NASA / TM-2000-210691

Construction and Utilization of a Beowulf Computing Cluster: A User's Perspective

Jody L. Woods, Jeff S. West
Lockheed Martin Space Operations – Stennis Programs
Stennis Space Center, MS

Peter R. Sulyma
NASA / John C. Stennis Space Center
Stennis Space Center, MS

October 2000

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 counterpart of peer-reviewed formal professional papers, but having 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 co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, 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***

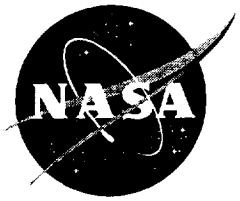
- E-mail your question via the Internet to help@sti.nasa.gov

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

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

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

NASA / TM-2000-210691



Construction and Utilization of a Beowulf Computing Cluster: A User's Perspective

Jody L. Woods, Jeff S. West
Lockheed Martin Space Operations – Stennis Programs
Stennis Space Center, MS

Peter R. Sulyma
NASA / John C. Stennis Space Center
Stennis Space Center, MS

National Aeronautics and
Space Administration

John C. Stennis Space Center
Stennis Space Center, MS 39529

October 2000

Acknowledgments

The authors would like to acknowledge the assistance of those individuals who aided in the construction of the Beowulf Cluster described herein. Kristen Riley (NASA - John C. Stennis Space Center) was solely responsible for locating and obtaining funding for the cluster's hardware and construction. Lester Langford (Lockheed Martin Space Operations – Stennis Programs) was instrumental in assembly of the cluster hardware and supporting infrastructure. The project was developed by Lockheed Martin Space Operations under NASA Contract No. NAS13-650 at the John C. Stennis Space Center

Available from:

NASA Center for AeroSpace Information
7121 Standard Drive
Hanover, MD 21076-1320
301-621-0390

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
703-605-6000

Abstract

Lockheed Martin Space Operations - Stennis Programs (LMSO) at the John C. Stennis Space Center (NASA/SSC) has designed and built a Beowulf Computing Cluster which is owned by NASA/SSC and operated by LMSO. A Beowulf Computing Cluster is a relatively new technology in which a collection of standard PC's operates as a single super-computer. This allows for super-computer performance to be achieved using off-the-shelf or commodity PC equipment and offers the best price to performance ratio available. The design and construction of the NASA-SSC/LMSO cluster are detailed in this paper. The cluster is currently used for Computational Fluid Dynamics (CFD) simulations. The CFD codes in use and their applications are discussed. Examples of some of this work are also presented. Performance benchmark studies have been conducted for the CFD codes being run on the cluster. The results of two of the studies are presented and discussed. The cluster is not currently being utilized to its full potential; therefore plans are underway to add more capabilities. These include the addition of structural, thermal, fluid, and acoustic Finite Element Analysis codes as well as real-time data acquisition and processing during test operations at NASA/SSC. These plans are discussed as well.

Contents

Figures

Tables

Introduction

Computational Fluid Dynamics (CFD) codes are used by the Test Technology Development Division in support of rocket engine testing at the John C. Stennis Space Center (NASA/SSC). One of their primary uses at this time is to support plume-induced environment studies. In the context of NASA/SSC operations, plume-induced environment problems are primarily situations in which the hot gases expelled from a rocket engine impinge on solid structures such as a flame deflector, concrete pad, or any other facility structure. CFD may be used in these situations to ensure that a particular flame deflector operates as desired, or that the concrete pad, etc. will not be damaged during engine testing. Another use of CFD is to determine optimum locations for sensors that measure flow properties of rocket exhaust plumes. CFD has also been used to ensure that NASA/SSC meets EPA guidelines by estimating the amounts of pollutants released into the atmosphere by rocket engine tests. The above examples are only a few of the many other interesting problems at NASA/SSC that have been investigated with CFD analyses.

CFD codes have become more and more useful at SSC in the past few years. However, CFD codes are computationally intensive and therefore expensive to use. As Test Technology Development's CFD capability matured, there was an evident need for more computing horsepower. To this end, a study was conducted in 1998 to determine the best approach for acquiring high performance computational capabilities. Both the traditional high performance computing solutions and the new distributed computing cluster technology, or Beowulf Computing Cluster [1,2], were investigated. In a Beowulf Computing Cluster, a collection of standard PC's operates as a single super-computer. This allows for super-computer performance to be achieved using off-the-shelf or commodity PC equipment and as confirmed in the study, offers the best price to performance ratio available. The study showed that it would cost much less to build a Beowulf cluster than to buy a traditional high-performance computing platform with similar performance. Figure 1 illustrates this fact. As a result of the study, the Beowulf approach made much more economical sense and was the adopted plan. It was decided to build a 48-node cluster with flexible upgrade options. Just to emphasize the good economics of Beowulf clusters, the projected cost of the proposed cluster at NASA/SSC was about $100K. The projected cost of a SGI ORIGIN of similar performance was about $1M.

