; exploration	338	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited