7X Performance Results – Final Report: ASCI Red vs. Red Storm

Joel O. Stevenson, Robert A. Ballance, Karen Haskell, and John P. Noe
Sandia National Laboratories

Dennis C. Dinge, Thomas A. Gardiner, and Michael E. Davis
Cray Inc.

Cray User Group 2008 Crossing the Boundaries, May 5-8, 2008

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 of Today's Discussion

- 7X: ASCI Red vs. Red Storm
  - Goal of 7X performance testing is to assure Sandia, Cray, and DOE that Red Storm will achieve its performance requirements.
  - 7X performance suite consists of ten applications. Describe applications and selection criteria.
  - Identify one or more problems for each application, run those problems at two or three processor sizes, and compare the results between ASCI Red and Red Storm - 25 cases under study. Discuss results.
- 7X: SN vs. VN Results on Red Storm
  - Each Red Storm compute node has dual core topology.
    - SN option – ignore the second processor – default mode.
    - VN option – treats each processor as a separate compute node.
  - During the course of executing 7X applications on Red Storm, results were collected in both SN and VN mode. Discuss results.

7X: ASCI Red vs. Red Storm

Red Storm Performance Evaluation

- Goal of 7X performance testing is to assure Sandia, Cray, and DOE that Red Storm will achieve its performance requirements.
- The 7X performance suite consists of ten applications and benchmarks that will be used in Red Storm performance testing and evaluation.
- Approach: Identify one or more problems for each application, run those problems at two or three processor sizes, and compare the results between ASCI Red and Red Storm - 25 cases under study.

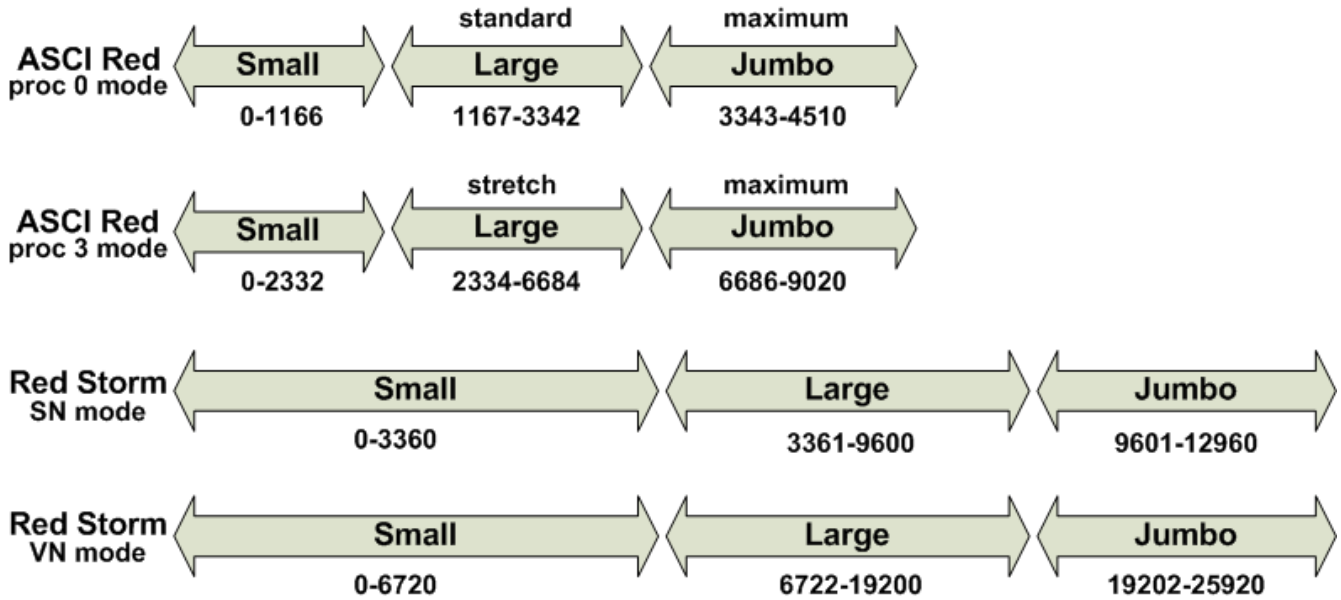
Application Selection Criteria

- Problem sets shall be “real”. The 7X testing effort represents production job behavior with actual input files and algorithms.
- The same calculations shall be run on ASCI Red and Red Storm. The primary metric is wall-clock time as measured by the elapsed time to execute the entire job script, including any pre and post processing.
- Calculations on ASCI Red and Red Storm should give equivalent answers.
- Problems should be chosen to use as many ASCI Red resources (processor, memory) as possible in order to place reasonable stress on Red Storm.
- Jobs run on ASCI Red should range from ~4-8 hours.

Application Selection Criteria (cont.)

- Simplified geometries are preferred in order to simplify input file creation and to avoid meshing problems during benchmarking.
- All applications should use standard production-use capabilities including I/O, checkpoint/restart, and visualization files.
- When an application can be run using alternative algorithms, such as Alegra with and without contact, that application may have more than one benchmark problem in the suite.
- We will test applications in three modes : standard, stretch, maximum.

Modes: Standard, Stretch, Maximum



ASCI Red: 4510 compute nodes (9020 processors). Proc 0 mode uses one processor per node and the full 256 MB of memory. Proc 3 mode uses two processors per node but only 128 MB of memory is available to each process.

Red Storm: Upgraded to 12960 compute nodes (25920 processors). Each node is dual core topology with minimum 2 GB of memory per node. Memory available to each process is halved when using two processors per node.

Modes: Standard, Stretch, Maximum

- Standard - the standard size should be easily run and accurately measured on both platforms. Standard will be used to calibrate the testing and to check for shifts in performance due to changes in the underlying system software.
  - Standard refers to “Large – proc 0” on ASCI Red and “Small” on Red Storm.
- Stretch - the stretch size will fully occupy the large configuration of ASCI Red. Problem sets will need to accommodate the reduced memory available in ASCI Red stretch mode.
  - Stretch refers to “Large – proc 3” on ASCI Red and “Large (SN)/Small (VN)” on Red Storm.
- Maximum - selected applications may also be run in maximum size that requires an operational configuration of ASCI Red’s entire compute node partition.
  - Maximum refers to “Jumbo – proc 0 or Jumbo – proc 3” on ASCI Red and “Large (SN)/Small (VN) or Large” on Red Storm.

Application Descriptions

- Alegra with Contact
  - Quasistatic electromechanics (QSEM) problem in which a curved impactor depoles a potted active ceramic element.
    - Standard - 2048 processors
    - Stretch - 6484 processors
- Alegra without Contact
  - QSEM problem identical to the contact problem except the boundary condition is a prescribed displacement rather than an impactor, eliminating the need for contact.
    - Standard - 2048 processors
    - Stretch - 6484 processors
- CTH
  - Shock physics (3D of a large conical shaped charge).
    - Standard - 2000 processors
    - Stretch - 6480 processors
    - Maximum - 9000 processors

Application Descriptions (cont.)

- ITS
  - Monte Carlo solution of linear time-independent coupled electron/photon radiation transport problems, with or without the presence of macroscopic electric and magnetic fields of arbitrary spatial dependence.
    - Standard - 3200 processors
    - Maximum - 4500 processors
    - Stretch - 6500 processors
    - Maximum - 9000 processors
- PARTISN
  - Sntiming problem - flux and eigenvalue convergence as monitored by Partisn (Parallel Time-dependant SN transport).
    - Maximum - 4096 processors
    - Stretch - 6480 processors
    - Maximum - 8930 processors

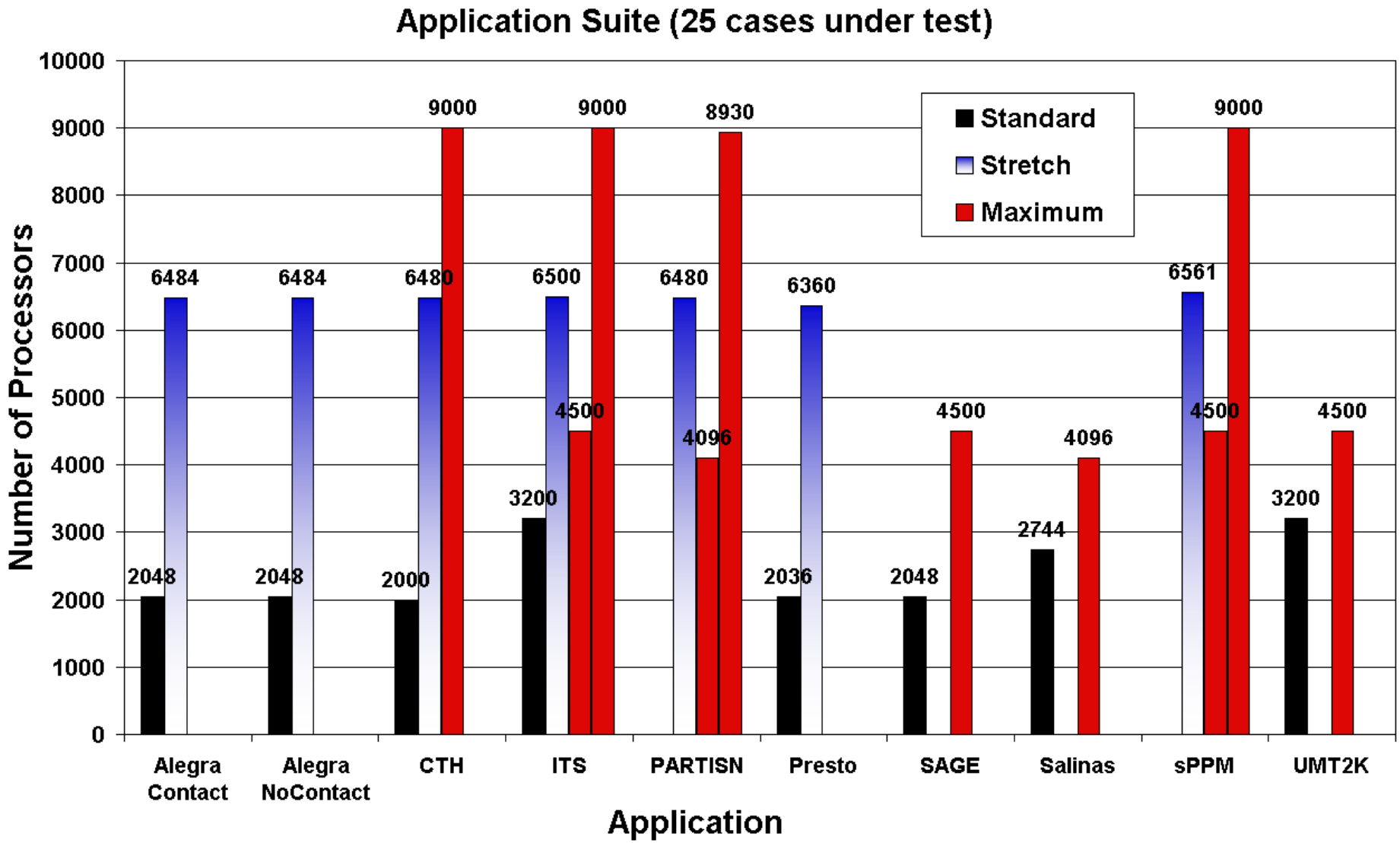
Application Descriptions (cont.)

- Presto
  - Rectangular bricks stacked in an alternating fashion in a plane to produce a wall three elements thick. Four walls are lined up in the thin direction and then given a sudden pressure loading such that they compress against each other.
    - Standard - 2036 processors
    - Stretch - 6360 processors
- SAGE
  - Asteroids simulation - 45 degree, 3D, asteroid impact into stratified medium of water, calcite, granite crust, and mantle.
    - Standard - 2048 processors
    - Maximum - 4500 processors
- Salinas
  - Transient dynamics problem – one unit cube model.
    - Standard - 2744 processors
    - Maximum - 4096 processors

Application Descriptions (cont.)

- sPPM
  - Shock physics - solves a 3D gas dynamics problem on a uniform Cartesian mesh, using a simplified version of 3D hydrodynamics code Piecewise Parabolic Method.
    - Maximum - 4500 processors
    - Stretch - 6561 processors
    - Maximum - 9000 processors
- UMT2000
  - 3D, deterministic, multigroup, photon transport code for unstructured meshes.
    - Standard - 3200 processors
    - Maximum - 4500 processors

Selected Applications/Benchmarks



Advanced Simulation & Computing™

Short Detour Before Presenting 7X Results

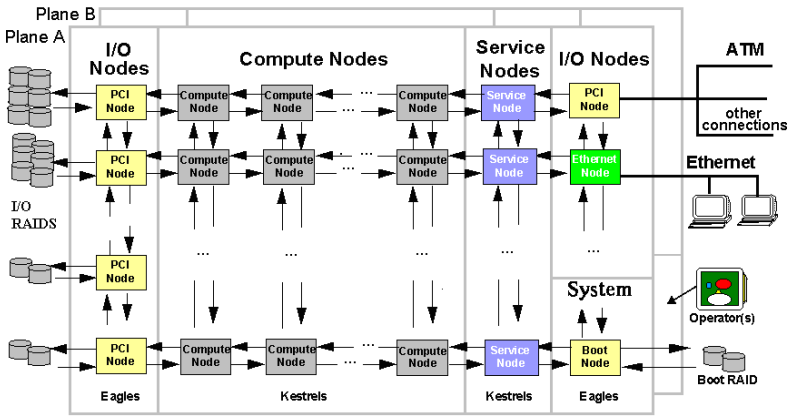
System Specifications
(ASCI Red vs. Red Storm)

NNSA National Nuclear Security Administration

Sandia National Laboratories

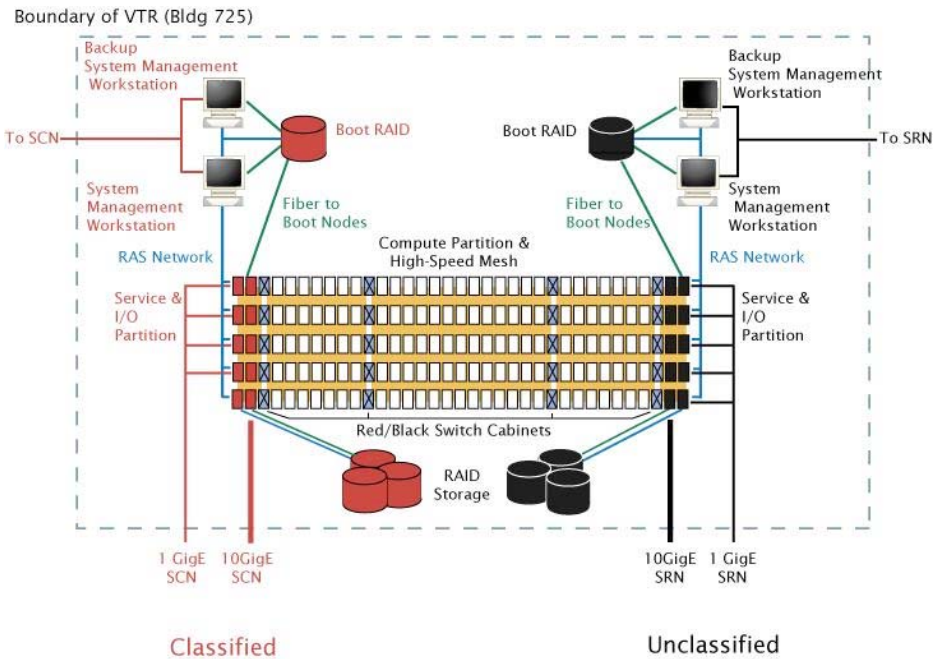
ASCI Red vs. Red Storm

ASCI Red



1168 nodes on the unclassified side and 1166 nodes on the classified side. The middle section contains 2176 nodes. Total number of compute nodes at 4510. Each compute node contains 2 processors, bringing total processor count to 9020.

Red Storm



5th row added bringing node count to 3360 on the unclassified side and 3360 on the classified side. The middle section contains 6240 nodes. Total number of compute nodes at 12960. Each compute node upgraded to dual core technology, bringing total processor count to 25920.

ASCI Red vs. Red Storm

	ASCI Red	Red Storm
Compute Nodes (Red/Center/Black)	4510 (1166/2176/1168)	12960 (3360/6240/3360)
Compute Processors (Red/Center/Black)	9020 (2332/4352/2336) PII Xeon 333Mhz	25920 (6720/12480/6720) Opteron Dual Core 2.4Ghz
Service Nodes (Red/Black) Disk I/O Nodes (Red/Black)	52 (26/26) 73 (37/36)	640 (320/320) Service and I/O partition (login, service, I/O, administrative nodes)
System Nodes (Red/Black)	2 (1/1)	RAS and System Management partition
Network Nodes (Red/Black)	12 (6/6) Ethernet ATM	100 (50/50) 10GigE to RoSE 20 (10/10) 1GigE to login nodes
Number of Cabinets	96 (76 compute/20 disk)	155 (135 compute/20 service and I/O)
Interconnect Topology	3-D Mesh (x,y,z) 38 x 32 x 2	3-D Mesh (x,y,z) 27 x 20 x 24

ASCI Red vs. Red Storm

	ASCI Red	Red Storm
Architecture	Dist. Memory MIMD	Dist. Memory MIMD
Theoretical Peak Performance	3.15 TF	124.42 TF
MP-Linpack Performance	2.38 TF	101.4 TF (2006) 102.2 TF (2007)
Total Memory	1.21 TB	39.19 TB
System Memory B/W	2.5 TB/s	78.12 TB/s
Disk Storage (Total / per Color)	12.5 TB / 6.25 TB	340 TB/170TB
Parallel File System B/W (Total / per Color)	2.0 GB/s / 1.0 GB/s	100 GB/s / 50 GB/s sustained disk transfer rate
External Network B/W (Total / per Color)	0.4 GB/s / 0.2 GB/s	50 GB/s / 25 GB/s sustained network transfer rate to RoSE

ASCI Red vs. Red Storm

	ASCI Red	Red Storm
Interconnect Bandwidth		
MPI Latency	**15 *u*s 1 hop, 20 *u*s max**	**~4.78 *u*s 1 hop, ~7.78 *u*s max**
Bi-Directional Link B/W	**800 MB/s**	**9.6 GB/s**
Minimum Bi-Section B/W	**51.2 GB/s**	**4.61 TB/s**
Full System RAS		
RAS Network	**10 Mb Ethernet**	**100 Mb and 1 Gb Ethernet**
RAS Processors	**1 for each 32 CPUs**	**1 for each 4 CPUs**
Operating System		
Compute Nodes	**Cougar**	**Catamount Virtual Node**
Service and I/O Nodes	**TOS (OSFI)**	**Linux**
RAS Nodes	**VX-Works**	**Linux**
Red Black Switching		
Switches	**2/row**	**4/row**

ASCI Red

- First computer in the ASCI program, built by Intel and installed at Sandia in late 1996. The design was based on the Intel Paragon computer.
- December, 1996, ASCI Red was measured at a world record 1.06 TF on MP LINPACK and held the record for fastest supercomputer in the world for several years, maxing out at 2.38 TF.
- ASCI Red decommissioned 2006, shortly after completing 7X runs.
- Distributed memory MIMD. The design provided high degrees of scalability for I/O, memory, compute nodes, storage capacity, and communications; standard parallel interfaces also made it possible to port parallel applications to the machine.
- Machine was structured into four partitions: Compute, Service, I/O, and System.

ASCI Red (cont.)

- ASCI Red used two operating systems, the Teraflops Operating System on the Service, I/O, and System Partition, and a Sandia developed lightweight kernel (Cougar) on the Compute nodes
- The Teraflops Operating System was Intel's distributed version of UNIX. It was a full-featured version of UNIX, used for boot and configuration support, system administration, user logins, user commands/services, and development tools.
- The operating system in the Compute Partition was Cougar, a very efficient and high-performance operating system providing program loading, memory management, message-passing support, some signal handling and exit handling, and run-time support for the supported languages.
- This combination of operating systems made it possible to specialize for specific tasks and standard programming tools to make the supercomputer both familiar to the user and non-intrusive for the scalable application. The machine provided a single system image to the user.

ASCI Red (cont.)

- In normal operation, disconnect cabinets divided ASCI Red into two sides; unclassified and classified. In this situation, each side appeared as a separate plane in the mesh topology.
- The configuration was 1168 compute nodes on the unclassified end and 1166 compute nodes on the classified end. The middle section consisted entirely of 2176 computational nodes, and could be switched from the unclassified end to the classified end and back again. Total number of compute nodes was 4510.
- While message passing was used between nodes, shared memory mechanisms were used to exploit parallelism on a node. Each compute node had two processors. The 7X testing was performed on ASCI Red in Proc 0 and Proc 3 modes only:
  - Proc 0 option with yod – ignore the second processor – default mode – entire system RAM on the node is available to the application (256 MB)
  - Proc 3 option with yod – this mode treats each processor as a separate compute node – virtual mode – the processors share memory so only half the system RAM is available to the application (128 MB)

Red Storm

- Red Storm, the follow-on computer to ASCI Red, was developed jointly by Cray and Sandia.
- Manufactured by Cray and installed at Sandia in early 2005 as an XT3 system.
- The architecture of Red Storm follows the model of ASCI Red: allows simultaneous usage on the unclassified (black) and classified (red) sides of the machine. Users interface to the system via a Linux operating system, and the compute nodes run a lightweight kernel.
- 2005, Red Storm was measured at 36 TF on MP LINPACK.
- 2006, Red Storm was measured at 101.4 TF on MP LINPACK.
- 2007, Red Storm was measured at 102.2 TF on MP LINPACK.
- Distributed memory MIMD. Combines commodity and open source components with custom-designed components to create a system that can operate efficiently at immense scale.

Red Storm (cont.)

- The basic scalable component is the node. There are two types of nodes. Compute nodes run user applications. Service nodes provide support functions, such as managing the user's environment, handling I/O, and booting the system. Each compute node and service node is a logical grouping of a processor, memory, and a data routing resource.
- Cray XT3 systems use a simple memory model – for applications distributed across numerous nodes, each instance of the application has its own processor and local memory – remote memory is the memory on the nodes running the associated application instance – there is no shared memory.
- The *system interconnection network* is the data-routing resource that Cray XT3 systems use to maintain high communication rates as the number of nodes increases. The system interconnection network enables the system to achieve an appropriate balance between processor speed and interconnection bandwidth.

Red Storm (cont.)

- In normal operation, disconnect cabinets divide Red Storm into two sides; unclassified and classified.
- 5th row added August-October 2006 bringing node count to 3360 on the unclassified side and 3360 on the classified side. The middle section contains 6240 nodes and can be switched from the unclassified end to the classified end and back again. Total number of compute nodes at 12960. Each compute node was upgraded to dual core topology, bringing total processor count to 25920.
- The 7X testing was performed on Red Storm in SN and VN mode:
  - SN option – ignore the second processor – default mode – entire system RAM on the node is available to the application.
  - VN option – each processor is a separate compute node – only half the system RAM on the node is available to each processor.

