SAND2008-2790P

CIS External Review 2008
Sudip Dosanjh, Senior Manager
Computer and Software Systems Overview

CONTEXT: This document describes Sandia's research and development in architectures, system software, geometry and meshing, and data analysis and visualization.

I. Architectures and System Software

I. 1 New Initiatives

Sandia has launched two major initiatives in computing this past year, The Institute for Advanced Architectures and Algorithms (IAA) and an Alliance for Computing at the Extreme Scale (ACES) with Los Alamos. The IAA was established by Congressional mandate in 2008 with Centers of Excellence in Albuquerque, NM and Knoxville, TN. Sandia and Oak Ridge will build upon their long history of collaboration (dating back to the Intel Paragon), strong ties to universities, and past successful industry collaborations (e.g., with Intel, Cray, AMD, Micron, and SUN). The IAA is focused on enduring research to overcome the architectural bottlenecks that limit supercomputer scalability and performance. This partnership is a premier example of a National Laboratory – University – Industry collaboration aimed at maintaining our global leadership in Science and Technology and future competitiveness.

The IAA will

- Perform R&D on key impediments to HPC, partnering with industry and academia.
- Foster the integrated co-design of architectures and algorithms
- Train future generations of computer engineers, computer scientists, and computational scientists.
- Deploy precompetitive prototypes to demonstrate technologies and to allow application developers and algorithm researchers to explore these advanced architectures.

The IAA's focus areas include interconnection network technologies, memory systems, supercomputer simulation, hierarchical algorithms, resilience and system software. Metrics for success include furthering DOE's missions in science, national competitiveness and national security as well as impacting the supercomputing industry.

The institute will host industry collaborators, visiting faculty members, graduate students, and Fellows on short-term assignments from other laboratories and government agencies.

The partnership with Los Alamos, ACES, is highly synergistic with the IAA and represents one important customer for the IAA's Research and Development efforts. Under ACES, a joint computer architecture office is being formed. This office will partner with computer vendors to design and develop production capability systems for NNSA applications. The architectures office will determine user requirements at the NNSA laboratories, interact with computer companies to be cognizant of industry roadmaps, develop and issue requests for proposals, conduct proposal selection reviews, and negotiate and execute contracts with supercomputer companies. Once a contract with a company is in place, the office will oversee work, participate in development, and partner to integrate, bring up and debug the supercomputer.

Once a few users are successfully executing large jobs, responsibility for running and managing the machine will be transitioned to the operations team. Supercomputers will be integrated and sited at

LANL's Strategic Computing Complex (SCC). Small initial delivery or prototype systems might be sited at Sandia if needed.

The goals of ACES are to provide capability computing for NNSA and national leadership in high performance computing. Production requires high reliability and usability and capability means that a single application must execute effectively across the entire machine (this places constraints on the time needed for remote memory access). Key strategies include:

- Aligning with and influencing industry roadmaps
- Co-architecting platforms, applications and algorithms recognizing that new architectures are likely to influence applications and algorithms
- Ensuring pragmatic migration of ASC codes to new platforms with significant performance gains
- Encouraging and fostering credible competition in the supercomputing industry (procurements will be open and competitive)
- Actively promoting public standards
- Focusing on a broad range of applications
- Impacting the supercomputing industry through market acceptance of designs and component technologies
- Being driven by cost, risk and benefit analyses
- Partnering with the DOE's Office of Advanced Scientific Computing Research and other government agencies (notably DARPA and other DoD agencies).

The material presented in this section is organized into two major categories. The first category covers our efforts to sustain, deploy or develop production oriented computing systems distinguished particularly by their scale – systems incorporating tens of thousands of micro-processors. The second category covers our research and development efforts targeted at enabling the next generation of such large scale systems, and exploring research themes relevant to computing at the petascale and beyond.

Red Storm, our flagship system straddles this conceptual structure. Hence projects relating to it appear in both categories.

I.2 High Performance Computing Platforms at Sandia

Within Sandia we now have three classes of distributed memory systems for running applications:

- Highest End Capability Platforms. The ***ASC Red Storm*** Supercomputer ("Red Storm") was co-developed by SNL and Cray and forms the basis of Cray's XT3/XT4/XT5 product lines. Red Storm contains 12,960 AMD Opteron dual-core compute nodes (25,920 cores) and has a peak speed of 124.4 Teraflops. It is a vertically integrated, single system architecture and supports both classified and unclassified computing via flexible reconfiguration in the manner of ASCI Red, and is shared among Sandia, Los Alamos and Lawrence Livermore national laboratories. Red Storm uses commodity components and strategically targeted custom capabilities, primarily the Seastar high-speed interconnect fabric, the light-weight kernel (LWK) system software and the fully integrated RAS subsystems. Red Storm entered general availability in March, 2007. Since its delivery in very late 2004, the system has been upgraded several times to increase and/or improve memory, processing, networking and capacity. This year, additional disk storage is being added and 65 of the 135 computation cabinets will be upgraded from dual-core to quad-core Opteron processors.

- Mid-range Capacity Platforms. The ***Thunderbird Cluster*** is a collaboration between SNL, Dell, Intel and Cisco to push the state of the art in scalable commodity Linux clusters. Thunderbird uses Dell servers based on Intel EM64T processors interconnected by Cisco Infiniband network interface cards and switches and leverages the rapidly evolving Linux, Infiniband-based system software stack, in particular OpenMPI and OpenFirmware. The system contains 8,960 processors and has a peak speed of 64.5 Teraflops and is dedicated to unclassified processing. This system has two important roles: 1) to provide the SNL institution with production computing cycles and 2) to foster a long-term collaboration among Dell/Intel/Cisco and SNL on the scalability issues associated with Infiniband clusters. This system entered general availability in March 2006.

- Capacity Platforms. SNL has acquired its capacity computing platforms over the last several years incrementally, based on available funding in any particular FY. Current platforms include: The ***Nuclear Weapons Compute Cluster*** (NWCC) systems are dual Intel EM64T node clusters based on Hewlett-Packard servers and the Myrinet interconnect architecture. There are two 1024 processor systems for New Mexico, one each for unclassified and classified environments, and one 384 processor system for California's classified environment. These systems are used for Defense Programs production capacity computing jobs and they were deployed in September 2004. The ***Institutional Compute Cluster*** (ICC) systems are dual Intel Xeon node clusters based on Hewlett-Packard servers and the Myrinet architecture. They are commercially procured platforms consisting of 1,200 compute processors configured in three physically distinct systems that were deployed in 2003. These ICC systems will soon be retired in order to provide space, power and cooling for a new set of ***Tri-Laboratory Linux Capacity Cluster*** (TLCC) systems. TLCC is currently under competitive procurement with initial deliveries targeted in 2008.

Sandia's computational systems need to be able to address a very broad spectrum of problem sizes – from smaller jobs of less than a few hundred processors that turn-key cluster systems can address, to jobs up to and beyond ten thousand processors. The ASC platform strategy is evolving and while performance is still a top driver, budget realities are also forcing a careful consideration of costs. The ASC strategy is to focus TLCC procurements on minimizing their total cost of ownership. Turnkey commodity Linux clusters will provide the lowest cost for servicing the computational demands of the lower end of our applications workload. The success of Sandia's Thunderbird system demonstrates the ability of the open source software community to address scalability issues with a minimal level of development effort on the part of Sandia and ASC. The rationale is to preserve our R&D resources to focus on Capability and Advanced Architectures technologies.

Sandia's three classes of production computing platforms provide essential infrastructure by delivering the cycles needed to conduct our applications work. They also impose or encourage a certain approach to the development and use of these applications, and thereby strongly influence the enabling technologies agenda and approach. A brief account of the status and direction of our efforts associated with these platforms follows. The numbering of the underlined topics below constitutes a reference system to correlate these descriptions with the project funding Table 2 and interaction matrix presented elsewhere in the review.

I.3 Architectures and System Software Research and Development

Introduction

A variety of R&D efforts are underway with multiple purposes:

- Foreseeing possible problems and growth areas in Red Storm, and proactively addressing them.
- Exploring possible architectures for practical petascale computing by 2010.
- Laying the foundation through research efforts for computing at yet higher scales.
- Investigating alternate technology such as Quantum Computing.

Research Project Synopses

1. Red Storm System

Red Storm is DOE/NNSA and Sandia's highest end capability MPP. This project is led by Jim Tomkins and Sue Kelly. It is currently in general availability and is also the subject of further joint development with Cray Inc. It is a distributed memory, MIMD, message-passing, general-purpose supercomputer. Sandia developed the initial architecture and specification for Red Storm and continues to engage with Cray on aspects of both hardware and software design of the system. The Sandia/Cray alliance resulted in a commercially-successful supercomputer. There are currently 35 systems deployed around the world at 20 different sites.

From a system software perspective, it has been a quiet year for Red Storm. We qualified multiple OS versions, but these were all bug-fix releases and provided very little new functionality. This is a good thing and provided a stable environment for the heavy production usage this past year. We did, however complete the integration of the Portals protocol offload software/firmware into the network interface chip. This software will be delivered later this year in V2.1 of the Cray software environment.

As mentioned above, Red Storm is being upgraded during FY08. New/additional disk drives are being added, increasing storage capacity from 340 TB to about 2 Petabytes total. Slightly less than half of the compute processors will be upgraded from dual core to quad core Opterons. Additional memory will be installed to provide 2 GB per processor core. The following table shows the upgrades to the system since its initial delivery in 2004. The easy with which this machine can be upgraded is a testimony to its well-conceived architecture.

