SF 1008-RA (FRONT 8-2003)
Supersedes (7-2003) issue

Information Release

SAND 2003-3124A+P **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.

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 Section 6 and to file it for future reference.
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).
For information on which R&A process to use, or additional R&A information:

- See the "Review and Approval for Communication" webpage: http://www-irn.sandia.gov/recordsmgmt/revapprov/revapprov.htm or
- Contact Linda Cusimano (NM), (505) 844-4980 ; Kelly McClelland (CA), (925) 294-2311

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 **ASCI Red Storm and Supercomputer Scalability**

Author or contact (Sandian) **Erik P. DeBenedictis** (Full First Name Full Middle Name Full Last Name) Signature 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** (Identifies funding source – will not be charged)

☐ Contract Author to Sandia. (Contractor's name and contract 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
☒ 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: **Conference on Supercomputations**
Place of Event: **VNIIEF, Sarov, Russia** Date: **10/6/03 thru 10/10/03**
Internet Address of Electronic Posting:

SECTION 4. Classification and Sensitivity of Information. Contact Classification Dept. 12225 (8511 - CA) for questions.

Indicate classification level and category of information release or whether information release is unclassified:
Classification of: Document Title U Document Abstract N/A The Document U

☐ Classified - Limited Release. Indicate additional access restrictions:
☐ NWD Sigma ☐ CNWDI ☐ NOFORN ☐ Program Designated Special Handling (Distribution Limitation) ☐ Other

☐ Unclassified - Limited Release. Indicate all Unclassified Controlled Information (UCI) categories access restrictions:
☐ Export Controlled Information (ECI) ITAR/EAR/
☐ 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 or DUSA Delegate: PAUL YARRINGTON (Name) Paul Yarrington (Signature) 9230 (Org.) 8/18/03 (Date)

(Continued on page 2)

2003 - 3124 A+P

SF 1008-RA (8-2003)
Supersedes (7-2003) issue

SECTION 5. Disclosure of Technical Advance

A Technical Advance is an original achievement or non-obvious progress in a scientific or engineering sense, including the creation of software. It may be protected by patent or copyright. 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-TD, been filed with the Sandia Patent and Licensing Center?
☐ Yes SD No. _____ ☐ No If No, please follow up with a TA form obtainable from:

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)	**1110**	**284-4017**		**8/11/03**
(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 USED FOR THE FORMAL REVIEW & APPROVAL PROCESS.

SECTION 7. To be completed by the Patent and Licensing Department (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 8-25-03

*** Only required for Technical and Scientific Information releases.**

SECTION 8. To be completed by the Classification and Sensitive Information Department (NM: 12225/MS 0175, CA 8511,.MS 9021).

Signature _____ Date 8/26/03

SECTION 9. To be completed by the Promotional Communications Departments (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). In addition, 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 a 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. To be completed by the Review & Approval Desk (NM: 9612/MS 0612, CA: 8511/MS 9021).

Approved under the following conditions:

1. **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.**
2. **That, if the release has been determined to be sensitive or classified, all the appropriate markings are on the released item;**
3. **That all edits requested by reviewers are completed before release;**
4. **That the final version is submitted to the Technical Library.**

Add the funding statement to the abstract

Signature Dorothy Martin Date AUG 27 2003

SAND 2003-3124P

ASCI Red Storm and and Supercomputer Scalability

Dr. Erik P. DeBenedictis
Sandia National Laboratories

Symposium on Supercomputations
Sarov, Russia

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.