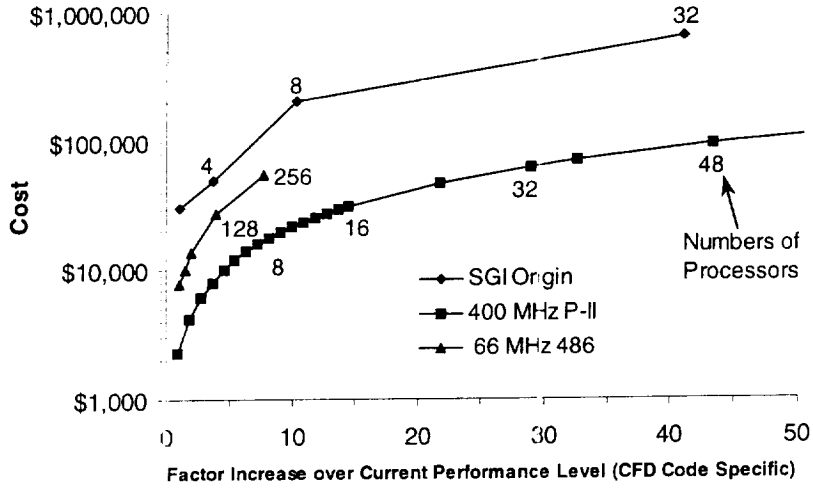


Figure 1: Cost Projections in 1998 of Various Hardware Options

The proposed Beowulf Computing Cluster was funded by NASA/SSC through the Space Shuttle Upgrades Program. It was then designed and built by Lockheed Martin Space Operations – Stennis Programs (LMSO). Currently LMSO uses and maintains the cluster in support of Test Technology Development and NASA/SSC test operations.

SSC Beowulf Cluster Description and Specifications

The hardware in the cluster consists of nodes, network switches, power conditioning equipment, and support equipment such as racks and ventilation fans. Each node is a separate PC with one or more network interface cards. There are different classes of nodes with different functions in this cluster. There are currently 50 compute nodes, a master node, a file-server node, and a gateway node. A summary of the node hardware and memory specifications is presented in Table 1.

Function	Quantity	CPU(s) / Architecture	Local Memory (MB)	Local Storage
Master Node	2	Dual Intel PIII-933 MHz	2048	30.0 GB EIDE
Compute Node	1	Intel PII-400 MHz	256	6.4 GB EIDE
Compute Node	48	Dual Intel PII-400 MHz	Various (128 - 1024)	6.4 GB EIDE
File Server Node	1	Intel Pentium-Pro 200 MHz	64	1.9 GB SCSI
Gateway Node	1	Intel PII-266 MHz	128	6.4 GB EIDE
Cluster Memory: 21.6 GB – currently expandable to 56 GB with no additional node purchasing Cluster Hard Disk Space: 400 GB				

Table 1: Summary of Node Specifications

The nodes communicate via Fast Switched Ethernet. The network currently consists of three 24-port 100 Mbit/sec 3-Com Super Stack Ethernet switches and a 16-port 1000 Mbit/sec Myrinet switch. The 3-com switches operate as a single unit such that any node has a dedicated 100 Mbit/sec connection directly to any other node. A schematic of the network topology is shown in Figure 2. All nodes have 100 Mbit/sec Intel network interface cards that are connected to one of the 3-Com switches. Sixteen of the nodes also have 1000 Mbit/sec Myrinet network interface cards that are connected to the Myrinet switch. The gateway node has an additional network interface card that is connected to the SSC Intranet. Two SGI workstations, a Windows PC, and a Mac PC are also networked to the cluster. These are used as terminals into the cluster and for pre- and post-processing, report generation, etc. As evident in Figure 2, with sixteen 100 Mbit/sec ports currently unused, room has been left for expansion before an additional Super Stack unit is required. The 3-Com Super-Stack allows for up to four 24 port switches to be joined into one Fast Switched Ethernet. One more Super Stack unit could be added and populated with 24 additional nodes without degradation of network performance. This gives the cluster a capacity of 96 nodes without major configuration changes. The total capacity of processors, RAM, and hard disk space would depend on the hardware makeup of additional nodes. Furthermore, any upgrade path is possible with major changes to the network topology.

