Information Power Grid: Distributed High-Performance Computing and Large-Scale Data Management for Science and Engineering

William E. Johnston, Dennis Gannon, and Bill Nitzberg
Numerical Aerospace Simulation Division, NASA Ames Research Center, Moffett Field, CA
wej@nas.nasa.gov, gannon@cs.indiana.edu, nitzberg@nas.nasa.gov

Abstract

The term "Grid" refers to distributed, high performance computing and data handling infrastructure that incorporates geographically and organizationally dispersed, heterogeneous resources that are persistent and supported.

The vision for NASA's Information Power Grid - a computing and data Grid - is that it will provide significant new capabilities to scientists and engineers by facilitating *routine* construction of information based problem solving environments / frameworks that will knit together widely distributed computing, data, instrument, and human resources into just-in-time systems that can address complex and large-scale computing and data analysis problems.

IPG development and deployment is addressing requirements obtained by analyzing a number of different application areas, in particular from the NASA Aero-Space Technology Enterprise. This analysis has focussed primarily on two types of users: The scientist / design engineer whose primary interest is problem solving (e.g., determining wing aerodynamic characteristics in many different operating environments), and whose primary interface to IPG will be through various sorts of problem solving frameworks. The second type of user if the tool designer: The computational scientists who convert physics and mathematics into code that can simulate the physical world. These are the two primary users of IPG, and they have rather different requirements.

This paper describes the current state of IPG (the operational testbed), the set of capabilities being put into place for the operational prototype IPG, as well as some of the longer term R&D tasks.

1 Introduction

"Grids" (see [1]) are an approach for building dynamically constructed problem solving environments using distributed and federated, high performance computing and data handling infrastructure that incorporates geographically and organizationally dispersed resources.

The overall motivation for most current Grid projects is to enable the resource interactions that facilitate large-scale science and engineering such as aerospace systems design, high energy physics data analysis, climatology, large-scale remote instrument operation, etc.

The vision for a computing, data, and instrument Grids is that they will provide significant new capabilities to scientists and engineers by facilitating *routine* construction of information based problem solving environments. That is, Grids will routinely – and easily, from the user's point of view – facilitate applications such as:

- coupled, multidisciplinary simulations too large for single computing systems (e.g., multi-component turbomachine simulation – see [2])

- management of very large parameter space studies where thousands of low fidelity simulations explore, e.g., the aerodynamics of the next generation space shuttle in its many operating regimes (from Mach 27 entry into the atmosphere to landing)
- use of widely distributed, federated data archives (e.g., simultaneous access to metrological, topological, aircraft performance, and flight path scheduling databases supporting a National Air Transportation Simulation system)
- coupling large-scale computing and data systems to scientific and engineering instruments so that real-time data analysis results can be used by the experimentalist in ways that allow direct interaction with the experiment (e.g. operating jet engines in test cells and aerodynamic studies of airframes in wind tunnels)
- augmented reality and virtual reality remote collaboration (e.g., the Ames / Boeing Remote Help Desk that will provide aircraft field maintenance personnel use of coupled video and non-destructive imaging to supply real-time data to a remote, on-line, airframe structures expert who uses this data to index into detailed design databases, and returns then 3D internal aircraft geometry imagery to the field for damage assessment)
- single computational problems too large for any single system (e.g. extremely high resolution rotocraft aerodynamic calculations)

IPG has the goal of providing significant new capabilities to scientists and engineers by facilitating the solution of large-scale, complex, multi-institutional / multi-disciplinary, data and computational based problems using CPU, data storage, instrumentation, and human resources distributed across the NASA community. This entails technology goals of:

- independent, but consistent, tools and services that support various programming environments for building applications in widely distributed environments
- tools, services, and infrastructure for managing and aggregating dynamic, widely distributed collections of resources - CPUs, data storage / information systems, communications systems, real-time data sources and instruments, and human collaborators
- facilities for constructing collaborative, application oriented workbenches / problem solving environments across the NASA enterprise based on the IPG infrastructure and applications. These constitute the primary science and engineering interface to Grids
- a common resource management approach that addresses, e.g., system management, user identification, resource allocations, accounting, security, etc.
- an operational Grid environment incorporating major computing and data resources at multiple NASA sites in order to provide an infrastructure capable of routinely addressing larger scale, more diverse, and more transient problems than is possible today

2 An Overall Model for Grids

Analysis of some specific requirements ([3]), of the work processes of the user communities, and for remote instrument operation, as well as some anticipation of where the technology and problem solving needs are going in the future, leads to a characterization of the desired Grid functionality. This functionality may be represented as a hierarchically structured set of services and capabilities which are described below, and who's interrelationship is illustrated in Figure 1.

Problem Solving Environments, Supporting Toolkits, and High-Level Services

A number of services directly support building and using the Grid problem solving environment, e.g., by engineers or scientists. These include the toolkits for construction of application

frameworks / problem solving environments (PSE) that integrate Grid services and applications into the "desktop" environment. For example, the graphical components ("widgets" / applets) for application user interfaces and control; the computer mediated, distributed human collaboration that support interface sharing and management; the tools that access the resource discovery and brokering services; tools for generalized workflow management services such as resource scheduling, and managing high throughput jobs, etc.

An important interface for developers of Grid based applications is a "global shell," which, in general, will support creating and managing widely distributed, rule-based workflows driven from a published / subscribed global event service. Data cataloguing and data archive access, security and access control are also essential components. The PSE must also provide functionality for remote operation of laboratory / experiment / analytical instrument systems, remote visualization, and data-centric interfaces and tools that support multi-source data exploration.

Programming Services

Tools and techniques are needed for building applications that run in Grid environments, cover a wide spectrum of programming paradigms, and must operate in a multi-platform, heterogeneous computing environments. IPG e.g., will require Globus support for Grid MPI [4] as well as Java bindings to Globus services. CORBA [5], Condor [6], Java/RMI [7], Legion [8], and perhaps DCOM [9] are all application oriented middleware systems that will have to interoperate with the Grid in order to gain access to the resources managed by the Grid.

Compilation environment management, distributed debugging and performance analyses are difficult and important areas that must also be addressed to facilitate the construction of applications. Tools are needed for converting and "wrapping" legacy codes for operation in Grids, and for incorporating legacy Fortran codes into CORBA environments that are used to support composing application components. Grid-enabled numerical solution libraries that can be optimized for distributed architectures are also important.

Grid Common Services: Execution Management

Several services are critical to managing the execution of application codes in the Grid. The first is resource discovery and brokering. By discovery we mean the ability to ask questions like: how to find the set of objects (e.g. databases, CPUs, functional servers) with a given set of properties; how to select among many possible resources based on constraints such as allocation and scheduling; how to install a new object/service into the Grid; and how make new objects known as a Grid service. The second is execution queue management, which relates to global views of CPU queues and their user-level management tools. Workflow management and global shells is the third category. The fourth category is distributed application management. The last category includes tools for generalized fault management mechanisms for applications, and for monitoring and supplying information to knowledge based recovery systems.

Grid Common Services: Runtime

Globus [10] has been chosen as the initial IPG runtime system and supplies basic services to characterize and locate resources, initiate and monitor jobs, and provide secure authentication of users. However, there are other runtime services that are needed, include checkpoint/restart mechanisms, access control, a global file system, and Grid communication libraries such as a network-aware MPI that supports security, reliable multicast and remote I/O.

High-speed, wide area, distributed data management services include global naming and uniform access, uniform naming and location transparent access to resources such as data objects, computations, instruments and networks that work through Grid-wide object brokers. This, in turn requires uniform I/O mechanisms (e.g. read, write, seek) for all access protocols (e.g. http, ftp, nfs, Globus Access to Secondary Storage, etc.) and richer access and I/O mechanisms (e.g. "application level paging") that are present in existing systems.

Data cataloguing and publishing constitute another needed class of services. These include the ability to automatically generate the meta-data about data formats, and management of use conditions and access control. The ability to generate model based abstractions for data access using extended XML and XMI [11] data models is also likely to be important in the complex and data rich environment of, e.g., aero-space design systems.

Of course, high-speed, wide area, access to tertiary storage systems will always be critical, in the science and engineering applications that we are addressing. In IPG we are using SDSC's Meta Data Catalogue / Storage Resource Broker ("MCAT/SRB") [12] to provide widely distributed access to tertiary storage systems, independent of the nature of the underlying mass storage system implementation. High-performance applications require high-speed access to data files, and the system must be able to stage, cache, and automatically manage the location of local, remote and cached copies of files. We are also going to need the ability to dynamically manage large, distributed "user-level" caches and "windows" on off-line data. Support for object-oriented data management systems will also be needed.

Services supporting collaboration and remote instrument control are needed. In addition, application monitoring and application characterization, prediction, and analysis, will be important for both users and the managers of the Grid.

Finally, monitoring services will include precision time event tagging for dispersed, multi-component performance analysis as well as generalized auditing data file history and control flow tracking in distributed, multi-process simulations.

Grid Common Services: Environment Management

The key service that is used to manage the Grid environment is the "Grid Information Service." This service – currently provided by Globus GIS (formerly MDS, see [13]) – maintains detailed characteristics and state information about all resources, and will also need to maintain dynamic performance information, information about current process state, user identities, allocations and accounting information.

Autonomous system management and fault management services provide the other aspect of the environmental services.

Resource Management for Co-Scheduling and Reservation

One of the most challenging and well known Grid problems is that of scheduling scarce resources such as a large instruments. In many, if not most, cases the problem is really one of co-scheduling multiple resources. Any solution to this problem must have the agility to support transient experiments based on systems built on-demand for limited periods of time. CPU advance reservation scheduling and network bandwidth advance reservation scheduling based on differentiated IP services are critical components to the co-scheduling services. In addition, tape

marshaling in tertiary storage systems to support temporal reservations of tertiary storage system off line data and/or capacity is likely to be essential.

Operations and System Administration

Implementing a persistent, managed Grid requires tools for deploying and managing the system software. In addition, tools for diagnostic analysis and distributed performance monitoring are required, as are accounting and auditing tools. An often overlooked service involves the operational documentation and procedures that are essential to managing the Grid as a robust production service.

Access Control and Security

The first requirement for establishing a workable authentication and security model for the Grid is to provide a single-sign-on authentication for all Grid resources based on cryptographic credentials that are maintained in the users desktop / PSE environment(s) or on one's person. In addition, end-to-end encrypted communication channels are needed in for many applications in order to ensure data integrity and confidentiality. This is provided by X.509 identity certificates (see [14]) together with the Globus Security Services.

The second requirement is an authorization and access control model that provides for management of stakeholder rights (use-conditions) and trusted third parties to attest to corresponding user attributes. A policy-based access control mechanism that is based on use-conditions and user attributes is also a requirement. Several approaches are being investigated for providing these capabilities.

Security and infrastructure protection are, of course, essential requirements for the resource owners. This area includes technologies such as IPSec and secure DNS to authenticate IP packet origin, secure router and switch management, etc. (see, e.g., [15]), and the plans are to deploy these in an IPG security testbed.

Services for Operability

To operate the Grid as a reliable, production environment is a challenging problem. Some of the identified issues include management tools for the Grid Information Service that provides global information about the configuration and state of the Grid; diagnostic tools so operations/systems staff can investigate remote problems, and; tools and common interfaces for system and user administration, accounting, auditing and job tracking. Verification suites, benchmarks, regression analysis tools for performance, reliability, and system sensitivity testing are essential parts of standard maintenance.

Grid Architecture: How do all these services fit together?

We envision the Grid as a layered set of services (see Figure 1) that manage the underlying resources, and middleware that supports different styles of usage (e.g. different programming paradigms and access methods).

However, the implementation is that of a continuum of hierarchically related, independent and interdependent services, each of which performs a specific function, and may rely on other Grid services to accomplish its function.

Further, the "layered" model should not obscure the fact that these "layers" are not just APIs, but usually a collection of functions and management systems that work in concert to provide the

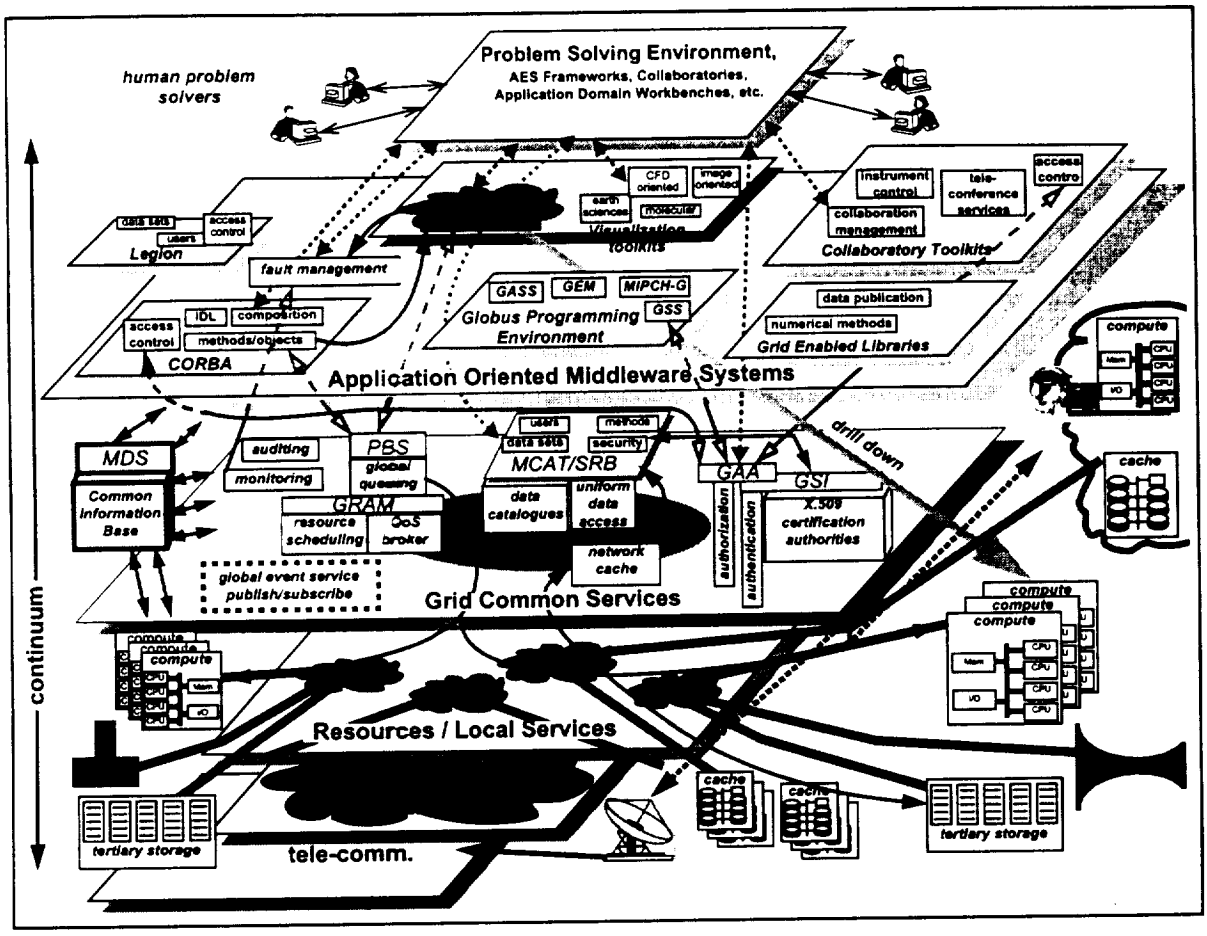

Figure 1 **A Representation of Grid Architecture**

"service" at a given "layer." The layering is not rigid, and "drill down" (e.g. code written for specific system architectures and capabilities) are easily managed by Grid services.

The arrows in the figure between several of the layers and services are intended to indicate how a real application involving a team working on a computational fluid dynamics ("CFD") based design problem might interact with Grid services, top to bottom.

3 How is IPG being accomplished?

Three main areas must be addressed in order to accomplish the goals of IPG:

1) new functionality and capability
2) an operational environment that encompasses significant resources
3) new services delivery model

The first area has already been discussed.

The second area, an operational system, is discussed below.

