SAND2008-2699C

Investigating the balance between capacity and capability workloads across large scale computing platforms

9th LCI International Conference on High-Performance Computing
April 29 - May 1, 2008
NCSA, Urbana, Illinois

Mahesh Rajan, Courtenay Vaughan, Robert Leland, Douglas Doerfler, Robert Benner
Sandia National Laboratories
P.O. BOX 5800, Albuquerque, NM 87185
mrajan@sandia.gov

This work was supported in part by the U.S. Department of Energy. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States National Nuclear Security Administration and the Department of Energy under contract DE-AC04-94AL85000.

Objective

- Investigate effectiveness of High-End Computing systems on meeting Sandia's capacity and capability simulation needs
  - Analyze application performance, to thousands of processors, on a large commodity InfiniBand cluster (Thunderbird (tbird)), and, on a large custom Cray XT3 (Red Storm(RS))
  - Use wall time and parallel efficiency to compare performance
  - Analyze parallel efficiency ratio between RS and tbird using a single parameter, namely, *communication time to computation time ratio*

Outline

- Capacity and capability computing and workload at Sandia
- Red Storm and Thunderbird overview
- Description and performance of the seven applications compared
- Scaling analysis
- Conclusions

Capability and Capacity Computing

- Capability Computing
  - Simulations that use a significant fraction of the total nodes installed
  - Simulations that require large memory, I/O, and storage
  - Simulations with stringent time-to-solution and short design cycle times
  - Some combination of the above analysis characteristics making it the only means of achieving the goal

- Capacity Computing
  - Typical analysis runs on tens to hundreds of PEs
  - Several runs to cover a range of parameter space for analysis like *uncertainty quantification*
  - Large user community with total workload constituting a large percentage of total computing cycle needs
  - Typical investment and operating costs are lower

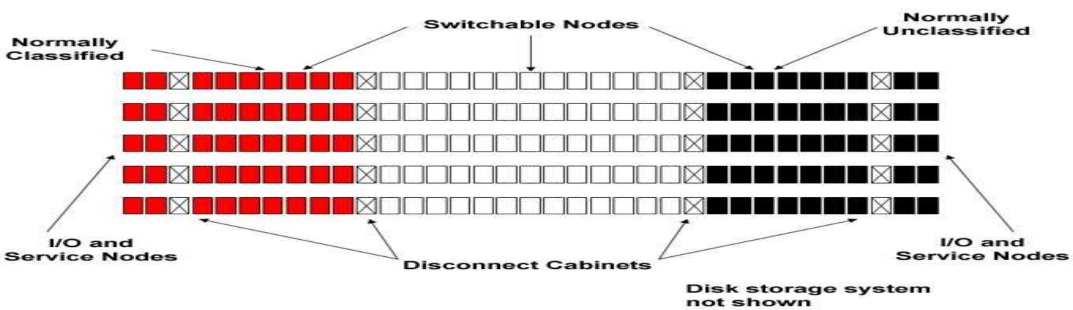
Workload

SNL application node-hour usage and projections

Code	Use	Numerical Method	Current Fraction	Future Fraction
Presto	Crash/ Solid dynamics	FEM, explicit time integration	34.4%	15%
Salinas	Vibration/ Structural dynamics	FEM, spectral analysis	15.8%	10%
LAMMPS	Molecular dynamics	FFT, sparse matrix methods	12.8%	10%
DSMC	Plasma dynamics	Discrete Simulation Monte Carlo	10.4%	10%
CTH	Penetration/ Hydrodynamics	Control volume, explicit time integration	7.4%	10%
ITS	Radiation transport	Monte Carlo	.08%	15%
SAGE	Hydrodynamics	Finite Volume	0.0%	TBD

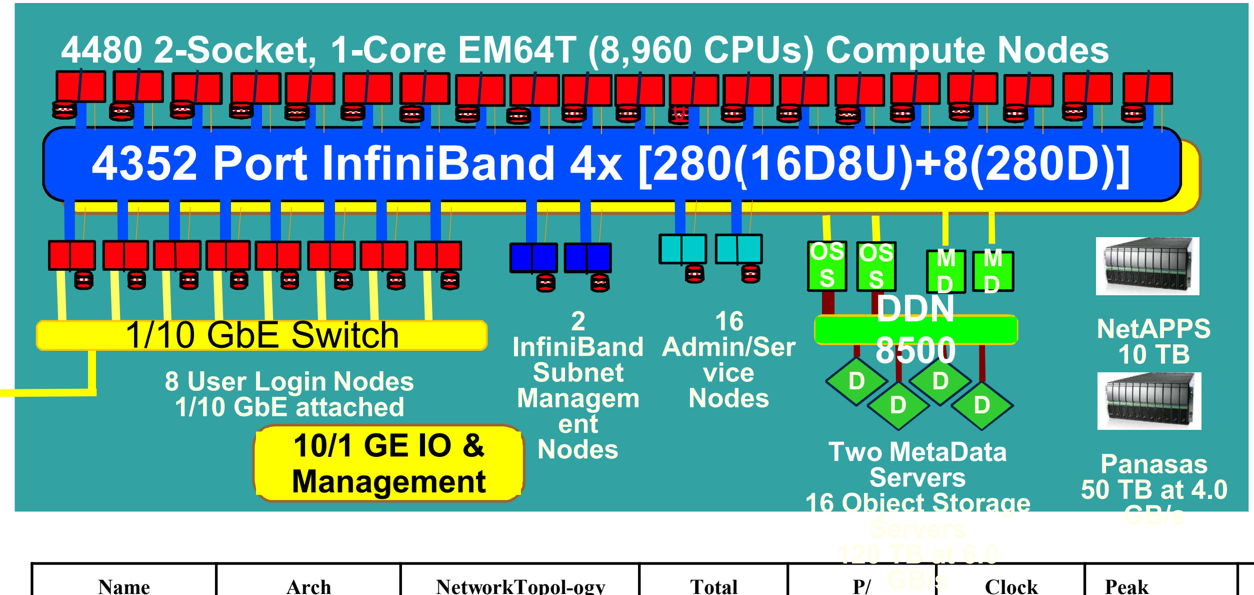
Red Storm and Thunderbird system characteristics

Red Storm Layout (post upgrade) (27 x 20 x 24 Compute Node Mesh)



