SAND2008-6293C

Red Storm: The Birth of a New Supercomputer

James Tomkins
Sandia National Laboratories

Cray Technical Workshop
University of Edinburgh, Edinburgh, Scotland
September 24-26, 2008

National Nuclear Security Administration

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

Advanced Simulation & Computing™

Sandia Perspective

National Nuclear Security Administration

Sandia National Laboratories

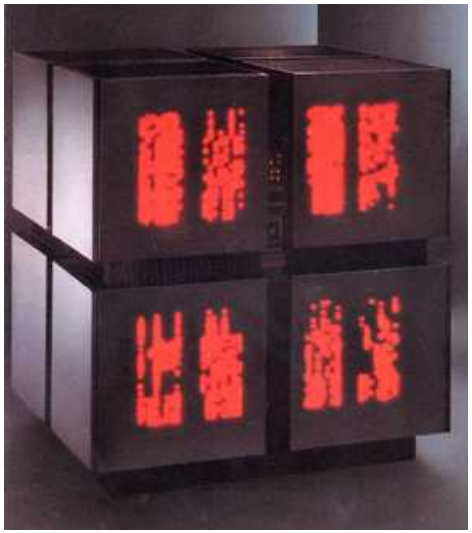
A Turning Point for Sandia

- Prior to 1987 Sandia was a follower in High Performance Computing
- In 1987 Sandia embarked on a path to leadership in High Performance Computing (HPC) through Massively Parallel Processing (MPP)

Sandia HPC History

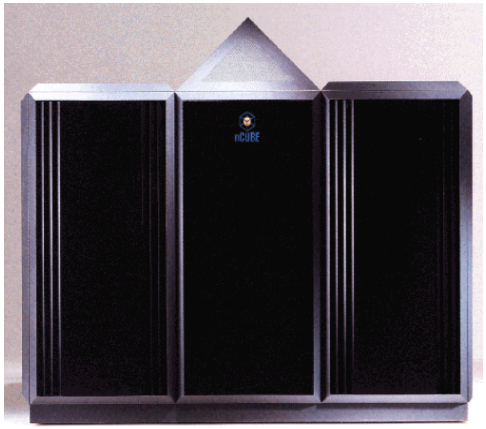
- 1987-- Sandia fielded first true MPP, 1024 node nCUBE 10
  - Won Karp Challenge
  - Won inaugural Gordon-Bell Prize
  - Awarded patents for nearly every aspect of MPP software

- 1988-- Sandia fielded a CM2 with 16K 1 bit processors and 512 32 bit floating point units
  - SIMD architecture
  - Later upgraded to CM-200



Sandia HPC History

- 1990-- Sandia fielded 2 1024-node nCUBE-2's, First true MPP supercomputers
  - Outperformed Cray vector computers @ ~1/7 the cost!
  - Sandia began research on Light Weight Kernel (LWK) operating systems
- 1993-- Sandia fielded first ~1850-node (~3900 processors) Intel Paragon
  - Sandia (Intel) Paragon is #1 on Top 500 list (first for Sandia)
  - Wins Sandia's second Gordon-Bell Prize
  - First use of Sandia developed Light Weight Kernel (LWK) Operating System software for production computing



Sandia HPC History

- 1997-- Sandia fielded Intel Tflops, world's first Terascale computer, 4600+ nodes (9200+ processors)
  - Ran Sandia developed LWK System Software
  - Number 1 on the "Top 500" list for 7 consecutive lists from June, 1997 through June, 2000, a record still unmatched!
- 1997-- Sandia began development of world's first Linux "Super-cluster", Cplant^{tm}
  - Sandia integrated DEC/Compaq HW with Myrinet network
  - Sandia developed all run-time, file system and messaging software
  - World's first terascale cluster-- Cplant grew to become the world's first terascale cluster and achieved ~1 TF on Linpack in 2003.

Advanced Simulation & Computing™

The Project

National Nuclear Security Administration

Sandia National Laboratories

Build on Sandia's Past Success

- Since 1987, Sandia has been a world leader in the successful use of large scale parallel computing. We have been able to achieve good parallel performance for nearly every major application type used by DOE across its missions.
- Sandia has received numerous national and international awards for work in HPC. These include 10 R&D100 Awards.
- Sandia has been granted over 40 patents in HPC.
- Sandia has led the Top-500 list more than any other institution.

Getting Started

- Getting Sandia Upper Management Support
  - Acceptance of Idea to Build a Custom Supercompter
  - Support for the Money - ~$90M project
  - Resource commitment - People and Facilities
  - Understanding Risk/Reward - What is Success
- Getting DOE ASC Support - Includes all of the Above
- Developing the detailed architecture requirements
- The Name - Red Storm

High Level System Drivers

- Capability Computing - The ability to efficiently run a broad set of complex engineering and scientific applications across the whole machine.
- Provide capability computing for both classified and unclassified work in a single flexible system.
- Provide a significant increase in computational power over existing resources without modifying the application codes.
- ASC Workload

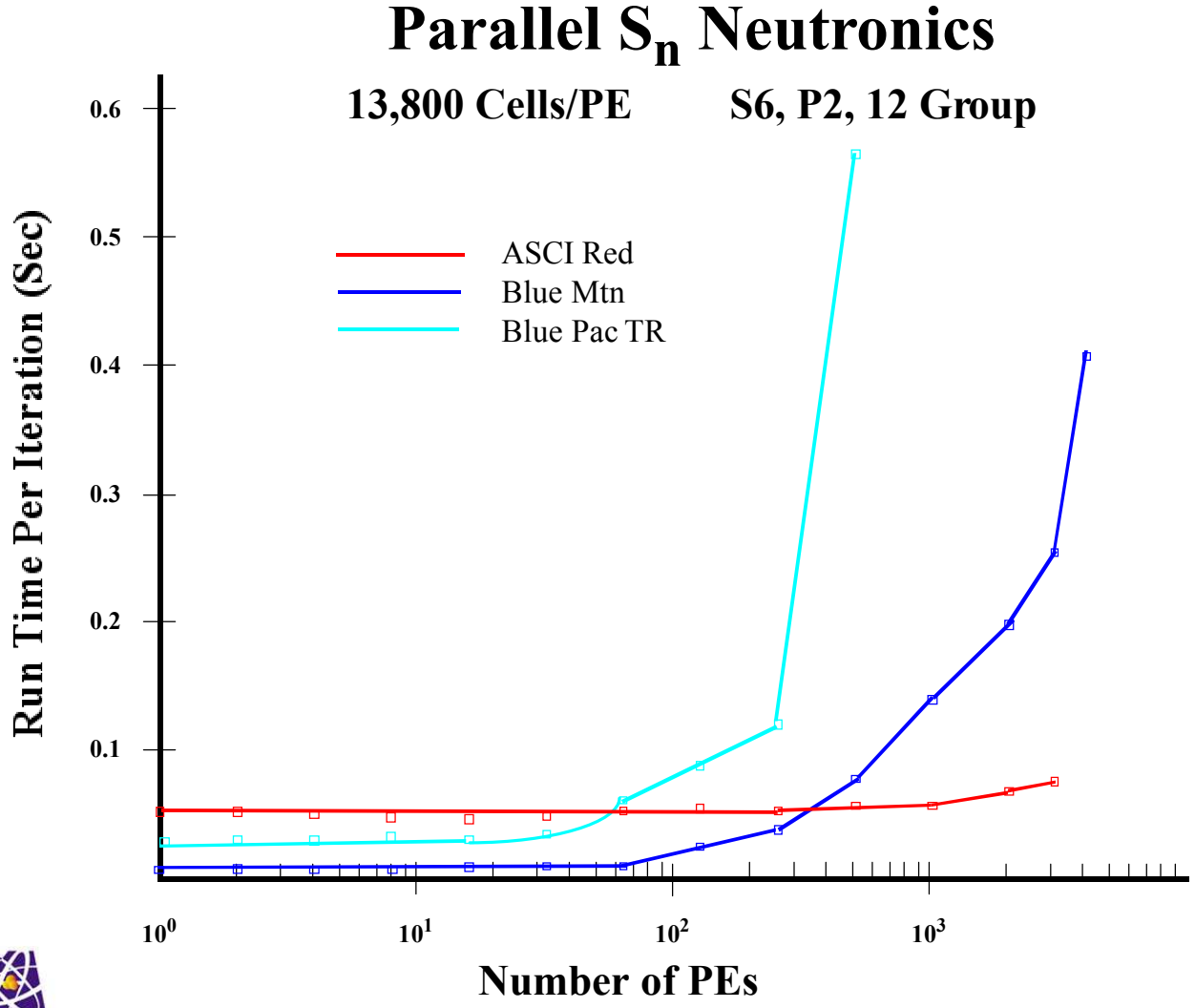
Architectural Goals

- Balanced System Performance: CPU, Memory, Interconnect and I/O
- **Scalability:** System Hardware and System Software scale, a single cabinet system to 32K node system
- Functional Partitioning: Hardware and System Software
- Reliability: Full system Reliability, Availability, Serviceability (RAS) designed into Architecture
- **Upgrade-ability:** Designed in path for system upgrade
- **Red/Black Switching**: Flexible support for both classified and unclassified Computing in a single system
- Custom Packaging: High density, relatively low power system
- Price/Performance: Excellent performance per dollar, use high volume commodity parts where feasible

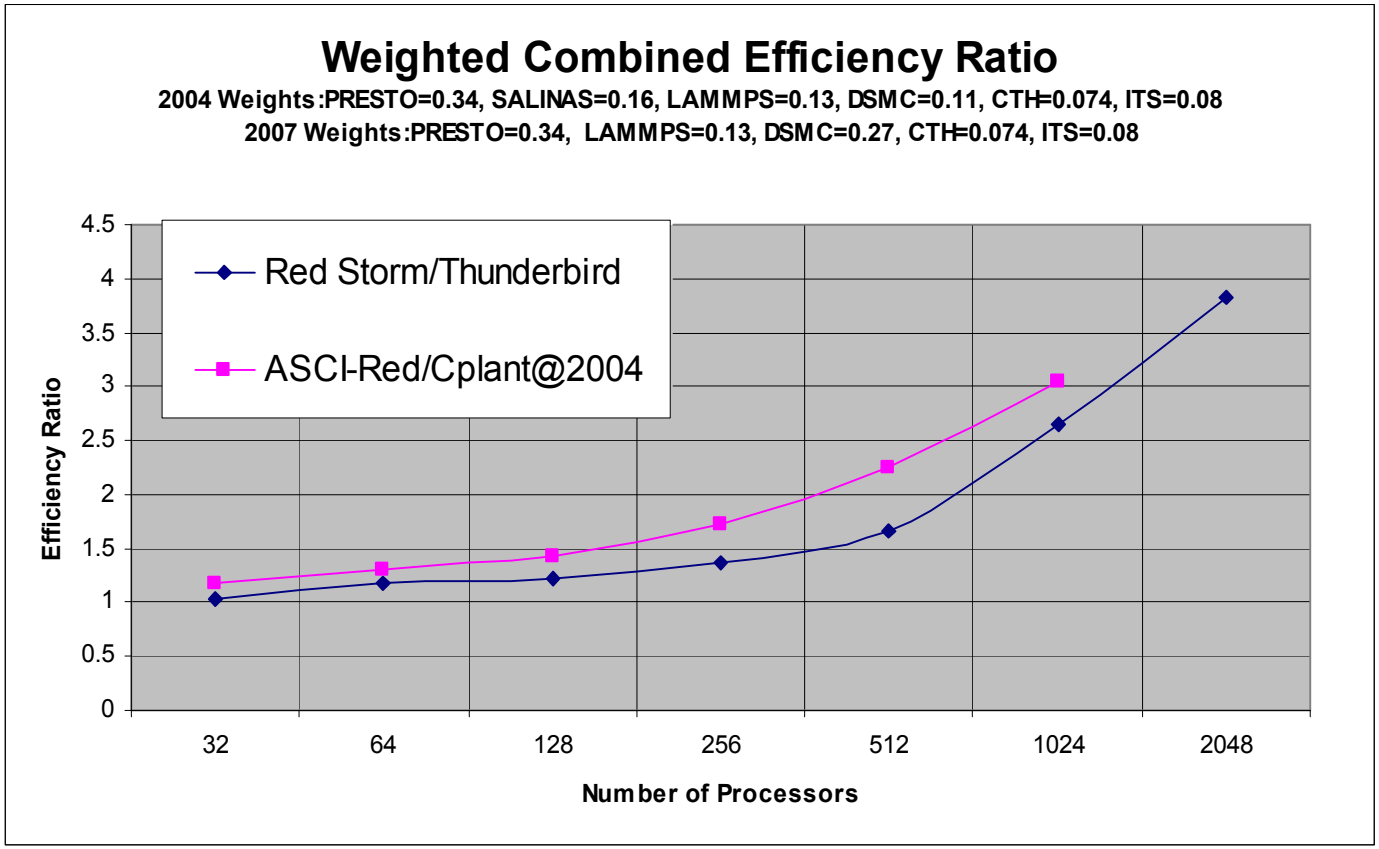
Steps Needed to Begin Procurement Process