Sandia National Laboratories

Outline

- Red Storm Overview
- Scalability
- Light Weight Kernel

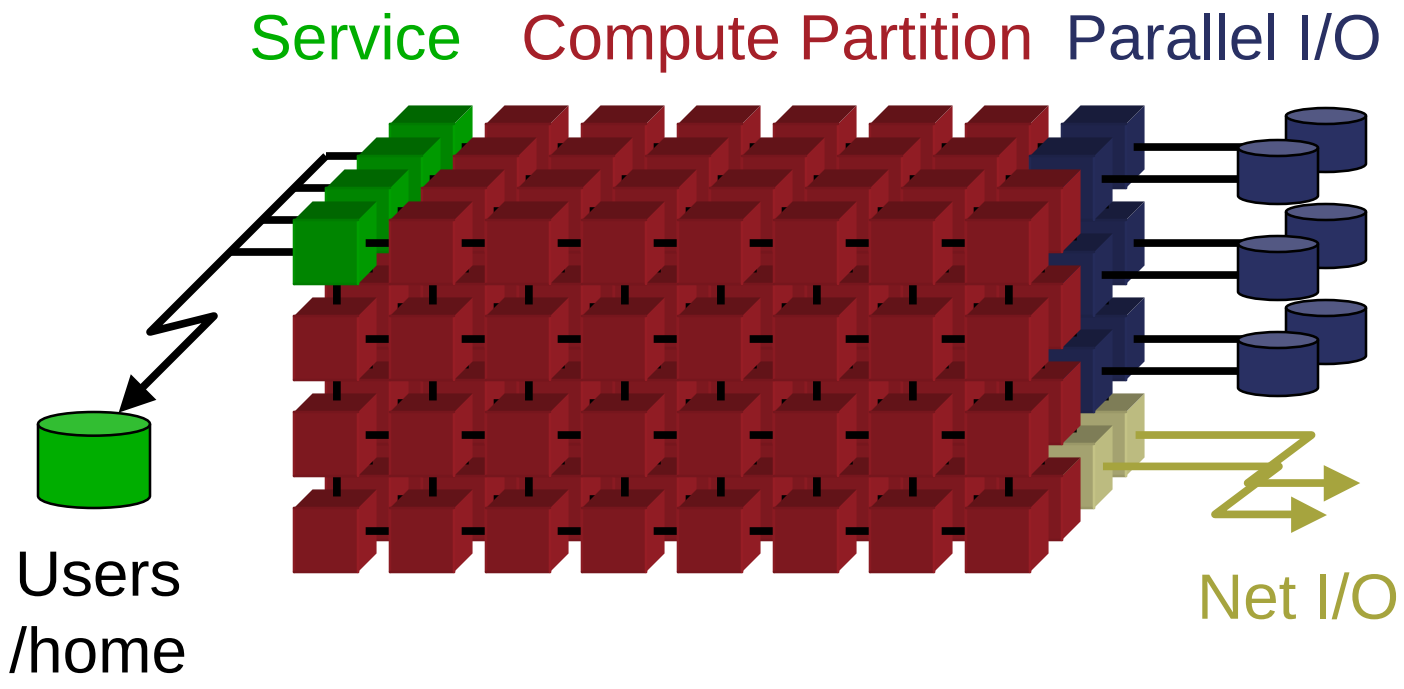
Sandia National Laboratories

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

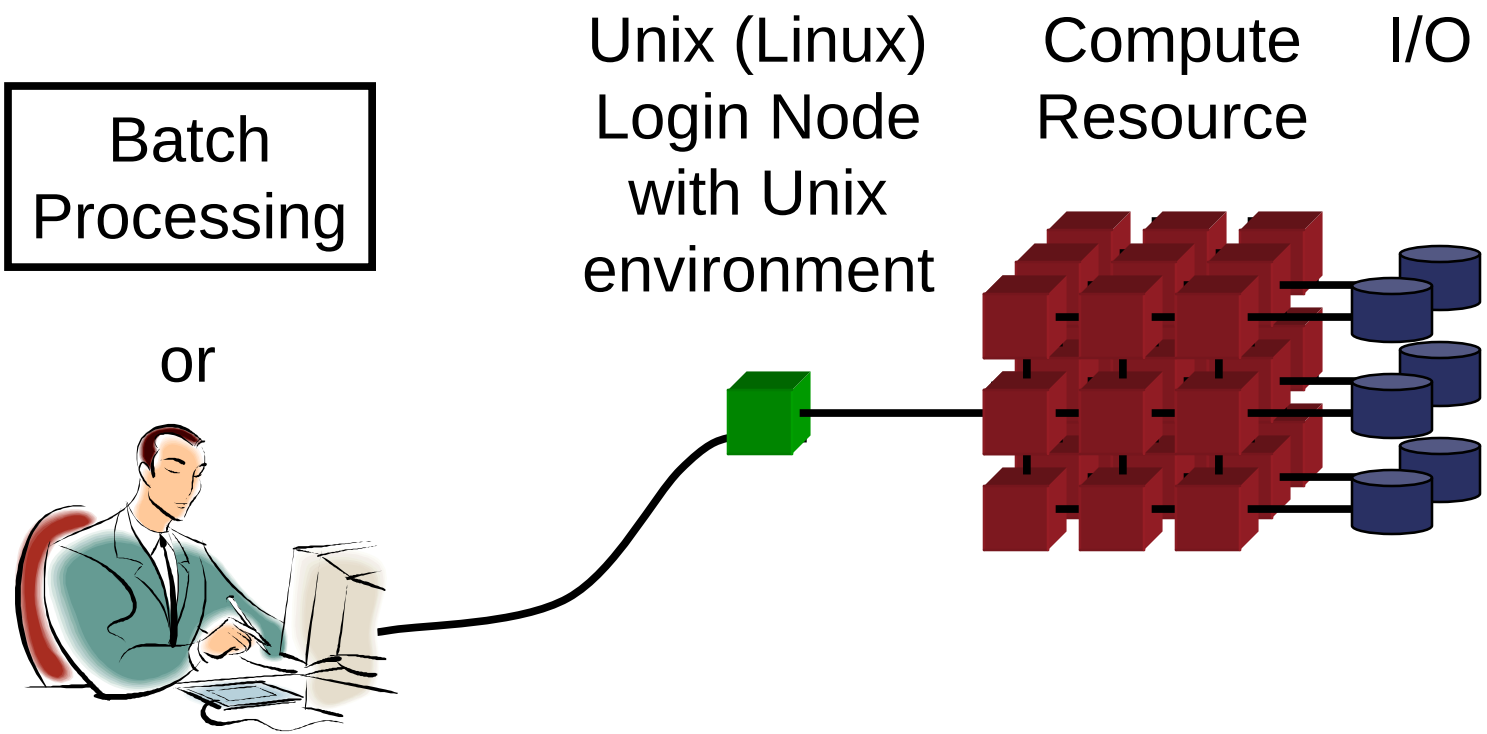
Sandia National Laboratories