Red Storm

- 124.42 teraOPS theoretical peak performance
- 135 compute node cabinets
- 20 service and I/O node cabinets
- 20 Red/Black switch cabinets
- 12,460 compute node processors, 320 + 320 service and I/O node processors
- AMD Opteron™ 2.4 GHz dual core processors
- 40 terabytes of DDR memory
- 340 terabytes of disk storage
- Linux/Catamount Operating Systems
- Cray SeaStar Interconnect

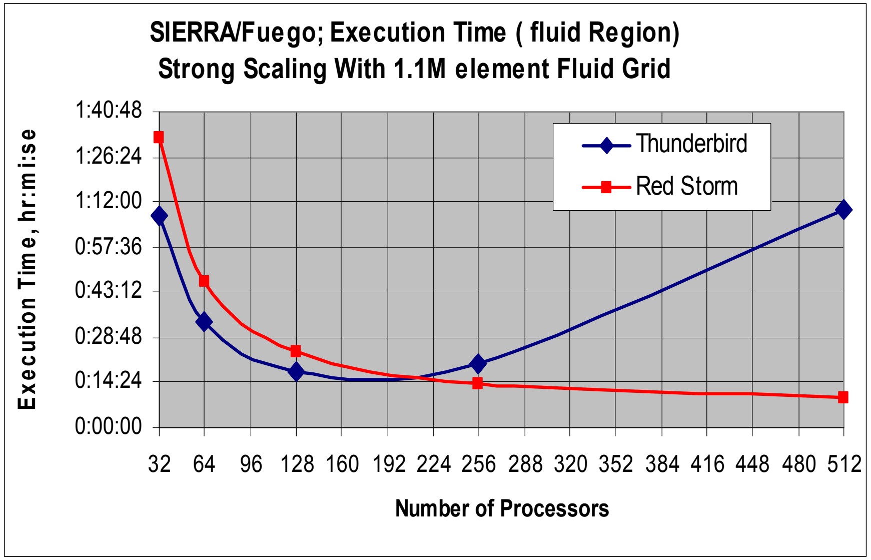


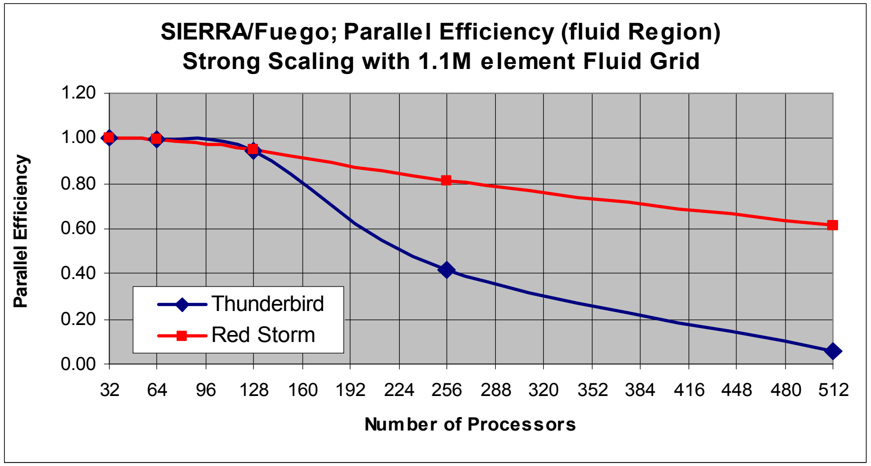
Thunderbird

- 64.5 teraOPS Peak
- 4480 compute nodes
- 9,000 InfiniBand ports
- Intel 3.6 GHz single core EM64T processors
- dual socket SMP nodes with 6GB DDR-2 400 SDRAM
- 26 terabytes of DDR memory
- 400 terabytes of disk storage
- RH Enterprise Linux OS

Name	Arch	NetworkTopol-ogy	Total P	P/ Node	Clock (GHz)	Peak (GF/s/P)	Streams BW(GB/s/P)	MPI Lat (µsec)	MPI BW (GB/s/P)
Red Storm	AMD Opteron	Mesh / Z-torus	25,920	2	2.4	4.8	2.5	5.4	2.1
Thunderbird	Intel EM64T	Fat tree	8960	2	3.6	7.2	3.8	6	0.468

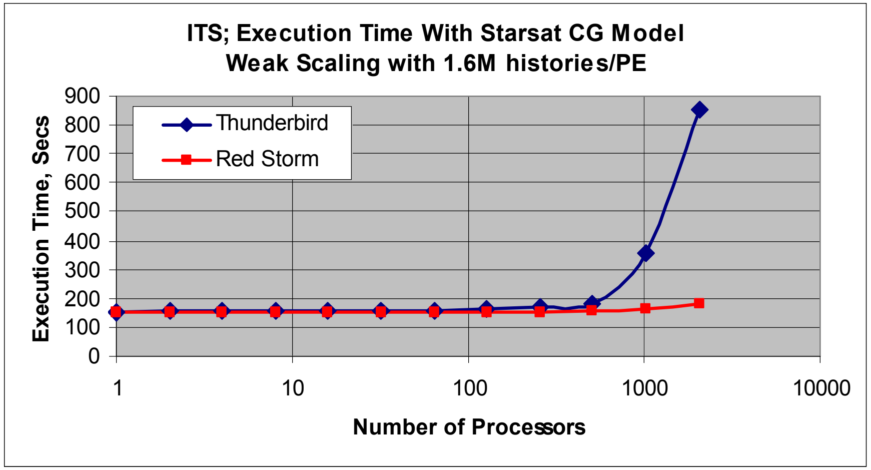
SIERRA/Fuego Fire Simulation

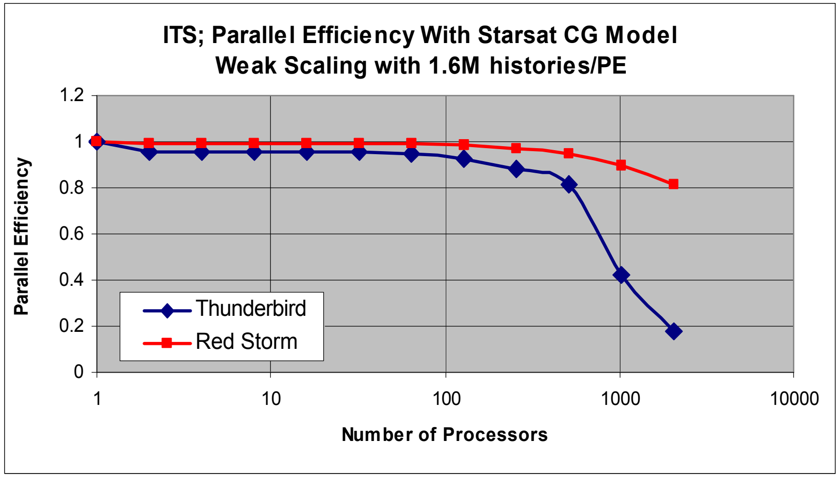




- Fluids, heat transfer, participating media radiation, multi-physics
- Model: weapon-like calorimeter with 1M element fluid mesh, 1M element radiation mesh, and small heat-transfer mesh
- Strong scaling analysis
- Scaling dominated by implicit fluid solves (ML)
- At 256 RS, tbird run times are close, but parallel efficiency is significantly different
- At 512 RS out performs tbird

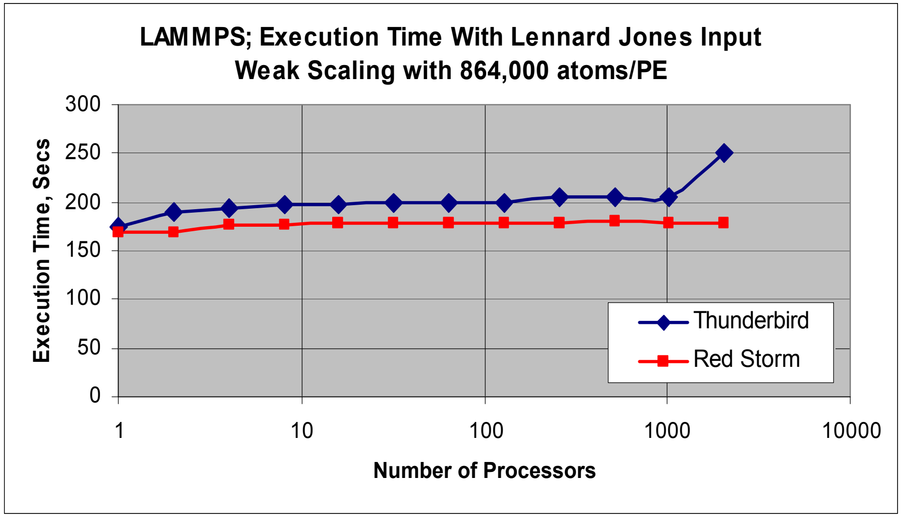
ITS – MC Radiation Transport

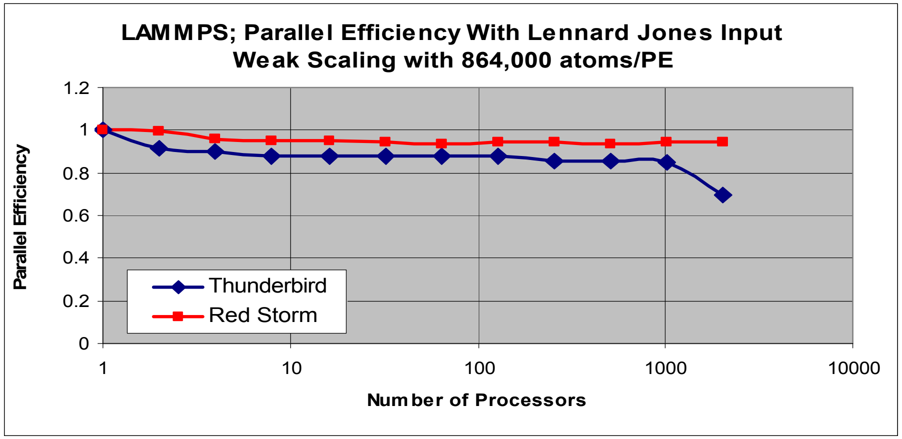




- ➢Monte Carlo Particle Radiation Transport
- ➢Model: STARSAT combinatorial geometry for the satellite electronic components; radiation dosimetery analysis with adjoint solves
- ➢Weak scaling analysis with 1.6 M histories per PE
- ➢ Scaling is only inhibited by the communication to the Master at the end of each batch of history computations from the worker processors
- ➢The communication time is a function of large message bandwidth. Red Storm has a 4X advantage in message bandwidth.

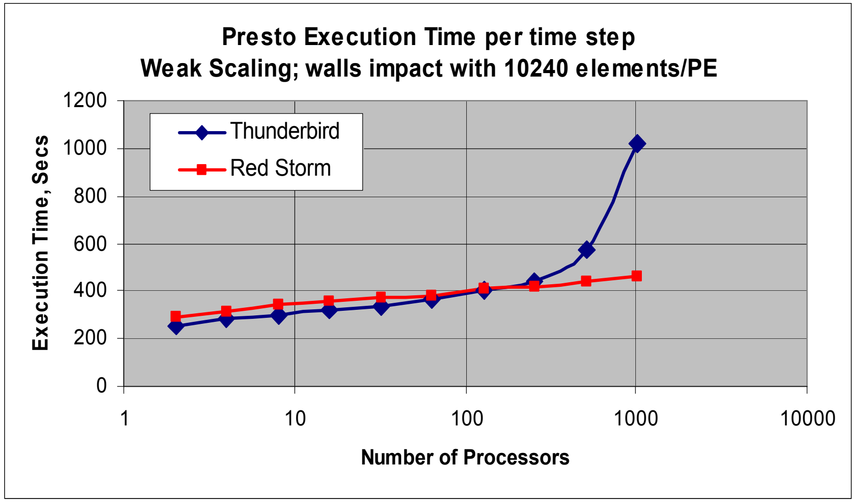
LAMMPS – Molecular Dynamics

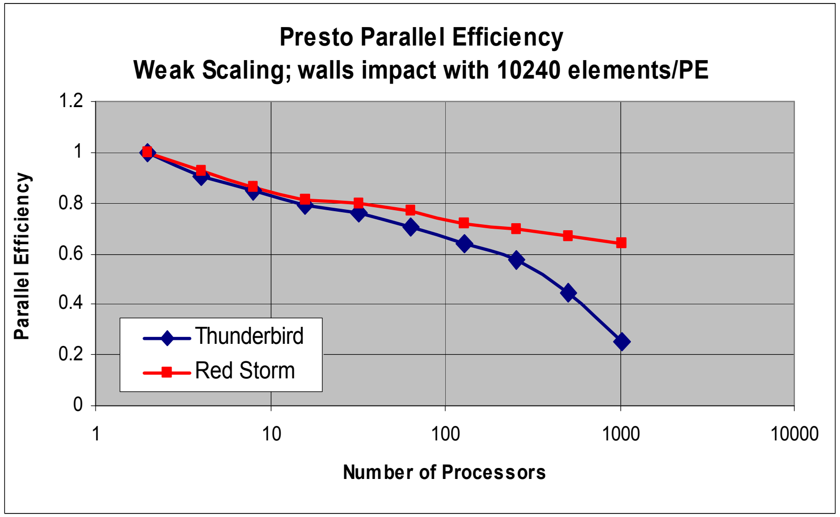




- ➢Classical molecular dynamics
- ➢Model: Lennard-Jones liquid benchmark
- ➢Weak scaling analysis with 864,000 atoms/PE
- ➢ LAMMPS divides the computational domain into three dimensional sub-volumes, and makes the sub-volumes as cubic as possible, The amount of data exchanged is proportional to the surface area of the sub-volume. This favorable volume to surface ratio leads to less than 3% MPI overhead for all the processor counts
- ➢The bump for tbird at 2048 is suspect to be due to OS jitter

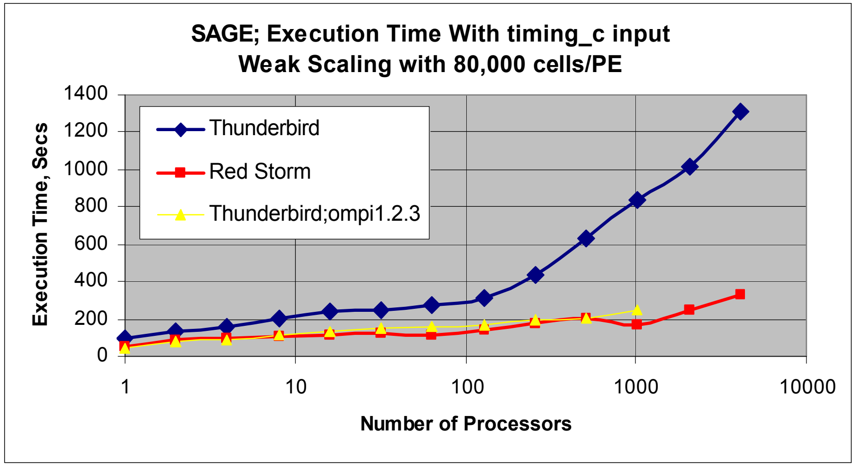
SIERRA/Presto Crash Dynamics

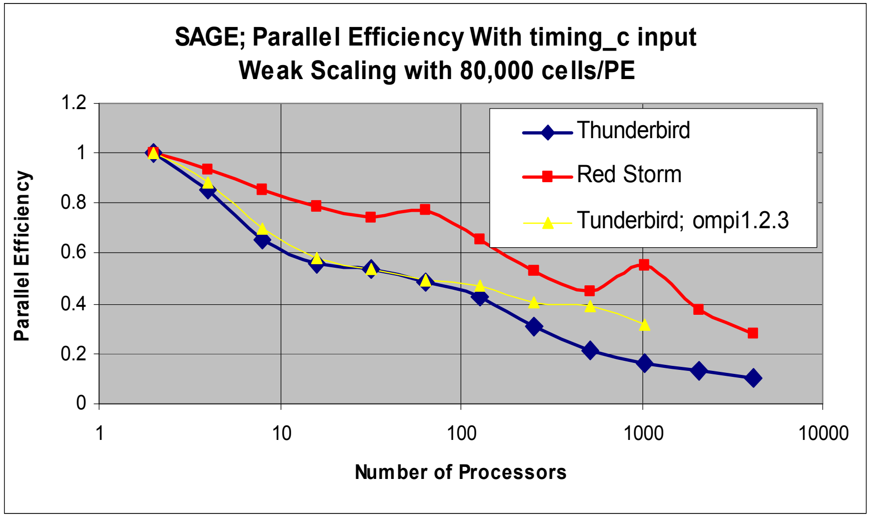




- ➢Explicit ‘crash’ Lagrangian transient dynamics
- ➢Model: Two sets of brick-walls colliding
- ➢Weak scaling analysis with 80 bricks/PE, each discretized with 4x4x8 elements
- ➢ Contact algorithm communications dominates the run time
- ➢ The rapid increase in run time after 256 processors on Thunderbird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

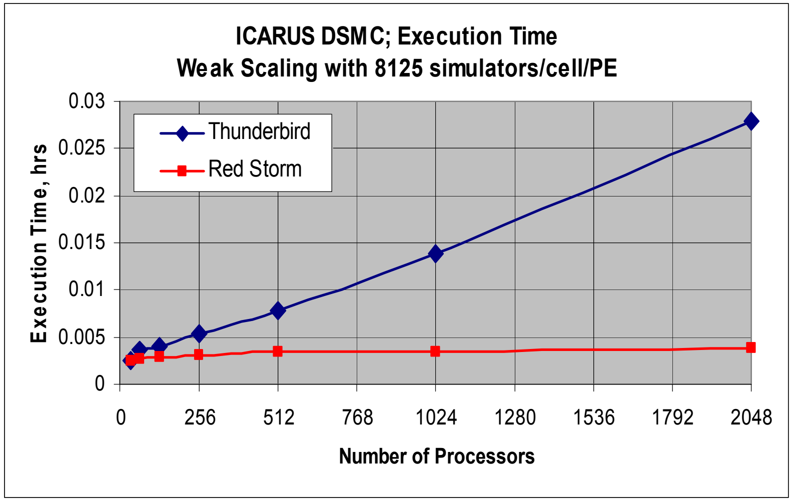
LANL SAGE Hydrodynamics

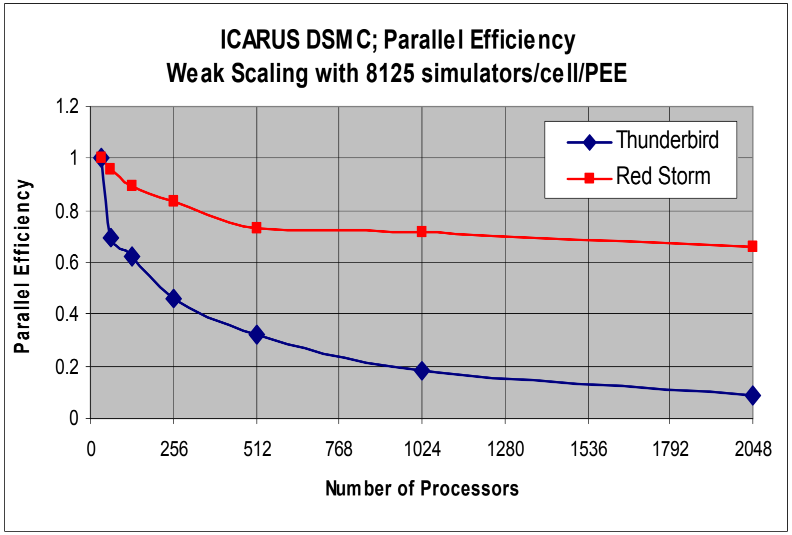




- ➢Multi-material Eulerian Hydro code with AMR
- ➢Model: timing_c hydro benchmark with high communication time to computation time ratio
- ➢Weak scaling analysis with 80,000 cells/PE
- ➢ Communication time is dominated by gather /scatter operations particularly in the z-direction exchanging boundary cell information and also by hundreds of MPI_allreduce at each time step
- ➢ On Tbird, using the OpenMPI 1.2.3 (much improved global Ops) the execution time is significantly better and the trend is quite similar to Red Storm
- ➢The 10%-20% performance advantage on Red Storm is mostly due to better bandwidth

DSMC/ICARUS non-continuum gas flow

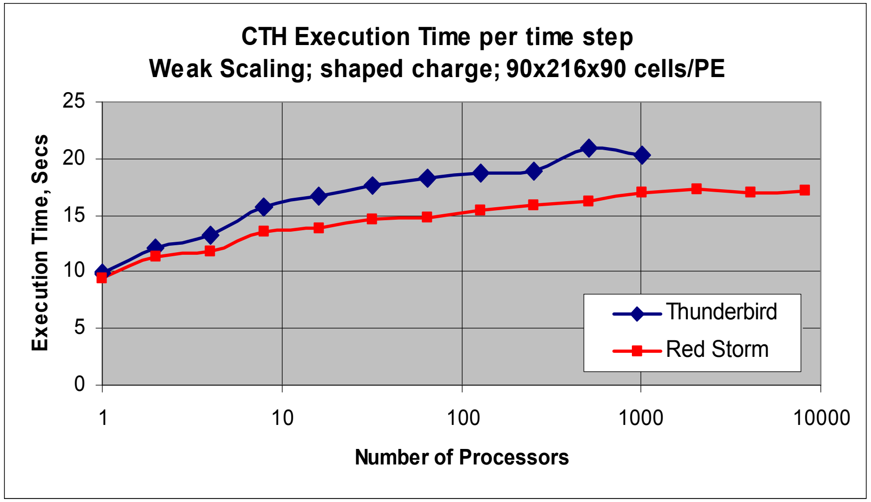


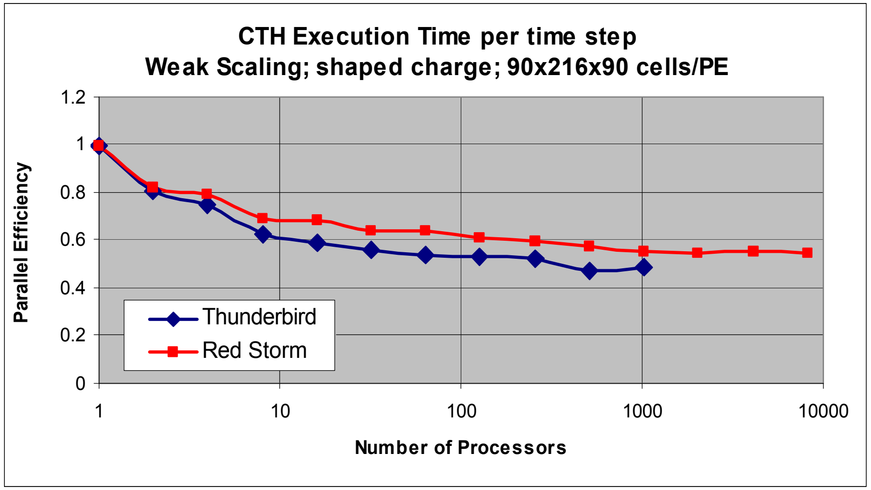


- Direct Simulation Monte Carlo low density flow code; uses computational molecules moving through space, reflecting from solid boundaries, and colliding with one another. By sampling the velocities of large numbers of computational molecules, the gas flow is determined.
- Model: two-dimensional micro-beam
- Weak scaling analysis with 8,125 molecules per PE
- The principal communication operations at each step are molecules position, velocity information to the target processor that has received these molecules
- Higher MPI overhead on tbird due to slower global Ops and lower message bandwidth

Num Procs	% Total time in MPI; Red Storm	% Total time in MPI; Thunderbird
64	14.6	37.9
256	26.6	56.0
1024	31.0	75.6

CTH-Sandia’s 3D Shock Hydrodynamics





- ➢CTH is used for two- and three-dimensional problems involving high-speed hydrodynamic flow and the dynamic deformation of solid materials
- ➢Model: shaped-charge; cylindrical container filled with high explosive capped with a copper liner.
- ➢Weak scaling analysis with 90x216x90 computational cells per processor.
- ➢Processor exchanges information with up to six other processors in the domain. These messages occur several times per time step and are fairly large since a face can consist of several thousand cells
- ➢Modest communication overhead with nearest neighbor exchanges

Efficiency ratio, Red Storm to Thunderbird

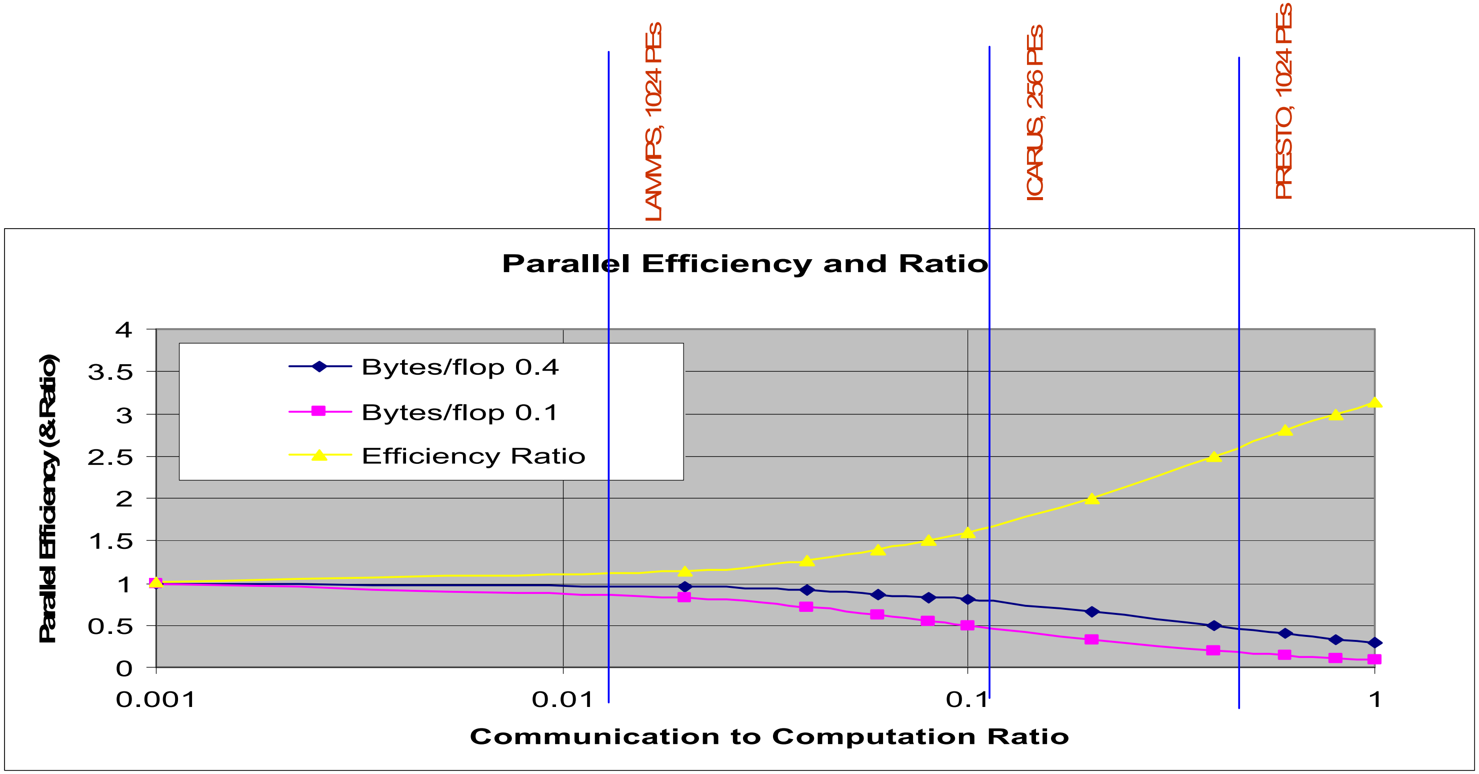
Apps.\ PEs	64	256	1024
ITS	1.048	1.101	2.121
SAGE	1.590	1.692	3.413
Fuego	0.999	1.933	10.133
DSMC	1.385	1.800	3.943
LAMMPS	1.074	1.109	1.108
CTH	1.183	1.135	1.136
Presto	1.091	1.214	2.563

Tbird Large Capability Run Inhibiting Factors

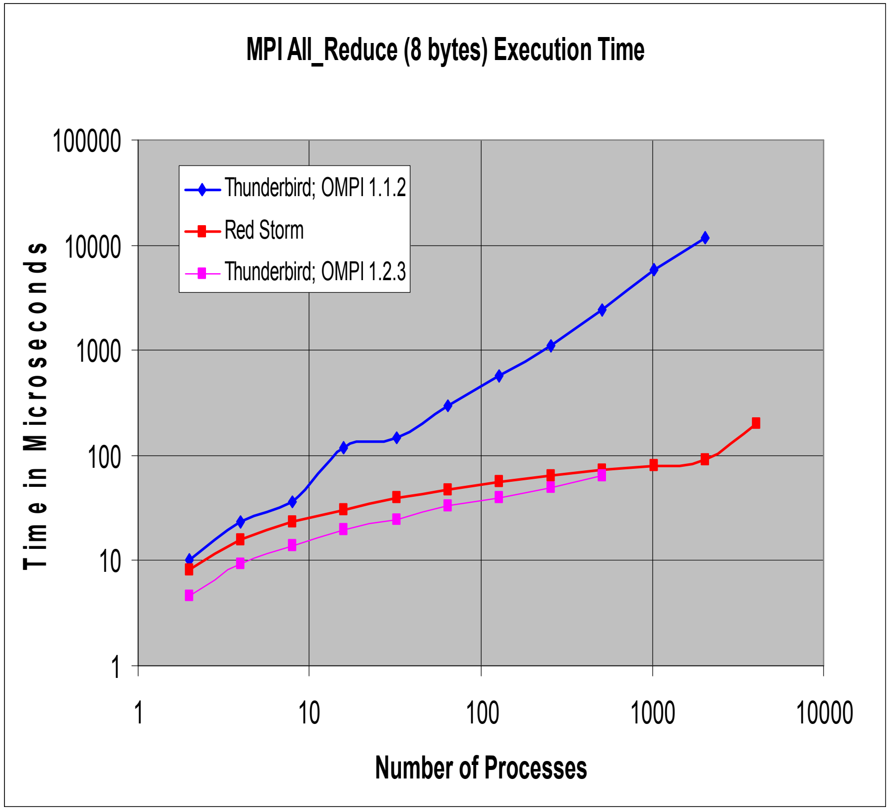
1. Large variation in message bandwidth rate across pairs of processors
2. Latency variation among communicating processors impacting application with short cycle time
3. 4X bandwidth disadvantage over Red Storm
4. OMPI 1.1.2 poor global ops performance (1.2.3 performance is much better)
5. OS noise effects
   a) 30% variation in run times vs. 2-3% on Red Storm
   b) 100 sec MATMUL loop in each PE shows 2.5% variation vs. 0.4% on Red Storm
   c) 100 sec MESSAGE_EXCHANGE between 50 pairs of nodes shows 42 % variation vs. 3% on Red Storm

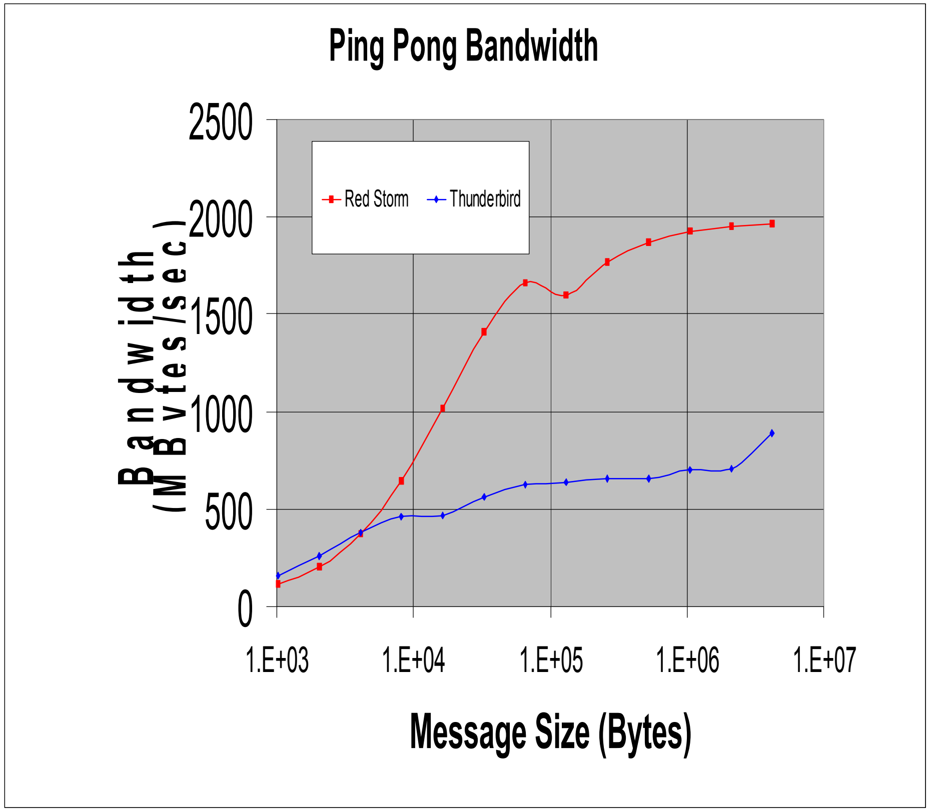
Parallel efficiency model; $E = 1 / (1 + f)$