- Develop Detailed Architectural and System Requirements - The Statement of Work (2001)
- Convince Potential Partners that it is a good idea
  - Companies want to sell what they have and are not always receptive to other ideas
  - How to ensure Competition
- How to Manage Risk
  - Financial
  - Schedule
  - Performance

Why Build a Custom Machine



Workload 'percentage weighted' parallel efficiency ratio for Red Storm/Thunderbird and ASCI-Red/Cplant

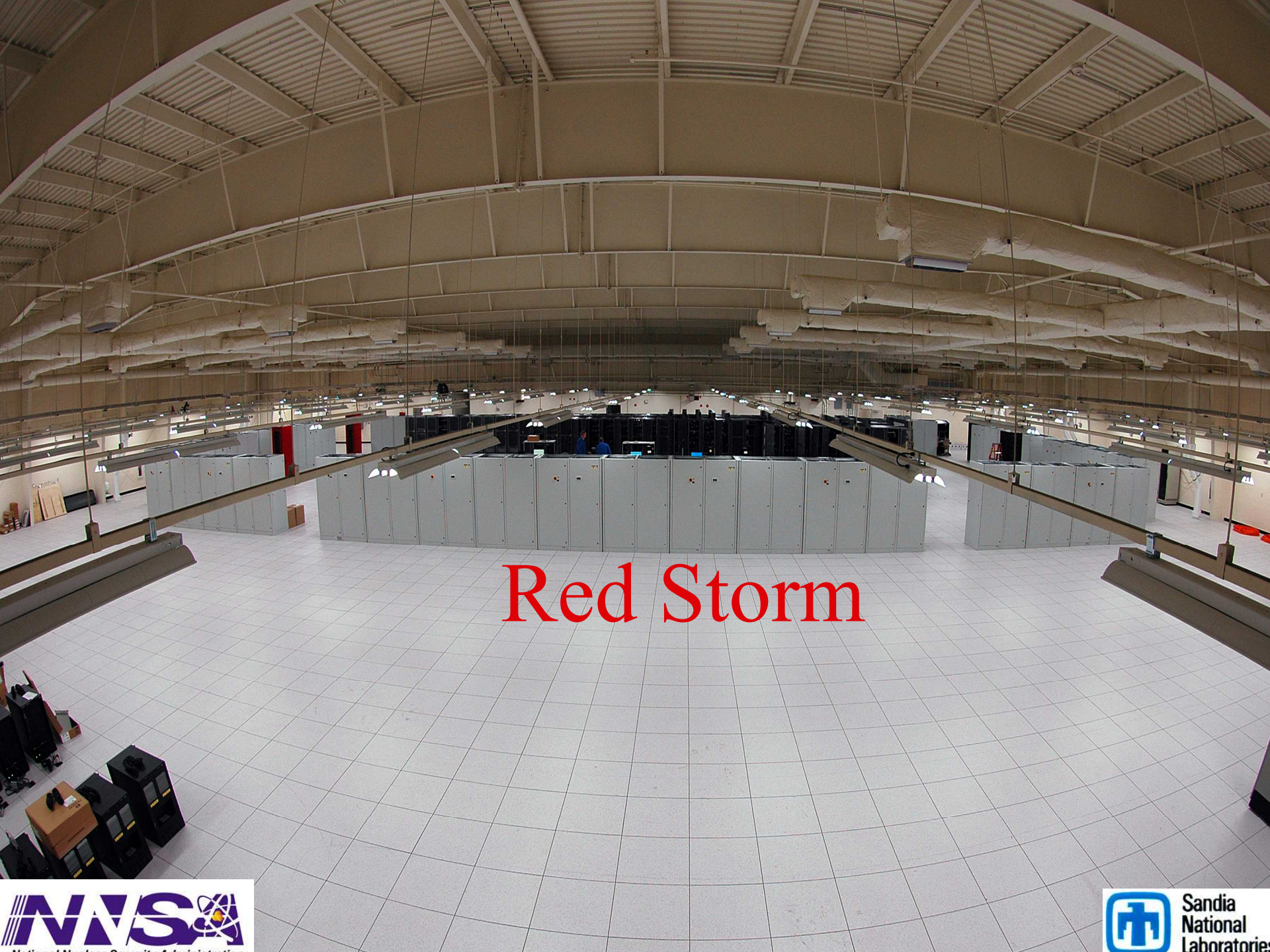


Creating a Partnership
Trials and Tribulations

- Choosing Cray - June 2002
- Approval to go ahead - DOE ASC, Sandia CFO
- The Letter Contract - June 2002
- The Full Contract - Sept 2002
  - The SOW
  - Commercialization
- System Installation - Early 2005
  - Hardware
  - System Software

The Partnership

- Sandia Architecture
- Cray Built Hardware
- Sandia and Cray Developed System Software
- Team Work



NNSA
National Nuclear Security Administration

Sandia National Laboratories

The Outcome

- Red Storm and the Cray XT product line is the outcome of the Sandia/Cray partnership:
  - Sandia Architecture
  - Sandia and Cray System Software
  - Cray Engineering and Manufacturing
  - Sandia Systems HW/SW Expertise

Red Storm Path

- Original Machine (Early 2005)
  - ~41 TF, 10,368 Single-Core Compute Nodes
- Added Fifth Row and Dual-Cores (Fall 2006)
  - ~125 TF, 12,960 Dual-Core Compute Nodes
- Disk Storage Upgrade (Summer 2008)
  - 2.36 PB of New Disk Storage
- Quad-Core and Memory Upgrade(Summer 2008)
  - ~284 TF, 6,240 Quad-Core Compute Nodes and 6,720 Dual-Core Compute Nodes - 2 GB of Memory per Core

Red-Black Switching

- Red Storm was designed from the start to support red/black switching
- Switchable components have no persistent state that can be written from the Classified side

• Switch Cabinets isolate sections of the machine, so that the service sections remain in production at all times

• Network infrastructure is balanced between black and red

- Design based on ASCI Red
  - Familiar to NNSA
  - Changeover O(1hr)
- RAS network is separable
- Advantages
  - Load balancing of work
  - Flexibility during upgrades and maintenance

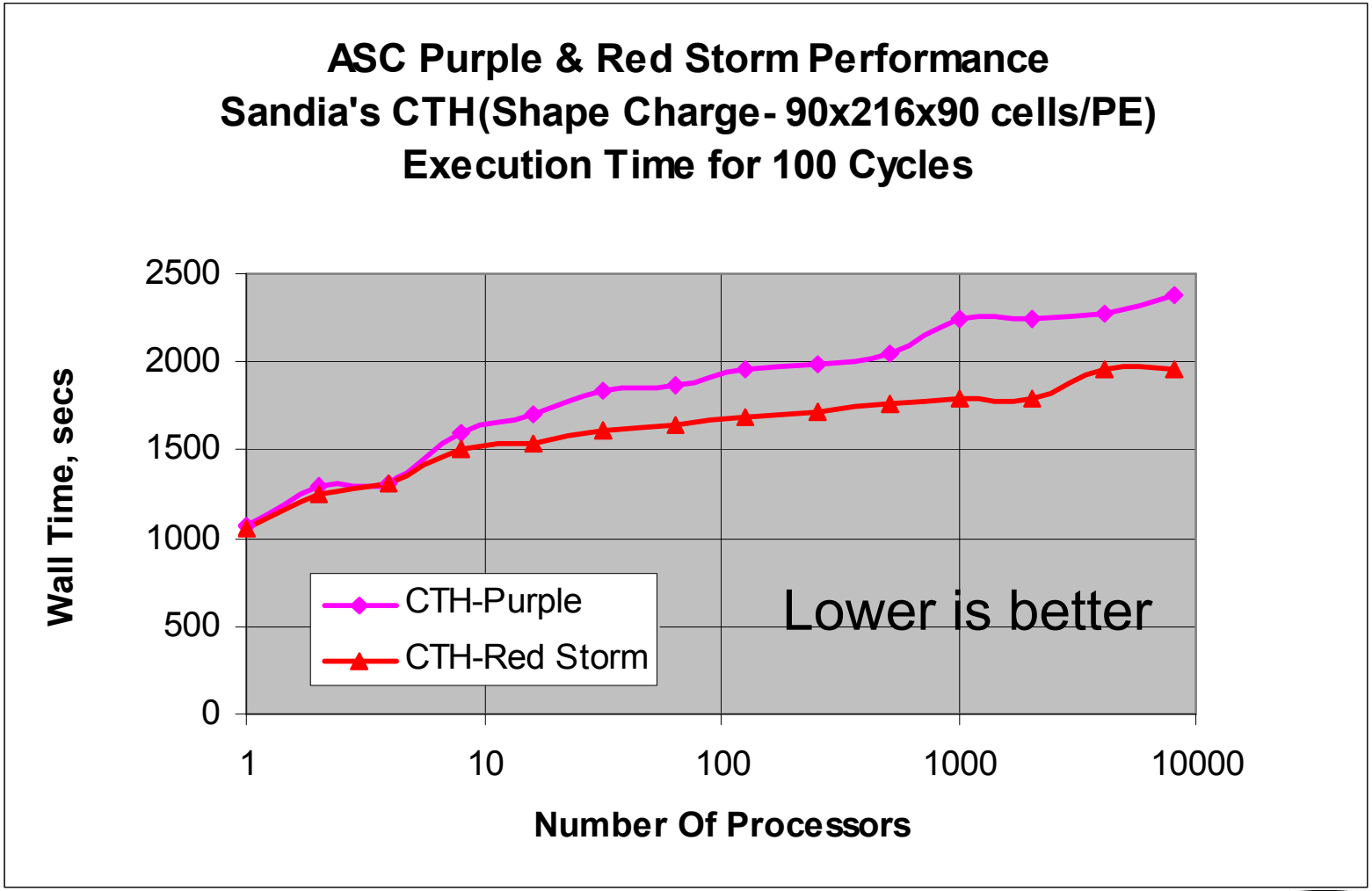
Unique Aspects of Quad-Core Upgrade

- Heterogeneous Machine
  - 70 Cabinets of Dual-Core
  - 65 Cabinets of Quad-core
- Operating System Support
  - Applications must be able to run on mix of nodes transparently
  - Required New Version of Catamount - CNW
    - 4 way SMP
    - Job scheduling by number of cores
    - CNL does not support mixed operation

Red Storm Performance

- Parallel efficiency exceeds Blue Gene and Purple on all applications studied:
  - –Sandia Nuclear Weapons engineering codes
  - –Climate-- atmosphere and ocean
  - –Molecular Dynamics
  - –LANL Nuclear Weapons physics codes

CTH - Shock Physics
Original Red Storm



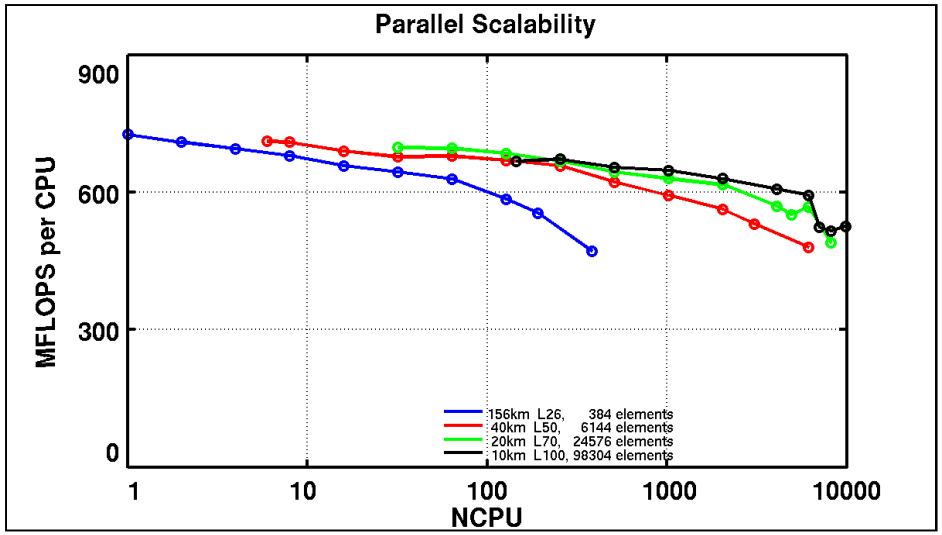
National Nuclear Security Administration

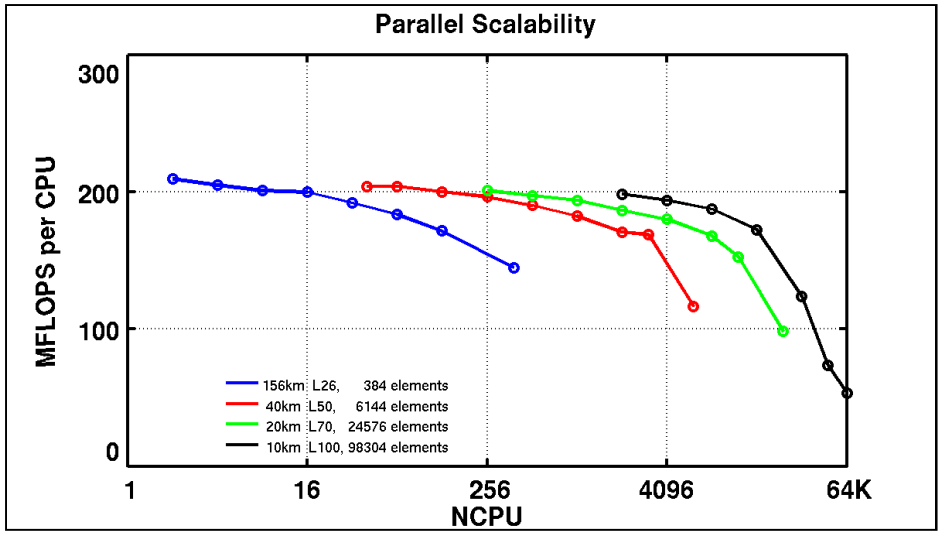
Sandia National Laboratories

SEAM Atmospheric Code
(Strong Scaling)

Red Storm – 5 TF max
Red Storm Peak: 41 TF

BG/L – 4 TF max
BG/L Peak: 360 TF



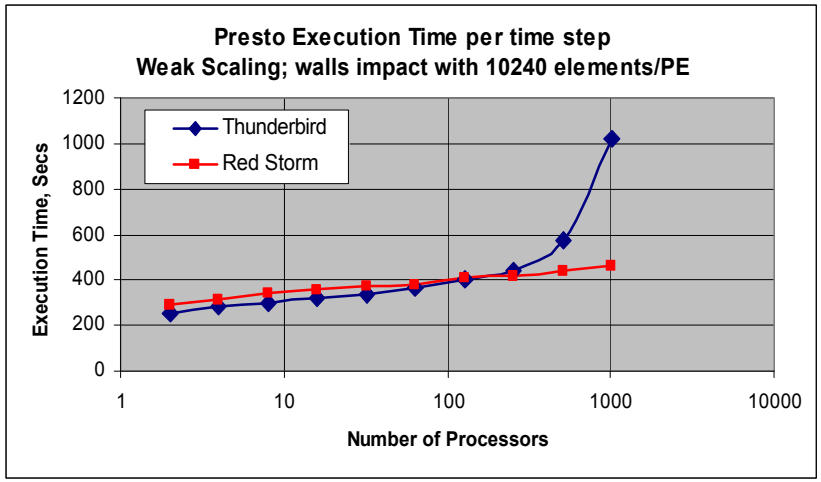


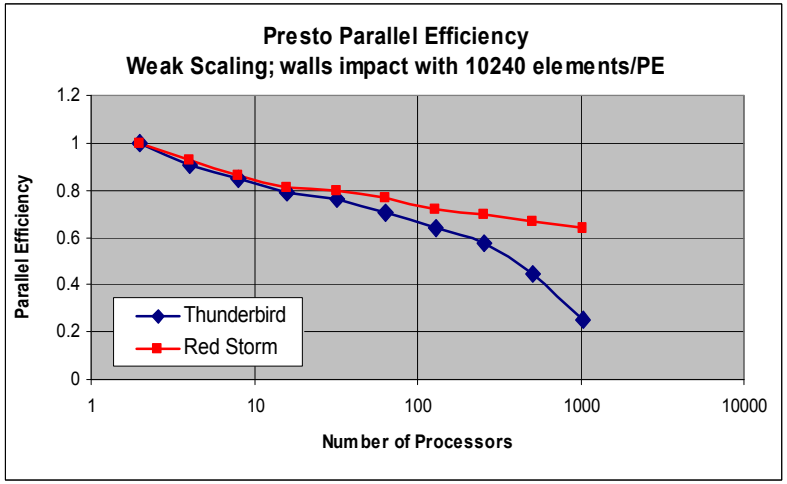
SEAM = NCAR's Spectral Element Atmospheric Model, POP = LANL's Parallel Ocean Program