Red Storm – System Software

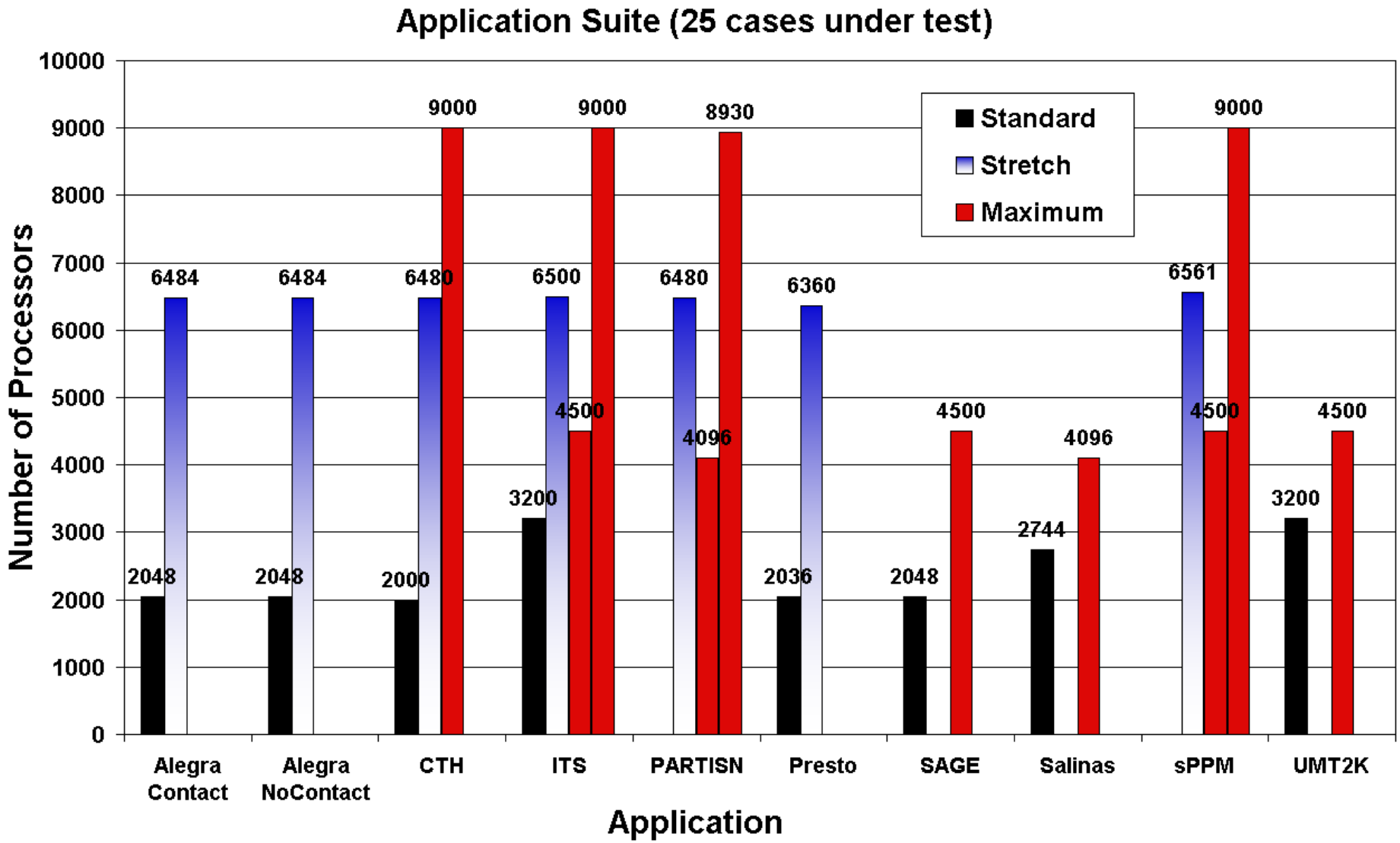
- Operating Systems
  - Linux on service and I/O nodes (SuSE Enterprise Server)
  - Catamount VN lightweight kernel on compute nodes
  - Linux on RAS monitors
- Run-Time System
  - Logarithmic job launch (yod)
  - Node allocator (CPA)
  - Batch system – MOAB
- File Systems
  - High performance file system (lustre)
- User Environment
  - PGI compilers – Fortran, C, C++
  - Libraries – MPI, I/O, Math, MPI-2
  - Showmesh
  - Debugger – Totalview
  - Performance Monitor
- Network
  - 50 x 10 GigE to RoSE, WAN
  - 10 x 1 GigE to login nodes
  - 1 GigE to Mgmt stations
- System Mgmt and Admin
  - Accounting
  - Red Storm Management System

