Early Evaluation of the Cray XD1
(FPGAs not covered here)

Mark R. Fahey
Sadaf Alam, Thomas Dunigan,
Jeffrey Vetter, Patrick Worley

Oak Ridge National Laboratory

Cray User Group
May 16-19, 2005
Albuquerque, NM

Acknowledgment

- This research was sponsored by the Office of Mathematical, Information, and Computational Sciences, Office of Science, U.S. Department of Energy under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

Evaluation of Early systems

- A project that attempts to evaluate *quickly* the promise of “early” (possibly immature) systems:
  - Verifying advertised functionality and performance
  - Quantifying performance impact of unique system characteristics
  - Providing guidance to (early) users
    - What performance to expect
    - Performance quirks and bottlenecks
    - Performance optimization tips

Early Systems

- **ORNL currently has three early systems**
  - **Cray XD1**
    - **144 processors installed in October 2004, upgraded to 1 12-chassis system just a week ago**
  - **Cray XT3**
    - **Currently 40 cabinets (3748 compute processors) running CRMS build of OS**
    - **See Jeff Vetter's talk at 2:30 in Taos**
  - **Cray X1E**
    - **X1E boards will start arriving in June**
  - **See Pat Worley's Interconnect talk at Wed at 2:00**

Evaluation Methodology

- Hierarchical evaluation
  - Microbenchmarks
  - Application-relevant kernels
  - Compact or full parallel application codes
- Open evaluation
  - Rapid posting of evaluation results
- Fair evaluation
  - Determining appropriate ways of using system, evaluating *both* traditional and alternative programming paradigms
  - Collecting data with *both* standard and custom benchmarks

Cray XD1

- **Each chassis has 12 2.2 GHz AMD Opterons**
  - **64K L1, 1M L2, 2 f-p inst. per cycle**
- **12 chassis in the rack**
- **4 GB of memory per processor**
- **Totals: 144 procs (peak 633 GFlops), 576 GB, 18 TB disk**
- **With recent upgrade, can run with up to 142 procs**
- **PBS Pro is our batch system**

Other test platforms

- **Cray XT3 at ORNL: 3748 2.4 GHz AMD Opterons in the compute partition, connected in a 10x16x24 grid**
- **Cray X1 at ORNL: 512 multistreaming processors**
- **Earth Simulator: 640 8-way SMP with a single stage crossbar interconnect**
- **SGI Altix 3700 at ORNL: 256-way SMP 1.5 GHz Itanium2 with NUMAflex fat-tree**
- **IBM p5 720 at ORNL: 4-way Linux SMP with 1.65 GHz POWER5**
- **IBM p690 cluster at ORNL: 27 32-way SMP 1.3 GHz POWER4, Federation Switch**
- **IBM SP at NERSC: 184 Nighthawk II 16-way SMP 375 MHz POWER3, SP Switch2**
- **HP AlphaServer SC at PSC: 750 ES45 4-way SMP 1GHz Alpha, Quadrics interconnect**

Outline for rest of talk

- **Sampling of**
  - **Microbenchmarks**
  - **Applications**

- **Much more at**
  - **http://www.csm.ornl.gov/evaluation**
  - **http://www.csm.ornl.gov/~dunigan/xd1/**

Caveats

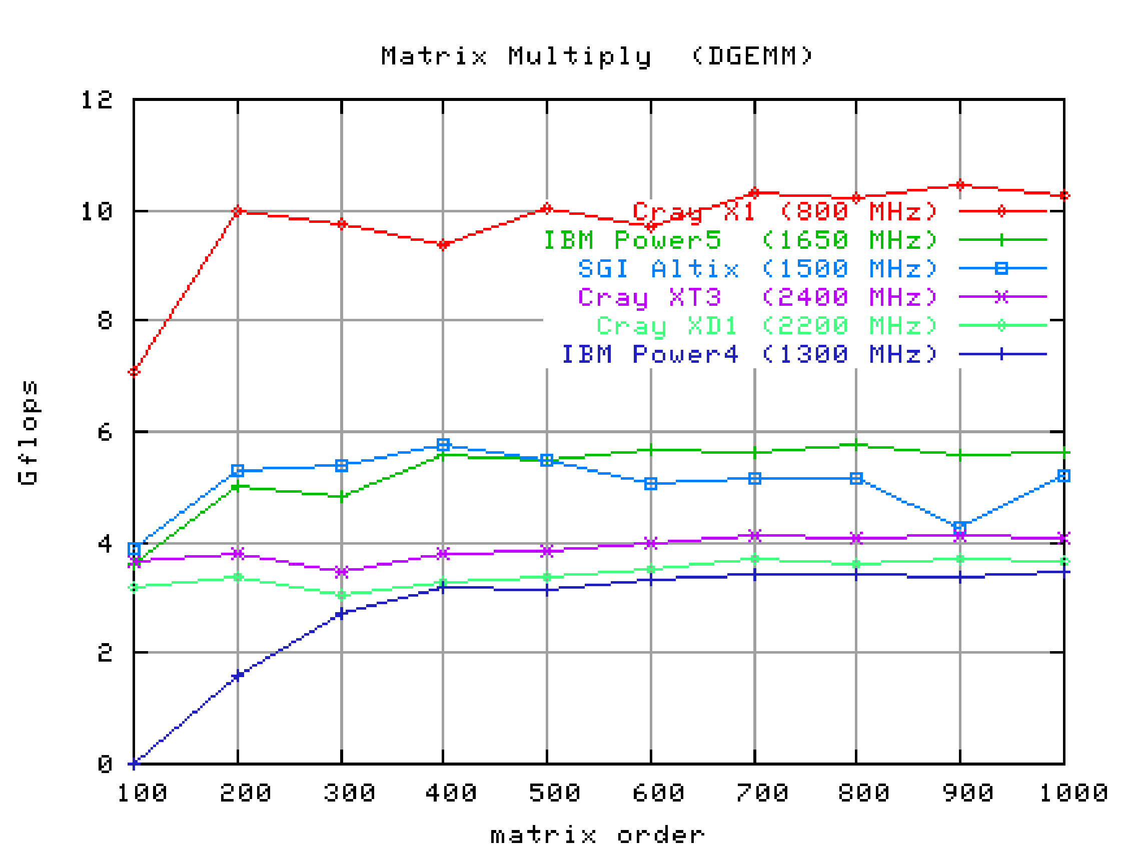
- These are early results resulting from sporadic benchmarking on evolving system software and hardware configurations
- Most results obtained on 64-processor configurations with main fabric only

Microbenchmarks

- **Latency between**
  - **nodes on a chassis 1.7 us**
  - **nodes on different chassis 2.2 us**
- **Bandwidth between**
  - **nodes on a chassis 1.3 GB/s**

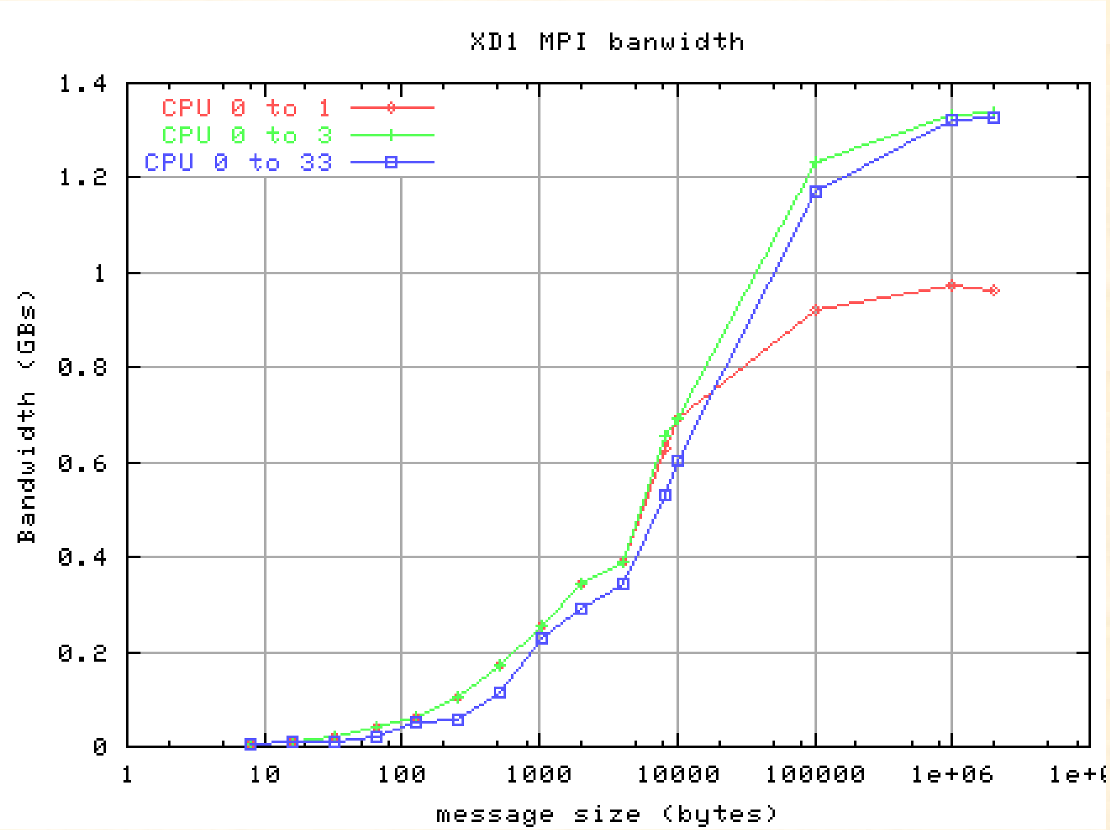
Matrix multiply

- **DGEMM using vendor optimized version**
- **XD1 84% of peak**

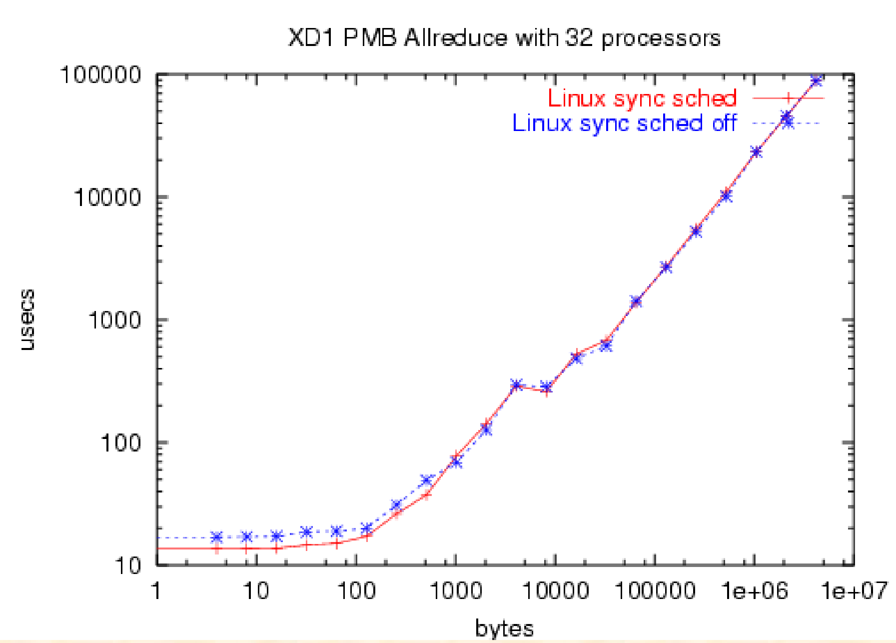


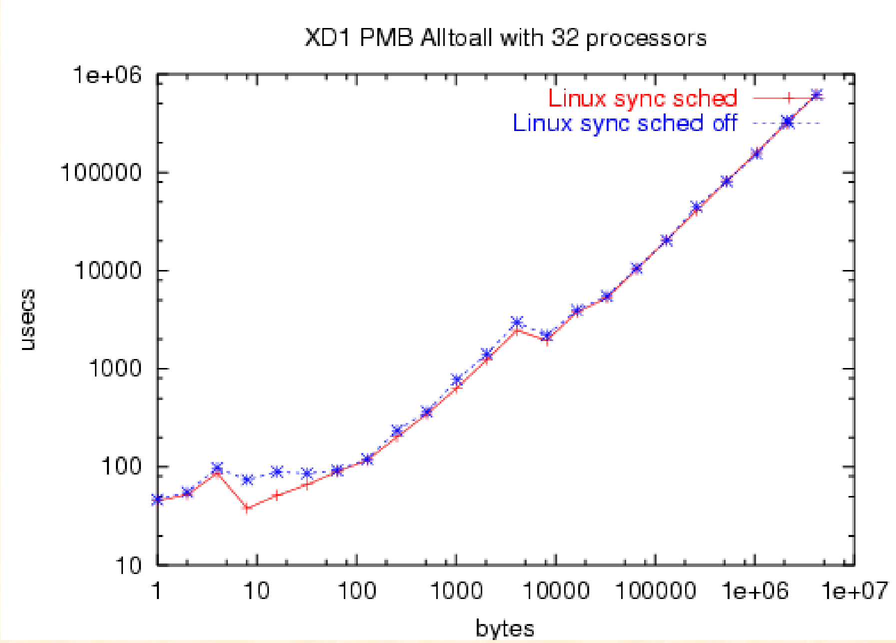
XD1 MPI Bandwidth

- **Internode bandwidth higher than intranode**
- **Little degradation as distance increases**