National Nuclear Security Administration

Sandia National Laboratories

Red Storm Comparison to Thunderbird SIERRA/Presto Crash Dynamics





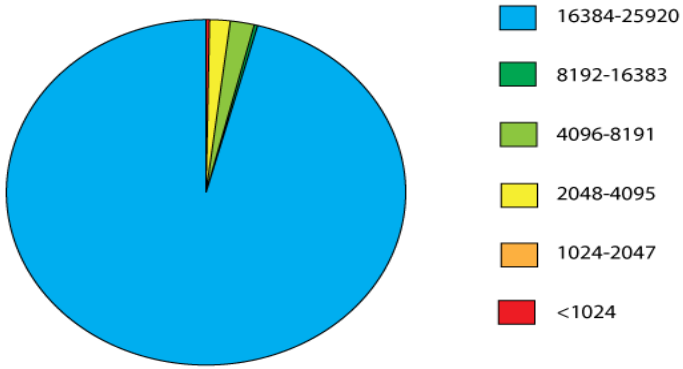
- Explicit 'crash' Lagrangian transient dynamics
- Model: Two sets of brick-walls colliding
- Weak scaling analysis with 80 bricks/PE, each discretized with 4x4x8 elements
- Contact algorithm communications dominates the run time
- The rapid increase in run time after 256 processors on Thunderbird is a consequence of the contact algorithm's sensitivity to latency

Early Work With Quad Cores

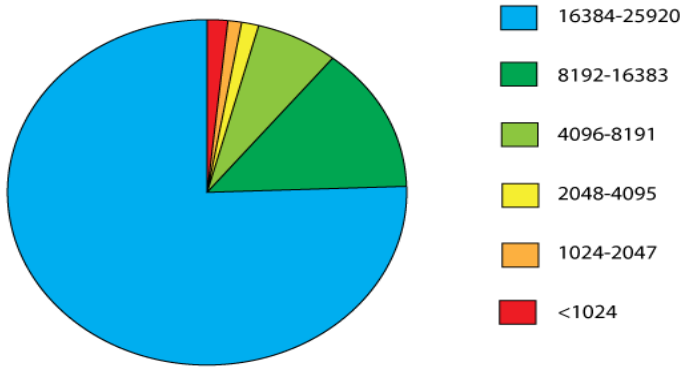
- We have done a lot of work looking at scaling on the AMD Quad-Core chips with a number of our codes. We have also done some work at scale with a limited number of applications. So far the results look good.
- CTH - Performance per core is slightly better for Quad-Core than Dual-Cores based on a comparison run on ~7800 cores of each. Looks good for mixed node runs.
- An LLNL application on 4096 cores was only slightly slower on Quad-Cores than Dual-Cores.
- LANL has had similar results.

Red Storm Capability Usage

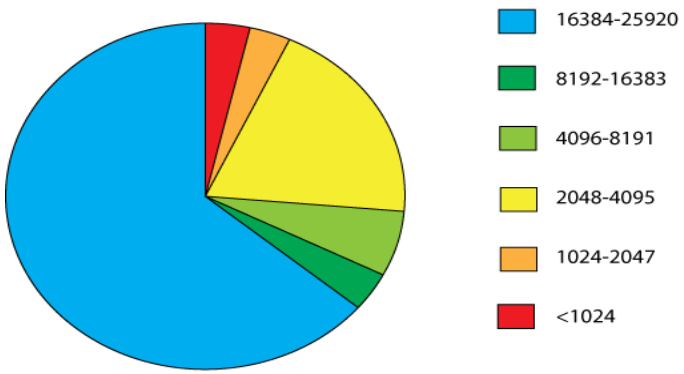
Red Storm Jumbo Mode, 12/8/07 – 3/11/08
System CPU Hours by Job Size



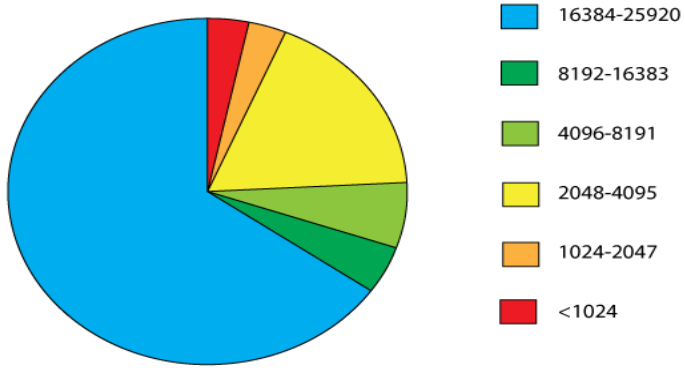
Classified
Opus Computation



Unclassified
All Computations



Classified
All Computations



All Computations

National Nuclear Security Administration

Sandia National Laboratories

Red Storm Project Impact

- Rebirth of the True MPP Supercomputer
- Cray's Signature Product
- A Major Step in the Path to PF Computing
- Important Work for National Security
- Cray and Sandia with the support of NNSA/ASC have created one of the most successful new supercomputers ever.

Additional Background

Catamount N-Way LWK OS

- Background
  - SUNMOS, research project on nCUBE2 that became production OS on Sandia Paragon Computer
  - PUMA, second generation LWK
  - Cougar, production version of PUMA on Sandia ASCI Red Computer
  - Catamount, production OS for initial Red Storm
  - Catamount VN, production OS for Dual-Core Red Storm
  - Catamount N-Way, production OS for heterogeneous (Dual-Core and Quad-Core) Red Storm

Quad Core Test System

- Cage system with 4 quad core nodes
  - XT4 cabinet
  - 2.2 GHz
  - 8 GB/node (2GB/core), DDR2 5300 DIMMs
  - Budapest nodes
- home directory is nfs mounted
  - no Lustre
- PGI 6.2.5 compilers with ACML 3.6.1
  - need 7.1 and 4.1 to utilize features of quad cores

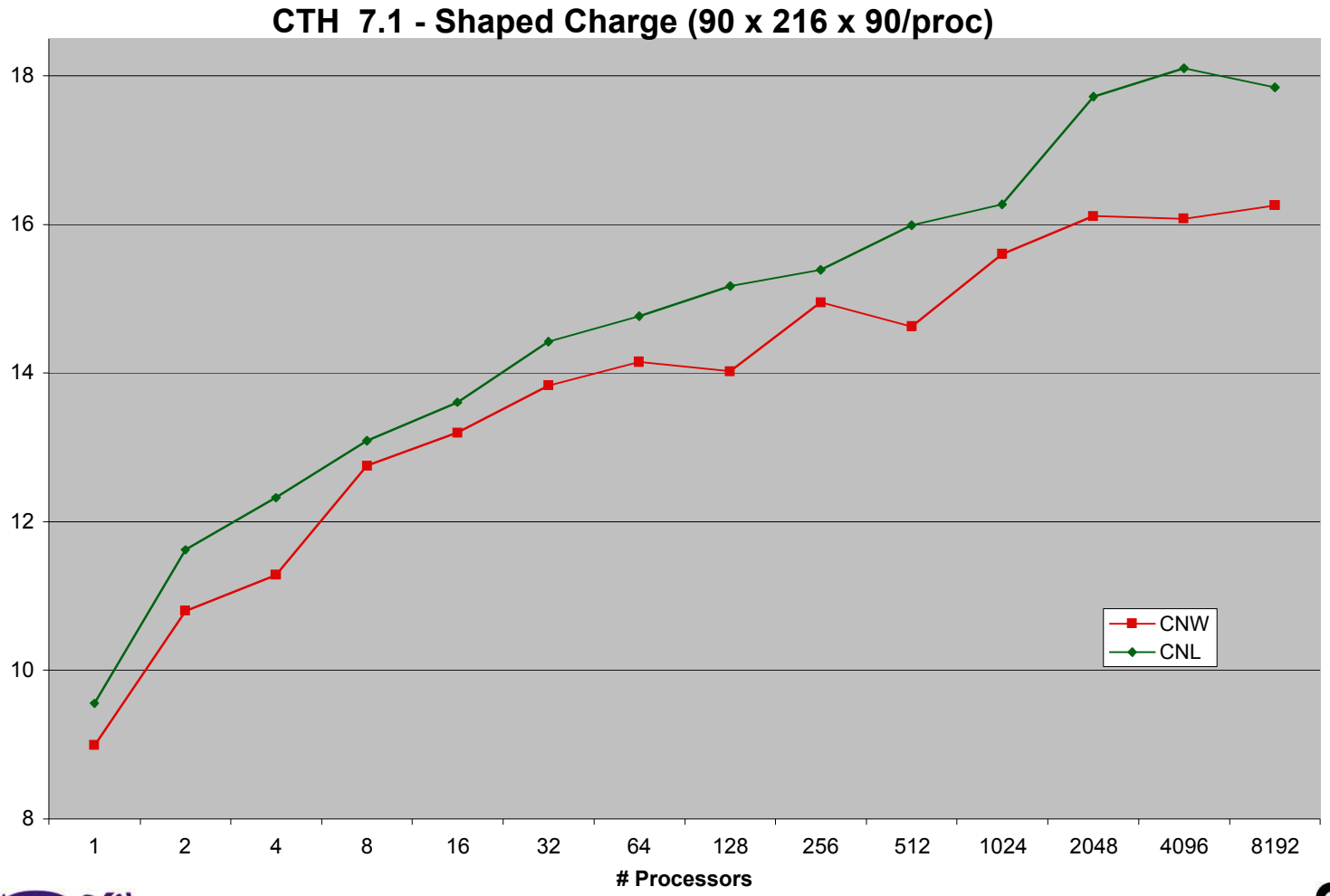
Catamount N-Way Testing
HPCC Quad-Core

HPCC Benchmark	Mumber of MPI Ranks	Cores per Node	CNL	CNW	CNL/ CNW
PTRANS	4	1	0.567	1.606	2.83
HPL	4	1	17.88	17.90	1.00
STREAMS	4	1	25.21	25.84	1.02
Random	4	1	0.0064	0.0118	1.83
FFT	4	1	1.609	1.646	1.02
PTRANS	4	2	0.488	1.551	3.18
HPL	4	2	17.78	18.03	1.01
STREAMS	4	2	16.45	18.11	1.10
Random	4	2	0.0061	0.0115	1.88
FFT	4	2	1.337	1.36	1.02

