

The Grand Challenge of Managing the Petascale Facility

Mathematics and Computer Science Division



About Argonne National Laboratory

Argonne is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC under contract DE-AC02-06CH11357. The Laboratory's main facility is outside Chicago, at 9700 South Cass Avenue, Argonne, Illinois 60439. For information about Argonne, see www.anl.gov.

Availability of This Report

This report is available, at no cost, at <http://www.osti.gov/bridge>. It is also available on paper to the U.S. Department of Energy and its contractors, for a processing fee, from:

U.S. Department of Energy

Office of Scientific and Technical Information

P.O. Box 62

Oak Ridge, TN 37831-0062

phone (865) 576-8401

fax (865) 576-5728

reports@adonis.osti.gov

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor UChicago Argonne, LLC, nor any of their employees or officers, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of document authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, Argonne National Laboratory, or UChicago Argonne, LLC.

The Grand Challenge of Managing the Petascale Facility

by
R.J. Aiken
aiken@mcs.anl.gov
Mathematics and Computer Science Division, Argonne National Laboratory

December 2006



UChicago ►
Argonne_{LLC}



Contents

| | |
|---|-------------|
| Preface | vii |
| Acknowledgments | viii |
| I. Executive Summary | 1 |
| II. Introduction | 6 |
| II.1 Scope of the Petascale Facility | 6 |
| II.2 Performance, Predictability, and Usability | 7 |
| II.3 Complexity | 7 |
| II.4 Productivity and Workflow | 8 |
| II.5 Petascale Management | 8 |
| II.5.1 Grand Challenge of Management | 9 |
| II.5.2 Local and Global Management | 9 |
| II.5.3 Management Areas | 9 |
| II.5.4 New Management Models | 10 |
| II.5.5 Instrumentation and Monitoring | 10 |
| II.5.6 Appropriate Use | 11 |
| II.5.7 Security | 12 |
| II.5.8 Morphable Infrastructure | 12 |
| II.5.9 Naming | 13 |
| II.5.10 Object-based Infrastructure | 13 |
| II.6 Summary | 13 |
| II.7 Structure of the Report | 14 |
| III. Challenges and Research Opportunities | 16 |
| III.1 Scaling of the Petascale Facility and Metafacility | 16 |
| III.2 Virtualization, Virtual Organizations, Virtual Facilities, and Grids | 17 |
| III.3 Architectures, Systems, Facilities, and Metafacilities | 19 |
| III.4 Networks | 22 |
| III.5 Data Management | 26 |
| III.6 Workflow | 28 |
| III.7 Security | 31 |
| III.8 Visualization and Analytics | 31 |
| III.9 Morphable Networks and System | 31 |

| | | |
|----------------------|---|-----------|
| IV. | Recommendations | 33 |
| IV.1 | Integrated Management | 33 |
| IV.1.1 | New Program for Petascale Facility Management | 33 |
| IV.1.2 | Performance-driven, Predictable, and Usable Resource Management System | 34 |
| IV.1.3 | Data Management | 34 |
| IV.1.4 | Workflow | 34 |
| IV.1.5 | Object-based Infrastructure | 35 |
| IV.1.6 | Naming and Addressing | 35 |
| IV.1.7 | Dynamic and Adaptable Infrastructure | 35 |
| IV.2 | Initial Steps | 36 |
| IV.3 | SciDAC Petascale Projects as Enabling Prototypes | 36 |
| Appendix AI. | Petascale Computing and Facilities | 38 |
| AI.1 | Leadership Capability Systems | 38 |
| AI.2 | Petascale-class Experimental Facilities | 40 |
| AI.3 | Capacity Computing | 41 |
| AI.4 | Petascale Storage and Data Management | 41 |
| AI.5 | Petascale Networks | 42 |
| AI.6 | Petascale Analysis and Visualization | 43 |
| AI.7 | Petascale Facility and Metafacility | 43 |
| Appendix AII. | Trends in Scientific Research | 44 |
| AII.1 | One-of-a-Kind Facilities | 44 |
| AII.2 | Computational Science and Simulation | 45 |
| AII.3 | Data Fusion and Integration | 46 |
| AII.4 | Integration of Scientific Databases and Data Management | 47 |
| AII.4.1 | Data Location | 47 |
| AII.4.2 | Data Searching and Analysis | 48 |
| AII.4.3 | Data Provenance | 48 |
| AII.4.4 | Policy | 49 |
| AII.4.5 | Object-based File Systems | 49 |
| AII.5 | Interdisciplinary Teams | 49 |
| AII.6 | Virtualization | 50 |
| AII.7 | Visualization | 52 |
| AII.8 | Workflow | 53 |
| AII.9 | Dataflow and Data Management | 56 |
| AII.10 | Persistence, Ubiquitous Computing, Nomadicity, and Remote Access | 59 |
| AII.11 | Summary of Trends in Scientific Research | 60 |

| | | |
|-----------------------|---|-----------|
| Appendix AIII. | DOE Facilities: Current and Future | 61 |
| AIII.1 | SciDAC | 61 |
| AIII.2 | NITRD and HEC Roadmap | 61 |
| AIII.3 | Current and Planned Leadership-class Capability, Capacity, and Cluster Systems | 62 |
| | AIII.3.1 Capability Systems | 62 |
| | AIII.3.2 Capacity Systems | 63 |
| AIII.4 | Storage and Data Management | 64 |
| AIII.5 | Petascale Experiments | 66 |
| AIII.6 | Networks | 69 |
| AIII.7 | Infrastructure | 69 |
| AIII.8 | Grids | 70 |
| | | |
| Appendix AIV. | Trends in Technology | 71 |
| AIV.1 | Technology Churn | 71 |
| AIV.2 | Computing and Moore’s Law | 72 |
| AIV.3 | Storage | 74 |
| | AIV.3.1 Challenge of Data Storage | 74 |
| | AIV.3.2 Tape Storage | 74 |
| | AIV.3.3 Disk Storage | 75 |
| | AIV.3.4 Parallel I/O, File Systems, and Data Formats | 76 |
| | AIV.3.5 Summary | 77 |
| AIV.4 | Interconnect Area Networks | 77 |
| AIV.5 | Networks | 79 |
| | AIV.5.1 Dark Fiber and Waves | 80 |
| | AIV.5.2 IAN, CAN, SAN, LAN, MAN, and WAN | 80 |
| | AIV.5.3 10 Gbs Building Blocks | 81 |
| | AIV.5.4 IP and Routing | 81 |
| | AIV.5.5 Optical Networking | 82 |
| | AIV.5.6 Transport Protocols | 84 |
| | AIV.5.7 Multimode Networks | 85 |
| | AIV.5.8 Hybrid Networks | 85 |
| AIV.6 | Security | 85 |
| AIV.7 | Architectures and Systems | 87 |
| AIV.8 | Visualization | 87 |
| AIV.9 | Naming and Addressing | 88 |
| AIV.10 | Virtualization | 88 |
| AIV.11 | Programming Environments | 89 |
| | | |
| References | | 90 |

Preface

This report is the result of a study of networks and how they may need to evolve to support petascale leadership computing and science. As Dr. Ray Orbach, director of the Department of Energy's Office of Science, says in the spring 2006 issue of *SciDAC Review*, "One remarkable example of growth in unexpected directions has been in high-end computation." In the same article Dr. Michael Strayer states, "Moore's law suggests that before the end of the next cycle of SciDAC, we shall see petaflop computers." Given the Office of Science's strong leadership and support for petascale computing and facilities, we should expect to see petaflop computers in operation in support of science before the end of the decade, and DOE/SC Advanced Scientific Computing Research programs are focused on making this a reality.

This study took its lead from this strong focus on petascale computing and the networks required to support such facilities, but it grew to include almost all aspects of the DOE/SC petascale computational and experimental science facilities, all of which will face daunting challenges in managing and analyzing the voluminous amounts of data expected. In addition, trends indicate the increased coupling of unique experimental facilities with computational facilities, along with the integration of multidisciplinary datasets and high-end computing with data-intensive computing; and we can expect these trends to continue at the petascale level and beyond. Coupled with recent technology trends, they clearly indicate the need for including capability petascale storage, networks, and experiments, as well as collaboration tools and programming environments, as integral components of the Office of Science's petascale capability metafacility.

The objective of this report is to recommend a new cross-cutting program to support the management of petascale science and infrastructure. The appendices of the report document current and projected DOE computation facilities, science trends, and technology trends, whose combined impact can affect the manageability and stewardship of DOE's petascale facilities.

This report is not meant to be all-inclusive. Rather, the facilities, science projects, and research topics presented are to be considered examples to clarify a point.

Acknowledgments

I thank the following colleagues for useful discussions relevant to this report and to the management of a petascale facility.

Argonne National Laboratory:

Ray Bair, Pete Beckman, Charlie Catlett, Remy Evard, Ian Foster, Ewing Lusk, Rob Ross, and Rick Stevens

CAIDA/ University of California San Diego UCSD:

Kim Claffy

Cal IT²/University of California San Diego (UCSD):

Tom De Fanti and Larry Smarr.

California Institute of Technology (Caltech):

Julian Bunn, Harvey Newman, Michael Thomas, Conrad Steenberg, Xun Su, and Yang Xia

Carnegie Mellon University(CMU):

Dave Farber

CERN:

Iosif LeGrand, Frank van Lingen, Dan Nae, and Sylvain Ravot

Cisco:

Javad Boroumand

Corporation for Network Research Initiatives (CNRI):

Bob Kahn

Department of Energy, Office of Science, ASCR:

Dan Hitchcock, Fred Johnson, Mary Anne Scott, and Michael Strayer

General Atomics:

David Schissel

Fermi National Laboratory (Fermilab):

Phil Demar, Don Petravick, Ruth Pordes, and Vicky White

Lawrence Berkeley National Laboratory (LBNL) ESnet:

Joe Burrencia and Bill Johnston

Lawrence Berkeley National Laboratory (LBNL) NERSC:

Francesca Verdier, Bill Kramer, Sandy Merola, Horst Simon, and Howard Walter

Lawrence Berkeley National Laboratory (LBNL):

Deb Agarwal, Ari Shoshani, and Brian Tierney

Lawrence Livermore National Laboratory (LLNL):

Dave Bader, Mike McCoy, and Dave Seager

Level 3 Communications, Inc.

Jack Waters

National Center for Supercomputing Applications (NCSA) / University of Illinois at Urbana-Champaign (UIUC):

Jim Myers

Oak Ridge National Laboratory (ORNL):

Al Geist, Nagy Rau, and Bill Wing

Pacific Northwest National Laboratory (PNNL):

Elsa Augustenborg, Andrew Cowell, Stephen Elbert, Deb Gracio, Earl Heister, Jerry Johnson, Moe Khaleel, George Michaels, Jeff Mauth, John McCoy, Jarek Nieplocha, Fabrizio Petrini, Rich Quadrel, Kevin Regimbal, Antonio Sanfilippo, T.P. Straatsma, and Troy Thompson

Qwest:

Wes Kaplow

Sandia National Laboratories - Livermore (SNLL) :

Helen Chen and Larry Rahn

Thomas Jefferson National Accelerator:

Roy Whitney

Stanford/SLAC:

Richard Mount

University California - Santa Barbara (UCSB):

Dan Blumenthal

University of New Mexico (UNM):

Barney Maccabe

University of Southern California (USC):

Alan Willner

University of Tennessee:

Micah Beck

University of Wisconsin:

Miron Livny

In addition to the personal meetings relevant to this report, I relied heavily on the work and findings of the Data Management Challenge Report of 2005, the NERSC Plan for 2006–2010, and the ESnet Science Requirements Report for current and projected application requirements and some of the high-end computing hardware technology trends. I also referenced conclusions, roadmaps, and goals in the 2006 SciDAC Report, along with the Interagency 2006 LSN, 2005 NITRD, and 2004 HEC Plan reports. Much of the Grid, distributed workflow, and resource management discussions are based on services, trends, and standards of Globus, Condor, TeraGrid, the Open Science Grid, and MonALISA.

I thank the following reviewers for their constructive comments: Deb Agarwal, Dave Bader, Ray Bair, Buddy Bland, Joe Burescia, Charlie Catlett, Steve Elbert, Ian Foster, Al Geist, Bill Johnston, Doug Kothe, Ewing Lusk, George Michaels, Richard Mount, Harvey Newman, Larry Rahn, David Schissel, Horst Simon, Rick Stevens, and Bill Wing.

I note that none of these reviewers has been asked to endorse the conclusions or recommendations of this report. The final content and message are solely my responsibility.

This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy, under Contract DE-AC02-06CH11357.

I. Executive Summary

A recent report published in *Nature* by a number of well-known computer scientists, computational sciences, and biologists discusses projections for the state of science by 2020. The report indicates major changes in the way science will be conducted, including intelligent interaction and information discovery, semantics of data, transformation of scientific communication, computational thinking, new kinds of virtual communities, synergy between biology and computer science, and the integration of sciences. The authors also note the development of new conceptual and technology tools such as prediction machines, “artificial scientists,” molecular machines, and new software models [1]. Many of these trends are already evident in multidomain and combined computational and experimental sciences and have their grass roots in the techniques, tools, and architectures currently used or identified in roadmaps by the Department of Energy Office of Advanced Scientific Computing Research (ASCR) for the INCITE and SciDAC programs, as well as other Office of Science flagship science programs. Petascale science, computers, experiments, data management, and infrastructure are key components for enabling the next generation of science.

Pressures on Infrastructure and Science

Cutting-edge science often takes its pioneers into new and challenging areas. Within a few years scientists supported by the Department of Energy’s Office of Science (DOE/SC) will be using petaflops capability computers. Petabyte-scale experimental science and data-intensive science are becoming the norm, and their integration with petascale computing not only will increase over time but will be required as the trend continues toward more multidisciplinary, multisite, multi-institutional, multi-researcher science.

The Spallation Neutron Source (SNS), for example, is planning to tightly couple capability computing and theoretical simulation with petascale experiments, both for model validation and for real-time to quasi-real-time steering of the experiments, thereby enabling more effective use of costly facilities and reducing the time needed to analyze the results. Similarly, Advanced Photon Source experiments are considering the coupling of experimental and computational facilities. Observational and experimental science projects such as KamLAND [2], the Supernova Factory [3], and the Large Hadron Collider [4] are further examples of the need for the coupled use of high-end computing.

We are entering a new era of science that will require *metafacility* interfaces and standards to better enable interfacility collaboration and the movement of the data and codes from one site to another. This top-down science trend is occurring simultaneously with a bottom-up technology trend toward more parallelism aimed at moving beyond the limits of Moore’s law (e.g., parallel file systems, application-specific integrated circuits, multicore processors). Another technology trend is the virtualization of systems, networks, services, and organizations. The combination of these bottom-up and top-down

trends will only increase the challenges that scientists will encounter as they move into the petascale level and beyond.

For effective petascale science, more than just an evolutionary approach will be needed. An innovative approach must be formulated that integrates local and global facilities whereby theory and computation are melded and tightly coupled with experimentation, observation, and data-intensive sciences.

Program Integration and a New Scientific Discipline

The DOE/SC ASCR high-end computing (HEC) programs are focused on the development and use of leadership capability computing. However, in order to better move DOE/SC science into the petascale era and beyond, a complementary program is needed to integrate observation, experimental, and data-intensive science (OEDIS) facilities as part of an overall DOE/SC petascale metafacility.

Some steps to better integrate OEDIS and HEC through SciDAC and other science-driven programs have already been initiated; however, they are mainly separate focuses (see Fig. I.1). The tight coupling of an OEDIS program to the HEC program in a complementary fashion is essential for the success of each. Together they will enable the next generation of science (see Fig. I.2).

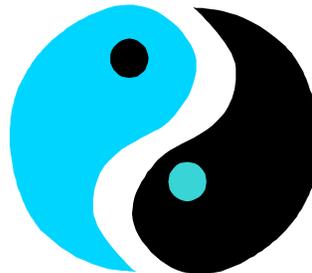


Figure I.1 Separate OEDIS and HEC programs



Figure I.2 Complementary and Integrated OEDIS and HEC programs

Both the HEC and OEDIS programs need to include a new focus on developing a *computing systems management science* (CSMS), similar to the development of computational science in the 1990s. This new science would combine systems operations, data management, distributed systems, networks, workflow systems, fault tolerance and autonomic computing, services, security, and computing sciences. The goal of the new

science would be to develop systems, standards, and tools to support the DOE petascale metafacility. Research and development in many of these areas has already been initiated in programs such as the Open Science Grid, MonALISA, Globus, Condor, and TeraGrid and in other workflow and distributed systems. DOE's proven leadership in integrating leading-edge science experiments, computing, and computational science (e.g., in SciDAC and INCITE) makes DOE/SC/ASCR an obvious choice to take the lead in developing this computing systems management science program. For example, DOE/SC/ASCR could prototype CSMS postdoctoral and graduate school opportunities at their capability petascale facilities, helping to build a talent pool for future high-energy computing.

Petascale Metafacility

The DOE/SC petascale metafacility is a combination of the leadership-class capability computers at Oak Ridge National Laboratory and Argonne National Laboratory, the capacity computing facilities at Lawrence Berkeley Laboratory/NERSC, petabyte-scale experiments such as the Large Hadron Collider and the Spallation Neutron Source, petabyte-scale to exabyte-scale storage systems, terabit networks capable of supporting petascale facilities, and petabyte-scale data management and analysis facilities. Operating systems, programming and scripting languages, programming development environments, middleware, Grids, workflow systems, system management architectures, and security are all crucial components of the petascale metafacility.

It is imperative that DOE treat all of these facilities and components as a metasystem with significantly more focus on their secure integration from a programmatic perspective, including the concurrent support of production and experimental/research infrastructure capabilities. Only in this way will researchers be able to migrate among the DOE/SC petascale facilities as their science requires.

Challenge of Scale and Complexity

Petascale computing facilities and systems will be made up of hundreds of thousands to millions of processor nodes. As the complexity of components and systems increases, so too do the probability of error and the difficulty in using the resources effectively. Moreover, the increased complexity introduced by combining a number of facilities into a metafacility increases the chances of security holes and opportunities for attacks. Combining one or more of these complex high-end computing systems with any other petascale storage, networks, and experiments—all of which are going through their own explosive virtualization and parallel growth—makes the challenge of securely managing these resources even more daunting. In addition, current systems and system software support static and coarse-grained allocation; to fully utilize these unique resources will require more dynamic and adaptive configuration and management.

Computing Systems Management

The term “management” in this report refers to the interaction and integration of *processes* (such as end-to-end security, monitoring, auditing, data analysis, visualization, workflow, data movement, and dynamic resource allocation), *techniques* (such as naming, object management techniques, schemas, autonomic computing, and the testing, benchmarking, verification, and integration of software, services, systems, and hardware), and *components* (such as instrumentation). The goal is not to develop a single management system; rather, it is to develop the ability to dynamically build virtual petascale facility and metafacility architectures and managements systems from a common set of standards-based management systems, subsystems, and core components to better support the next generation of science.

Recommendations

ASCR should create a program that integrates *observation, experimental, and data-intensive science* (OEDIS) with *high-end computing* (HEC) facilities. The combined program should foster the development of a new *computing systems management science* (CSMS) that integrates the architectures, systems, protocols, instrumentation, and tools necessary for the secure management of DOE’s petascale facilities and metafacility.

This program should identify and develop the key architectural elements, services, systems, and management capabilities needed to enable the SciDAC, INCITE, and other flagship applications of the future. The program should leverage and codify advances made in the SciDAC program and combine them into an overall architectural vision and roadmap suitable for supporting science of the future.

To this end, ASCR should select a program lead and enlist a team of experts to evaluate existing architectures, tools, and technologies and to identify research and development efforts needed to address shortcomings in the current state of the art. ASCR should convene a workshop to further refine the proposed OEDIS/HEC program and to determine an implementation plan. The program should then fuse and evolve the best components, coupled with relevant computer and computational research, into a secure management capability that melds both local and global architectures and perspectives and that provides for the appropriate combinations of subsets of management systems relevant to specific science disciplines and resources.

An aggressive roadmap would be as follows:

- Definition of a program within twelve months,
- A potential curriculum and graduate-level laboratory-based opportunities within three years,
- An overarching management and data management framework and system fully implemented by 2011, and
- An advanced, combined OEDIS/HEC petascale infrastructure by 2016. The infrastructure would allow for the secure and manageable plug-and-play

combination of individual standards-based management systems and components at both the local and global level into one or more virtual integrated management systems, to support the application researcher and the supporting systems scientists.

II. Introduction

DOE/SC/ASCR facilities are currently operating in the terascale range. Leadership-class computing centers at Argonne and Oak Ridge National Laboratories are scheduled to break into the petascale computing level within the next few years and into the exascale and yottascale level with respect to data in 10 to 20 years. Moreover, since researchers typically develop, test, and debug a code first on a smaller system before porting the code to a leadership-class computer, DOE/SC will need to use the full continuum of computational and experimental facilities as an integrated metafacility in order to successfully support science at the petascale level and beyond. Much as a race car's very powerful engine needs to be complemented by racing-caliber transmissions, suspension systems, and tires, the leadership-class systems need to be complemented by capability-level networks, storage, data management, I/O, and programming environments. Ignoring any one component may well make that component a gating factor that could impede the effective utilization of individual petascale facilities or the metafacility and the data generated.

II.1 Scope of the Petascale Facility

The 2004 Federal Plan for High-End Computing report notes that a major goal is to make high-end computing easier and more productive to use:

Emphasis should be placed on time to solution, the metric of value for the high-end computing users. Time to solution includes: time to cast the physical problem into algorithms suitable for high-end computing; time to write and debug the computer code that expresses these algorithms; time to optimize the code to the computer platforms being used; time to compute desired results; time to analyze those results; and time to refine the analysis into improved understanding of the original problem that enables scientific or engineering advances.

The HEC Task Force [5] limited its scope to the top high-end computing systems and noted that “networking, grid computing, visualization, general security issues, and applications-specific software were considered outside the scope of this planning effort.” DOE/SC, however, does not have the luxury of treating these as individual and disjoint programs. As DOE/SC plans for the deployment and management of its petascale and leadership-class facilities, it will need to address the fact that data, specialized software, and one-of-a-kind computing and experimental facilities will need to be integrated to support future petascale science endeavors.

Such integration will mean that what is included in the DOE/SC metafacility is far more than just a petaflop machine. This fact does not diminish the importance placed on leadership-class capability computer systems. Rather, it acknowledges the reality that such systems generate large amounts of data and that the researchers collaboratively use these and other facilities from many different time zones and geographic locations. In addition, the state-of-the-art capability systems of today are expected to become the

capacity systems of tomorrow; and therefore a full continuum of computer, storage, and network facilities will continue to exist as new technologies and architectures churn their way through the continuum.

II.2 Performance, Predictability, and Usability

The DOE research community has excelled at taking leading-edge, and sometimes bleeding-edge, technologies and integrating them into productive science systems. The community also has excelled at developing powerful and effective software and hardware when such was not commercially available. The goal has been to ensure access to systems that concurrently provide performance, predictability, and usability (PPU). This same goal will need to be addressed in the petascale era, especially as part of the “time to solution” issue identified by the HEC committee. The scaling and management issues associated with moving to the petascale will be challenging.

II.3 Complexity

Two trends in science and technology are evident. The first trend is the extensive and increasing use of parallelism to address the challenges of scaling as we approach the limits of Moore’s law for processing and other computing-related functions. Increments in capability and speed are not keeping pace processor gains. The second trend is the strong movement toward virtualization at the technology, resource, systems, science and organizational level. The dynamic creation and configuration of processors into a virtual machine or system, virtual packet and lambda networks, virtual routers, virtual science projects, virtual facilities, and virtual organizations all contribute to increased infrastructure complexity.

This move to parallelism and virtualization in all aspects of technology will bring with it a large rise in the complexity of doing business with respect to the use, management, and proper stewardship of these resources. Moreover, the move to massive component-based software systems further increases the probability of system failure. Paul Horn, IBM director of research, has said, “The obstacle [to the next era of computing] is complexity. ...Dealing with it is the single most important challenge facing the IT industry.”

Increased complexity has the potential of adversely affecting the performance, predictability, and usability of petascale facilities if it is not addressed appropriately as part of a well-managed and designed infrastructure. Highly available and predictable resources, along with fault-tolerant software and hardware systems, will require the intelligent engineering and architecture of supporting infrastructures and services, in addition to the development and deployment of adaptive and autonomic systems [6] (i.e., intelligence embedded throughout the infrastructure). Many efforts in this area, sometimes competing and uncoordinated, are being pursued. However, there remains the need for developing an overall secure management framework based on standardized protocols, management, tools, systems, and services. Equally important is an object- and attribute-based infrastructure with a unified and ubiquitous name space that can support

the creation of virtual management systems comprising diverse management systems and subcomponents to support the petascale systems of the future.

II.4 Productivity and Workflow

In order to enable science at the petascale and beyond, not only must we support the concept of “time to solution” as defined in the Federal Plan on HEC [5], but we must go one step farther and support this concept from the metafacility workflow level all the way to individual systems. DOE/SC has a continuum of computing, networking, and data management resources that can be used to enable a streamlined “time-to-solution” environment. To fully exploit these resources, however, we need to develop a secure virtual management system that integrates workflow, dataflow, data management, I/O, file/object and storage management, program development environments, collaboration environments and tools, petascale experiment management and steering, and network management, all as a metasytem (i.e., at both the local and global level). DOE/SC’s top applications are very demanding with respect to infrastructure and can provide the mission-driven focus in defining a flexible architecture that leverages current tools and systems as well as support the R&D for the next generation of petascale architectures.

II.5 Petascale Management

In the FY2007 NITRD Supplement to the President’s Budget, benchmarking and performance modeling are identified as areas of focus to enhance the productivity of high-end computing systems. Given the current trends toward the integration of science domains and databases, the integration of experimental and computational facilities, and the international distributed nature of science facilities, it is imperative that we develop, enhance, and integrate *management* capabilities—not only for leadership-class computers, but also for the overall workflow, data, data management, data-intensive computing, and metafacility that support petascale research.

The term “management” in this context refers to the overall management of a petascale facility and metafacility and to the especially large challenge of distributed and integrated petascale facilities with respect to resource management as a system (i.e., processor, storage, network, etc.), including policy, priority, security, data integrity, reliability, availability, instrumentation, auditing, analysis, management data analysis, monitoring tools, scheduling, (de)allocation, pre-emption, addressing, and naming. It also includes the macroscopic-level management commonly seen in various workflow and dataflow architectures and systems. An ancillary yet important dimension of the petascale computing and experimental facilities will be the responsibility of ensuring the appropriate provenance and management of data and information as well as ensuring appropriate access to the data by both the public and science community.

Achieving this objective will require an enhanced and evolutionary approach, as well as setting the bar high for a long-term science infrastructure to better enable science at the petascale and exascale levels and beyond. Areas worth investigating include a unified and simplified name space and a secure, distributed, object-based architecture.

II.5.1 Grand Challenge of Management

During the evolution of supercomputer centers over the past few decades, we have developed the expertise, tools, and techniques to manage advanced computing facilities. However, we are now approaching a stage and level of computation and experimentation (i.e., at the petascale level with rampant parallelism and virtualization and the increased use of large numbers of software components and services) that will sorely tax our abilities to effectively manage these facilities. *The grand challenge of petascale management is the management of the interactions and interdependencies* at the software and hardware level and on both a local and global scale. The overall petascale infrastructure, much like the major experimental facilities, needs management systems that can integrate individual components, systems, and facilities into metafacilities by means of virtual management systems built from a common foundation of standards-based management systems and components. Effective management must ensure that the facilities perform at acceptable levels and that the availability and use of these facilities is predictable.