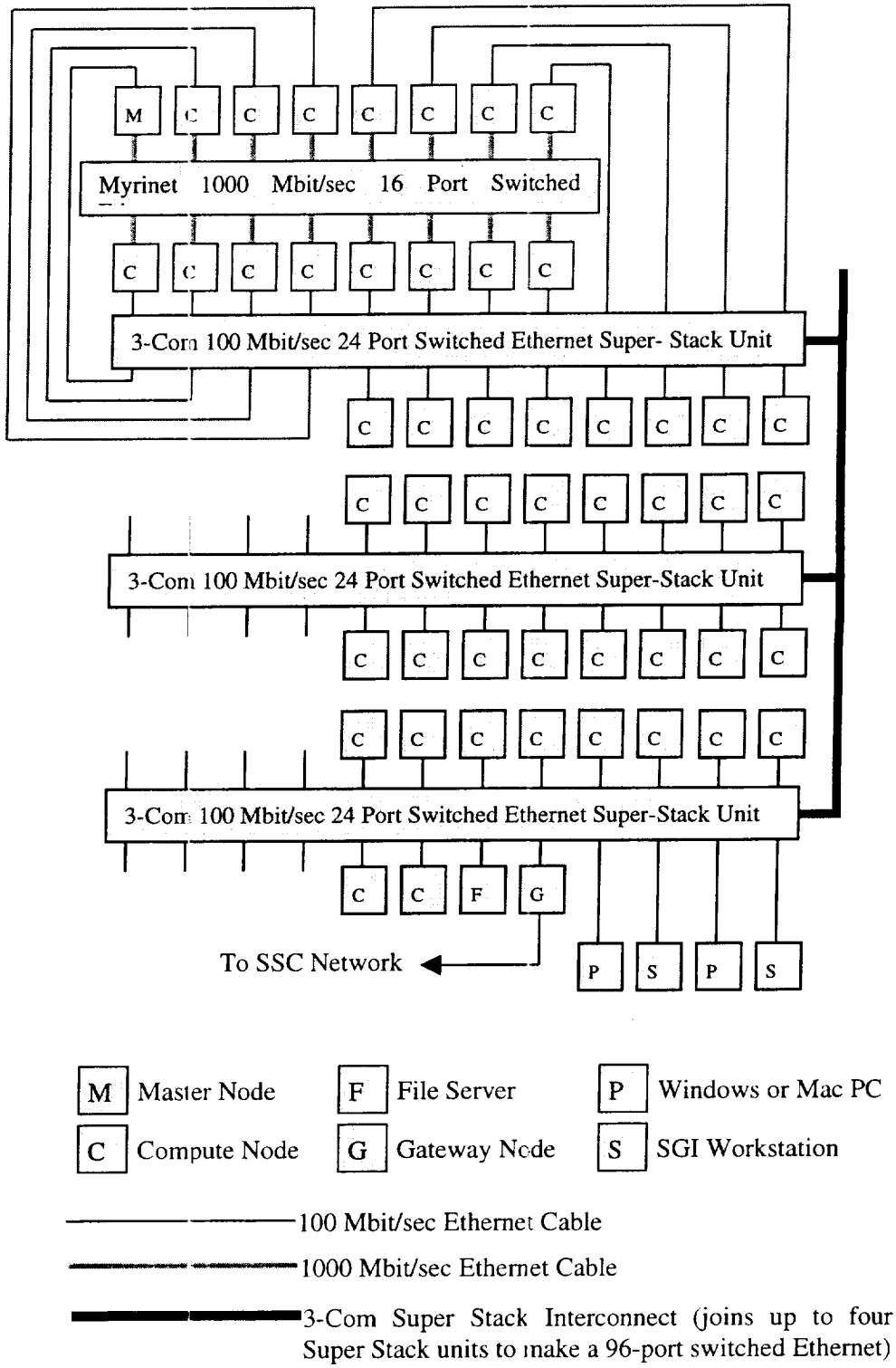


Figure 2: Cluster Network Topology

The cluster is housed in a dedicated 10-ft by 20-ft air-conditioned room. A photo of the room is shown in Figure 3. The 48 homogeneous compute nodes are housed in medium-tower cases that sit on two movable racks of shelves with additional fans for ventilation. The remaining nodes, housed in medium-tower cases as well, are located on tables in the room. Power conditioning and battery backup are provided by three high-capacity uninterruptable power supplies (UPS) and three smaller UPSs. The two SGI workstations and the Windows and Mac PCs that are networked to the cluster are in different locations in the same building as the cluster room.