Table 1: Comparison of Red Storm Specifications through each Upgrade

Attributes	Red Storm 04	Red Storm 06	Red Storm 08
Full System Operational	December 2004	October 2006	September 2008 (target)
Compute Nodes' Theoretical Peak	41.47 TF	124.42 TF	284.16 TF
MP-Linpack Performance	36.19 TF	101.4 TF	TBD
Architecture	Distributed Memory MIMD	Distributed Memory MIMD	Distributed Memory MIMD
Number of Compute Nodes	10,368	12,960	12,960
Processors per Node	1	2	2 (on 6720 nodes) and 4 (on 6240 nodes)
Number Service & I/O Nodes	256 (each end)	320 (each end)	320 (each end)
Processor	AMD Opteron @ 2.0 GHz	AMD Dual Core Opteron @ 2.4 GHz	AMD Opteron (Dual cores at 2.4 GHz; Quad cores at 2.2 GHz)
Compute Node Memory	10.4 TB	39.19 TB	75 TB
Memory per Compute Node	1.0 GB	2-4 GB	4-8 GB (2GB per core)
System Memory B/W	57.9 TB/s	78.12 TB/s	123.14 TB/s
Disk Storage	120 TB each end	170 TB each end	440 TB - 1.67 PB
Parallel File System B/W	50 GB/s each end (targeted)	50 GB/s each end (targeted)	50-60 GB/s each end (targeted)
External Network B/W	25 GB/s each end	25 GB/s each end	25 GB/s each end
Interconnect Topology (X,Y,Z)	3-D Mesh, 27 x 16 x 24	3-D Mesh, 27 x 20 x 24	3-D Mesh, 27 x 20 x 24
Interconnect Performance			
MPI Latency	6.59 □s 1 hop, 9.59 □s max	4.78 □s 1 hop, 7.78.0 □s max	TBD
Bi-Directional Link B/W	9.6 GB/s (3.2 GHz LVDS)	9.6 GB/s (3.2 GHz LVDS)	9.6 GB/s (3.2 GHz LVDS)
Minimum Bi-Section B/W	3.69 TB/s	4.61 TB/s	4.61 TB/s
Full Machine RAS System			
RAS Network	100 Mbit Ethernet	100 Mbit Ethernet	100 Mbit Ethernet
RAS Processors	1 for each 4 CPUs	1 for each 4 CPUs	1 for each 4 CPUs

Operating Systems Service and I/O Nodes Compute Nodes RAS Nodes	LINUX Catamount (Cougar) LINUX	LINUX Catamount Virtual Node LINUX	Linux Catamount N-way Linux
Red/Black Switching (proc. counts)	2688 – 4992 - 2688	6720 – 12480 - 6720	6720 - 24960 - 6720
System Foot Print	~ 3000 sq ft	~ 3500 sq ft	~ 3500 sq ft
Power/Cooling Requirement	1.7 MW	2.5 MW	2.6 MW
Software Tools Debugging Performance Modeling	TotalView Apprentice	TotalView Apprentice	TotalView Apprentice

2. Catamount Risk Mitigation

Closely aligned with the Red Storm research project is work funded by the Office of Science to modify Catamount to support quad-core Opteron processors. Led by Sue Kelly, we extended the virtual node concept to generically support ***N*** cores per socket where *N* can be set to 1, 2, or 4. The code is designed to support higher values of ***N***, but these are currently untestable. Again, the memory is equally divided among each CPU core and they must share the NIC. By combining this work with the NIC-based protocol processing work, we expect to retain reasonable system scalability even with 4 cores on each node. This project completed in March, 2008. Delays in quad core availability allowed the software to be tested on only 4 quad-core processors. The results are encouraging with eleven different applications successfully tested. The NIC-based protocol processing was introduced as planned. At this scale, it is not possible to make any general statement about scalability.

3. A Light Weight Operating System for Multicore Capability Class Supercomputers

Led by Kevin Pedretti, the goal of this project is to create an open-source Lightweight Kernel (LWK) operating system (OS) with Hypervisor functionality for capability class supercomputers made up of multi-core processors. In order to accelerate development and provide a more familiar environment for collaborators, the new LWK, called Kitten, is being built from the ground up and is leveraging Linux kernel code and structure where appropriate. Critical LWK subsystems such as memory management are being custom developed. The current open-source (GPL) alpha release boots on Red Storm systems as a drop-in replacement for Compute Node Linux (CNL). A networking stack is currently under development. Our goal is to implement the necessary portions of the Linux ABI so that CNL applications, including those using OpenMP, will run unmodified under the new LWK. Having an open source LWK platform is already starting to be beneficial. We have begun working with external collaborators Patrick Bridges (University of New Mexico) and Peter Dinda (Northwestern University) to incorporate a lightweight hypervisor that they are developing as part of a NSF-funded project into the Kitten codebase.

A significant part of this project is focused on understanding current and future multi-core architectures, emerging programming models, and exploring new OS capabilities to better support this new environment. We are currently pursuing a novel intra-node shared memory mechanism that allows direct address-space to address-space copies without kernel intervention that will allow more efficient MPI, PGAS, and OpenMP implementations. We are exploring in conjunction with Mike Heroux (SNL) ways to leverage this capability to support mixing MPI and threads models in the same application. Another

OS area we are exploring is the virtual to physical memory mapping. One counterintuitive disadvantage of the physically contiguous virtual memory mapping provided by Catamount is that it can potentially lead to low performance for applications that stream through several large arrays with unit strides. We have identified this to be due to conflicts in the memory subsystem for particular access patterns. A more random virtual to physical mapping, such as what Linux provides, avoids this problem at the cost of best-case memory performance and more complexity in the network stack. We are developing a LWK mechanism to deterministically interleave the memory mapping, which we hope will retain the current network stack simplicity and demonstrate increased performance.

4. Portals Research and Development

This project, led by Ron Brightwell, addresses research in several areas of low-level network programming interfaces: (a) collective optimizations for the Red Storm network, with particular emphasis on increasing the performance of applications whose performance and scaling are dependent on collective communication, (b) implementation and analysis of non-blocking collective communication, (c) implementation and analysis of one-sided communication operations, e.g. remote-memory gathers and persistent broadcasts, (d) lightweight mechanisms for supporting atomic memory operations needed by global address space programming models. This year we have also been exploring operating system enhancements to enable more efficient intra-node data movement. We have enhanced the Catamount lightweight kernel to allow for several data movement optimizations, including lowering intra-node latency by 81% and increasing intra-node small message rate by more than 5x. We expect similar improvements to collective operations on quad-core processors, and we anticipate this new capability will be a significant win for the upcoming quad-core upgrade to Red Storm. We have developed prototype implementations of MPI using this new capability. This year we also completed an initial prototype implementation of the next-generation Portals API on Red Storm. This was a proof-of-concept implementation using modified SeaStar firmware and was the focus of a quick start CRADA with Intel. We also continued to explore InfiniBand scalability issues by starting development of an implementation of the Open Fabrics API on top of Portals to allow for an in-depth analysis of the scalability limitations of this API.

5. Scalable I/O Research and Development

Led by Lee Ward, this project endeavors to enhance the IO software stack in the supercomputer. At present, there are many ongoing activities:

- The SYSIO library is a POSIX-like, user-level virtual file system with many extensions for high performance IO. While research and work is ongoing, stable versions continue to be integrated by Cray and are deployed at all production Cray XT3 sites.
- As part of an Institute funded by the DOE Office of Science SciDAC program, we are attempting to instrument the SYSIO library in order to profile applications. This work will provide file systems researchers with critical application IO traces at scale.
- A collaborative effort with Los Alamos National Laboratory, Argonne National Laboratory, Oak Ridge National Laboratory, and Ohio Supercomputer Center has been funded by DOE Office of Science FAST-OS program will fund an effort to address predicted IO limitations with Petascale and Exascale machines. Credible projections of the number of components required by peta-scale machines, and beyond, suggest that current and known file system solutions will be unable to directly service the clients. The contemplated solution draws heavily on the Sandia Systems' group history in light-weight designs but targets all current solutions from IBM's BlueGene, to Sandia's SUNMos/Puma/Catamount, to Linux.
- A collaboration with researchers at Sandia's Livermore CA site, funded by a significant LDRD grant, scheduled to begin at the start of FY '08, will allow us to investigate and mitigate

deficiencies in the Linux operating system that manifest as very slow write-IO operations when remote DMA (rDMA) methods are used. Solving this problem is critical to a successful future for the NFSv4 parallel and direct IO extensions or other rDMA-enabled third-party transfers using very high speed network interfaces.

- In collaboration with many researchers from multiple DOE labs, universities, and industry, we are working on a small set of proposed extensions to the POSIX standard that will allow portable access to efficient, scalable IO services for distributed groups of cooperating processes. Prototype implementations of a few of the proposed enhancements show promise and have been accomplished by a joint effort between Sandia and Argonne National Laboratory.
- A long, background, effort to restart significant research into high-speed, parallel IO stacks in academia has borne fruit. We are now a part of an advisory group to the High End Computing University Research Alliance's Interagency Working Group (HECURA/IWG). In FY 07, over eleven million dollars of multi-year research monies were awarded by the National Science Foundation to academic institutions. We helped to create and run a workshop from which was crafted the NSF call and have visited the newly funded institutions in order to link the academic research to real world issues. This work is ongoing and other NSF calls have and will include our suggestions for issues and goals.

Credible projections of the number of components required by peta-scale machines, and beyond, suggest that current and known file system solutions will be unable to directly service the clients. In collaboration with Los Alamos National Laboratory, Argonne National Laboratory, Oak Ridge National Laboratory, and the Ohio Supercomputer Center we have architected a potential solution. This solution draws heavily on the Sandia Systems' group history in light-weight designs but targets all current solutions from IBM's BlueGene, to Sandia's SUNMos/Puma/Catamount, to Linux. We are avidly pursuing funding in order to craft the implementation and investigate impact on performance and scalability in a petaflops environment.

6. Light Weight File System

Led by Ron Oldfield, in collaboration with Arthur B. Maccabe at the University of New Mexico, the Lightweight File Systems (LWFS) project takes a fresh approach to file system services guided by the same philosophies that generated Sandia's lightweight operating system series. The LWFS project is not a risk-mitigation effort for currently planned systems. It is a long-term research effort that targets expected I/O problems in next-generation architectures. An early prototype exhibits scalability comparable to best-of-breed solutions, and researchers outside of Sandia are interested in adapting the scalable security design in their file system efforts.

The key features of the LWFS are a strong and scalable security model and a service-oriented model that allows extreme flexibility with respect to policies for metadata management and application-specific I/O optimizations. This flexibility led to a number of innovations in how to perform I/O for current and next-generation MPP systems. In particular a paper accepted by IEEE MSST describes specific optimizations to dramatically improve performance of checkpoint/restart operations. Another paper in preparation for USENIX FAST conference illustrates through experiments on Red Storm how to significantly reduce traffic, by an order of magnitude or more, to centralized metadata servers. Such issues are critical in performance

This past year focused on performance tuning and scalability analysis on Red Storm. Analysis of large-scale results (up to 10K nodes) is still underway, but early results show I/O performance achieving nearly 80% of peak hardware rates for standard I/O benchmarks. Detailed trace analysis of I/O server behavior shows potential for even better performance. Application performance analysis of CTH and Sage, using the libsysio LWFS driver, is underway.

We are also continuing to develop key functionality with a number of academic partners at the University of New Mexico, Georgia Institute of Technology, and the University of Texas at El Paso. In particular, we are developing support for distributed transactions, overlay networks, off-line support for manipulation of raw storage data, and support for InfiniBand networks.

7. High Performance Computing RAS R&D

Led by Jim Laros, this project has researched microcontroller capabilities and programming methods as they relate to both a generalized RAS (Reliability, Availability, Serviceability) system implementation and to increase our knowledge of the Red Storm RAS implementation. This research has directly impacted our ability to be successful with Red Storm, during both the initial Risk Mitigation effort and the IO Risk Mitigation effort. Resulting research from this effort has fed numerous projects and proposals to date. Recent work in this area has focused on the emerging issue of power consumption of advanced architectures specifically multi and many-core architectures.

8. Increasing Supercomputer RAS via Informatics

This effort is led by Jon Stearley and is aimed at utilizing informatics to proactively isolate the root cause of system problems in large parallel computer systems. With the specific goal of increasing supercomputer RAS, we are working to create a machine-learning system which enables content-novice analysts to efficiently understand evolving trends, identify anomalies, and investigate cause-effect hypotheses in large multiple-source log sets. In 2008, the Thunderbird and Spirit systems were setup to use Sisyphus (Red Storm began using it in 2007 and use is ongoing). Jeff Ogden (9328) has extended the tools to perform additional unique analyses (e.g. submit a query for "the most informational logs produced during job id X"). Los Alamos has deployed Sisyphus on one of their systems, and is including it in their new monitoring system to be deployed on Roadrunner and all systems thereafter. Sisyphus was presented at a Red Storm quarterly meeting, an invited presentation at IBM TJ Watson research center (http://www.research.ibm.com/pac2/), and an invited presentation at Oak Ridge National Laboratory. Collaboration with Stanford has continued, resulting in two publications this year: "Bad Words: Finding Faults in Spirit's Syslogs" (IEEE International Symposium on Cluster Computing and the Grid), "Alert Detection in System Logs" (ACM SIGKDD International Conference on Knowledge Discovery and Data Mining) - co-authored with Adam Oliner (PhD student) and Alex Aiken (Faculty). The logs analyzed in these papers have been downloaded by a dozen institutions (8 academic, 2 industry, 2 national laboratories), and the Sisyphus tools have been downloaded 85 times between July 1, 2007 and March 19, 2008, bringing the total number of downloads to 365).

9. Defining and Measuring Supercomputer Reliability, Availability, and Serviceability (RAS)

Also led by Jon Stearley, the goal of this TriPOD (SNL, LANL, and LLNL) effort is to drive HPC RAS improvements that enhance system productivity. A draft specification for the definition and measurement of RAS for the TLCC systems was written, but work towards implementing it has been blocked by more operations-critical issues regarding Moab. TriPOD reorganization has resulted in this group being a sub-group of the TriPOD Monitoring and Metrics working group (for which Bob Ballance is the SNL lead). Efforts on TLCC RAS metrics may resume when hardware is delivered in mid-2008, depending on the priorities set by Bob Ballance.

10. Advanced Systems

Led by Doug Doerfler, its purpose is to investigate new high performance computing architectures and system solutions addressing future SNL and NNSA platform needs. It will achieve this by developing strategic industrial, government and academic partnerships and leverage the widely recognized leadership SNL has provided in high performance, scalable architectures. The programs current focus includes: a) investigations into chip level multiprocessing architectures and their applicability to SNL's application workload, b) investigations into advanced memory subsystems, led by Rich Murphy and c) researching alternative high speed networking technologies, led by Scott Hemmert. These efforts involve interactions with numerous industry and government agencies. This work has formed the basis for the DOE Institute for Advanced Architectures project described elsewhere in this document.

11. Scientific Performance Modeling for Application and System Characterization

Led by Doug Doerfler, this effort has an objective to develop models of application performance that tie back to characteristic system performance parameters. This effort measures and models the performance of Sandia's most used application codes on MPP, cluster and vector architectures and draws conclusions regarding the cost effectiveness of these architectures as a function of problem scale, thus providing feedback to program managers and analysts on how best to map their computing demands onto our portfolio of available computational resources. In concert with other Sandia research teams we are pursuing the development of compact applications that are representative of major Sandia applications that are complex and difficult to port and benchmark on new systems. We also seek to understand architectural issues, such as multi-core compute nodes, to mitigate their impact on current applications and architectures, and minimize their impact on future applications and architectures. These results are used to improve application, solver, and library performance at large scale and are prominent in discussions internal to DOE on acquisition strategy. Knowledge gained from this effort will also help form the basis for interactions with potential strategic partners and form the technical requirements for future platform acquisitions. This effort has become a project of Sandia's Institute for Advanced Architecture (IAA).

12. Structural Simulation Toolkit (SST)

Led by Arun Rodrigues, since he worked at Sandia as a student intern, the Structural Simulation Toolkit (SST) is evolving into a strategic capability. For several years, Sandia has been building SST into an infrastructure for architectural evaluation and exploration. The Structural Simulation Toolkit is a modular and extensible tool which has been used by several CSRF and LDRD projects to explore network, memory, and processor architectures as well as perform application characterization. The SST has been used to simulate advanced network and MPI enhancements, microprocessor modifications to support scientific applications, Processor-In-Memory systems, and application studies in support of the advanced systems project. The SST has been used to foster microprocessor research with the University of Wisconsin, systems and power research with the University of Notre Dame, transactional memory research with Oak Ridge National Laboratory, compiler research at Rice University, and programming model research with Louisiana State University.

Recently, the SST has been expanded to include a SeaStar module by Michael Levenhagen to model the Red Storm interconnection network. The code base has been released under an open source license to foster external collaborations. Under the CSRF Next Generation Systems support, Will McLendon has ported the SST to 64-bit x86 Linux and added a regression test suite, as well as improving portability. Daniel Barnette is working to improve the documentation as well as creating a new website. Uzoma Onunkwo is working to integrate the PTLSim x86 simulator into the SST framework.

A collaboration with Georgia Tech has produced a prototype parallel version of the SST. We plan to expand on this work to create a parallelized version of the simulator capable of simulating thousands of nodes on hundreds of real-world nodes. Additionally, a collaboration with the University of Maryland has begun to improve the DRAM subsystem by incorporating the DRAMSim tool into the SST framework.

13. Supercomputer System Design through Simulation (Seshat)

Led by Rolf Riesen, the goal of this project is to enable design of a supercomputer by modeling architecture, interconnect, runtime environment, and applications. In the process we are creating analytical capabilities to address various issues. Current capabilities are embodied in a simulation code named Seshat. It runs an application inside a virtual-time framework and simulates the message-passing characteristics of a system. Its output is timing and quantitative information about the message-passing behavior of an application, such as how many messages and bytes were sent from node A to node B. Its focus is on understanding the impact of machine features on application performance. The parameters describing the machine can be easily modified. This allows answering questions like: What is the application impact of doubling the network bandwidth or speed at which a collective operation can be done. It is easy to adapt to new machine parameters. No knowledge of application is needed; only object code to link against. It provides quick answers within the running time of the application on a given problem. However, it is not as detailed a simulation of microprocessor components or network interface as SST (described elsewhere in this report).

An effort is underway to test cross-platform modeling; i.e., simulate the Red Storm network on the Thunderbird cluster. Since the CPU and memory architecture of these two systems differ, Seshat has to compensate for the discrepancy. We are currently investigating whether a simple speedup factor will let us roughly model the performance of a Red Storm node on Thunderbird.

Working with a Ph.D. Student at UNM, we are using Seshat technologies to simulate a PowerPC cluster using a model of our Red Storm network. We use the Mambo PowerPC cycle-accurate simulator and run several instances of it on nodes of a PC cluster. Because we are modeling the network instead of simulating it, this approach scales. We are working on a paper describing this effort and are planning large scale experiments this summer.

14. Advanced Interconnect Research

This research, led by Scott Hemmert with participation by Arun Rodrigues, Brian Barrett, Kevin Pedretti, and Ron Brightwell, focuses on network interface architecture to handle bandwidths of over 30 GB/s (per direction), latencies of under 750 ns through the MPI layer, and a message rate of 10 to 12 million messages per second per direction for each acceleration engine in the NIC. Such capabilities are required to enable scalability on next generation MPP platforms. In past years we have developed individual programmable hardware units (associative list processing unit, list management unit and microcoded match unit) to accelerate MPI matching. Faster matching improves both messaging rate and MPI latency. We are currently working to simulate the combination of the individual hardware components into a complete “queue processor”. Concurrently, we are designing send side enhancements to allow message injection rates to match the matching rates of the queue processor. We are also proceeding with smaller subprojects which are researching the impact of network bandwidth and latency on application performance, as well as work on understanding the best way to support the rendezvous protocols which are typically used for large messages. The data from these studies will allow us to further refine the NIC architecture.

The past year has also seen continued interest from industry in this work. Specifically, final details for a non-exclusive license are being negotiated with Intel for inventions developed as part of this research. These inventions include the associative list processing unit (ALPU), which allows MPI matching to proceed in parallel for a small number of entries in the MPI lists, and the list management unit and match accelerator, which are used to accelerate matching for list items which do not fit in the ALPU. In addition, Keith Underwood, the former lead on this project, is taking a leave of absence from Sandia to work at Intel where he is working to commercialize some of the technology from this project under a license agreement between Sandia and Intel. Keith's work is expected to be just one activity under a much broader long-term collaboration to develop a next generation supercomputer for the 2010-12 timeframe. Other companies have also expressed interest in these technologies and we are currently working with them to determine how they might be incorporated into future products.

15. Storage-Intensive Supercomputing (SISC)

In support of Lawrence Livermore National Laboratory's storage-intensive supercomputing (SISC) effort, Craig Ulmer at SNL/CA is investigating how new, solid-state storage technologies can be utilized to provide significant improvements in large-data applications. The Sandia portion of this work focuses on adapting sparse graph algorithms to take advantage of the unique performance characteristics of emerging mass-storage products that employ flash memory. In addition to having random access times that are an order of magnitude better than traditional disks, these mass-storage products utilize FPGA hardware that can be reprogrammed with application-specific hardware. As such, we are developing a flash-memory controller for the FPGA that is customized to the data access patterns found in graph applications. These customizations will enable us to shave overheads and increase performance.

In 2007, Sandia joined the Netezza Developer Network. The Netezza system is designed to provide efficient parallel "Structured Query Language" queries over distributed databases for business applications. At the 2007 Netezza User Conference, Netezza announced their intention to support development of non-SQL analysis of large-scale distributed datasets, and backed this up with a competition to grant 4-node systems to developers with interesting new data analysis applications. Sandia was granted a development system to develop analysis applications to make quantitative comparisons of scientific simulations with experiments and among ensembles of simulation results. This development system will be a useful complement to the 56-node system we already have and formalize interactions with Netezza's applications development team.

16. Enhancing microprocessors to support scientific applications (LDRD funded) completed last FY

This project, also initiated by Keith Underwood, is now led by Arun Rodrigues with participation by Richard Murphy, seeks to enhance processor micro-architecture to better support Sandia's applications. With each generation of microprocessor, it has become increasingly challenging to sustain performance with scientific applications. Analysis of Sandia's code base reveals that our applications' integer and floating point usage is significantly different than many industry standard benchmarks. This research currently has two aspects - an effort to re-architect the floating-point unit to provide better performance (including collaboration with the University of Notre Dame) and an effort to enhance integer unit performance (as part of a collaboration with the University of Wisconsin). We have already demonstrated a new class of instructions which allows floating point instructions to better use cache and memory bandwidth, yielding performance improvements of up to 20%. We are currently refining ideas for a new complex floating-point unit as well as exploring ideas for accelerating integer instructions.

17. Multi PetaFlops Supercomputing

Led by Erik DeBenedictis, this cluster of projects leads the trend to multi-core computers and beyond. It is now evident that microprocessor clock rates have flat-lined and that power issues will become more important until halting progress in computer performance in a couple decades. The objective of these efforts is to identify architecture and software R&D efforts for the final phases of CMOS-based computing and to start research on CMOS replacements and their implications.

- Zettaflops Activities
  Erik and Thomas Sterling of LSU have produced two "Zettaflops" workshops, the latest in October 2007. These would appear to be the most future oriented workshops that engage the computational science user community. By sponsoring the 2005 and 2007 Zettaflops workshops, Sandia has become one of the lab supporters of emerging Government research initiatives such as E3, DARPA Exascale, and NSF EMT. The 2007 Zettaflops workshop has resulted in several speaking follow-ons for Erik on the future of computing. There is a paper and a monograph planned.

- Embedded and Space Computing
  The Zettaflops activities have gathered quite a bit of interest from mobile computing (e. g. spacecraft, UAVs, avionics) and an embedded equivalent is planned for May 2008. These mobile computing applications have similar demands as supercomputers in terms of power efficiency and architecture, although they would deploy systems of much smaller absolute size. Erik is working with Sandia space systems organizations and center 1700 (which has expertise in rad hard electronics) to leverage 1400's expertise in parallel systems to this related application area.

- Nanotech
  We also participate in the cross-Sandia Nanotech activities with the MESA center. Our effort is to understand the architectural and computer reliability implications of emerging nanotech devices. Past efforts have been in quantum dots and nanowires. There is current set of activities raising the maturity level of Carbon Nanotube research at Sandia to become a capability in the MESA fab. 1400's contribution is to assist in integrating physical sciences research and represent the digital electronics application area.

- Quantum Information Program (LDRD & Externally funded)
  In the last three years, Sandia has gone from a standing start to having a $5M+/year program in Quantum Computing. Now a program, it was originally a grassroots initiative started with coordinated projects in physical sciences and theory. The physical sciences efforts were centered on the development of ion trap and solid state quantum dot quantum computer components, both in center 1700. Our Center (1400, CCIM) contributed the connection to applications in cryptanalysis and physical science simulation and the architecture of quantum computer components (e. g. Quantum Error Correction). In the last year, 1400 has added the use of supercomputers to simulate quantum computer components (electron wave functions and NEMO) under the direction of Rich Muller. The LDRD Grand Challenge initiated in this FY by center 1700 yet based on an equal mixture of theory and experiment. External WFO funding has begun to flow in the last year also with a 50-50 mixture of goals between experiment and theory.

- Extending Moore's Law
  Led by Erik, we have several projects aimed at understanding the limits of Moore's Law for CMOS, and devising strategies to assure continuity in the growth of computing potential. To this end, Erik is organizing a second "Zettaflops" workshop for October 2007. This workshop will develop a technical vision for post-Petascale computing that joins emerging exascale initiatives

from DOE, DARPA, NSF, and other agencies. Of these, we also participate directly in DOE's Exascale for Energy, Ecological Sustainability and Global Security (E3SGS).

Erik is currently an organizer of the architecture section of the Emerging Research Devices (ERD) section of the International Technology Roadmap for Semiconductors (ITRS). This is the officially sanctioned roadmap for the future of Moore's Law for industry and Government. This committee membership offers Sandia a forum to influence industry direction and feeds back to the Sandia community an inside look at technology trends.

We also participate in the NNEDC Nanotech initiative. This is a joint physical and computer sciences activity with participants in the MESA center. Our effort is to understand the architectural and computer reliability implications of emerging nanotech devices.

18. Programming Model Research

Led by Zhaofang Wen, programming model research at Sandia is complementary to other efforts in the HPC community. We are in collaboration with Syracuse University and Oak Ridge National Labs. We take an evolutionary approach towards advanced parallel programming models for future architectures. As a first step in this approach, we introduced the Bundle-Exchange-Compute (BEC) model, focusing on two issues un-addressed by the HPCS and GAS (Global Address Space) communities: (1) smooth migrations of the legacy MPI applications and their programmers, and (2) efficient support of unstructured applications, especially those that involve high-volume, random, fine-grained communication. BEC provides a virtual shared memory (a.k.a. global address space (GAS)) environment which automatically and dynamically bundles up (unstructured) fine-grained messages for coarse-grained communications. BEC includes a light-weight runtime library on top of the message-passing layer (e.g., MPI), and optionally a minor language extension (e.g., C) for the syntax of shared data.

Using a BEC prototype, we completed a few unstructured applications and demonstrated clear advantages over using MPI alone. For example, in a linear system solver using the Conjugate Gradient (CG) method for sparse matrices, compared with an MPI version highly-tuned by an expert programmer, the BEC version achieved same performance and scalability, but with much simpler code. (For example, in the communication part, BEC has 11 lines of code and MPI has 277 lines.) We released BEC Version 1.1 to a few Sandia users last Dec.; and we plan to release BEC Version 1.2 to the users at Sandia and some other DOE labs in April 08. We are developing more applications using BEC. Going forward, we will enhance the BEC model and extend the idea to a broader scope, including the development of a programming model for the multi-core for HPC applications.

19. PIM (Processor In Memory) Simulation Project (LDRD Funded)

Led by Richard Murphy, the Processing-In-Memory (PIM) project is investigating the use of very simple, highly threaded processors near to DRAM memory for accelerating two classes of applications of relevance to Sandia: first, traditional scientific and engineering applications that represent the core of our work; and second, emerging informatics problems that are extremely challenging for conventional MPP supercomputers to execute at scale. These chip architectures are being examined in homogeneous all-PIM system configurations and augmenting a conventional processor in a "PIM-enhanced" Red Storm-like architecture. In the first year, we have shown that both classes of applications are memory bound in performance, and that the emerging informatics codes are even more memory bound than traditional applications. We have quantitatively compared these to the standard SPEC benchmark suite used by architects and shown that SPEC demonstrates significantly higher spatial and temporal locality, and an almost trivially small data set size. The proposed PIM architecture has also demonstrated over an order of magnitude performance improvement at significantly lower clock rate than conventional machines for the informatics codes. This simulation has also beaten a version of the Cray MTA scaled to higher clock rate and memory bandwidth. Finally, we have developed a strong collaboration with Micron to examine a new type of PIM implemented by 3D stacking of DRAM on top of a logic chip. This architecture leverages what we believe could become a commodity DRAM part, allowing for higher-yield, lower-cost PIM implementations than what have previously been explored elsewhere. This collaboration has impacted other on-going advanced memory efforts at Sandia.

20. Infiniband Development for High Performance Computing

This project is now led by Curtis Janssen in SNL-CA. The OpenFabrics and InfiniBand overview presented by Matt Leininger during the 2007 CIS EAB review provided an example of how active engagement between the DOE/NNSA laboratories and industry can change the high performance computing market. This work was essential to the development of a robust and efficient InfiniBand software stack for HPC, and now that OpenFabrics is a self-sustaining success, we have embarked on new paths on which we at SNL/CA, in unison with SNL/NM, can impact high performance computing. There are two primary direct follow-on to the InfiniBand work. The first is a set of studies comparing 10 Gigabit Ethernet to InfiniBand. A particular feature of this work is the use of 10GigE switches that support adaptive routing. We have demonstrated that the routing algorithms employed are superior to the oblivious routing method used in InfiniBand. This work is done in conjunction with John Naegle at SNL/NM. The second follow-on project is the development of new architecture simulation technologies. Work in this area grew out of the need to demonstrate to InfiniBand vendors that congestion control and adaptive routing would improve application performance. Departing from traditional network simulation techniques using statistical network traffic patterns, we have begun implementing a simulator that can generate traffic patterns based on data gathered from actual applications. It is not specific to InfiniBand and it will impact capability in addition to capacity computing. The simulator allows long run times on a large number of simulated processors and nicely complements the Sandia SST clock cycle accurate simulator. SNL/CA is engaged with SNL/NM and the Institute for Advanced Architecture and Applications to develop and integrate our simulator and other technologies into a complete multiple scale/fidelity architecture simulator.

21. Scalable Programming Tools

Led by Curtis Janssen, this project is concerned with providing programmers tools to aid them in making their code reliable and efficient. These tools will be designed to be useful for understanding and utilizing performance data as well as for collecting application data for use in architecture design and tuning. Programming tools for the large scale machines have unique requirements including coexistence with MPI, scalability to tens of thousands of processors, and operability on ASC platforms, which, in some cases, run operating systems unique to high performance computing platforms. This work is done in collaboration with LANL, LLNL, as well as external parties. Major near-term deliverables include development of tools that provide programmers with intuitively interpretable information about application performance and adaptation of performance tools to be an aid in architecture simulation.

22. Research and Application of Stochastic Approaches for RAS

Research and Development of algorithms and frameworks for parallel, scalable intelligent monitoring and analysis of large-scale computational clusters is led by Jim Brandt and Ann Gentile. This effort is based on research and development of statistical and stochastic methods for anomaly detection within an ensemble of statistically-similar devices (e.g., compute nodes, network elements). The core of this approach relies on characterizing normal behavior of these devices via automatic learning of probability distributions for device state variables, which accounts for non-homogeneous machine environment and allows identification of abnormalities as events with low probability. This methodology enables proactive cluster management through earlier warnings and more precise characterization of cluster state than traditional monitoring methods. A U.S. patent application on this new approach was submitted in 2006. We have developed OVIS, a framework supporting this methodology; it was released as a serial version in 2006. A parallel version addressing scalability and robustness to failures will be released later this FY.

Table 2 below summarizes the costs by project consistently with the numbering scheme used throughout this document. Note that the FTE counts are determined from expenditure using an average loaded labor rate for technical staff, or, in the case of Red Storm project management a higher average labor rate more reflective of the senior personnel involved.

Table 2: Project Cost Summary

		Funding ($M)				
Project/Program Area	**FTEs**	**ASC**	**ASCR**	**CSRF**	**LDRD**	**Total**
1a. Red Storm Program Management	0.8	0.25				**0.25**
1b. Red Storm System Software Development	3.0	1.00				**1.00**
2. Catamount Risk Mitigation	1.0		0.33			**0.33**
3. A Light Weight OS for Multicore Capability Class SC	1.5			0.10	0.36	**0.46**
4. Portals Research and Development	2.4	0.56		0.25		**0.81**
5. Scalable I/O Research and Development	3.0	0.65	0.32		0.17	**1.14**
6. Light Weight File System	2.0			0.30	0.30	**0.60**
7. High Performance Computing RAS R & D	0.5	0.20				**0.20**
8. Increasing Supercomputer RAS via Informatics	0.25	0.10				**0.10**
9. Defining and Measuring Supercomputer RAS	0.25	0.08				**0.08**
10. Advanced Systems	2.0	0.30		0.38		**0.68**
11. Perf Modeling for Application & System Characterization	2.2	0.75				**0.75**
12. Structural Simulation Toolkit (SST)	1.8	0.17		0.20	0.20	**0.57**
13. Supercomputer System Design through Simulation (Seshat)	0.3			0.15		**0.15**
14. Advanced Interconnect Research	1.2				0.40	**0.40**
15. Storage-Intensive Supercomputing	1.5	0.30		0.20		**0.50**
16. Enhancing microprocessors for scientific applications	0.5			0.05	0.10	**0.15**
17. Multi Petaflops Supercomputing	1.0				0.44	**0.44**
18. Programming Model Research	0.5			0.15		**0.15**
19. PIM (Processor In Memory) Simulation	1.5			0.22	0.25	**0.47**
20. Infiniband Development for High Performance Computing	1.5	0.45				**0.45**
21. Scalable Programming Tools	3.0	1.0				**1.00**
22. Research & App of Stochastic Approaches for RAS	2.0	0.6				**0.60**
	37.8	**7.43**	**1.88**	**2.16**	**1.73**	**13.20**

II. Geometry and Mesh Generation

II.1 Introduction

Sandia has a long history of research and development in geometry and mesh generation. Beginning in the early 1990's with the first automatic unstructured quadrilateral mesh generation algorithm, Paving, and development of the CUBIT Geometry and Mesh Generation Toolkit, the program has grown to be a world leader in this area. The Computational Modeling Sciences (CMS) Department at Sandia currently employs 7 FTEs, 7 contractors, 3 student interns and an active research program at 2 universities. The primary strategic objectives of the group include the following

1) **Decrease time to analysis**: The principal bottleneck in the computational modeling process remains the geometry and meshing process. Through the research and development activities of the department, we address specific solutions and algorithms that attempt to stream-line this process and reduce the time it takes to provide a suitable model for analysis. This includes development of new automated geometric tools for preparing CAD models for meshing, new algorithms for automatic mesh generation, mesh quality improvement and modification through automated and interactive tools as well as improved interface design to reduce the learning curve for complex software.

2) **World-class product**: The ultimate measure of success for our work will be the adoption of the software tools we have developed by engineers and analysts in their every-day work. While Sandia is the principal customer for these tools, adoption by the NWC and DOE Laboratories as well as military, academic and commercial users demonstrates the success of the tools. The CUBIT Meshing and Geometry Software Toolkit is the principal mechanism that these meshing and geometry tools are introduced. Building, distributing and maintaining a quality, world-class software product is one of the principal roles of the department.

3) **Technical Expertise**: The staff of the Computational Modeling Sciences Department is a highly skilled group of researchers, and software developers in the area of geometry preparation and mesh generation. We seek to lead the world in meshing and related geometry research. Through internal projects such as DART as well as external work-for-others projects such as the Goodyear CRADA and the Navy, our technical staff is relied upon to provide world-class solutions to current, relevant problems.

4) **Technical Leadership**: Leadership in the technical community is an important aspect of the department through innovative and groundbreaking research, sponsorship of the annual International Meshing Roundtable as well as participation in many conferences, symposia and publications. We have maintained world-class leadership in the area of geometry preparation and mesh generation.

II.2 CUBIT Geometry and Meshing Toolkit

The CUBIT Geometry and Meshing Toolkit is a software tool that is both an engineering tool and a research platform. Activity surrounding the research, development and production of CUBIT is the central focus of the Computational Modeling Sciences Department. The following is a summary of CUBIT and its role as a preprocessor for computational simulation.

Input: A user will typically import a CAD model that has been developed by a designer in another commercial tool such as Pro/Engineer or Solidworks. CUBIT also provides tools for basic geometry creation and manipulation.

Capability: A typical process would involve manipulation and editing of the geometry to obtain a model suitable for meshing. This would involve several diagnostic and geometric editing tools. Automated meshing tools for generating 2D and 3D meshes would then be employed using the prepared geometry. Mesh quality diagnostics and improvement methods can then be used to verify correctness of the mesh. This is all done with interactive 3D graphics using familiar graphical user interface tools.

Output: A file describing the finite element mesh along with boundary conditions ready to be used in a computational simulation is usually the objective of a CUBIT session.

Distribution: Sandia distributes CUBIT for a minimal handling fee for academic and government use. A non-exclusive commercial distribution agreement has also been established. Hundreds of licenses have been issued for CUBIT to government and academic users since its inception, 250 within the last three years.

User Metrics: Usage statistics are automatically tracked for CUBIT. Each month a unique user that uses CUBIT at least twice during the month is recorded. Recent average usage (unique repeat users within a month) has been over 400 per month, up substantially from about 50 users only 3-4 years ago.

Quality Assurance: CUBIT maintains standards and procedures prescribed by the ASC SQE document. Some of these procedures and practices include: regular software releases, a dedicated testing team, support manager, defect tracking system, project management process, etc.

Platforms: CUBIT is supported on Windows, Mac, and Linux, platforms for both 32 and 64 bit architectures.

Current Release: CUBIT version 11.0 was released November, 2007.

Support: An active customer support policy is key to the success of CUBIT. It includes a user website (cubit.sandia.gov), extensive online documentation, email lists, monthly user meeting, telephone and personal support as well as custom development for Sandia users.

II.3 Project Areas

The following is a brief description of the projects currently underway in the CMS Department.

1) **DART Integration**: DART is the overarching program at Sandia focused on improving the entire design through analysis process. CUBIT is one of the main applications contributing to the DART program at Sandia. The program has been focused on moving the tools forward (e.g. the CUBIT Geometry Tolerant milestone) and supporting the deployed integrated tools in a manner that users could be encouraged to begin using the suite. Darryl Melander has been instrumental in ensuring all the applications are integrated well, in collecting and reviewing milestone information and supporting the technical needs of the DART integration.

2) **CREATE MAGIC (Meshing and Geometry Innovation Collaboration)**: Ted Blacker, who is currently on temporary assignment in Washington DC with the Navy's HPCMP (High Performance Computing Modernization Program) office, has established a DoD effort to develop a meshing and geometry capability to support the modeling and simulation needs within the DoD. Begun this year, this program will seek to utilize the CUBIT code base as its foundation and bring in partners from industry and academia through a series of CRADA agreements to enhance it capabilities. Working

directly with DoD customers, the CUBIT team will establish specific requirements and provide direction to a DoD funded team who will develop the majority of the infrastructure for the MAGIC project. This strategic partnership is intended to build on the current investment DOE has made in geometry and meshing by dramatically expanding the contributors to the CUBIT code base as well as the overall scope and customer-base within the government. Intended to be a long-term (10-15 yr) project, Ted Blacker will be returning from DC in August of this year and will be continuing his leadership of the CREATE MAGIC program while at Sandia.

3) **Geometry Tolerant Meshing**: Prior to execution of any mesh generation algorithm, geometry issues with the CAD model must be resolved in order for the meshing algorithm to perform adequately. These include eliminating sliver surfaces, gaps and overlaps in the geometry which can be time consuming to detect and resolve. This project is intended to automatically resolve many of these problems without user interaction. The approach taken is to utilize a facet-base representation of the solid model. Facets are then manipulated and modified to resolve local geometric problems. With the new facet-based geometry representation, the surfaces can then be meshed with existing tri or quad meshing techniques which can in turn be sent to a solid meshing algorithm. This project, led by Mike Brewer, is part of a level 2 ASC milestone that is intended to demonstrate a dramatic decrease in overall meshing time on models that contain typical geometry errors.

4) **Tolerant Imprint for Assemblies**: Computational models at Sandia typically involve complex assemblies. These models often involve errors where tolerance problems have introduced gaps or overlaps between parts. To ensure a conformal mesh between parts, the imprint and merge process is employed to essentially join parts together so they share a common interface. This process can be error prone, often taking substantial time for the analyst to resolve problems prior to meshing. This project, led by Brett Clark, provides a comprehensive set of diagnostics and tools for recognizing imprint problems along with geometric tools for resolving the problems. Built on CUBIT's ITEM tool introduced in 2007, and part of the ASC level 2 milestone, it utilizes the diagnostic-solution approach to resolve the imprint/merge for assemblies problem prior to meshing. Once a consistently imprinted and merged assembly model has been developed, the geometry tolerant meshing approach described above can be used on individual parts of the model, essentially providing a fully automatic meshing solution.

5) **Geometry-less mesh representation**: One of CUBIT's strengths is its ability to efficiently develop meshes on complex CAD models. This approach assumes an initial geometry representation, upon which a mesh is generated. While effective for large scale mesh generation, CUBIT has been less effective at making smaller scale modifications to an existing mesh. The main issue has been the need to associate mesh with an existing geometry representation. Led by Darryl Melander and supported by Karl Merkley, this project dramatically expands the scope of CUBIT's capabilities to allow an existing mesh to be imported and manipulated within CUBIT without the need for geometry associativity. This new paradigm opens up a wide range of capabilities that will have immediate impact on the Sandia community. For example, capabilities such as local mesh manipulation, joining of meshes, boundary condition editing, mesh smoothing and improvement are now feasible without the necessity for an associated geometric model. It also supports the ASC level 2 milestone by facilitating a new paradigm on which a geometry-tolerant mesh can be developed.

6) **All-hex meshing R&D:** Primarily funded through CSRF, Jason Shepherd leads the effort to develop new automatic methods for all-hexahedral mesh generation. Jason recently returned from a doctoral studies program at the University of Utah where he developed new theory with respect to the generation of all-hexahedral mesh generation. Following on from this research, this past year Jason has implemented a new sheet insertion approach to mesh generation. Sheet insertion begins with a base mesh, typically a Cartesian grid or swept mesh, and inserts topological sheets or layers of hexes

to enforce geometric conformity where needed. This technique is completely general in the geometries it may address, but currently may require user interaction to provide the best base mesh and locations for sheet insertions. The objective of the research is to completely automate the sheet insertion procedure to generate a high quality hexahedral mesh for arbitrary solid models.

7) **Goodyear Support:** The Goodyear CRADA requires substantive modeling support. This support has the added benefit of directly impacting our own internal needs, including strengthening our geometry editing capabilities and our mesh algorithm development and research.

   a. **CUBIT Insertion into CATIA**: An important aspect of the Goodyear X3D project is the insertion of CUBIT directly within the CATIA environment. Byron Hanks and Corey Ernst have contributed to this Goodyear funded project which now enables CUBIT to be the principal corporate meshing solution for Goodyear Tire Corporation. It will also provide valuable additional capability to CUBIT's Common Geometry Module (CGM), with several Sandia customers anxious for the integration.

   b. **Mesh generation for tire cross-sections**: Tire cross sections represent a highly constrained two-dimensional quadrilateral mesh generation problem that Goodyear is relying on Sandia to completely automate. Building on the unconstrained paving work from last year, Bob Kerr is developing an alterative automatic quad meshing algorithm that automatically selects from a set of existing meshing schemes by recognizing specific tire cross section topology information. A fully automatic meshing approach that consistently delivers high quality quad elements for the wide range of Goodyear products is the objective of this development effort and Bob Kerr continues to directly work with Goodyear engineers to make this a reality.

   c. **Mesh generation for tire treads:** While cross section meshing requires a 2-D mesh, the objective of tread meshing is to deliver a fully automatic 3-D mesh on the tread portion of the tire represented by a CAD model. With the objective of developing an automatic tread meshing capability within the CATIA environment, the tread CAD models are transparently imported into CUBIT where new automatic meshing and geometry algorithms are employed. Because of the complexity of the tread CAD models, the standard pave and sweep approach has not proven completely effective. Instead a method is under development whereby a quality base mesh is first generated that temporarily ignores tread features. Following this, features are then inserted into the existing mesh in a systematic and conformal manner. Byron Hanks is leading the effort on this project that is currently under development which is scheduled to be delivered in September of this year.

8) **University Research:** The CMS department currently maintains research contracts at two universities, Carnegie Mellon University and Brigham Young University. For a relatively small investment, graduate students are trained and become part of the large software development team providing valuable experience to the students and developing a potential pool of new employees. Graduate student researchers at Brigham Young University, mentored by Jason Shepherd are developing new techniques for conformal all-hexahedral refinement and coarsening and have contributed significantly to the recent hex refinement capability in the recent CUBIT 11.0 release. A PhD student at Carnegie Mellon University, mentored by Mike Brewer has been developing new techniques for hybrid element boundary layer mesh generation for CFD. In addition to contributing to the CUBIT code base students will publish and base their thesis or dissertation on work they have done in CUBIT.

9) **CUBIT Maintenance:** Since CUBIT usage has increased 8 fold in the past three years, ongoing maintenance and strong support of the customer base is required.

a. **Support**: Customer support activities account for a significant amount of time and effort in the department. Apart from their assigned projects, each developer is expected to devote up to 30% of their time to customer support related activities such as answering questions, fixing bugs, teaching tutorials and maintaining platforms. Kevin Pendley, a full-time contractor directs these activities. In addition, Sara Richards, another contractor is employed part-time to enhance and maintain the CUBIT documentation.

b. **Enhancement Requests**: One of the most valuable assets that CUBIT provides to Sandia is the ability to turn around custom enhancement requests in a relatively short amount of time that can impact critical milestones. CUBIT staff and contractors interact regularly with engineers in center 1500 and across the laboratory, gathering requirements and implementing new custom capabilities into CUBIT that would otherwise be too costly or impossible to develop within a commercial software tool.

c. **Testing and QA**: As CUBIT's user base continues to increase, testing and quality assurance continues to be of prime importance. Kevin Pendley, a full-time on-site contractor supports these efforts. He mentors a group of 3 part-time student interns. They are in charge of developing and implementing the test plan for each CUBIT release and provide ongoing testing for the program.

II.4 Funding Allocation

Table 3 below summarizes the approximate costs by project consistently with the numbering scheme used throughout this document. Note that the FTE counts are based on the actual individuals working on each project. Amounts will change based on whether FTE is an independent contractor, student intern, Sandia SMTS or PMTS.

Table 3:

		Funding ($K)				
Project/Program Area	**FTEs**	**ASC/ DART**	**GY**	**CSRF**	**Other**	**Total**
1. DART Integration	0.6	200				
2. CREATE MAGIC	0.4				150	
3. Geometry Tolerant Meshing	1.0	350				
4. Tolerant Imprint for Assemblies	1.0	350				
5. Geometry-less mesh representation	1.0	350				
6. All-hex meshing R&D	1.0			290		
7. Goodyear Support	3.5		1000			
8. University Research		60		50		
9. CUBIT Maintenance	5.5	750				
	12.0	2060	1000	340	150	**3550**

II.5 Community Activity

International Meshing Roundtable: The CMS department has continued as sponsors of the International Meshing Roundtable (IMR). Now in its 16th year, the IMR is the premier technical conference in the meshing area. Last year Mike Brewer, a CUBIT project team member, was the IMR co-chairman. This year Brett Clark, another CUBIT team member, serves on the roundtable board. The roundtable will be held in Pittsburgh, Pennsylvania in October of this year. A hard-bound conference proceedings published by Springer-Verlag in which articles are thoroughly peer reviewed is a central focus of the conference.

External Review: Held as part of the International Meshing Roundtable, the CMS Department for the past four years, has invited experts from the community to evaluate its current research and development activities. Each member of the department presents current research progress and challenges to a panel of about 6 external leaders in the community. Open to the IMR community this activity has generated considerable interest and has become a feature of the IMR. In addition it has provided valuable feedback to the each of the members of the department and in some cases helped adjust our research direction.

MeshTrends: The 6th Trends in Unstructured Meshing Symposium was held in conjunction with the World Congress on Computational Mechanics in Los Angeles 2007. Steve Owen, Cubit project lead, was co-chair of this symposium with Mark Shephard from RPI. Steve has co-chaired this symposium since its inception in 1997. A special issue of Engineering with Computers from the last MeshTrends symposium was recently published with Steve Owen as guest editor.

EM08 (Inaugural International Conference of the Engineering Mechanics Institute): A special session of EM08, to be held at the University of Minnesota, will be chaired by Jason Shepherd this year. This session will involve invited speakers and submitted papers and abstracts from international experts in mesh generation and geometry.

III. Data Analysis and Visualization

III.1 Introduction

Sandia's Data Analysis and Visualization group is a world leader in scalable visualization technology, with a long history of creating both hardware and software for visualization of large, complex data. The group serves a broad range of customers inside and outside Sandia, so our research and development must address a wide variety of use cases. Thus, we develop both foundation technologies, such as those we contribute to the open source project ParaView, as well as specialized applications such as Prism, a Sandia tool which addresses the large data requirements of a particular scientific domain. The group's technology has broad external exposure through contributions to open source projects such as ParaView, a leading open source scalable visualization tool.

Historically strong in traditional scientific visualization, the group is leveraging this expertise to work in the emerging area of scalable informatics, creating scalable foundation technologies, engaging in research, and delivering end-user applications used daily by intelligence analysts.

The primary strategic objectives of the group include the following:

1. **Advance understanding of complex data:** Our customers routinely handle terabyte datasets, and near term planning calls for us to handle their petascale data. Though we have made significant progress in scalable visualization and data analysis, the requirements of petascale data will require fundamentally different approaches in data management, analysis and visualization. Our customers need to explore and understand complex sets of data – families of simulation runs, experiment-to-simulation comparison, and optimization data. How does one explore hundreds of simulations at once? How does one compare sparse data from sensors to highly detailed simulations of the same event? We pursue both research and development of fundamental technologies to address these needs, focusing on informatics, information visualization, multiple linked views, and optimizations in I/O for large data.

2. **Develop leading-edge applications in partnership with domain experts:** Our team has worked diligently over the past several years to build expertise in powerful re-usable software components that promote Rapid Prototyping – a development model in which teams quickly create working software that customers can touch and use. This results in a tightly integrated software development process in which our researchers and developers can work closely with experts in a variety of domains. By quickly delivering working software to our customers, we break down barriers in development, deliver innovative solutions quickly, and build lasting partnerships with a variety of domain specialists. This approach to development has

resulted in successful partnerships internal Sandia customers, as well as external customers such as NASA, and the National Ground Intelligence Center (NGIC).

3. **Deliver world-class research within production-quality software:** In a typical research environment, it is difficult to deliver advanced research to end users, as most code is not developed in rigorous software environment. Most projects begin as efforts in an individual developer's sandbox, and are difficult to bring to production use. Our group is committed to decreasing the effort it takes to deliver research in production-ready end user tools. Through our expertise in both VTK and ParaView, we are able to do advanced research in algorithms, analysis, and user interfaces, and deliver those to users in regular releases of ParaView. In addition, each software project is developed under strict design guidelines, in accordance with high-quality software practices, coding standards, and testing. All software is subjected to nightly building and testing on Windows, Linux, Mac operating systems.

4. **Provide World-class technical leadership in analysis of large data:** Sandia's scalable rendering technology, ICE-T, has been adopted in production tools in both the US and abroad. Large data analysis software such as ParaView, Lawrence Livermore Lab's VisIt and Commissariat à L'énergie Atomique's (CEA) LOVE tools depend upon ICE-T's advanced algorithms and software in production environments. In addition, through such venues as SC (Supercomputing), IEEE Vis, and other conferences, Sandia leads discussions of technical interest to the visualization and supercomputing communities. This year, we will host tutorials at Supercomputing that detail how tools such as ParaView can be utilized to explore large data.

Our efforts this year have focused on delivering end user tools for large data analysis, and technologies to address petascale data in the near future. Sandia has lead the design and development of ParaView 3.0, which includes a re-designed client/server architecture, advanced application capabilities (multiview, etc.), and a scalable python client. This new version of the software allows us to build advanced solutions to address the needs of specific groups of customers, while at the same time delivering capability that can be used by a wide range of problem domains. The ability to adapt a stable scalable software stack to address the needs of specific customers is a crucial strategic capability.

Our main focus areas include:

1. **Advancing scientific visualization with informatics**. Our data is growing in size and complexity. Our customers need to be able to compare, analyze and understand many types of data at the same time. In addition, data analysis needs for Petascale and beyond require fundamentally different capabilities for analysis and visualization. Our goal is to provide a suite of tools that allow analysts to explore, manipulate and understand large datasets – whether the data is generated by a simulation, or captured from experiments. Our tools must be responsive, regardless of the size of the data, and this requires advanced work in compression, feature detection, and rendering.

2. **Partnering in Scalable Analysis, with a Verification and Validation (V&V) focus:** Utilizing the scalable Python client available in ParaView 3.0, Sandia is collaborating with internal customers to design and deploy scalable analysis tools that directly impact Sandia's V&V mission. Sandia's analyst community has begun developing software that performs both verification and validation of codes and results, and we expect this to be a major focus of the coming year as well. The crucial aspect of this work is collaboration; working with customers to co-develop software that directly addresses their specific needs, while at the same time developing a common software layer that can be used by many different

application areas. See Fig. 2 for a detailed view of the analysis capabilities we have recently delivered in collaboration with Sandia analysts.

3. **Scalable Informatics**. ThreatView^{TM} (see Fig. 1) is Sandia's tool for scalable analysis of information. Based on our experience with the successful LDRDView tool, we are developing a scalable solution to information visualization that promotes abstract exploration (utilizing visualizations such as landscape, graph, and geographic views), as well as fast querying of data. ThreatView is being released to external customers for feedback, and will be in use on real analysis problems by the end of FY07.

III.2 Project Summaries

Scalable Analysis

Last year, we released ParaView 3.0, a major new version which significantly changed the software architecture, including a re-designed client/server architecture, advanced application capabilities (multiview, etc.), and a scalable python client. This new version of the software allows us to build advanced solutions to address the needs of specific groups of customers, while at the same time delivering capability that can be used by a wide range of problem domains. The ability to adapt a stable scalable software stack to address the needs of specific customers is a crucial strategic capability.

Our efforts this year have focused on tools and technologies to address extremely large data, including V&V efforts for important large scale problems coming off of Red Storm. Because of the flexible architecture in the new version of ParaView, we can quickly address specific problems at scale, providing unique capabilities on demand. The ability to develop specific capabilities for problems at scale, in collaboration with analysts, is an important distinguishing capability of this technology. In addition, our commitment to developing in an open source application ensures that the community benefits from solutions developed at Sandia, and that we can collaborate with academic, industry and government partners from around the world.

Titan Informatics Toolkit

The Titan Informatics Toolkit is Sandia's Open Source Software informatics toolkit. It is a critical technology for the Network Grand Challenge (NGC), where it provides the integration and visualization framework for a broad range of work. In addition, Titan is a critical technology in addressing the visualization and analysis needs of ASC analysts working on extreme scale data. By providing a common informatics framework for both scientific and non-scientific data analysis, Titan is a powerful and effective development toolkit for a range of Sandia needs. Titan expands upon our prior scientific visualization work, applying the same parallel client-server infrastructure underlying the ParaView application. Titan is enabling us to develop a unique capability to analyze and render data at sizes and complexities that are impractical to process with serial applications.

Massive Graph Visualization LDRD

Current tools for the visualization of graph structures are limited to serial computing. The goal of this LDRD is to jumpstart the field of graph visualization to scalable tools that run on distributed memory machines and that can handle data collections of enormous size. The research fostered by this project is directly contributing to the Titan/Infovis framework and the ThreatView project.

Electrical Visualization CSRF

Sandia's parallel circuit simulation software includes a new analysis approach that uses multi-time partial differential equations (MPDEs) to simulate devices that operate on vastly different time scales (differing by many orders of magnitude). MPDE encodes the problem using a complex data model that is not intuitive for circuit designers and analysts, who have traditionally analyzed simulation results using waveform plots. Unfortunately, for problems of this kind, where period information is changing over long timescales, waveform plots that include a long enough time period to discern a trend become so dense that the detail showing the periodicity is lost. Viewing simulation results directly as an MPDE grid has the advantage that both macroscale and microscale information can be viewed simultaneously. However, both views are valuable for revealing different aspects of the solution and of the data model. This project will create a novel analysis capability through the development of an informatics tool that interactively correlates waveform plots with the more abstract and information-rich representation provided by MPDE grids.

Large Data Comparison CSRF

Our work in data comparison this year focuses on one central problem: there is simply too much data for an analyst to sort through by hand. Moreover, only a tiny portion of the data is at all interesting. Since analysts' time is a scarce and precious resource, our efforts center around tools and algorithms to concentrate that time where it is most needed. We are developing methods that enable the computer to identify and efficiently extract interesting regions from petascale ensembles of data comprising hundreds or thousands of individual simulation results. Broadly speaking, this work involves a database-like paradigm to allow an analyst to ask specific and quantitative questions about the data as well as an access framework to enable automated detection of interesting areas among data sets of differing structure and origin. This work crosses a full spectrum of tools and methods ranging from low-level algorithms for petascale data interpolation to statistical display metaphors for deeper understanding of ensemble behavior.

ThreatView application

ThreatView is a parallel data analysis and visualization tool designed for use within the intelligence community. It expands upon our prior scientific visualization work, applying the same parallel client-server infrastructure underlying the ParaView application to the needs of intelligence analysts. Because ThreatView is based on this existing architecture, we are rapidly developing a unique capability to analyze and render data at sizes that are impractical to process with serial applications. ThreatView is the application within which we deliver production-ready capability from a range of research efforts, including graph analysis, layout algorithm development, and text analysis capabilities. Because it is used daily by analysts, and we interact with a group of internal users to help determine how features and capabilities will be delivered.

Scalable Solutions for Processing and Searching Very Large Document Collections

Intelligence analysts currently must answer questions of national security under extreme time pressure by evaluating petascale document collections. Analysts need to explore the data iteratively by testing various "what-if" scenarios, requiring quick turnaround and visual representations that allow them to evaluate conceptual content without having to read all of the documents. Through the Titan toolkit, this project provides scalable ingestion and exploration of large document collections by developing a set of robust, parallel methods for processing, searching, and visualizing information contained in unstructured text documents.

SciDAC Institute for Ultra-Scale Visualization

The SciDAC Ultra-Scale Visualization Institute, established on September 15, 2006, is a 5-year research and outreach effort sponsored by the DOE SciDAC program. Its mission is to address the upcoming peta-scale visualization challenges facing computational science and engineering. The Institute will revolutionize the very process of scientific discovery by equipping scientists with tools that shed light on the knowledge hidden in previously incomprehensible datasets. Sandia's part in the institute is to ensure that a general purpose visualization tool is available for peta-scale visualization and beyond.

Table 4: Staffing levels

		($M) Funding				
Project/Program Area	**FTEs**	**ASCI**	**Other**	**CSRF**	**LDRD**	**Total**
1. Scalable Analysis	3.0	3.0				**3.0**
2. Titan Informatics Toolkit	1.5				1.5	**1.5**
3. Massive Graph Visualization LDRD	1.0				1.0	**1.0**
4. Electrical Visualization CSRF	0.5			0.5		**0.5**
5. Large Data Comparison CSRF	0.5			0.5		**0.5**
6. ThreatView Application	0.5		0.5			**0.5**
7. Very Large Document Processing	1.0				1.0	**1.0**
8. SciDAC Institute for Ultra-Scale Viz.	1.0		1.0			**1.0**
	9.0	**3.0**	**1.5**	**1.0**	**3.5**	**9.0**

IV. Community Activity

Our publication record for the period (July 2007 to March 2008) since the last external review is summarized in Table 6 on an annualized basis. For a bibliographical listing of publications, a collated publication list is provided. In response to suggestion from the external review, we have substantially enhanced our publications in quantity and quality. This year our total number of publications has increased by 34%, Journal papers by 40%, conference papers by 89%, and technical reports by 21%.

Table 6: Publications Summary Annualized

Publication Type	**Quantity**
Journal papers	**28**
Conference and workshop papers	**83**
Short papers	**4**
Technical reports	**23**
Total	**138**

In community efforts, we have continued as Co-organizers of the Salishan Conference which had another successful year. Our staff also serves as the co-organizers of the SOS workshop series. SOS12, hosted by the Swiss Institute of Technology, delved into large science applications. Our staff continues to serve as organizers, steering committee members, program committee members, and reviewers in almost all major conferences in the high performance computing area. Overall, we are very active in community affairs in high performance computing.

RESPONSES TO EXTERNAL REVIEW PANEL QUESTIONS ON SYSTEMS FOR YEARS 2005, 2006, AND 2007

By Sudip Dosanjh
Sr Mgr, 1420
Computer and Software Systems

This document compiles our replies to comments and suggestions made by the External Review Panel in 2005, 2006, and 2007 that pertain to Sandia's efforts in systems and enabling technologies.

2007 External Technical Review Computers and Information Sciences Research Foundation, July 24-26, 2007

Computer & Software Systems, Sudip Dosanjh, page 8

While noting SNL-NM focus on capability systems highlighted by our R&D in lightweight operating system, NIC architecture, and memory subsystems, you pointed out two issues: the need for long-term academic and industrial partners and the need to better balance strategic focus (e.g., next LWK) with tactical service (e.g., ORNL risk mitigation). We are addressing both these issues. We continue to expand our academic and industrial partnerships. As you know we have a CRADA under negotiation with Intel and we are expanding our partnerships with Micron, AMD, Cray, Sun and IBM. We hosted two workshops: one on runtime and programming environments, and the other on memory architectures. Some of the Panel members attended these. We are actively cultivating partnerships with other agencies including Office of Science, DARPA and NSA). We have undertaken a strategic drive toward an open source lightweight kernel while fulfilling our contractual obligation for dual core and quad core capabilities in Catamount. We will be reporting on these developments in the future.

Open Fabrics & InfiniBand, Matt Leininger, Page 9

We appreciate your commendation for SNL/CA role in catalyzing an open source community effort around InfiniBand, particularly in the early days of InfiniBand. We agree that the work described had a large public service and infrastructure component and did not emphasize research or quantitative analysis. You noted that the presentation did not elucidate scalability issues. You reiterated a research focus on a good characterization of InfiniBand performance, scalability bottlenecks, and comparison with alternatives (e.g., 10G Ethernet).

Matt Leininger overview provided an example of how active engagement between the DOE/NNSA laboratories and industry can change the high performance computing market. This work was essential to the development of a robust and efficient InfiniBand software stack for HPC, and now that OpenFabrics is a self-sustaining success, we have embarked on new paths on which we at SNL/CA, in unison with SNL/NM, can impact high performance computing. There are two primary direct follow-on to the InfiniBand work. The first of these is a set of studies comparing 10 Gigabit Ethernet to InfiniBand.

A particular feature of this work is the use of 10 Gigabit Ethernet switches that support adaptive routing. We have demonstrated that the routing algorithms employed are superior to the oblivious routing method used in InfiniBand. This work is done in conjunction with John Naegle at SNL/NM. The second follow-on project is the development of new architecture simulation technologies. Work in this area grew out of the need to demonstrate to InfiniBand vendors that congestion control and adaptive routing would improve application performance. Departing from traditional network simulation techniques using statistical network traffic patterns, we have begun implementing a simulator that can generate traffic patterns based on data gathered from actual applications. It is not specific to InfiniBand and will impact capability in addition to capacity computing. The simulator allows long run times on a large number of simulated processors and nicely complements the Sandia SST clock cycle accurate simulator. SNL/CA is engaged with SNL/NM and the Institute for Advanced Architecture and Applications to develop and integrate our simulator and other technologies into a complete multiple scale/fidelity architecture simulator.

System Software, Ron Brightwell, page 10

We appreciate your praise for: our LWK work, risk mitigation, move to open source, virtualization and composable operating system to support customization for applications and future architectures. We agree that rigorous characterization of OS noise, however good and needed, may not be elegant research. You suggested a study to understand OS requirements of informatics workloads (“Google OS”). We have begun analyzing informatics workloads both from OS and architecture perspectives under a Grand Challenge LDRD effort. We will be reporting on it in future.

2006 External Technical Review, Computer and Information Sciences (CIS) Centers, 1400, 8900, June 7-9, 2006:

System Session, Sudip Dosanjh, General Comments, slide 10, pages 32-35

We do appreciate your high praise for Red Storm. We’ve worked hard on a post-Red Storm strategy. NNSA’s decision to consolidate supercomputing to two sites has given this effort even greater urgency. We are pursuing two complementary initiatives to ensure that Sandia continues to impact supercomputing architectures. The U.S. Congress established the Institute for Architectures and Algorithms in 2008 with centers of excellence in Albuquerque, NM and Knoxville, TN. We have also entered a strategic partnership with LANL to architect, design, develop, procure, deploy and operate supercomputers for NNSA. The IAA helps ensure the health of our research and the LANL/SNL partnership is an important avenue for impacting DOE’s mission and the supercomputer industry.

We are targeting resilience and fault-tolerance in this year’s LDRD call. We have written proposals on hardware virtualization technology to the Office of Science and Sandia’s LDRD program but have not yet achieved success.

Making Red Storm a Success, Sue Kelly, page 11

We appreciate your kind words on Red Storm success in Limited Availability mode, and our joint efforts with Cray for needed improvements. Upgrading to dual core and the next Seastar NIC and improving the Lustre file system has led to significant performance boosts as predicted by the panel.

Geometry and Mesh Generation, Matt Staten, page 12

We appreciate very much your kind words on CUBIT Geometry and Mesh Generation Toolkit, CUBIT as the poster child for enhancing our collaboration with academia, and industry, etc. Your specific suggestion to accelerate work on AMR on hexahedral meshes toward significantly reducing computation time for complex problems is welcome. Our progress is as follows:

Since the 2006 CIS review, three specific areas have been perused in AMR: conformal hex refinement, coarsening and new R-adaptive techniques. A student thesis was completed and publication accepted in the area of conformal hex refinement that introduced several new strategies for locally refining arbitrarily shaped regions that will maintain a convex refinement shape and reduce the number of required elements. Implementation of this new technique in the CUBIT software also demonstrated major efficiency improvements over the previous implementation. Following on from this work, conformal hexahedral coarsening work has made significant progress. Utilizing new algorithms developed for sheet insertion and extraction, promising results have been demonstrated showing controlled local coarsening of a hexahedral mesh. A publication has also been accepted for this work and a patent is pending. The third area of work involved working closely with the Alegra team to develop a new r-adaptive algorithm to adaptively adjust node locations to maintain quality anisotropic quadrilateral elements through a series of Alegra analysis runs. This work supported the Z-Pinch Grand Challenge and a journal article was published on the method.

Visualization with ParaView, Brian Wylie, page 13

We appreciate your recognition of ParaView as a successful open source toolkit that is used by 60 Labs and universities and is the only tool able to handle the largest simulation runs on Red Storm. You suggested that we compare it to other visualization tools like VxInsight, and report on user feedback.

The question asks about ParaView in comparison to tools such as VxInsight, but we assume that the intended question was to compare ParaView with EnSight, a commercial visualization product.

ParaView is an Open Source Software (OSS) project developed by Sandia's Visualization R&D team. With this open source project we can develop visualization and analysis capabilities for

ASC scale datasets, in arenas that would not be practical for a commercial product like EnSight. In addition, we can customize ParaView for specific problem domains, delivering capabilities for crucial problems specific to the Sandia mission. ParaView has proven invaluable in the postprocessing of some of the largest simulations in the world generated by the Red Storm platform. Given the success of ParaView, both at Sandia and the world at large, our research often influences capabilities in commercial products. In addition, ParaView, as an Open Source Software project, attracts contributors from around the world, fostering a rich community of researchers that contribute back to the code base.

Supercomputing System Design through Simulation, Rolf Riesen, page 14

In hindsight, we agree with you that our presentation was premature. We have taken your suggestions very seriously. There is a much greater synergy between this effort (Seshat) and the Performance Modeling effort. Our internal management reviews have helped ensure that there is no duplication of efforts, and we are targeting unification at a future date. In the meantime the teams are working collaboratively. We have launched a concentrated broader HPC performance modeling and simulation toolkit that will be presented this year.

General Comments on HPC Performance Modeling, page 15

We appreciate your recognition of HPC Performance Modeling as becoming an essential tool for guiding next generation system specification, making tradeoffs configuration, and tuning application codes. We also appreciate your realization that it needs critical mass, broad coverage, and integration of multiple architectures, methodologies, and a believable Verification and Validation strategy. In particular, you asked us to consider DOE tri-Labs and DoD initiative in this area. We agree in principle, and we are exploring collaboration with DoD. We wish to address these in a newly funded Institute for Advanced Architecture for which we are working with ORNL.

2005 External Technical Review, Computers and Information Sciences (CIS) Centers, 1400, 8900, August 10-12, 2005:

Introduction and Overview by Sudip Dosanjh, page 14

We appreciate the kind words of the Panel on our collaborations with universities, national Labs, corporations, and at large, and in our publication efforts. We have provided better context for the detailed talks in subsequent years.

Red Storm, Jim Tomkins, pages 15 and 16

We appreciate your praise on Red Storm, its contribution to the state of the art in capability systems, and our enabling desirable changes in Cray. Sandia was able to stabilize Red Storm and is now planning a second upgrade. Many large calculations have been run across the entire machine.

Performance Modeling, Robert Benner, page 17

In line with Panel's appreciation and suggestion, we have published performance tuning and modeling papers and presented results at workshops. We have not yet accomplished the ultimate goal of incorporating performance modeling into applications development. However, results are conveyed to the applications on a regular basis.

FPGA Based Processing Architecture, Keith Underwood, page 18

We are grateful for your appreciation of our contributions in improving the double precision FPU in FPGAs. In line with your suggestion, we did not expand this project to a full architecture.

Lightweight File System Project, Ron Oldfield, page 19

Suggestions from the Panel have been incorporated in development of the LWFS in subsequent years. The project was focused more on the research aspects of filesystems and was not used as a risk-mitigation strategy for Red Storm. Close work with CFS helped improve the performance and reliability of Lustre. The PI did survey the I/O patterns of important applications. A related project has developed an I/O profiling tool. Currently, LWFS is undergoing large scale deployment tests on Red Storm as described in the Overview section.