II.5.2 Local and Global Management

Management includes the management of resources ranging from individual local facilities to global metafacilities. Local systems (e.g., local clusters, leadership-class computers, I/O, and storage systems) are often located on the same floor in one building, or sometimes on a couple of floors within one building. Global metafacilities are normally located in separate buildings, sites, nations, or even continents. The challenge of managing these systems includes the integration of local and global facilities—a sort of unified field theory of facility and system management—along with the overarching issues of security, policy, resource management and scheduling, workflow, dataflow, and data management, which intersect, integrate, and aggregate all of the relevant individual facilities and system components into a multipolicy virtual metafacility, depending on the requirements of the researchers.

II.5.3 Management Areas

Management encompasses the following major areas, often on a multidomain, multisite, and multidisciplinary basis:

Management of research programs

e.g., MICS, BER, BES, FES, HEP, and NP

Management of experimental facilities theoretical programs

e.g., LHC, SNS, EMSL, CEBAF, RHIC, NERSC, and ESnet

Management of physical hardware and systems

e.g., LCCCs, NERSC, ESnet, storage, networks, and experiments

Management of software

e.g., operating systems, file/storage systems, I/O systems, and networks

Management of data

e.g., provenance of data, metadata, analysis, transformation, reduction, integrity, security, and making scientific results readily available to the public and science community.

Management of facilities

e.g., Argonne's and ORNL's leadership-class facilities, ESnet, and LHC

Management of the multidomain metafacility

e.g., the integration of leadership-class computers, experiments, capacity computing, data-intensive computing, data management, visualization servers, and networks

Management of the management architecture and tools themselves

All of these areas must be capable of being integrated into virtual management systems to provide the researcher—who will most likely be part of a virtual distributed research team—with the performance, predictability, and usability of the required facilities and systems in a secure fashion.

II.5.4 New Management Models

Many large physical experiments, such as ITER and SNS, have multiple programs and committees to oversee the development, deployment, operation, and management of the experiments. As we move into the reality of unique computational facilities, experimental facilities, metafacilities, and virtual facilities, as well as the integration of the individual petascale facilities into a metafacility in a Gridlike fashion, we may need to explore how to enhance or create new management approaches at the programmatic level to ensure that these facilities are responsive to the requirements of the DOE/SC science programs, especially for petascale science.

Just as important, however, or maybe even more important, is management at the technical level and the need for the development and implementation of a common standards-based management framework from which virtual end-to-end management systems can be built and combined with the appropriate tools and capabilities necessary to manage the OEDIS and HEC individual facilities as well as their integration into global metafacilities or virtual facilities. The issue of power management for HEC capability systems at the local system level, as well as their inclusion as a key component of the metafacility, needs to be addressed.

An ancillary yet important management issue is the responsibility by the one-of-a-kind facility or its science community for ensuring the availability and publication of scientific results to both the public and colleagues.

II.5.5 Instrumentation and Monitoring

As steward of these unique and costly leadership resources, the DOE/SC needs to develop systems, architectures, and tools that will enhance the current levels of protection against attacks such as denial of service, hijacking, and other security threats, as well as

enhance the accountability associated with the use of petascale facilities and resources. Specifically needed is the capability to instrument, monitor, audit, analyze, and manage the facilities, components, services, and codes that run on the facilities. The objective is not only to enhance the containment and treatment of unintentional errors but also to provide better insight into the use and management of the systems, codes, software components, and services, as well as their integration with experiments and simulation. Both security and system management systems focus on tracking resource usage and identifying anomalies. These capabilities will need to be enhanced and merged to determine “good” versus “bad” agents, codes, services, and “bots.” Many current systems have some form of management and monitoring capabilities, including checkpointing and deadlock prevention; however, management systems need to span all metafacility resources at a workflow level. As an example, deadlock can occur at the workflow or dataflow level depending on the high-level workflow and the integration of databases and global resources as part of the program flow and allocation of resources. Another aspect of management that requires a metalevel approach is the integrity of data as it is moved across networks, stored, and manipulated, especially when integrating multidomain, multidiscipline, and petascale resources, databases, and facilities.

System management and the codes that run on petascale facilities need to become smarter and more aware of the metafacility or virtual facility in which they reside. Achieving this goal will require appropriate middleware and APIs, as well as integrated monitoring, auditing, analysis, and management systems that work on an end-to-end basis from the application to the network and to the most sophisticated computational or experimental facility. Many individual monitoring and management systems currently exist on a per system basis, and work in under way on developing tools for the management of distributed, collaborative, integrated facilities and Grid resources, as evidenced in Condor, Globus, MonALISA, TeraGrid, and others. Many of these are still under development, however, and there is no de facto system—although use of Globus components and services is fairly common. There still exists the need for a DOE/SC consensus-based set of standard components and capabilities from which multiple virtual management systems can be built and can support the integration of the multiple set of services (e.g., disk, file, network, data movement) and management services into an easy-to-use and appropriately instrumented virtual management system that can securely combine local systems as well as the various metafacility, workflow, and data management systems. Moreover, the management of local resources needs to be more tightly coupled and integrated as part of global metafacility virtual management systems.

II.5.6 Appropriate Use

Management also includes the appropriate use of petascale facilities, not only to ensure that only authorized personnel and users have access to them, but also to ensure that the jobs are matched to the appropriate computational facilities. Effective management will include the dynamic configuration and (de)allocation of processors, I/O, networks, and other resources to a job. To do any of these tasks well will require a more detailed understanding of how petascale applications and codes interact. Much of what we know today with respect to code and the use of resources at both a local and global level will

most likely not be sufficient as we move into an era of leading-edge metafacility and petascale computation-based science.

II.5.7 Security

Security has often been a challenging aspect of advanced infrastructures and computing facilities. Because of the visibility of the DOE leadership-class computing facilities and metafacilities, they may become targets for hackers; therefore, security must be an important and effective component of the management system, without undue cost or impediment. The trend of multiresearcher, multisite, multifacility, multidataset science, the growing number of virtual institutions, the increased use of “bots” and agents in workflow systems, the coupling of OEDIS and HEC facilities, dynamic configuration and autonomic computing, and numerous naming systems and data units will increase the challenge of providing a secure end-to-end environment for science. The recent TeraGrid security attack occurred even after the TeraGrid had integrated security into its architecture. Enhanced security capabilities are needed in firewalls, intrusion and anomaly detection, sitewide authorization, access control, and auditing as integral components of an overall management architecture and system. Object-based architectures and data systems, coupled with policy-based attributes and capabilities, may provide some help in this area. But these will, in turn, require R&D to develop the right architectures, management systems, and granularity at which to implement security features, as well as a manageable number of naming and associated resource location, management, and security systems.

II.5.8 Morphable Infrastructure

Another challenge facing the DOE/SC community as it moves into petascale computing and beyond is that it will require new technologies, capabilities, and architectures at both the local system and metafacility level. Advanced technology churn occurs on a recurring two- to three-year cycle. An infrastructure is needed that enables the programmer and researcher to “test drive” new technologies while keeping some part of the environment at production quality so that, if a problem arises through the use of the new technology, the researcher can revert to a safe production system. The concept for this type of an infrastructure was initially identified in MORPHnet (1997) and has since been incorporated as a basic premise of the National Lambda Rail and other networks, as well as being integrated as a major tenet of the Internet 2’s next-generation network.

This concept of morphing resources and systems between research and production modes needs to be extended not only into the local area, cluster, and storage area networks but also into the end systems themselves, including storage systems, I/O systems, and the computers. The ability to use a dynamically morphable infrastructure will be crucial for accelerating scientific use and the evolution of petascale facilities and beyond, through the introduction of new technologies and architectures to better match the needs of science. The management of such an adaptable and morphable system is itself a research challenge; but, if successfully tackled, it will accomplish much by rapidly moving computational science into petascale and even exascale environments and beyond.

II.5.9 Naming

Another challenge facing the successful management of petascale systems is the naming and management of objects. We currently deal with a plethora of naming domains, architectures, and systems as found in the networks, operating systems, workflow systems, dataflow systems, and file systems. The ideal goal, albeit most likely unattainable, is to have one naming and addressing system that covers the whole facility and metafacility, from memory and disk blocks to network packets and object-based file systems. Given the grand challenge of management facing DOE with respect to its petascale facilities, this may be the time to revisit the notion of building the future metafacility on the basis of objects and attributes or capabilities as the building blocks, albeit focused on high-end scientific use, with a single global and scalable name space upon which one can add metadata and attributes to any component of the metafacility including subcomponents such as processors or ports, as well as file systems and whole systems of hardware, software, and data. Such a system would further enable better security, monitoring, and management of the resources.

II.5.10 Object-based Infrastructure

A recent trend in the scientific community has been the use of object-based infrastructures and services. The secure management of the petascale system and metafacilities will be enhanced by having at least the major components of the systems identified as unique objects with relevant metadata and attributes associated with the system objects to control access to the components, as well as support other security, policy, and management functions. Examples of relevant system-level objects are files, processors, nodes, network ports, disks, tape drives, firewalls, services, software components, and file systems. Other examples include packets, lambdas, virtual networks, virtual routers, virtual firewalls, multicomponent objects and files, and blocks of memory or storage. An object-based system needs to be able to express relationships among objects, pedigree, level of assurance, and other relevant interobject attributes, as well as a unified standards-based naming system. Such a system would better enable the addition of attributes for security, auditing, context, and other relevant metadata necessary for the management of the metafacility and its subcomponents, as well as the data.

II.6 Summary

DOE/SC's long-term goal for petascale, exascale, and yottascale computing and science should be a smartly instrumented, secure, performance-focused, predictable, usable, adaptive, morphable, and autonomic metafacility in 15 years. The roadmap should establish a program and associated budget within DOE/SC/ASCR to initiate a new focus on the area of *management*, to include data management, visualization, data-intensive computing, and the proactive and dynamic management of resources, as well as management of the integration of observational and experimental facilities with computational facilities and metafacilities. A crosscutting program aimed at managing

petascale facilities should leverage the monitoring and management work started by Globus, MonALISA, Condor, TeraGrid, OSG, and other workflow and management systems.

An initial step would be for DOE/SC to sponsor one or more workshops with experts from all relevant fields to focus on this grand challenge. The workshop participants would identify the best processes, architectures, and tools for managing and processing the resulting scientific data and results. The participants would also identify current activities that can be used as a base for starting the evolution toward a smarter set of management architectures and systems for supporting petascale facilities and the petascale metafacility. In addition, the workshop participants would identify longer-term computer science research (e.g., object-based petascale architectures and systems) needed to support the goal of petascale science and beyond.

ASCR should also consider creating a *computing systems management sciences* program and discipline focused on developing and supporting HEC systems and their integration with each other, as well as with leading-edge petascale and exascale observational, experimental, and data-intensive science facilities in a secure and manageable manner.

II.7 Structure of the Report

Part I is the executive summary.

Part II is the introduction and overview of a petascale facility, related challenges, and the definition of management with respect to petascale facilities..

Part III is a compilation of relevant challenges associated with the petafacility of the future, along with potential relevant research areas.

Part IV contains recommendations for a DOE/SC program on managing petascale facilities and identifies high-level research areas.

Appendix AI is a brief description and definition of the terms for leadership computing, storage, networking, and experiment facilities as used in this report.

Appendix AII provides an overview of trends in scientific research, which in turn affect both current and projected supporting infrastructure based on these requirements, especially at the petascale level.

Appendix AIII provides a brief overview of current and future DOE petascale facilities, highlighting their capabilities and requirements and the impact on supporting infrastructure and computer science research.

Appendix AIV provides an overview of technology trends in processor, computing architectures, storage, I/O, file systems, and networking, with the focus on potential

impact of future petascale systems and metafacilities.

III. Challenges and Research Opportunities

The old Chinese saying “May you live in interesting times” is applicable to today’s computing environments. The move into petascale computing will only make the times more interesting.

III.1 Scaling of the Petascale Facility and Metafacility

As the age of petascale science arrives, we will need to address the challenges of scaling, supporting, using, and managing petascale facilities in order to ensure their productive use by researchers. Current techniques, architectures, and tools for monitoring and managing these facilities will most likely not scale, and today’s resource allocation and management systems are just beginning to be developed and deployed.

In addition to the move into petascale facilities, there are an increasing number of multisite collaborations combining both computation and experimentation, all of which require a secure workflow and distributed computing infrastructure. More and more researchers are collaborating on a single research project is increasing, and research projects are becoming more international. Consequently, a better understanding of the evolving social science organization mechanisms, along with the evolving technical OEDIS and HEC facilities, is needed in order to develop, engineer, and architect an enabling infrastructure. As the scale and complexity of the petascale facilities and components increase, there will be an increased rate of use of smart “agents,” “bots,” and “actors” as part of a workflow and management system. For this type of autonomic computing and facility support to become a reality and provide added value, further development and implementation are needed focusing on next-generation monitors and instrumentation at the component, system, facility, and metafacility level.

With the growing trend toward service-oriented architectures, an increasingly important component of a management system is the *definition and characterization of the various services* that make up a facility and metafacility, so that such services can be combined, scheduled, and used as part of the researcher’s workflow and job management process. As an example, movement of data will most likely include a local storage service at both the source and sink sites (i.e., disk systems and file systems), a data-generating experiment or capability computer, (potentially) a visualization service, and the various network services (WAN, MAN, LAN, etc.), each of which has different characteristics with respect to latencies, jitter, and bandwidth and each of which may have different management, naming, and security policies.

The metafacility is really just one type of a service-oriented architecture; and each of the services just noted needs to be characterized and instrumented such that a metafacility resource management system at the workflow level can determine whether the appropriate resources and services can be made available and combined to accomplish the function requested. An example of a metasystem tool that could be useful is a systemwide “ping” and traceroute for tracking the flow of the data on an end-to-end basis (i.e., memory or storage block to memory or storage block) across a network. Initial work

toward this end has been done by the SciDAC Scientific Data Management Center based on timed syslogs for file transfers, and other “pings” exist at the distributed system level. These systems and concepts need to be expanded and evolved.

Another example of the trend toward a service-oriented architecture and the outsourcing of services is ESnet. An experimental prototype process is being supported by ESnet where the ownership and management of a wavelength-based “layer one” network are being outsourced to another R&E organization. This may be the first step of outsourcing the other layers of networking (e.g., the support of multiple IP and policy networks on separate lambdas) and may even portend the outsourcing of other scientific infrastructure. Amazon, Sun, and others are already offering capacity storage and computing services. The important relevant issue with these trends is that, in order to demonstrate appropriate stewardship of DOE/SC-supported infrastructure and resources, especially with outsourced services, considerably more effort will need to be expended in the definition of observable and enforceable service level agreements, which in turn requires better characterization and instrumentation of the infrastructure, along with appropriate monitoring, verification, auditing, and analysis tools.

All of the individual and metafacility management and analysis systems will require a solid base of instrumentation, monitoring, analysis, management, and application of policy in order to ensure the proper stewardship of these petascale capability and metafacility resources.

A major challenge in the evolution to petascale computing and beyond will be balancing the need to reduce the complexity and overabundance of systems with the wish to encourage a Darwinian R&D approach to the evolution of these same critical technologies and architectures, where one can continue to experiment with new technologies and techniques.

In summary, all of these trends require a full-scale effort focused on a new program for the management of infrastructure and facilities for petascale science and beyond. The rest of this chapter further delineates some of the research challenges and opportunities in various technical areas that are relevant as part of this recommended program on management. Each subsection identifies technologies, middleware, and architectures directly related to the functioning of the petascale facility and metafacility.

III.2 Virtualization, Virtual Organizations, Virtual Facilities, and Grids

The trend toward multisite, multidiscipline, multifacility, multidataset, multiresearcher science is a reality. *Virtual organizations* (VOs) arose as a result of this trend. People who are part of a VO, whether or not they are of the same science or discipline, are working on the same project and research. There is also a trend toward the development of the *virtual facility* (VF) (called a metafacility in this report and often referred to as a *Grid*). The virtual facility is a combination of some subset of experimental and computational facilities, or any combination of computational, experimental, data intensive computing, storage, visualization, and networking facilities. The location and

placement of facilities and resources may become dynamic in the future to better support the research being undertaken. For example, data storage and caching services could be located in the network or spread across numerous sites. These types of virtualization will require an enhanced ability to identify and define the resources as well as track and manage them in a secure and usable manner. Since there will be many dynamic and virtual overlay architectures, organizations, systems, and facilities, there will be the need to support multiple integrated virtual management systems. These management systems should be derived from a varied set of standards-based management services and components, with appropriate application-accessible control points and interfaces to enable the management of multiple systems and services.

A researcher needs to be able to locate, request, and use facilities at multiple sites and administrative domains. Grid and workflow systems of today, much like the job control systems of yesteryear, are trying to address these requirements; yet today there remains a plethora of naming systems, services, and domains that a researcher has to navigate across all systems. The result often is confusion, and the potential for error is high. An ideal goal would be a single, unified naming system and object-based architecture upon which one can attach attributes and capabilities for the appropriate and secure matching of codes and data to resources. A more realistic approach would, at the very least, involve reduction of the naming domains and systems to a small, manageable number. Initial techniques and services applicable to the location, allocation, and monitoring of use of distributed resources are being investigated now as part of the Grid work in Globus, Condor, MonALISA, TeraGrid, and other such systems. Many of these are still in development and will require more support, investigation, and development to provide for a robust and flexible management system that securely integrates and manages the local site resources, as well as the metafacility workflow level, with the appropriate level of instrumentation, monitoring, auditing, and analysis. Of particular interest for VO and VF security and stewardship is the area of intersite and interdomain policy definition and enforcement. This includes the integration of local and metalevel resource allocation, management, and monitoring systems. Initial work in this area, along with attribute-based access control, is being undertaken by Globus Toolkit 4 (GT4) and other distributed access control and authorization services.

Specific trends in technology virtualization include the use of parallel WAN network transfers acting as one (e.g., GridFTP), the use of multiple network switches acting as one switch (load balancing), the use of virtual private networks at layers 2 and 3, extending cluster area networks (e.g., Infiniband, Ethernet, fiber channel) across an IP-based WAN, parallel I/O acting as one transfer, and striping I/O over multiple devices. Virtualization on commercial routers includes the deployment and support of virtual and logical routers complete with their own policy and routing databases. Virtualization on computing systems includes the ability to configure the number of compute and I/O nodes on a per run basis, as well as supporting multiple OS images.

The challenge of managing these virtual components, especially the transient nature of many virtual facilities, will grow as the speed of these components and the number of systems grows. A major challenge in managing petascale virtual systems will be

determining the right type, scale, placement, granularity, and frequency of the instrumentation and monitoring tools, to better understand and manage the resources without adversely affecting the individual system or metafacility.

Early work in sensor networks and ad hoc networks explored the use of an adaptable and morphable set of monitoring systems and resources that dynamically determines what it needs to monitor and where to enhance the manageability of networks. This concept should be investigated to see whether it can be extended to virtual facility and systems management. In addition, research is needed on the feasibility of applying lessons learned from the human immune system and biological networks to the dynamic placement, monitoring, and analysis of petascale facilities and metafacilities..

III.3 Architectures, Systems, Facilities, and Metafacilities

The secure management and analysis of petascale facilities (leadership-class computers and petascale experiments) present a daunting task; but the integration and use of multiple scientific datasets, the coupling of simulation and experimentation, and the general distributed collaborative nature of international science make the challenge even greater, especially as administrative and scientific provenance boundaries are crossed.

The term “management” in this case not only refers to the instrumentation, monitoring and analysis associated with managing data, systems, and components but also includes the security associated with the use of these facilities, as well as the secure identification, location, allocation, accounting, auditing, and management of individual resources. In addition, more work needs to be done on generating consensus on a common set of taxonomies and ontologies for describing petascale resources. For instance, the term “network” can refer to a system backplane interconnect, cluster interconnects, LANs, MANs, WANs, VPNs, and more. The term “global address space” is often used to mean the address space that covers the thousands of processors and memory found within a leadership-class computer, yet “global” is also used to describe distributed resources located in many different countries. The Grid Laboratory Unified Environment (GLUE) schema work for system elements is a good beginning for a facility schema.

The interagency Large Scale Networking committee noted in the NITRD Report the goal of “end-to-end network performance monitoring and measurement” [7]. Given that the network is much more than the WANs, MANs, and LANs normally envisioned when using the term networks—it also includes cluster area networks (CANs) and storage area networks (SANs)—a combined management approach needs to address the end-to-end integrity, movement, and management of data across all types of networks: memory to memory, application to application, and storage to storage (disk to disk).

Many monitoring and management systems exist for storage, file systems, high-performance computers, and the various types of network. In addition, many efforts are under way to develop workflow and Grid-level metafacility processes, data management, and monitoring capabilities. Python, GRAM, MonALISA, Kepler, SRM, INCA, and other Grid and workflow-related systems are all positive moves toward a usable

heterogeneous scientific programming environment. However, the wide variety of such management systems and subsystems, along with their associated security models and naming systems at both the local and global level, makes it challenging for anyone trying to administer or use these resources. A major focus is needed to identify and develop a suite of standards-based core management capabilities and services, along with interfaces and control points, which can be combined to create a virtual end-to-end management system to support the petascale facilities and metafacilities. These monitoring and management systems also need to support easy-to-use, secure APIs for monitoring long-running codes, which often span petascale facilities and administrative domains. This type of support, along with checkpointing and recovery mechanisms, needs to be an integrated part of both individual facilities and the metafacility.

Another growing trend is the insertion of more intelligence into the networks, systems, storage, file systems, and workflow systems. With advances in ASICs and FPGAs, as well as multicore multipurpose nodes, we can expect this trend to continue and will need to address the challenges associated with it. However, the adoption and use of these technologies and capabilities as part of an overall management plan may also provide some of the help necessary to support a “smart petascale facility” and “smart metafacility.” More research is needed on how to leverage the work in nanotechnology, ASICs, and FPGAs for use in the instrumentation of and monitoring of facility components and services (e.g., I/O and storage) for the purposes supporting efficient resource allocation, striping, and autonomic computing.

Increasingly, “agents,” “bots,” and “actors” are being deployed as part of metafacility and Grid management systems, as well as autonomic facilities, at both the local and global level. Many of these are specifically designed to support the parallelization and virtualization of resources. The trend to larger and larger numbers of processors, network interfaces, I/O interfaces, and components at all levels of the petascale facility and metafacility will bring with it an increased probability of component failure; therefore, more focus is needed on the design and architecture of an overall system to achieve high availability. An integrated monitoring and management system is needed that goes beyond single leadership-class systems and includes support for metafacilities, workflow, dynamically configured facilities, virtual architectures, autonomic self-monitoring and self-healing systems, and self-describing programs (i.e., identify hardware and software resources the program requires, much like the old JCL days), which can then trigger appropriate monitoring and allocation systems.

A major challenge facing petascale researchers and computer scientists will be developing a better understanding and control of the resources in individual facilities, metafacilities, and virtual facilities. Such an understanding is paramount for effective system performance tuning and monitoring, system benchmarking and analysis, detection of system anomalies, application performance tuning, and the dynamic allocation and configuration of resources. Research into providing dynamic configuration and use of resources (e.g., morphable compute and I/O nodes or dynamically allocating I/O and network pipes) will better address the researchers’ requirements to tune, debug, and run codes efficiently and effectively, as well as enabling experimentation with new hardware

and software capabilities. For example, Fermilab has instrumented its disk-based storage system and built a monitoring system to better understand and manage the use of its storage system. This type of capability needs to be coupled with appropriate middleware and management systems to integrate information about disk, storage systems, computer systems, and networks to support end-to-end resource management and allocation.

A major aspect of managing any facility and infrastructure is ensuring its integrity, availability, and security. “Integrity” refers to the integrity of the data, the facilities, and all of the components and subcomponents. “Availability” covers the availability of resources from an allocation and scheduling perspective as well the detection of resource failure and recovery. “Security” includes authentication, authorization, access control, intrusion detection, and prevention of denial of service. Security, resource management, and system management all need to be integral components of the petascale facility management system.

Another major challenge facing the designer, administrator, or user of these complex, distributed, and often heterogeneous facilities is the need for an integrated policy management system that incorporates the security, network, workflow, system, and facility management into one system such that the facilities are secure and yet enable the researcher to pursue their science without being impeded by conflicting and competing policies or overly constraining the use and performance of the facilities. The management of the sometime conflicting policies of security, QoS, VPNs, load balancing, resource allocation, resource schedulers, and system management plays an important factor in the availability and integrity of resources. The policy system needs to address who controls and can modify resources, in addition to when and where.

Given the current trend toward higher levels of complexity in high-end computing facilities and the e-science workflow trend to integrate and combine various types of capability and capacity computing, storage, and networking systems, it is imperative to develop new and evolve current techniques and metrics for benchmarking and analyzing systems and facilities. Enhanced information and analysis of infrastructure capabilities will help us measure utilization and identify bottlenecks associated with I/O, file system, CANs, and network systems and services. As the need for parallel I/O at all levels increases, so will the need for better monitoring tools and architectures, data analysis tools, metrics, and techniques.

The interagency HEC team is supporting development of enhanced benchmarking capabilities and tools for leadership-class computers. This is a good first step; but more attention needs to be focused on overall system-level and workflow-level benchmarking and analysis with a focus on time to solution and wall clock time, even if the initial step is to develop better benchmarks, metrics, and analysis tools for each facility subsystem (e.g., I/O, file system, storage, interconnects, cluster area networks) and then combine these systems into an integrated virtual system. To do so may require a new definition of metrics and resource units as well as service definition models for networks, storage and other facilities. Service definition work has been started as part of the Global Grid Forum,

Condor, and other Grid and workflow architectures and needs to be continued and evolved.

An aggressive goal would be to have a standards-based core set of management systems capabilities by 2016, that is, local and global systems that could be combined to form virtual management systems capable of monitoring, auditing, managing, resource (de)allocation, securing, and using any and all of the DOE/SC-supported petascale facilities and metafacilities.

III.4 Networks

Networks are the glue that interconnects all types of scientific components and facilities into a metafacility. Such networks will become even more important for petascale science. The dependence on high-performance, predictable, and usable networks is increasing; yet our ability to manage and control the networks has not evolved much since their inception over 20 years ago. Many of today's network monitoring capabilities and tools are fairly crude; for example, SNMP, ping, and traceroute are used for debugging and monitoring. Some of the active network performance tools such as PerfMon inject traffic into the network and thus can affect the system they are trying to measure. In addition, given the nature of a best-effort IP network, the most recently acquired network status may not be valid by the time an application chooses to make use of the resource. This challenge of determining network status will only get worse as the speed and bandwidth of the networks get faster, the parallelism increases, and virtualization takes hold. DOE/SC needs to focus effort on next-generation network allocation, monitoring, management, and analysis tools.

Although vendors are making 40 Gbs ports and links and are testing 100 Gbs clear channel optics in the labs, and the IEEE committees are already engaged in 40 Gigabit Ethernet and 100 Gigabit Ethernet standards development, the "economic sweet spot" for networks for at least the next four to five years will be at the 10 Gbs level. Anything above 10 Gbs will most likely be short-range optics or will be accomplished by integrating multiple 10 Gbs waves into a chip or running parallel striped 10-gigabit links. Subsequently, there will be a need to develop mechanisms for the allocation, management, and monitoring of these striped networks, which will be used to move large amounts of scientific data. Challenges exist for developing open source passive monitoring capabilities that can be used as part of a management system for networks at the 10-gigabit level and above. OC-48 mon, an evolution of OC-3 mon and OC-12 mon, was developed to provide monitoring capabilities on research networks. It is no longer supported because of the move to 10 Gbs networks. The National Science Foundation has just instructed the National Laboratory for Applied Network Research (NLANR [8]) to disperse with the old OC-3 mon and OC-12 mon equipment.

Work is being done on an OC-192 monitor; but at the current time it is not funded at the level necessary to achieve success in the near future. Even more important is the need to have capabilities for both OC-x and 10 Gigabit Ethernet monitoring to enable the analysis of networks on an end-to-end basis (i.e., LAN, WAN, and MAN). NLANR and other

researchers are investigating the use of lambdaMONs for passive network monitoring of dense wavelength division multiplexed (DWDM) optical networks. These monitors are architected to enable the collection and real-time analysis of IP packet data from any one of the wavelength carriers on a DWDM optical network.

DOE/SC should support the development and deployment of network-based instruments, monitors, and capabilities that could be used to gather information with respect to the use of networks on an end-to-end basis. DOE/SC also should support the integration of relevant network monitoring and security tools with tools for the management, analysis, and visualization of the data generated by such a management system. These tools and capabilities will be crucial for supporting the proper stewardship and accountability for services, such as the outsourced network services envisioned for ESnet and its ancillary Science Data Network.

The NITRD Report [7] specifies end-to-end agile networking as one of its goals. It also specifies QoS and GMPLS as important capabilities at the network level to support the requirements of the sciences. Many challenges need to be addressed to enable such agile networks. End to end really needs to be application to application, which means that any management and allocation system deployed needs to integrate cluster area networks, storage area networks, local area networks, metro area networks, and wide area networks. A crucial first step is for better instrumentation in the networks, coupled with more powerful APIs for both the applications and middleware to gather information about the network. CAIDA [9] and its predecessor NLANR both focused on network analysis, mainly with respect to peering points and attached storage for WANs. However, more R&D needs to be supported to better understand the true end-to-end network traffic, especially as we move toward the use of virtual networks and hybrid networks that support both packets and circuits.

The second challenge for enabling an agile end-to-end network is the integration of the network as an equal resource into an overall secure metafacility and workflow management system. This integration will require defining attributes and characteristics (e.g., component and system metadata), along with implementing the appropriate instrumentation, for networks ranging from cluster area networks through wide area networks. In an ideal world this would also include interconnect area networks, backplanes, and I/O nodes..

The lure of lambdas and the recent trend toward the use of circuit-based wavelength services has occurred mainly from the perception that they are a more predictable and deterministic network resource than “best-effort IP” (BEIP) networks, where the link is often shared with other applications. Current BEIP networks have been cited as not being well adapted to the efficient movement of large datasets and flows often associated with the sciences. There also exist the challenges of BEIP-related latency and jitter. These are challenges in today’s terascale science support networks and will only get worse as scientific applications move into the petascale. Even when used on circuit-based and dedicated wavelength services, traditional TCP and its sliding windows algorithms can often impede the full optimal use of the circuit by large scientific flows. As a result, many

science projects that routinely ship large amounts of data look to use circuit-based wavelength services and alternatives to TCP as a transport protocol. TCP and derivative transport protocols such as FAST, XCP, and HS-TCP address some of these shortcomings, even at 10 Gbs. However, continued research is needed to ensure effective evolution of TCP and IP for high-speed networks.

Hybrid networks generally refer to those that incorporate BEIP and circuit-based wavelengths in one infrastructure, albeit currently in a parallel and not an integrated manner. ESnet is pursuing this model and is augmenting its IP network with a wavelength-based circuit network called the Science Data Network (SDN). These circuits are initially planned to be statically set up in parallel. The initial SDN deployment consists of dedicated circuits between a small number of high energy physics applications for the purpose of supporting the large volumes of data that will be moved as part of the LHC experiment. The objective is to provide network services with scalable cost. Once the other DOE/SC science programs also start generating and moving petabytes of data, ESnet will need to support a fairly dense mesh of dedicated wave circuits to the DOE/SC supported sites and laboratories or will need to delve more deeply into the dynamic management of layer 2 and lambda-based circuits. Plans for future generations of ESnet call for additional waves; but the determination of the required density of the mesh to support all science requirements is not yet clear. Financing a full “n by n” mesh is most likely not going to happen. Such a solution will require more than the current GMPLS protocols and reservation systems currently being developed and prototyped by the network community. Further research, development, and experimental deployments are needed in this area, especially to support multidomain management; such efforts will need to leverage and integrate the concepts of policy-based resource management developed as part of the original timesharing operating systems with respect to priority, preemption, allocation, and time division multiplexing. This work will need to include the concepts of prescheduling, dynamic scheduling, and preemption of network resources based on policies and priorities developed by the DOE/SC ASCR office and its community. Moreover, this work will need to integrate programmatic oversight of allocations as well as interdomain policy servers and resource managers.

DOE/SC should support the development and use of a dynamic hybrid network that would concurrently support BEIP services, wave circuit and layer 2 circuit services for large dataflows, and a network tailored for low latency and jitter to support visualization and collaboration-intensive research. The initial implementation could start as a wave dedicated to each of these types of network service (i.e., each site has three wavelengths or subwavelengths) and a policy server and resource manager that would take requests from client applications, where the application states which type of services it desires. Alternatively, the network could monitor the traffic and dynamically select the type of network circuit based on attributes or metadata associated with the data objects being transmitted.

Another interesting R&D area in networking involves the scope and use of the optical plane versus the router plane, as well as their interaction and coexistence. Research is needed on the current trend to move the intelligence of wave management out of the

optical switch and optical control plane and place it in a router via a router blade and integrated layer 1 and layer 3 control plane; the optical switch acts as a dumb optical mux/demux, and “alien waves” are managed through these at the request and control of the router. In addition, more research is needed to enhance IP over DWDM (e.g., the alien waves) with respect to issues of latency and jitter at the IP router level, mainly with respect to buffering, queuing, and nonblocking flow control. DOE/SC should support the investigation, development, and deployment of these different types of architectures and technologies and develop the appropriate instrumentation and monitoring capabilities to better understand which will scale to support the requirements of petascale science.

Both the federal HEC Plan and the NITRD/LSN reports identify as a long-term goal the development of optical interconnects. At the cluster, site, and distributed levels the Optiputer [10] project is combining multiple sites with an IP over optical switch. The Optiputer can dynamically (de)allocate 10 Gbs waves to applications. DOE/SC should investigate Optiputer and other optical switch-based architectures to be integrated as components of its petascale infrastructure. DARPA is supporting the development of all optical data routers to run at greater than 100 Tbs. UCSB has demonstrated 40 Gbs and is expecting to demonstrate 100 Gbs by the end of the decade. DOE/SC will need to evaluate how to integrate and manage this new class of optical data routers as part of its overall petascale metafacility architecture.

Given the trend toward convergence and merging of interconnect area networks, cluster area networks, and local area networks as they evolve to the 10 Gbs level and beyond, as well as the trend to more offload processing for IAN (e.g., TCP offload engines, aka TOE) and CAN protocols, it is important that DOE/SC understand the impact on applications with respect to latency and jitter, so that one can map appropriate technologies to codes and encourage standards-based solutions. DOE/SC should experiment with, benchmark, and better understand how the various IAN, CAN, and LAN protocols interact and affect the efficiency and capabilities of petascale facilities and metafacilities. To this end, DOE/SC should support a multisite institute focused on the evaluation of IAN, CAN, LAN, and WAN protocols, along with their tunneled counterparts such as IB over IP and E-RDMA, so that the appropriate network technologies can be used to support the applications.

The recent development of virtual and logical routers, if they are appropriately instrumented and managed, can help support the efficient and effective use of network resources, as well as provide for individually managed virtual networks by the user or program. Research on the management of these virtual resources is needed, as well as R&D with respect to monitoring tools and data gathering and analysis tools for both real and virtual network systems and resources.

IAN, CAN, LAN, MAN, and WAN protocols each have their own naming and addressing systems and techniques for addressing latencies and jitter. An ideal goal, albeit probably a holy grail, would be to have one network protocol data unit (PDU) that could be used on an end-to-end basis across these networks. Minimizing the number of protocols supported in these networks would help reduce some of the complexity of the

system. Each time a PDU is “touched,” examined, or managed, there is an increase in the consumption of power and the generation of heat in each device involved. An ideal goal would be to reduce the number of times the data or PDU is “touched,” as opposed to switched or routed, and thus reduce the complexity of the system, the potential of error for every transformation or handling, and the power consumed and the heat generated.

In addition to the naming and PDU challenges associated with the current plethora of network protocols, another unmet need is a systemwide “ping” and traceroute that could navigate and traverse the varied technologies and protocols across the various networks.

III.5 Data and Information Management

Data management will remain one of the top challenges in managing the petascale facility. The NITRD Report [11], page 11) identifies the need for a new focus on data management and revolutionary file systems. The DOE/SC Data Management Report [12] not only outlines the many challenges facing the scientific community with respect to the management of scientific data but also identifies multiple areas where additional R&D needs to be supported. Given that this report is only two years old and had forecast the movement of science into the petascale area, we will not duplicate that work here. Instead, we will briefly reference those definitions, references, and recommendations where appropriate to reinforce the need for DOE/SC to support the report’s recommendations.

Petascale to exascale I/O, storage, and file systems will be required to support petascale capability computing, petascale experiments, and petascale data management facilities. To this end, petascale I/O and file systems need to scale and work both within local onsite facilities and at a metafacility distributed level. The trend of integrated databases and sciences does not lend itself technically, strategically, or politically to all of the petascale capability resources being collocated in one place. Consequently, petascale science infrastructure will need to address the issues of naming, locating, accessing, and managing file and object systems on a truly global scale. NERSC already has in production the NERSC Global Filesystem (NGF), a high-performance parallel global file system that is accessible from all of its multivendor computing systems; Argonne National Laboratory is continuing the development of the Parallel Virtual File System (PVFS [13]); and other file systems for high-end computing systems are under development. A golden opportunity exists for DOE/SC to take the lead in defining and implementing a petascale capability metafacility-wide file system that is capable of supporting all of the requirements of the various DOE/SC science program requirements and sites.

In addition to file systems, more effort is needed to develop the parallel I/O and parallel network synchronization necessary to support a petascale facility. Another ancillary challenge is being able to effectively move data from the leadership-class computers and experiments to storage. Data management encompasses data provenance, placement, transformation, movement, transmission, storage (temporal, long term, archival), reduction, analysis, location, query, and access. More analysis and research are needed in

support of the dynamic allocation of I/O and network systems and nodes as well as protocols and architectures, in addition to morphable and dynamic I/O and striping architectures.

Challenges and opportunities in the management of data for petascale science are bountiful. They include data movement via the use of networks and I/O (transmission, switching, routing), the merging of movement and storage functions (i.e., caching, RAM disks, optical disks, optical networks as temporal storage), distributed storage systems, the intelligent and timely staging of data in hierarchical storage, intelligent data reduction and transformation, compression, indexing, location and search, the annotation of the data itself (commonly captured in the term metadata), and the overall movement and management of data as incorporated in files/object systems, visualization systems, dataflow and integration, and workflow systems. Security is an essential component for all of these. In summary, petascale science will require systems that can ensure that the data arrives at the right place, at the right time, secure and intact, and that it can be easily used by the researcher.

The challenges associated with the transmission and movement of data are tightly coupled with the challenges associated with the combined use of naming systems, file/object systems, networks, and I/O. Other research areas include the integration of intelligence at diverse levels, from storage device, I/O and network device, to system, in order to handle “on the fly” analysis, annotation, transformation, and storage of data. An area that has not been well addressed is the architecting and engineering of autonomic, intelligent, dynamic, adaptable, and network-based data management and movement systems.

Additional development and effort in the area of metadata schemas and global naming systems would also help establish a manageable and usable infrastructure. Rather than one single schema or preferred way of defining relevant metadata, a small number of science domain-specific schemas and techniques should be identified. In addition, research is needed on the treatment and management of data from the perspective of object management. Most data management systems currently work with files. Often, hundreds to thousands of files and associated names are used for one job. The successful support for evolving and enabling an object-based storage system (e.g., LUSTRE) to address the DOE/SC science requirements would enhance the ability to manage and analyze petascale datasets.

The appropriate annotation of data and data subcomponents, coupled with the use of policy and security attributes attached to data objects, infrastructure elements, and systems, could better aid the management of data. Naming is an important aspect of data and system management. A unified object-based naming system (e.g., one based on the Digital Object Identifier system) applied to the networks and file/storage/data management systems could enable a more secure system. The long-term goal for facilities and metafacilities at the petascale and beyond should be an infrastructure based on objects (services, hardware components, software, data, packets, blocks, metadata, files, ports, processors, memory, etc.). Achieving this goal would enable a higher level of

integrated and securely managed programming environments for the user, with the combined benefit of a unified name space and interdomain public key system.

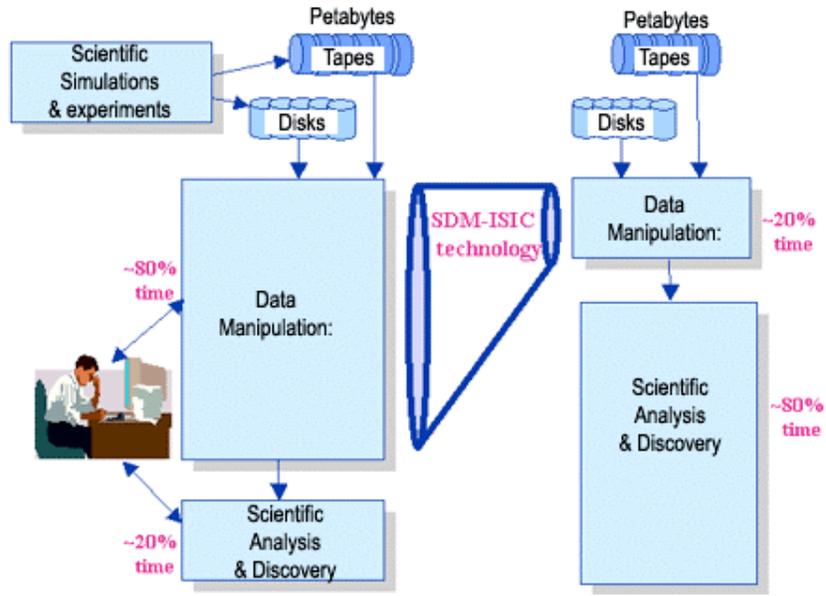
The term “object based” also includes the concept of large-scale objects such as the component-oriented architecture [14] framework developed at LLNL. The infrastructure should be focused on the principle of movement, computation, storage, analysis, switching/routing of objects, with the lower-level embedded systems having the intelligence and ability to deal with the slicing and dicing of bits, bytes, blocks, packets, and circuits. The use of objects needs to remain at a high enough level to avoid impeding performance or introducing an overly cumbersome granularity. A research challenge is determining the “sweet spot” of definition for an object-based infrastructure, that is, the appropriate granularity for the definition of objects at the data, device, service, system, and facility level to enhance functionality.

A long-term R&D program with respect to data management should be focused on developing an end-to-end object-based network and data management/file/storage system. The network would route, switch, and move the data objects based on the attributes and metadata associated with the data and relevant system and infrastructure elements, which would subsequently enable more intelligent policy-based infrastructures with the potential for automated use of parallel resources (e.g., an object-based SRM), path and QoS selection of transmission, and the appropriate level of caching, reuse, placement, and storage of data. An opportunity exists to develop a science-focused object file system complete with a consensus-based set of infrastructure taxonomies and ontologies.

III.6 Workflow

Workflow is a superset of scheduling and management systems associated with each of the facilities and its subcomponents (i.e., operating systems, file systems, and networks). The DOE/SC research community needs to develop a set of capabilities for an advanced secure and effective management system and framework that integrates capability and capacity systems, both production and experimental infrastructures, local- and global-level systems, policy management, a unified field theory of naming and addressing, job control and scheduling, and both the static and dynamic (de)allocation of resources. The ideal goal would be to provide the researcher with a standard management control plane system, protocols, tools, and infrastructure that would enable use of the capability and capacity petascale facilities, without dealing with a plethora of competing or complementary models, systems, protocols, or the like.

Workflow systems have been under development by the Global Grid Forum, MonALISA, TeraGrid, and others. A Storage Resource Manager [15] has been developed by the Scientific Data Management [16] team (Figure III.6.1) as part of a SciDAC project. It focuses on dynamic space and file management at a metafacility and Grid level to enable the smart movement of data as part of an overall workflow system.



**Figure III.6.1 Storage Resource Manager
(courtesy of LBNL SDM website)**

Figure III.6.2, included from the Data Management Report [12], depicts a generic science workflow system. The report notes many R&D gaps and opportunities with respect to dataflow and workflow systems. One example is the need to address the scalability limitations of graphical representations of control and dataflow in scientific workflows as the number of components becomes large. Another example is the need to define a scientific workflow language that can describe inputs and outputs for each component, workflow metadata, granularity of tasks and subworkflows, task invocation, and the human tasks of notifications and alerts, dataflow streaming granularity, and performance expectations. The study also notes the need for more research into the analysis of workflows and workflow patterns similar to the effort done in the past for software patterns on single systems.

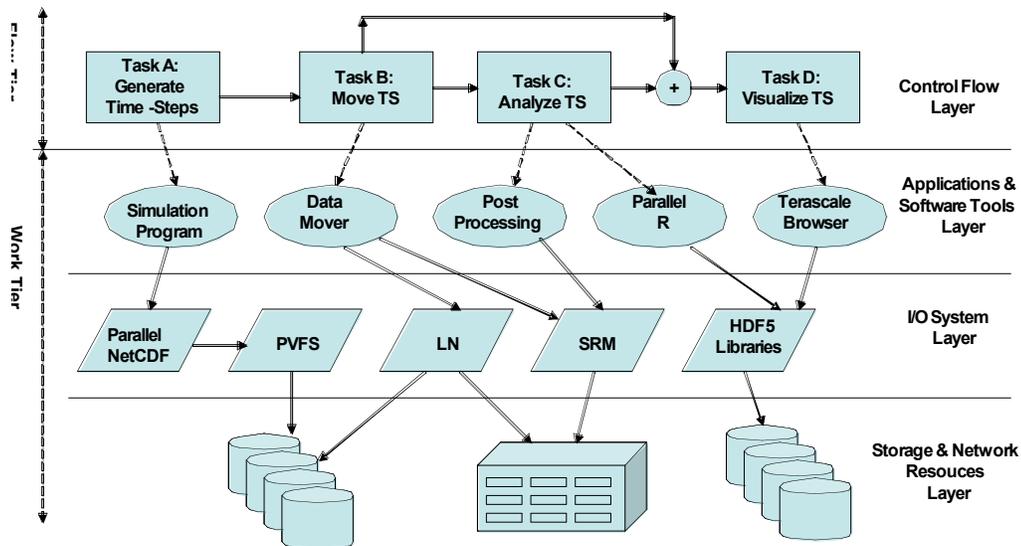


Figure III.6.2 Anatomy of a scientific workflow-management.

Control Flow Layer – execution ordering of tasks by using different views of sequencing, branching – different constructors that permit flow of execution control; **Application & Software Tools Layer** – invoked applications and software tools used by the workflow tasks; **I/O System Layer**– the I/O systems that allow efficient read and write operations by the applications (predicted data volumes and their characteristics can also be described in this layer); **Storage and Network Resource Layer** – information about physical devices used by the tasks during their executions (e.g., required data transfer rates).

With regard to the handling of security in the workflow management system, the report makes two recommendations. The first recommendation is that the workflow management system “provide access controls to the scientists and their collaborators that limit access to the scientists’ specific workflow. Only the collaboration group should be able to create, modify, and monitor their workflow.” The second recommendation is that the workflow be able to “hold the credentials of one or more of the collaborators to enable the various workflow components to access the necessary resources. The workflow management system will need to protect the credentials while they are in the custody of the system.” These recommendations require the deployment and support of an interdomain, federated, trust-based public key infrastructure.

Other areas of workflow requiring more investigation and research include the monitoring and management of long running codes and jobs, especially across multiple facilities and metafacilities; self-registration and request of resources by applications as a means to better ensure resource availability and handle error recovery; management system seamlessly working across local and global facilities including system, storage, networks, interconnects, cluster area networks, and file systems; use of intelligent and directed prefetch and caching of data at the workflow level; the availability and use of resources based on priority; deterministic and schedulable movement of data; the management and monitoring of the workflow systems management systems themselves

such as workflow level check pointing and restart/recovery; and end-to-end benchmarking (e.g., wall clock time and time to solution). Many efforts are under way at the individual system and metafacility level (e.g., GARA, GRAM, MonALISA, Condor, TeraGrid), which need to be leveraged and extended.

III.7 Security

Security needs to be a major integral component of any petascale facility and metafacility. The NITRD Report identifies security as a strategic priority, citing the need for high-end computing systems to function as “network-centric multi-domain enterprise with ubiquitous secure collaboration.” Security must be extended to the concept of workflow and must be built and integrated into all major system, facility, and metafacility components.

Areas needing to be addressed as part of the management of a petascale facility and metafacility include dynamic configuration of firewalls; location of firewalls as part of an overall architecture; federated trust as part of a multidomain, multisite, multinational metafacility; high-speed monitoring and intrusion detection systems that can differentiate between good and “hostile” agents and bots; scalable security tools and architectures; scalable PKI (with revocation of CERTS); integrated security and system management monitoring at system, facility, metafacility, and workflow levels; integrated and coordinated security, network, and system policies and services; and a coordinated auditing and analysis system that integrates logged data all sources and services involved in a metafacility.

III.8 Visualization and Analytics

Visualization and analytics is a young area, although one that is often included in the term “data-intensive sciences.” The visualization process involves fetching the data (which may come from many sources), mapping it, rendering it, and then displaying it. This process may be iterative as the researcher uses the visualization services to steer the experiment or computation. The research and opportunities associated with visualization include the location and integration of visualization servers or services as part of petascale research (i.e., many times the visualization server needs to be collocated with the capability computer or experiment to ensure the shared use of the file system and low latency for computational steering); remote access; enhanced annotation and indexing systems to support more efficient access and exploration of the data; and the need to have virtual visualization and analytic agents (i.e., virtualize the cognitive processes of viewing data and recognizing patterns) to peruse the vast volumes of scientific data .

III.9 Morphable Networks and Systems

Ian Foster is quoted in *Nature* as noting, “All scientists will be adept at applying existing computational techniques, but they will also understand that progress in their own fields will require innovation in computing technology.” Application scientists, computer scientists, and system administrators will all need the ability to experiment with new

technologies, protocols, and architectures as they develop and evolve their codes for the petafacility of the future. The NLR network is an example of the concurrent support of production and research in a backbone. This network morphability capability needs to be extended into the MANs, LANs, and CANs to support an adaptable network on an end-to-end basis. Current capabilities of nodes to be either I/O nodes or compute nodes might be harnessed to dynamically balance the percentage or number of I/O or compute nodes based on the dynamic requirements of a program during a run. Research into the adaptability and morphability of systems and metasystems needs to be done to better make use of these unique and expensive facilities and capabilities and encourage researchers to test and adopt new technologies. Just as we moved from statically allocated memory partitions and processor allocations to timeshare systems with virtual memory, disks, and processing in the 1970s and 1980s, we now need to move to the next level of virtualization and adaptable petascale systems.

IV. Recommendations

This section is divided into three parts. The first part presents a new program for petascale facility management, highlighting the areas discussed earlier in this report. The second part outlines the initial steps needed to make the program a reality. The third part briefly discusses the role of SciDAC as a model.

IV.1 Integrated Management

DOE/SC/ASCR has a long and successful history of developing leading-edge systems comprising integrated subsystems and components. It is therefore well suited to take the lead in addressing the grand challenge of petascale facility management.

IV.1.1 New Program for Petascale Facility Management

“Management” in this case includes the proactive, collective, and integrated management of objects and resources and the melding of secure allocation, scheduling, and management of resources, along with workflow, data-intensive computing, the instrumentation and monitoring of petascale capability computing, storage, data management, I/O, visualization, and network facilities—all in the support of the predictability, performance, and usability of those resources by the researcher.

A New Program. To address the petascale management situation, ASCR should create a *new hybrid computer science and infrastructure program* that crosscuts all of ASCR’s existing research programs, including SciDAC, and integrates relevant petascale research and infrastructure areas. To this end, ASCR should work with other programs within DOE as well as NITRD agencies. It should also work closely with leading vendors in networking, interconnect technologies, system architectures, storage systems, operating systems, object and data management systems, visualization systems, data-intensive computing systems, security systems and tools, and collaborative middleware tools and systems to develop the next generation of management architectures and tools necessary to support the petascale facility and metafacility.

A New Discipline. DOE/SC/ASCR should develop and support a *computing systems management sciences (CSMS) discipline* that will integrate the architectures, systems, protocols, instrumentation, and tools necessary for the secure integration of observation, experimental, and data-intensive science with high-end computational-based facilities.

A New Set of Integrated Management Systems. DOE should investigate the development of a new set of integrated management systems comprising management components, subcomponents, and interfaces from which virtual management systems can be built. At a minimum this set should focus on networks (interconnects, clusters, LANs, MANs, WANs), I/O, storage and file systems, computer architectures and models (shared memory and message passing), systems and facility integration (including workflow, data management, and metafacility management), security (authentication, access control,

auditing, identity management, firewalls, etc.), instrumentation and monitoring, error detection and recovery (e.g., checkpointing and roll-back capabilities), resource scheduling, and the enforcement of multipolicy management systems.

Project Leverage. This management program should leverage the initial work in this area by projects and programs such as Globus, Condor, MonALISA, CORBA/OMG, and TeraGrid and should focus on the virtualization of these facilities and the workflow and data management systems and processes used as underpinnings for such a venture

IV.1.2 Performance-driven, Predictable, and Usable Resource Management System

A performance-driven, predictable, and usable resource management system is crucial to the future of petascale science. This includes a focus on the appropriate level of instrumentation of facilities, systems, and components; monitoring the facilities, jobs, and data; gathering and disseminating data associated with managing these systems and facilities; matching appropriate codes to architectures; analyzing local and global systems and how their interaction affects workflow and time to solution; and allocating local and global resources securely. Also of major importance is the development of an overall “capability” system that includes leadership capability computing, storage, I/O, and networks.

IV.1.3 Data Management

The management of data, both scientific and systems-related, is a major function of all large-scale computing facilities. A concentrated effort in this area is crucial to the successful deployment and use of the petascale systems. Areas of data management requiring focus are networks (optical, interconnects, hybrids, etc.), I/O, data and information provenance, caching, data placement, storage (hierarchical, distributed, etc.), analysis, visualization; metadata and schemas, object management, and the workflow and dataflow systems associated with the movement and management of data. All of these need to be secure and trusted services. This work should leverage the excellent baseline of capabilities and challenges noted in the 2004 Data Management Report.

IV.1.4 Workflow

The integration of leadership capability computing systems with capability-level experimental facilities for steering and analysis requires the timely synchronization of resources, including the allocation of network and data storage and movement facilities. Similarly, integration of datasets from various scientific disciplines requires the same synchronization of supporting facilities capacity and capability systems. All of these “integrations” of facilities and datasets require a secure workflow and dataflow management system coupled with a multicontext, multisite policy framework, similar to the work started in Globus and MonALISA. A basic component of the workflow system is a set of services to support resource and services request management; reaction and trending capabilities for workflow systems; and the collection, interpretation, and access to the management systems information for better enabling the control plane to do its job.