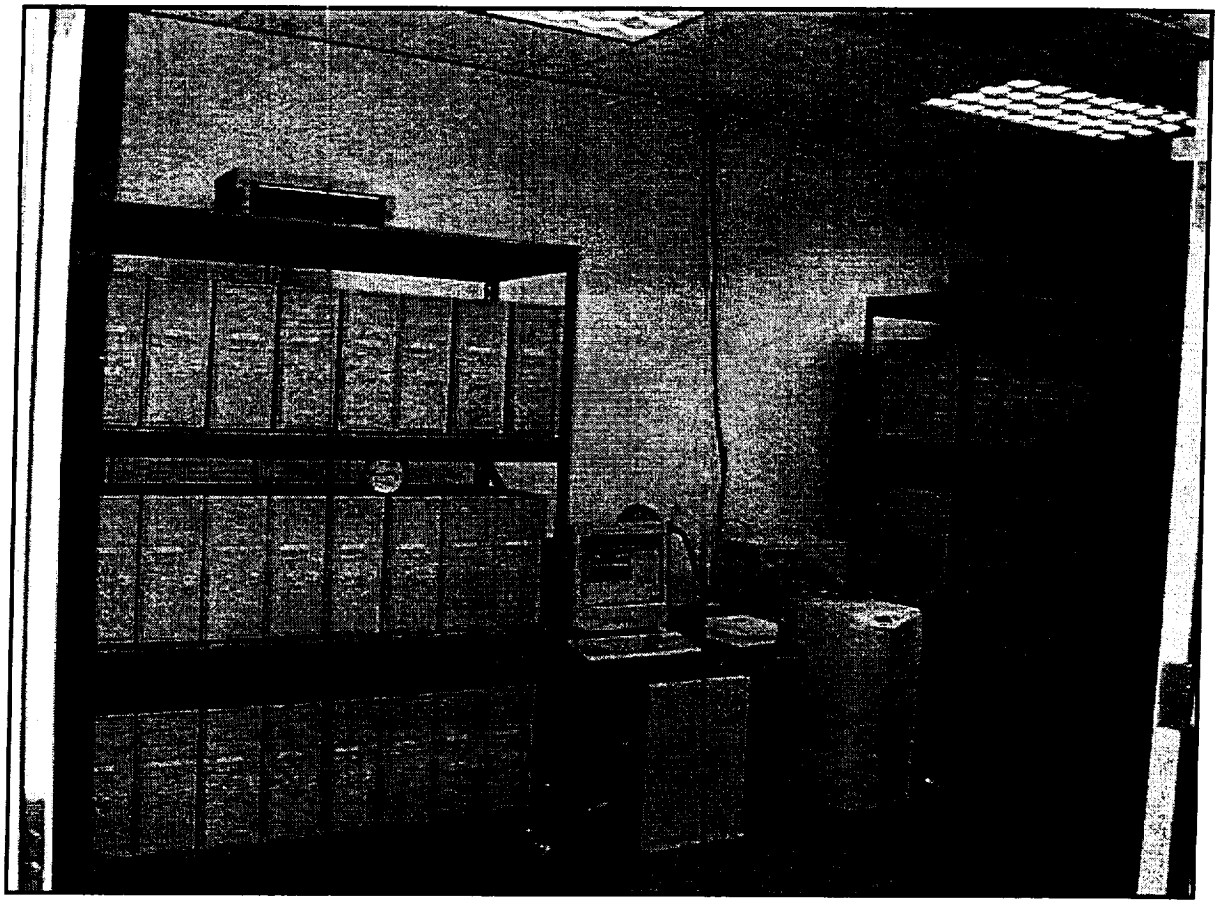


Figure 3: Cluster Room

Each of the nodes is running the Red Hat LINUX 5.2 operating system [3] with a cluster-optimized kernel from its local drive. The nodes also have local copies of the clustering software Parallel Virtual Machine (PVM) [4,5] and Message Passing Interface (MPI) [5,6,7]. The application codes that are run on the cluster are located on the file server node and exported to the other nodes using NFS. These include CFD codes, acoustics prediction codes, and FORTRAN and C/C++ development environments. User accounts are maintained on a single node and the account information is mirrored on the other nodes using scripts. The gateway node handles security for the cluster-to-outside-world connection. The security measures implemented include the use of TCP wrappers and only allowing connections from within the NASA/SSC Intranet which has its own comprehensive security policy in place.