Red Storm – Lightweight Kernel

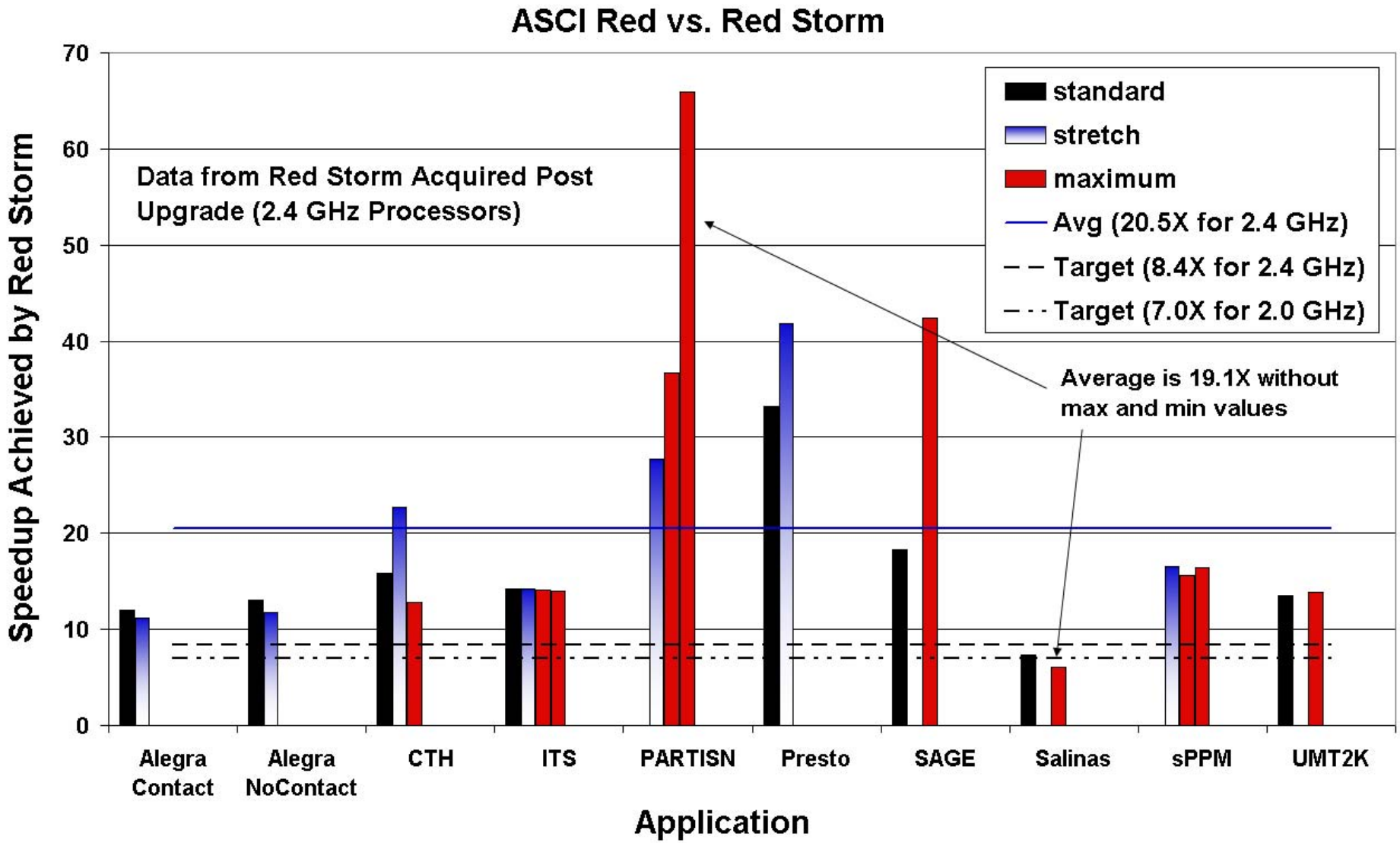
- Lightweight compute node OS is fundamental to the Sandia architecture. It is essential for:
  - Maximizing CPU resources
    - Reduce OS and runtime system overhead
  - Maximizing memory resources
    - Small memory footprint, large page support
  - Maximizing network resources
    - No virtual memory, physically contiguous address mapping
  - Increasing reliability
    - Small code base, reduced complexity
  - Deterministic performance
    - Repeatability
  - Scalability
    - OS resources must be independent of job size

How Much Faster is Red Storm on the 7X Applications?

Just a Reminder of the 7X Suite…



Red Storm is ~20X Faster

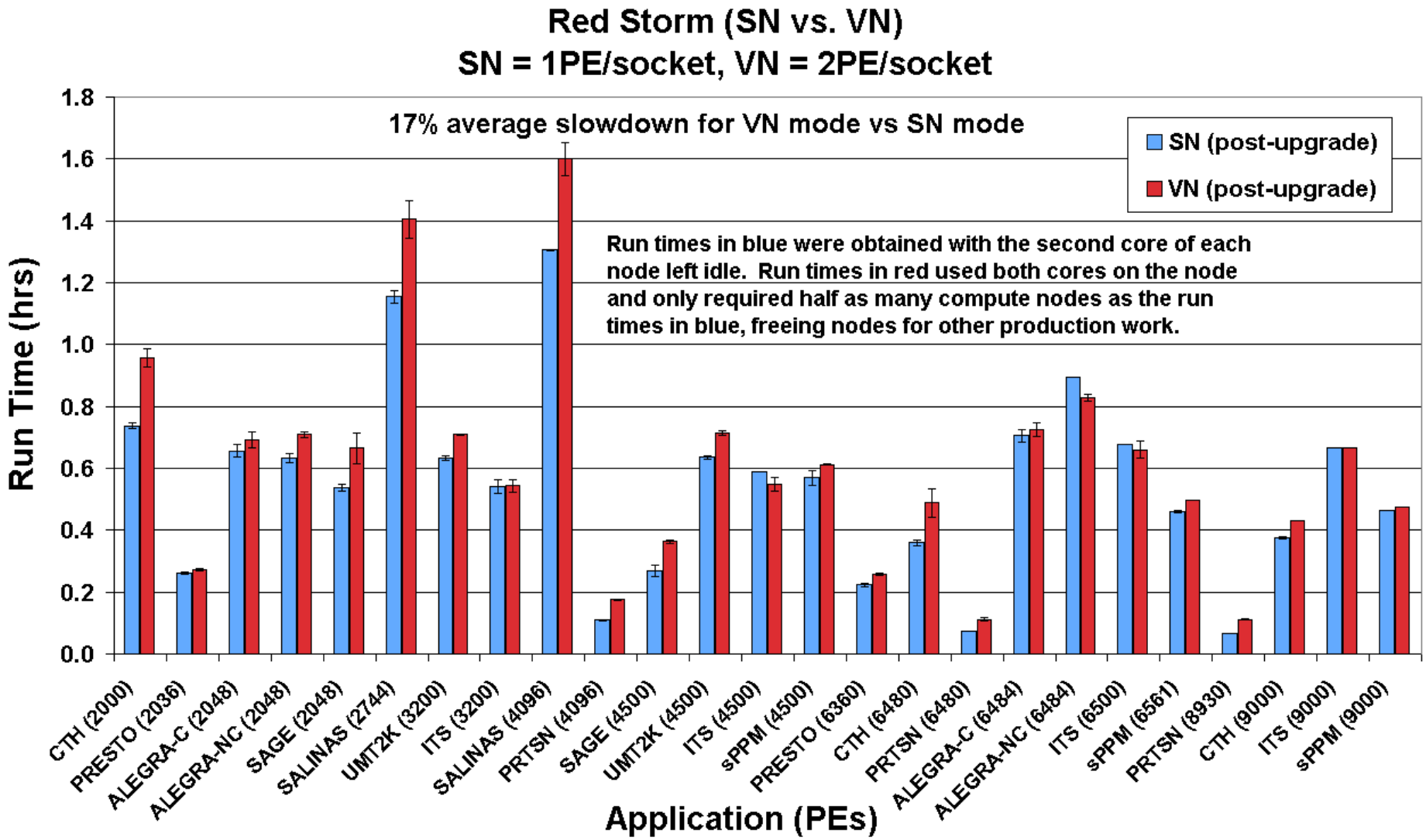


How do the 7X VN and SN Results Compare on Red Storm?

SN vs. VN

- The recent upgrade of Red Storm to dual core sockets has provided the option of specifying either one or two cores (processors) per socket when launching an application.
- If applications can run efficiently in VN mode, this would free up sockets for other applications.
- 7X application testing performed on Red Storm in SN and VN mode:
  - SN option – ignore the second processor – default mode – entire system RAM on the node is available to the application.
  - VN option – each processor is a separate compute node – only half the system RAM on the node is available to each processor.
- Compare the results in terms of execution time.

VN Performs Well on 7X Applications



Summary

- 7X: ASCI Red vs. Red Storm
  - The 7X test suite consisted of 10 applications and benchmarks that were used in ASCI Red vs. Red Storm performance testing.
  - One or more problems were identified for each application, and those problems were run at two or three processor sizes, comparing the results between ASCI Red and Red Storm – 25 cases studied.
  - Red Storm has achieved its requirement of 7X performance over ASCI Red, posting an average speed-up of ~20X.
- 7X: SN vs. VN Results on Red Storm
  - VN performed well on the 7X applications - average efficiency drop was 17%.
- Future Work
  - Impending quad core upgrade on Red Storm (summer 2008).
  - Upgrade will provide another opportunity to demonstrate the usefulness of the 7X suite to track performance across single, dual and quad core processors.

Acknowledgments

- The authors thank Courtenay Vaughan, Bob Benner, John Van Dyke, Sue Goudy, Mahesh Rajan, and Hal Meyer for their assistance with compiling, configuring, and troubleshooting on ASCI Red and Red Storm. Many thanks also to the ASCI Red (Frank Jaramillo, Paul Sanchez, Mike Martinez, Sean Taylor) and Red Storm system administrators and support staff for their assistance. Thanks also to Mark Hamilton for assistance in setting up the Sourceforge repository.
- The authors thank Cray engineers Paul Burkhardt, Doug Enright, and Ron Pfaff for their assistance with compiling and optimizing the application codes for Red Storm benchmark runs.
- Sue Goudy, Sue Kelly, Mike McGlaun, Jim Tomkins, and Courtenay Vaughan have all provided help, suggestions, and guidance as the predecessor to this report was assembled (The 7X Cookbook).
- We also thank the application code developers for their assistance: Brian Franke (ITS), Garth Reese (Salinas), Riley Wilson (Salinas), Galen Gizler (SAGE), John Daly (SAGE), Kevin Brown (Presto), Arne Gullerud (Presto), Allen Robinson (Alegra), Rich Drake (Alegra), and Josh Robbins (Alegra).