DOE/SC should increase funding and support of network infrastructure, bandwidth, and capabilities, in addition to data, information, and workflow systems and middleware, in order to properly support the development and use of the distributed petascale-level experiments, capability and capacity computing systems, storage, and visualization.

IV.1.5 Object-based Infrastructure

The best current approaches of cobbling systems and components together will become even more challenging as we continue our move into the petascale level. The number and variety of components and facilities will dramatically increase. This is the time to (re)examine the use of objects as the basic building blocks for the petascale facility and metafacility. By treating ports, processors, storage devices, I/O nodes, files, data, and even packets as objects, we can define and attach policy and security attributes to them to enhance the management, secure access, and use of the resources and data, as well as to enable better data governance and provenance.

IV.1.6 Naming and Addressing

In addition to the need to attach attributes to data, facilities, and components for the appropriate and responsible management of resources and data, it is important to try to reduce the large number of naming and addressing systems used throughout of DOE facilities. Every naming system conversion brings with it an increased chance for error or security breach, as well as confusion for the user. An ideal goal would be to have a *unified naming system* that could be used in networks, storage systems, computing systems and architectures (interconnects, I/O, processors, etc.). The goal should be an object-based framework and system that is conducive to petascale science and yet avoids the overly prescriptive or overly fine-grained use of objects to the point where it impedes performance and usability.

A unified naming system, coupled with an object-based distributed architecture, could truly revolutionize the way in which science is carried out. A higher order of abstraction may increase usability, reuse, services, monitoring, analysis and the effective use of resources. In addition to better enabling the petascale facility and metafacility, this approach, much like http and the Web in the early 1990s, could positively change the nature of the Internet architecture in general. We should be moving, routing, analyzing, computing, and visualizing objects—not packets, bytes, and bits.

IV.1.7 Dynamic and Adaptable Infrastructure

A dynamically adaptable infrastructure will be an important aspect of the petafacility of the future. Not only will it be necessary for addressing the R&D and support for the autonomic error detection and recovery aspect of the facility, but it will also be necessary for matching appropriate resources (e.g., processors or I/O nodes) to the requirements of the codes. In addition, to encourage the development of new petascale codes, the evolution of current terascale codes to petascale and beyond, and the experimentation

with new technologies and architectures by petascale science, researchers must have access to production and experimental networks, storage, and computing facilities. The amount of virtualization and parallelism already evident in systems will increase, as will the need for adaptable infrastructures to utilize and manage them. A MORPHnet-like infrastructure [17] that extends end to end through networks, clusters, storage, I/O, systems, and facilities will be a necessary ingredient for supporting these aggressive goals. The management of a morphable infrastructure is in itself a research project that needs to be explored and undertaken. Investment and use of facilities such as the NLR backbone, which were designed to support this capability, are a good start, but this capability needs to be supported at the sites and end systems, as well as by investments into SC/ASCR network infrastructures such as ESnet and UltrascienceNet. Alternative networks and architectures also need to be investigated; such studies must include best-effort IPs, layer 2 VPN networks, and latency-sensitive and QoS-driven network capabilities.

IV.2 Initial Steps

Efforts are under way in system management, workflow, object management, architecture and system instrumentation, data management, and distributed resource management that can be leveraged as a basis for a petascale management program. An initial step would be to have a blue-ribbon committee provide a survey and status of current systems, tools, and capabilities. Then ASCR should host one or more workshops where the experts from these various fields can be assembled to help identify specific research and infrastructure requirements necessary for the success of such a management program.

An ancillary approach to the support of the management program would be to identify a small number of institutes, preferably multisite, that could focus on the development and support of specific architectures and analysis tools relevant to the goal of managing the petascale facility and metafacility. One example is an institute focused on the benchmarking, analysis, and potential service definitions of I/O and interconnect technologies on an end-to-end basis. Another example is an institute devoted to the development of the metadata, monitoring, and tracking techniques and tools required to manage a secure end-to-end petascale network transfer from storage device to storage device at two sites; this would need to include the characterization and instrumentation of the components, services facilities, and workflow systems to enable their coupling to support the virtual metafacility. NLANR and CAIDA are examples of such institutes that have been successful in the past.

IV.3 SciDAC Petascale Projects as Enabling Prototypes

The DOE SciDAC program has been very successful in bringing together teams of scientists and computer scientists to focus on overall infrastructures and collaborative tools in support of specific science projects. The SciDAC model should be extended to this program of petascale management and in fact is a natural evolution of the multidisciplinary work and teams initiated by SciDAC. DOE should chose three to four

flagship SciDAC-2 projects that are planning to use petascale facilities, datasets, infrastructure, and computing and then create R&D teams made up of scientists, architects, and experts in file system, I/O, security, data management, and networks and focus on them on a specific scientific petascale challenge. These people would be bound together as a dedicated team for the life of the SciDAC project (i.e., three to five years) and would work on specific but complementary technologies and architectures necessary for the success of a SciDAC project. The chosen SciDAC projects, infrastructure, and tools should complement each other and could be combined such that they could be used to build a complete secure virtual management system that could be used by the SciDAC researchers.

Appendix AI. Petascale Computing and Facilities

AI.1 Leadership Capability Systems

The federal HEC Plan defines leadership and production systems as follows:

- **Leadership-class Systems:** the leading-edge high capability computers that will enable breakthrough science and engineering results for a select subset of challenging computational problems. These are problems that have been unsolvable with currently available computing resources.
- **Production Systems:** computers that address the challenging computational problems that require high-end computational resources but do not require access to the extraordinary leadership-class systems.

Leadership-class systems, also known as “capability systems,” are currently mainly terascale level, with movement into petascale within a couple of years. They usually support a smaller number of users than capacity systems. “Production systems,” also known as “capacity systems,” refer to today’s class of cluster and parallel systems, with the top end of these computers sometimes referred to as high-end capacity systems or “near-capability” systems. It is safe to assume that leadership-class machines will be at least in the petascale range for the next five to ten years

Petascale computing involves more than just the petaflops one uses on a leadership-class capability computing system (LCCC). These machines consume and generate large volumes of data (i.e., petabytes) as part of the program’s and project’s normal dataflow and workflow processes. In the future the data will scale to exabytes and more. Some capability computing users develop and debug their codes first on local site compute clusters before they port their codes to an LCCC system, so as to reduce the time of trial and error on such an expensive resource. Subsequently, all LCCCs need to be an integral component of a petascale metafacility, which itself contains a continuum of computing power and would include not only the LCCC and other computational resources but also the data and file systems as well as the networks necessary to enable the migration between the class of facilities more easy and error free. Figure AI.1.1 is a graphic depiction of a continuum of computing capabilities as noted in the NERSC 2006-2010 plan. Where certain systems fall on such as continuum is not as important as the recognition of the existence of a continuum and its value to science. The metafacility needs to include the appropriate supporting security, collaborative tools, middleware, and other relevant services (e.g., a unified name service) to allow the researchers to collaborate and share information.

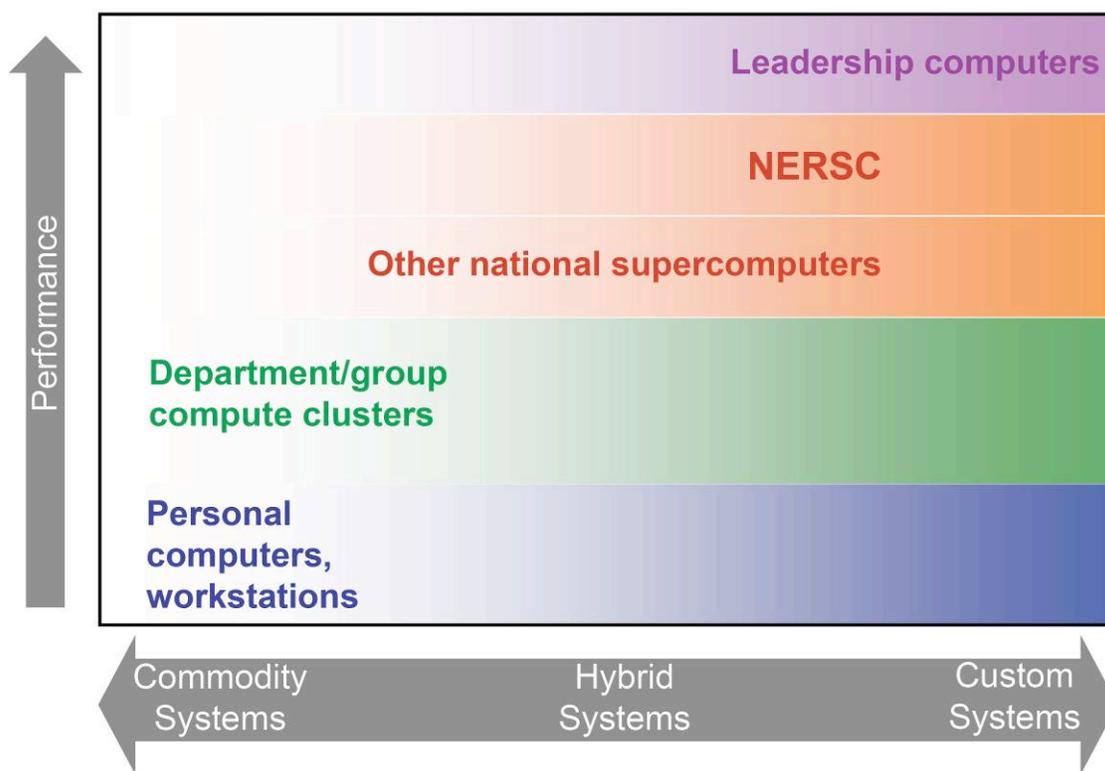


Figure AI.1.1: The continuum of scientific computing systems.

In summary, in order to get appropriate applications to run on these petascale, researchers need an appropriate environment—namely, a “less than LCCC” production system—on which to define, develop, and debug their programs, which can then be migrated to the LCCC system. If a problem is encountered, the researcher can then go back to the production system(s). Also, in order to encourage the next generation of petabyte applications to be developed for the next generation of petabyte computers, researchers need to be able to experiment with currently available or soon to be available LCCC systems. Moreover, researchers need the ability to experiment with new technologies and various combinations of petascale infrastructure capability systems such as networks, visualization, non-human-based analytics, autonomics, storage, and I/O.

The MORPHnet concept, introduced in 1997, specifically argued for these types of experimental computing environments. The National Lambda Rail is a more recent attempt at providing this type of capability at the WAN level. It is important that the MORPHnet concept be extended to include the LAN and local resources as well as support the concept of the dynamic facility where researchers can test various types of production and experimental next-generation hardware and software as they probe the limits of their codes and how new technologies or different architectures (e.g., different ratio and number of I/O nodes or compute nodes coupled with various SAN, CAN, LAN, and WAN technologies) either enhance or impede their codes and science.

AI.2 Petascale-class Experimental Facilities

Today's experiments are mainly terascale, but petascale experiments are anticipated in the near term future. Some experiments are already generating petabytes of data. One such example is the TeraScale Supernova Initiative (TSI), which does three-dimensional simulations of stellar explosions. TSI is currently producing data at the staggering rate of 5 TB per day, and the data produced is expected to rise in the next few years to hundreds of terabytes per simulation.

Experiments that are expected to break into the petascale level with respect to data management, or that have already done so, include the following:

- CEBAF at JLab, RHIC at BNL, and the Tevatron at Fermilab, which have all broken the petascale level for data.
- The Large Hadron Collider (LHC) at CERN [18], complemented by matter-antimatter “factories” at the Stanford Linear Accelerator Center and KEK [19] and the current world's highest energy collider at the Fermi National Accelerator Laboratory. Collisions detected at the LHC will have a raw information content of close to a petabyte per second. Even though much of this data will not be retained, it will still culminate in petabyte datasets.
- The High Energy Physics International Linear Collider, which is in design now and will both complement and work with the LHC with respect to research on the Higgs boson and “map the unified ‘electroweak force.’”
- The Spallation Neutron Source (SNS), a new nanoscience facility at Oak Ridge National Laboratory, which will provide orders of magnitude more neutron flux and larger detector arrays than predecessor facilities, with concomitant increase in data volume. At full capacity, SNS expects to have 24 instruments and to generate petabytes of data a year. It will require coupling with petascale computing and petascale visualization for real-time steering of experiments.
- Next-generation confocal microscopes for support of biological research (e.g., proteomics at Pacific Northwest National Laboratory), which will generate petabytes of data a year. Biological research is undergoing a transformation from a qualitative, descriptive science to a quantitative, predictive science as a result of the availability of high-throughput, data-intensive “omics” technologies such as genomics, transcriptomics, proteomics, and metabolomics, together with the advance of high-performance computing.
- Climate research, which is generating (by both measurements and model simulations) datasets that range in size from a few megabytes to tens of terabytes. Examples include raw measurements from satellite instruments, data from in situ observation networks such as the DOE Atmospheric Radiation Measurement program sites, and the output of three-dimensional global coupled climate models such as the Community Climate System Model, which produces 7.5 TB of data in a single 100-year integration.
- Combustion research, which is just beginning to simulate laboratory-scale turbulent flames using massively parallel computers combined with emerging models and codes. Current computations generate about 3 TB of raw data per simulation, posing

new data storage and movement challenges and requiring a new paradigm for data analysis. Adaptive steering and subsetting of data as it is computed are needed in order to enhance discovery and to enable further analysis and visualization of events whose occurrence was not known a priori.

- Nuclear energy experiments and data management processes, which can peak at data generation rates of tens of megabytes per second, with the major programs generating on order of one petabyte per year. The data analysis environments include tens to hundreds of scientists simultaneously accessing refined datasets of tens of terabytes.
- Fusion research, which is developing increasingly realistic simulations that will result in large and diverse data that in turn will demand powerful data-management frameworks and tools. In particular, ITER's plasma production effort, planned to be operational around 2015, will generate an enormous amount of data, which will need to be collaboratively analyzed and managed during experimental operations by a worldwide scientific community.

AI.3 Capacity Computing

Capacity computing can range from desktop computers to local site clusters to mission-dedicated clusters such as those used at SLAC and Jefferson Lab and can range from the low-teraflops class to the leadership capability computing level. These capacity computing facilities are often used as pre- or postprocessing engines, as well as debugging and development resources, for leadership computing as well as to support data-intensive computing and leadership-class experiments. All of these environments generate or analyze petabytes of data and thus are included in the DOE/SC continuum of petascale facilities. (Because of their very nature as capacity systems, these facilities are often acquired with more of a focus on memory and processing issues than with the aim of building a leadership capability facility.)

AI. 4 Petascale Storage and Data Management

As we move into petascale computing and science, we need to address the fundamental challenge of moving data to and from leadership-class computers, as well as the associated general data management and provenance issues. This challenge will be especially acute in the I/O and storage systems where systems are already challenged to keep pace with the raw computational capabilities that exist today. The HEC Task Force notes in its plan that "one of the neglected areas of modern high-end computing is the ability to stream computed data to storage systems, i.e., parallel I/O."

Sites such as NERSC that once were more of a source for data are now large sinks for data, and the future explosion of DOE petascale facilities will generate the need for even more storage systems and facilities. Petascale storage and data management systems are currently in the range of hundreds of terabytes to petabytes and can be found at almost all of the DOE/SC labs

AI.5 Petascale Networks

Petascale networks are currently gigabit-scale networks (10s to 100s of gigabits per second) and will eventually become terabit-scale networks in order to support the petascale facilities. The networks will need to support architectures and protocols that in turn support large file and dataset transfers. The Large Scale Network (LSN) subgroup notes as a goal in the NITRD report of 2007 ([11], page 12) the need to “enable near-real-time petabyte and above data transfers, by 2008, to support science cooperation and modeling.” The subgroup also notes the need for Infiniband and single-stream flows to be supported over WANs, in addition to the “need to develop protocols to move massive amounts of data.” These goals are all focused mainly on the ability to move large datasets and files in support of science. This means that petascale capability networks will need to support a wide range of services including best effort IP (BEIP), high-speed TCP implementations, layers 2 and 3 virtual networks, tunneled cluster and interconnect protocols over the WAN (e.g., Infiniband, Ethernet, and fiber channel), lambda switching, and eventually optical packet switching. The National Lambda Rail (NLR) was the first large-scale, nationwide, multiple 10-gigabit wavelength multilayer network. The NLR, Internet 2, ESnet, and the regional networks need to scale up appropriately to support petascale facilities of the future. Petascale networking needs to be supported on an end-to-end basis in the WAN, MAN, and LAN. An example of relevant work at the site and LAN level is Fermilab’s Lambdastation. ESnet’s planned dual mode network to support large data transfers is depicted in Fig. AI.5.1.

ESnet New Architecture to Address Science Requirements: IP Core+Science Data Network Core+Metro Area Rings

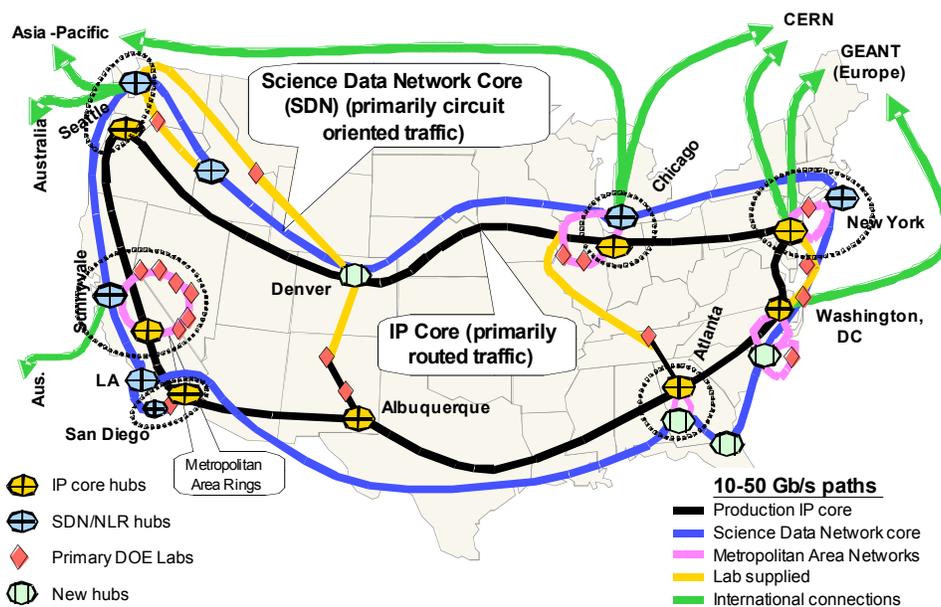


Figure AI.5.1: ESnet’s dual-mode network

AI. 6 Petascale Analysis and Visualization Facilities

Table AI.6.1, taken from the NERSC 2006-2010 Plan, identifies the data-intensive sciences. All of these science programs will be working with petascale datasets and will augment and support both petaflops computing and petascale experiments. The majority of the visualization servers are terascale systems and are predominantly distributed shared-memory architectures. These systems are often closely located to the petascale computing and experimental facilities to enable real time simulation as an integral component of experimental and computational steering processes.

Table AI.6.1: Applications and Algorithms Matrix

| Science Areas | Multi-physics & Multi-scale | Dense Linear Algebra | Sparse Linear Algebra | FFTs | AMR | Data intensive |
|---------------|-----------------------------|----------------------|-----------------------|------|-----|----------------|
| Nanoscience | X | X | X | X | | |
| Climate | X | | | X | X | |
| Chemistry | X | X | X | X | | |
| Fusion | X | X | X | | X | X |
| Combustion | X | | X | | X | X |
| Astrophysics | X | X | X | X | X | X |
| Biology | X | X | | | | X |
| Nuclear | | X | X | | | X |

AI.7 Petascale Facility and Metafacility

Dr. Orbach notes on page 8 of the SciDAC Review [20], “One remarkable example of ‘growth in unexpected directions’ has been in high-end computation. This is now one of the most important facilities.” For the purposes of this report it is important to realize that a petascale facility can be any of the individual systems or petascale facilities; but more important is the fact that these facilities can be combined in many combinations to make a virtual metafacility. Some of those resources may be located locally and on site, while others can be located remotely on other continents. The concept of metasystems was alluded to by Dr. Michael Strayer (then director of SciDAC and associate director for DOE/SC’s Advanced Scientific Computing Research and now director of ASCR) when he noted in the SciDAC Review [20], “In computational science, as in other sciences, the area of international collaboration is the forum for science for the 21st century. The SC is heavily committed to multiple international collaborations – including the International Thermonuclear Energy Reactor, the International Linear Collider, and the Large Hadron Collider, to mention a few. SciDAC, together with other ASCR facilities, could provide powerful resources and the nexus for a new global village for computing that could take computational science and scientific discovery to wholly new peaks.”

Appendix AII. Trends in Scientific Research

AII.1 One-of-a-Kind Facilities

The cost and complexity of one-of-a-kind experimental facilities have already generally resulted in there really being only a single such facility per scientific focus. The LHC at CERN, SNS at ORNL, and ITER at Cadarache, France, are just a few examples. The same trend is occurring in capability-scale high-end computing. DOE/SC is investing in two leadership-class capability systems that are intended to evolve to petascale-level computing facilities. One will be an IBM Blue Gene and the other a CRAY XT3. Each of these will have unique architectures. The cost and power requirements associated with running a capability machine will keep the numbers low. The future may well bring with it high-end capability machines focused solely on a particular class of problems, for example, Monte Carlo codes. All of these unique petascale experiments and computers will require petascale networks, I/O systems, and storage systems.

In order to support these capability experimental and computational facilities, there has been a recent trend to move back to a centralized system architecture, where the majority, or all, of the resources are located at one site. Access to cheaper energy and affordable networks may also affect the future location of capability petascale computing and experiments. A centralized system will need excellent data management and analysis, networking, and collaboration infrastructure to support the secure access and participation by international researchers. Specifically, the community will require not only enhanced tools and capabilities to manage the individual petascale systems but also the ability to handle metafacility and workflow processes associated with remote and internationally distributed use of these facilities. The provenance and management of the data associated with all of these petascale facilities will tax any current and planned infrastructures and architectures.

Another feature associated with one-of-a-kind facilities is that since they are truly one of a kind, the analysis, instrumentation, and management of the systems and their individual components are very challenging. There is no large base on which to draw for expertise for the support of these systems as there is for commodity and off-the-shelf systems. DOE/SC and the high-end computing community need to develop their own teams and tools for properly supporting and managing these facilities. The intellectual capital and resources necessary for managing and supporting these types of facilities should not be underestimated.

Another relevant trend that affects the support of these unique facilities is that the science community is expanding the number of collaborations in which it engages. The number of researchers collaborating on any given research project continues to increase. Many research papers now use two pages to list the co-researchers on a project. The roles of the researchers are also evolving. Often a researcher is a principal researcher on one project, a collaborator on a second project, and a consultant or minor contributor on many other

projects. Hence, not only does the infrastructure need to efficiently support the larger number of collaborations, access, and use of these facilities, but it also needs to support the many different roles of the researcher with respect to access, use, control, and management of the facilities. This last challenge is one that is especially relevant to open science facilities (i.e., DOE/SC facilities), as opposed to the NNSA, which supports a reduced and more controlled set of researchers using their facilities.

The challenges associated with all of these trends are the need to address the dynamic creation and management of virtual organizations, role-based security, data management, and enhanced facility, metafacility, and workflow instrumentation and management.

AII.2 Computational Science and Simulation

As Dr.Orbach noted, we are already in a new era with respect to the use of high-end computers to perform science-based simulations that could not have been previously done. Simulation and computational science continue to gain respect and are now being coupled with experiments for steering purposes. The Data Management Report [12] notes the value of increased use of simulations to steer a variety of science projects:

Sometimes the data to be analyzed is streamed directly from an instrument or simulation (e.g., when monitoring an observation or experiment in progress). By analyzing the data as it is being generated, scientists can detect anomalies and error conditions and can steer the experiments or simulations to focus on interesting events. If the anomaly is known, a signature-based method can be used. Alternatively, the “normal” data can be modeled and deviations from that model can be flagged. Research is needed to expand existing capabilities in real-time algorithms, approximate algorithms, robust sampling techniques, and time-constrained queries, in order to handle massive and complex data.

The climate modeling community is heavily involved in the development of coupled simulations. In addition to the current coupled atmosphere-land-ocean-sea ice simulations of the physical climate system, future Earth system models will incorporate a true carbon cycle component, which has models of biological processes, such as land vegetation and ocean microorganisms. Model-data fusion, whereby simulations are coupled with satellite and surface-based multi-instrument data streams, are part of the research regimen. Satellite instruments scheduled for deployment will monitor a wider range of geophysical variables at higher resolutions, which will be used to evaluate climate model simulation suitability for a wide range of targeted research and practical applications.

The Data Management Report [12] identifies three categories as the central workflow components of simulation-driven science: data movement and reorganization, data analysis, and visualization. All involve data-management challenges. The report also notes that simulation scientists desire a change from the current batch mode to interactive capabilities to better enable data management, visualization, and analysis.

The interesting challenges that arise from these trends will be the need to focus on the integration of these facilities, coupled with either petascale capability computers or petascale experiments, into a metafacility. Of specific interest will be the use of simulation for the quasi-real-time (e.g., in minutes or tens of minutes) steering of experiments. Figure AII.2.1 shows the SNS architecture, which integrates and incorporates simulation on capability systems with visualization and petascale experiments.

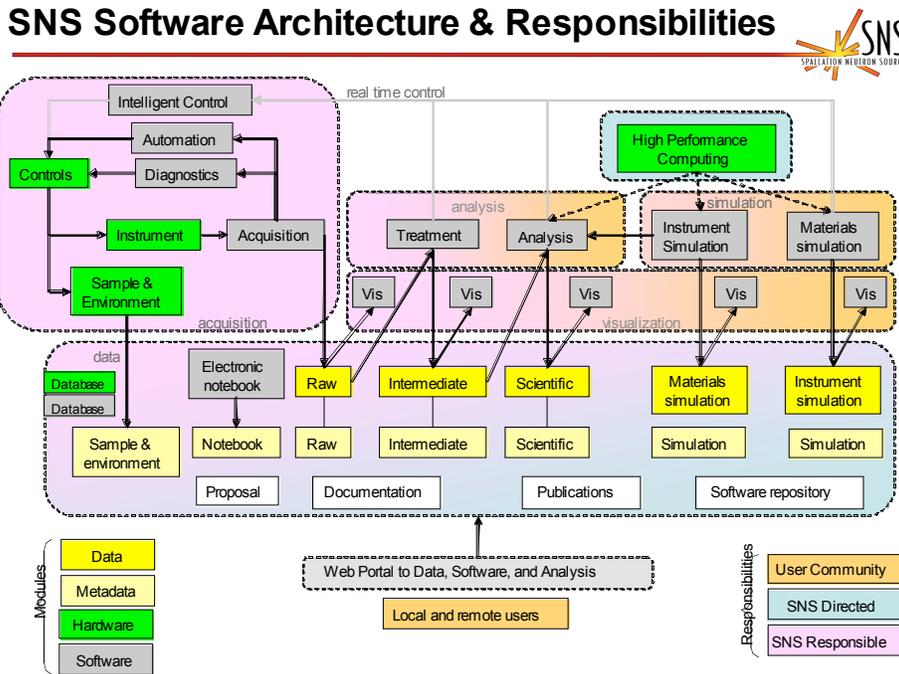


Figure AII.2.1: SNS software architecture (courtesy of ORNL)

AII.3 Data Fusion and Integration

In addition to the integration of simulations, computation, and experiments, there is a trend toward more interdomain science collaboration, specifically with respect to the integrated use of datasets from multiple science domains. The Data Management Report [12] summarizes the general process of data integration:

Data integration requires resolving the differences and inconsistencies in the data management systems (e.g., different vendors), in the data models (e.g., relational, network, ER, object-oriented), in the query and data manipulation languages, in the data types (e.g., text, graphics, multimedia, hypermedia), in the format (e.g., structured, semi-structured, specialized formats), and in the semantics. The ability to manipulate data requires both a characterization of the internal structural and semantic properties and a characterization of the relationship of the dataset to the

associated material. Moreover, to achieve the vision of data integration, one must have the technology to describe the data models, data structure, data format, and data semantics.