In the third area, Grids, such as IPG, effectively represents a new business model for operational organizations delivering large-scale computing and data resources. Grids require that these

services be delivered in ways that allow them to be integrated with other widely distributed resources controlled, e.g., by the user community. This is a big change for, e.g., traditional supercomputer centers. Implementing this service delivery model requires two things: First, tools for production support, management, and maintenance, of integrated collections of widely distributed, multi-stakeholder resources must be identified, built, and provided to the systems and operations staffs. Second, organizational structures must be evolved that account for the fact that operating Grids is different than operating traditional supercomputer centers, and management and operation of this new shared responsibility service delivery environment must be explicitly addressed.

4 What is the State of IPG?

Point 1), above, is being addressed by a detailed examination of requirements generated by several NASA application communities, both in terms of specific capabilities identified by the applications community, and as the result of analysis of the requirements and desired operating environments by computer scientists. These requirements are documented in the IPG implementation plan (see [3]).

Addressing point 2), the two year (11/2000) IPG goal is an operational and persistent, "large-scale" prototype-production Information Power Grid providing access to computing, data, and instrument resources at NASA Centers around the country so that large-scale are more easily accomplished than is possible today.

The first major milestone toward this goal (11/1999, see [3]) is a baseline Grid system (Figure 2) that includes:

- approximately 600 CPU nodes in half a dozen SGI Origin 2000s at four NASA sites
- several workstation clusters
- 30-100 Terabytes of uniformly accessible mass storage
- wide area network interconnects of at least 100 mbit/s
- a stable and supported operational environment

Addressing point 3), the NAS Division at NASA Ames is identifying the new services that will be delivered by IPG, and is creating groups that will develop (as necessary), test, deploy, and support these services. In addition to new local organizational structure and local R&D, NAS is coordinating related activities at the NSF supercomputer centers [16], and at several universities, to provide various components of the new operational model.

Current progress is reflected in the IPG Engineering Working Group tasks (see [3]): 30+ tasks have been identified as critical for the baseline system, and groups have been organized around the major task areas:

- identification and testing of computing and storage resources for inclusion in IPG
- deployment of Globus ([10]) as the initial IPG runtime system
- global management of CPU queues, and job tracking, and monitoring throughout IPG
- definition and implementation of a reliable, distributed Grid Information Service that characterizes all of the IPG resources
- public-key security infrastructure integration and deployment to support single sign-on using X.509 cryptographic identity certificates (see [14])
- network infrastructure and QoS