Red Storm is a Massively Parallel Processor

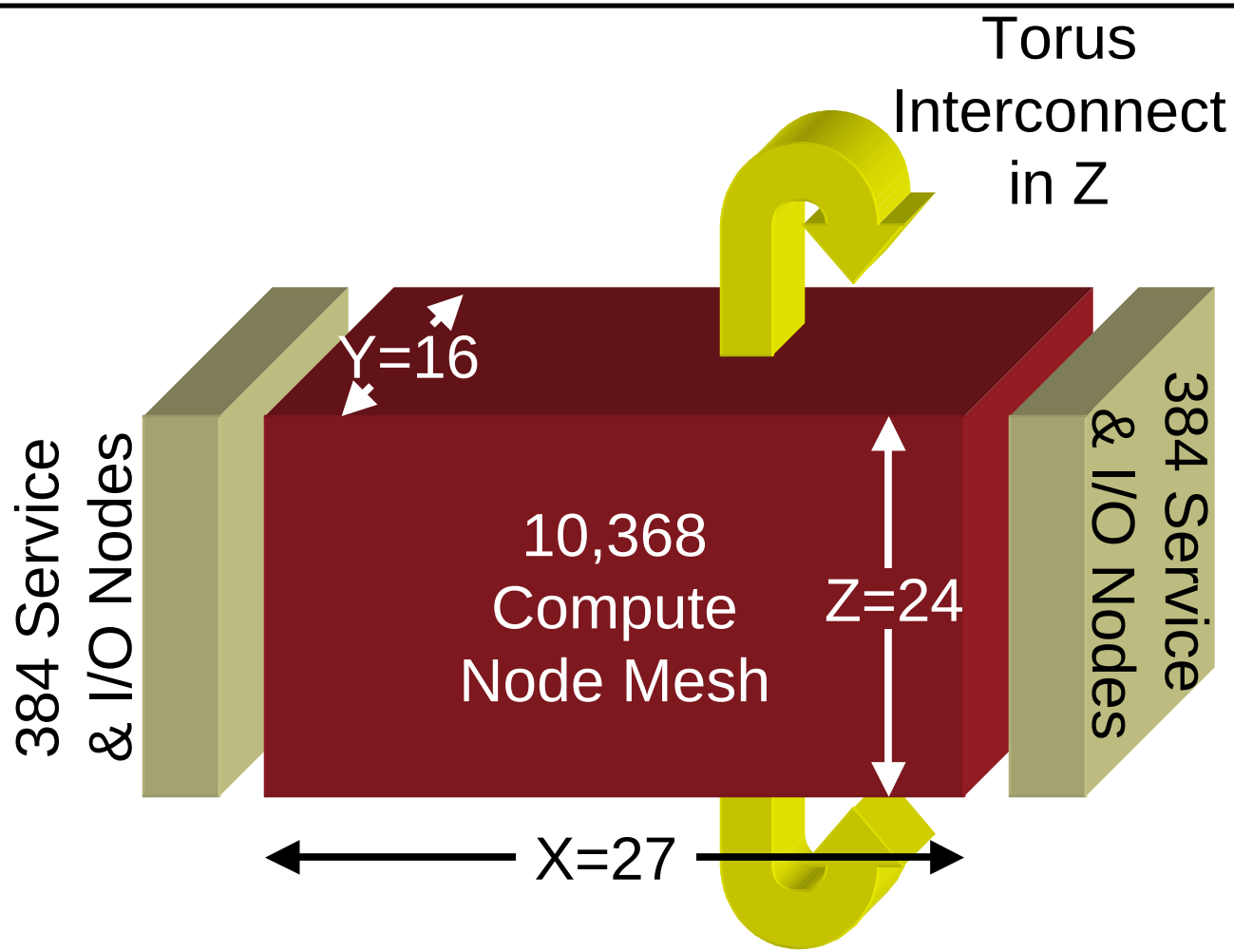


Sandia National Laboratories

Usage Model


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

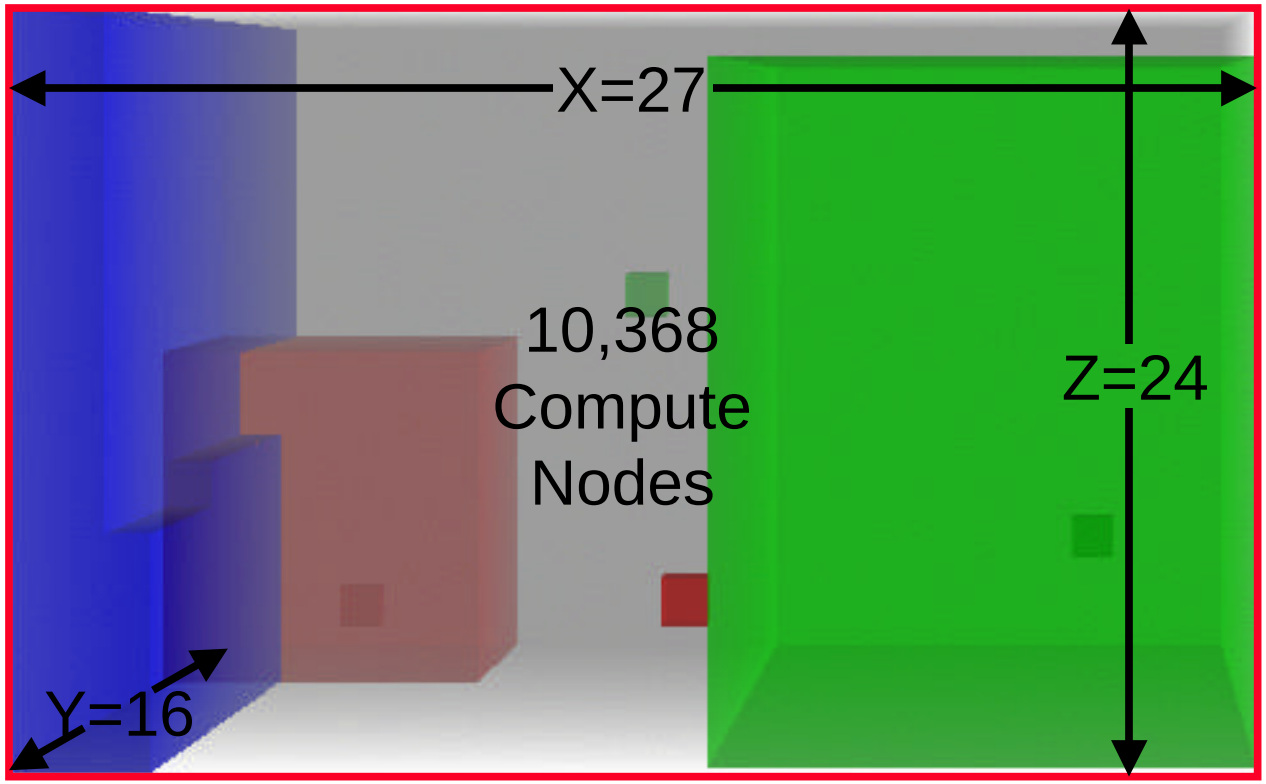
27×16×24 3D Mesh/Torus + I/O



Sandia National Laboratories