The final cost breakdown for the design and construction of the NASA/SSC cluster is presented in Table 2. As discussed in the introduction, the projected cost for the cluster was around $100K. The actual cost of $127K is in line with this projection. The cost breakdown for this cluster should be typical of any other small to moderately sized Beowulf cluster built from off-the-shelf components. If a very large cluster was being built, commodity components could be purchased thus lowering the percentage of total cost for computer hardware as well as the overall cost per node.

Description	Cost	Percent of Total
Computer Hardware	$ 107,430	84.5%
Supporting Infrastructure	$ 13,607	10.7%
Labor	$ 6,069	4.8%
Total Cost	**$ 127,106**	

Table 2: Actual Costs

Cluster Utilization and Performance

Currently, the only applications in use that take full advantage of the cluster's computing power are two government owned CFD codes. These CFD codes are highly advanced programs that numerically solve the 3-dimensional Navier-Stokes equations for multi-species, compressible, reacting flows. They are implemented to run in parallel on Beowulf clusters.

Figure 4 presents the results of a CFD analysis of a generic plume impingement problem. This is a visualization of the flow-field of a plume impinging on a flat plate oriented perpendicular to the plume centerline. The Mach Number contours are shown in the figure. The impact pressure and shear forces imposed upon the plate and the cold-wall heat flux to the plate can be determined from this analysis.

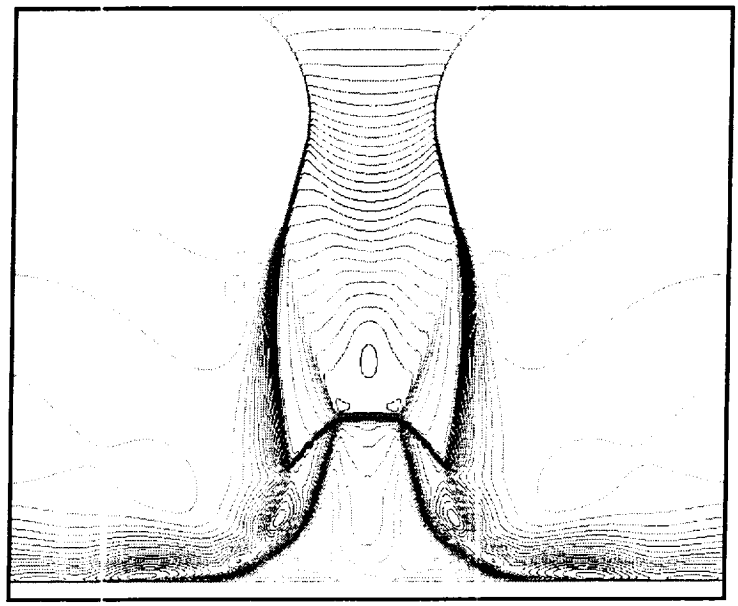


Figure 4: Example of CFD Plume Impingement Analysis

CFD codes that are optimized for parallel execution achieve near linear speed-up behavior. As an example, Figures 5 and 6 illustrate the reduction in analysis time obtainable by increasing the number of processors used in a typical but relatively small problem analyzed using CFD. The problem consisted of a 100,000 point computational grid. In this case, the grid was ideally optimized for parallel execution. This was done by dividing the grid into equally sized zones, one for each processor used. Dividing a grid into multiple zones for parallel execution is referred to as domain decomposition. In this scheme, each processor works on its own zone. The processors must also communicate zone boundary information to each other throughout the analysis. Figure 5 shows the number of iterations executed each second as the number of processors used is increased. Figure 6 shows the time required to perform 200 iterations as the number of processors is increased. A typical flow-field analysis may require anywhere from 500 iterations to 20,000 or more depending on the physical phenomena involved and the desired results.