Data integration also encompasses the integration of data types, formats, and systems. The Data Management Report [12] notes that “data sources are physically distributed and heterogeneous in how information is stored, organized and managed. Second, they reside on heterogeneous hardware platforms with diverse software interfaces. Third, the data is of different types (e.g., text, video, images, audio) and formats (e.g., netCDF, HDF, SILO) as well as dynamically changing in both content and form.” The report also notes that the goal of a data integration system is “to provide users with a uniform interface to access, relate, and combine data stored in multiple, geographically distributed, and possibly heterogeneous information sources. It enables users to focus on specifying what they want, rather than thinking about how to obtain the answers.”

The obvious challenge with respect to these trends, and which go beyond the normal coupling of data storage and movement systems into a metafacility or virtual facility, is the need to address the various schemas, ontologies, and taxonomies associated with the various science domain datasets and to enable an inter-working set without imposing one standard..

III.4 Integration of Scientific Databases and Data Management

The evolution of petascale science will incorporate new trends of data integration and data intensive computing. Advances in the science domain will no longer be just derived from simulations alone or from either experimental or observational sources. Scientific exploration is now utilizing datasets from multiple science domains as well as relying solely on data generated or reduced as a result of other simulations and data-intensive computing endeavors, that is, simulations are now using the results of other simulations as source data.

Richard Mount of SLAC has noted that within the HEP community pure computational high-end computing is evolving to a combination of computation with data-intensive high-end computing (i.e., using data generated from other sources). The biology and chemistry science domains and data are being combined as part of the research in bio-fuels. There are efforts to combine climate simulations and observational data as part of the overall research in climate modeling. SNS nanoscience is coupling work with biology datasets. Chemistry is coupling quantum chemistry and molecular dynamics simulations in chemistry, combustion, geochemistry, biochemistry, and environmental studies.

III.4.1 Data Location

Data management encompasses multiple facets including data placement, location, replication, tracking, caching, file and storage systems, architectures, transfer, metadata, and provenance of both the data and processes associated with data management. Other attributes associated with data management include short- vs long-term properties,

hierarchies of caching (e.g., NERSC writes caches to disk immediately), and location of the data with respect to local, remote, or distributed storage, even including transient storage in the network. Of particular interest will be the need to support long-term archives and efficient access to them.

AII.4.2 Data Searching and Analysis

A major challenge facing petascale science is dealing with the vast volumes of data generated by petascale facilities. New and enhanced techniques and tools are required for enhancing the scientists' ability to annotate and attach attributes and metadata as the data is being generated and moved, so that appropriate sets of the data can be located and processed later. DOE/SC-funded projects such as FASTBIT [21] and ScalaBLAST [22] support indexing, searching, and analysis techniques aimed at large scientific datasets, but more much needs to be done. The trend toward more intelligent and cognitive techniques for analyzing datasets will continue as the volume of data outstrips our ability to effectively utilize humans and human visualization techniques to visualize and analyze the data. The user interface of choice for the management and control of data currently is the Web. As the number of scientific datasets generated increases and as the demand for access to those datasets increases, DOE/SC and its community will be faced with more and greater challenges with respect to data management and provenance. Of particular interest will be generating consensus on a limited yet workable set of schemas for all of the science domains and enhanced analytical tools including visualization and virtual visualization.

AII.4.3 Data Provenance

A good definition of data provenance is found in the Data Management Report [13].

Provenance information is meta data that describes the logical organization of data in terms of its origins, including the original conditions under which an ancestor dataset was produced, the sequence of transformations applied to produce the derived data, and the people and software involved in performing these transformations. Provenance includes description at the level of science (dataset A is the Fourier transform of dataset B) and engineering (the transform was done with version 2.3 of software package X on a specific compute resource). Provenance meta data, particularly engineering-level information, is most easily collected directly from applications and workflow systems and can be used to create new, related workflows, for example by using the provenance of one analysis pipeline to instantiate a parameterized analysis of additional datasets.

An example of a workflow management system that captures provenance information for planning and execution in Grids is the Pegasus [23] stem, which was developed as part of the GriPhyN (Grid Physics Network [24]) project. The Pegasus system takes a high-level definition of a desired workflow and schedules the tasks in the workflow on available resources, based on the requirements of the tasks and the availability of resources in the Grid. Pegasus tracks the original abstract workflow, the input files, and the output files

that are generated as products of the workflow execution. Pegasus highlights the fact that datasets are not the only entities that will require metadata descriptions; it will be necessary to describe hardware and software tools with information about their provenance and the data they are capable of processing.

Another example is the MyGrid system [25], which enables a biologist to dynamically compose workflows and quickly discover sequences of interest among the thousands returned by curated databases for an investigation.

II.4.4 Policy

Another important area of data management relates to who controls the data, has access to it, and can manipulate it, namely, who manages the policy of the data regardless of location. This will become even more of an issue as the datasets become larger and are managed at remote sites and data storage sinks, and specifically includes the issue of who owns the intellectual property rights, who can reference the data (and how and when), and who controls the placement and state of the data. An ancillary challenge will be the integration and bridging of local/site and global/meta-facility policies in the secure management of the data, that is, the combined policies associated with data, security, and systems. In addition, the current trend is to have visualization and analysis servers closely collocate with petascale computers and experiments so that they can effectively share the same file and storage system. Effective and secure remote access to these visualization servers is imperative for their effective use and will require addressing multisite and multinational security and data policies.

II.4.5 Object-based File Systems

High energy physics at SLAC, biology at PNNL, and other sciences have recently experimented with the use of object based file systems. Some high-energy physicists found the Lustre file system could not effectively handle large numbers of small files, and hence the scientists developed their own minimalist object-based file system, called “ROOT” [26]. PNNL and others continue to use Lustre. The use of object-based databases in the science arena will continue to grow, and if an object-based database or file system is developed that addresses the range of science requirements with respect to access and transactions, its use could become predominant. As an example, some experiments generate data segments or objects on the order of 100 bytes. An object-based database that would support these types of datasets along with their relevant operations would greatly enhance the data management prowess of the DOE/SC sciences. A related challenge will be the development of an efficient and effective metafacility-wide object-based file system.

II. 5 Interdisciplinary Teams

A major goal of the interagency High Performance Computing and Communications and the Next Generation Internet programs in the 1990s was to encourage interdisciplinary teams to tackle the grand challenges of science. SciDAC took that concept and made it a

reality through its programs and projects by combining domain scientists, computational scientists, and a variety of computer scientists into project focused teams.

A current trend in science is the combination of scientists from various science domains, such as biology and chemistry or biology and climatology. As the Data Management Report [12] notes, “Many simulation scientists collaborate in small groups in most stages of the scientific process. Increasingly, however, scientifically important problems require large, multidisciplinary teams. In these instances, the need to access distributed data and resources is the rule rather than the exception. Scientific discovery requires that we ultimately create distributed environments that not only facilitate access to data but also actively foster collaboration between geographically distributed researchers.” The challenge, then, is to ensure that these scientists have the appropriate supporting infrastructures, architectures, and tools so that they can work on an interdisciplinary basis.

AII.6 Virtualization

The number of international scientists working together on interdisciplinary teams is increasing. As a result, virtual organizations (VOs) are created to provide policy management and oversight as well as support collaboration on major projects. A VO is often associated with some set of facilities and a project. People no longer give a second thought to collaborating with colleagues around the world on either an extemporaneous or long term ongoing basis. The VOs themselves follow normal socioanthropological groupings into teams and some form of bureaucracy, each with its own set of dynamics, to support the research project and its subcomponents. A side effect of this trend is the need to match the social organizations and research processes with the placement of and access to resources, as well as the requirement for virtual collaboration middleware, tools, and infrastructure as part of an integrated architecture. Therefore, the architecture and placement of petascale facility and metafacility resources will need to address these distributed and remote requirements. A 2005 OECD workshop report [27] on Grids makes specific note of the challenges associated with the virtualization of organizations and resources as part of the evolution of science inquiry. The workshop report also notes that the resources that make up a Grid are dynamic. This same characteristic will apply to petascale facilities and metafacilities.

Another dimension of the virtual organization will be the move toward more use of codes, data, and services from other science domains. The Service Oriented Architecture (SOA), incorporated in Globus and other Grid architectures, offers users virtual codes and systems. Commercial vendors are starting to deploy and support virtualized SOA-based virtual systems. One example is HP's SOFTUDC concept. It is a virtual system that will virtualize not only the data center, blades, and servers but also the entire system, namely, virtual Ethernet based VPNs and operating systems. When utilizing SOA services, users will need the capability not only to request but also to validate the integrity of the codes and services they intend to use, as well as the ability to monitor and track usage of those services for accounting purposes.

At the technical level, virtualization is even more widespread. We have had virtual memory, with all of its layers of prefetch and caching techniques, and virtual operating systems for many years. The many processors on a massively parallel machine act like one virtual system. We have virtual disks and tape drives, virtual networks (tunneled and VPNs), virtual routers (similar to virtual operating systems but focused on routing), virtual switches (“n” switches act as one), virtual firewalls (i.e., software based), GridFTP (a virtual FTP where multiple parallel FTPs are used but the user sees only one), and now the advent of service-oriented architectures, Grids, and metafacilities where many individual facilities, resources, and systems are combined into a virtual system. In addition, next-generation science will depend on virtual researchers, that is, teams of researchers, acting as one, will attempt to address the tsunami of scientific data needing to be analyzed and managed via the use of agents, “bots,” and services to search, locate, analyze, visualize, and manage data and information.

The management of virtual resources presents numerous challenges. These include the verification of codes and services, cross-domain security, and trust, as well as the instrumentation, monitoring, and management of both the virtual and real resources. All of these are important issues that need to be addressed. For example, a user may have a different virtual machine configuration every time the user runs a code on a leadership-class computer, that is, a different number of processors or ratio of I/O nodes and compute nodes, as well as potentially a different operating system loaded with each run and even virtual network configurations and file systems. These often transient virtual environments and states will require management tools to monitor and analyze the use of the resources as well as supporting the configuration of the system either before or during run time.

The Globus Toolkit, one of the most widely known and implemented set of distributed systems, support software, and services, has been under development for many years and has been working to address a wide range of management, allocation, and monitoring services for the support of virtual organizations and the use of distributed resources. Globus provides a workspace management service (Fig. AII.6.1) comprising a broad set of service implementations and infrastructure services.

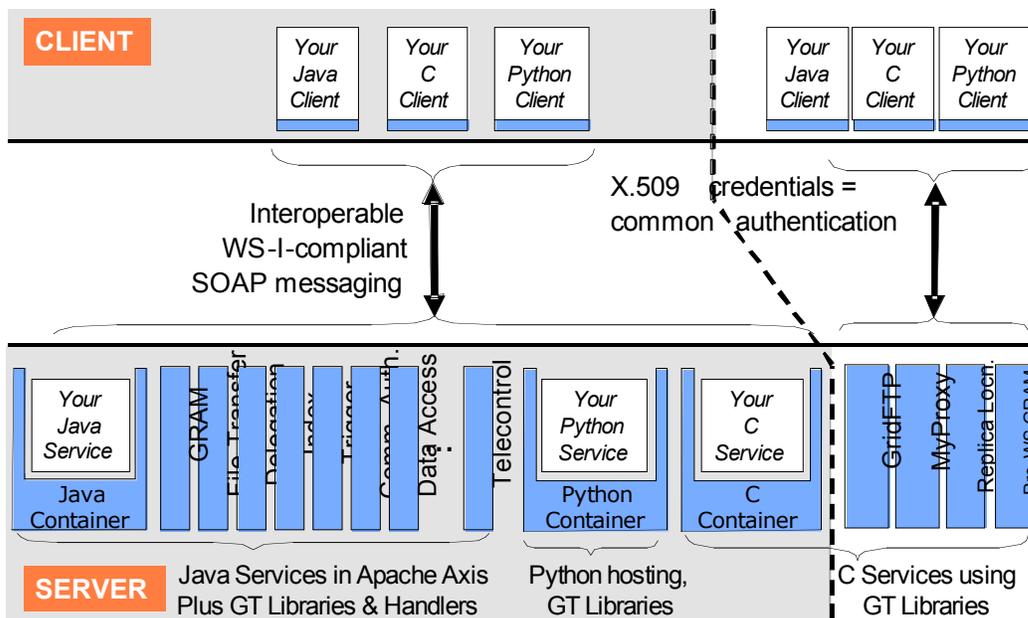


Figure AII.6.1: Selected GT4 components and interactions. Shaded boxes are GT4 code and white boxes user code.

The fourth version of the Globus Toolkit, GT4 [28] (Fig. AII.6.2), specifically provides standards-based services and mechanisms, as well as associated XML-based resource properties and attributes with resources, as part of an overall monitoring system that provides for two aggregator services for collecting information as well as capabilities to push (subscription) or pull (query) relevant data.

AII.7 Visualization

The Data Management Report [13] describes scientific visualization as follows:

...the transformation of abstract data into images that are more readily comprehensible than the data itself. It is the primary means by which scientists “see” their data, and it forms a central part of most, if not all, scientific processes. Visualization techniques can be applied to raw data as well as to the results of analyses; and, to a large degree, the challenges raised by increased data size and complexity for visualization mirror those of analysis tools. New visualization techniques are needed to make the patterns in large, complex datasets stand out. Visualization algorithms applied to raw data and, increasingly, to large subsets derived by analyses will need to shift from serial to parallel designs, including parallel transfer of data to large displays.

Visualization is coming of age as part of the scientific inquiry and analysis process. The use of Powerwalls, CAVEs, ImmersaDesks, and powerful desktop graphical engines are used to generate visual representation of complex, multidimensional and multimodal data. Visual comparative analysis techniques and interactive data exploration are

common tools now for scientists. Analytical tools and techniques such as FASTBIT [21] and ScalaBLAST [22] are used for indexing and searching data, which is then often graphically displayed. Researchers use visual comparative analysis and interactive visual data exploration techniques as part of their analytical process. As the scale of data increases, we will need even better visualization and analytical tools.

Two major challenges face the increased incorporation of visualization as part of the scientific process. First and foremost is the fact that the amount of data is outpacing scientists' ability to view the results. Projects will not be able to afford to hire hundreds to thousands of students and postdocs to view the graphical output associated with these petascale systems; and even if they could, coordinating hundreds of people to visually analyze graphical output doesn't scale. Petascale science requires other than "human only" visualization tools, that is, nonhuman visual cognitive visualization and analytical agents and bots to aid in the analysis, identification, selection, and reduction of the data, which then can be viewed by humans as part of the process. This is a data-intensive process. Many DOE/SC-supported labs are pursuing the research and development of new visualization techniques and approaches. As an example, PNNL has developed innovative visualization tools [29] such as Starlight and InSpire to analyze many types of scientific data and processes. The development of visualization centers (e.g., PNNL's National Visualization and Analytics Center, <http://nvac.pnl.gov/>) is also beginning to emerge.

The second challenge is also one of management: the management of the visualization process and service. Integrated data analysis and visualization environments data and process models are "coupled with integrating programming and graphical interfaces, to simplify common tasks, automate the mechanics of using advanced data management technologies and enable reuse of analysis, visualization, and other technologies" [12]. Visualization is as much a service as it is a system and includes information and data analytics (i.e., computing), graphical hardware and software, networks, and access to file and storage systems. A visualization service/server needs to be integrated into an overall petascale system management system so that it can be used in the real time process of steering of petascale computing and experiment steering as well as for post computation or observation analysis.

AII.8 Workflow

The remote use of one-of-a-kind facilities and the increased use of virtual metafacilities that combine multiple resources and facilities, both computing and experimental, from multiple sites requires the movement of large amounts of data as well as the need to manage the processes associated with that data movement and processing. The researchers want a performance-based, predictable, and usable facility and metafacility. They want to be able to specify when and where a program will run, ensure that it doesn't fail, have various levels of checkpointing and recovery, have the data available at the right place and time for the analysis or computation, and schedule any other relevant resources such as networks and storage. This desire fuels the push for faster and better technologies and systems. The scientific community has recently shown an increased

interest in workflow and data management as it pertains to their use and management of scientific resources. The first international workshop on workflow in e-science was held in May 2006 and the second in May 2007 [30] (see Figure AII.8.1 for a graphical depiction of a workflow process).

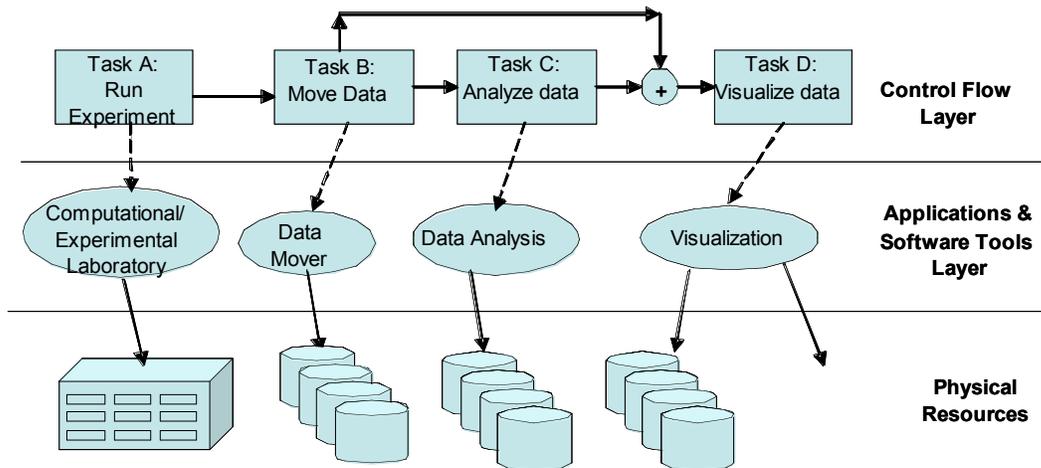


Figure AII.8.1: Example of a workflow created in the scientific investigation process, showing the three layers: control flow, applications and software tools, and physical computer hardware (courtesy of the Data Management Report [12])

Workflow in some sense is no more than the time honored tradition of job control as developed in the 1970s with respect to the timeshared use and management of computing resources (i.e., processing, memory, and storage). There are many challenging facets to the workflow process, the secure movement of data (i.e., networks and I/O), the integration of numerous files systems, the intersection of naming systems, and multiple site and interdomain resource management, policies, and security. Many of these issues have been identified by Globus, Condor, and other Grid systems; and initial work has started in these areas. Another additional aspect of the workflow process that needs to be addressed is the need for data, schema, and workflow transformations and peering. In addition, the workflow process requires its own metadata and services to describe, monitor, audit, and manage its processes and flow. An example is the need for metadata for workflow components and to both capture and represent the relationships between the data, systems components, system status, time, and other relevant variables.

There currently exist more than a few workflow-oriented systems whose focus is to address either a component of the overall workflow process, for example, the movement of data, or the total workflow process. Python is a popular scripting language, famous for its concise syntax and clear semantics, which many programmers use to control workflow. The SDM is focused on the management and movement of data. The European Data Tag program developed the Grid Laboratory Unified Environment (GLUE [31]) schema provides not only service definition schemas and templates but also templates for data and resources such as compute nodes, storage, and end nodes. Globus,

Condor, and MonALISA are focused on the management and allocation of distributed resources.

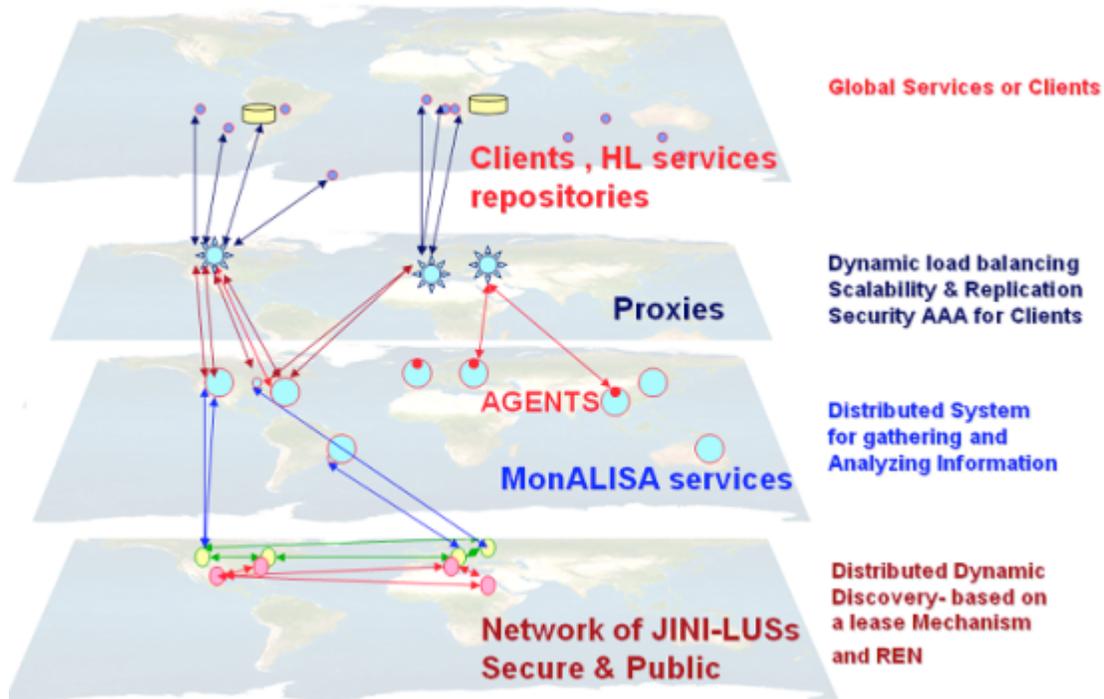


Figure AII.8.2: The MonALISA architecture

MonALISA (Figure AII.8.2) is an agent-based workflow system used by the international and distributed HEP community for support of its scientific collaborations. We describe the MonALISA system as just one example of a workflow system that incorporates many essential scientific workflow architectural capabilities. MonALISA, which stands for Monitoring Agents using a Large Integrated Services Architecture, has been developed over the past four years by Caltech and its partners with the support of the U.S. CMS software and computing program. The framework is based on Dynamic Distributed Service Architecture and is able to provide complete monitoring, control and global optimization services for complex systems. The MonALISA framework is a fully distributed service system, that is, distributed registration and discovery for services and applications, with no single point of failure, and it provides for the monitoring of all aspects of complex systems. For example, it monitors system information for computer nodes and clusters, network information (traffic, flows, connectivity, topology) for WAN and LAN, monitoring the performance of applications, jobs or services, and end user systems, as well as end-to-end performance measurements. It can interact with any other service(s) to provide in near-real-time customized information based on monitoring information, and it provides for the secure, remote administration for services and applications. Agents are used to supervise applications, to restart or reconfigure them, and to notify other services when certain conditions are detected. The agent system can be used to develop higher-level decision services, implemented as a distributed network of communicating agents, to perform global optimization tasks. MonALISA supports

graphical user interfaces to visualize complex information and global monitoring repositories for distributed virtual organizations.

A major challenge exists with respect to workflow and resource management systems, namely, the need to leverage the initial work in the numerous existing workflow systems in order to develop an integrated workflow framework complete with associated metadata, schemas and processes that can be used as a common set of standards-based services and enable an object-based infrastructure from which an integrated monitoring, security, and management system can be built to support the various types of architectures and facilities.

AI.9 Dataflow and Data Management

Not only is science moving into petascale computing and experiments, it is also moving into data-intensive computing where the data may be analyzed, moved, reduced, transformed, and visualized more than once. The amount of data being generated and analyzed only continues to increase. As a result, more attention is being placed on dataflow and data management. Data management is an important component of the workflow process and is focused on the placement, searching, movement, and management of data. Two important aspects of data management that many are trying to address are its temporal nature, namely, how long it is kept, and its access patterns, both of which can affect what type of device the data should be stored on. For example, NERSC automatically copies data off fast disk to slow disk and tape for integrity purposes, while SLAC is experimenting on RAM disks as part of an overall architecture to enhance access. The management of data also requires the ability to appropriately describe the data and its relationship to other data and processes with the right amount of granularity and relevant information. Another important architectural issue with respect to data management is the separation of metadata from data for easier traversal, search, control, and management of the workflow and data management processes.

The Data Management Report notes that metadata descriptions are used “to discover, interpret, evaluate, and transform the data. This additional description is often referred to simply as ‘metadata,’ and managing such information is considered ‘semantic engineering,’ or ‘knowledge engineering.’” The report (page 47) also points out that “tools such as problem solving environments, portals, and electronic notebooks can also document aspects of workflow, but they are more directly involved in the logical organization of data into project and experiment hierarchies. These tools can also be used to support a wide range of structured and unstructured annotations, such as a similarity between a gene in one organism and one in another, information about a detected feature, reviews of data and assertions about data quality, or simply some text about an idea for a new experiment triggered by current work.” The report cites as an example the SAM-based Electronic Laboratory Notebook, which allows text, drawings, images, equations, and arbitrary files to be associated with data and organized into electronic chapters and pages. These tools and capabilities need to be an integral component of science workflow systems.

The Data Management Report (page 49) points out that effective indexing schemes and techniques are crucial to enhancing data analysis. It notes that “the traditional indexing techniques, such as B-trees and hashing, are inefficient for datasets with a large number of searchable attributes. Even multidimensional indexing techniques, such as R-trees, are efficient only for datasets with no more than 10 or 15 attributes. If there are more attributes or if the user query involves only a small number of the indexed attributes, a brute-force scan is more efficient than these indexing schemes.” Research and development of better indexing techniques is currently being undertaken at LBL (FASTBITS), PNNL (ScalaBLAST), and other centers.

One aspect of data management that can be very important to the effectiveness of any facility, metafacility, and workflow systems is data placement. The Data Management Report page 55) describes data placement as consisting of two elements: “selection of an available storage location capable of holding the data ... and the actual transfer of the data into this location.” The report notes that “regardless of whether it is an end user, an application, a middleware component, or a low-level system function, the entity that triggers a data placement request must be able to influence when, where, and for how long the data should be stored. The decision can be based either on the properties of the target storage unit (e.g., proximity to the current data location, reliability, throughput) or on a set of goals or intentions (e.g., keeping the checkpoint data until the end of the simulation run). Providing data placement services requires an appropriately layered design.” An example of the importance of placement is found in the TSI experiment

Replication of data is another aspect of data management that is very relevant to sciences and the distribution and access to data generated by petascale facilities. The Data Management Report (page 56) notes that “replica management and cache management are closely related, but replica management focuses on the particular issues that arise in the management of geographically distributed copies of datasets. In geographically distributed computing environments, computational tasks may be performed at locations that are far away from necessary datasets.” This issue will become even more pertinent with the current trends of petascale science integrating datasets from multiple science disciplines. The report defines replication as follows:

Replication involves creating multiple copies of identical files or portions of files in order to increase data locality and fault tolerance and to reduce the latency of data access in a wide-area, distributed computing environment. Traditional replica management for transactional database management systems keeps track of table updates and synchronizes the changes among the database replicas. In scientific applications, most datasets are read-only after they are published, and data access is predominantly file-based; these characteristics simplify replica management because update synchronization is not needed.

