Information Release
REVIEW & APPROVAL FORM

Originating organization: Please complete Sections 1 - 6. Print or type all information. See attached instruction sheets for additional information.

This form is used to review and approve information releases before they are released outside of Sandia.

Most public releases must go through the Formal R&A Process, in which case this form must be completed through Section 10. For releases going through the Organizational R&A Process, organizational management is encouraged to complete this form through at least Section 6 and to file it for future reference. For information on which R&A process to use, or additional R&A information, see:

2. Contacts: Linda Cusimano (NM), (505) 844-4980 or Kelly McClelland (CA), (925) 294-2311
   - This form can be used (through Section 6) to document organizational approval of information.
   - This form can also be used (through Section 6 & Section 8) to document approvals when an information release changes distribution limitations (e.g., when Internal Distribution Only becomes Unlimited Release).

SECTION 1. Protecting Sandia and Partnership Interests.

Is this release the result of
- [ ] a CRADA?
- [ ] Work for Others?
- [ ] Other partnership, MOU agreement, funding source, or understanding OF ANY KIND?
- [ ] Information controlled by other agencies

☐ If NO: Go to Section 2.
☐ If YES: Agreement Number is

Has your partner, non-SNL information owner, or funding agency given approval for this release?  [ ] Yes  [ ] No

SECTION 2. Document Title and Author Information:

Full title of document: Red Storm Update HPC User Forum

Author or contact (Sandian): Erik P. DeBenedictis

Phone No. 284-4017  E-Mail Address: epdeben@sandia.gov

Org. No. 9223 Mail Stop No. 1110

Project number 27355 Task number 03.02.04.01

☐ Contract Author to Sandia. (Contractor's name and contact no.)

SECTION 3. Document Format and Release Event Information. Indicate the planned format(s) of the information release, as well as information about the release event.

Document Format(s):
- [ ] SAND Report
- [ ] Abstract
- [ ] Conference Paper
- [ ] Journal Article
- [ ] Sandia Open Network (External)
- [ ] Computer Software
- [x] Publication (all other types of publications including reports, graphs, posters, exhibits, displays, videos, brochures, internal memoranda, newsletters, factsheets

Release Event: Indicate the name of the conference, meeting, or publication, the sponsoring organization, and the place and date of event. If this release is an electronic posting, provide the current viewing address and intended posting location.

Name of Conference / Journal / Book: HPC User Forum  Sponsor: IDC (International Data Corp.)

Place of Event: Sundance, Utah  Date: 4/9/03 thru 4/9/03

Internet Address of Electronic Posting: N/A

SECTION 4. Classification and Sensitivity of Information. Contact Classification Dept. 3132 (8511 - CA) for questions.

Indicate classification level and category of information release or whether information release is unclassified:

Classification of:  Document Title: UNC  Document Abstract: N/A  The Document: UNC

☐ Classified - Limited Release. Indicate additional access restrictions:
- [ ] NWD Sigma
- [ ] CNVDI
- [ ] NOFORN
- [ ] Specified Dissemination
- [ ] Other

☐ Unclassified - Limited Release. Indicate all Unclassified Controlled Information (UCI) categories access restrictions:
- [x] Export Controlled Information (ECI) ITAR/EAR/_______
- [ ] Internal Distribution Only (IDO) or Internal Use Only (IUO)
- [ ] Non-Sandia Proprietary Information
- [ ] Official Use Only (OUO) Exemption No. ________
- [ ] Patent Caution

☐ Unclassified - Unlimited Release. Information is unclassified with no access restrictions, i.e., distribution may be made worldwide.

☐ DUSA Exemption - This information is released under DUSA Exemption ________ (Mark appropriate sensitivity above. Section 8 review not required.)

Derivative Classifier (DC) who is knowledgeable of information sensitivity: ____________________________  ____________________________  9230 4/1/03

Name  Signature  Org.  Date

(Continued on page 2)
SECTION 5. Disclosure of Technical Advance
A Technical Advance is an original achievement or nonobvious progress in a scientific or engineering sense, including the creation of software. It may be protected by patent, copyright, or as Sandia Commercially Valuable Information. The Originators of a Technical Advance may be inventors or authors.

Does the subject of this Information Release represent a Technical Advance as defined above?
☐ Yes  ☒ No  If No, go to Section 6.

If Yes, has a Disclosure of Technical Advance (TA), Form SF 1155-G, been filed with the Sandia Patent and Licensing Center?
☐ Yes  ☒ No  If No, please follow up with a TA form obtainable from:
(1) Patent & Licensing Center; paper or PC or Mac diskette ((505) 845-8958 or e-mail: patents@sandia.edu); in California, paper or Mac diskette ((925) 294-3690),
(2) Sandia's Internal Web <http://www.patents.sandia.gov/patents>, or
(3) Sandia Line ((505) 845-6789, Quick Dial Code 1057).

SECTION 6. Line/Program Signatures and Approvals. Print or type all author information; obtain appropriate signatures from next-level manager. Where concurrence is obtained in case of multiple authors, approval need only go through the principal author's line organization.

Authors' Names (Print or type)
(Full First, Middle, & Last Name)

Erik P. DeBenedictis

Org. No./Mail Stop
9223/1110

Phone No.
284-4017

Next Level Manager's Signature

Date
4/10/2003

Program Manager's Name and Signature

THE FOLLOWING SECTIONS ARE NORMALLY TO BE USED FOR THE FORMAL REVIEW & APPROVAL PROCESS.*

Copyright Interest? ☐ Yes  ☒ No  If Yes, copyright may be asserted, subject to DOE approval.
Patient Interest? ☐ Yes  ☒ No  If Yes, TA form has been or should be submitted.
Patient Caution? ☐ Yes  ☒ No  If Yes, TA form has been or should be submitted and dissemination will be limited.

Patent Attorney’s/Agent’s Signature

Date

SECTION 8. Classification and Sensitive Information Review (NM: 3132/MS 0175, CA 8511, MS 9021).

Signature

Date

SECTION 9. Review of Promotional Communications/Review of Products Using Color Printing (NM: 9612/MS 0612, CA: 8815/MS 9021). Promotional Communications must be reviewed for adherence to Corporate “Common Look and Feel” guidelines by PR & Communications Center 12600, MS 0619 (NM) or by Communications & Public Affairs Dept., 8528, MS 9131 (CA).
Also, all publications (including SAND Reports and fact sheets) distributed outside of Sandia that use color (ink) printing, as well as all internal products of two or more colors that use color printing, must go through this Section 9 review. NOTE: SAND Reports and internal release products that use color copying do not need a Section 9 review. (Contact Printing & Duplicating Dept. 12630 to get a head start on DOE approval.)

DOE approval received

Date

Approved under the condition that the following statement is printed on the document:

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.
Red Storm Update
HPC User Forum

Erik P. DeBenedictis
Outline

• Project Organization
• Processor
• Network and Network Topology
• Light Weight Kernel (LWK)
• Reliability, Availability and Serviceability (RAS)
Project Organization

- ASCI Red was very successful
- Red Storm RFQ very nearly ASCI Red sped up by Moore’s Law (7x)
- Cray is selling Red Storm to Sandia as a custom product
  - However, Sandia is supplying key expertise for this specific architecture to Cray, and
  - Sandia supplying a major part of the systems software to Cray for integration into Cray’s product
- This organization is working
• Sandia did not specify a processor, but concurs with Cray that the Opteron is a very good choice
• Sandia conducted an evaluation of many available processors
  – Considered overall ability of a processor to integrate into a system
  – Specifically considered FLOPS, memory bandwidth, I/O bandwidth, power consumption
  – Ran benchmarks of top Sandia/ASCI codes
Processor Specifics

• Processors
  – AMD Sledgehammer (Opteron)
  – 2.0 GHz
  – 64 Bit extension to IA32 instruction set
  – 64 KB L1 instruction and data caches on chip
  – 1 MB L2 shared (Data and Instruction) cache on chip
  – Integrated dual DDR memory controllers @ 333 MHz
  – Integrated 3 Hyper Transport Interfaces @ 3.2 GB/s each direction

• Node memory system
  – Page miss latency to local processor memory is <140 ns
  – Peak bandwidth of ~5.3 GB/s for each processor
Network and Network Topology

• Sandia has had very good experiences with the mesh topology
  – ASCI applications tend to be physical in nature. Mapping a 3D problem to a 3D machine preserves locality and maximizes use of fast “nearest neighbor” links.
  – Space-shared batch processing creates a communications locality that matches meshes very well
  – Works well with Red/Black switching
• Meshes look very promising for the future
  – The longest wire in the network determines performance
  – Meshes need no long wires
Red Storm Topology

• Red Storm RFQ specifies a 3D mesh, Sandia and Cray concurred on specific topology
• Compute node topology:
  – 27 x 16 x 24 (x, y, z)
  – Mesh in x & y, torus in z
  – Red/Black split: 2,688 – 4,992 – 2,688
• Service and I/O node topology
  – 2 x 8 x 24 (x, y, z) on each end
  – 192 full bandwidth links to Compute Node Mesh (384 available)
Red Storm Topology

Normally Classified

Switchable Nodes

Normally Unclassified

I/O and Service Nodes

Disconnect Cabinets

I/O and Service Nodes
Advantages and Disadvantages

+ Works well for space-shared batch processing

- An application crossing the narrowest point of the mesh has a “bisection bandwidth” constraint
  - Not sure Sandia has any of these
Interconnect Performance

• Interconnect performance
  – MPI Latency <2 μs (neighbor), <5 μs (full machine)
  – Peak link bandwidth ~3.0 GB/s each direction (sustained 1.8 GB/s each direction)
  – Minimum bi-section bandwidth 1.5 TB/s

• I/O system performance
  – Sustained file system bandwidth of 50 GB/s for each color
  – Sustained external network bandwidth of 25 GB/s for each color
Light Weight Kernel

• Sandia has had very good experiences with LWK
  – Sandia-University of New Mexico Operating System (SUNMOS)
  – Cougar
  – Puma
  – Now Catamount (tell story about name)

• Why?
  – Timing stability
  – Maturity
LWK & Musical Rehearsal

• N musicians Rehearsing 2 Minute Pieces

<table>
<thead>
<tr>
<th></th>
<th>Song 1</th>
<th>Song 2</th>
<th>Song 3</th>
<th>Song 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 min</td>
<td><img src="image1.png" alt="Rehearsal" /></td>
<td><img src="image2.png" alt="Rehearsal" /></td>
<td><img src="image3.png" alt="Rehearsal" /></td>
<td><img src="image4.png" alt="Rehearsal" /></td>
</tr>
<tr>
<td>2 min</td>
<td><img src="image1.png" alt="Rehearsal" /></td>
<td><img src="image2.png" alt="Rehearsal" /></td>
<td><img src="image3.png" alt="Rehearsal" /></td>
<td><img src="image4.png" alt="Rehearsal" /></td>
</tr>
<tr>
<td>4 min</td>
<td><img src="image1.png" alt="Rehearsal" /></td>
<td><img src="image2.png" alt="Rehearsal" /></td>
<td><img src="image3.png" alt="Rehearsal" /></td>
<td><img src="image4.png" alt="Rehearsal" /></td>
</tr>
<tr>
<td>6 min</td>
<td><img src="image1.png" alt="Rehearsal" /></td>
<td><img src="image2.png" alt="Rehearsal" /></td>
<td><img src="image3.png" alt="Rehearsal" /></td>
<td><img src="image4.png" alt="Rehearsal" /></td>
</tr>
</tbody>
</table>
Musical Rehearsal with Breaks

- 2 Minute Pieces with Asynchronous Breaks

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Song 1</td>
<td>Wait</td>
<td>Song 2</td>
<td>Song 3</td>
<td>Wait</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 min</td>
<td>2 min</td>
<td>4 min</td>
<td>6 min</td>
<td>8 min</td>
</tr>
</tbody>
</table>

Note: The diagram shows the schedule for musical rehearsal with breaks, including 'Song 1', 'Wait', 'Song 2', 'Song 3', and 'Wait' periods, each with specific durations.
Breaks in MPP Systems Software

- Unix, Linux, any OS
  - Kernel memory allocation
  - TCP/IP backoff calculations
  - Routing tables
  - Clock synchronization
  - Scheduler
  - Etc., full list unknown, but has been extremely problematic with DOE labs

- Light Weight Kernel
  - [Intentionally blank]
Run Time Impact of Unix Systems Services

• Say breaks take 50 $\mu$s and occur once per second
  – On one CPU, wasted time is 50 $\mu$s every second
    • Negligible .005% impact
  – On 100 CPUs, wasted time is 5 ms every second
    • Negligible .5% impact
  – On 10,000 CPUs, wasted time is 500 ms
    • Significant 50% impact

• Red Storm will be 10,000 CPUs, but will not have asynchronous services
Red Storm Systems Software

• Operating Systems
  – LINUX on service and I/O nodes
  – LWK (Catamount) on compute nodes
  – LINUX on RAS nodes
• Run-Time System
  – Logarithmic loader
  – Node allocator
  – Batch system – PBS
  – Libraries – MPI, I/O, Math
• Parallel File System
  – Several file systems are being evaluated
Reliability, Availability, and Serviceability

• Red Storm RFQ specifies 100 hour MTBI
  – You would take a PC back to Best Buy if it crashed every 4 days
  – However, Red Storm must be able to continue operating while nodes fail and get replaced just to meet this standard

• Red Storm will have a separate RAS network and system of 2500 Unix processors to manage the main machine
  – Will be able to pause running programs, reconfigure hardware, and continue
RAS Network

• RAS Workstations
  – Separate and redundant RAS workstations for Red and Black ends of machine
  – System administration and monitoring interface
  – Error logging and monitoring for major system components including processors, memory, NIC/Router, power supplies, fans, disk controllers, and disks

• RAS Network: Dedicated Ethernet network for connecting RAS nodes to RAS workstations

• RAS Nodes
  – One for each compute board
  – One for each cabinet
Red Storm Performance

Peak of ~ 40 TF

Expected MP-Linpack performance >20 TF

Aggregate system memory bandwidth - ~55 TB/s

Interconnect
  Aggregate sustained interconnect bandwidth > 100 TB/s
  MPI Latency - 2 μs neighbor, 5 μs across machine
  Bi-Section bandwidth ~2.3 TB/s
  Link bandwidth ~3.0 GB/s in each direction

Disk and External Network I/O
  Sustained 50 GB/s each color parallel disk I/O
  Sustained 25 GB/s each color external network I/O
Red Storm Hardware Status

- 24 Boards
- 96 Operton™ Processors
- EMI containment
- Vertical Air Cooling
Red Storm Hardware Status
## Comparison of ASCI Red and Red Storm

<table>
<thead>
<tr>
<th></th>
<th>ASCI Red</th>
<th>Red Storm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full System Operational Time Frame</strong></td>
<td>June 1997 (Processor and Memory Upgrade in 1999)</td>
<td>August 2004</td>
</tr>
<tr>
<td><strong>Theoretical Peak (TF)</strong></td>
<td>3.15</td>
<td>41.47</td>
</tr>
<tr>
<td><strong>MP-Linpack Performance (TF)</strong></td>
<td>2.379</td>
<td>&gt;20 (est)</td>
</tr>
<tr>
<td><strong>Architecture</strong></td>
<td>Distributed Memory MIMD</td>
<td>Distributed Memory MIMD</td>
</tr>
<tr>
<td><strong>Number of Compute Node Processors</strong></td>
<td>9,460</td>
<td>10,368</td>
</tr>
<tr>
<td><strong>Processor</strong></td>
<td>Intel P II @ 333 MHz</td>
<td>AMD Opteron @ 2.0 GHz</td>
</tr>
<tr>
<td><strong>Total Memory</strong></td>
<td>1.2 TB</td>
<td>10.4 TB (up to 80 TB)</td>
</tr>
<tr>
<td><strong>System Memory B/W</strong></td>
<td>2.5 TB/s</td>
<td>55 TB/s</td>
</tr>
<tr>
<td><strong>Disk Storage</strong></td>
<td>12.5 TB</td>
<td>240 TB</td>
</tr>
<tr>
<td><strong>Parallel File System B/W</strong></td>
<td>1.0 GB/s each color</td>
<td>50.0 GB/s each color</td>
</tr>
<tr>
<td><strong>External Network B/W</strong></td>
<td>0.2 GB/s each color</td>
<td>25 GB/s each color</td>
</tr>
<tr>
<td><strong>Interconnect Topology</strong></td>
<td>3-D Mesh (x, y, z) 38 X 32 X 2</td>
<td>3-D Mesh (x, y, z) 27 X 16 X 24</td>
</tr>
</tbody>
</table>
## Comparison of ASCI Red and Red Storm

<table>
<thead>
<tr>
<th></th>
<th>ASCI Red</th>
<th>Red Storm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interconnect Performance</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPI Latency</td>
<td>15 $\mu$s 1 hop, 20 $\mu$s max</td>
<td>2.0 $\mu$s 1 hop, 5 $\mu$s max</td>
</tr>
<tr>
<td>Bi-Directional Link B/W</td>
<td>800 MB/s</td>
<td>6.0 GB/s</td>
</tr>
<tr>
<td>Minimum Bi-section B/W</td>
<td>51.2 GB/s</td>
<td>2.3 TB/s</td>
</tr>
<tr>
<td><strong>Full System RAS</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAS Network</td>
<td>10 Mbit Ethernet</td>
<td>100 Mbit Ethernet</td>
</tr>
<tr>
<td>RAS Processors</td>
<td>1 for each 32 CPUs</td>
<td>1 for each 4 CPUs</td>
</tr>
<tr>
<td><strong>Operating System</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compute Nodes</td>
<td>Cougar</td>
<td>Catamount (Cougar)</td>
</tr>
<tr>
<td>Service and I/O Nodes</td>
<td>TOS (OSF1)</td>
<td>LINUX</td>
</tr>
<tr>
<td>RAS Nodes</td>
<td>VX-Works</td>
<td>LINUX</td>
</tr>
<tr>
<td><strong>Red Black Switching</strong></td>
<td>2260 - 4940 - 2260</td>
<td>2688 - 4992 - 2688</td>
</tr>
<tr>
<td><strong>System Foot Print</strong></td>
<td>~2500 sq ft</td>
<td>~ 3000 sq ft</td>
</tr>
<tr>
<td><strong>Power Requirement</strong></td>
<td>850 KW</td>
<td>1.7 MW</td>
</tr>
</tbody>
</table>