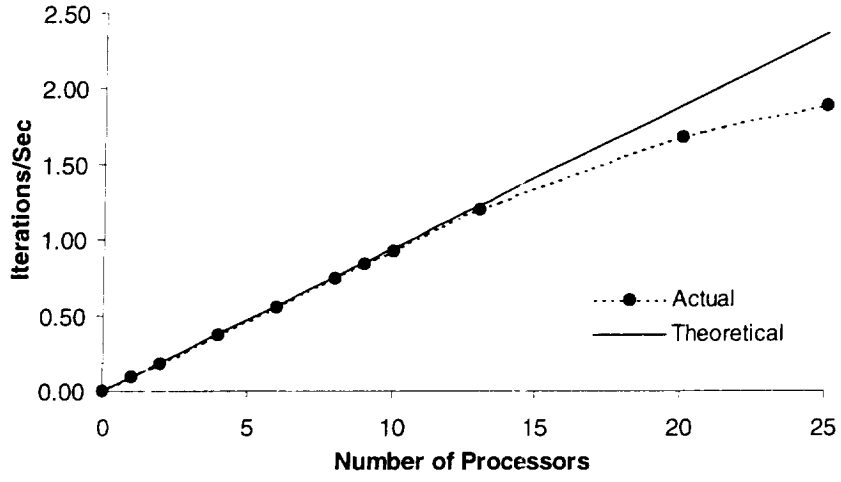


Figure 5: Parallel Speedup for an Ideal CFD Grid: Iterations per Second

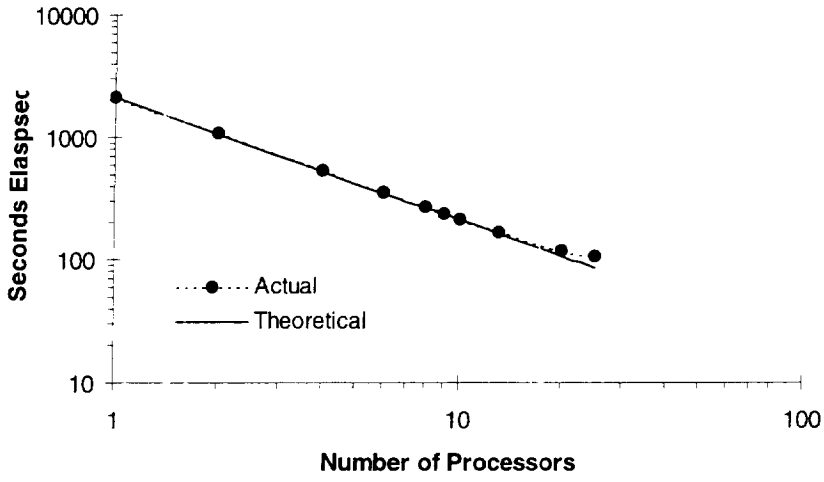


Figure 6: Parallel Speedup for an Ideal CFD Grid: Time for 200 Iterations

Figures 5 and 6 exhibit speedup that matches the theoretical speedup for most of the data points. In this context, theoretical speedup means that with twice the number of processors, the analysis time should be halved, or the number of iterations per second should be doubled. There is a limit to this trend that in this case, occurs between 15 and 20 processors as evident in the figures. Here, the speed at which data is communicated between the processors can not keep up with the speed at which the data to be communicated is produced. In this case, the communication overhead slows the overall computational speed. If this analysis were extended to many more processors, there would be a point when no more speedup could be achieved and the performance would actually begin to degrade.

The results of another performance study using a more typical grid size are shown below. This study was done using a 1.07 million point CFD grid. Figures 7 and 8 show the results of the study. Unlike the first example, the grid here was not ideally optimized for parallel execution. The zones were different sizes and could not be applied equally to multiple CPUs. The computational grid consisted of 26 zones of 30,000 grid points, 13 zones of 21,000 grid points, and one zone each of 10,000 and 7,000 grid points. This is typical of CFD grids in general. It is hard to equally divide a grid describing a complex geometry.