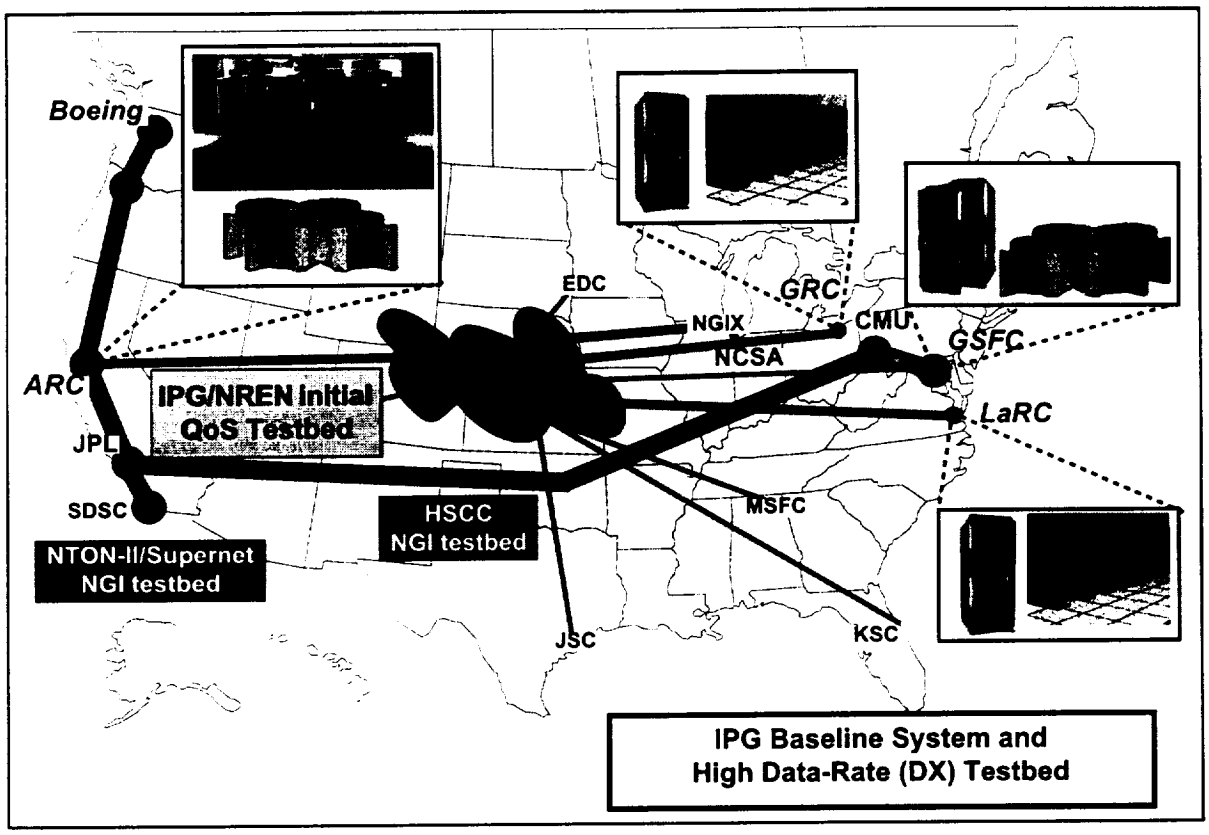

Figure 2 **First Phase of NASA's Information Power Grid**

- tertiary storage system metadata catalogue and uniform access system (based on MCAT/SRB - [12])
- operational and system administration procedures for the distributed IPG
- user and operations documentation
- account and resource allocation management across systems with multiple stakeholders
- Grid MPI [4], CORBA [5], and Legion [8] programming middleware systems integration
- high throughput job management tools
- distributed debugging and performance monitoring tools

5 What Comes Next?

There are many challenges in making Grids a reality, in the sense that they can provide new capabilities in production quality environments.

While the basic Grid services have been demonstrated, e.g. in the IPG prototype demonstrated at the Supercomputing 1999 conference, and the GUSTO testbed ([17]) demonstrated in 1998, a general purpose computing, data management, and real-time instrumentation Grid involves many more services. One challenge is to identify the minimal set of such services. In many cases, these services exist in some form in R&D environments, as described in this paper, however, then the challenge is to convert these into robust implementations that can be integrated with the other services of the Grid. This is hard, and is one of the big challenges for NASA's IPG.

The architecture described above is being implemented in several Grid projects, including NASA's IPG and the DOE ASCI distributed computing program ([18]). Projects such as these will refine the Grid concepts and implementation so that Grids will meet the challenge of how to accelerate routine use of applications that:

- require substantial computing resources
- generate and/or consume high rate and high volume data flows
- involve human interaction
- require aggregating many dispersed resources to establish an operating environment:
  - multiple data archives
  - distributed computing capacity
  - distributed cache capacity
  - "guaranteed" network capacity
- operate in widely dispersed environments.

In addition to the NASA and DOE projects, Grids are being developed by a substantial and increasing community of people who work together in a loosely bound coordinating organization called the Grid Forum (www.gridforum.org - [19]). From efforts such as this, Grids will become a reality, and an important component of the practice of science and engineering.

6 Acknowledgements

Almost everyone in the NAS division of the NASA Ames Research Center, numerous other people at the NASA Ames, Glenn, and Langley Research Centers, as well as many people involved with the NSF PACIs (especially Ian Foster, Argonne National Lab. and Carl Kesselman, USC/ISI) have contributed to this work. The NASA Research and Education Network (NREN) has played a critical role in building the initial IPG. We would also like to specifically thank Bill Feiereisen, NAS Division Chief, and while the NASA HPCC Program Manager the initiator of IPG, Alex Woo, NAS Research Branch Chief, Bill Thigpen, NAS Engineering Branch Chief. IPG is funded primarily by NASA's Aero-Space Enterprise, Information Technology (IT) program (http://www.nas.nasa.gov/IT/overview.html).

7 References

[1] Foster, I., and C. Kesselman, eds., *The Grid: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman. Morgan Kaufmann, Pub. August 1998. ISBN 1-55860-475-8. http://www.mkp.com/books_catalog/1-55860-475-8.asp

[2] NPSS - see http://hpcc.lerc.nasa.gov/grndchal.shtml

[3] "Information Power Grid." See www.nas.nasa.gov/~wej/IPG for project information, pointers, and the IPG implementation plan.

[4] Foster, I., N. Karonis, "A Grid-Enabled MPI: Message Passing in Heterogeneous Distributed Computing Systems." Proc. 1998 SC Conference. Available at http://www-fp.globus.org/documentation/papers.html

[5] Otte, R., P. Patrick, M. Roy, *Understanding CORBA*, Englewood Cliffs, NJ, Prentice Hall, 1996.

[6] Livny, M, et al, "Condor." See http://www.cs.wisc.edu/condor/

[7] Sun Microsystems, "Remote Method Invocation (RMI)." See http://developer.java.sun.com/developer/technicalArticles//RMI/index.html .

[8] Grimshaw, A. S., W. A. Wulf, and the Legion team, "The Legion vision of a worldwide virtual computer", *Communications of the ACM*, 40(1):39-45, 1997.

[9] Microsoft Corp., "DCOM Technical Overview." November 1996. See http://msdn.microsoft.com/library/backgrnd/html/msdn_dcomtec.htm .

[10] Foster, I., C. Kesselman. Globus: A metacomputing infrastructure toolkit", *Int'l J. Supercomputing Applications*, 11(2);115-128, 1997. (Also see http://www.globus.org)

[11] "XML Metadata Interchange" (XMI). See "XML News and Resources" http://metalab.unc.edu/xml/

[12] Moore, R., et al. "Massive Data Analysis Systems," San Diego Supercomputer Center. See http://www.sdsc.edu/MDAS

[13] Fitzgerald, S., I. Foster, C. Kesselman, G. von Laszewski, W. Smith, S. Tuecke, "A Directory Service for Configuring High-Performance Distributed Computations." Proc. 6th IEEE Symp. on High-Performance Distributed Computing, pg. 365-375, 1997. Available from http://www.globus.org/documentation/papers.html .

[14] Public-Key certificate infrastructure ("PKI") provides the tools to create and manage digitally signed certificates. For identity authentication, a certification authority generates a certificate (most commonly an X.509 certificate) containing the name (usually X.500 distinguished name) of an entity (e.g. user) and that entity's public key. The CA then signs this "certificate" and publishes it (usually in an LDAP directory service). These are the basic components of PKI, and allow the entity to prove its identity, independent of location or system. For more information, see, e.g.,RSA Lab's "Frequently Asked Questions About Today's Cryptography" http://www.rsa.com/rsalabs/faq/, *Computer Communications Security: Principles, Standards, Protocols, and Techniques*. W. Ford, Prentice-Hall, Englewood Cliffs, New Jersey, 07632, 1995, or *Applied Cryptography*, B. Schneier, John Wiley & Sons, 1996.

[15] "Bridging the Gap from Networking Technologies to Applications." Workshop Co-sponsored by HPNAT & NRT (High Performance Network Applications Team & Networking Research Team of the Large Scale Networking (Next Generation Internet) Working Group). NASA Ames Research Center, Moffett Field, Mountain View CA. Moffett Training and Conference Center, August 10 - 11, 1999. To be published at http://www.nren.nasa.gov/workshop_home.html ("HPNAT/NRT Workshop")

[16] The NSF PACIs are the Alliance/NCSA (http://www.ncsa.uiuc.edu/) and NPACI/SDSC (http://www.npaci.edu/).

[17] "Globus Ubiquitous Supercomputing Testbed Organization" (GUSTO). At Supercomputing 1998, GUSTO linked around 40 sites, and provides over 2.5 TFLOPS of compute power, thereby representing one of the largest computational environments ever constructed at that time. See http://www.globus.org/testbeds .

[18] "Distance Computing and Distributed Computing (DisCom2) Program." See http://www.cs.sandia.gov/discom .

[19] Grid Forum. The Grid Forum (www.gridforum.org) is an informal consortium of institutions and individuals working on wide area computing and computational grids: the technologies that underlie such activities as the NCSA Alliance's National Technology Grid, NPACI's Metasystems efforts, NASA's Information Power Grid, DOE ASCI's DISCOM program, and other activities worldwide.