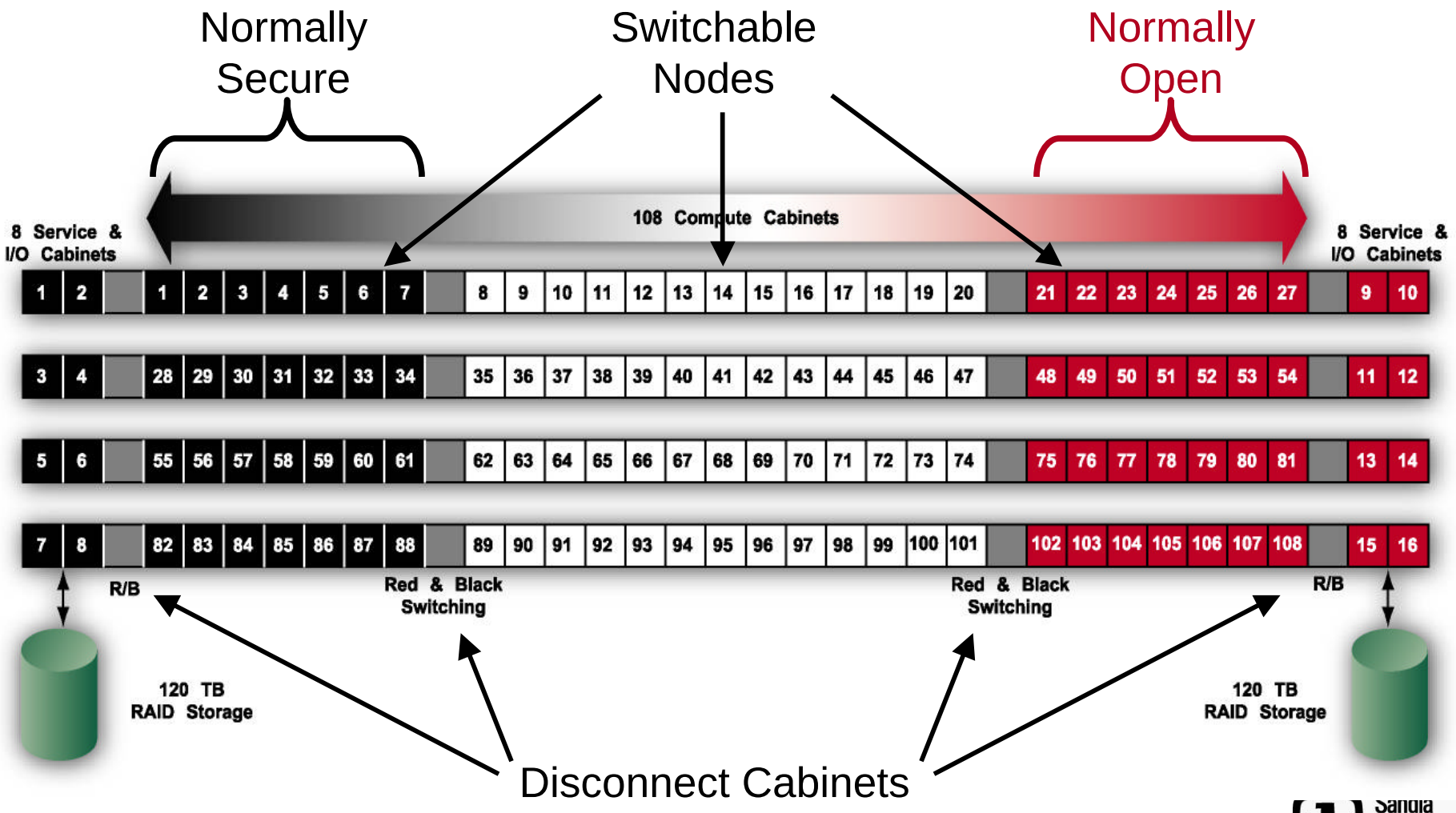
Space Sharing of Jobs

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



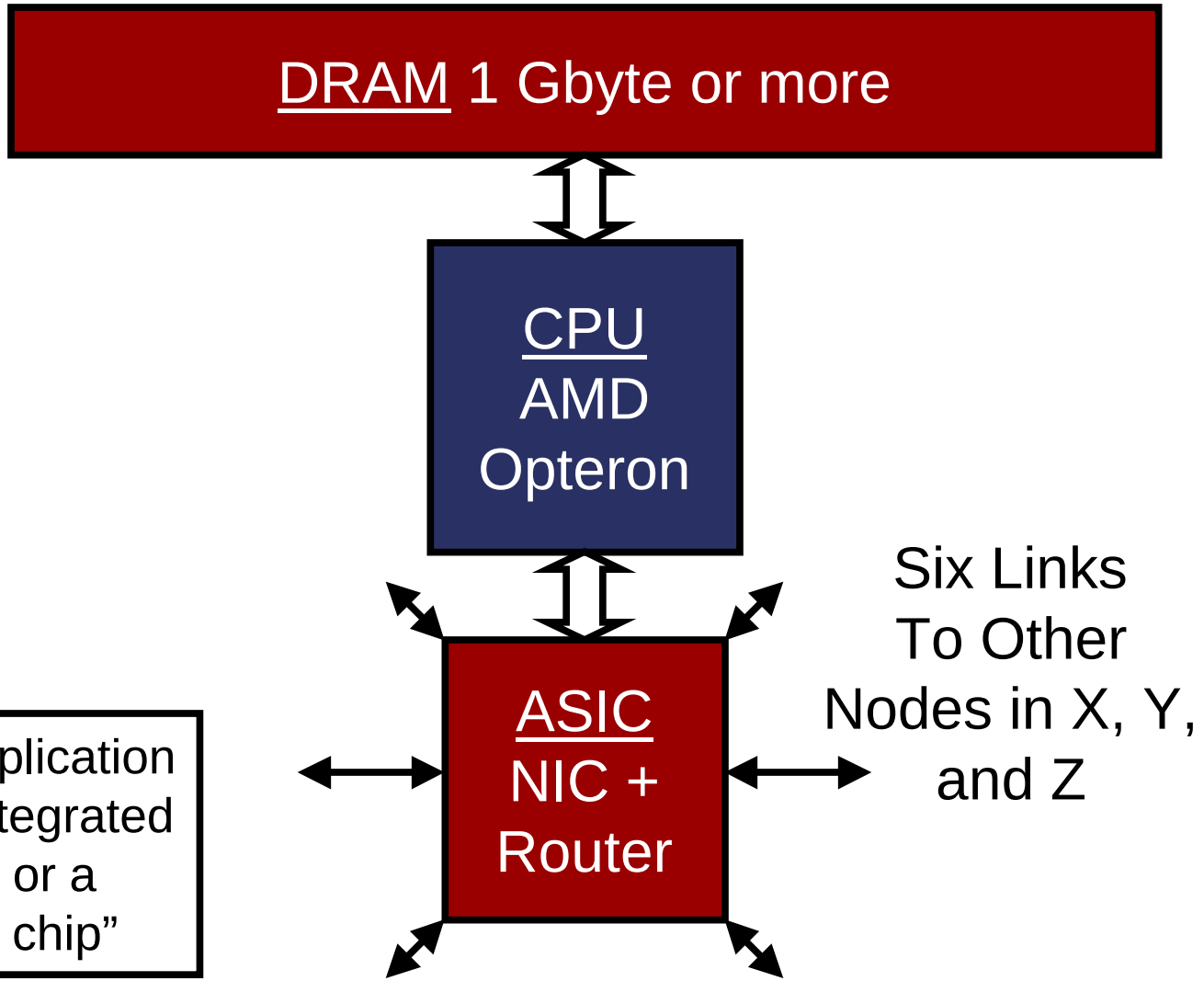
Sandia National Laboratories

Red Storm Hardware Overview



Sandia National Laboratories

Node Architecture



ASIC = Application Specific Integrated Circuit, or a "custom chip"

Sandia National Laboratories

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?

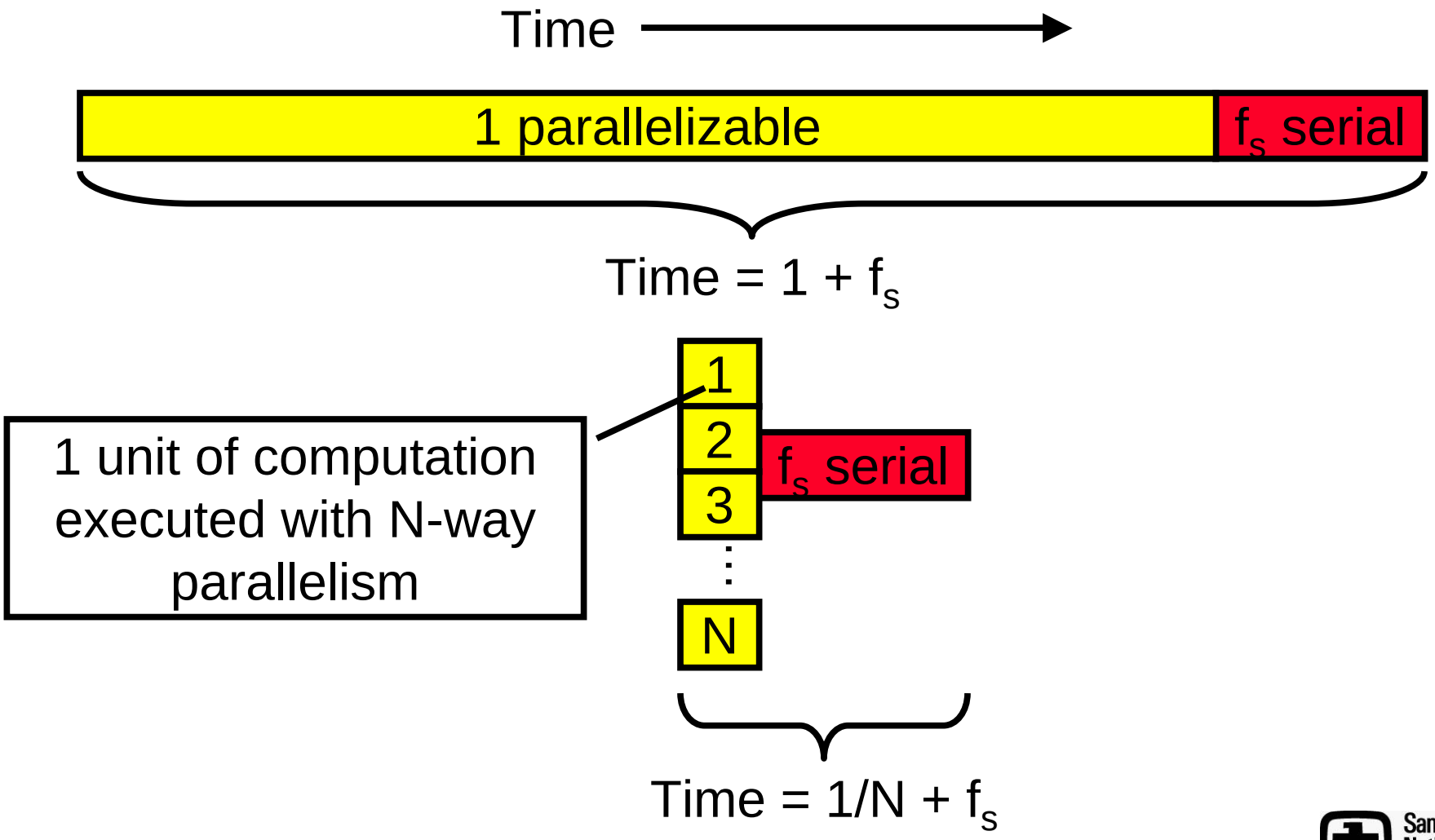
Amdahl's Law

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

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.

Amdahl's Law Picture



Sandia National Laboratories

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

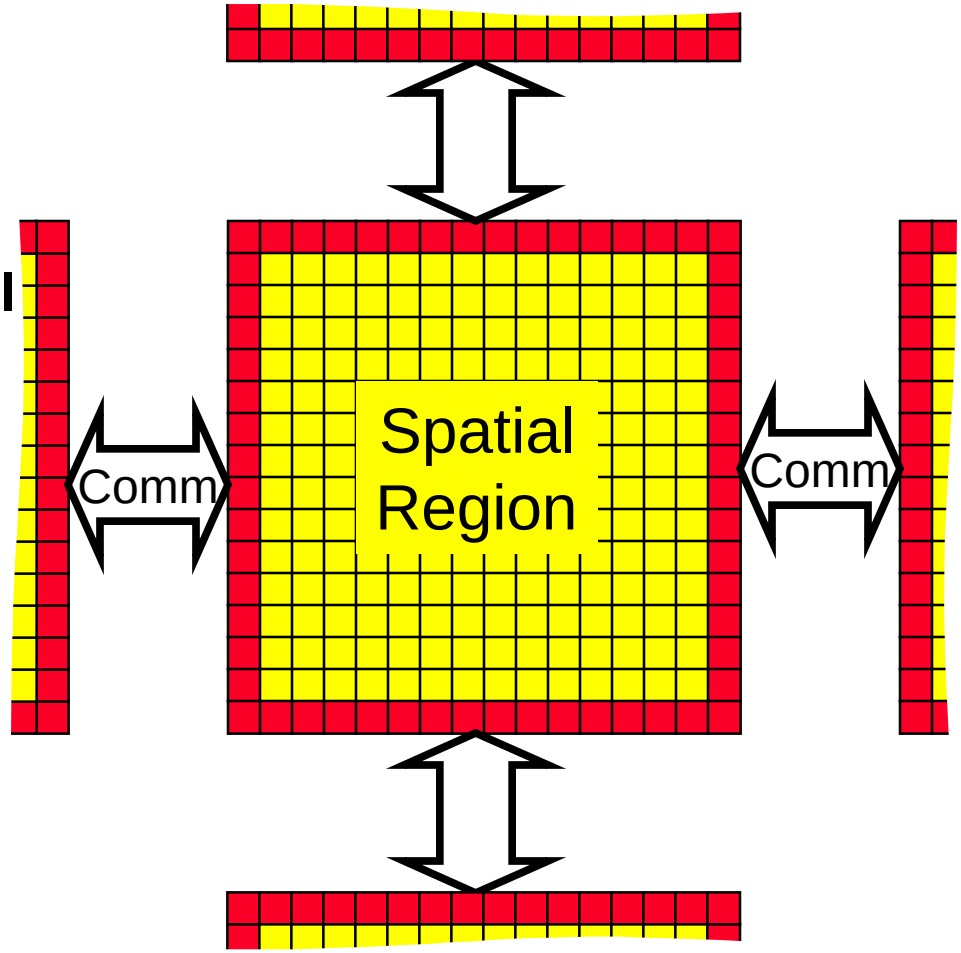
$$S(N) \sim S_{Amdahl}(N)/[1 + f_{comm} \times R_{p/c}],$$

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**

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



Sandia National Laboratories

Implications of Realistic Amdahl's Law

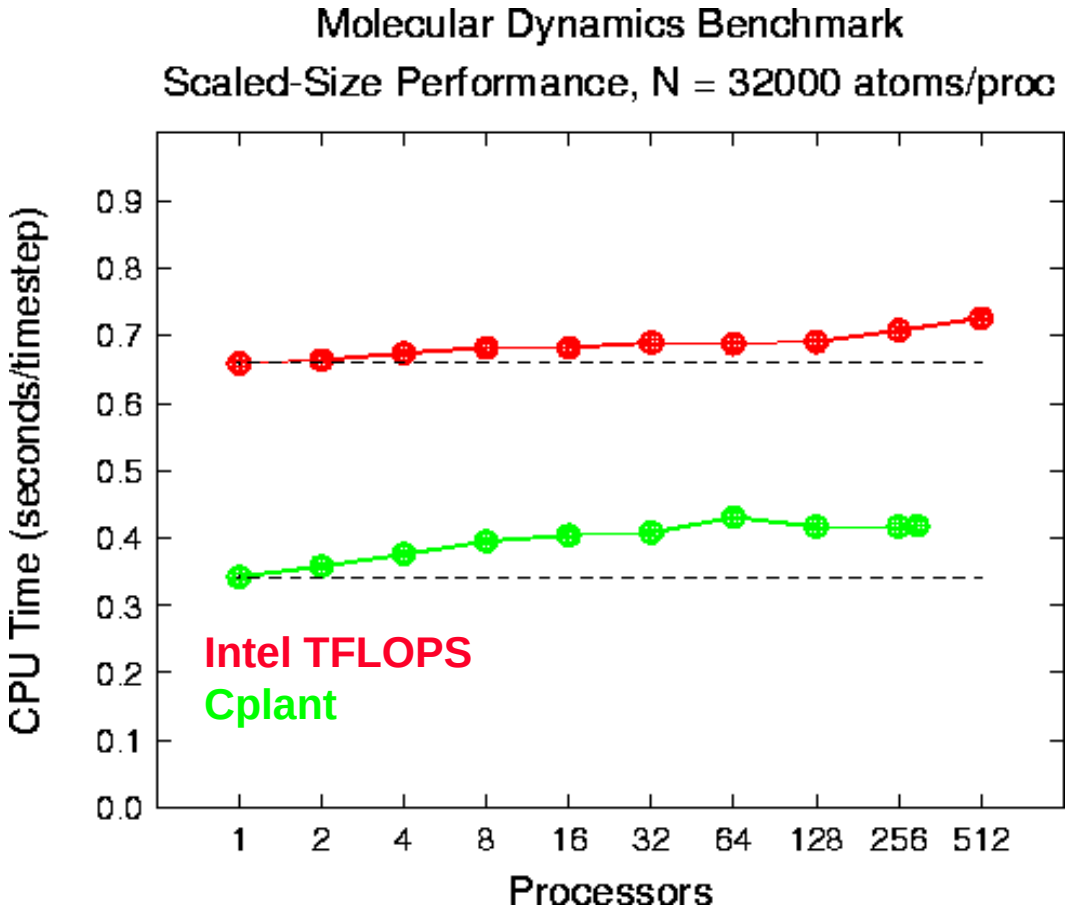
- Let's consider three cases on two computers:
  - The two computers are identical except that one has
    - $R_{p/c}$ = 1 Byte/FLOP (fast communications)
    - $R_{p/c}$ = 0.05 Byte/FLOP (not so fast communications)
  - The three cases are
    - f_{comm} = 0.01,
    - f_{comm} = 0.05, and
    - f_{comm}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

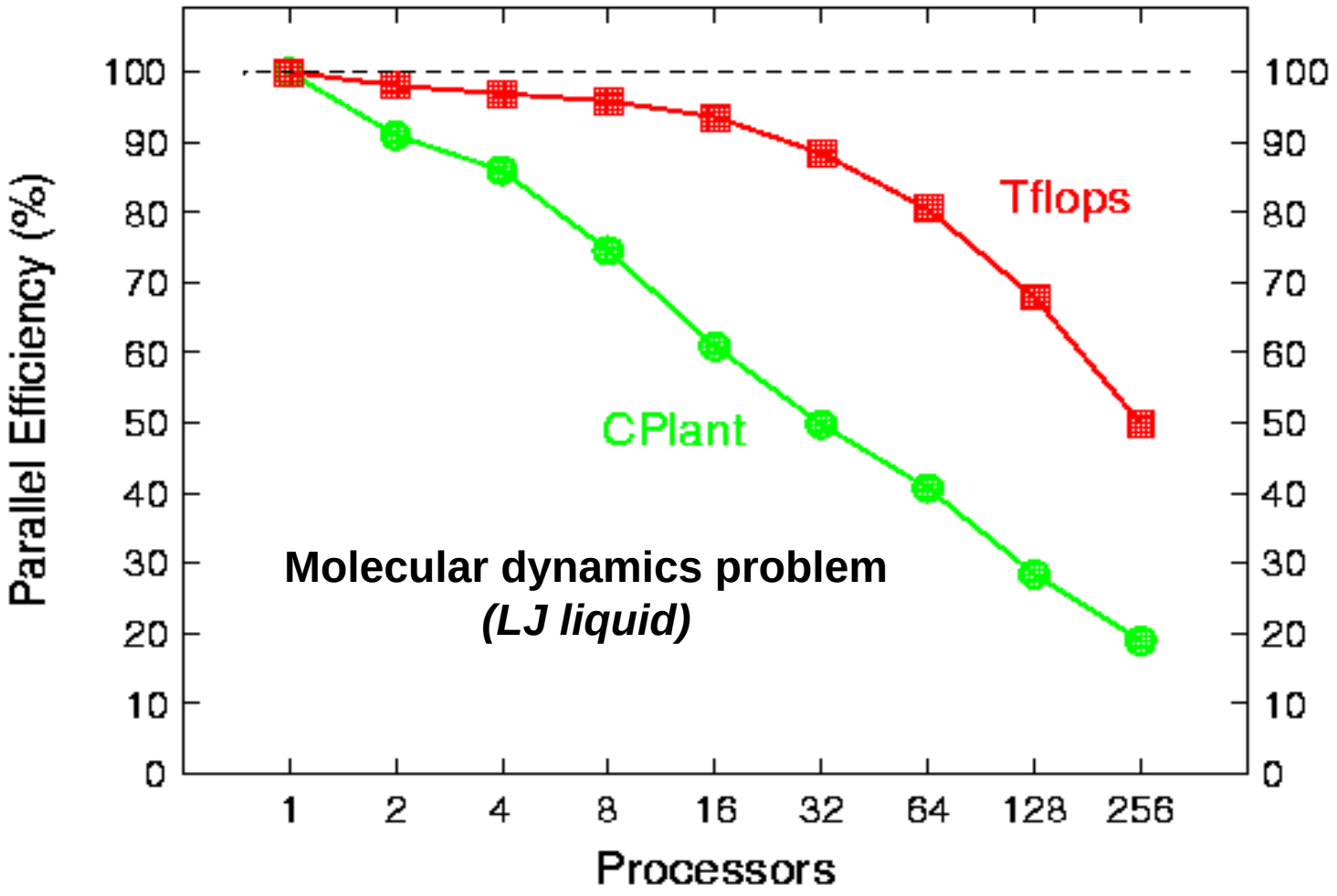
Sandia National Laboratories

Sandia Experience with $R_{p/c}$



Sandia National Laboratories

Sandia Experience with $R_{p/c}$



Sandia National Laboratories

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

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**

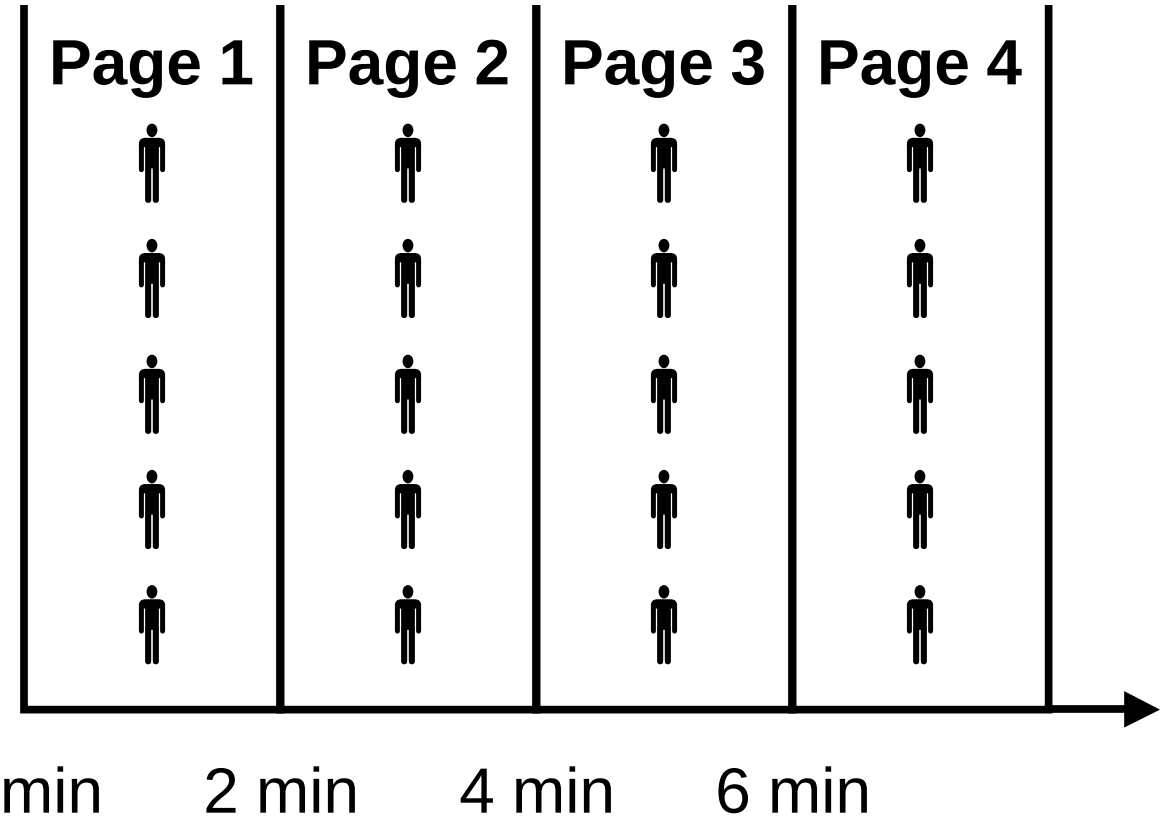
Sandia National Laboratories

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

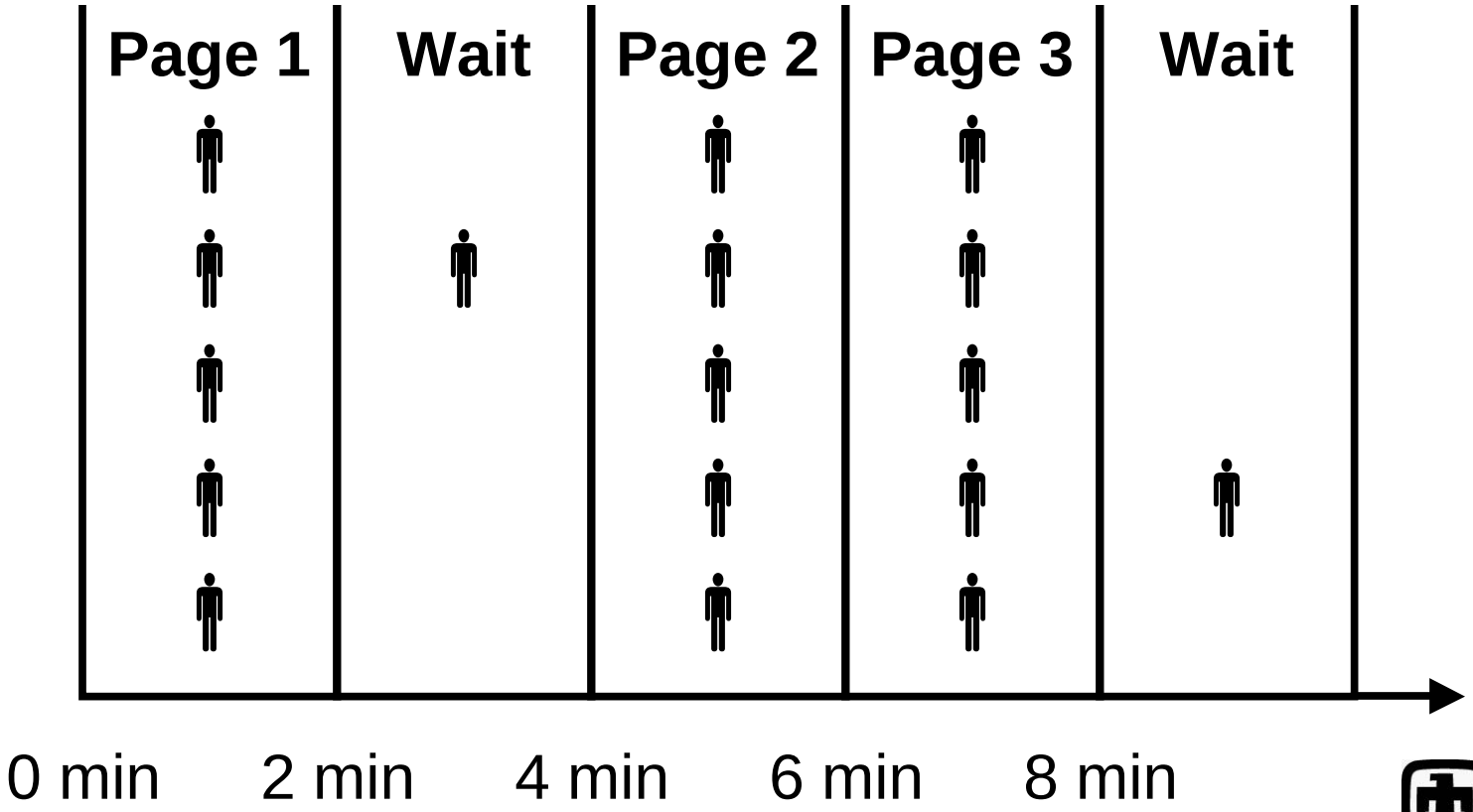
- **N musicians Rehearsing 2 Minute Pages of Music**



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

Conclusions

- Red Storm is under construction as a 40 TFLOPS supercomputer
  - Delivery in about one year
- Built on engineering principles of ASCI Red
  - Expected to perform 7x as efficiently
- Performance analysis indicates that the architecture can be scaled considerably beyond Red Storm