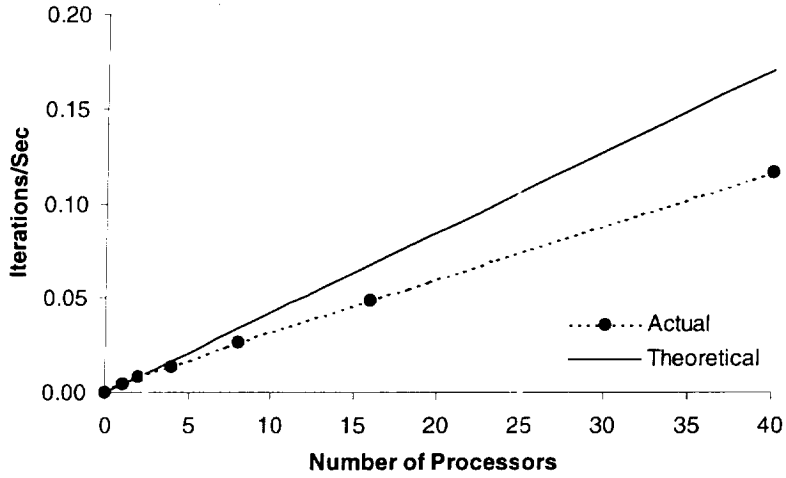


Figure 7: Parallel Speedup for a Non-Ideal CFD Grid: Iterations per Second

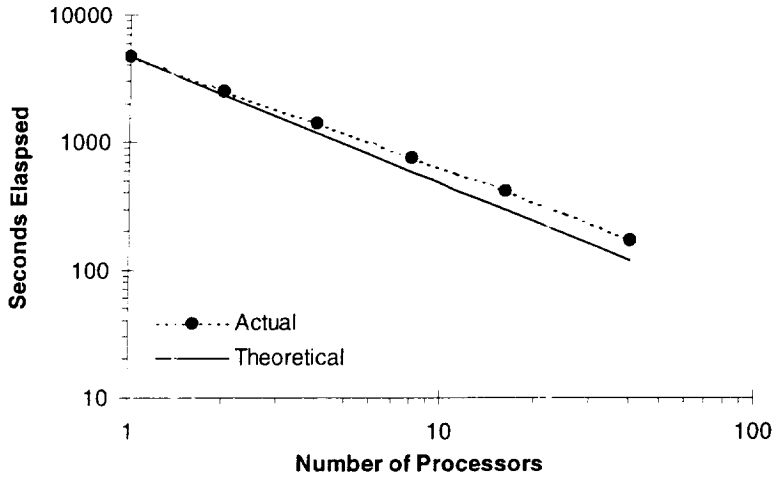


Figure 8: Parallel Speedup for a Non-Ideal CFD Grid: Time for 20 Iterations

The result of using non-ideal grid zoning is evident in the figures above. Here, the actual speed-up no longer matches the theoretical speed-up. This is because the processors working on smaller zones finish each of their iterations before those working on larger zones and must wait on the zone boundary information from the processors working on the larger zones before the start of each new iteration. The effect of communication overhead evident in Figures 5 and 6 as the number of processors is increased beyond 15 is not evident here because of the total number of grid points is an order of magnitude larger. However, if the number of processors continued to be increased, there would be a point at which performance would begin to degrade.

It should be noted that in these examples of CFD code parallel performance benchmarking, the analyses were run using the 100 Mbit/sec Ethernet. The 1000 Mbit/sec Myrinet is a recent upgrade that has not been fully tested. The use of the faster Myrinet, for these examples, may significantly increase the number of processors that could be used before loosing parallel efficiency.

The examples of CFD code parallel performance benchmarking presented here illustrate certain concepts that the analyst must keep in mind in order to use a Beowulf cluster efficiently. A balance should be maintained between the computation speed and communication speed. In other words, there must be a balance between the problem size and the number of processors used. The problem should also be equally divided between the processors. These intuitive concepts apply to the use of a Beowulf cluster in general, not just to CFD analyses.

Future Plans

Although the individual CFD analyses conducted using the cluster make efficient use of the cluster's resources, the cluster itself may sit idle for a great deal of time. Therefore, other areas of interest are being investigated to more fully utilize the cluster in support of Test Technology Development and NASA/SSC operations.

One of the many areas of interest that are being considered to better utilize the cluster is the addition of structural, thermal, fluid, and acoustic FEA codes. These are all areas of interest that must be supported at NASA/SSC. FEA codes are currently being used at NASA/SSC for structural and thermal problems. However, the codes being used do not run in parallel. The very nature of the Finite Element Method makes FEA codes very amenable to running in parallel via domain decomposition. There are already a few commercially available FEA codes that run on Beowulf clusters; the MARC and NASTRAN codes are two examples. However, their prices are prohibitively high at this time. Since this is a new area of consumer interest, more commercial and government FEA codes will be ported over to run on Beowulf clusters as the demand for such a capability increases.

One of the other areas that are being investigated for cluster utilization is real-time data acquisition and processing during or shortly after rocket engine or rocket engine component tests. Volumes of data are acquired during each rocket engine or rocket engine component test at NASA/SSC. NASA/SSC then provides the data to customers such as Marshall Space Flight Center, Boeing, or TRW. The data is usually processed at NASA/SSC based on the customer's requirements. For example, the customer may want the frequency domain information extracted from time domain data. Currently, this process is a manual one. The data is recorded from many channels in the field onto one media type. It is usually transferred to another media type and then brought into a data processing program. Various things may then be done to the raw data to get it into the form that the customer desires. Furthermore, the customer

may receive their data hours or even days after the test. A Beowulf cluster could be used to automate and greatly speed up this process. If the channels carrying the raw data were networked to a Beowulf cluster and properly engineered software was implemented, the customer's deliverable could be produced in real or near-real time. The Test Technology Development cluster has been proposed for use as a development platform for this capability. The development of this real-time data acquisition and processing capability could be accomplished using pre-recorded data so as not to interfere with ongoing testing. A dedicated Beowulf cluster could be built for real-time data acquisition and processing when the capability is production ready.