Data replication, and even the caching of scientific data, can have a positive effect on an overall system, depending on the architecture. The HEP LHC program uses a tiered replication and storage system for distributing LHC data on an international basis. The data is replaced at the replication sites. The location of the sites is important, and these

sites will require much more networking bandwidth and storage capabilities as a result. The size of the datasets being moved is normally very large and requires the use of parallel transfers to accomplish the movement. The GridFTP protocol and Reliable File Transfer Service are two well-known services employed to accomplish these transfers.

The Data Management Report notes seven issues that need to be addressed (page 56):

(1) specifying the source files to be copied and registered; (2) specifying the target directory or locations for the data; (3) specifying the catalogs in which new replicas should be registered; (4) coordinating copy and registration operations; (5) identifying and recovering from failures; (6) considering the state of resources, including network performance, existing replica locations, and the availability and performance of storage systems and computational resources; and (7) considering policy issues, including security and resource management policies that define which groups and applications have permission to access particular datasets, storage systems, and computational resources and what priorities are assigned to different requests.

The Scientific Data Management Center has developed the Storage Resource Management System (SRM [15]) to handle these functions (see Figure AII.9.1 for a description of the SciDAC-funded SRM system). SRM uses Global Grid Forum standards, such as GridFTP, but also provides front-end services to other files systems such as HPSS, and manages both local and global resources. Fermilab uses SRM as a front-end and management system for its local disks systems.

Storage Resource Managers: Essential Components for the Grid

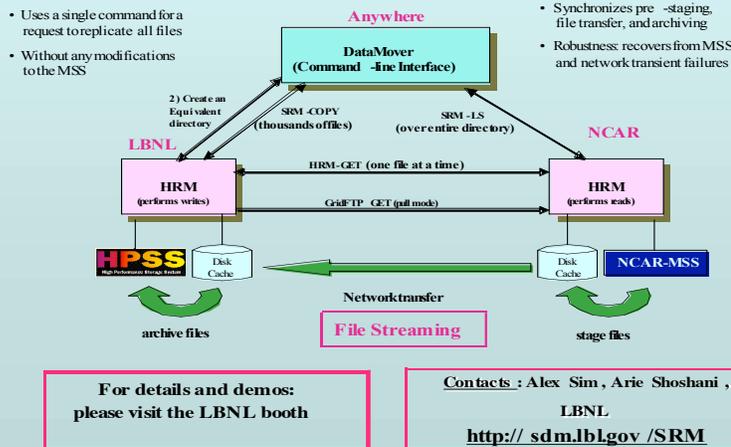


SRM Functionality

- Accepts multi-file request, manages a queue, allocates space
- Invokes GridFTP to get/put files from/to remote sites
- Pins files for a limited lifetime, accepts early release of files
- Queues staging and archiving to MSSs
- Provides automatic garbage collection
- Monitors and recovers from failures of file staging, archiving, and transfer
- SRMs can be invoked by clients, other middleware, and other SRMs

SRMs provide robust multi-file replication

Replicating Thousands of Files Robustly between
Two different remote Mass Storage Systems
used in production at ESG and PPDG -STAR



-2-

Figure AII.9.1: Storage Resource Management (courtesy of the SciDAC SDM Center)

AII.10 Persistence, Ubiquitous Computing, Nomadicity, and Remote Access

Scientific work environments, both technical and social, have evolved to the point that when scientists are on business travel or are telecommuting from their homes, they expect to have access to some portion of their work environment and data. Unified messaging systems, VOIP, IP-based PDAs, VPNs, WI-FI, and other technologies allow the scientists to stay connected, including online real-time tracking of experiments as well as personal and collaboration or team communications. Nomadic access, ubiquitous computing, and PDAs need to be securely treated as an integral component of any petascale facility infrastructure and incorporated as part of the overall facility and metafacility from both a security and management perspective. An evolution of persistent presence will see the use of intelligent agents and bots regardless of whether the scientist is really on line at that time. Researchers will be able to use these agents, services, and bots to monitor and manage their research and computation in the background while they attend to other tasks.

Ubiquitous computing and nanotechnology are just emerging, and their total impact on the scientific process is yet to be seen; but they have the potential for providing additional

capabilities for monitoring and managing codes, components, systems, services, and facilities, as well as enabling a more secure and smart facility and metafacility.

AII.11 Summary of Trends in Scientific Research

All of these trends will expand the use of distributed petascale resources, one-of-a-kind facilities, and leadership-class computers by an even more distributed set of international researchers and virtual organizations. All of this, combined with the virtualization occurring at the technical level, will stress the ability to manage from an allocation and operational perspective petascale computing, resources, and networks, which need to be managed as an integrated system.

APPENDIX AIII. DOE Facilities

This appendix provides a brief overview of current and planned DOE/SC computing facilities. The purpose of the appendix is to highlight the breadth and depth of computational and experimental capabilities located at DOE/SC laboratories and the value of using the full continuum of these capabilities with the leadership-class capability computers and leading-edge capability experiments into a DOE/SC petascale metafacility. The broad spectrum of facilities supported by DOE/SC, coupled with the imminent data tsunami, further underscores the need to support a petascale facility and metafacility management program with a strong focus on security, data, resource, workflow, and operational management

AIII.1 SciDAC

Although SciDAC is not a facility per se, SciDAC has engendered the creation and use of many virtual facilities because it integrates facilities, science, and researchers. As a result it has become an important cornerstone for DOE/SC science. SciDAC is mainly a collaborative multidisciplinary approach to very challenging science problems. As Dr. Michael Strayer notes on page 62 in the SciDAC Report, “One of the key successes of the SciDAC program has been its ability to integrate diverse interdisciplinary groups that are focused on scientific discovery. The individual investigator approach has evolved into a tripartite partnership between discipline scientists, applied mathematicians, and computer scientists.” Many DOE/SC scientists note that the way they now pursue their scientific research has forever been changed in a positive fashion because of the team-based collaborative approach enabled and supported by the SciDAC program. This new modus operandi, which builds teams of scientists and uses distributed facilities and resources, will be the norm for future scientific research and subsequently be dependent on the combined and integrated use of petascale facilities.

AIII.2 NITRD and HEP Roadmap

The federal plan for HEP advanced the concept of leadership high-end computing systems, also known as “capability” systems. DOE/SC is currently deploying and upgrading leadership systems at both Argonne and ORNL, as well as aggressively upgrading the near-capability capacity production system at NERSC. In addition to how SciDAC has changed the way DOE/SC researchers pursue science, it is as equally important to note the role that the full continuum of computation, network, and storage facilities plays at the various laboratories and universities as part of the overall petascale metafacility to support DOE/SC science

The NITRD FY2007 Report (page 2) states that as part of the FY2007 goal and focus of the joint agency program on petascale leadership-class capability computing, ORNL will upgrade its leadership-class facility to over 250 TF, Argonne will acquire a 100TF IBM Blue Gene/P system, and NERSC will acquire the 100-150 TF NERSC-5 as its next

generation of advanced capacity systems. The 2005 NITRD report (page 11) sets a long-term goal for HEC capability systems in the range of 10 to 100 sustained petflops. The systems are to be adaptable and self-tuning.

As part of the supporting infrastructure for leadership capability and capacity systems, DOE/SC has defined its networking goals in the NITRD 2005 report (page 48) to include the ability to transfer 3 petabytes per year by 2009-2010, move 1 petabyte in 24 hours using secure file movement over a 160 to 200 Gbps best effort service, support the cell biology community with shared immersive environments that include multicast and latency QoS guarantees using 2.4 Gbps links with strong QoS guarantees, and support real-time exploration of remote datasets using secure remote connectivity over 1 to 10 Gbps links to the desktop with modest QoS services.

AIII.3 Current and Planned Leadership-class Capability, Capacity, and Cluster Systems

Dedicated and specific mission-focused systems are supported at Argonne's Jazz machine, the SLAC and FERMI HEP systems, and the QCD resources at Jefferson Lab, PNNL, and BNL. BNL also supports RHIC. General facilities and systems that support multiple programs and projects are at Argonne, NERSC, and ORNL. Below is a brief summary of the computational facilities at various DOE/SC labs. We note that the Argonne and ORNL systems are leadership-class capability facilities and subsequently will be expected to serve a small and selective number of petascale science problems or projects; however, a large amount of processing and analysis capabilities is required for other research, as well as a component of petascale metafacilities.

AIII.3.1 Capability Facilities

Oak Ridge National Laboratory currently has the following resources:

- Cray X1E Vector system with 1,024 processors, 18.5 TF peak, 2 TB shared memory, and 45 TB of disk storage;
- Leadership-class Cray XT3 Opteron-based 25 TF peak, 5,294 processors with 2 GB per processor (10.5 TB total memory) and 120 TB disk storage;
- General storage consisting of 4 HPSS STK silos (STK 9840 and 9940 tape drives) with approximately 4 PB of tape and 10 TB disk cache as part of HPSS that front ends the silos;
- Opteron-based visualization cluster with 64 dual-processor nodes with quadrics interconnect and NVideo graphics cards

In the near term (mid-2006) ORNL plans to have a Cray XT3 Opteron system evolving to dual core processors upgrading to 50 TF and 21 TB RAM, and by November 2006 to expand the system to 100+ TF with 46 TB RAM and 900 TB disk. By December 2007 the system will be upgraded to 250+ TF with 92 TB RAM. December 2008 will bring a 1 PF machine to ORNL. Visualization services will be accomplished on the original 56 cabinets.

Argonne National Laboratory currently has the following:

- Leadership-class IBM BG/L with 1,024 processors, 5.7 TF peak, .5 TB RAM, and 14 TB parallel file system as an evaluation system
- Jazz, a 350-node, 1.6 TF, 1.05 TB RAM system used by Argonne scientists and engineers
- Pentium-3 Cluster (“Chiba City”) with 256 nodes and 512 Pentium 3 500 MHz processors for a total of 256 GF

The primary visualization resource is an NSF-supported visualization cluster of 96 nodes/192 XEON processors at 2.4 GHz with 384 GB RAM.

Argonne plans to have a leadership-class 100 TF IBM BG/P by late 2007.

AIII.3.2 Capacity Facilities

DOE/SC has a few “just less than capability” capacity facilities that are important components of the petascale computing portfolio.

NERSC

NERSC current has a 9TF SP3, a 3TF NCS-A Infiniband cluster, and a 7 TF NCS-B capacity system. The NERSC-5 is planned for operation in 2009 and will initially have 35 TF (grow to 100-150 TF) with 5,000 processors, 1000 to 2000 nodes, 25 TB of memory, 350 TB of disks [32]. NERSC will keep its current NERSC-3 IBM SP with 6000 processors, 7.8 TB memory, and 48 TB disks. NERSC specific networking capabilities are currently projected to be 40 Gb by 2008 and 100 Gb by 2010. Current NERSC visualization services include a SGI ONYX 3400 with 12 processors, 24 GB memory, and 5 TB of disk storage. NERSC plans to upgrade in 2009 to a visualization processing server with 100+ processors, 2 TB of memory, and 50 TB of disk storage.

SLAC

SLAC supports mission-dedicated systems to focus on data-intensive batch processing. SLAC’s cluster has 1700 nodes, 4 GigaSpecInt2000s (approximately 9.73 TF), 1 GB RAM per core, and a 275-node file server with 670 TB capacity, and a HPSS mass storage (Storage Tech Tape system) with the capacity for 6 PB (2.5 PB stored now). SLAC also supports MPI clusters with 256 processors, usually arranged in 4-64 node clusters with 0.25 TF per 64-node cluster. Miscellaneous disk storage is about 200 TB. Visualization services are supported through an SGI Altix with 72 processors and a combined total of 0.43 TF and 192 GB RAM. SLAC also supports the PetaCache RAM disk project prototype with 64 nodes and 1 TB RAM.

PNNL

The Molecular Science Computing Facility (MSCF) within the Environmental Molecular Science Laboratory (EMSL) at PNNL has an HP cluster of 1960 Itanium 2 processors

with 6.8 TB of memory a peak performance of 11.8 TF and 45 TB of local scratch disk space.

BNL, Fermilab, and Thomas Jefferson Lab (JLab)

The BNL, Fermilab, and JLab facilities are focused on the support of QCD computational science.

BNL has a 9.6 TF system physically divided into four machines of 3.2, 3.2, and 1.6, 1.6 TF. This is a custom machine, similar to the Blue Gene/L but with single core 12000 custom processors based on PowerPC, 400 MHz. Both BNL and Fermilab are primary participants in the HEP LHC program.

Fermilab has a 4 TF system with approximately 1000 dual core 2 GHz Opteron processors each with 1/2 GB/core. By the fall of 2006 Fermilab will also have a 3 TF Xeon system with 512 nodes with 1 GB per node. By March 2007 JLab will have a 5 TF dual core (Opteron/Xeon to be determined). Fermilab currently has more than 3,000 dual-processor computers in its data-processing farms and analysis clusters, more than 4 PB of scientific data in its storage systems, and more than 100 TB of disk storage. Fermilab will add more than 500 additional nodes over in 2007 to support LHC and QCD requirements. To support the data distribution needs of the supported research communities, Fermilab currently has two 10Gbs WAN connections in production use and will have eight in use by the end of 2006. Many of the existing Fermilab experiment facilities are now interfaced to FermiGrid, which provides an organization wide Campus Grid infrastructure supporting the sharing of resources, policy-driven scheduling of compute jobs and processors, and common monitoring and management services.

JLab currently has a 3.3 TF Pentium-D, 280 processor, 3 GHz, 1/2 GB per core system, 2.1 TF Xeon, 384 processor with 3.0 GHz nodes and 1/2 GB per node, and a 1.3 TF Xeon system with 256 processors at 2.66 GHz with 1/4 GB per node. Overall storage currently associated with these systems is approximately 80GB.

III.4 Storage and Data Management

This section highlights a few petascale storage facilities not addressed in the prior section. In 2003 NERSC turned a corner when it started to receive more data than it exported. Figure III.4.1 shows the trends of NERSC's petabyte storage facilities and the projection of archival capacity being outpaced by the data stored in 2010. Table III.4.1 shows the associated network and I/O requirements to handle the expected movement of the data to and from NERSC's storage and archival systems. By 2010 NERSC expects to move 117 TB of their 39 PB system, an effort that will require a network that can move 1356 Mbs, which is 10.85 Gbs, and subsequently needs a minimum of 40 Gbs of network capability.

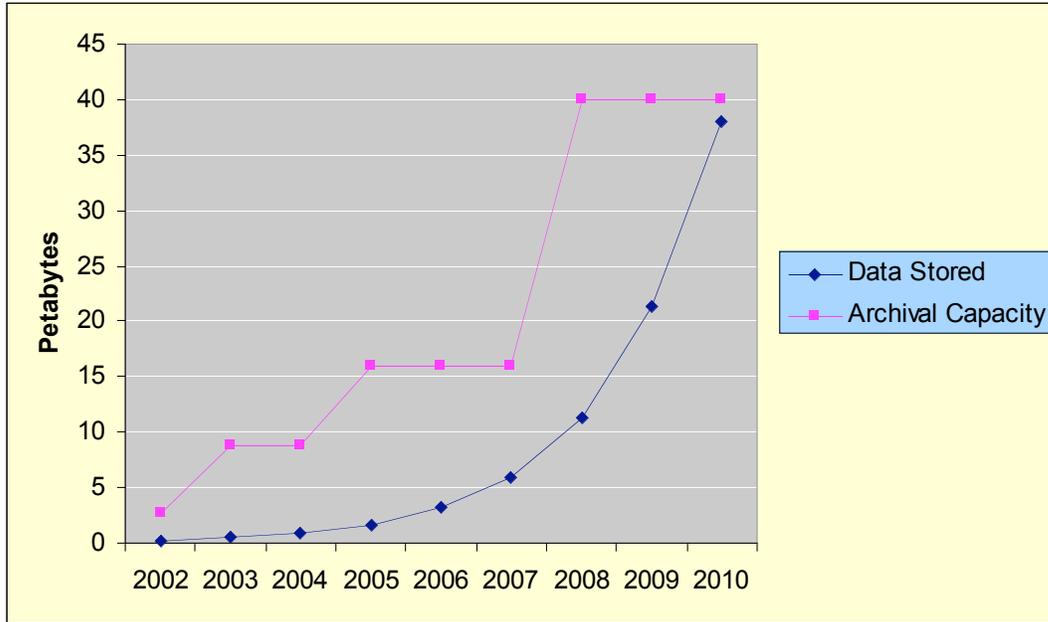


Figure III.4.1: Projected growth of stored data and archival capacity at NERSC (page 27 of the NERSC 2006-2010 report)

Table III.4.1 : NERSC Projected Data Growth and Bandwidth Requirements, 2005–2010

| Year | Total Archived Data | Data Transfers per Day | Transfer Rate |
|------|---------------------|------------------------|---------------|
| 2005 | 1.5 PB | 6 TB | 60 MB/s |
| 2006 | 2.9 PB | 10 TB | 120 MB/s |
| 2007 | 5.0 PB | 20 TB | 231 MB/s |
| 2008 | 11.0 PB | 33 TB | 392 MB/s |
| 2009 | 21.0 PB | 64 TB | 749 MB/s |
| 2010 | 38.0 PB | 117 TB | 1356 MB/s |

These projections are very important, given that many other sites also act as storage sinks and database sites. Moreover, new data sites are expected that will be required to support LHC, SNS, ITER, climate, biology, and other petascale experiments.

The architecture of the storage, database, and file systems is a crucial facet of a petascale storage facility. The LHC project is replicating and distributing its data to Tier 1 data sites (Fermilab and BNL in the United States), which in turn will distribute the data to Tier 2 sites, and so on. This is a distributed and replicated data management and storage system.

The TSI researchers also addressed the storage placement and management challenge by replicating data at multiple sites. The TSI researchers, as noted on page 32 of the SciDAC Report devised ways to transmit and store terabytes of data among multiple national sites

to better enhance access to the data. The report notes that “the TSI team worked with computer scientists at the Logistical Computing and Internetworking Laboratory, UTK, and used their Logistical Runtime System to move data on parallel streams, across multiple Internet paths [33]. This tactic increased data transfer rates by 10-20 times. Moreover, data are now stored on a community Linux cluster at NCSU which yields interactive access to all team members. The team also worked with networking researchers at ORNL to move data from the ORNL Cray X1E to NCSU, using the bearer channel protocol.”

The climate modeling researchers have multiple databases located at different international sites. LLNL currently hosts data output from 16 different international climate modeling groups which are archived at that single location, while grid based technologies are used to combine archives from three sites where the CSM model is run into a single distributed database. Current plans are to create a multi-site, multi model international model output database over the next few years.

Another integral and important component of storage facilities is the file system or database architecture. HPSS, PVFS, and other systems are widely used today. Will they be able to scale to effectively handle the amount of data generated by the petascale facilities? NERSC has in production the NERSC Global Filesystem (NGF), a high-performance parallel global file system that is accessible from all of the multivendor computing systems in use by NERSC. NERSC is also using IBM’s GPFS as the basis for its NGF and plans to integrate it with HPSS in 2007. Argonne and many other sites are using the Parallel Virtual System (PVFS). Lustre, an object-based file system, was used by the HEP community but has largely been replaced by the object-based ROOT system. However, Lustre is being successfully used at PNNL and LLNL. Both have expressed interest in seeing Lustre evolve to handle the challenges of petascale science data management. The LLNL Lustre file system currently supports one petabyte of data.

III.5 Petascale Experiments

Petascale experiments are briefly listed below, not only because many of these are being integrated with HEC and thus are part of the capability-based virtual petascale metafacility, but also they attest to the large amount of data that each facility is expected to generate and that will need to be moved, stored, and analyzed. The figures quoted below are a combination of information derived from discussions with scientists, as well as information from the Science-Driven Network Requirements for ESnet report by Eli Dart of ESnet [34] and the Data Management Report.

Astronomy, Astrophysics, and Cosmology

The Large Synoptic Survey Telescope (LSST) obtains exposures of the entire night sky every 2–3 days to search for transient objects such as supernovae. It is expected to generate 20 TB of data per night. If one assumes at least 100 such nights, this results in 2 PB a year.

High Energy Physics

The Large Hadron Collider is a one-of-a-kind facility focused on research of the basic objects of the universe. It will generate a raw data rate of a petabyte per second, much of which is to be reduced on the fly; however, it is still expected to distribute tens of petabytes of data per experiment approximately four times a year, for a yearly total of about 40 PB. As noted earlier, the data will be replicated at Tier 1 sites, which then provide access to Tier 2 sites, and so forth. LHC expects to be moving and storing thousands of petabytes within five years. Both BNL and FNAL are expecting to have 30-40 Gbs network capability by then to help support this data movement. The LHC is complemented by matter-antimatter factories at SLAC and KEK and the world's highest-energy collider at Fermilab. Fermilab is developing a distributed LHC control room capability to support in situ experimental control and steering from a few controlled international sites.

Fusion

Magnetic fusion experiments operate in a pulsed mode producing plasmas of up to 10 seconds in duration and acquiring around 2–3 GB per pulse every 10 to 20 minutes. Decisions to change the next pulse are informed by data analysis conducted within the roughly 20 minutes between pulses. This mode of operation requires rapid data analysis that can be assimilated in near-real-time by a geographically dispersed research team. Assuming that all the data needs to be moved to remote analysis centers, this translates into an approximate 16–24 Gbs network capacity. In the 5+ year timeframe, there will be hundreds of terabytes of simulation data in addition to the data generated by the International Thermonuclear Experimental Reactor, a burning plasma experiment that is expected to produce data on the order of petabytes per year.

Macromolecular Crystallography

Macromolecular crystallography will generate tens of GB per experiment, with an expectation of approximately 10 experiments per day, for a total of 1 TB per day, or about 300–360 TB per year. BNL estimates it will need to transfer a 1 TB file per day.

Spallation Neutron Source

The SNS at full capacity expects to have 24 instruments operational and will support 12 experiments per day, about 200 days per year, and will generate an average of 160 GB of data per day for a total of 32,000 GB per year, or 32 TB per year. Real-time data mapping will require 2Gbs as part of a distributed computer and experiment network

Nuclear Physics

The Relativistic Heavy Ion Collider (RHIC) experiment currently generates 365 TB per year of STAR data and 400 TB per year of PHENIX data for a total of 765 TB per year. RHIC researchers predict their networking requirements to be on the order of 12 Gbs to support the movement of data. By 2011–2012, the STAR experiment will generate 2,610 TB per year (2.61 PB) of data, and the PHENIX data will be on the order of 1,599 TB per year (1.5 PB/yr) for a total of 4,110 TB per year (4.11 PB/yr). The scale of the data handling issues is characterized by experiments having peak data generation rates of tens of megabytes per second, the major programs generating on the order of 1 PB per year,

with data analysis environments supporting tens to hundreds of scientists simultaneously accessing refined datasets of tens of terabytes.

Climatology

The Data Management Report (page 10) notes that “the datasets generated by both measurements and model simulations for analysis by climate researchers range in size from a few megabytes to tens of terabytes. Examples include raw measurements from satellite instruments, data from in situ observation networks such as the DOE Atmospheric Radiation Measurement (ARM) program sites, and the output of three-dimensional global coupled climate models such as the Community Climate System Model (CCSM).” Currently a single climate model of a 100-year integration generates approximately 7.5 TB of data. The Earth System Grid integrates supercomputers with large-scale data and analysis servers located at numerous national labs and research centers to create a powerful environment for next-generation climate research and for the dissemination of model data and scientific analysis. Climate modeling will generate 400 TB of data (0.4 PB) by 2007. The data repositories at NCAR have 180 TB, at NERSC 76 TB, and at ORNL 75 TB. By 2011 there will be at least a petabyte of data per year at NCAR alone and after that 5–10 PB.

Chemistry

Chemistry simulations and computations currently generate about 3 TB of raw data, with a near-term expectation of datasets of 1 to 30 TB being generated twice a year. These simulations, which layers data, models, and both simulation and analysis tools, will require a petaflop-scale machine over a one- to two-month period. In 2011, 3D simulations will generate datasets in the 100 TB to 1 PB range twice a year. After 2011, multidisciplinary simulations will be generating hundreds of terabytes to petabyte datasets as well.

Biology

PNNL will be using remote steering of its new confocal microscopes. There will be two sensors per microscope, with an expected 10 microscopes operational by 2010. Each sensor generates 39 MB/s. Therefore with two sensors it will generate 78 MB/s (624 Mb/s) per microscope. With 10 microscopes there will be approximately .78 TB/s (6.24 Tb/s) of high-resolution video generated. The PNNL proteomics mass spectrometers currently generate 50 TB of data. The Genomes to Life program will start sometime after 2012: the genomics GTL program will be able to analyze hundreds of samples per day and therefore generate petabytes of data per year within the next decade.

AIII.6 Networks

DOE/SC network facilities currently comprise the 10 Gbs LAN best effort IP ESnet, which also supports 10 Gbs MANs in four metropolitan areas, and the multi-10 Gbs wave UltraScienceNet network, which is experimenting with lambda switching.

UltraScienceNet uses multiple 10 Gbs lambdas from ORNL to Chicago, Seattle, and other peering sites. Plans are under way for an addition to the ESnet network, namely, the Science Data Network (SDN), which is expected to become operational by 2007. The SDN will be a layer 2 VPN service layered over 10 Gbs waves acquired from other R&E sources. The initial implementation of this network will be dedicated to support the movement of HEP large datasets required by the LHC project. The main reason for the second network is to provide a predictable, scalable cost, and performance-based network environment for the “big science” community for large file and dataset transfers. The SDN network will provide for multiple 10 Gbs waves to support Fermilab’s and BNL’s role as Tier 1 LHC data centers and to support large data transfers between them and both the LHC and Tier 2 sites. As other experimental and computational petascale facilities (e.g., SNS) start generating the large amounts of data they are projecting, it will be necessary to provide for a robust hybrid network and architecture that can concurrently support best effort IP and VPN services at layer 3, layer 2 VPN services, and multiples of waves or lambdas with the future ability to switch and allocate the latter based on policy and priority. Initial R&D on this front is being performed on GMPLS, UltrascienceNet wave allocation, and OSCARS dynamically provisioned virtual circuits, to address some of these network management and control issues; the latter is projected to use GMPLS in the future. However, much more R&D and financial support needs to be invested to develop a viable multimode petascale network to effectively support these expanded requirements.

AIII.7 Infrastructure

DOE facilities are currently used either strictly for production or for experimentation and research. This strategy has been the norm for the past 10–15 years and in some sense has hindered the R&D of new network and network-based distributed systems research necessary to support next-generation petascale metafacilities. DOE scientists require production-quality infrastructure in order to perform their research. However, if we expect to have new petascale applications and evolve current terascale science to petascale science, DOE/SC will need to develop and integrate new technologies into the distributed petascale programming environment. In order to do this, applications, researchers, and systems (OS, storage, network) will need to develop and experiment with new technologies. The mere integration of multiple production technologies into a complex system often turns the metafacility into an experimental infrastructure at the system level until the kinks are worked out of the system. A morphable infrastructure that extends beyond the network backbone into the end systems will be necessary for enabling the next generation of SciDAC teams (domain scientists as well as storage, OS, network, file and programming environment researchers) to push into the petascale level of science, and beyond.