Impact of communication to computation ratio of different applications

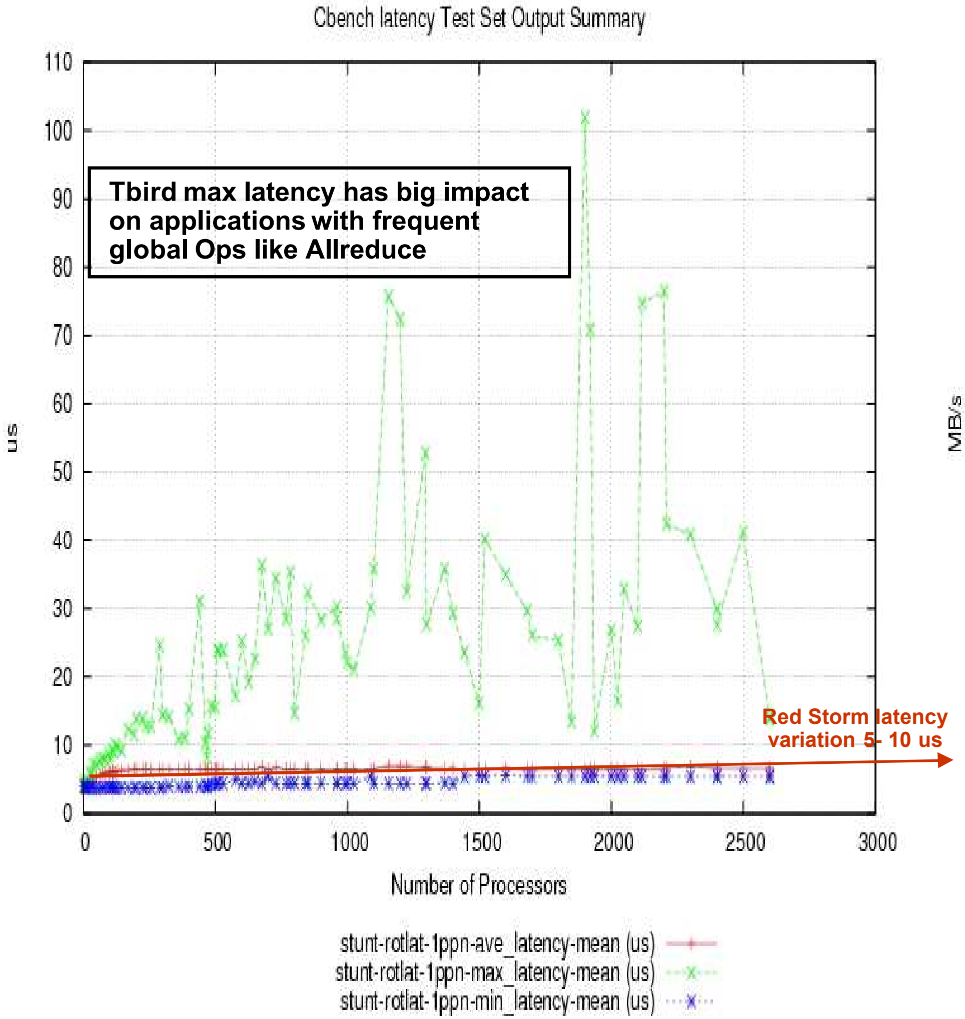


MPI Allreduce and ping-pong performance comparison



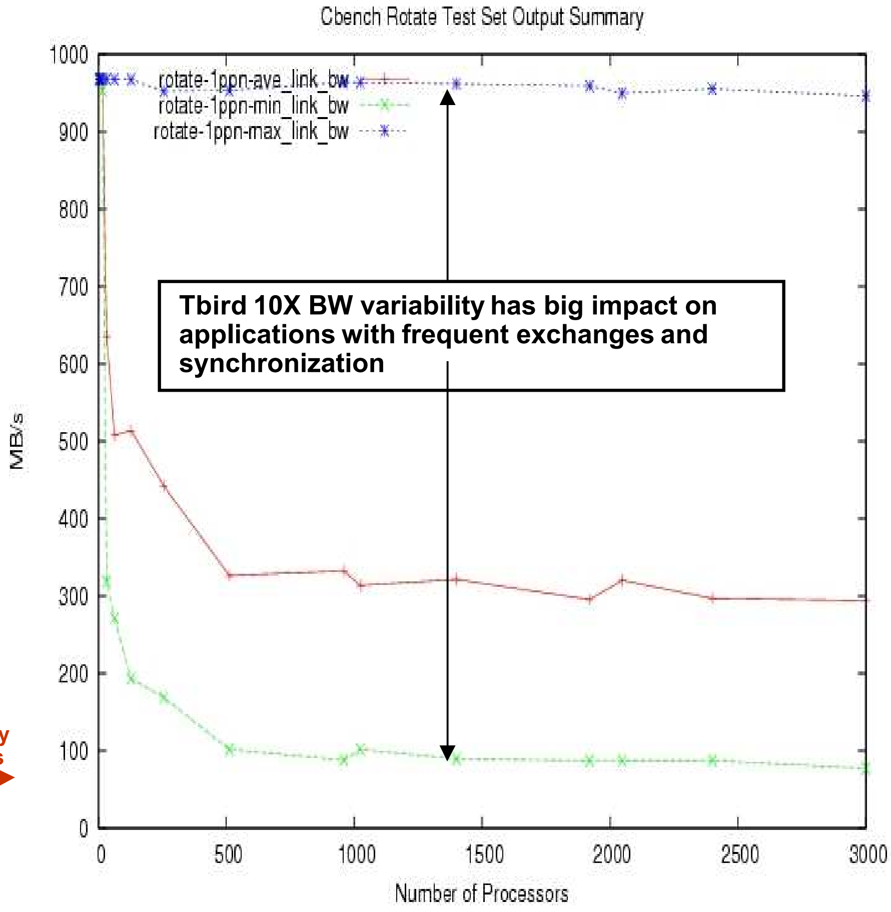


CBENCH; Thunderbird (IB) Bandwidth, Latency Variation Impact



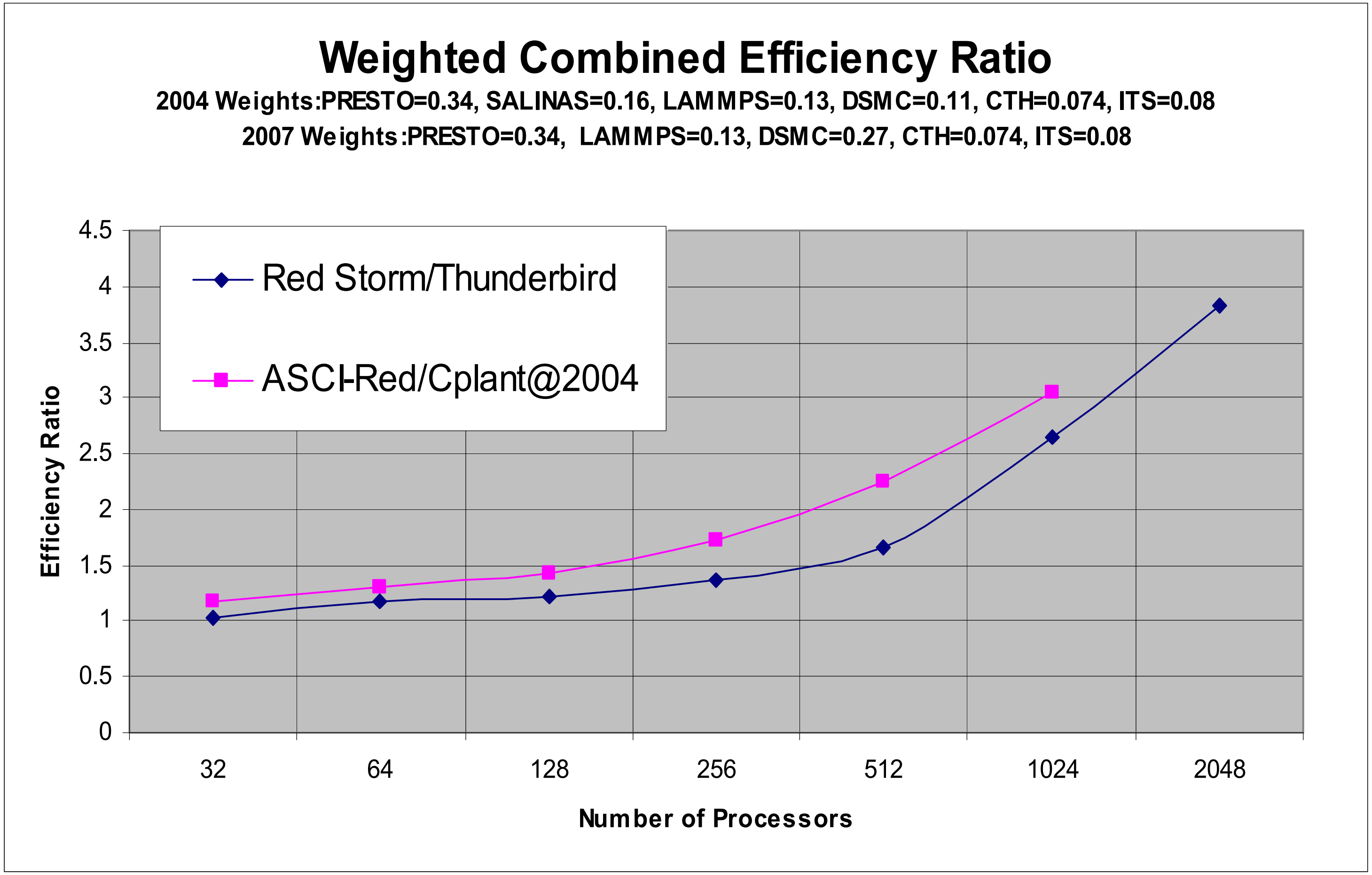
Tbird max latency has big impact on applications with frequent global Ops like Allreduce

Red Storm latency variation 5- 10 us



Tbird 10X BW variability has big impact on applications with frequent exchanges and synchronization

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



Conclusions

- Application performance characteristics constituting the workload has been measured on Red Storm and Thunderbird
- Used parallel efficiency ratio as a simple measure to compare capability and capacity system performance
- Applications show a factor of 2 to 10 better performance on Red Storm at 1000 processors
- Principal factors limiting commodity clusters (due to low parallel efficiency) for capability class simulation are:
  - At thousands of processors latency and and bandwidth significantly degrade (CBENCH data)
  - For applications with short cycle times, the fraction of communication overhead grows significantly (non-linearly) due to contention and OS Jitter
  - Application with short cycle time, requiring several global operations after each cycle, showed poor scalability, with early releases of OpenMPI. Global operations performance improved with OpenMPI1.2.3.
- Work remains to identify precise causes of performance differences seen between Red Storm and Thunderbird