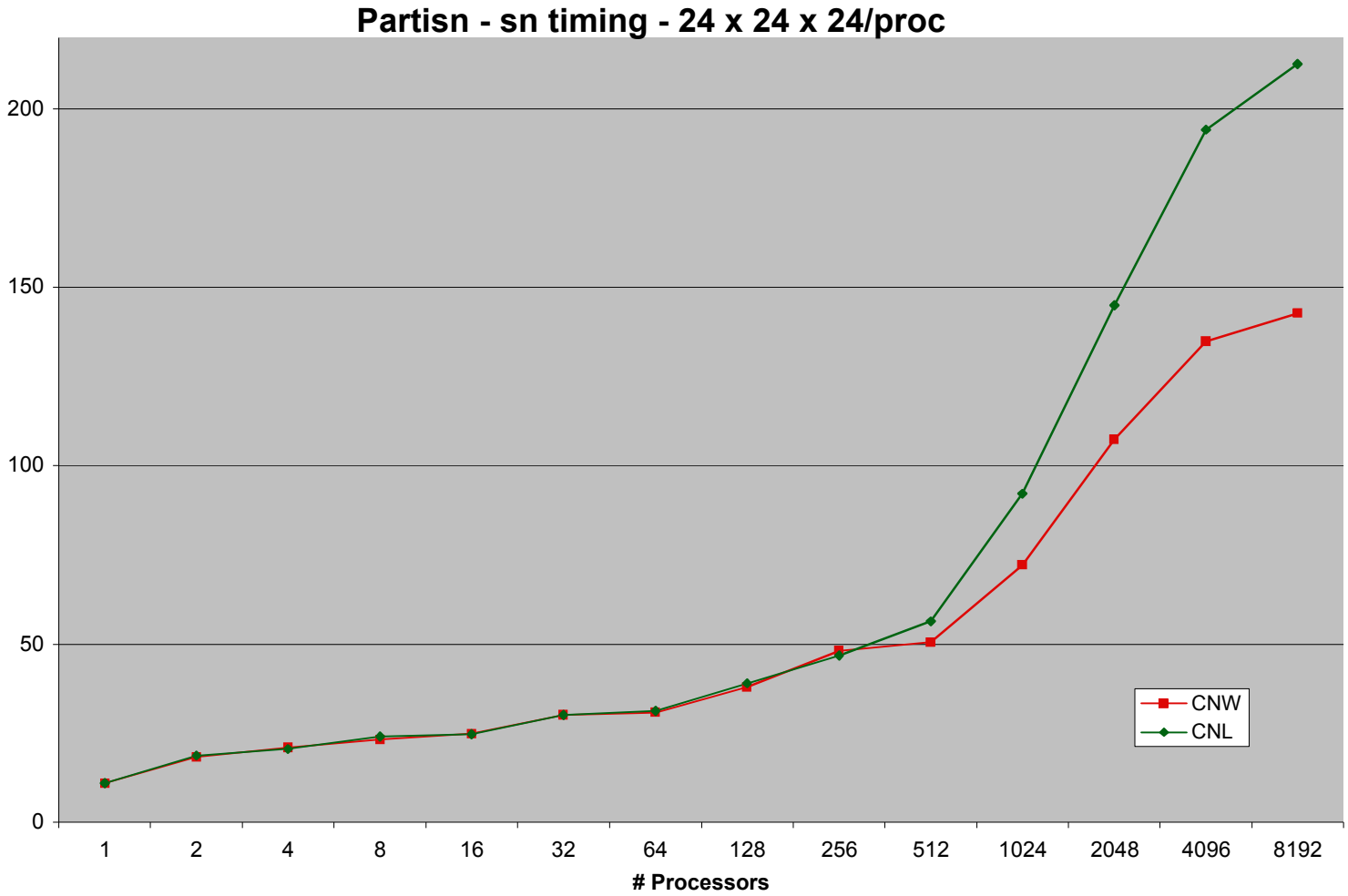
Catamount N-Way Testing
HPCC Quad-Core

HPCC Benchmark	Mumber of MPI Ranks	Cores per Node	CNL	CNW	CNL/ CNW
PTRANS	4	4	0.287	1.244	4.33
HPL	4	4	17.59	17.72	1.01
STREAMS	4	4	7.85	9.95	1.27
Random	4	4	0.0060	0.0115	1.92
FFT	4	4	0.902	0.959	1.06

Catamount N-Way Testing
Large Scale Dual-Core - CTH



Catamount N-Way Testing
Large Scale Dual-Core - Partisn



Quad-Core Testing
CTH

- Shock hydrodynamics
- Shaped charge problem with explosives and meso-scale problem with a large number of materials
- VN2 mode takes 10% to 13% more time than SN mode
  - Consistent with dual core results
- VN4 mode takes 39% to 52% more time than SN mode

time on 4 cores	SN mode	VN2 mode	VN4 mode
shaped 4	8.385	9.338	11.832
shaped 2	4.534	5.056	6.901
meso 2	12.312	13.863	18.023
meso 1	6.482	7.076	8.953

Quad-Core Testing
Sage

- hydrodynamics code
- timing_c problem with 250000 cells/PE
- Took 5.8% longer in VN2 mode than SN mode
  - Consistent with dual core results
- Took 26.3% longer with VN4 mode than SN mode

4 cores	SN mode	VN2 mode	VN4 mode
CC/sec/PE	6181.6	5823.6	4556.2

Quad-Core Testing
Partisn

- Time-Dependent, Parallel Neutral Particle Transport Code
- Reports a Transport and Diffusion Grind Time in nanoseconds
- Takes 28% to 46% longer in VN2 mode than in SN mode
- VN4 mode takes 2.2 to 2.8 times as long as SN mode

grind times 4 cores	SN mode	VN2 mode	VN4 mode
184 MB/core	32.56 / 28.23	47.55 / 39.45	90.78 / 74.07
367 MB/core	41.73 / 29.11	53.52 / 38.51	92.02 / 71.02
1444 MB/core	34.70 / 26.93	48.43 / 35.37	89.45 / 65.53

Quad-Core Testing
Small Pages vs Large Pages

- Have 128 TLB entries for large pages (2MB) with quad core instead of 8 on our original Opterons
  - Number for small pages (4KB) still 512
- Most applications run slightly (< 3%) better with large pages
  - Change from original Opterons
- Some benchmarks ran significantly better with large pages
  - Random Access - 14% to 23% better
  - PTRANS - 2.8 to 3.2 times better

Effective Use of Cores

Application	Utilization of each core	Cores effectively used
CTH	71%	2.84
PRONTO	79%	3.18
SAGE	74%	2.95
UMT2K	91%	3.62
PARTISN	40%	1.60

Performance Comparison
Dual-Core vs Quad-Core

	Time in Seconds		
Processor (Socket)	2.4 GHz Dual (940)	2.2 GHz Quad (AM2)	2.6 GHz Dual (AM2)
CTH SN	1938.5	1949.3	1715.9
CTH VN	1357.0	1184.3	1092.4
Pronto SN	3504.0	3503.0	3198.2
Pronto VN	1674.7	1594.4	1456.6
Partisn SN	619.3	425.6	513.7
Partisn VN	297.1	172.2	283.9
UMT2K SN	7108.8	5517.6	6488.4
UMT2K VN	3933.8	2939.9	3512.9

Red Storm Configuration Over Time

	Red Storm (04)	Red Storm (06)	Red Storm (08)
Theoretical Peak Performance	43.52 TF	130.56 TF	290.30 TF
MP-Linpack Performance	36.19 TF	102.2 TF	TBD
Total Memory	33.4 TB	39.2 TB	78.75 TB
System Memory B/W	57.99 TB/s	78.14 TB/s	126.29 TB/s
Disk Storage (Red □ Black)	170 TB □ 170TB	170 TB □ 170 TB	1.5 PB □ 700 TB
Parallel File System B/W (Red □ Black)	100 GB/s 50 GB/s □ 50 GB/s	100 GB/s 50 GB/s □ 50GB/s	160 GB/s 100 GB/s □ 70 GB/s
External Network B/W (Red □ Black)	25 GB/s □ 25 GB/s	25 GB/s □ 25 GB/s	25 GB/s □ 25 GB/s