AIII.8 Grids

Grids are virtual facilities that support distributed research. One example of a Grid being used by the DOE/OSC community is the Open Science Grid [35]. The OSG Consortium has grown out of a collaboration between a SciDAC-1 project working with peer NSF-supported projects such as iVDGL and GriPHyN and is specifically aimed at working with the stakeholder application and computer science communities. The OSG is a distributed facility with sites spanning the United States. It operates as a loosely coupled yet coherent virtual facility. The individual facilities, including the DOE HEP laboratories and DOE- and NSF-sponsored university sites, maintain autonomy over the use of their local resources while providing remote access to shared storage and processing for all the communities participating in the OSG Consortium. Europe's DataGrid and the U.S. TeraGrid are additional examples of the use of Grids to support distributed scientific research.

APPENDIX IV. Trends in Technology

This appendix highlights some of the top technology trends that are directly relevant to the development and support of infrastructures to support petascale science. It is not meant to be an exclusive or inclusive list; rather, the trends are noted in order to further highlight the increasing challenges associated with managing petascale facilities and metafacilities. This appendix borrows generously from the NERSC and Data Management reports with respect to processor and storage hardware growth trends. The NITRD reports were used to document the interagency projections associated with HEC.

AIV.1 Technology Churn

The turnover rate for new technologies is moving at a rapid pace with many sites planning on a three-year turnover rate for capacity and desktop computing environments. Networking technology churn tends toward a three- to five-year turnover rate. Even leadership-class computers are overtaken by new architectures and technologies at a fast rate. Rarely do the top supercomputers remain at the top of the 500 list for more than one or two years. The technology churn rate is a very important variable because it means that there will always be a continuum of older to newer computational, storage, and networking facilities and technologies represented in the DOE/SC petascale computing continuum. Even more important is the fact that this creates a big challenge with respect to the ongoing need for developing the appropriate monitoring, debugging, engineering, analysis, and management tools within a timeframe that will support these technologies, especially for HEC and leadership-class capability systems.

Another challenge faces providers and supporters of HEC infrastructure and facilities. Many large vendors of computing and networking equipment are very conservative with respect to how quickly they will develop, adopt, ship, and support the newest technologies. They often leave that work to smaller, innovative, and more nimble startup companies and then acquire those companies or technologies after the technology has been proven and there is a very visible large market share associated with that technology. This situation often means that new leadership technologies and architectures are not developed by larger vendors. Convergence in the markets via acquisitions and mergers also affects the availability of choice with respect to new and advanced computing and networking technologies. Venture capitalists have also become conservative, and as a result there are fewer new innovative startups. The decision to use conservative, older, and open standards-based technologies is a safe choice; but such technologies often cannot compete with the startup or proprietary technologies in satisfying the cutting-edge requirements of the leading-edge sciences.

The advantage of the cutting-edge approach is obvious: When operational, it usually brings remarkably increased capabilities to bear in support of the science. IBM's Blue Gene and the CRAY XT3 both use proprietary interconnects for performance reasons, and the widely deployed Myrinet started as a proprietary cluster interconnect technology.

Smaller startup network vendors such as CASPIAN, Calient, Cienna, Force10, Infinera, and Juniper have brought innovations to market quicker than have larger competitors and were subsequently aligned with the leading-edge requirements of the science community.

The disadvantages of this approach include a lack of a broader community expertise normally achieved with open standards-based technologies, as well as the potential risk of having a “niche” startup company, or the “startup” department or product line of a large company, fail or fall behind in the ability to supply or support its equipment. Another issue is the ability of a niche, startup, or smaller vendor to remain profitable and solvent because of proprietary protocols, a small installed base, or the challenge of developing the next generation of technology while kick-starting and supporting its first product lines. The latter point is exemplified by Myricom’s recent decision to build 10 GE networking equipment to remain competitive in an evolving marketplace and IBM’s use of Infiniband to connect Power Series machines. The decision ultimately comes down to a risk versus potential benefit tradeoff analysis; but grand challenge science normally requires grand challenge-capable technologies, which are more often delivered by niche startup companies or niche departments in larger companies.

The reason that these technical business trends are noted is that it is important to realize both the pros and cons when choosing cutting-edge technologies. If the requirements of the science mandate the use of cutting-edge or niche technologies and facilities, then DOE/SC/ASCR needs to anticipate and provide for the additional support infrastructure and personnel necessary to appropriately manage such a facility. The grand challenge of the future will be the management of these technologies and the systems in which they are deployed.

AIV.2 Computing and Moore’s Law

The expectation is that Moore’s law with respect to computing will continue for the next decade; however, this will be accomplished mainly through the use of parallelism—multicore compute nodes, multiple and multifunction ASICs and FPGAs, parallel striping of I/O, storage, and networks, and the use of parallelism in the middleware such as GridFTP. The NERSC Report (page 16) notes:

We expect these performance trends to continue in the next decade with two major differences. First, as opportunities for additional instruction-level parallelism wane, chip manufacturers will exploit smaller feature sizes by increasing the number of processors per chip. Indeed, some experts suggest that the number of processors per chip is likely to double every few years. Second, vendors of high performance computing systems will increasingly implement processor and network accelerators. These two factors will expose more parallelism to applications and potentially require significant code and algorithm changes to achieve high performance, placing an even higher burden on applications programmers.

This move to multicore processors and self-contained systems on a chip will most likely need to be accompanied by agile and adaptive systems and software, which will subsequently increase the challenge associated with managing and utilizing these systems for petascale science and beyond.

The NERSC Report (page 16) further notes that “the peak performance of high-end computing platforms has increased by a factor of 1.8x annually over the past decade, with two trends contributing equally to this growth: (1) individual processor performance has grown by about 1.4x annually, due to both higher clock speeds and the increased use of on-chip parallelism; (2) the number of processors in high-end machines has increased by an average factor of 1.3x annually.” Terascale systems of today currently have 128, 256, 1024, and 2048 processor nodes. The top-end terascale and near-term future petascale systems will have anywhere from 1,024 to 65,000 processors. The systems eight to ten years from now will have hundreds of thousands of processors. The NITRD Roadmap (page 11) for the long term calls for self-tuning HEC systems that can sustain 10 to 100 PF. These systems will have hundreds of thousands of processors. There will be self-organizing and self-reliant chips, nodes, processors, and systems. In order to be able to benchmark, tune, monitor, and manage these systems, we need to better understand the system and the many interconnected dynamics and components that affect the system, namely, memory, processors, I/O for storage, interconnection area network, and even external network and storage data I/O for file and data movement.

There is a growing gap between processor performance and memory system performance. Increase of memory bandwidth of commodity microprocessor memory systems has slowed to about 23% per year, as noted on page 17 in the NERSC 2006-2010 Report, which itself references an NRC report on the future of supercomputing. Memory accelerators are one way to try to address this memory performance challenge (see page 18 of the NERSC Report [36] for a more detailed description), but in essence rely mainly on prefetching the data. Work will continue in this and other areas of prefetching and caching of data to address both the memory and the I/O latency issues.

Another significant hardware trend is the increased use of application-specific integrated Circuits (ASICs) and Field Programmable Gate Arrays (FPGAs) in systems to support specific functions and offload the general processing units. Both ASICs and FPGAs are useful for applications and systems with special requirements (i.e., instructions or function). IBM’s Blue Gene uses the same ASIC for both computation and I/O. The function is determined by configuration control at boot time. We can expect to see increased use of both ASICs and FPGAs as integral parts of systems in the future, specifically in support of particular applications. Given the ability to select the mode of the ASIC at boot time, as is the case for the IBM Blue Gene, it is not a stretch to envision the use of dynamic ASICs that could morph between I/O and computation during the life of a job for enhanced matching of function to dynamic job requirements (i.e., as the code moves into a heavy I/O phase the ASIC changes to an I/O node). In addition, today’s high-end computers and I/O systems are statically configured much like the operating systems and architectures of the 1960s and 1970s. The use of ASICs and FPGAs may be

of help in evolving the current capability petascale computers, storage, and networks to a new dynamic age that molds itself to the applications being run.

AIV.3 Storage

AIV.3.1 Challenge of Data Storage

Data storage will remain a major challenge for DOE scientists and facilities. The volumes of data continue to grow in all science domains. In 2003, the NERSC center went from being a data source to a data sink. It now holds approximately 30 years of archived data, and it expects that the data needing to be archived and stored will catch up with storage capacity and capabilities by 2010. LLNL's HPSS petabyte file system is expected to soon become a 10-petabyte system. Data provenance not only includes the mechanics of storing the data but also needs to ensure both the integrity and safety of the data while stored and in transit. For example, NERSC uses a three-phase data management plan to address the issue of data integrity and safety. Data safety is implemented on many levels. As the NERSC report notes, "The first level of storage hierarchy is a high performance HPSS cache RAID5 and RAID3 disk. Data is written to tape immediately after it arrives on HPSS cache disk. Our tape environment is fully automated, and storage tape silos are strong, resistant to damage, and equipped with fire suppression. The storage system meta data is backed up offsite for disaster recovery every three months." Each of these levels of storage will be sorely challenged to handle the onslaught of data that will be generated by the petascale capacity and capability facilities. Many feel that holographic storage is still very far off in the future, with little evidence of any recent documented advances: in other words, holographic storage remains a holograph itself.

AIV.3.2 Tape Storage

Tapes and tape drive systems continue to remain viable for archival and longer-term storage purposes. Many scientists expect tapes to still be in use 10+ years from now. The density and speeds of tape drives continue to increase. Although tapes are not very good for efficient access to small random access objects and are very costly for high-speed capabilities, they will be an important part of the storage solution for some time to come, given the sheer capacity challenge facing science. As the Data Management Report (page 68) notes, "Tape storage becomes expensive, however, if the data must be accessed at high speed or, even worse, at high speed with an unpredictable access pattern. Buying 100 Mbytes/s streaming throughput from an array of tape drives costs 40 to 100 times as much as from a disk array. Today's robotic tape systems support efficient random access to objects of 10 gigabytes or larger but are expensive and inefficient solutions for smaller objects."

The NERSC Report (page 29) notes that it is using tape striping with HPSS. Current tape technology streams data at 30 MB/s, and by striping across three tape drives one can achieve a 90 MB/s transfer rate to better handle large file transfers. For example, a 372 GB file would take 3 to 4 hours using only one tape drive but about an hour with three tape drives. Tape technology also continues to evolve and increase in speed. NERSC will

upgrade its current tape drives to ones that can stream data at 120 MB/s. Hence, tape technology will remain a viable and integral component of DOE/SC's storage and archival portfolio; but, as noted in the Data Management Report, tapes are no longer viable for random access for small files.

AIV.3.3 Disk Storage

The Data Management Report (page 65) notes that magnetic disk transfer rates are not keeping pace with computational capacity. This mismatch is also the case with respect to the increases in rotation rates of the disks for both accesses and seeks. This in essence means that in order to provide for the demanding data and I/O requirements of terascale and petascale applications and leadership-class computers, there will be even more use of parallel I/O, virtualization techniques such as the prefetching and caching of data, and innovative architectures and technologies. The NERSC Report (page 29) notes that “using a four-way disk stripe configuration and multiple network interfaces, the network speed for data was increased from 80 MB/s to 200 MB/s. With the upgrade to 10 GigE, each individual stripe of data will be capable of full disk bandwidth, raising transfer capability to 800 MB/s.” This is but one example of the move toward parallelism to address the growing I/O requirements of terascale and soon petascale applications and data requirements. The Parallel Virtual File System also uses parallel I/O to address the mismatch between the computing and storage systems. Many sites are forced to add additional disks and controllers to their disk farms, not primarily to add storage capacity, but to provide better parallel I/O access to keep pace with the output of the computations.

These approaches are necessary, but they increase the complexity factor of the system, making it harder not only to benchmark I/O and storage systems but also to monitor, debug, and manage these systems. Another challenge is that as the increasing disk capacities make it harder to get truly random access to data for smaller objects. The Data Management Report (page 69) notes with respect to high-capacity disks, “If used to store thousands of 10-megabyte objects, today's disks may still be considered to support random access to these objects.” However, it goes on to note, “Many scientific applications require the retrieval of much smaller objects, often at or below the kilobyte level, resulting in retrieval rates that can be dominated by disk access time.” Although these trends may seem to argue for more R&D on architectures and solutions that incorporate prefetching and caching of data, the Data Management Report notes that “caching on a modest scale cannot be expected to eliminate the problem. High-transaction-rate commercial database systems address this issue with a memory cache equal to the size of the database. The future challenge will be to exploit large-market technologies to create in-memory scientific databases that are cost-effective”. The DOE/SC/ASCR-supported RAM disk project at SLAC is one approach that may be effective [37]. It requires data-cache memories of 10–100% of the data size, depending on the scientific field. In order to better understand and address these challenges, enhanced and better instrumentation, monitoring, benchmarking, and analysis capabilities of I/O systems, interconnect technologies, caching, disks, and networks are required.

AIV.3.4 Parallel I/O, File Systems, and Data Formats

With the explosive growth in the number of processors and parallelism, more attention will need to be exerted in the development of parallel I/O and parallel file systems as an integral component of a capability HEC system. The Data Management Report (page 68) notes, “Near-future applications require access speeds in excess of 10 Gbytes/s. Current ‘hero’ file I/O benchmarks are in the 1 to 10 Gbytes/s range using as many as hundreds of disks in parallel. Translating these benchmark results into comparable end-to-end I/O performance has been difficult and will become more difficult as more disks and compute processes are added to the system.” The challenges will grow larger with respect to benchmarking, monitoring, and managing these systems as the number of processors, ASICs, and FPGAs used in systems increases.

MPI is one model of parallel I/O where a system approach to the management of the I/O benefits the researcher. Other popular parallel I/O libraries are netCDF, Parallel netCDF, Panda, and HDF5. HDF5 and netCDF are two higher-level I/O libraries that provide the ability to abstract away the details of file layout as well as standard portable file formats and metadata descriptions of contents. UIUC/NCSA developed the parallel I/O Panda library. Panda provides for wrappers around HDF so that the user does not have to know about physical I/O. In addition, data-parallel model languages and programming environments such as Fortran 90, High Performance Fortran, Global Arrays, and Global Addresses support various types of parallel I/O.

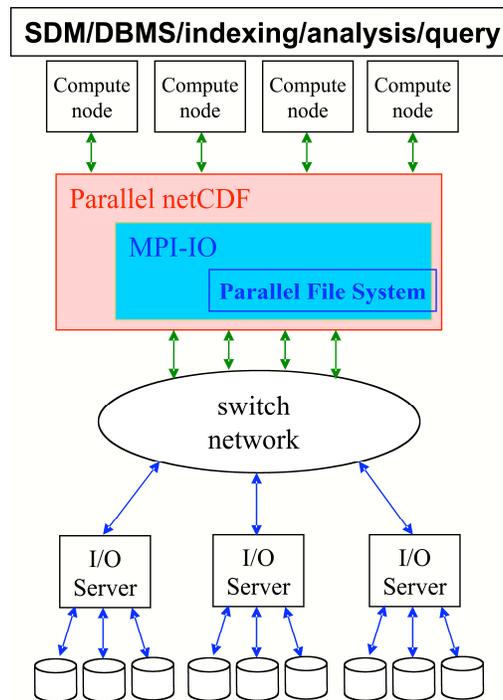


Figure AIV.3.4.1: MPI parallel file system

As noted in the Data Management Report (page 68), the MPI model works in terms of a cloud of parallel compute processes accessing a cloud of file servers (see Fig. AIV.3.4.1).

In order to provide a convenient model for access and high throughput to storage, a collection of I/O components is used, which consists of three distinct layers: high-level I/O libraries (e.g., PnetCDF and HDF5), I/O middleware (e.g., MPI-IO), and parallel file systems (e.g., PVFS2, GPFS, Lustre). For future applications to use this I/O model, the performance of these components must be improved, particularly in the area of throughput and scalability. One way of improving the stack as a whole is to tune how components communicate with one another. For example, implementing a richer language for describing I/O accesses to the parallel file system and using this language in MPI-IO can provide significant performance gains.”

In addition to the physical challenges associated with storage and I/O hardware systems, as well as the various parallel I/O libraries, there exists the issue of the data formats. There seems to be no shortage of data formats being used by the science community (e.g., HDF, HDF5, netCDF, gif, jpeg, XML, and the GGF data format description language). HDF5, the successor to HDF, supports multithreading and files over 2 GB; netCDF is a machine-independent self-describing data exchange standard.

AIV.3.5 Summary

Some believe that the development of optical interconnects will help address some of the aforementioned I/O challenges; but optical interconnects by themselves provide no panacea for addressing this I/O to storage challenge, specifically because of the media challenges (i.e., tapes, disks) associated with keeping pace with the data tsunami. These challenges to scaling will need to be addressed through parallelism and innovative architectures for the short and medium term. However, experimentation with new technologies, techniques, and architectures should also be pursued now, in order to try to address the longer-term challenge of effective data provenance. One potential example is the combination of optical interconnects with first-level storage RAM disks as part of the system interconnect; combined with morphable I/O and compute nodes, it might help address some of these parallel I/O issues. R&D is also being supported in using fiber as a storage media (i.e., CANARIE’s optical disk [38]) as well as work done at the University of Tennessee using the network for transient data storage and management. SLAC is also working on a DOE/SC/ASCR-funded project to use RAM as disks. Moreover, Google and Akamai have focused on content management and delivery for a variety of Web-based commercial and public delivery systems. Similar attention needs to be placed on the long-term management and placement of scientific data in a distributed metafacility. There may be lessons learned for the R&D community from the Google, AKAMI, and Amazon models.

AIV. 4 Interconnect Area Networks

The NRC report (page 17) notes that “the cost and power of providing bandwidth between chips, boards, and cabinets is decreasing more slowly than the cost and power of providing logic on chips, making the cost of systems bandwidth dominated by the cost of global bandwidth.” The NERSC report (page 19) adds that the gap between processor speed and network latency is growing more quickly than the gap between processor

speed and memory latency. These trends indicate that the interconnect is losing the race relative to the increases in speed of processors and memory, and will most likely become a bottleneck in petascale computing. Latency, speed, and jitter remain a major concern with any interconnect technology and architecture. The management goals of performance, predictability, and usability remain very relevant.

The NERSC report (page 19) makes the following comments:

The problems of bisection bandwidth scaling, especially as the number of processors scale, has led to an increasing interest in interconnect topologies with less than full bisection bandwidth, such as torus and related mesh topologies. For relatively small systems, such interconnects perform well, especially because point-to-point bandwidth is over provisioned, resulting in reasonable bisection bandwidth. It has not yet been demonstrated that these topologies can be effective on NERSC's diverse workload at large system sizes (thousands of nodes). Various forms of network accelerators are being considered by researchers and HPC vendors to address the network performance gap. Support for direct access to remote memory has been available in custom supercomputing networks for many years, but at a high cost. The introduction of standard interconnects such as Infiniband has marked a trend in networking—these networks have a broader market than high performance computing, and still support direct access to remote memory. There is still a latency advantage to more highly customized networks that connect through a memory interface, rather than an I/O bus, but the performance difference is smaller.

With respect to the cluster interconnect options, Myrinet, which was a proprietary standard, was the original technology of choice for HEC systems. There has been a recent surge by Infiniband in this niche market. These two protocols were specifically designed for HEC clusters and have predominated over Fast and Gigabit Ethernet, which were designed for more general inter nodal communications, because of the latter's superior latency and bandwidth; but with the coming of age of 10 GigE and TCP offload engines (TOE), which moves the processing of TCP off board much as Infiniband and Myrinet perform off board processing of their network protocols, the 10 GigE and TOE architecture might become just as acceptable as the others for high-end computer systems. Figure VI.4.1 provides a stack comparison of 10GE and TOE versus Infiniband. Prices for 10 GigE and TOE will come down as the technology matures; and if Gigabit Ethernet is any indication, the mass production of 10 GigE in the next few years will make it very cheap compared to proprietary protocols.

The majority of the IAN and CAN protocols rely on remote direct memory access and offloading the node processor with an intelligent NIC that handles the protocol stack processing without involving the node processor(s). Traditional Ethernet has been known for causing a lot of overhead and interrupts for the host node. Myrinet, Infiniband, and 10 GigE/TOE all make use of RDMA and offloading, much like the CDC 6600's peripheral processing units. Some vendors are seeking to converge and combine some of these competing technologies; and Myrinet is starting to support the standards-based 10 GigE.

The NCSA/UIUC MPI Pallas benchmarks show that with respect to bandwidth and latency [39], Infiniband is better than Myrinet, which is better than TCP over Gigabit Ethernet, and the 64-bit processor tests showed SHMEM to be better than Infiniband. These tests are now a couple of years, however, with no comparisons to 10 GigE and TOE. It is important that the petascale infrastructure architects and users better understand the minimum latency and bandwidth required to support various sets of applications and architectures so that they can do a proper cost-benefit tradeoff comparison between the alternatives. Given the interplay between interconnect and cluster area networks with parallel I/O libraries and file systems, it is critical not only to understand the low-level protocols but also to better understand what the potential ideal number of I/O and file server nodes for an application coupled with the characteristics of the interconnect technologies. The petascale research community needs more and enhanced benchmarking, monitoring, and analysis for all interconnect technologies, networks and I/O. Given the lack of much empirical data for today’s terascale systems, it will be even more important to know how MPI, GA, Infiniband, 10 GigE and TOE, Myrinet, and other technologies perform as we move from terascale to petascale systems with hundreds of thousands of processors and hundreds of I/O nodes [40]. The establishment of one or more multisite benchmarking and performance institutes would help address this challenge.

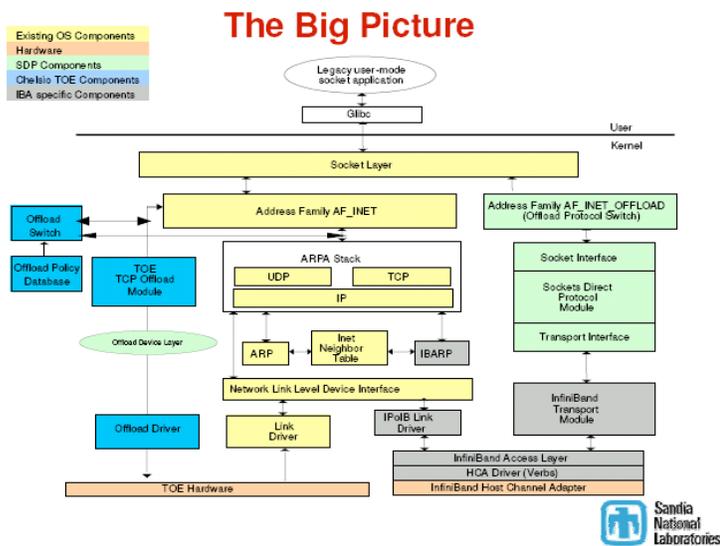


Figure AIV.4.1: Comparison of 10 GigE/TOE and Infiniband architectures (courtesy of Sandia National Laboratories)

AIV.5 Networks—Enterprise or ISP

One of the challenges faced by the DOE/SC science programs is that the networks supporting the sciences is often considered an enterprise network from the perspective of management and acquisition, but in reality these networks are more like ISPs and in fact

use ISP equipment in their networks. As a result, science networks often require ISP-class high-performance routers and switches with enterprise-focused “high touch” feature support such as QoS. Yet many vendors consider this a niche market and often do not provide such features for high-end ISP equipment, or these features are added very late in the life cycle of the router/switch. This has been and will continue to be a challenge with supporting the international science community. The requirements of commercial content providers such as Comcast, Google, and Amazon mirror some of those of the science community; and in those areas where they do intersect, they may have some effect on the availability of certain technologies.

AIV.5.1 Dark Fiber and Waves

Many of the bandwidth advances in telecommunications over the past five years were accomplished with the adoption of DWDM and more specifically with the adoption of dark fiber by the R&E community. CANET, NLR [41], SURFNET, FLR, UltraScienceNet, I-Wire, and CENIC are just a few examples of networks that have acquired dark fiber. With the recent spate of telecom mergers and consolidations, we can expect to see less dark fiber available to the community. The Telcos will provide wave services; but, as part of their financial recovery from the early 2000s dot.com bust, they will be seeking to generate more profit, and therefore it remains to be seen how competitively the Telcos will continue to price these wave services. Those possessing long-term dark fiber IRUs may have a financial advantage five plus years from now; however, the real cost of a network needs to also include the cost of the technology needed to light, terminate, and mux the fiber, as well as the technologies needed for moving and switching the data. Petascale facilities are dependent on multiple 10 Gige wave services, and the cost of services and the quality of the support of those waves thus is of paramount importance.

AIV.5.2 IAN, CAN, SAN, LAN, MAN, and WAN

Networking has evolved from the simple local area network (LAN) connected to the wide area network (WAN). We now have interconnect area networks (IANs) that provide for the interconnection of massively parallel high-end systems, cluster area networks (CANs) that interconnect processors in loosely or tightly coupled clusters, storage area networks (SANs), local area networks (LANs), and metropolitan area networks (MANs).

The MANs were built to better serve intermetropolitan distribution and connectivity and are often an edge aggregator that connects to WANs. ESnet currently supports four MAN services (BAMAN, LIMAN, CHIMAN, JLAN) connected to its WAN. The SANs (e.g., fiber channel) arose to address specific requirements such as QoS and latency among local system and storage systems, which the predominant Ethernet LANs at that time could not do. The CANs (e.g., Myrinet and Infiniband) were developed specifically to address the low-latency, high-performance cluster environment. IANs were developed mainly as a high-speed backplane network to interconnect tightly coupled processors. Current IANs are mostly proprietary technology. Some of these networks (such as CAN, SAN, and LAN) are converging (e.g., Infiniband), and many varieties of CAN, SAN, and

LAN protocols are being tunneled over an IP WAN. The MAN and WAN networks are now mainly SONET and DWDM based.

As part of moving data to or from a petafacility and as part of a petascale computation, experiment, visualization, or analysis, the data may traverse an IAN, SAN, CAN, LAN, MAN, and WAN, often each with its own protocols, protocol data units (PDUs), architectures, naming, and addressing. The more protocols and different technologies that the data has to pass through, the more likely that error, delay, and jitter will be introduced. In addition, a lot more energy is consumed because each time a device needs to perform a “high-touch” action such as framing, transformation, PDU (de)fragmentation, or encapsulation. Another issue associated with the data traversing the many different types of networks is the ability to coherently allocate, monitor, and manage all of the network resources.