Summary and Conclusions

The design and construction of the NASA/SSC Test Technology Development Beowulf computing cluster has been discussed above. The cluster was: 1) designed as a high-performance computing platform for CFD simulations; 2) built for much less money than a traditional high-performance computing platform; and 3) built with flexible upgrade options in mind.

The addition of the Beowulf computing cluster to Test Technology Development has been very beneficial to NASA/SSC. It has been used to greatly expand their CFD analysis capabilities. There are many other potential uses for this technology. Other areas are being investigated in which the cluster can be utilized in support of NASA/SSC operations. The Test Technology Development cluster, with its many potential uses and flexible upgrade options, should remain useful as a high-performance computing platform at NASA/SSC for many years to come.

References

[1] D. H. Spector, *Building LINUX Clusters: Scaling LINUX for Scientific and Enterprise Applications*, O'Reilly and Associates (ISBN 1-56592-625-0), 2000.

[2] T. L. Sterling, J. Salmon, D. J. Becker, D. F. Savarese, *How to Build a Beowulf: A Guide to the Implementation and Application of PC Clusters*, MIT Press (ISBN 0-26269-218-X), 1999.

[3] R. Petersen, *Red Hat LINUX: The Complete Reference*, Osborne/McGraw-Hill (ISBN 0-07212-537-3), 2000.

[4] A. Geist, A. Beguelin, J. Dongarra, *PVM--Parallel Virtual Machine: A User's Guide and Tutorial for Network Parallel Computing*, MIT Press (ISBN 0-26257-108-0), 1994.

[5] "Recent Advances in Parallel Virtual Machine and Message Passing Interface", Proc. 7th European PVM-MPI Users' Group Meeting, Sept. 1997.

[6] W. Gropp, E. Lusk, A. Skjellum, *Using MPI: Portable Parallel Programming with the Message-Passing Interface*, MIT Press (ISBN 0-26257-132-3), 1999.

[7] "Second MPI Developer's Conference", IEEE Computer Society Press (ISBN 0-81867-533-0), July 1996.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-10-2000	NASA/TM	Oct 2000

4. TITLE AND SUBTITLE
Construction and Utilization of a Beowulf Computing Cluster:
A User's Perspective

5a. CONTRACT NUMBER
NAS13-650

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jody L. Woods; Jeff S. West
Lockheed Martin Space Operations

Peter R. Sulyma
NASA/John C. Stennis Space Center

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lockheed Martin Space Operations
Stennis Programs
John C. Stennis Space Center
Stennis Space Center, MS 39529-6000

8. PERFORMING ORGANIZATION REPORT NUMBER
SE-2000-10-00017-SSC

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
NASA
VA34
John C. Stennis Space Center
Stennis Space Center, MS 39529-6000

10. SPONSORING/MONITOR'S ACRONYM(S)
NASA

11. SPONSORING/MONITORING REPORT NUMBER
NASA/TM-2000-210691

12. DISTRIBUTION/AVAILABILITY STATEMENT
UNCLASSIFIED- UNLIMITED

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Lockheed Martin Space Operations - Stennis Programs (LMSO) at the John C. Stennis Space Center (NASA/SSC) has designed and built a Beowulf computer cluster which is owned by NASA/SSC and operated by LMSO. The design and construction of the cluster are detailed in this paper. The cluster is currently used for Computational Fluid Dynamics (CFD) simulations. The CFD codes in use and their applications are discussed. Examples of some of the work are also presented. Performance benchmark studies have been conducted for the CFD codes being run on the cluster. The results of two of the studies are presented and discussed. The cluster is not currently being utilized to its full potential; therefore, plans are underway to add more capabilities. These include the addtion of structural, thermal, fluid, and accoustic Finite Element Analysis codes as well as real-time data acquisition and processing during test operations at NASA/SSC. These plans are discussed as well.

15. SUBJECT TERMS
Beowulf computing cluster; high performance computing; parallel computing; LINUX

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19b. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Jody L. Woods
					19b. TELEPHONE NUMBER *(Include area code)*
UNCLAS	UNCLAS	UNCLAS	UNLIMITED	16	(228) 688-3583