Red Storm Usage Model
Version 1.12

Karen L. Jefferson and Judy E. Sturtevant

Sandia National Laboratories

Red Storm Usage Model
Version 1.12

Karen Jefferson (ed.) and Judy Sturtevant (ed.)
Sandia National Laboratories
P. O. Box 5800
Albuquerque, New Mexico 87185-9152

With Input and Review from
ASC/S&CS

Abstract

Red Storm is an Advanced Simulation and Computing (ASC) funded massively parallel supercomputer located at Sandia National Laboratories (SNL). The Red Storm Usage Model (RSUM) documents the capabilities and the environment provided for the FY05 Tri-Lab Level II Limited Availability Red Storm User Environment Milestone and the FY05 SNL Level II Limited Availability Red Storm Platform Milestone. This document describes specific capabilities, tools, and procedures to support both local and remote users. The model is focused on the needs of the ASC user working in the secure computing environments at Los Alamos National Laboratory (LANL), Lawrence Livermore National Laboratory (LLNL), and SNL.

Additionally, the Red Storm Usage Model maps the provided capabilities to the Tri-Lab ASC Computing Environment (ACE) requirements. The ACE requirements reflect the high performance computing requirements for the ASC community and have been updated in FY05 to reflect the community's needs. For each section of the RSUM, Appendix I maps the ACE requirements to the Limited Availability User Environment capabilities and includes a description of ACE requirements met and those requirements that are not met in that particular section. The Red Storm Usage Model, along with the ACE mappings, has been issued and vetted throughout the Tri-Lab community.

Acknowledgements

We thank Jim Ang for providing the initial framework for the document, assistance with the initial call for information, and the review of Peculiarities of the System. We also want to thank Jim Handrock, the owner of the Red Storm User Environment Milestone, who gave us the organizational framework and managerial support to ensure the completion of this document and the success of the milestone.

Contributors

Bob Ballance
Jim Brandt
Doug Brown
Brian Carnes
Milt Clauser
Scott Cranford
Mike E. Davis
Kristin Dion
Martha Ernest
Rena Haynes
Mike Heroux
Ann Hodges
Lisa Ice
Curtis Janssen
Barbara Jennings
Wilbur Johnson
Jeff Jortner
Sue Kelly
Glenn Machin
Susan McRee
John Naegle
Bill Rahe
Randal Rheinheimer
Patti Sanchez
Greg Sjaardema
Judy Sturtevant
Larry Tolendino
Jim Willenbring
Brian Wylie

Change Table

Version	Date	Comments
0.01	6/21/04	First working draft distributed among the RS UM team
0.02	6/30/04	Reworded Headings in Section 5 for Schematic/Netlist setup, updated DisCom reference and some minor edits.
0.03	7/15/04	Inputs from Bob Ballance, Ann Hodges, Curtis Janssen, and Judy Sturtevant
0.04	7/22/04	User support, account administration, access. BJJ Section 1 – Section 8.1. 8.2
0.05	8/2/04	RAB – merge of BJJ comments and additional edits
0.06	8/17/04	MJE – merge LT Section 3 and comments and MJE comments/edits. Additional edits by JES and JAA.
0.07	11/22/04	Address comments from JES and Glenn Machin. Added discussion of focus on the timeframe of the 4QFY05 Level 2 milestones. Dropped discussion of system administrator concept of operation from Section 0.
0.08	11/22/04	Address comments from Sue Kelly
1.0	1/6/05	Creation of version 1.0 to support LA Environment milestone.
1.0	2/16/05	Modified to address tri-lab concerns and questions and SNL internal comments. Contains mapping to ACE version 8.0
1.01	3/23/05	Included modifications to ACE mappings and clarifications by Ballance and Kelly
1.02	3/29/05	Updated Section 3.3 (Tools for transferring files) to reflect discussions held with Haynes, Rahe, McCree, and Clauser
1.03	4/18/05	Updated Introduction to include a definition of Limited Availability and the Limited Availability perspective of the Milestone.
1.04	4/20/05	Updated Section 3.3 (Tools for transferring files) to reflect discussions held with Naegle and Machin.
1.05	5/11/2005	Updated Section 3 with mods from Glenn Machin.
1.06	6/29/2005	Updated Section 3 with mods from Susie McRee. Deleted all Appendices except the ACE Mapping. Added Acknowledgements.
1.07	6/29/2005	Version 1.06 with all changes accepted.
1.08	8/26/2005	Modified to reflect known milestone functionality.
1.09	8/31/2005	Modifications to reflect known milestone functionality reviewed by KJ & JS.
1.10	9/7/2005	Updated ACE mappings to reflect known milestone functionality.
1.11	9/12/2005	Add SEACAS to Section 4; change title for Section 3.
1.12	10/12/2005	Editing changes for SAND report.

Intentionally Left Blank

Contents

Figures

Tables

Intentionally Left Blank

Acronyms and Abbreviations

AHP	(Sandia) Analyst Home Page
ACE	ASC Computing Environment
ASC	Advanced Simulation and Computing
CEMS	(Sandia) Cyber Enterprise Management System
CPA	Compute Processor Allocator
CUP	(ASC) Capability Usage Performance
CLIK	Collaborative Learning, Information, and Knowledge
DART	Design through Analysis Realization Team
DMZ	Demarcation Zone (the boundary between the network external to SNL and SNL's internal network)
DSA	Data Services Archive
EPR	(Tri-Lab) Expedited Priority Run
HPC	(Sandia) High Performance Computing Network
LANL	Los Alamos National Laboratory
LLNL	Lawrence Livermore National Laboratory
NWIS	(Sandia) Network Information System
PFTP	Parallel File Transfer Protocol
RSUM	Red Storm Usage Model
PBS	Portable Batch System
SARAPE	Synchronized Account Request Automated Process
SC	Sandia Classified
SCN	Sandia Classified Network
SNL	Sandia National Laboratories
SRN	Sandia Restricted Network
SSCS	Sandia Scientific Computing Support
WAN	Wide Area Network
WebCARS	Web Computer Account Request System

Intentionally Left Blank

Red Storm Usage Model
Version 1.12
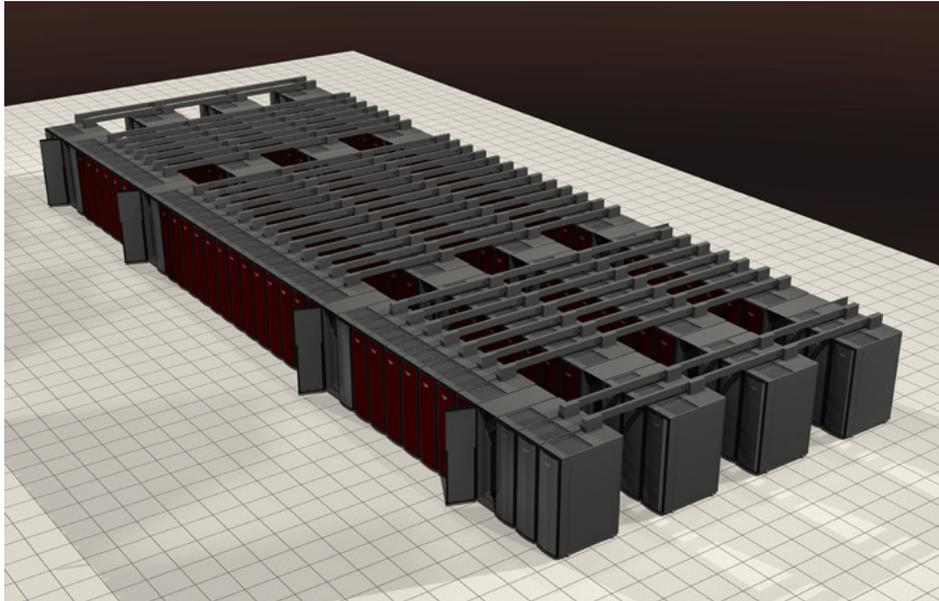
For ASC Program Elements to identify capabilities that are expected to be available for their respective Red Storm 4QFY05 Level 2 milestones

For Red Storm Technology Developers to understand how their technologies are integrated with others' into the Red Storm User Environment

For Red Storm Administrators and Computational Scientists to describe recommended procedures and tools for the Red Storm User Communities

For Red Storm Users to understand the technical capabilities that are available to support their usage of Red Storm

Introduction

The Red Storm Usage Model (RSUM) documents the capabilities and the environment provided for the FY05 Tri-Lab Level II Limited Availability Red Storm User Environment Milestone and the FY05 SNL Level II Limited Availability Red Storm Platform Milestone.

The goal of the FY05 Tri-Lab Level II Limited Availability Red Storm User Environment Milestone is as follows:

Demonstrate a user environment that provides application development and execution, data analysis and visualization, and distance computing in accordance with ASC Red Storm and application requirements as defined in the Red Storm Usage Model. PSE, together with VIEWS and DisCom, will build on the limited environment developed in FY04 to deliver both a local and distance-computing environment for a limited user base.

The goal of the FY05 SNL Level II Limited Availability Red Storm Platform Milestone is as follows:

> Transition the Red Storm computer system to a status that provides Limited Availability service to the TriLab NWC customers. The system will be accessible to a small number of ASC code teams, on the order of 2 or 3 teams per lab. Access will be granted for up to 30 users per laboratory to work on programmatic deliverables (i.e., ASC Level I or Level II milestones). The set of code teams having access in Q4FY05 is to be determined by the ASC management team at each laboratory. The Expedited Priority Request (EPR) process will coordinate usage requests among these teams. The system will be operating in both classified and unclassified mode. Authenticated access to the system from the DisCom WAN will be available to LANL and LLNL customers. File transfers to and from classified systems at LANL and LLNL will be supported.

In reference to the milestones and this document, limited availability extends to the capabilities provided by the system, system availability, and the number of users. A number of capabilities that users may expect to be available may not yet be in place, as well as capabilities limited in performance and functionality. Follow-on work will continue on these capabilities and is indicated in this document as "after FY05". While attempts will be made to schedule downtimes, it may not always be possible to schedule these with adequate lead time for users to plan their work schedule with confidence. It will be necessary for users to coordinate their major calculations with the computer center to enhance chances of a successful outcome. Each laboratory is requested to put forth applications for the Limited Availability milestones, with the intent to restrict the number of users and applications for the milestones.

Users (and System Administrators) of Red Storm are expected to engage in up to eight activities:

1. Getting started (learning about the system, gaining access etc.),
2. Setting up the work environment,
3. I/O and Data migration,
4. Application and system code development,
5. Problem setup,
6. Running the application to solve the problem,
7. Processing simulation output,
8. Tri-Lab coordinated operational support.

This document describes specific capabilities, tools, and procedures to support both local and remote users. The model is focused on the needs of the ASC user working in the secure computing environments at LANL, LLNL, and SNL.

1.   Getting Started (Learning About the System, Gaining Access, etc.)

1.1.   Platform Map

Topology - Physical Layout
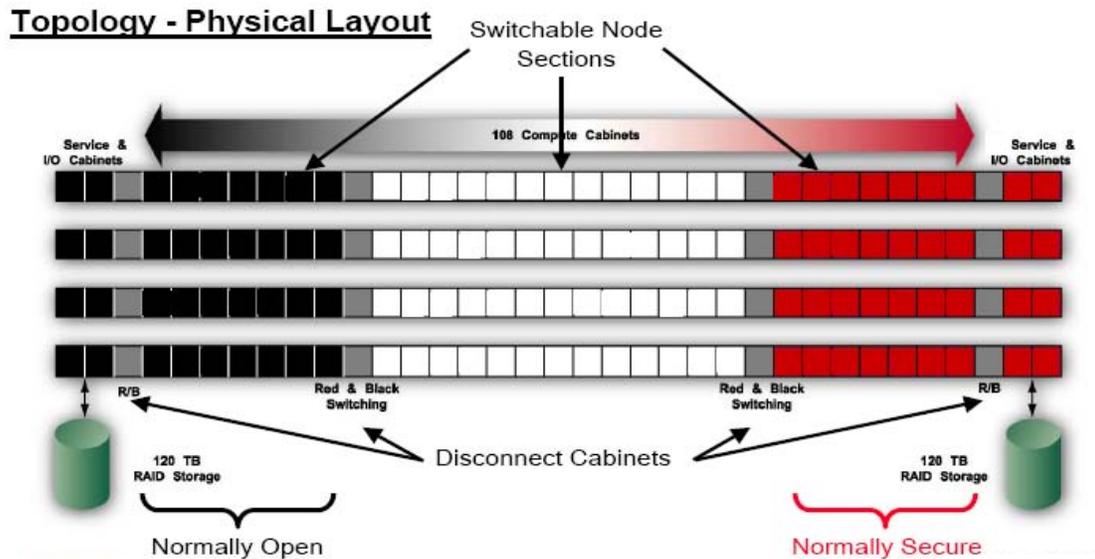
Figure 1. Red Storm Physical Layout.

1.1.1.   Service Nodes

Each of the service nodes has a very specific function. In FY05, service nodes include the following types of nodes:

- **Login Nodes.** Users log into these nodes and perform all of their interactive work; compiling, debugging, etc.,

- **Network Nodes.** Nodes with 10Gb Ethernet interfaces that are solely for moving data on and off of the machine,

- **Disk Nodes.** Control the Fibre Channel disk storage with the Lustre file system; no connectivity to the outside world, and

- **Boot, Admin, System Database (SDB), Syslog, etc.** Various nodes that control and operate the machine.

After FY05, service nodes may include Visualization Nodes, which contain graphics cards for rendering images.

Service nodes are preconfigured and do not change their function. The actual number of each type of node is still being determined.

1.2.   Learning about the System

For the purposes of using the Red Storm system and the associated environment at SNL, web based, written, and on-line system documentation and training are provided. This information provides both a high level overview and detailed descriptions of how to utilize the Red Storm environment. Web-based documentation for Red Storm can be divided into four categories: informational pages, on-line support knowledge bases, links to printable materials, and status information.

1.2.1.   Web-based System Documentation

Consistent with High Performance Computing systems available throughout the Tri-Lab community, the Red Storm web pages provide detailed user information for policies, getting started, systems status, hardware and software environment, code development, running jobs, and technical assistance contacts. Information on obtaining user accounts and accessing the system can also be obtained. Web pages are located via the Red Storm Home Page URL http://.redstormweb.sandia.gov/ and an alias of rsweb.sandia.gov on the Sandia Restricted Network (SRN). Availability on the Sandia Classified Network (SCN) may occur after FY05. These pages conform to the ASC Web page standards, and where applicable, to the Sandia Classified (SC) Web Page Standards.

The public or open network web pages provide sufficient information about Red Storm to establish its role among the ASC platforms. The pages provided may be used to establish whether Red Storm would be an appropriate computing platform for prospective users, projects, or applications. These web pages will be available after FY05.

Inclusive searching of all on-line documentation is provided through the web search capability of the Collaborative Learning, Information, and Knowledge (CLIK) tool. The CLIK tool is available on the Sandia Restricted Network only at http://clik.sandia.gov and also accessible from the Red Storm Home Page. The CLIK tool combines an on-line search capability with a problem submission and tracking capability.

Both the Red Storm web pages and CLIK tool provide access to the Sandia Analyst Home Page (AHP). The Analyst Home Page provides links to information required by the Sandia Engineering Sciences community, and as such, contains links to documentation of the Red Storm web pages. The AHP site, available only on the SRN, provides access to information about the Sandia analyst tools, codes, and servers for analysis projects. Topics include FAQ, policies, how to information, job submission, and hardware status. Information on the AHP is Sandia-centric and may not be useful to the non-Sandia employee or contractor.

The Cray documentation product, CrayDoc™, is available and populated with the full set of Red Storm documentation. Users have access to this documentation set through the Red Storm Home Page.

1.2.2.   On-line System Information

Real time and static machine information is available to the user community on the SRN only. Information of interest includes current configurations, planned activities (center section swings,

dedicated time, preventive maintenance), queued and running jobs, and the general state of the machine, i.e., number of cpus running or free.

Problem submission and tracking is available on-line through the CLIK tool. Subsets of this information may be limited to Cray and SNL developers.

Current and scheduled system configuration changes (red/black switching, maintenance, software upgrades) are visible from the Red Storm Home Page.

1.2.3. Written System Documentation

Cray will be providing an extensive set of written documentation for Red Storm. Access to their documentation is provided via the Red Storm Home Page and may be searched using the CLIK tool. SNL-specific documentation is provided via this same home page. To help users get started, a first draft of the ASC "Red Storm Primer", formerly known as "Red Storm for Dummies", is provided in the same form as the previously published documentation for ASC Red and ASC White.

All Red Storm written user documentation is accessible as PDF files from the web pages or through the searching functionality in the CLIK tool.

1.2.4. Training

Potential users, after FY05, must complete the Red Storm "Driver's License" course to enable accounts. This on-line course is available to users after they have been approved for an account on Red Storm. The information covered in this class includes security requirements, file system configuration, system usage for submitting jobs, job management instruction, and required usage protocols. This training also introduces the user to the Red Storm knowledge and support community.

Cray's training strategy does not currently include online training. The focus is on classroom training, either on-site (Albuquerque, NM) or at Cray facilities. Where third-party vendors supply on-line training, Cray will assist SNL to integrate the training into the Red Storm infrastructure. The following classes are planned for the Limited Availability environment: a Cray developed end-user performance and optimization class; and a third-party vendor class on the Lustre file system.

After FY05, an Access Grid Training facility at SNL may be installed, which allows SNL to provide distributed training. On-site training at other locations can be obtained by contracting directly with Cray.

1.2.5. Consulting

Sandia Scientific Computing Support provides three mechanisms to access support for Red Storm: the Web, email, and phone. Web assistance can be obtained via the CLIK tool, which is accessible from the Red Storm Home Page. Assistance through email may be requested via redstorm-help@sandia.gov. After FY05, phone assistance can be accessed at 505-284-SSCS. This hotline will be fully staffed from 7 AM to 6 PM Mountain Time. Requests are responded to

by a team of experts (expert team TBD). The CLIK tool provides access to information regarding external vendor resources, documentation, and the corporate knowledge base. Additional services, such as intersite account requests, training, sample codes, code porting, and application tutorials, can be found at the Red Storm Home Page 24 hours a day.

Cray provides three on-site computational scientists for Tri-Lab application support. They will be available for consultation by July 2005.

1.3. Gaining Access to the System (In General)

1.3.1. Account and Password Management (Authorization)

To obtain a Red Storm account, SNL employees and contractors require authorization from their immediate management. The request, as well as authorization, is routed via the workflow provided by the Web Computer Account Request System (WebCARS) (https://cfwebprod.sandia.gov/cfdocs/WebCars/templates/webcars.cfm). Additional approvals by the SNL milestone coordinators are required for accounts during Limited Availability.

Non-SNL employees and contractors (intersite users) are required to submit Intersite Guest Account Requests electronically via the Synchronized Account Request Automated Process (SARAPE) (sarape.sandia.gov) application. SARAPE requests require Guest Processing and Host Processing Agents approval. Additional approvals by the Tri-Lab milestone coordinators are required for accounts during Limited Availability. For intersite users, SARAPE provides additional accounts, SNL SecurIDs, etc., as appropriate for the platform requested (e.g. a visualization account provided when a Red Storm account is requested).

Synchronization of user accounts on SNL computers occurs through SNL's Network Information System (NWIS), and as such, updates Red Storm's /etc/passwd file nightly. In this way, SNL ensures standard account administration across all of the SNL supercomputers. Nightly update and synchronization with corporate NWIS information allows for timeliness of account and access status for users of the system.

1.3.2. Gaining Access to the System (Authentication)

Gaining access to RedStorm can be broken down into the following elements:

- SARAPE
- WebCARS
- NWIS
- Account provision through the Kerberos Database (SNL accounts only)
- Account provision through the Kerberos Cross-Cell Database (Tri-Lab classified accounts only)
- Account provision through the SecurID account (Tri-Lab unclassified accounts only)
- Account provision on RedStorm

The account creation process involves SARAPE, WebCARS, and NWIS. The Kerberos provisioning elements are managed by SNL Corporate processes and are not specific to Red Storm.

1.3.2.1.   Gaining Access to the Unclassified System (authentication)

In FY05, access to the unclassified portion of Red Storm varies for local versus remote users. Those local users with access to the SRN may access the system directly from the SRN. Tri-Lab personnel, who do not have access to the SRN, are required to use a SNL SecurID card in order to provide two-factor authentication. ASC Alliance users are similarly required to use a SecurID card. In FY05, ASC Alliance and Tri-Lab users enter the Red Storm local network via the same entrance. After FY05, ASC Alliance and Tri-Lab users will enter the Red Storm local network via separate entrances.

Users with direct access to SNL's SRN can connect to Red Storm (unclassified) using ssh. The SSH client is required to be SSH version 2 compatible.   It is desirable that the SSH client also support gssapi-mic authentication to allow for Kerberos ticket authentication. If the SSH client does not support gssapi-mic authentication, the user will be required to supply a Kerberos username and password. Users may either use Kerberos tickets generated at their local site, where they are forwarded to and honored by the Red Storm system, or provide their fully qualified Kerberos username (i.e., username@dce.sandia.gov) and Kerberos password when prompted by ssh.

To access the Red Storm unclassified environment, ASC Alliance users and Tri-Lab personnel are required to use a SNL issued SecurID or CryptoCard token card to first gain access to the SRN. These users ssh to a SNL gateway, specifying the ssh login name to be their site Kerberos username, and validate their SecurID or CryptoCard pass code.   Once validated on the gateway, the user is prompted for either telnet or ssh, the host destination (System Name, e.g. redstorm.sandia.gov), and Kerberos username and password. Tri-Lab users must be able to provide their site's Kerberos username, in the form of username@site-kerberos-realm, and Kerberos password. Alliance users must use the Kerberos username and password issued to them.

An example Tri-Lab access session to unclassified Red Storm is listed below:

```
ssh -l gmachin@dce.sandia.gov srngate.sandia.gov
gmachin@dce.sandia.gov's password: (Enter passcode from Sandia
issued SecurID or CryptoCard)

Welcome gmachin to the ssh gateway

1 -  telnet session
2 -  ssh session
Q -  Quit

Enter a selection from above: 2

System Name: redstorm.sandia.gov
Username: gmachin@dce.sandia.gov
```

```
(The section below – bounded by [] will only occur the first time
the user goes to the destination)
[Host key not found from database.
Key fingerprint:
xetir-lukef-fokuf-konyh-livik-fukad-nokal-pupiv-bopeg-byhyk-nexax
You can get a public key's fingerprint by running
% ssh-keygen -F publickey.pub
on the keyfile.
Are you sure you want to continue connecting (yes/no)? yes
Host key saved to
/home/telnet_home/gmachin/.ssh2/hostkeys/key_22_redstorm.sandia.g
ov.pub
host key for redstorm.sandia.gov, accepted by gmachin Wed Dec 08
2004 17:19:48]
gmachin@dce.sandia.gov's password:(Enter Kerberos password)
```

1.3.2.2. Gaining Access to the Classified System (authentication)

Access to the classified portion of Red Storm can only be obtained via routes that provide network access to the SCN.

Secure SNL users and Tri-Lab remote users connect to the Red Storm (classified) using ssh. The SSH client is required to be SSH version 2 compatible. It is desirable that the SSH client also support gssapi-mic authentication to allow for Kerberos ticket authentication. If the SSH client does not support gssapi-mic authentication, the user will be required to supply a Kerberos username and password. Users may either use Kerberos tickets generated at their local site, where they are forwarded to and honored by the Red Storm system, or provide their fully qualified Kerberos username (i.e., username@dce.sandia.gov) and Kerberos password when prompted by ssh.

1.3.3. System Availability and Scheduling Information

System availability and scheduling information is available via links from the Red Storm Home Page. The information posted may vary between the classified and unclassified networks. Static information includes the schedule for preventative maintenance, center section swings or red/black switching, dedicated system time, dedicated application time, upgrades of software and hardware, queue configuration, and expected network outages.

Web based status and monitoring services allow Red Storm users to view queued and running jobs and the general state of the machine, i.e., number of cpus running or free. Additionally, Portable Batch System (PBS) provides the *qstat* command to display the current queue and job status. Users having ssh access can also query the state of the queues from the command line.

Web based system availability and scheduling information, as well as status and monitoring information is available only on the SRN for FY05.

Alert information from Red Storm is integrated with the Sandia Cyber Enterprise Management System (CEMS) to provide corporate production 24×7 monitoring capability. CEMS monitors the health of the networks and systems to support trouble shooting and response.

Some status information may require the use of authenticated web pages. In this case, every effort will be made to use the same login and password as used for access to Red Storm itself.

1.3.4.   Requests for Priority Jobs

The PBS-Pro queuing system is used to implement a fair share scheduling/allocation policy to manage allocations by laboratory and/or alliance center. The expectation is that Red Storm will need to supplement this PBS management system with a process for handling expedited requests for priority access. Requests can include such considerations as: increased priority over other jobs from the requestor's institution, increased priority over jobs from all institutions, ability to exceed job duration or job size limits, dedicated application time, etc. These requests will be discussed at the weekly Tri-Lab Expedited Priority Run (EPR) video-teleconference meeting. These weekly meetings are used to address similar requests for priority jobs/users on ASC White and Q. Contacts include Jim Lujan (LANL), Brian Carnes (LLNL), and Jim Ang (SNL).

Once a priority request is approved, the requesting user or group is granted rights to a PBS queue configured to meet their needs. Access to such queues can be controlled on the basis of access control lists. Once the queue is activated, a user who is allowed to use that queue may submit and run jobs without further interactions with the system administrators.

2.   Setting Up the Work Environment

2.1.   File System Usage, Paths, Environment Variables, etc.

The Red Storm directory structure, documented in Section 3.4 Addressing I/O: Effective Use of the File Systems, Serial and Parallel I/O Libraries, is available for the Limited Availability milestone. To the extent that ACE standards for path naming conventions exist and are documented, the standards are met in this directory structure.

2.2.   Setting up User Groups, Environment Variables, Modules, etc.

Red Storm is delivered with Cray's module system for managing your programming environment. Modules are used to manage environment variables and paths. In general, one needs to initialize the module system by designating your preferred shell. The default user login files will correctly initialize the modules environment. The module command loads the modules you require. The modules system also provides a way to automatically load modules when you login. The Red Storm environment supports modules for SNL specific configurations and ASC software, as well as user-defined modules. PrgEnv, a key module, establishes the environment for compiling jobs.

The Red Storm supported shells are the Bourne-Again Shell (*bash*) and the Tenex C-Shell (*tcsh*). When invoked as *sh*, *bash* restricts itself to conventional Bourne-shell semantics, and the *tcsh* shell is backwards compatible with *csh*. The Korn shell (*ksh*) is installed to support extant scripts and applications, but *ksh* is not supported as a login shell.

Table 1. Supported Shells.

Shell	Links to	Supported Login Shell
/bin/bash		Yes
/bin/sh	/bin/bash	Yes via bash
/bin/tcsh		Yes
/bin/csh	/bin/tcsh	Yes via tcsh
/bin/ksh		No

Default user login shells and groups can only be modified by Red Storm system administrators. Default login shell and groups are maintained in a read-only file system; write privileges are limited to the Red storm system administrators. Users can change their login shell and default group by sending mail to redstorm-admin@sandia.gov. Default skeleton files are available for all supported user shells.

Groups are managed using Sandia's 'Metagroups' utility (https://metagroups.sandia.gov). The 'Metagroups' utility provides a web-based interface for managing groups and group memberships. A central information repository daily propagates group configuration changes to Red Storm. The 'Metagroups' utility is available only to Sandia employees. Groups for Tri-Lab users must be created and managed by a Sandian. Classified groups must be created by a Sandia manager and can be managed by one or more Sandia employees.

3. I/O and Data Migration

ASC efforts are directed at developing and providing tools and services with the goal of enabling Distance Computing. Below are the results of these efforts in early FY05.

3.1. Red Storm-ASC Network Architecture Map

A goal of SNL, working closely with the DisCom program element, is to provide adequate tools and high-performance throughput for both local and remote users. High performance data movement over the Wide Area Network (WAN) requires common solutions based on the DisCom network architecture. Throughput issues between sites can be resolved only through coordination and cooperation among the sites.

For the Limited Availability Milestone, the system service partitions are connected to the classified and unclassified restricted environments. While the configuration for both of these environments appears to be similar, there are distinct differences. The remainder of this section emphasizes the differences and capabilities of each environment.

3.1.1. Classified Red Storm-ASC Network Architecture

The DisCom program element continues to provide 2.5 Gigabits/second (Gb/s) dedicated links connecting LLNL, LANL, and SNL. The WAN utilizes parallel path architecture to achieve a maximum 2 Gb/s aggregate throughput linking the Tri-Lab secure environments. Data movement directly between Red Storm and the visualization and HPSS servers at LANL and LLNL is an integral requirement of the Tri-Lab secure environment. The WAN bandwidth is appropriate for these modes of usage.

Data transfer to and from Red Storm is realized via the High Performance Computer Network (HPC). The HPC is based on 10 Gb/s Ethernet technology and is interconnected to the DisCom WAN by a specialized network router, which can accommodate the speed disparity between the HPC and the DisCom WAN. Normal file transfer data movement is capable of data rates of 100 MegaBytes/second (MB/s), while specially configured data transfers are capable of 200 MB/s. Performance of specially configured data transfers depend on each of the remote Tri-Lab partners providing hosts that are capable of receiving an aggregate of 200 MB/s from Red Storm. The current DisCom Bandwidth (or better) must also be available.
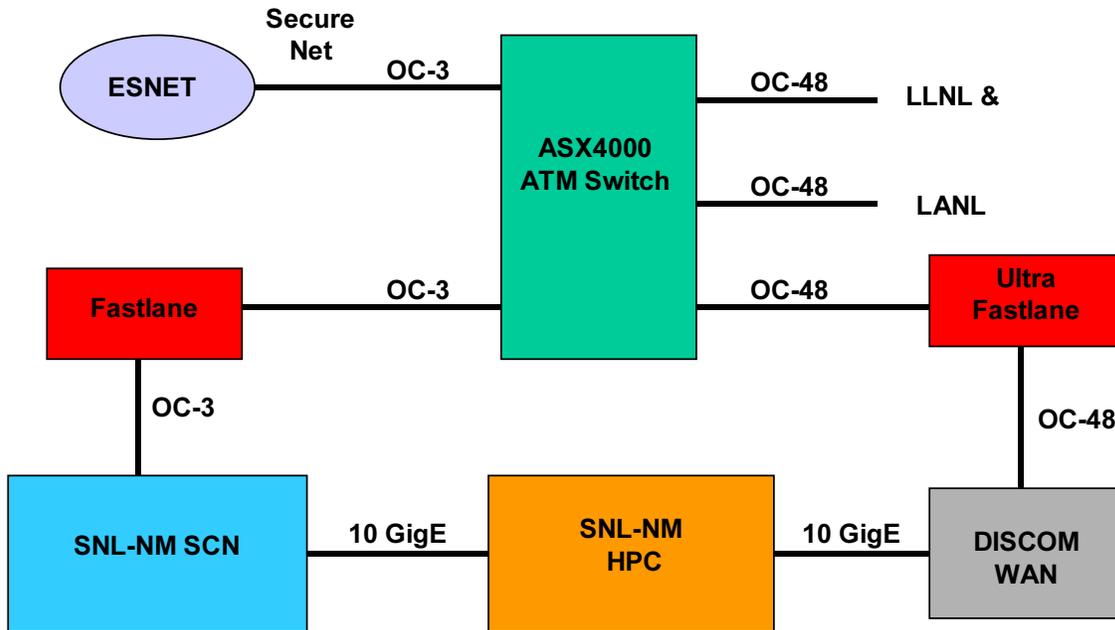
Figure 2. Classified High Performance Computer Network Connection to the DisCom WAN.

Secure, Tri-Lab remote users connect to the HPC/Red Storm system using Kerberos authentication, the process currently used for remote, secure computing. Users may use Kerberos tickets generated at their local site and forwarded to the Red Storm system where they are honored or provide their fully qualified Kerberos username (i.e., username@dce.sandia.gov) and the Kerberos password when prompted by ssh.

3.1.2. Unclassified Red Storm-ASC Network Architecture

Connectivity between LANL/LLNL, the ASC Alliance Centers, and the restricted unclassified Red Storm is different. For ASC Alliance Centers, LANL, and LLNL, connectivity is provided by SNL's standard ISP, ESNet, at 100 Megabits/second (Mb/s) maximum. As seen in Figure 3, all users outside SNL access the HPC by going through the Demarcation Zone (DMZ), the boundary between the network external to SNL and SNL's internal network. In FY05, this is accomplished via a Sandia SecurID card, with access to the HPC through SNL's SRN. To enhance performance, access directly from the DMZ to the HPC will be implemented after FY05. SecurID, or some variant, will still be required to access the DMZ.
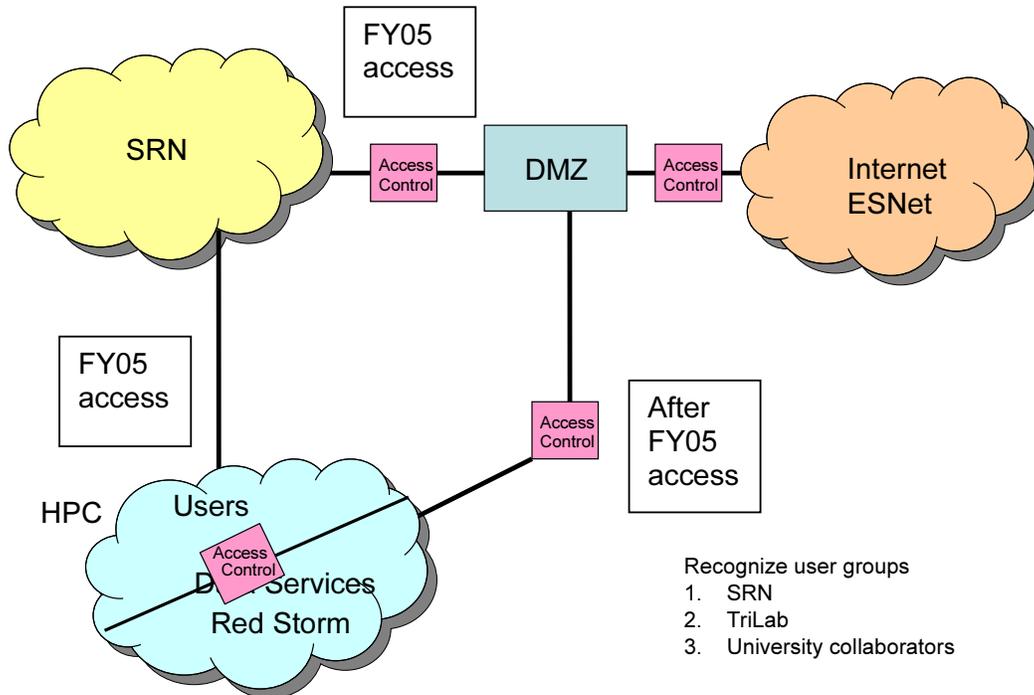
Figure 3. Unclassified Access Controls Exist for Multiple User Groups; SNL SRN, Tri-Lab, and University Users.

For LANL, LLNL and SNL/CA users, we plan to change connectivity to the higher speed DisCom WAN links after FY05. Authentication would then be done locally and the Kerberos tickets forwarded to Red Storm. Connectivity for ASC Alliance users will not change.

3.2.   Red Storm and RoSE Network Architecture

In FY05, visualization and data analysis for Red Storm are supported by the Red RoSE and Feynman clusters. Data is transferred to/from Red Storm to/from Red RoSE or Feynman using Parallel File Transfer Protocol (PFTP) (see Section 3.3 Tools for Transferring Files). The Red RoSE and Feynman clusters provide three principal capabilities in support of Red Storm: (1) Archival (tape) storage of Red Storm output; (2) Visualization and other analyses of Red Storm output (and input); (3) After FY05, direct, shared access to RoSE's large, high performance parallel file system by other mid-range computing systems at Sandia.

The RoSE clusters are designed to provide the capability to transfer output files from Red Storm's disks to RoSE's disks at sustained rates of 25 Gigabytes/second (GB/s). To achieve this, both the network communications bandwidth and the parallel file system bandwidth of each RoSE cluster are designed to be 25 GB/s or greater.

3.2.1.   Classified Red Storm to Red RoSE Networking

In FY05, classified visualization and data analysis capability is provided by the 264 visualization nodes of Red RoSE, served by a high-bandwidth Lustre parallel file system. (For additional system details, see Section 7.2.2.1 Classified Support Cluster – Red RoSE.) In FY05, network

communications capabilities will provide 25 GB/s of network bandwidth between Red RoSE and Red Storm.

3.2.2. Unclassified Red Storm to Feynman Networking

In FY05, unclassified visualization and data analysis capability is provided by the 128 visualization nodes and 100 compute nodes of Feynman, served by a high-bandwidth parallel file system. (For additional system details, see Section 7.2.2.2 Unclassified Support Cluster – Feynman.) In FY05 the network communications capabilities provide 8 GB/s bandwidth between Feynman and Red Storm, a third of the full design capability. After FY05, Feynman will be replaced by Black RoSE.

3.3. Tools for Transferring Files

The ASC Tri-Lab community supports PFTP (at LLNL, this is synonymous with the Tri-Lab and HPSS derived FTP client GSS-PFTP) for high performance parallel data transfer via the classified network. Since DFS is being phased out, a NFS/DFS gateway may be deployed after FY05 to provide similar functionality to DFS across the Tri-Lab environment. When transferring entire files locally or remotely, the user can expect the best performance from PFTP. Furthermore, PFTP performs the authentication if a valid Kerberos credential is available.

Scripts are provided to transfer files to systems between sites. These scripts check for a Kerberos credential, and if present, run PFTP to the appropriate remote machine using the credential. Therefore, interactive authentication is not required (e.g., entering a user name and password). If a credential is not found, the script returns an error. At SNL on Red Storm, scripts are installed in */usr/local/bin* that simplify the commands for users. Example scripts include pftp2b21, pftp2b22, pftp2lanl, pftp2edison, and pftp2llnl (e.g., LANL B21, LANL B22, LANL HPSS, SNL Edison, and LLNL HPSS).

PFTP for classified Red Storm includes a multi-node client and a single-node server. Future capabilities (after FY05) include a multi-node server. The PFTP Client and Server provide high performance data transfer between classified Red Storm, the classified RoSE cluster, and other Tri-Lab classified systems. The PFTP Client and Server also provide high performance data transfer between unclassified Red Storm and the unclassified RoSE cluster. It may not be possible to transfer data between unclassified Red Storm and unclassified Tri-Lab systems.

A fast data connection exists between Red Storm and RoSE/Feynman. Data can be transferred from Red Storm to/from RoSE/Feynman using the PFTP Client on Red Storm and the DisCom PFTP Server on RoSE/Feynman. The reverse is also true – data can be transferred from RoSE/Feynman to Red Storm using the PFTP Client on RoSE/Feynman and the DisCom PFTP Server on Red Storm.

The Sandia Data Tools and Components project supports two tools, available on RoSE/Feynman, for transferring files between Red Storm and RoSE/Feynman. The *dsacp* tool provides transfer of directory hierarchies of data to and from the SNL Data Services Archive (DSA) HPSS tape drives (STK 9940B) attached to Red RoSE and Feynman. The *zephyr* tool provides transfer of directory hierarchies of data between supported systems. Initial supported classified systems

include Red Storm, Red RoSE, and edison. Initial supported unclassified systems include Red Storm, tesla, and Feynman. The *dsacp* and *zephyr* tools must be executed on RoSE/Feynman.

A *dsabrowse* utility provides ftp-based browsing and management of data on the SNL DSA HPSS. The *dsabrowse* utility must be executed on RoSE/Feynman. Information on the *dsacp* and *dsabrowse* tools is available at http://www-irn.sandia.gov/DSA/.

3.3.1.   PFTP Classified Data Transfer Tools

On Red Storm, there is a PFTP Client that can connect to any PFTP Server, via the classified network, including both an HPSS PFTP Server and a DisCom PFTP Server. The PFTP Client can connect to both types of PFTP Servers, via the classified network, at LANL and LLNL. The PFTP Client can connect to DisCom PFTP Servers on other classified SNL systems.

On Red Storm, there is also a DisCom PFTP Server. Users on classified systems at LANL, LLNL, and on other classified SNL systems can use a local PFTP Client to connect to this server on Red Storm.

On Red RoSE, there is a PFTP Client which can connect to any PFTP Server, via the classified network, including both an HPSS PFTP Server and a DisCom PFTP Server. The PFTP Client can connect to both types of PFTP Servers on classified systems at LANL and LLNL. The PFTP Client can connect to DisCom PFTP Servers on other classified SNL systems.

On Red RoSE there is also a DisCom PFTP Server. Users on classified systems at LANL, LLNL, and on other classified SNL systems can use a local PFTP Client to connect to this server on RoSE.

3.3.2.   PFTP Unclassified Data Transfer Tools

On the unclassified Red Storm, there is a PFTP Client that can connect to DisCom PFTP Servers on other unclassified SNL systems. It may not be possible to transfer data between unclassified Red Storm and unclassified Tri-Lab systems.

On the unclassified Red Storm, there is also a DisCom PFTP Server. Users on other unclassified SNL systems can use a local PFTP Client to connect to this server on Red Storm.

On Feynman, there is a PFTP Client which can connect to any PFTP Server on other unclassified SNL systems, including both an HPSS PFTP Server and a DisCom PFTP Server. It may not be possible to transfer data between Feynman and unclassified Tri-Lab systems.

On Feynman, there is also a DisCom PFTP Server. Users on other unclassified SNL systems can use a local PFTP Client to connect to this server on Feynman. It may not be possible to transfer data between Feynman and unclassified Tri-Lab systems.

3.3.3.   SNL DSA HPSS Data Transfer Tools

Sandia's DSA HPSS is attached directly to RoSE/Feynman. The parallel file systems on Red RoSE and Feynman provide staging areas that allow data to be transferred rapidly from Red Storm, and then more slowly to tape. The only way to transfer data to and from the SNL DSA

HPSS, is using *dsacp*. The user can login to RoSE/Feynman and use *dsacp* to transfer data between Red Storm or RoSE/Feynman and SNL DSA HPSS.

To transfer data from the classified SNL DSA HPSS to classified systems at LANL or LLNL, the user can login to RoSE and use *dsacp* to retrieve the data onto RoSE. Then PFTP can be used to move the data to LANL or LLNL.

To transfer data from the unclassified SNL DSA HPSS to other unclassified systems at SNL, the user can login to Feynman and use *dsacp* to retrieve the data onto Feynman and then move the data to the target system using PFTP. It may not be possible to transfer data between unclassified Feynman and unclassified Tri-Lab systems.

The syntax of *dsacp* is similar to *scp*. For example, to transfer data to the SNL DSA HPSS:

- *dsacp* source-host:source-directory-or-filepath hpss:hpss-directory-or-filepath

And for example, to retrieve data from the SNL DSA HPSS:

- *dsacp* hpss:hpss-directory-or-filepath dest-host:dest-directory-or-filepath

- A "-checksum" option is available; the default is no checksum calculation.

The SNL DSA HPSS tape drives attached to Red RoSE and Feynman each have one silo. Capability of the silos after FY05 includes a capacity of 2 petabytes (1 PB each) and a bandwidth from disk to tape of about 0.5 GB/s. Red RoSE and Feyman have one "mover" node for each two tape drives. The mover nodes have dual 2.0-GHz Xeon processors with 2.0 GB of memory and two 2-Gb/s Fibre-Channel connections to their respective tape drives.

3.3.4.   Long Term Storage Policies

Users on Red Storm accessing SNL's DSA HPSS data storage require network bandwidth to DSA HPSS, and storage capacity on DSA HPSS. The network bandwidth has been provided based on the requirements of the Red Storm machine. The DSA HPSS capacity requirements are established by the number of users and how they use DSA HPSS. SNL is adopting the Tri-Lab policy to allow remote Tri-Lab users to access and use SNL's DSA HPSS. A resource management policy will be developed and publicized after FY05.

3.4.   Addressing I/O: Effective Use of the File Systems, Serial and Parallel I/O Libraries

3.4.1.   Effective Use of the File Systems

Red Storm's file systems are not shared with any other computing platform. High-speed file transfers that exploit Red Storm's network connectivity to other systems, such as the RoSE clusters, are managed by scripts and commands provided by Sandia's Data Tools and Components project (see Section 3.3 Tools for Transferring Files).

There are no limitations in Catamont for the number of files that a process can open. Under Linux, the limit will initially be set to at least 600,000. The determining factor will be the Lustre file system performance.

After FY05, Red Storm may import files via NFS from other systems. NFS-mounted file systems will be visible only to Red Storm service nodes. NFS-mounted files will not be visible to processes running on the compute nodes. Data from NFS-mounted file systems must be moved into local Red Storm file systems prior to using them during a parallel computation.

Table 2. File System Usage.

	Home	Projects	Lustre Parallel FS
Path	/home	/projects	/scratch*
Availability	Login, compute	Login, compute	Login, compute
Parallel I/O	No	No	Yes
Large data, many files	No	No	Yes
Small data, few files	Yes	Yes	Yes
Backups	Yes*	Yes*	No
Purge schedule	None	None	TBD
Maintenance	Yes	Yes	Yes

The Lustre parallel file system is available for /scratch* as shown in Table 1. /home and /projects are served via NFS. After FY05, temporary snap shots or backups may be created for the /home and /projects file systems in the classified environment upon user request (Yes* in Table 2). These backups will not leave the machine and are only provided for convenience. Long-term backups are the responsibility of the user. The purge policy for the /scratch* file system will be published at a later date. These resources are available to applications running on Red Storm and not to be used as long or even medium term storage. Hence, when a problem completes, the user is expected to copy the files elsewhere and then remove the majority of their files from Red Storm.

After FY05, there may be NFS directories, mounted from other systems and accessible only from login nodes. There is no current plan to mount the /home and /projects directories on any machines other than Red Storm. The /scratch* parallel file system performs best with large buffer (4MB or larger) writes and reads. In the future (after FY05), there may be one or more alternative parallel file systems tuned for different types of usage.

Documentation of the Red Storm file systems and I/O libraries will be available on Red Storm Home Page by July 2005. Cray has no plans to provide a specific online tutorial on I/O, although some of the material in the I/O documentation may be tutorial in nature. In the future (after FY05), special-purpose tutorials and papers may be developed and posted on the Red Storm web site on an as-needed basis.

3.4.2. Parallel I/O Libraries

After FY05, the CRAY MPICH MPI-IO library, which implements the MPI-IO portions of the MPI-2 standard, will be provided on Red Storm. Also after FY05, parallel HDF5 libraries will be provided on Red Storm. Parallel HDF5 and MPI-IO libraries support the Sets And Fields (SAF) and other data formats. In FY05, the SNL Engineering Analysis Code Access System (SEACAS) I/O performance was optimized; after FY05, SAF I/O performance will be optimized.

4. Application and System Code Development

Users will have access to the Red Storm system for development, debug and smaller production runs (all considered "routine" work) in preparation for capability computing (considered to be "priority" work). The user activity of submitting jobs is described in Section 6.1 Submitting the Job.

Applications and projects that require their own project-managed storage will be provided with directories in the /projects file system. These 'project' directories will be created using normal user and group permissions that will allow designated users to manage their own subdirectories. Application teams can use their project directories to install and update their codes. Projects that provide applications to the Red Storm user community are expected to manage their installations using the modules facilities of Red Storm (see Section 2.2 Setting up User Groups, Environment Variables, Modules, etc.).

CVS access can be achieved from Red Storm via *ssh*. The user must set two environment variables, "CVS_RSH" and "CVSROOT". For example:

```
export CVSROOT=:ext:<system name>:<directory path of CVS
repository>
export CVS_RSH=/usr/bin/ssh
```

4.1. Peculiarities of the System

Red Storm is modeled on ASC Red, Sandia's first large ASC computing system. Like ASC Red, Red Storm features Red/Black switching, a UNIX programming environment, parallel and sequential file systems, diskless compute nodes, MPI-based message passing, and the use of a lightweight kernel on compute nodes. Unlike ASC Red, Red Storm features a large service partition composed of nodes running Linux. Service partition nodes are assigned to one of several different classes based on their hardware configuration and their intended use. For instance, login nodes provide network connectivity to the outside networks (the SCN) via local network interfaces and are configured to provide compilation, job management, and debugging tools. Users will be able to log into different login nodes, and some commands like 'ps' will report on all login node processes. In contrast, the I/O disk nodes are configured with high-speed disk connections, and are configured to run as Lustre file system components. End users will not have direct access to the I/O disk nodes.

4.1.1. Lightweight Kernel

SNL has a long tradition of building and supporting massively parallel computing systems that use diskless compute nodes and a lightweight kernel. As system scale increases, the need for stable and reliable operation requires that the individual components be as simple and therefore as minimal as possible. By eliminating virtual memory, computationally intensive applications can obtain the highest memory access and network performance.

Lightweight kernels are minimal operating systems that are more akin to resource managers. Properly written, a lightweight kernel provides only enough services and requires only enough resources to allow the application to run with maximal performance. This philosophy implies

that some conventional services provided by UNIX or Linux are omitted by a lightweight kernel. For example, the Red Storm lightweight kernel does not support

- fork(), exec(), system(), or dynamic process management,

- TCP/IP protocol stacks (MPI is used instead),

- UNIX-style interprocess communication (pipes, fifos, etc.),

- daemon-based programming models (programs that wait for system events, especially I/O or network events other than MPI),

- Dynamic linking (job startup will not scale),

- User logins to compute nodes.

In return, the application is granted access to the processors, memory, and interconnects where it runs; memory allocation rewards large block allocations; and once your job launches it will run in a predictable period of time.

The lightweight kernel on Red Storm is called 'Catamount'. It is an updated release of 'Cougar', the kernel used for ASC Red. Writing code for a lightweight kernel means being cognizant of the library calls and services that the kernel supports. Most POSIX calls, aside from dynamic linking and process management, are supported by Catamount. Full documentation can be found on the Red Storm Home Page.

4.1.2.   Linux in the Red Storm Environment

Service nodes on Red Storm run a diskless incarnation of SuSE Linux. The initial deployment of Red Storm will use a pre-release variant of SuSE Enterprise Server 9.0. Without a disk, processes on a Red Storm Linux node will not be able to use swap space other than that available as a portion of main memory. Memory exhaustion, then, is a real possibility. The various classes of Linux nodes have been configured with this limitation in mind. Other than the lack of a local disk, service nodes are essentially full, standard distributions of SuSE Linux.

Parallel job launch on Red Storm differs from a conventional Linux cluster because the nodes involved have to be allocated from different pools. Every PBS job requires at least one service node. A service node will be designated by PBS to run a process (yod) which launches executables onto other compute and service nodes and manages the parallel job. After FY05, users can request multiple service nodes if their job requires them.

The architected model for allocating service nodes to logins relies on a dynamic allocation of login nodes to users. The allocation is based on minimizing the workload across the set of login nodes. Service nodes, including login nodes, may be shared with other users. Service nodes, other than the one running the yod command, are allocated by the Compute Processor Allocator (CPA).

The CPA allocates compute nodes and service nodes, other than the distinguished service node allocated by PBS. The CPA is consulted by PBS and by yod. This design echoes the environment on ASC Red. The list of compute nodes and service nodes allocated to a job can be retrieved by running the utility, xtshowmesh.

Users can log into the Linux (service) nodes involved in the job, but not the compute nodes. In addition, PBS will report only the service nodes in the usual PBS node lists; the list of compute nodes can be retrieved by running the utility, xtshowmesh. After FY05, external access to Red Storm will ordinarily pass through a load-balancing allocator. As a result, a user will ssh into one of many login nodes. Once logged in, any user can ssh to other login nodes over the high-speed mesh. Users can redirect STDIN and STDOUT to a file.

4.1.3.   Parallel Applications on Red Storm

Parallel applications can use several architectures. Overall Red Storm is tuned for:

- A single parallel application (managed by a single yod) is launched onto one or more compute nodes. In this scenario, a single executable, compiled and linked to run on Catamount is run on one or more compute nodes. All of the compute nodes run the same executable. The application instances communicate with each other using MPI. Standard I/O (i.e., stdout, stderr) is redirected back to the service node from which the job launched, where that job's yod process is running. In contrast, file I/O is routed directly to Lustre I/O servers.

Other application architectures, available after FY05, include:

- A single parallel application (managed by a single yod) launches different executables on different compute nodes.

- A single parallel application consists of several Linux processes running on different service nodes. In this case, Red Storm would look very much like a Linux cluster, except for low-level differences in job launch. In this case, the processes can use Linux IPC, TCP/IP over the high-speed mesh, or MPI over the high-speed mesh. While not intended for end-user computation, Red Storm system software such as PFTP exploits this capability. Since there are relatively few service nodes able to run user processes (login nodes, and planned classes of nodes such as visualization and dedicated 'yod' nodes), the size of end-user parallel jobs is limited.

  Do not mistake the service partition of Red Storm for a conventional Linux cluster! While the two architectures share many similar attributes, the Red Storm environment will be sufficiently different that porting effort will be required. Sandia is committed to making any such porting effort as simple and clean as possible.

  For example, normal mpi-based scripts such as mpirun or mpiexec will not be supported. Instead, the *yod* command will be used for parallel execution. Similarly, to accommodate the large sizes of Red Storm jobs, the PBS interface to shell scripts has been altered to avoid creating large files of node names (e.g. PBS_NODEFILE is not set). Finally, most

of the Linux-based service nodes in Red Storm have well defined and tightly architected roles in the system; these nodes will not be available for general user computations.

- A single parallel application consists of multiple Linux processes (on service nodes) and multiple lightweight kernel processes running on compute nodes, all communicating via MPI. The Linux-based processes can also use Linux IPC.

- Multiple yods, possibly launching different executables, share a set of nodes. The processes on each compute node are independent. They do not share the same MPI communication world. Initially, each compute node can support up to four user processes, subject to memory availability. The intention to share each compute node must be specified at job launch.

4.2. Parallel Programming Models and Run-time Systems

MPI is supported on Red Storm via the MPICH-2 implementation running over a Cray-developed Portals channel. Some functions introduced in the MPI-2 standard are not implemented for FY05. Unsupported functions include those in Chapter 5 (Process Creation and Management) of the MPI-2 document:

```
int MPI_Close_port();
int MPI_Comm_accept();
int MPI_Comm_connect()
int MPI_Comm_disconnect();
int MPI_Comm_get_parent();
int MPI_Comm_join();
int MPI_Comm_spawn();
int MPI_Comm_spawn_multiple();
int MPI_Lookup_name();
int MPI_Open_port();
int MPI_Publish_name();
int MPI_Unpublish_name();
```

The README for MPICH-2 mentions that the external data representations for the I/O functions are currently not supported. There are also known limitations – for example, the maximum number of outstanding posted receives and maximum number of unexpected messages.

4.3. Third Party Libraries and Utilities

SEACAS is a collection of structural and thermal codes, data format and I/O libraries, and utilities used by analysts at SNL. The system includes pre- and post-processing tools, analysis codes, database translation tools, support libraries, shell scripts, and an installation system. The support tools for data pre- and post-processing and data manipulation include Algebra, Aprepro, Blot, Gen3D, Grepos, Grope, Mapvar, nem_slice, nem_spread, and epu (nem_join), to name a few. Support libraries include Chaco, ExodusII data model, and NetCDF I/O library. Additional information may be found at http://jal.sandia.gov/SEACAS.

4.3.1. Math Libraries

Library	Description
libm	Math intrinsics from glibc
libacml	ACML: AMD Core Math Library containing BLAS levels 1, 2, and 3; LAPACK; and FFTs in -, double-, single-complex and doublecomplex data types
ScaLAPACK	Distributed memory LAPACK, based on BLACS
SuperLU	Distributed memory sparse solvers
Trillinos	SNL mathematical solver library. The Trilinos Framework provides the infrastructure and tools to support research, development, collaboration, delivery and ongoing support of mathematical solver libraries at Sandia. It provides tools for software quality engineering processes including source management, issue-tracking and collaboration tools and also provides software libraries for basic distributed memory linear algebra computations and APIs for advanced solver capabilities.

4.3.2. Networking and other Libraries

Library	Description
MPICH2	Message passing library implementing MPI-2 functions. (MPI-IO functionality to be delivered after FY05)
JDK	Java Developer Kit available on service nodes.
JRE	Java Runtime Environment available on service nodes.

4.4. Compilation

Compiler	Description
PGI compilers for C, C++, F77, F90	Compliant with their respective language standards. Compilation to lightweight kernel nodes (compute partition) is supported. Compilation to linux nodes (service partition) is supported. PGI compilers are accessible via the modules environment (see Section 2.2 Setting up User Groups, Environment Variables, Modules, etc.). A cross compilation environment is available for users of the PGI compilers and compute nodes.
Python	Installed on Linux nodes of the service partition.
Perl	Installed on Linux nodes of the service partition.

4.5. Debugging and Correctness Testing

TotalView is the primary debugger for Red Storm, and will be supported by Cray. TotalView provides a full featured parallel debugging environment. Full online documentation is available on the Web (www.etnus.com). In FY05, TotalView is capable of debugging an application running on 64 nodes. After FY05, TotalView will be capable of debugging an application running on 1024 nodes.

Valgrind is an open-source memory debugging tool that is popular for providing detailed memory correctness checking on Linux/IA32 platforms. Work has begun to port Valgrind to AM64 platforms. In late FY05, work on a lightweight kernel port and MPI awareness will begin. After FY05, it should be possible to use Valgrind on Red Storm.

4.6.   Performance Measurement, Analysis, and Tuning

Application performance measurement analysis and tuning can be broken down into two major areas: CPU utilization profiling and MPI utilization profiling. Three methods for performing MPI profiling are planned for Red Storm. Two of these enable full profiling of the MPI timeline: Vampir and Apprentice. Apprentice is a Cray tool that Cray will port to Red Storm and maintain. Vampir is owned by Intel, and its availability for Red Storm is subject to Intel licensing and porting constraints. We will continue to make an effort to make Vampir available for Red Storm for the following reasons: (1) It is available for other SNL platforms such as ICC, as well as ASC platforms at the other labs; (2) Users already have experience with it; and (3) ASC has funded enhancements to Vampir that improve its usability on large scale machines. The third MPI profiling tool is mpiP and will be available on Red Storm after FY05. mpiP collects aggregated information about MPI usage.

CPU profiling involves collecting information about which lines, or regions of code, are using processor resources.   Time-based profiling is the most fundamental technique, and this can be supplemented by hardware event counter based profiling to give a more detailed look at the code. Red Storm implements the Performance API (PAPI), which can be used directly by applications to collect hardware performance counter information. After FY05 the SNL application, VProf, should work on Red Storm by using PAPI to easily allow users to see profiling information organized by lines, files, and routines.

In summary, after FY05, the PSE supported tools, VProf, mpiP, and Vampir, will be deployed on Red Storm. Cray supported tools, PAPI V2, and Cray Apprentice 2 are available in FY05. After FY05, PGI profiling tools will be available, probably for Linux only.

5.   Problem Setup

Problem setup includes generating meshes, domain decomposition, materials databases and constitutive models, model/assembly management, visualization of meshes, and archiving of solid models. At SNL, the Design through Analysis Realization Team (DART) focuses on issues relating to the problem setup environment. While the Limited Availability user environment provides system resources and network connectivity for the generation of meshes/netlists/ schematics, domain decomposition, material databases, and application databases, the deliverables of the DART effort are not tied to the Red Storm Limited Availability Milestone. For specific information on the DART deliverables, refer to the SNL DART Milestone. Additional information on DART activities can be found at the Analyst Home Page (www-irn.sandia.gov/analyst/).

5.1.   Domain Decomposition

Prior to running a simulation on a parallel computer, the simulation must first be decomposed into tasks assigned to different processors. Efficient use of the machine requires that each processor have approximately the same amount of work to accomplish and that the quantity of inter-processor communication remains small. Finding an optimal decomposition is very difficult, but due to its practical importance, a great deal of effort has been devoted at SNL to developing heuristics for this problem. A variety of algorithms for graph partitioning have been researched, developed, and implemented, and they include *Chaco*, *Zoltan*, and *Metis*. SNL specific codes using these libraries are *nem_slice*, *nem_spread*, and *yada*.

6.   Running the Application to Solve the Problem

If the problem is expected to use a significant portion of the Red Storm system or receive a higher service priority than dictated by the user's fair-share allocation (what we are calling a priority calculation), policies must be followed to first get permission to acquire those resources (see Section 1.3 Gaining Access to the System). The problem can then be submitted as a batch job as described in Section 6.1 Submitting the Job.

The application run-time environment for the Red Storm system offers many tools (e.g., MPI) and resources. The tools, resources, recommended usage strategy, and characteristics of the Red Storm system, are documented on the Red Storm Home Page. InFY05, a simple queue structure is provided. After FY05, as more is learned about loads and machine characteristics, the queue structure will evolve to meet the requirements.

6.1.   Submitting the Job (Local and Remote)

SNL provides two mechanisms for submitting jobs: PBS-Pro in FY05 and ASETS after FY05. Red Storm is delivered with PBS-Pro, which has a number of enhancements over PBS. This is the basic tool for submitting jobs, and enforces the policies for running "routine" jobs at SNL. Documentation for PBS Pro can be found on the Red Storm Home Page. For detailed information on PBS, please consult the following sources: the Cray Xt3 Programming Environment Users Guide, which provides an overview of PBS usage, and the PBS Pro 5.3 User's Guide, which provides the details of usage for a standard PBS Pro 5.3 installation.

After FY05, the Red Storm queuing structure enforces the ASC Capability Usage Performance (CUP) policy requirement on Red Storm. The CUP policy states that 80% of the node-hours of utilization must be allocated to jobs that run on 40% or more of the system. Red Storm enforces the CUP policy through the use of four job queues listed in order of increasing priority:

- **Standard.** Any user may submit jobs to the Standard queue. Jobs submitted to the standard queue must use less than 40% of the available nodes. The standard queue will be configured as the "default" queue.

- **Express.** The express queue exists to fast-track ordinary jobs under exceptional circumstances. Only users who have been added to an access control list for the express queue may submit jobs to this queue. Access is granted by the Red Storm system management team. Jobs submitted to this queue must also use less than 40% of the available nodes. All jobs in the express queue will be considered for execution by the PBS scheduler before any job in the Standard queue.

- **Large.** Any user may submit jobs to the Large queue. Jobs submitted to the Large queue must use at least 40% of the section's nodes. All jobs in the Large queue will be considered for execution before any job in the Express queue.

- **Expedited Priority.** The Expedited Priority queue is for any job which is deemed to be extremely urgent. Admission to the Expedited Priority queue is granted by the Expedited Priority Run committee. Jobs submitted to this queue are not subject to any job size or job duration limits. All jobs in the Expedited Priority queue will be considered for execution before any job in the Large queue.

Table 3. Queue Structure.

Summary of Queue Structures			
Queue	**Who may submit?**	**Job Size**	**Gatekeepers**
Standard	Any user	<40% of partition	None
Express	Approved users or teams	<40% of partition	Red Storm Managers
Large	Any user	>=40% of partition	None
Expedited Priority	Approved users or teams	No limit	Tri-Lab EPR Committee

As more knowledge about machine loads and characteristics is gained, the queue structure will evolve to meet those requirements.

After FY05, the DART Job Submission team may provide an alternative solution called ASETS, based on LSF. The ASETS system is a distributed framework for the construction of analyst-centric simulation execution and control applications, called services. The goal of ASETS is to assist the analyst in running simulation codes in the Red Storm environment by normalizing access configurations (login profiles), providing automated task sequencing of analyst activities, where possible, and by allowing the analyst to manage the running of simulation codes through a simple-to-understand interface tailored to their specific visual model of the simulation process.

Tri-Lab use of ASETS, based on interest and funding, will be evaluated after full deployment to the SNL users on Red Storm (classified and unclassified) after FY05. Use of the LSF command line interface does not require the purchase of an LSF license.

6.2. Monitoring Job Status

To monitor job status, *xtshowmesh* and *qstat* are installed and operational for the Limited Availability Milestone. *xtshowmesh* diplays a list of the compute and service nodes allocated to a job. To check the status of a job from the command line, use the *qstat* command. *qstat* by itself lists those jobs in the queue that you have submitted. The output of *qstat* indicates the state of the job, one of Exiting (E), Held (H), Queued (Q), Running (R), in transition (T), Waiting for its requested start time (W), or Suspended (S). Other options can be found in the *qstat* man page.

As specified in Section 1.2.2 On-line System Information, job status will be available on the SRN via the web for Red Storm users, without the need to log into the Red Storm System. Some elements of web based monitoring may not be available until after FY05.

Upon job submission through PBS, the standard output and standard error file descriptors are directed to files located in a directory visible only to PBS. At the conclusion of the job, those files are copied back into the directory from which the job was launched. If a user submits a job, and that job writes to the standard output or standard error without redirecting them, the results

will not be visible until completion of the job. However, if a user submits a job, and within the job submission script standard input or standard output are redirected to files, then the contents of those files are visible as the script is running. The contents of all other files are visible during job execution.

6.3.   Stopping the Job

PBS Pro provides command-line job management tools. From the command line, use the *qdel* command to remove a job from the queue. This command should remove a job from the queue, even if it is running. However, for running jobs, you may need to send a kill signal to the job via the *qsig* command. Both commands use the PBS job id that was assigned when you submitted the job. The job id can be found by using *qstat*.

After FY05, ASETS may provide for the stopping of jobs through the use of the *bkill* LSF command and from the graphical interface system. *bkill* supports the use of a PBS job id to identify which Red Storm job to stop. The graphical interface has an implied job id defining which job to stop, so that the analyst will not have to provide one. For both methods, ASETS can determine which PBS method (*qdel* or *qsig*) to use on behalf of the analyst.

6.4.   Interactive Use

After FY05, a small interactive partition of approximately 100-200 nodes will be provided for quick turnaround time to allow users the ability to run short jobs to accomplish code testing, input validation and verification, interactive debugging sessions, etc. These jobs will likely be limited to ~16 nodes and 1-2 hour job durations. Large interactive jobs are supported via dedicated application time. A policy will be in place for requesting and obtaining dedicated time for large-scale debugging by requesting dedicated application time via the Expedited Priority Run committee described in Section 1.3.4 Requests for Priority Jobs.

6.5.   Adapting the Job for Expected System Reliability

All applications on Red Storm that run for more than 2 hours should provide a restart capability. Restart allows you to better cope with expected machine outages, especially early in the lifetime of the resource. Restart also allows the operations staff opportunities to better manage the machine in case jobs need to be preempted.

Guidance for adapting jobs to Red Storm will be part of the online documentation.

7.   Processing Simulation Output

Visualization and data analysis for Red Storm are supported by the Red RoSE and Feynman clusters. After FY05, Feynman will be replaced by the Black RoSE cluster. These clusters provide three principal capabilities in support of Red Storm: (1) Archival (tape) storage of Red Storm output; (2) Visualization and other analyses of Red Storm output (and input); (3) After FY05, direct, shared access to RoSE's large, high performance Lustre parallel file system by other mid-range computing systems at Sandia.

In FY05, classified visualization and data analysis capability is provided by the 264 visualization nodes of Red RoSE, served by a high-bandwidth Lustre parallel file system. (For system details, see Section 7.2.2.1 Classified Support Cluster – Red RoSE.) In FY05, Red RoSE's network communications capabilities provide 25 GB/s bandwidth, about a third of the full design capability.

In FY05, unclassified visualization and data analysis capability is provided by the 128 visualization nodes and 100 compute nodes of Feynman, served by a high-bandwidth parallel file system. (For system details, see Section 7.2.2.2 Unclassified Support Cluster – Feynman.) In FY05, the network communications capabilities provide 8 GB/s bandwidth, a third of the full design capability.

After FY05, RoSE's I/O rates should provide the capability to transfer output files from Red Storm's disks to RoSE's disks at sustained rates of 25 GigaBytes/second (GB/s). To achieve this, both the network communications bandwidth and the parallel file system bandwidth of each RoSE cluster are designed to be 25 GB/s or greater. After FY05, RoSE's visualization capabilities should enable highly interactive visualization of output data from a 300-million cell simulation. After FY05, each cluster will have 264 dual-processor visualization nodes, each with PCI-Express graphics cards, and these visualization nodes will have direct access to the 25-GB/s file systems.

7.1.   Prepare Data for Analysis

In FY05, it is recommended that data manipulation be performed on the Red RoSE and Feynman clusters. To work on the Red RoSE and Feynman clusters, data is transferred to/from Red Storm to/from Red RoSE or Feynman using PFTP (see Section 3.3 Tools for Transferring Files). The manipulation of data may take place before, after, or during job computation. Examples of this include changing the format of the data, segmentation of the data into smaller pieces, recombining the data, etc. The need for data manipulation is dependent upon the code being run and the size of the job.

After FY05, if available, data manipulation may also be performed on Red Storm visualization nodes.

SEACAS is a collection of structural and thermal codes, data format and I/O libraries, and utilities used by analysts at SNL. The system includes pre- and post-processing tools, analysis codes, database translation tools, support libraries, shell scripts, and an installation system. The support tools for data pre- and post-processing and data manipulation include Algebra, Aprepro,

Blot, Gen3D, Grepos, Grope, Mapvar, nem_slice, nem_spread, and epu (nem_join), to name a few. Support libraries include Chaco, ExodusII data model, and NetCDF I/O library. Additional information may be found at http://jal.sandia.gov/SEACAS.

A Data Services Toolkit (*dstk*) is available in FY05 on the RoSE clusters. *Dstk* is a python scripting-based tool for query, output, extraction, subset and manipulation of results data. *Dstk* supports SAF and ExodusII database formats. Information on *dstk* is available at http://www-irn.sandia.gov/DSTK/.

7.2. Analyze the Results

7.2.1. Visualization

The visualization capabilities of Red Storm vary depending on the existence of visualization nodes, service nodes with full OS, or no nodes with full OS. In FY05 the one possibility for visualization is:

a) No Red Storm nodes for visualization

Data generated on Red Storm is moved at high speeds to the RoSE clusters where EnSight and Para View are both available as servers.

After FY05 ParaView and EnSight may have the additional capability of using viz nodes with graphics cards to render images:

b) Red Storm with viz nodes containing graphics cards

ParaView runs in client/server mode on these types of nodes and streams geometry **or** images to the desktop/tile display. EnSight runs in client/server mode on these types of nodes and streams geometry to the desktop/tile display.

7.2.2. Support Clusters for Visualization and Data Analysis

7.2.2.1. Classified Support Cluster – Red RoSE

In FY05, visualization and data analysis capability is provided by the 264 visualization nodes of Red RoSE, served by a high-bandwidth Lustre parallel file system running over InfiniBand. The visualization nodes each have dual 3.6-GHz Xeon EM64T processors, 4.0 GB of 800-MHz memory and a PCI-Express graphics processor with an additional 256 MB of memory. The aggregate processing power of the 528 Xeon processors is 3.8 teraflops. All of Red RoSE's nodes are interconnected with an InfiniBand-4X (1+1 GB/s) network. Red RoSE's operating system is Red Hat Enterprise Linux. Maui/OpenPBS is used for scheduling. Intel compilers are the default; Gnu and PGI compilers are also available. TotalView provides a parallel debugging capability. MVapich and mpiexec provide the message passing capability.

Red RoSE's Lustre parallel file system is central to the operation of Red RoSE and is used in all the major functions. In the archival storage of output from Red Storm, it provides a staging area that allows data to be transferred rapidly from Red Storm, and then more slowly to tape.

In FY05, Red RoSE's network communications capabilities provide 25 GB/s bandwidth, about a third of the full design capability. This capability enables rapid transfer of datasets from (and to) Red Storm, minimizing the time users must wait for their data to become available on Red RoSE. High-bandwidth network communications are also used for simultaneously delivering visualization output to multiple users.

After FY05, Red RoSE should provide the capability to visualize output from a 300-million cell simulation at interactive rates. It will have about a third of the I/O capability, and half the tape storage capability planned for the complete system. Completion of the system depends on need and future funding availability.

7.2.2.2. Unclassified Support Cluster – Feynman

In FY05, visualization and data analysis capability is provided by the 128 visualization nodes and 100 compute nodes of Feynman, served by a high-bandwidth parallel file system. The visualization nodes have dual 2.0-GHz Xeon processors, 1.0 GB of memory and a graphics processor. The compute nodes have dual 2.4-GHz Xeon processors with 2 GB of memory. These nodes are interconnected with a Myrinet-2000 network. Feynman's operating system, like Red RoSE, is Red Hat Enterprise Linux. Maui and OpenPBS are used for scheduling. Intel compilers are the default, Gnu and PGI compilers are also available. TotalView provides a parallel debugging capability. Mpich and mpiexec provide the message passing capability.

As in Red RoSE, the parallel file system is central to the operation of the cluster and is used in all the major functions. In the archival storage of output from Red Storm, it provides a staging area for rapidly transferring data from Red Storm. In visualization and data analyses, very large data sets utilize the large capacity and high bandwidth of the file system.

In FY05, the network communications capabilities provide 8 GB/s bandwidth, a third of the full design capability. This capability enables rapid transfer of datasets from (and to) Red Storm reducing the time users wait for their data to become available for visualization and analysis. High-bandwidth network communications are also used for simultaneously delivering visualization output to multiple users.

In FY05, the Feynman cluster provides initial support for Red Storm, until the Black RoSE cluster is deployed in FY06. This initial capability provides a quarter of the planned visualization and data analysis capability, a quarter of the planned I/O capability, and half the planned tape-storage capability. After FY05, Black RoSE should provide the capability to visualize output from a 100-million cell simulation at interactive rates.

7.2.3. Data Exploration and Feature Extraction

Data exploration and feature extraction capabilities are not available in FY05.

7.3. Hand Off the Results to Another Code

After FY05, a *Give* capability will be deployed on Red Storm.

7.4. Archive the Results of the Simulation

See Section 3 I/O and Data Migration.

8.   Tri-Lab Coordinated Operational Support

Operational support for this machine accommodates the needs of both local and remote users. This requires integrated hotline support, timely dissemination of operational information (e.g., scheduled and unscheduled machine or WAN downtime), training and documentation that meets the needs of local and remote users, as well as the availability of common Tri-Lab web pages with consistent format and information. This goal can only be accomplished by continued integration and coordination efforts among the Labs. This includes, but is not limited to, the Face-to-Face Tri-Lab Coordinated User Support Meetings (three yearly – one hosted by each lab), the Expedited Priority Runs video conference meeting (weekly), and the following weekly teleconferences: (1) Tri-Lab Web Group, (2) Tri-Lab/NWC configuration changes, Tri-Lab HPC support (metrics), NWC Group (helpdesk to helpdesk), and (3) Tri-Lab Status and Monitoring, and Tri-Lab User Training.

8.1.   User Support

As described in Section 1.2.5 Consulting, Sandia Scientific Computing Support provides three mechanisms to access user support for Red Storm: the Web, email, and phone. Phone support will be available after FY05.

8.2.   Trouble shooting (In Depth Consulting)

Tri-Lab agreements exist to contact experts (TBD) at each lab to provide trouble-shooting in the following areas: file transfers, networking, web pages, etc. HPC environment services are tested in an end-to-end approach to provide status and performance information for the users as well as the service providers. Within the Tri-Lab the information is being integrated with local status and monitoring capability. At SNL, the HPC environment status and availability information is being integrated with the local production Cyber Enterprise Management System (CEMS).

Intentionally Left Blank

Appendix: ACE Version 8 Mappings to RSUM Version 1.0

Black indicates no change in requirements from Version 7s to Version 8.0, Green a change in the requirement from Version 7s to Version 8.0, and Pink a new requirement in Version 8.0.

1 Getting Started (learning about the system, gaining access, etc)
1.1 Platform Based Environment

ACE requirements met in the RSUM Version 1.0

- Shall provide machine policy information this will include but not be limited to limitations on interactive use.
- Shall clearly specify platform security policies these shall include but not be limited to export control policies.

ACE requirements NOT met in the RSUM Version 1.0

- Verify the environment by running a set of application based regression tests to exercise changes to the operating system, compilers, libraries, etc.

1.2 Learning about the system

1.2.1 Web-based system information

ACE requirements met in the RSUM Version 1.0

- Shall provide web-based system information.
- The web-based information system shall include open, secure, non-password protected pages that are available to the Tri-Lab.
- The web-based information system shall include links to platform-specific web pages.
- Shall provide current information about the system that is both open and secure.
- Shall provide web pages that are searchable.
- Shall provide a single, well-known point of entry for system information. This point of entry shall provide links to policy information.
- Shall provide information about the main compute platform. It shall also provide information about supporting platforms this shall include but not be limited to information about development, visualization and open computing.
- The ACE system information shall cover utilities for moving data between machines and within a platform group.
- System information shall be kept current and shall include a last modification date and email address for the personnel responsible for the page content.
- Shall provide a user feedback system.
- Shall provide links to application project web pages that provide information about what applications run on what systems. (*This requirement may only apply to secure environments given security policies.)*
- Shall maintain list of FAQs.
- Shall provide access to current system configuration information.
- Shall provide a current set of man pages for each platform to include locally developed software. (This requirement is in section 1.2.2 of the ACE document but information regarding man pages is included in this section of the RSUM).
- Shall provide access to current system configuration information.
- Shall publish a "good neighbor" policy to advertise the policy for submitting jobs to the queues. (From ACE 6.1)
- Shall publish "best practices" for performance. (From ACE 4.9)

- Shall provide a history of machine reliability to assist users in determining frequency of restart dumps. (From ACE 6.5)
- Shall provide users with information about planned system down time. (From ACE 6.5)
- Single location (per platform) (e.g. a web site.) with pointers to commonly known or encountered deviations from industry HPC standards – could be implemented via an updated FAQ as items are identified, via release notes, or a bulletin board (From ACE 4.2)

ACE requirements NOT met in the RSUM Version 1.0

- ASC platform status and configuration information shall be maintained and accessible (both push and pull model) to users. This information   includes the current, historical, and planned (1) operating system and system libraries, (2) compiler and embedded, (3) MPI library, (4) other "standard and customary" libraries, and (5) system status e.g., maintenance down time, dedicated mode, etc. Attributes shall identify the entity including name, location, version, and patches; shall include statement regarding backward compatibility with previous versions; shall include the dates planned for delivery, actually validated, planned for withdrawal, and actually withdrawn. (From ACE 4.x Configuration Control)

1.2.2 On-line system information (on the machine)
   ACE requirements met in the RSUM Version 1.0
   ACE requirements NOT met in the RSUM Version 1.0
1.2.3 Written system documentation
   ACE requirements met in the RSUM Version 1.0
- Shall provide a current set of (easily printable) documentation for each platform, especially compilers, libraries, MPI-IO and MPI that are available locally to users.
- Shall provide a "quickstart" guide for users of new platforms. (From ACE 4.2)
   Could be ASC Red Storm for Dummies.

   ACE requirements NOT met in the RSUM Version 1.0
1.2.4 Training
   ACE requirements met in the RSUM Version 1.0
- Shall provide vendor and locally developed training information.
- Shall provide training materials that are electronically accessible. (Note: Driver's License Training only)
- Shall provide trainer contact information.
   ACE requirements NOT met in the RSUM Version 1.0
- Shall provide periodic tailored training for remote users at their sites. Targeted to specific groups like end-users and developers.

1.2.4 Consulting
   ACE requirements met in the RSUM Version 1.0
   ACE requirements NOT met in the RSUM Version 1.0

1.3 Gaining access to the system (general)
1.3.1 Account and password management (authorization)
   ACE requirements met in the RSUM Version 1.0
- Shall provide the ability to request accounts on-line. This shall include instructions and criteria for getting accounts.
- Shall provide information about tri-lab access policies and procedures. This shall cover all ASC computing resources with links to platform specific web pages.

- Shall provide site-specific access policies and procedures that specify differences from tri-lab standards. The information system shall also provide links to tri-lab and platform specific web pages.
- Shall provide guidance to users concerning which platform they should consider applying for an account based on the machine maturity and how their intended usage.
- Shall provide clear guidance on the authorization approval criteria.
- Shall provide a tri-lab service to process account requests, with user transparent links to site-local authorization mechanisms.

ACE requirements NOT met in the RSUM Version 1.0

1.3.2 Gaining access to the machine (authentication)

ACE requirements met in the RSUM Version 1.0

- Shall minimize the number of authentications users are required to make while meeting security requirements.
- Shall support cross-laboratory honoring of credentials.

ACE requirements NOT met in the RSUM Version 1.0

- Automated authentication support for batch jobs (e.g. cron), for nightly builds, regression testing, etc. (including documented solutions)
- Provide the ability for multiple users to have access control of a common running application.

1.3.3 System availability/scheduling information

ACE requirements met in the RSUM Version 1.0

- Shall provide an estimated uptime when the system is unavailable.
- Shall provide the capability for users to determine the size of a job that can be run.
- Shall provide users with current maintenance and upgrade schedules.
- Shall provide accurate information about resources available to the user, to include disk and node resources (memory, CPU, etc.). It is expected that the system will have removed any transient resources allocated the previous user.
- Shall limit access to secure job status information.
- Shall be able to determine the state of the current queue and job limits.
- Every user should be able to get an interactive login (at least one) (with modules) for compiles or whatever, but there may have to be a limit on the number they can have.
- Shall offer a single web site to provide information about platform status and availability to the tri-lab environment. This site shall also provide recommendations about which platforms to avoid due to other commitments.
- Provide resources configured to facilitate code development (system and application) *(From ACE 4.1)*

ACE requirements NOT met in the RSUM Version 1.0

- Shall support use of lightweight scripts access to real-time availability information. *(e.g. Perl, CSH)*
- A utility that returns the set of queues (authorized for user) that meet resource requirements (nodes, memory, etc.)
- Shall provide real-time information via the web about system availability and system loading.

1.3.4 Requests for Priority Jobs

ACE requirements met in the RSUM Version 1.0

ACE requirements NOT met in the RSUM Version 1.0

2 Setting up the work environment

2.1 File system standards and documentation

ACE requirements met in the RSUM Version 1.0

- Shall use standardized path names for the location of applications and project areas. *(e.g. /projects/<project-name>/) (Note: To the extent standards are documented)*
- Shall use configurations for home directories, scratch areas and archival storage that are available and consistent across platforms.
- Shall provide current documentation on configurations of home directories, scratch space and parallel file systems.
- Shall provide appropriate information for all file systems. This shall include but not be limited to back-up policy, quota and purge policy.
- File names should implement similar policies. (e.g. scratch, netscratch, vizscratch.)

ACE requirements NOT met in the RSUM Version 1.0

- Shall provide tri-lab access to common disk space to enable a remote build capability.

2.2 Setting up User Groups Environment Variables, Modules, etc.

ACE requirements met in the RSUM Version 1.0

- Shall provide a documented and maintained skeleton template cshrc file and others that provides the basics for setting up the user environment.
- Shall have a standard module capability on all platforms to manage software dependencies.
- Shall standardize the system environment variable across platforms (to the degree possible) to enable portable user scripts.
- Shall provide paths to common software that are standardized across platforms. (Note: To the degree that paths are standardized and documented)
- Shall provide documented processes that allow tracking additions and deletions to groups. (From ACE 2.3)
  - Shall provide a service for group owners to set up and manage groups. (From ACE 2.3) (Note: metagroup utility available only to Sandia employees. Tri-Lab must request assistance for group creation and management.)
  - Include a standard architecture-specific environment initialization capability. *(e.g. cshrc.aix, .cshrc.tru64, etc.)*
  - Provide other initialization scripts (termcap, emacs, vim)
  - Shall provide the capability to set up groups for export control purposes. *(From ACE 2.3)*

ACE requirements NOT met in the RSUM Version 1.0

- Shall provide home directories that are portable across multiple architectures
- Shall provide a hierarchical group capability. (Used to expand the number of groups that a user can belong to.) *(From ACE 2.3)*
- Shall have a single gid space across the complex to enable imported tar files to map to the correct group. *(From ACE 2.3)*
- Shall provide a user level mechanism to change the default group. *(From ACE 2.3)*
- Shall allow users to change the login shell without requiring sys admin intervention. (Note: with caveats.)

3 I/O and Data migration

3.1 Red Storm-ASC Network Architecture Map

ACE requirements met in the RSUM Version 1.0
- Shall provide the infrastructure to support high speed remote bulk data transfer that can be executed unattended and is resilient.
- At least one file system shall be provided backup for source code.

ACE requirements NOT met in the RSUM Version 1.0
- Shall provide local files systems that are accessible over the WAN.
- Shall support all combinations of high-speed (parallel) data transfer from local parallel file system/archive to remote parallel file system/archive on both the classified and unclassified networks. This shall include: local to local, local to remote, remote to local, and remote to remote.

3.2 Red Storm RoSE Network Architecture

ACE requirements met in the RSUM Version 1.0
ACE requirements NOT met in the RSUM Version 1.0
- Shall support all combinations of high-speed (parallel) data transfer from local parallel file system/archive to remote parallel file system/archive on both the classified and unclassified networks. This shall include: local to local, local to remote, remote to local, and remote to remote. (*From ACE 3.1*)

3.3 Tools For Transferring Files

ACE requirements met in the RSUM Version 1.0
- Shall provide tools for high performance and scalable data transfer. (*currently pftp) (From ACE 3.1)*
- Shall provide system environment components that facilitate development of common procedures for setting up similar types of problems on the platforms. *(e.g. scripts, common file movement facilities, etc.)* (*From ACE 3.2*)
- The archive implementation shall provide high reliability. Provide disaster recovery mechanisms for user identified files to attain even higher reliability. (From ACE 7.4)
- The archive implementation shall provide high availability. Provide failover for archive write. (From ACE 7.4)
- The archive implementation shall provide data security. To include access control lists, portable across HPSS platforms/sites implementations. (From ACE 7.4)
- The archive implementation shall provide the capacity sufficient to track user needs. (From ACE 7.4)
- The archive implementations shall provide common functionalities. (From ACE 7.4)
- Shall provide a common (across systems and tri-lab) user interface to the archive. (From ACE 7.4)
- Shall provide utilities to consolidate many small files into a single file. *(currently htar).* (From ACE 7.4) (Note: tar, but not htar, is available.)

ACE requirements NOT met in the RSUM Version 1.0
- The archive performance (bandwidth delivered to a single file archive) shall track user needs for archival storage as determined by problem size and expectations for time to archive. *(e.g. a constant time to archive that tracks problem size).* (From ACE 7.4)

- Shall provide a persistent archiving capability with seamless transition across archival systems. (From ACE 7.4)
- Provide archive data management and organizational services (to include audit capabilities), consistent with security policies (e.g. GREP, FIND, LS, Head, Tail, Bulk file management) (From ACE 7.4)

3.3.1 Archival Storage Policies

ACE requirements met in the RSUM Version 1.0

ACE requirements NOT met in the RSUM Version 1.0

3.4 Addressing I/O: Effective Use of the Files systems, serial and parallel I/O

ACE requirements met in the RSUM Version 1.0

- Shall provide a location where libraries and tools associated with an application can be maintained. *(From ACE 3.2)*
- Shall provide guidance for code developers to make effective use of serial and parallel file systems, including the main file systems primarily to include but not limited to large parallel file system-topology. (From ACE 4.4)
- Shall provide links to information about the system parallel file system. (From ACE 4.4)
- Shall allow for the number of concurrent open files (including system files) equal to approximately ten times the number of process (From ACE 4.4)

ACE requirements NOT met in the RSUM Version 1.0

- Shall provide platform specific single global parallel file system visible from all nodes with homogeneous node IO performance to the file system. (From ACE 4.4)
- Shall provide a single parallel file system that spans platforms. (From ACE 4.4)
- Shall provide optimized and standard compliant parallel I/O libraries: MPI IO, HDF5, PNetCDF, and UDM (From ACE 4.4) (Note: MPI IO and HDF5 will be supported after FY05)
- Shall provide IO libraries that are thread safe with scalable performance. (From ACE 4.4)
- Need an API to determine whether a file system is local or global, and serial vs. parallel (From ACE 4.4)

3.4.1 Parallel I/O Libraries

ACE requirements met in the RSUM Version 1.0

ACE requirements NOT met in the RSUM Version 1.0

4 Application and System Code Development

ACE requirements met in the RSUM Version 1.0

- Shall provide access to source repository (e.g. SourceForge, Razor, CVS, etc.) (From ACE 4.x Configuration Control) (Note: Red Storm will reside in the SRN and SCN environments (physical separation via a physical red/black switch). If the source repository is accessible from those environments, it should be accessible from Red Storm.)

ACE requirements NOT met in the RSUM Version 1.0

- Shall provide the capability for developers and support personnel to maintain products without requiring sys admin privileges. (From ACE 4.5)

4.1 Peculiarities of the System
   ACE requirements met in the RSUM Version 1.0
   ACE requirements NOT met in the RSUM Version 1.0. (From ACE 4.2)
- Shall document deviations from POSIX and ANSI/ISO standards. (From ACE 4.2)
- Shall document deviations from compliance with language and run-time standards. *(e.g. MPI, I/O, Cray pointers etc.).* (From ACE 4.2)
- Shall publish deviations from IEEE arithmetic. (From ACE 4.2)

4.2 Parallel Programming Models and Run-Time Systems
   ACE requirements met in the RSUM Version 1.0
- Shall provide the ASC standard compliant MPI library (currently 1.2) optimized for both on-box and across-box messaging. (From ACE 4.3)
- Shall make it clear to users which MPI library is loaded, especially when wrapped compiler commands are used. (From ACE 4.3)
- Shall provide resources for the installation and support of locally developed libraries that can be centrally accessed . (e.g. LA-MPI, UPS, HYPRE). (From ACE 4.3)
- Shall provide documentation with guildelines for running OpenMP and MPI programs (especially for threadsafe code.) *(e.g. how to tune buffer sizes, etc.)* (From ACE 4.3) (Note: There is no thread support on the compute nodes. There is no support for OpenMP. The list of MPI environment variables will be published, to enable MPI tuning.)
- Need a mechanism to force a specific version of MPI to be linked into the code (From ACE 4.3) (Note: The module capability can be used to specify a specific MPI version.)
- Need a standard interrupt propagation mechanism to tasks associated with a parallel application. (From ACE 4.3) (Note: Signals can be sent to yod. SIGTERM, SIGUSR1, and SIGUSR2 are catchable by the application. All others, including SIGINT (ctrl-C), will automatically kill the application.)
- Consistent integration with the batch system, and access to resource availability in order to build the MPI run command. (From ACE 4.3) (Note: There is integration between the batch system and the resource availability information. xtshowmesh and the qstat commands can be used to determine current and queued resource requests. PBS Pro does not support negotiation of resource requests and allocations among the job, the scheduler, and the resource manager.)
   ACE requirements NOT met in the RSUM Version 1.0
- Shall provide POSIX compliant thread package. (From ACE 4.3) (Note: Catamount on the compute nodes does not support threads; Linux, on the service and I/O nodes, supports threads.)
- Shall provide support for hybrid OpenMP/MPI parallel programs. (From ACE 4.3)
- Provide mechanisms in MPI for enhanced reliability – e.g. survive NIC failures (From ACE 4.3)

4.3 Third Party Libraries and Utilities
ACE requirements met in the RSUM Version 1.0
- Shall manage changes to center supported third party libraries via a change control procedure. (From ACE 4.5)
- Shall provide information about different versions of libraries that are on the system and guidance to the user as to which one to use including but not limited to MPI. (From ACE 4.5)
- Shall provide system documentation that includes a POC for each center supported third party product on the system. (From ACE 4.5)
ACE requirements NOT met in the RSUM Version 1.0

4.3.1 Math Libraries
ACE requirements met in the RSUM Version 1.0
- Shall provide support for standard sparse matrix operations. (From ACE 4.5.1)
- Shall provide support for parallel sparse linear solvers. (From ACE 4.5.1)
- Shall provide optimized BLAS and LAPACK. (From ACE 4.5.1)
- Shall provide system documentation that point to what solvers are installed and the support POC. (From ACE 4.5.1)
ACE requirements NOT met in the RSUM Version 1.0

4.3.2 Networking and other Libraries
ACE requirements met in the RSUM Version 1.0
- Shall support libraries that are callable from C++. (Note: If important please supply which libraries.) (From 4.5.2)
- Shall support secure interprocess communications such as sockets, HTTPS, and/or ssh. (From 4.5.2) (Note: Available only on the Linux service and I/O nodes, not on the compute nodes.)
- Shall support the deployment of client/server socket-based parallel tools (including visualization) on platforms. (From 4.5.2) (Note: Support for client/server socket-based parallel tools, not including visualization, is available only on the Red Storm Linux service and I/O nodes, not on the compute nodes. Graphics hardware for visualization is available only on the support clusters, Red RoSE and Feynman.)
ACE requirements NOT met in the RSUM Version 1.0

4.4 Compilation
ACE requirements met in the RSUM Version 1.0
- The ACE system information shall provide current documentation for compilers, libraries and development tools. (From ACE 4.2)
- Shall provide user recommendations for recommended compiler options. (From ACE 4.6)
- Shall provide information about different compiler versions resident on the system and guidance to the user as to which one to use. (From ACE 4.6)
- Shall provide compiler support for F77 codes including support for Cray pointers. (From ACE 4.6)
- Shall provide common scripting languages. (e.g. perl, python, and Tcl/Tk, sh, csh,tcsh,ksh, expect) (From ACE 4.6)
- Shall provide JAVA support. (editorial note – this is an attempt to clarify a previous requirement): JAVA – via standard JDK toolkit (From ACE 4.6)

- Shall provide the latest 64-bit versions of the gnu utilities. (*e.g. make, g++, gawk, gdb, gawk, etc.).* (From ACE 4.6) (Note: Available only for the Linux service and I/O nodes, not the compute nodes Typically, gnu utilities will be updated whenever the Linux OS distribution is updated.)
- Shall provide standard compliant compilers. C, C++, and FORTRAN9x. (From ACE 4.6) (Note: There are several standards and many versions of each of these standards. The PGI compilers do conform to some of these, but the Portland Group is not being very forthcoming with details. PGI does not conform to the de facto standard of the gnu-specific compiler features.)

ACE requirements NOT met in the RSUM Version 1.0
- Shall provide compiler support for OpenMP. C & OpenMP, C++ & OpenMP, and Fortran9X & OpenMP. (From ACE 4.6)
- Shall provide parallel make and compilation in build. (From ACE 4.6)
- Shall provide support for dynamic/shared libraries. (From ACE 4.6) (Note: Not supported on compute nodes; supported on service and I/O nodes.)

4.5 Debugging and Correctness Testing

ACE requirements met in the RSUM Version 1.0
- Shall provide the ASC standard parallel debugger. *(currently TotalView). (From ACE 4.7)*
- Shall provide machine resources to be used for code debugging – (this requires dedicated interactive use of the resource - and make it clear to users which resources are to be used for this purpose. *(From ACE 4.7)*

ACE requirements NOT met in the RSUM Version 1.0
- Shall provide a usable debugging capability for jobs that span across 1/3 of the system. (From ACE 4.7)
- Shall provide tools for static analysis of source. *(e.g. flint). (From ACE 4.7)*
- Shall provide tools that measure testing code coverage for all supported standard languages. *(From ACE 4.7)*
- Shall provide common tools for parallel aware memory checking including user callable routines to track memory use and availability. (e.g. Valgrind type functionality) (From ACE 4.7)

4.6 Performance Measurement, Analysis and Tuning

ACE requirements met in the RSUM Version 1.0
- Shall provide loop level application profiling tools. *(From ACE 4.8)* (Note: Cray is providing a suite of profiling and performance analysis tools.)
- Shall provide user level access to hardware performance counters, preferably via the PAPI API. (*From ACE 4.8*)
- Shall make it clear which tools can be used together. *(e.g. compiler and memory tool, or debugger and mpi). (From ACE 4.8)*

ACE requirements NOT met in the RSUM Version 1.0
- Shall provide tools to characterize the IO performance of a code. (*From ACE 4.8)*
- Shall provide a message tracing tool that is usable for jobs that span across 1/3 of the system. *(e.g. Vampir). (From ACE 4.8)*
- Shall provide a message profiling interface *(e.g. mpiP). (From ACE 4.8)*
- Shall provide tools to access thread performance. (From ACE 4.8)
- Shall provide tools to analyze cache performance. (*From ACE 4.8)*

5 Problem Setup
ACE requirements met in the RSUM Version 1.0
- Shall provide resources and connectivity to users for the generation of meshes/netlists/schematics, material databases, and necessary data storage for application data bases. (From ACE 5.0)

ACE requirements NOT met in the RSUM Version 1.0

5.1 Domain decomposition
ACE requirements met in the RSUM Version 1.0
ACE requirements NOT met in the RSUM Version 1.0
- Shall provide real-time information to support smart domain decomposition to allow near optimal mapping onto the machine, taking into account non-homogeneous system resources. (From ACE 5.2)

6 Running the application to solve the problem
6.1 Submitting the job (local and remote)
ACE requirements met in the RSUM Version 1.0
- Shall provide users with a common queue syntax and semantics across the tri-lab. (i.e., "large queue" means the same across the tri-lab) To the extent that a common queue syntax and semantics for the tri-lab exist and are documented.
- Shall provide a common user interface to access different underlying resource management system resource information. (Note to achieve tri-lab common interface – may require change to your scripts.) To the extent that a common user interface for the tri-lab exists and is documented.
- Shall provide the user the ability to specify the approximate requirements of their job. (e.g. the ability to specify the resources that are needed for the job.)
- Shall provide the user with a sorted list of resources available to handle their job. (e.g. How many processors are available, In which queue, Estimate of time to start-up, Estimate of run time) (Note: On Red Storm, nodes are not ordinarily assigned to specific queues, PBS qstat commands can be used to view the state of the queues.)
- Shall make the remote job submittal process the same across the complex. *(Note: To the extent that a common job submittal process exist)*
- Shall provide resource management mechanisms that support accurate and reasonable time limit and node allocation requirements.
- Need quick turnaround for small regression tests. (Via the dedicated partition or by using the 'interactive' mode of PBS.)

ACE requirements NOT met in the RSUM Version 1.0
- Shall support dependent jobs.
- Shall provide automatic job submission e.g. for regression testing. *(e.g. cron)*
- Shall provide job schedulers that place processes so as to optimize use of the platform topology or specialized resources. *(e.g for compilation, I/O, visualization, etc)*

6.2 Monitoring the job
ACE requirements met in the RSUM Version 1.0
- Shall allow users to examine the contents of files being used by an executing job. (with STDIN and STDOUT caveats)

- Shall be able to attach to a running job to control, steer and monitor it. (e.g. run proxy.) (Note: Users will be able to attach to a running job to steer and monitor with TotalView.)

ACE requirements NOT met in the RSUM Version 1.0
- Shall be able to monitor memory and cpu use for load balancing purposes.

6.3 Stopping the job
ACE requirements met in the RSUM Version 1.0
- Shall allow a user to initiate a kill of a job causing an immediate termination, to include an optional flushing of IO buffers - this capability needs to work in the presence of resource managers. *(e.g. the signal propagates through the various layers of daemons to the user processes).*

ACE requirements NOT met in the RSUM Version 1.0
- Shall provide a well-defined signal interface to be sent to all user processes prior to the system going down. The signal shall be sent out with sufficient notice to allow graceful shutdown and cleanup.
- Shall make it possible for the user to inform the resource management system of the appropriate time delay between signal issue and process termination for their application.

6.4 Interactive use
ACE requirements met in the RSUM Version 1.0
- Shall provide some portion of the system to be available to support timely interactive use.
- Shall provide dedicated resource allocation for interactive use. (e.g. for computational steering or debugging)

ACE requirements NOT met in the RSUM Version 1.0

6.5 Adapting the job for expected system reliability
ACE requirements met in the RSUM Version 1.0
- Shall provide tools and systems to provide a reliable environment for users. (e.g. lock out unreliable or nodes with other problems) (From ACE 6.6)

ACE requirements NOT met in the RSUM Version 1.0
- Shall provide the capability to transparently migrate user processes if a node goes down.
- Shall provide procedures for "automatic" job restart after system interruption. (From ACE 6.6)

7. Processing Simulation Output
7.1 Prepare Data for Analysis
ACE requirements met in the RSUM Version 1.0
- Shall provide specialized resources for the manipulation (e.g. combining) of visualization data - may require more memory than compute nodes, etc. (Note: Support clusters have a limited number of nodes with 12 GB memory per node.)

ACE requirements NOT met in the RSUM Version 1.0

7.2 Analyze the Results
7.2.1 Visualization
ACE requirements met in the RSUM Version 1.0
- Shall document what viz tools are available.

- Shall provide a full featured scalable parallel visualization tool. This requirement currently met by Ensight and VIsIt.
- Shall provide visualization from the desktop.

ACE requirements NOT met in the RSUM Version 1.0

- Shall establish policies that are established for utilization of compute platform visualization nodes.
- Shall provide the ability to visualize the data in place without the need to move data to another system.
- Shall provide computing (hardware) resources to be used for data extraction and visualization as part of an interactive visualization process does not require the movement of data.
- Shall provide data analysis capabilities. IDL is the preferred tool due to a link to experimentalists' data.

7.2.2 Data Exploration and Feature Extraction
ACE requirements met in the RSUM Version 1.0
ACE requirements NOT met in the RSUM Version 1.0

7.3 Hand off the results to another code
ACE requirements met in the RSUM Version 1.0
ACE requirements NOT met in the RSUM Version 1.0

- Shall provide a command to transfer a file to another user. Must be a standard interface and a standard location. (e.g. "give" command)

7.4 Archive the Results of the Simulation
ACE requirements met in the RSUM Version 1.0
ACE requirements NOT met in the RSUM Version 1.0

8 Tri-Lab coordinated operational support
8.1 User support
ACE requirements met in the RSUM Version 1.0

- Shall publish and maintain phone numbers, email addresses, and hours of support hotline and help desk.
- Shall provide contact information for reporting problems outside working hours.
- In the ACE, the local user support team shall act as a clearinghouse for local user support related to both local and remote platform usage.

ACE requirements NOT met in the RSUM Version 1.0

- Shall have the capability for open and secure problem reporting and tracking for each system. CLIK is not scheduled to be available on the secure side.
- Shall provide adequate local support staff that is available during tri-lab working hours to answer user questions either in person or via the telephone (single access number/site.)

8.2 Trouble shooting (in depth consulting)
ACE requirements met in the RSUM Version 1.0

- Shall provide experts to be available to resolve complex system problems in areas such as IO, compilers, visualization, parallel programming, performance, debugging, platform specific issues, etc.
- Shall provide enhanced system support for early machine problems and milepost calculations.

ACE requirements NOT met in the RSUM Version 1.0
- Shall have the capability for open and secure problem reporting and tracking for each system. Clik will not be available on the Secure side.

Distribution:

1		Los Alamos National Laboratory
		Attn: Randal Rheinheimer, B294
		PO Box 1663
		Los Alamos, NM 87545
3		Lawrence Livermore National Laboratory
		Attn: Brian Carnes, L-067
		7000 East Avenue
		Livermore, CA 94550

1	139	Paul Yarrington, 1902
1	372	Kristin Dion, 1526
1	807	William Patrick Collins, 4328
1	0139	Arthur L Hale, 01900
1	0139	Robert K Thomas, 01904
1	0316	Sudip S Dosanjh, 01420
1	0382	Greg Sjaardema, 01543
1	0431	Ann Hodges, 00232
1	0470	Patricia M Sanchez, 01905
1	0799	Martha J Ernest, 04333
1	0806	John A Larson, 04330
1	0806	John H Naegle, 04336
1	0806	Lawrence F Tolendino, 04334
1	0806	Leonard Stans, 04336
1	0806	Tan Chang Hu, 04338
1	0806	Thomas J Pratt, 04338
1	0807	Barbara J Jennings, 04328
1	0807	John P Noe, 04328
5	0807	Judy E Sturtevant, 04328
1	0807	Michael E Davis, 04328
1	0807	Robert A Ballance, 04328
1	0807	Susan McRee, 04328
1	0807	Timothy J Preston, 04328
1	0807	William H Rahe, 04328
1	0813	Glenn D Machin, 04312
1	0817	James Ang, 01422
1	0817	Suzanne M Kelly, 01422
1	0822	Brian Wylie, 01424
1	0822	Constantine Pavlakos, 04326
1	0822	David White, 01424
1	0822	Lisa G Ice, 04326
1	0822	Rena A Haynes, 01424
1	0823	John D Zepper, 04320
1	0847	Milton Clauser, 04326

1	1109	James L Tomkins, 01420
1	1110	James Willenbring, 01414
1	1110	Lee H Ward, 01423
1	1110	Michael A Heroux, 01414
1	1110	Ruth Ann Klundt, 01423
1	1138	Charles Michael Williamson, 06225
1	1138	Harvey Ogden, 06225
1	1138	Wilbur R Johnson, 06225
1	9151	Chuck Oien, 08940
1	9151	James L Handrock, 08900
1	9152	James Brandt, 08963
1	9152	Jeffrey N Jortner, 08963
1	9152	Jerrold A Friesen, 08963
5	9152	Karen L Jefferson, 08964
1	9158	Curtis Janssen, 0891
1	9158	Scott Cranford, 08961
1	9159	Michael F Hardwick, 08960
2	9018	Central Technical Files, 8945-1
2	0899	Technical Library, 4536

Intentionally Left Blank