AIV.5.3 10 Gbs Building Blocks

We are about to enter the time with respect to 10 Gbs technology components and manufacturing processes, at both the LAN and WAN level, that market analysts often call the “sweet spot.” This means that the availability and cost of 10 Gbs is at a point to encourage wide deployment for the next four to five years. This is especially true for 10 GE. Desktops are now seeing 10 Gbs network interface cards (NICs). In 2002, the NLR made the commitment to the first large-scale national 10Gbs network, 10GE to be exact. The NLR was operational in 2005. ESnet is now at 10 Gbs, as are many other R&E networks. Some vendors are already providing 40 Gbs optical switches, but this is done by striping 10 Gbs waves with proprietary technology. Infinera combines 10 waves onto a chip and drives all 10 waves with one transponder. The 40Gbs rates of some high-end routers are accomplished through parallel striping of multiple 10 Gbs waves backed by an inverse optical mux. Some vendors are already experimenting with 100 Gbs in their research labs, but these are short reach optics. As the speed is increased past 10 Gbs, the reach is reduced. In addition to increasing the bandwidth through various means, some vendors are also concentrating on extending the distances of 10 Gbs links to multiples of thousands of kilometers, which would reduce the cost of supplying and supporting a 10 Gbs service because of the reduction of costly regeneration. R&D is also being done on combining Ethernet labels with G.8709 wrappers to try to develop effective and less costly technologies. In summary, the science community can expect to be relying on 10 Gbs building blocks for its networks for the next five years. Higher speeds will be attained through striping 10 Gbs waves, that is, parallel pipes.

AIV.5.4 IP and Routing

The telecoms and carriers are moving toward an IP core for both data and voice. They are deploying multiaccess routers at the edges of their networks and use these multiaccess routers to combine ATM, Frame Relay, and IP services onto an IP core. There is also a convergence of layers 2 and 3. Layer 2 switches are often indistinguishable from layer 3 routers with respect to their function and capabilities. MPLS is often used in conjunction with IP to provide for layer 3 path-based VPNs and QoS.

One direct result of the move toward an IP-based core is that the core IP-based routers need to be the very powerful and capable. The highest-end routers such as Juniper's T series routers and Cisco's CRS-1 router are themselves massively parallel, multichassis, multiprocessor, "capability" HEC systems dedicated to the task of moving data. These routers are capable of aggregates of 640 gigabits and more and are used, as is MPLS-capable devices, at the core as well as the edge of the network. A disadvantage associated with these high-end routers is their cost, as well as footprint, cooling, and power requirements. Hence, some network architects are trying to reduce the number of high-end backbone routers in the networks and use optical switches and cross-connects in their place.

Another trend in core, edge, and site-based routers is the use of "blades" and network service processors (NSPs) in the switches and routers. These are often used by the vendors to support high-touch ancillary features such as QoS, security, IDS, DOS, and network management. Many of the blades support some flavor of Linux or Unix. An opportunity exists for collaboration between the science community and willing router/switching vendors to investigate the use of generic blades in routers and switches that could then be programmed to address scientific data management challenges such as caching and temporary storage, RAM disks, prefetching, in situ metadata engines and repositories, GMPLS, and other relevant functions associated with the movement and management of scientific data.

The router versus switched lambda debate often focuses on the need for performance based, predictable, and usable network services. Many proponents of the optical wave solution and circuit-based VPNs point to the fact that IP is not predictable under heavy load and does not provide for QoS, mainly because of the way TCP works with IP, as well as the way buffer and queue management is implemented in many routers. Many QoS techniques are available at the IP level (e.g., diffserv), and their lack of deployment is often due to the lack of an appropriate inter-domain policy and resource management service rather than a lack of technical capability. Lambda switching faces this same challenge. As a counterexample of the move away from IP-based networks for predictable performance, Caspian Networks supports QoS and flow-controlled IP services. Other trends with respect to routing include the support of logical and virtual routers that have their own routing databases and can be managed separately.

AIV.5.5 Optical Networks

We noted that the bandwidth and speeds that we have today are a result of the move to an optical infrastructure, specifically DWDM. The debate between circuit- and packet-switched networks has raged for decades and will most likely continue. BEIP is just what it says, "best effort," albeit with MPLS one can get a version of QoS. Circuits, specifically DWDM waves/lambda, are now being favored for the movement of large datasets and files by the science community, mainly because of the end user's goal for determinism, predictability, and performance (i.e., if there is but one application on the lambda or circuit, it will not have to contend with other applications). A circuit-switched

lambda service can be coupled with one of the TCP variants that support large dataflows (e.g., FAST or HS-TCP), for a predictable high-performance service that allows for the full use of the circuit by one application.

The majority of today's lambda networks are more of a timeshared multiplexed lambda service whereby a prefixed and preallocated set of lambdas is shared with a set of networks and applications. Such sets are often shared without any accounting and authentication (e.g., GLIFF) and are made available to the applications on a scheduled time-shared basis. There is some experimentation with lambda switching and automated optical patch panels. GlimmerGlass and Calient fall into the latter category and essentially are used to switch the directions of light signals without regard to framing. The Optiputer project dynamically (de)allocates private optical paths to form a distributed virtual computer.

Work is also under way on incorporating optical transponders into router blades. This strategy gives the router, working with the appropriate optical switches (e.g., Cienna's Core director) and dynamic optical patch panel systems (e.g., Glimmerglass and Calient), the ability to signal the layer 1 optical infrastructure to affect lambda switching; that is, the optical switch is treated as merely an optical mux/demux by the router. This technique works for enterpriselike deployments where one administrative domain controls both the layer 3 routers and the layer 1 optical infrastructure, since it needs to ensure consistency and trust between the various layer management systems. The management, control, and business model aspects associated with this cross-plane management of optical infrastructure is one reason that interdomain lambda switching will remain a challenge and will probably not be supported by the carriers. A new trend is in the use of transponders in routers, and optical switches, to convert the ITU grid frequencies of 1550 and 1310 nanometer signaling into DWDM lambdas and then use optical switches merely as optical mux/demuxes. These are often referred to as "alien waves." Many vendors are experimenting with and investigating alien wave support.

GMPLS is a protocol being developed to support the signaling for lambda switching. The management and business model challenges associated with lambda switching needs to be solved to make inter-domain lambda switching viable, otherwise the community may have to rely on layer 2 VPNs to affect the predictability, performance, and usability envisioned with the adoption of lambda switching. Lambda switching has the same interdomain business, policy, and resource management challenges that face the deployment of interdomain QoS-based IP services, and it is a very similar with respect to the challenges encountered in deploying ATM virtual circuits in the 1990s. If the waves are statically allocated and configured among predetermined sites, as will be the case in the first iteration of ESnet's Science Data Network, which will provide multiple 10 Gbs waves between LHC tier 1 sites at Fermilab and BNL, then in order to grow the network capacity to satisfy the other petascale facilities as they start generating petabytes of data, DOE/SC will need to support a much denser mesh of waves and circuits between all sites. This could be costly and hard to manage. The other option is lambda switching or the development of a true layer 3 IP QoS capability. All require relevant policy and

management systems. The OSCARS capability is designed to address the allocation of VPNs.

Optical technology, as noted above, is currently in the 10 to 40 Gbs range and is being developed and deployed by vendors. DARPA is supporting two all optical high-end router research projects as part of its DoD-N program. UCSB's LASOR project is an example of an all-optical routing project. The prototypes should be available by 2010 and will be capable of 40 to 100 Gbs of all optical packet switching. One of the challenges involved with moving into this realm is that the optics itself needs to be more self-monitoring and correcting. Tunable lasers are being deployed now for easier set up and maintenance; however, as Alan Wilner notes, the optical network is a fragile infrastructure and will become even more so as we increase the speeds and bandwidths [42]. For example, wave leakage is more likely to occur at these higher speeds and can even become an unintentional DOS attack. Wilner argues for enhanced operating, administration, and maintenance systems.

The CAL-IT² program at UCSD is supporting research on quantum optical networks. A five-node terabit network could be built by 2008, but much more research needs to be supported in the application of information theory to achieve any serious advances in quantum optical networks.

AIV.5.6 Transport Protocols

One of the reasons researchers are abandoning BEIP for other more predictable solutions involves the use of TCP and the problems associated with the effect of the TCP sliding window mechanism on the ability to predictably make full use of the bandwidth for moving large datasets. The ESnet Requirements Report (page 27) notes the challenges with using TCP:

For example, in order to fill a one-gigabit path which has a 100-millisecond round-trip time and a packet size (MTU) of 1,500 bytes, a TCP stream would have to have of the order of 8,000 packets in flight continuously—the equivalent of 12 megabytes. A 100 millisecond round-trip time is approximately an East-West coast transfer. The TCP protocol is designed with the premise that the random loss rate in network components is insignificant compared to the loss rate due to congestion. Thus, the TCP sending rate is a function of the packet loss rate (assumed congestion) in the network. The packet sending rate drops dramatically as a response to a congestion event (packet loss); then the sending rate increases slowly until the next congestion event is encountered. The ESnet network must be designed to minimize end-to-end single stream TCP packet loss as well as support alternatives to TCP.

A lot of effort is being focused on the evolution of TCP. The FAST version of TCP was used at SC2006 and was instrumental in winning the bandwidth speed challenge. Other TCP revisions include High Speed TCP (HS TCP), XCP, and the tuning of end systems

as part of the solution (e.g., the WEB100 project). Considerable work needs to be done on making TCP scale to tera- and petascale networks as well as investigating alternatives.

AIV.5.7 Multimode Networks

The NITRD report (page 44) notes the need for multimode networks: “However, controlled use networks or dual use networks will be required to assure that researchers doing research for network development (with a high risk of bringing down the network) do not interfere with applications researchers who require highly reliable networks.” The NLR, based on the MORPHnet concept, was designed to provide this capability. DWDM was the liberating variable for enabling the NLR and MORPHnet to become a reality since it provided cost-effective layer 1 “physical” separation. In addition to the various types of production services incorporated into a hybrid multimode network, the network also needs to include an experimental capability to encourage researchers and applications to experiment with and evaluate new technologies such as skinny transport protocols for the movement of large data files, alien waves, and lambda switching. This capability needs to extend past the backbone and into the local, storage and interconnect networks.

AIV.5.8 Hybrid Networks

When one weighs the pros and cons of packet switching versus circuit switching and TCP versus other transport protocols, it is apparent that hybrid networks, such as those planned for ESnet4, will be used for some time to come. DWDM wave-based services are valuable because they can support each one of these services on one or more lambdas while keeping them physically separate. Another necessary component of the hybrid network will be the support of an experimental network infrastructure, for example, the use of one or more waves to support the investigation and analysis of various transport protocols or lambda switching in support of high-end science.

AIV.6 Security

What most people refer to as security—authentication, access control, and auditing—is really a subset of any organization’s overall management of their resources and people. It is not surprising, then, that both network and system management systems have many intersections and overlaps with the frameworks, technologies, and approaches supported and used by security personnel and systems support staff. For example, network and security management systems both build systems aimed at detecting anomalies, although the security systems usually monitor for intentional attack, theft, or interference with resources whereas network and operating systems have traditionally sought to detect and protect against unintentional errors. These two areas are converging, and systems now need to protect themselves against both intentional and unintentional attacks. Both networks and systems keep some form of “logs.” It behooves the community to integrate and synchronize these logs into an overall management system and capability.

NERSC, as do other sites, implements a variety of security measures. NERSC, however, does not use a traditional border firewall but instead utilizes internal and external intrusion detection systems based on Bro (<http://bro-ids.org/>). Network subnets and various flavors of internal firewalls segment services and critical systems such as user account databases and system control workstations are used in combination with a single sign-on, workload profiling, and auditing. PNNL has architected its network to push the firewalls to as close to the system as possible to allow for a finer grained security. For example, the EMSL cluster has its own firewall. Another trend that accompanies the trend of virtualization of networks and routers is the virtualization of firewalls. There can also be a tension between site and facility/meta-facility security policies and access due to the global nature and subsequent international remote use of the facilities. The continued challenges to all firewalls and intrusion detection systems indicate a need for all security policies and network/system policies to be folded into an enhanced overall management system that integrates the potentially conflicting policies

Deb Agarwal of LBL has suggested that more R&D be focused on the following security areas: auditing and forensics, dynamic host firewall port management, identity management, and secure middleware. The petascale facility of the future will require high-speed (10 Gbs and up) network encryption and monitoring, certificate management (i.e., certificate revocation), interdomain and federated trust and certificate management, multisystem and multifacility security, and interdomain management and auditing systems. Regardless of the advances made in infrastructure security, the petascale infrastructure of the future will require an enhanced focus on security as a major component of an overall policy driven management architecture.

Ian Foster, Tom Scavo, and Frank Siebenlist have a soon-to-be-published implementation approach for enabling attribute-based authorization for the TeraGrid [43] and other distributed systems. The approach leverages work done by AKENTI [44], Shibboleth [45], VOMS [46], and other access control and attribute management systems and combines the best of these into the Globus Toolkit 4 [47]. Figure AIV. 6.1 shows the use of policy information points and policy decision points as key components of this integration in the GT4 authorization architecture.

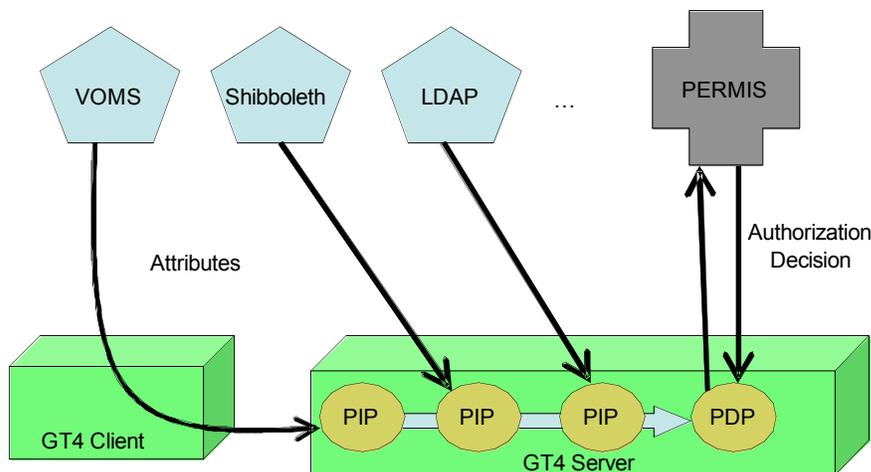


Figure AIV.6.1: GT4 attribute-based authorization architecture.

The visibility and importance of HEC facilities may make them attractive targets for hackers; and if the protections are breached, not only will DOE/SC lose the use of a petascale facility, but the infected facility can be used to launch massive denial of service attacks against other facilities.

AIV.7 Architectures and Systems

DOE has long excelled at using a variety of advanced technology “serial one” leadership computers and combining them into a cutting-edge system and facility to support advanced science. DOE/SC is again facing this challenge as it moves into petascale science. The systems and facilities may often require collocation because of the performance requirements and the need to share local I/O and file systems as well as to address latency and jitter. At the same time many DOE labs support their own local capacity clusters. The recent trend toward the distributed integration of facilities and multidiscipline datasets will create tension with this other trend toward centralized capability architectures. This tension between centralized and distributed systems will never go away. An approach that combines the two models will best support the researchers. Petascale facilities, metafacilities, and Grids all depend on integrated middleware, networks, security, auditing, resource allocation, scheduling, object and data management, and naming.

Relevant architectural work has been pioneered by Globus, Condor, Legion, TeraGrid, and most recently Optiputer and MonALISA. The Grid-based systems of Globus, Condor, and Legion have done a lot to advance the concept and formative implementation of virtual facilities and virtual organizations through shared resource management, scheduling, and allocation. The TeraGrid is a multisite experiment that allows the user to shop for local resources, much like the Grid systems; and like the Grid systems the TeraGrid has focused on a common set of software to be supported at the participating sites. The Optiputer system is a layer 1 optical lambda/wave system that has private optical paths at both the local- and wide-area basis and that can be dynamically set up on demand and combined with end resources to form a distributed virtual computer. All of these require policy-based management. The Simple Network Management Protocol (SNMP) with its associated MIBs became the standard management platform for a broad variety of network technologies. An effort should be made to investigate the development and implementation of a standards-based Simple Systems Management Protocol (SSMP), which could be used to support the management of petascale facilities, metafacilities, systems, and Grids. The GLUE schema work by HEP community, object-based architecture work sponsored by OMG and CORBA, and the resource definition languages of Globus, Condor, and other relevant technologies are a good base from which to start.

AIV.8 Visualization

Visualization is a data-intensive operation that utilizes visualization servers, computational clusters, networks, storage systems, and file systems. Currently, many of

these visualization services/servers are collocated with the leadership-class computers or leadership experiments. The NERSC Report (page 32) notes:

Over the years and into the foreseeable future, the most suitable systems for data-intensive computing are scalable shared-memory systems. In the late 1990s and early 2000s, the SGI Onyx platform was the best choice for providing raw I/O bandwidth, a scalable shared-memory architecture, support for diverse parallel programming models, and cache-coherent non-uniform memory access (ccNUMA). Especially in the case of interactive work, the load on the machine goes from near idle to 100% capacity across all processors and I/O channels in response to the “bursty” nature of interactive visualization.

A lot of work has been done on visualization display technologies, including desktop computers, ImmersaDesks, CAVEs, and Powerwalls. One of the challenges associated with some of these technologies is the scheduling of these room-based technologies and the desire of researchers to perform their research from the comfort of their offices. For the appropriate project, however, these visualization facilities provide immeasurable benefits. Many labs and universities are working on visualization tools and software to take advantage of these physical capabilities. PNNL [48], for example, has developed Starlight, an object-oriented approach to do exploratory information analysis, and Inspire, which provides tools for exploring textual information, including query, subset, and trend analysis tools. The open source distributed computation model Paraview is a widely used application designed to support the visualization of large datasets [49].

AIV.9 Naming and Addressing

Naming and addressing are crucial elements of any local or global infrastructure. There is currently a tower of Babel of naming schemes and infrastructures, which cover a wide range of facilities including networks, the Web, file systems, storage systems, parallel I/O systems, global addresses, global arrays, and massively parallel systems. There are URIs, URNs, URLs, e.164 addressing, SIP, Handle system, DNS, IPv4, IPv6, NFS, HPSS, netCDF, HDF5, X.500, MPI, Akenti, and many other naming systems.

AIV.10 Virtualization

As noted many times in other sections of this report, virtualization of resources and organizations is happening at a very quick pace, at the system, network, storage, and facility levels. Virtualization has taken hold in commercial data centers; and HP is delivering, or will in the near future, virtual computation, network, and storage resources. EMC supports VMware for x86 compatible computers. One aspect of virtualization is the use of parallel resources in a seamless fashion such that the combined resources appear as one to the user. A challenge associated with the use of parallel resources is determining at what level the use of parallel resources introduces too much overhead with respect to communications and synchronization such that the system is no longer benefited by the additional breakdown of tasks into subtasks. The Mythical Man Month noted that there is a similar level of task breakdown with people.

System virtualization started with the IBM S/360 and VM. VMware and XEN are two more recent supporters of virtualization of systems. Barney Maccabe, at the University of New Mexico, and other leading researchers have been researching whole system virtualization as well as hypervisor capabilities in HEC systems. Current high-end routers support logical and virtual routers and VPNs exist at all levels of the network. In order to better understand when to use parallelism or virtualization in systems or networks and to support their use, new monitoring, benchmarking, and analysis tools and capabilities will need to be developed.

AIV.11 Programming Environments

Many of the research agencies are supporting R&D focused on the next generation of operating systems, languages, and libraries necessary for utilizing the next-generation capability machines. Language-based programming models include Co-array from Rice, the unified parallel code (UPC) from Berkeley and LBNL, Titanium from Berkeley, and the DARPA-supported HEC language development of FORTIS with Sun, Chapel with CRAY, and X10 with IBM. These languages all incorporate more of an abstract approach to using HEC facilities. In addition to the languages, there are various library-based programming models such as the message-passing MPI-2 library, the CRAY XT3 SHMEM shared-memory library, the Global Arrays library developed at PNNL, and the Global Address Space (GAS). Projects supported by the Forum to Address Scalable Technology for runtime and Operating Systems (FAST-OS) include HEC and K42, Scalable Fault Tolerance, ZeptoOS, Petascale SSI, Modular Linux and Adaptive Runtime (MOLAR), Dynamic Adaptivity in Support of Extreme Scale (DAiSES), Config OS, and Colony. An ongoing challenge will be whether to add new capabilities in the OS, language, or library.

Given the differences in approaches by both languages and libraries, as well as the capabilities of the operating system, it would be helpful to have better benchmarking, instrumentation, monitoring, analysis, and management capabilities to better compare and contrast all of these with a variety of applications. One of the overarching issues is that the applications using leadership-capability HEC systems will generate a lot of data and will subsequently use a broad set of I/O, networking, data management, visualization, and computing technologies and systems. Therefore it is important to provide an environment that supports a metafacility view of programming and management. Another issue that needs to be addressed across the board is that these codes tend to run for a long time and subsequently will need appropriate monitoring, checkpointing, recovery, and restart on a metafacility and workflow basis, not just for the HEC system. A complementary issue is the need to provide for enough hooks, tools, and APIs to allow for a workflow, system, or application to release resources as soon as possible in case of failure to avoid the wasteful use of the unique petascale facilities.

References

- [1] <http://www.nature.com/nature/focus/futurecomputing/index.html>
- [2] http://www.nersc.gov/news/annual_reports/annrep05/research-news/06-whispers.html
- [3] http://www.nersc.gov/news/annual_reports/annrep03/advances/1.3.supernovas.html
- [4] http://lhc.web.cern.ch/lhc/general/gen_info.htm
- [5] “Federal Plan for High-End Computing, Report of the High-End Computing Revitalization Task Force, Executive Office of the President, Office of Science and Technology, May 10, 2004
- [6] <http://www.research.ibm.com/autonomic/>
- [7] “The Networking and Information Technology Research and Development Program: Supplement to the President’s Budget for FY2007,” A Report by the Subcommittee on Networking and Information Technology Research and Development, Committee on Technology, National Science and Technology Council, February 2006
- [8] <http://www.nlanr.net/>
- [9] <http://www.caida.org/home/>
- [10] <http://www.optiputer.net/>
- [11] “Guide to the NITRD Program FY2004-FY2005: Supplement to the President’s Budget for FY2007,” A Report by the Interagency Working Group on Information Technology Research and Development, Committee on Technology, National Science and Technology Council, December 2004
- [12] “The Office of Science Data-Management Challenge, A Report from the DOE Office of Science Data-Management Workshops of March-may 2004,” May 2004
- [13] <http://www.pvfs.org/>
- [14] <http://www.llnl.gov/CASC/components/overview.html#mgmt>
- [15] <http://sdm.lbl.gov/srm-wg/index.html>
- [16] <http://sdm.lbl.gov/sdmcenter/>
- [17] R. Aiken et al., “Architecture of the Multi-Modal Organizational Research and Production Heterogeneous Network (MORPHnet),” ANL-97/1, Argonne National Laboratory, Argonne, IL, Jan. 1997.

- [18] CERN: European Laboratory for Particle Physics, Geneva, Switzerland. The CERN LHC program involves major U.S. participation.
- [19] KEK: High Energy Accelerator Research Organization, Tsukuba, Japan.
- [20] “Advanced Computing: Simulating the Inaccessible, Discovering the Unknown,” SciDAC Review, Number 1, Spring 2006
- [21] <http://crd.lbl.gov/~kewu/fastbit/>
- [22] <http://picturethis.pnl.gov/>
- [23] Pegasus: <http://pegasus.isi.edu/>
- [24] GriPhyN: <http://www.griphyn.org/>
- [25] MyGrid: <http://www.mygrid.org.uk/>
- [26] <http://root.cern.ch/>
- [27] <http://www.oecd.org/dataoecd/30/36/36213997.pdf>
- [28] I. Foster, “Globus Toolkit 4: Software or Service Oriented Systems,” Journal of Computing Science and Technology, July 2006, Vol. 2, No. 4, pp 513-520
- [29] <http://www.pnl.gov/infoviz/about.html>
- [30] <http://staff.science.uva.nl/~zhiming/iccs-wses/>
- [31] <http://infforge.cnae.infn.it/glueinfomodel/>
- [32] NERSC, “Science-Driven Computing: NERSC’s Plan for 2006-2010,” NERSC/LBNL-57582
- [33] <http://loci.cs.utk.edu/>
- [34] Science-Driven Network Requirements for ESnet, Eli Dart, ed., LBL report, 2006
- [35] <http://www.opensciencegrid.org/>
- [36] National Research Council Committee on the Future of Supercomputing, *Getting Up to Speed: The Future of Supercomputing*, S. L. Graham, M. Snir, and C. A. Patterson, eds. (Washington, DC: National Academies Press, 2004).
- [37] R. Mount, “A Leadership-Class Facility for Data-Intensive Science,” http://www-user.slac.stanford.edu/rmount/leadership/HighEndComputingProposal--4_9_04.doc

- [38] www.viagenie.qc.ca/en/obgp_wdd/wdd/canarie_workshop-wdd-20011129.pdf
- [39] N. Biebaum, H. Chen, J. Decker, E.V.D. Vreugde, "Infiniband and 10 Gigabit Ethernet for I/O," Cluster Symposium, July 2005
- [40] <http://vmi.ncsa.uiuc.edu/>
- [41] <http://www.nlr.net/>
- [42] 2006 MIT CIPs Annual Meeting, May 2005
- [43] <http://www.teragrid.org/>
- [44] <http://dsd.lbl.gov/Akenti>
- [45] <http://shibboleth.internet2.edu/>
- [46] <http://vdt.cs.wisc.edu/VOMS-documentation.html>
- [47] <http://www.globus.org/toolkit/>
- [48] <http://www.pnl.gov/infoviz/technologies.html>
- [49] <http://www.paraview.org.HTML/Index.html>

Distribution List for ANL/MCS-07/5

Internal:

| | | |
|---------------|------------|---------------|
| R. Aiken (10) | E. Lusk | MCS Files (2) |
| R. Bair | G. Pieper | |
| P. Beckman | R. Ross | |
| C. Catlett | R. Stevens | |
| I. Foster | TIS Files | |

External:

| | | |
|---------------------|-------------------------|----------------------------------|
| D. Agarwal, LBNL | D. Hitchcock, DOE SC | B. Sanford, ORNL |
| D. Atkins, NSF | P. Jain, Nuova | D. Schissel, Gen. Atomics |
| D. Bader, LLNL | B. Johnston, LBNL/ESnet | A. Shoshani, LBNL |
| M. Beck, U. Tenn. | B. Kahn, CNRI | H. Simon, LBNL/NERSC |
| D. Blumenthal, UCSB | W. Kaplow, Qwest | L. Smarr, UCSD |
| B. Boston, Cisco | B. Kramer, LBNL/NERSC | E. Stechel, SNLL |
| V. Cerf, Google | M. Livny, U. Wisc. | M. Strayer, DOE SC (25) |
| H. Chen, SNLL | B. Maccabe, UNM | T. Thompson, PNNL |
| J. Chen, SNLL | M. McCoy, LLNL | J. Waters, Level 3 Comm. |
| K. Claffy, UCSD | G. Michaels, PNNL | V. White, Fermilab |
| D. Crawford, NSF | R. Mount, SLAC | R. Whitney, T. Jeff. Nat. Accel. |
| S. Elbert, PNNL | J. Myers, NCSA/UIUC | A. Willner, USC |
| T. De Fanti, UCSD | H. Newman, Caltech | B. Wing, ORNL |
| D. Farter, CMU | W. Polansky, DOE SC | T. Zacharia, ORNL |
| A. Geist, ORNL | R. Pordes, Fermilab | ANL Library |
| C. Giancarlo, Cisco | L. Rahn, SNLL | |



Mathematics and Computer Science Division

Argonne National Laboratory
9700 South Cass Avenue, Bldg. 221
Argonne, IL 60439-4844

www.anl.gov



UChicago ►
Argonne_{LLC}



A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC