SF 1008-RA (FRONT 3-2003)
Supersedes (11-2002) issue

SAND 2003-1601P

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:

1 The Sandia Web page called "Review and Approval for Communication": http://www-irn.sandia.gov/recordsmgmt/revapprov/revapprov.htm
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 **The Red Storm Computer Architecture and Its Implementation**

Author or contact (Sandian) **Erik P. DeBenedictis** Signature
(Full First Name Full Middle Name Full Last Name)

Phone No. **(505) 284-4017** E-Mail Address **epdeben@sandia.gov** Org. No. **9223** Mail Stop No. **1110**

Project number **3289** Task number **2.01**

☐ Contract Author to Sandia. (Contractor's name and contract no.)
(Identifies funding source – will not be charged)

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 ☒ Publication (all other types of publications including reports, vugraphs, 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: **CCGrid 2003**
Place of Event: **Tokyo, Japan** Date: **5/12/03 thru 5/15/03**
Internet Address of Electronic Posting:

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 **Unclassified** Document Abstract ~~**Unclassified**~~ N/A The Document **Unclassified**

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

☐ Unclassified - Limited Release. Indicate all Unclassified Controlled Information (UCI) categories access restrictions:
☐ 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
☐ Protected CRADA Information (Release date)
☐ Program Designated Special Handling (Distribution Limitation)
☐ Unclassified Controlled Nuclear Information (UCNI)
☐ Other (specify)

☒ 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: Paul Yarrington 7230 5/9/03
Name Signature Org. Date

(Continued on page 2)

2003-1601 9

[...]ssue

[...]5. Disclosure of Technical Advance

[...]al Advance is an original achievement or nonobvious progress in a scientific or engineering sense, including the creation of software. [...]e protected by patent, copyright, or as Sandia Commercially Valuable Information. The Originators of a Technical Advance may be inventors or [...]s.

[...]s 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 SD No. ______ ☐ No If No, please follow up with a TA form obtainable from:

(1) Patent & Licensing Center; paper or PC or Mac diskette ((505) 845-9536 or e-mail: patents@sandia.gov); 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)	Org. No./ Mail Stop	Phone No.	Next Level Manager's Signature	Date
Erik P. DeBenedictis (Full First Name Full Middle Name Full Last Name)	9223/1110	284-4017	Neil Pundit	5/9/2003
(Full First Name Full Middle Name Full Last Name)				
(Full First Name Full Middle Name Full Last Name)				
(Full First Name Full Middle Name Full Last Name)				
(Full First Name Full Middle Name Full Last Name)				
(Full First Name Full Middle Name Full Last Name)				

Program Manager's Name and Signature ______

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

SECTION 7. Patent and Licensing Review (NM: 11500/MS 0161 CA: 11600/MS 9031).

Copyright Interest? ☐ Yes ☒ No If Yes, copyright may be asserted, subject to DOE approval.

Patent Interest? ☐ Yes ☒ No If Yes, TA form has been or should be submitted.

Patent 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 5-9-03

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

Signature Ronald Williams Date 5/9/03

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 Signature ______ Date ______

SECTION 10. Review & Approval Desk (NM: 9612/MS 0612, CA: 8511/MS 9021).
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.

Signature Ramon [illegible] Date 5/9/03

SAND 2003-1601P

The Red Storm Computer Architecture and its Implementation

Dr. Erik P. DeBenedictis
Sandia National Laboratories

CCGrid 2003
Tokyo, Japan

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.

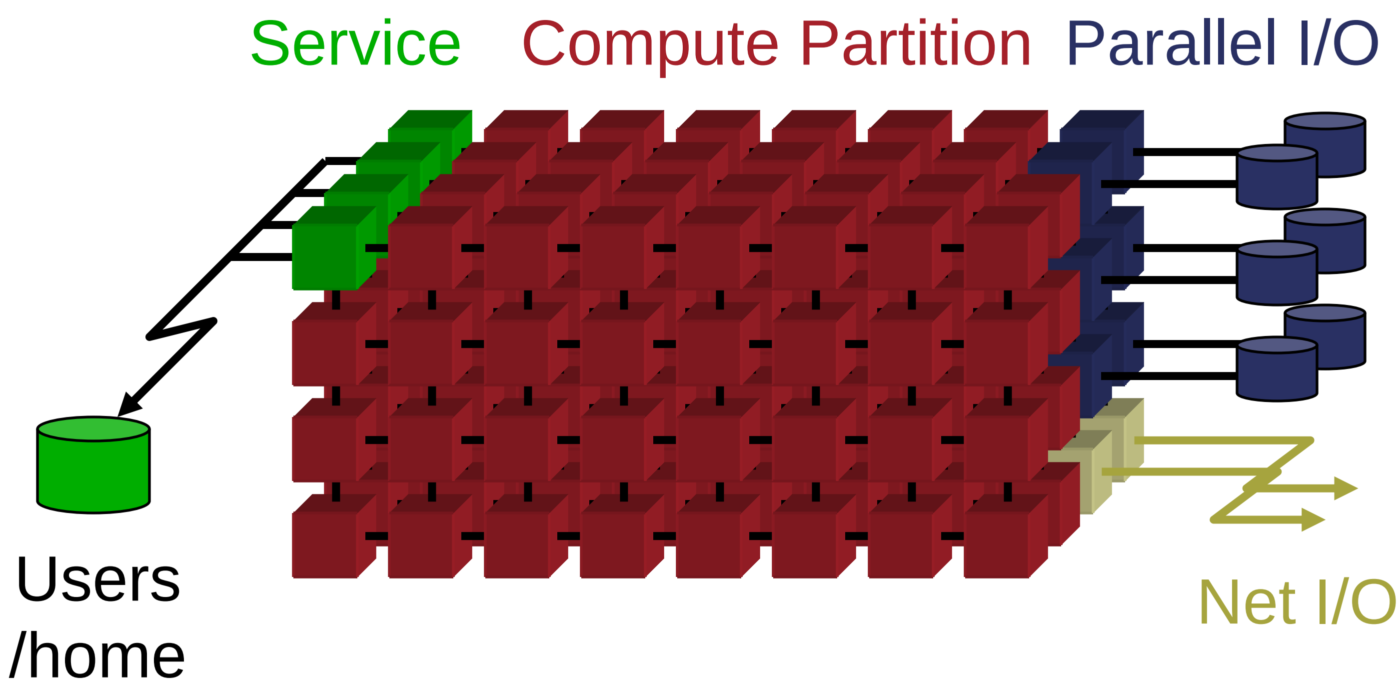
Outline

- Red Storm Overview
- Scalability
- Interconnect
- Reliability
- Light Weight Kernel
- Economy
- Selected Specifications

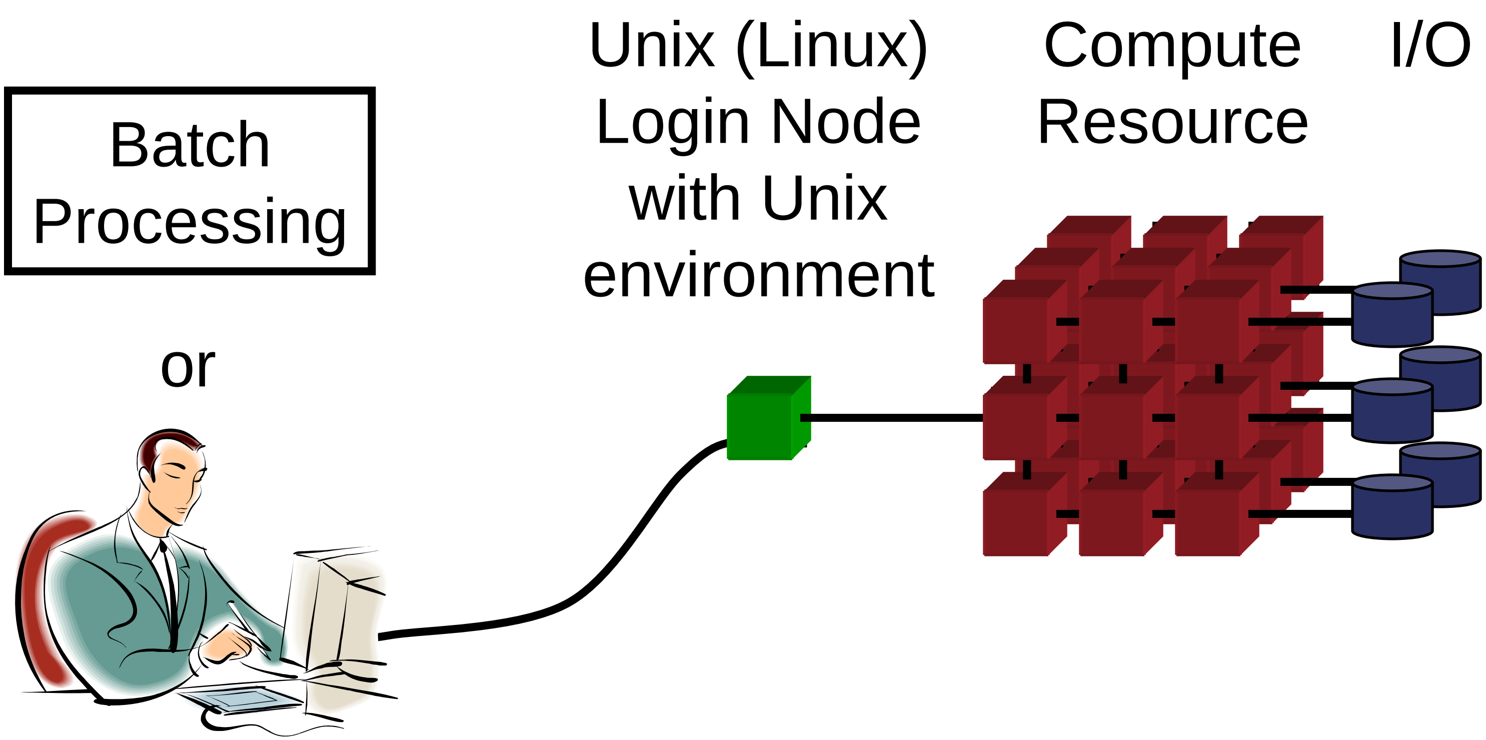
Project Overview

- Red Storm is a nominally 40 TFlops supercomputer that is part of the Advanced Simulation and Computation (ASCI) program
- Red Storm was specified by and is being procured by Sandia National Laboratories
- Red Storm is being manufactured by Cray, Inc.
- Initial delivery to Sandia is scheduled for May, 2004

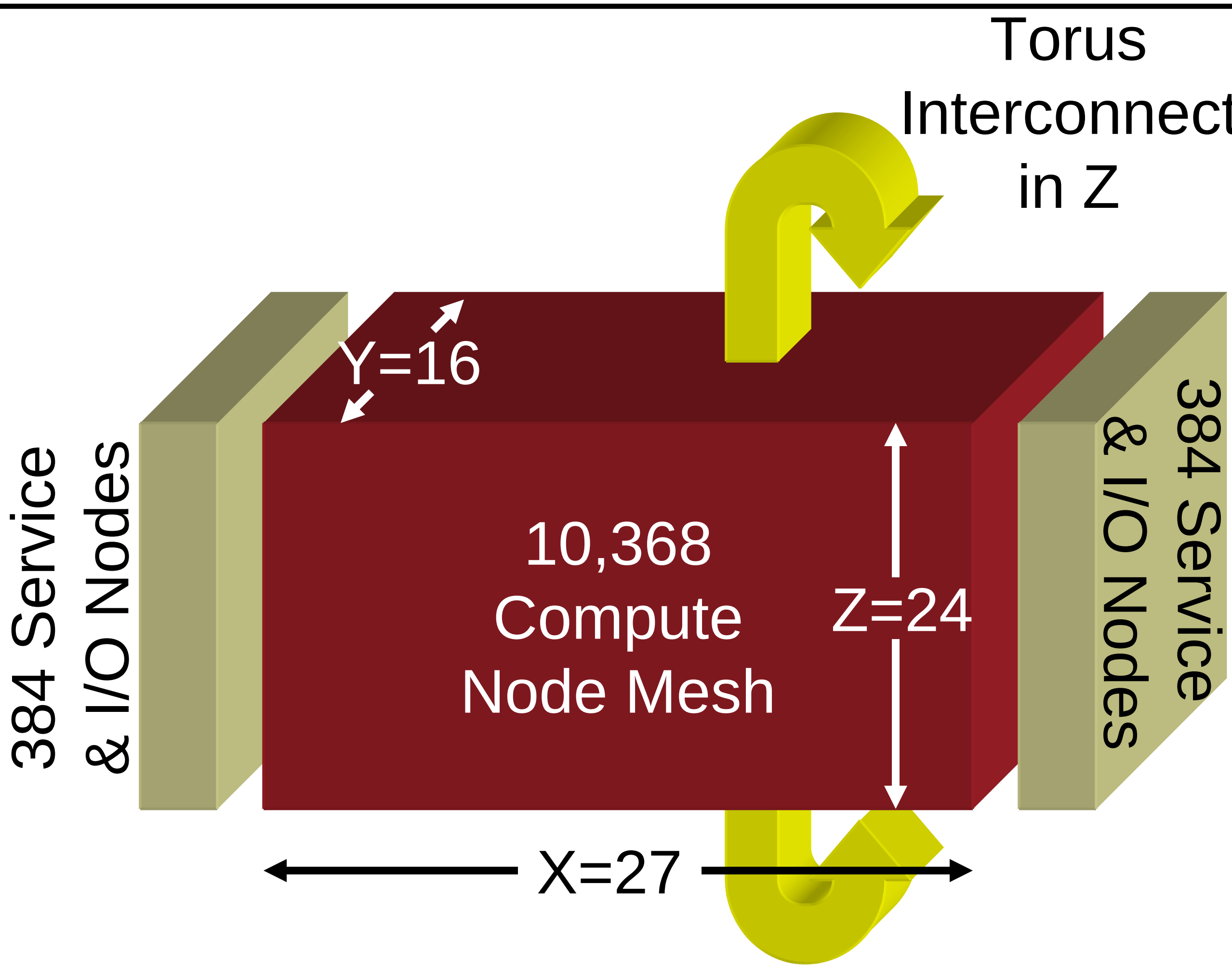
Red Storm is a Massively Parallel Processor



Usage Model

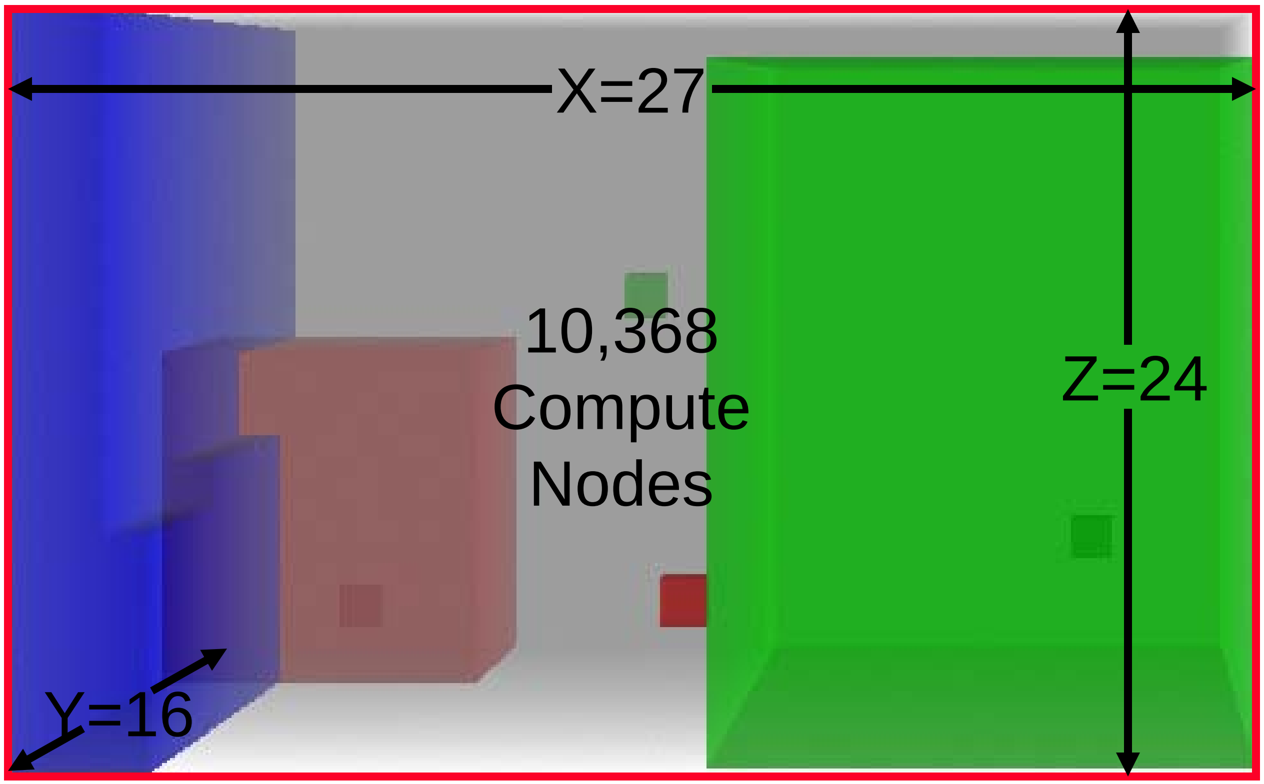

Batch Processing
or
Unix (Linux) Login Node with Unix environment
Compute Resource
I/O

27×16×24 3D Mesh/Torus + I/O

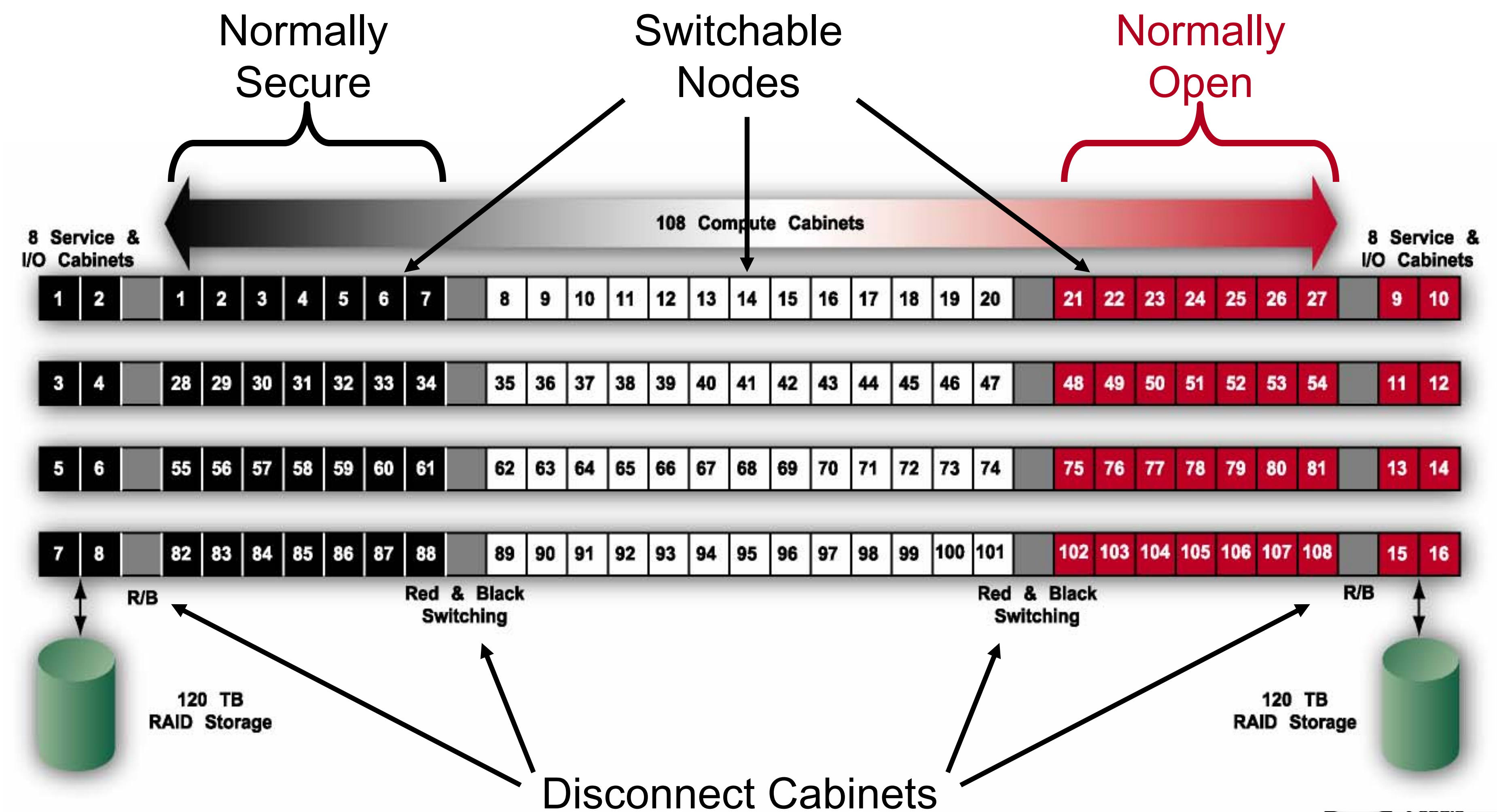


Space Sharing of Jobs

- Jobs occupy disjoint regions simultaneously
- Example – red, green, and blue jobs:

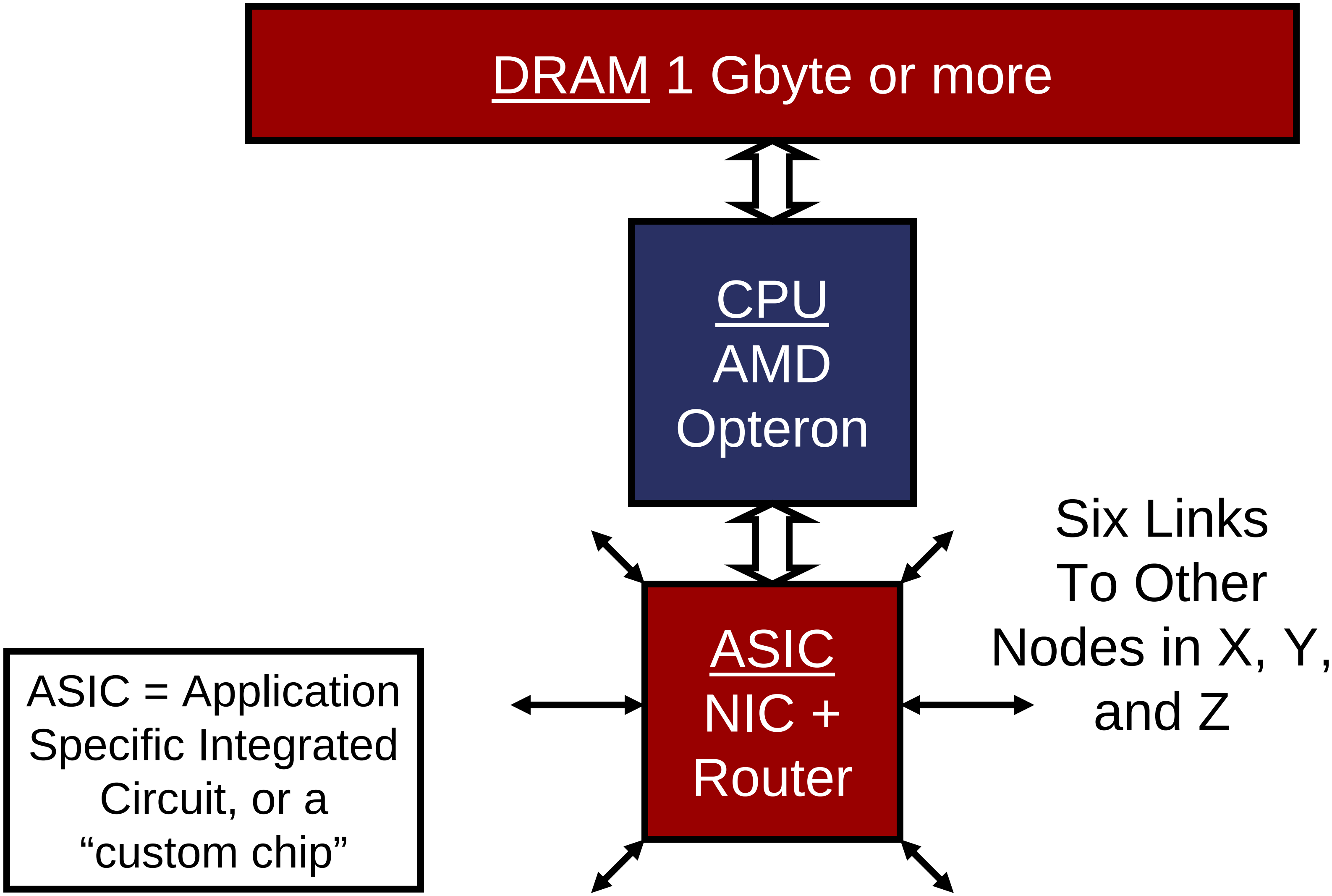


Red Storm Hardware Overview



Sandia National Laboratories

Node Architecture



Software Environment

- Operating System
  - Linux on Login node (where the user logs in)
  - Catamount (Light Weight Kernel) on compute nodes
- Programming Paradigm
  - Message Passing/MPI (no shared memory)
- I/O
  - Initial release: PVFS with Cray enhancements
  - Final release: Lustre

Scalability

- Communications is the key concern
  - Amdahl's Law limits the scalability of parallel computation…
  - but not due to serial work in the application

- Why?

Amdahl's Law

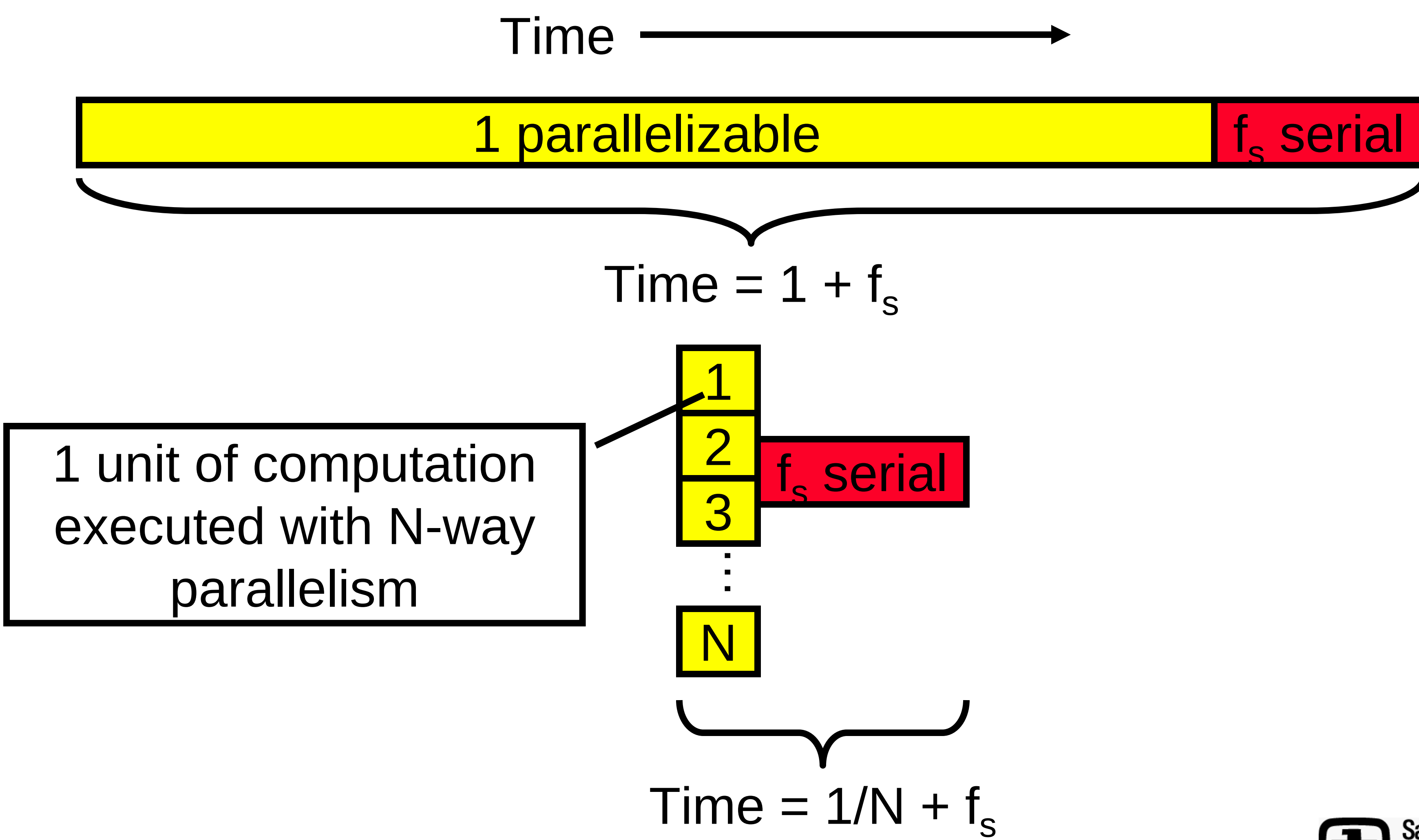
$$S_{Amdahl}(N) = [1 + f_s]/[1/N + f_s]$$

where S is the speedup on N processors and f_s is the serial (non-parallelizable) fraction of the work to be done.

Amdahl says that in the limit of an infinite number of processors, S cannot exceed $[1+ f_s]/f_s$. So, for example if $f_s = 0.01$, S cannot be greater than 101 no matter how many processors are used.

Sandia National Laboratories

Amdahl’s Law Picture



Amdahl's Law

Example:

How big can f_s be if we want to achieve a speedup pf 8,000 on 10,000 processors (80% parallel efficiency)?

Answer:

f_s must be less than 0.000025 !

Amdahl’s Law

Contrary to Amdahl & most folks’ early expectations, well designed codes on balanced systems can routinely do this well or better!

However in applying Amdahl’s Law, we neglected the overhead due to communications.

A Realistic View of Amdahl's Law

The actual scaled speedup is more like

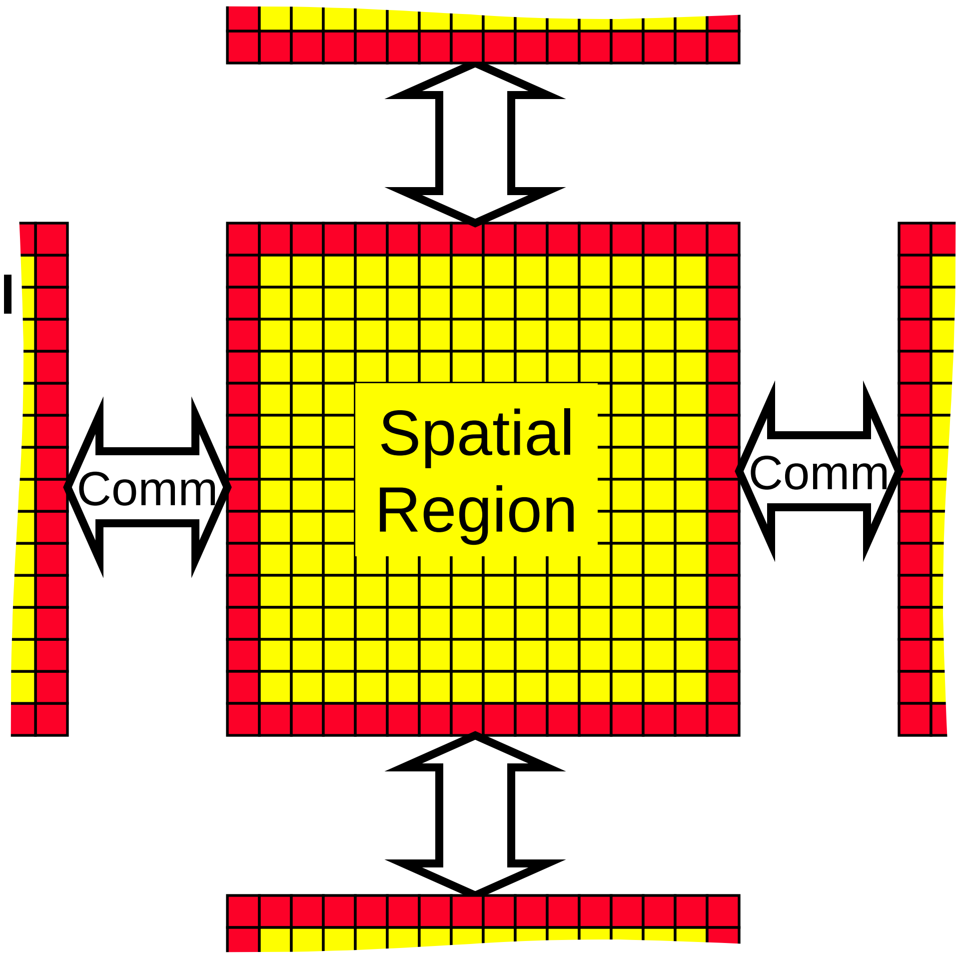
$$S(N) \sim S_{Amdahl}(N)/[1 + f_{comm} \times R_{p/c}],$$

where f_{comm} is the fraction of work devoted to communications and $R_{p/c}$ is the ratio of processor speed to communications speed.

Realistic Picture of Amdahl's Law

- **Problem is a physical simulation in two dimensions**
- **Ratio of boundary (■) to all points (□+■) is f_{comm}**
- **Boundary runs at slower due to communications, say ratio of $R_{p/c}$**
- **Communications will slow execution by factor of**

$$\frac{1}{1 + f_{comm} \times R_{p/c}}$$



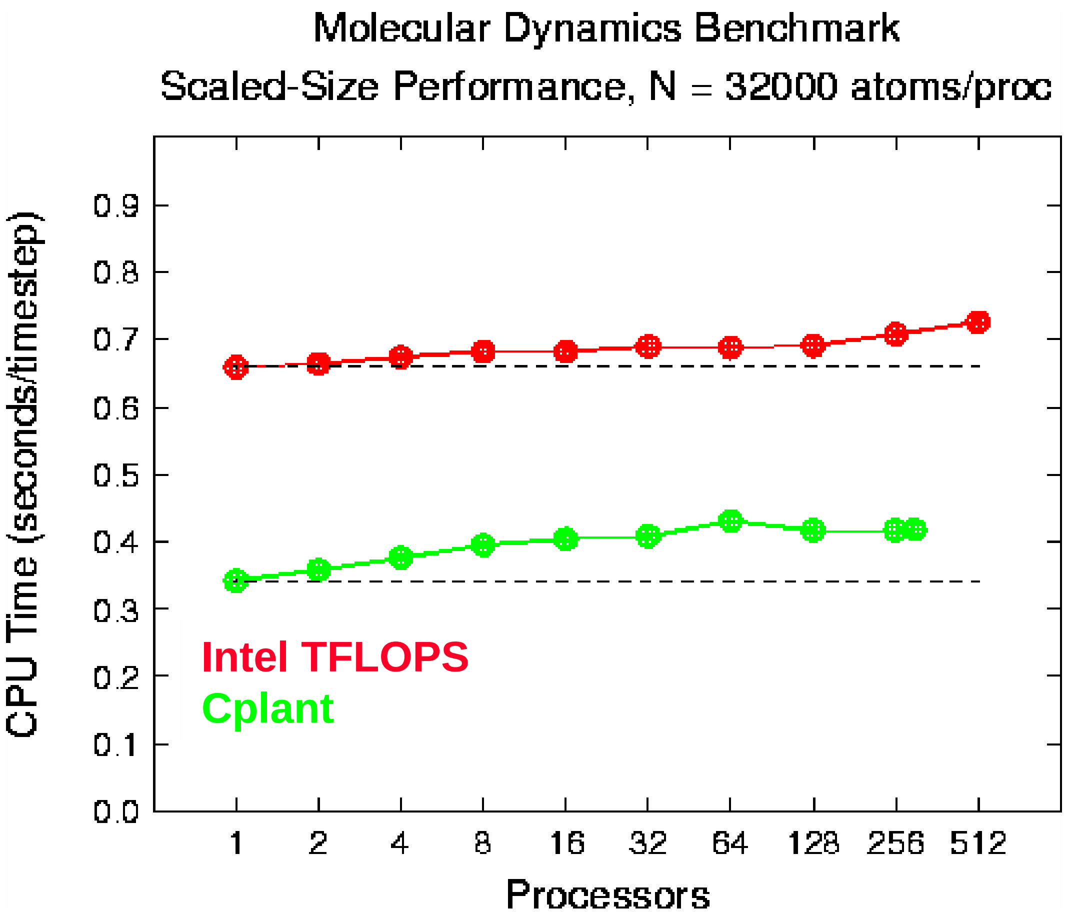
Implications of Realistic Amdahl's Law

- Let's consider three cases on two computers:
  - The two computers are identical except that one has
    - $R_{p/c}$ = 1 Byte/FLOP (fast communications)
    - $R_{p/c}$ = 0.05 Byte/FLOP (not so fast communications)
  - The three cases are
    - f_{comm} = 0.01,
    - f_{comm} = 0.05, and
    - f_{comm} = 0.10

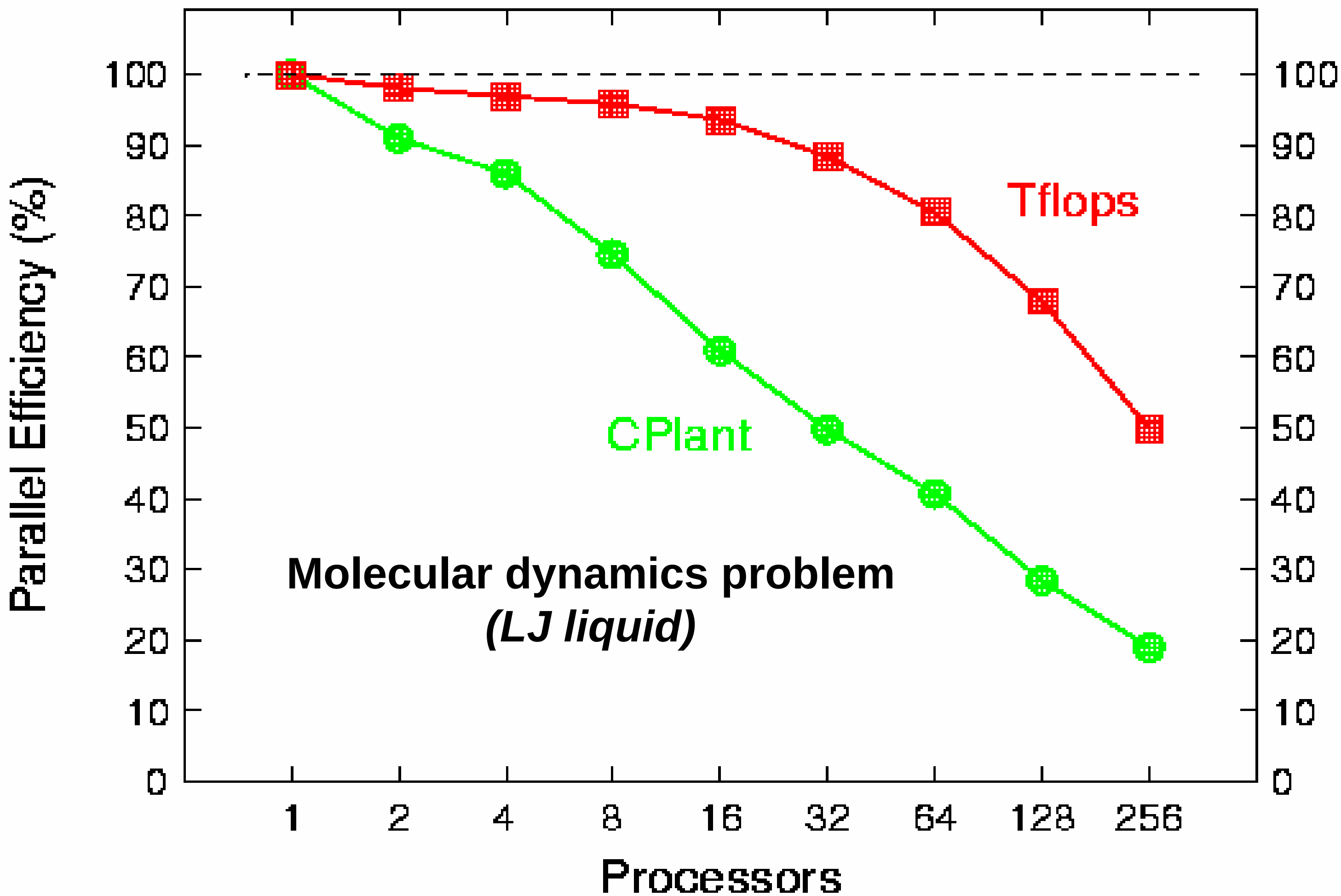
Real Amdahl's Law Efficiency

Efficiency	F_{comm} = .01 99% comp. dominated	F_{comm} = .05 95% comp. dominated	F_{comm} = .1 90% comp. dominated
$R_{p/c}$ = 1 Time to send a number ≈ time for an op on it	99% Efficient	95% Efficient	90% Efficient
$R_{p/c}$ = 0.05 Time to send a number ≈ time for 20 ops on it	83% Efficient	50% Efficient	33% Efficient

Sandia Experience with $R_{p/c}$



Sandia Experience with $R_{p/c}$



Importance of Balanced Communications

- A “well-balanced” architecture is nearly insensitive to communications overhead

- By contrast a system with weak communications can lose over half its power for applications in which communications is important

- <u>Red Storm has been designed with $R_{p/c} \approx 1$</u>

Sandia National Laboratories

Comparisons of Communications Balance

Machine	Node Speed Rating(MFlops)	Link BW (Mbytes/s)	Ratio (Bytes/flop)
ASCI RED	**400**	**800(533)**	**2(1.33)**
T3E	**1200**	**1200**	**1**
ASCI RED**	**666**	**800(533)**	**(1.2)0.67**
Cplant	**1000**	**140**	**0.14**
Blue Mtn*	**500**	**800**	**1.6**
BlueMtn**	**64000**	**1200 (9600*)**	**0.02 (0.16*)**
Blue Pacific	**2650**	**300 (132)**	**0.11 (0.05)**
White	**24000**	**2000**	**0.083**
Q*	**2500**	**650**	**0.2**
Q**	**10000**	**400**	**0.04**

Interconnect

- Connection Choices
  - PCI/PCIX based processor connections– adequate
  - Memory sub-system based connections – much better (e.g. Marvel interconnects and AMD Hypertransport Layer)

- Switch Fabric
  - Commercial networks:
    - Myrinet (cheaper, fairly fast)
    - Quadrics (more costly; currently faster)
    - Gigabit Ethernet (cheap, not a good idea for scaling to 10^4 nodes)
  - Custom interconnects:
    - IBM; ASCI Red; T3E;SGI; Cray, …

Interconnect Tradeoffs

- Commercial networks:
  - Quadrics can get within a factor 2-4 of the latency requirements and within a factor of 4 of the bandwidth targets for Red Storm.
  - Cabling costs may be higher than for custom interconnects.
- Custom interconnects:
  - Easily meet the BW and latency requirements for Red Storm.
  - Need to pay the NRE costs somehow; takes 24-30 months to bring it to production

Red Storm Interconnect Choice

- Custom interconnects, if possible
  - If cost & schedule can be controlled, this is the best solution
    - should permit rolling upgrades
    - meets all scaling targets
- Quadrics (with modifications) might be an acceptable alternative

Sandia National Laboratories

Interconnect Topology

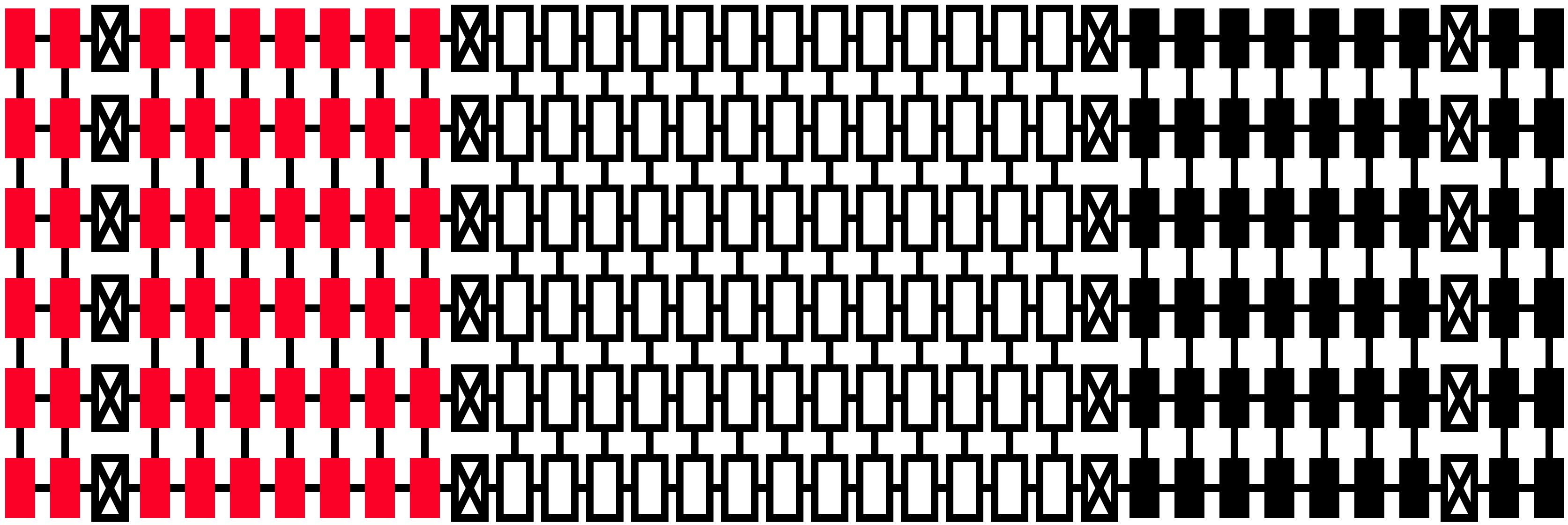
- Large Switches
  - Full Xbar (Some folks' Holy Grail)
  - IBM Colony & follow-on
  - Quadrics Fat Tree
  - Myricom Clos Switch
- Mesh or Mesh-like
  - e.g., Cplant, ASCI Red; T3E; Cray SV-2*, …

Red Storm Topology Choice

- Switch Topology (ignoring photonic switches)
  - Large Switches
    - These are excellent for modest-size clusters.
    - Their cost grows faster than linearly and the cabling issues grow enormously difficult for large systems
  - 3D meshes
    - Cost is linear in both switches and in cables.
    - For our applications on a large system, this is by far the best choice.

How to Expand a Mesh

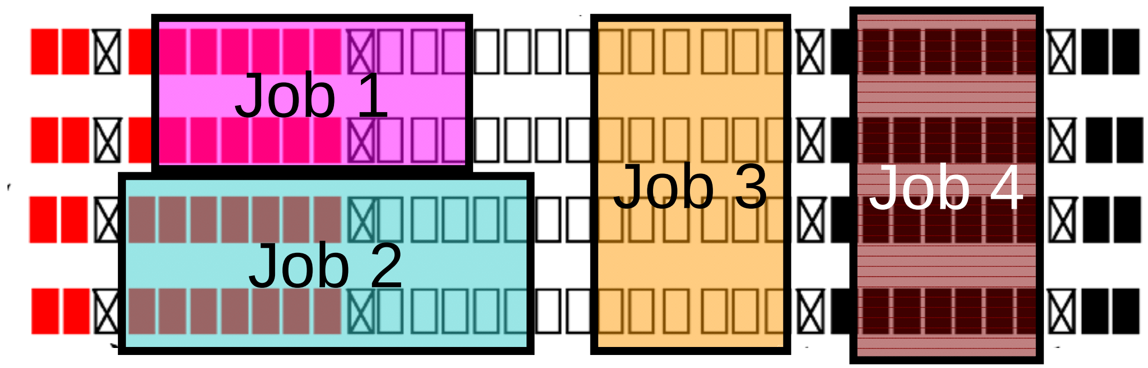
- **Mesh topology is the same as the machine room topology; just put more down**



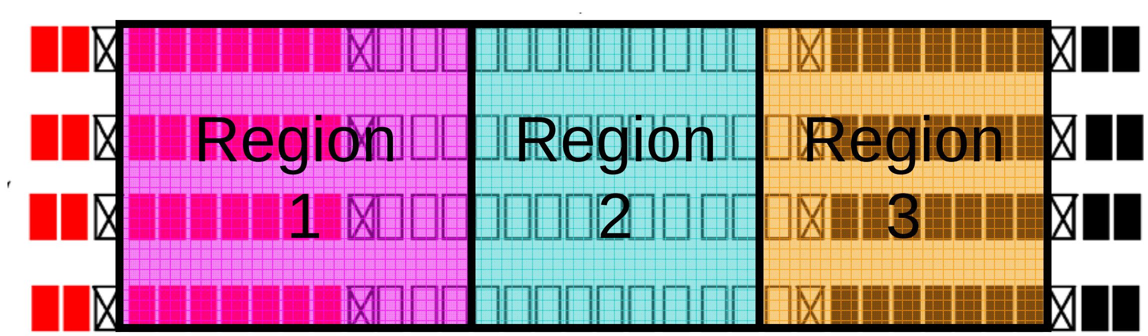
Red Storm Cabinet Layout in Machine Room

Communications Locality in Sandia's Jobs

- Space-shared capacity load



- Problems of a physical origin

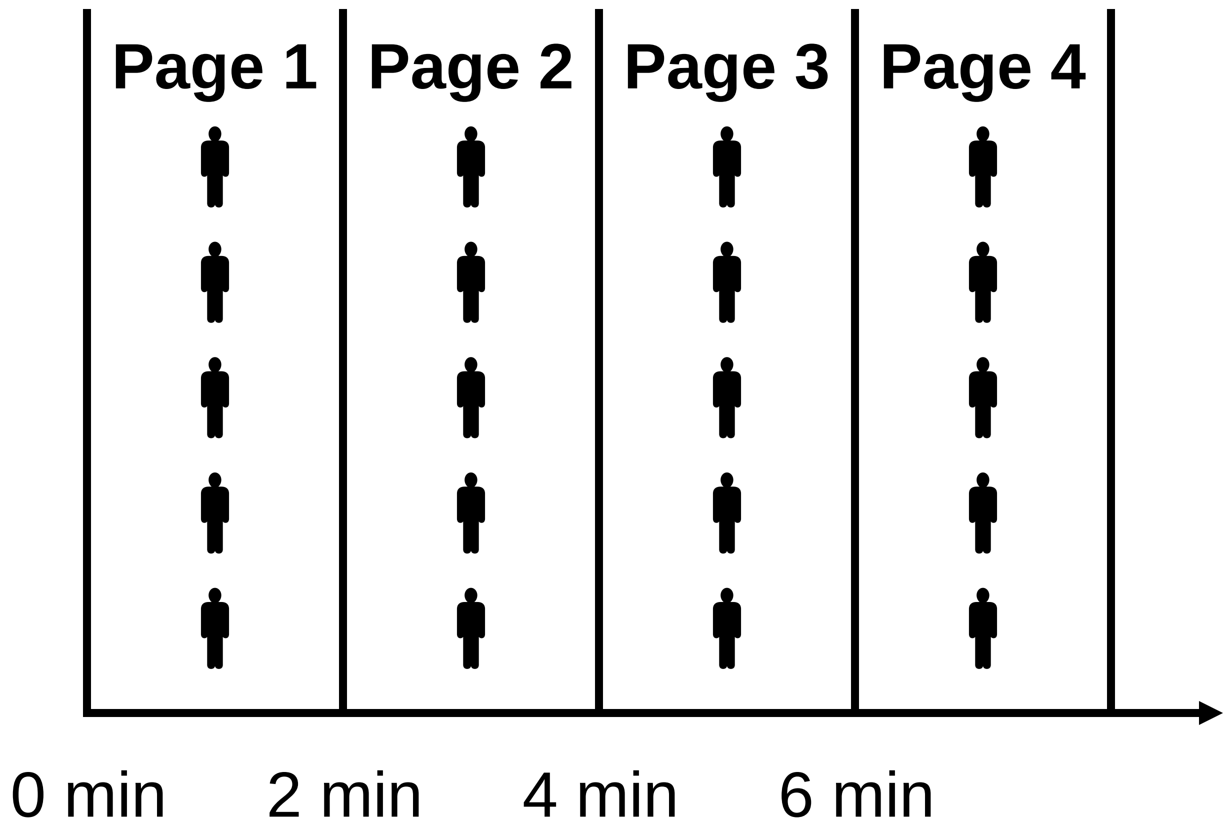


Light Weight Kernel

- 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

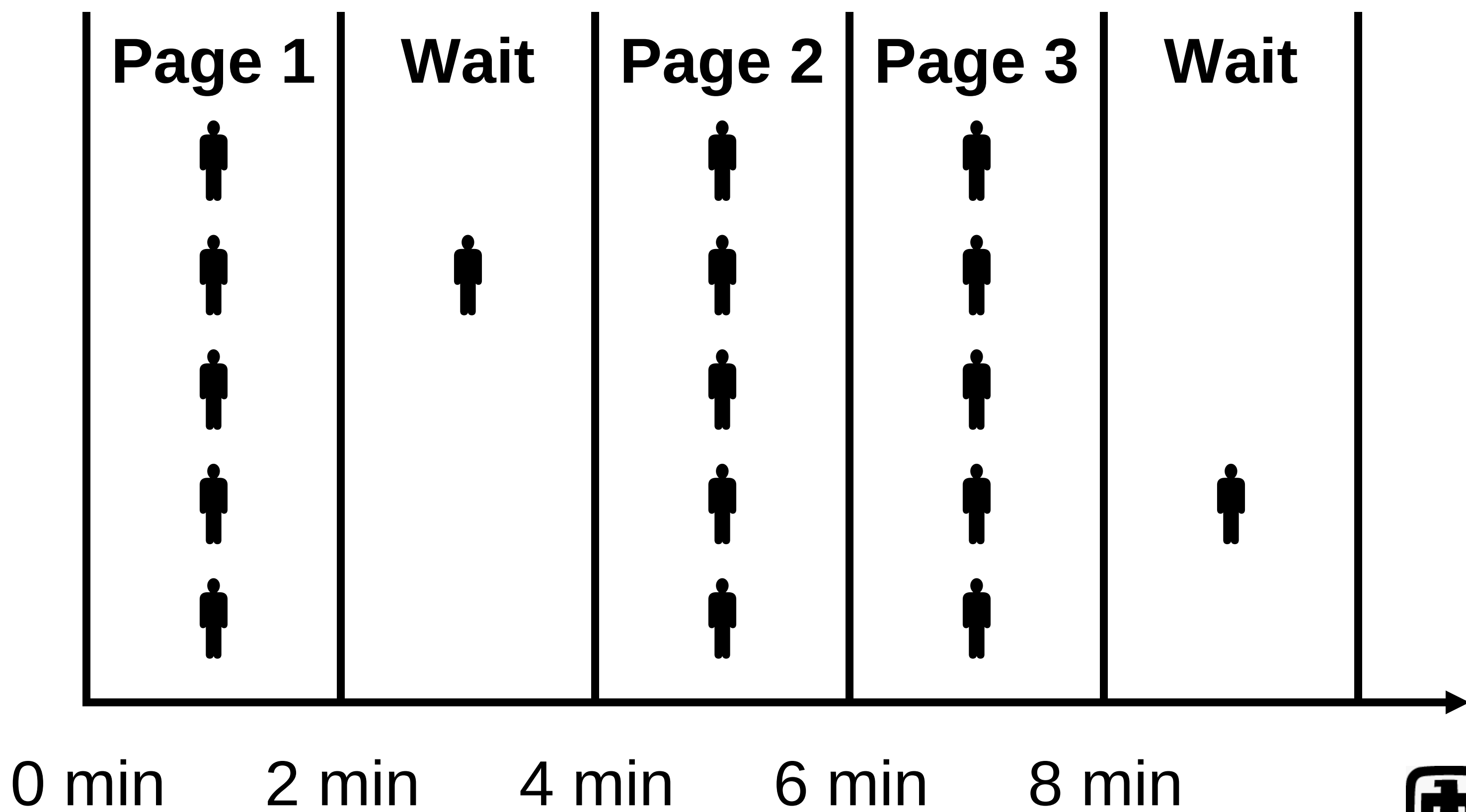
- N musicians Rehearsing 2 Minute Pages of Music



Sandia National Laboratories

Musical Rehearsal with Breaks

- 2 Minute Pieces with Asynchronous Breaks



Sandia National Laboratories

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
  - None

Run Time Impact of Unix Systems Services

- Say breaks take 50 μS and occur once per second
  - On one CPU, wasted time is 50 μs 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 have 10,000 CPUs, hence LWK approach important

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

Sandia National Laboratories

Reliability

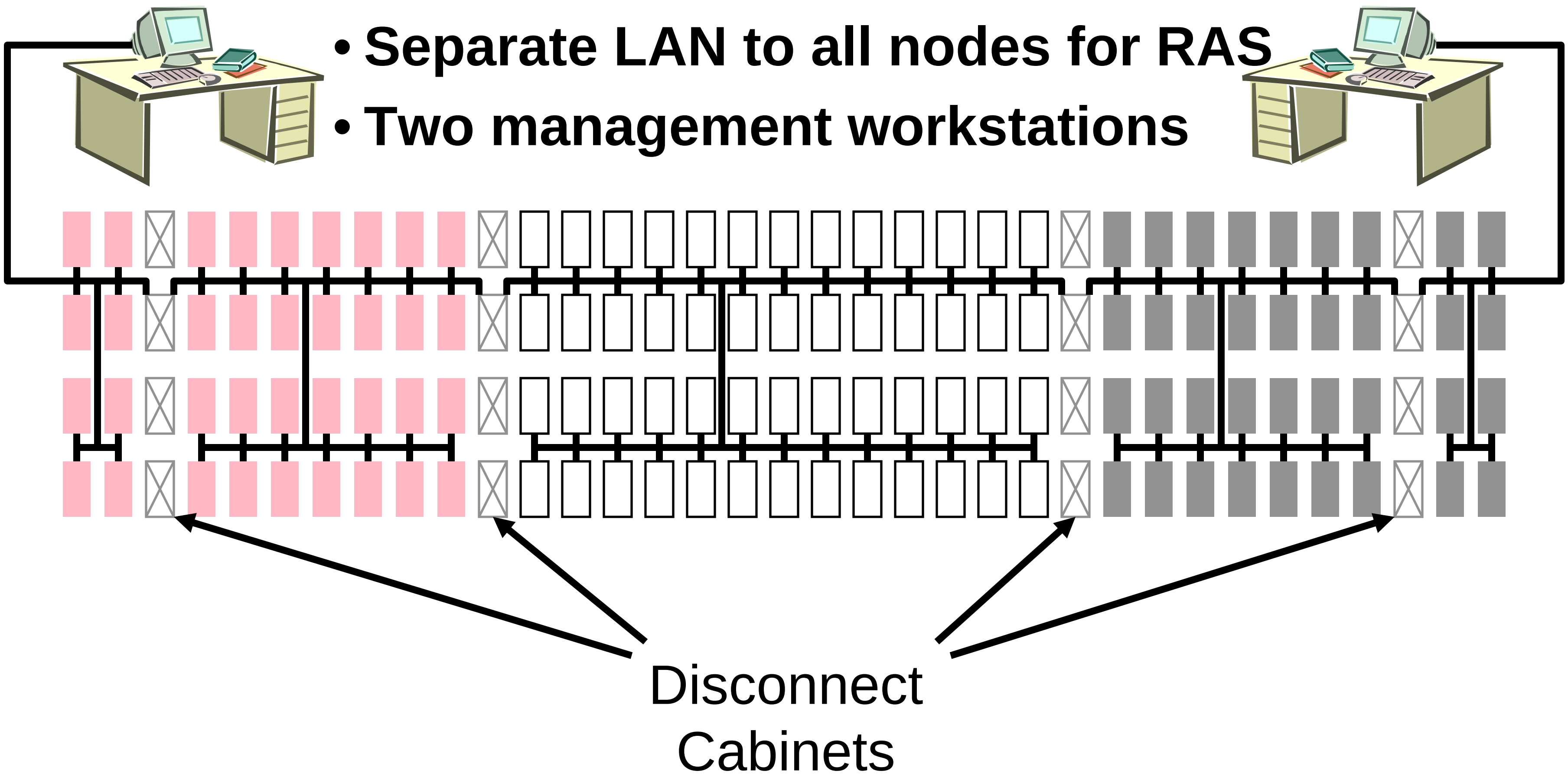
- What is Reliability for Scientific Applications at Sandia
  - High Mean Time Between Interrupts (MTBI) for hardware and system software
  - High Mean Time Between Errors/Failures (MTBF) that affect users
- What it is not
  - High availability

Reliability Features for Red Storm

- Light Weight Kernel (LWK) O. S. on compute partition
  - Less code fails less often
- Monitoring of correctible errors
  - Fix soft errors before they become hard
- Hot swapping of components
  - Overall system keeps running during maintenance
- Redundant power supplies & memories
- RAS System

Reliability, Availability, and Serviceability

- **Separate LAN to all nodes for RAS**
- **Two management workstations**



RAS Workstation Capabilities

- Separate and redundant RAS workstations for secure and general availability portions of the machine
- Error logging and monitoring for major system components including processors, memory, NIC/router, power supplies, and disks
  - Shows diagram of machine and highlights positions of nodes/boards requiring service
- Configure/deconfigure any board
  - Hot swap any Field Replaceable Unit (FRU) without disturbing running applications

Economy

- Red Storm leverages economies of scale
  - AMD Opteron microprocessor & standard memory
  - Air cooled
  - Electrical interconnect based on Infiniband
  - Linux operating system
- Selected use of custom components
  - System chip ASIC
    - Exceptionally important to mission
  - Light Weight Kernel
    - Truly custom, but we already have it

Sandia National Laboratories

Selected Specifications

- 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

Selected Specifications

- 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

Selected Specifications

- Balance
  - Peak of ~40 TF
  - Aggregate system memory bandwidth ~55 TBytes/second
    - 2.375 bytes per peak flops
  - Aggregate sustained interconnect bandwidth > 100 TBytes/second
    - 2.5 bytes per peak flops
  - Link Bandwidth ~3.0 GBytes/second each direction
    - ~1.5 Bytes/ peak flops

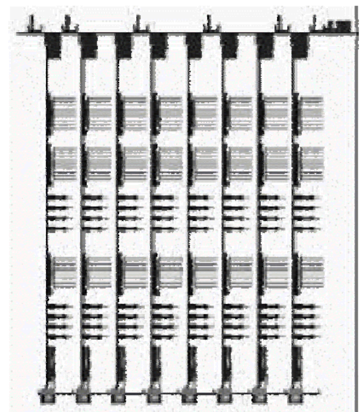
Selected Specifications

- Disk Subsystem
  - 240 TBytes total storage (120 secure + 120 general availability)
  - 100 GBytes/second sustained I/O rate (50 secure + 50 general availability)
- High Speed Networking
  - 50 GBytes/second sustained network I/O rate (25 secure + 25 general availability)

Extra Slides

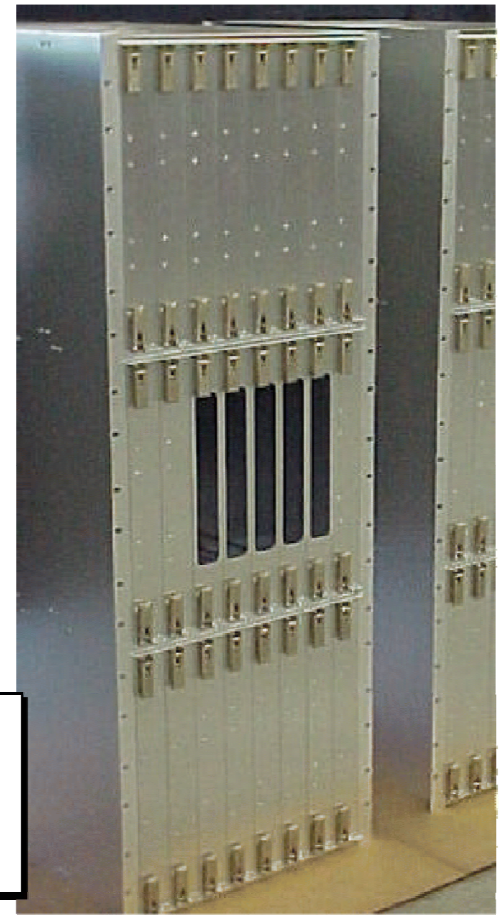
A Building for Red Storm

Red Storm Hardware Status



Card Layout - Top View

- **24 Boards**
- **96 Operton™ Processors**
- **EMI containment**
- **Vertical Air Cooling**



Red Storm Hardware Status

Pre-Prototype Racks

Comparison of ASCI Red and Red Storm

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

Comparison of ASCI Red and Red Storm

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

Sandia National Laboratories