PMB allreduce and alltoall





- **LSS works; about ~15-20% improvement for smaller messages**
- **Bump going from 4096 to 8192 – not understood**

Applications

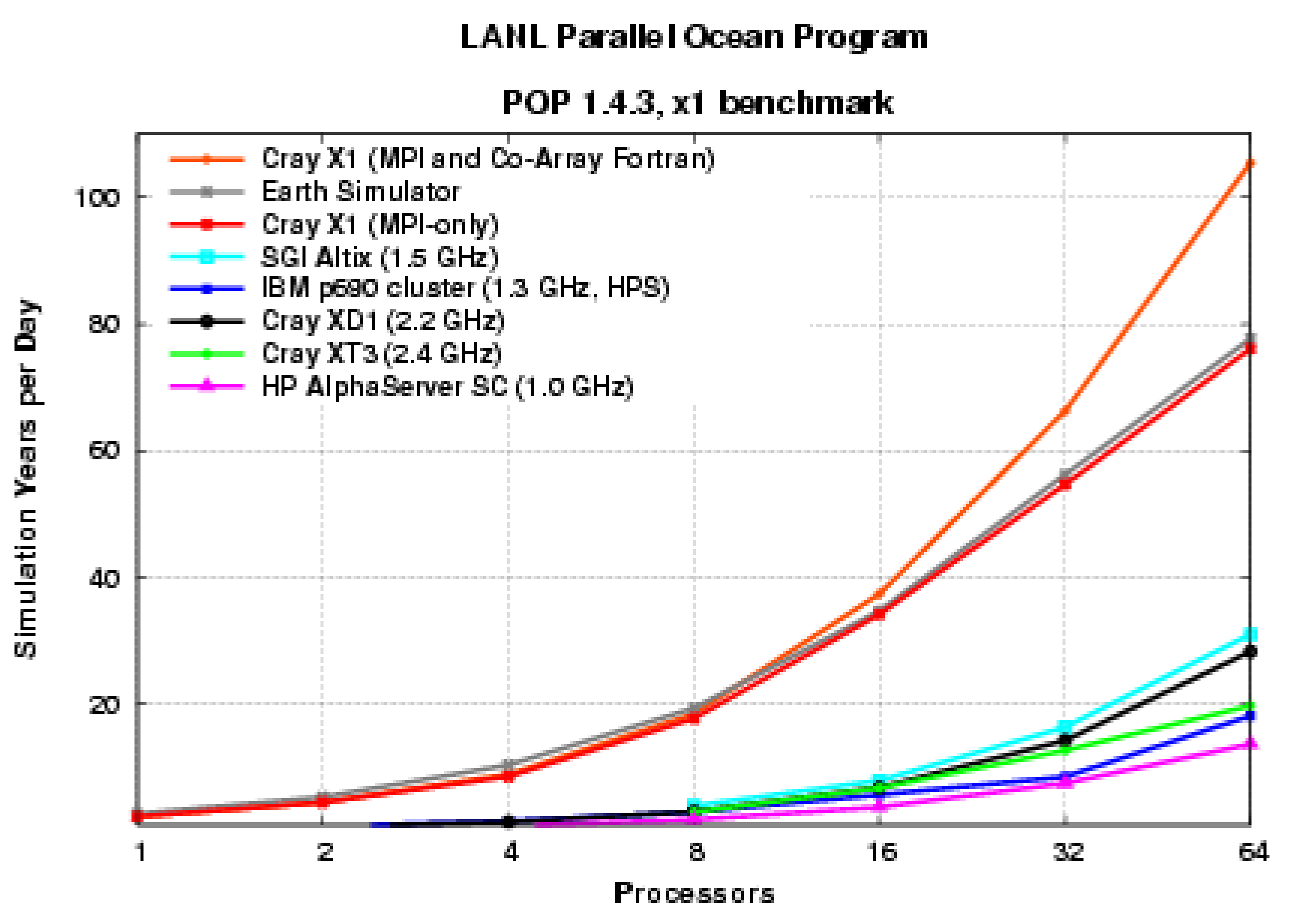
- **POP (climate)**
- **CAM (climate)**
- **GYRO (fusion)**
- **VASP (materials)**
- **sPPM (hydrodynamics)**

POP

- **Parallel Ocean Program**
  - **Ocean component of the Community Climate System Model (CCSM)**
  - **Developed and maintained at LANL**
  - **Finite-difference scheme of 3-d flow equations**
  - **Used the “x1” benchmark**
    - **Relatively coarse resolution, similar to current coupled models**
    - **1 degree resolution**
    - **40 vertical levels**

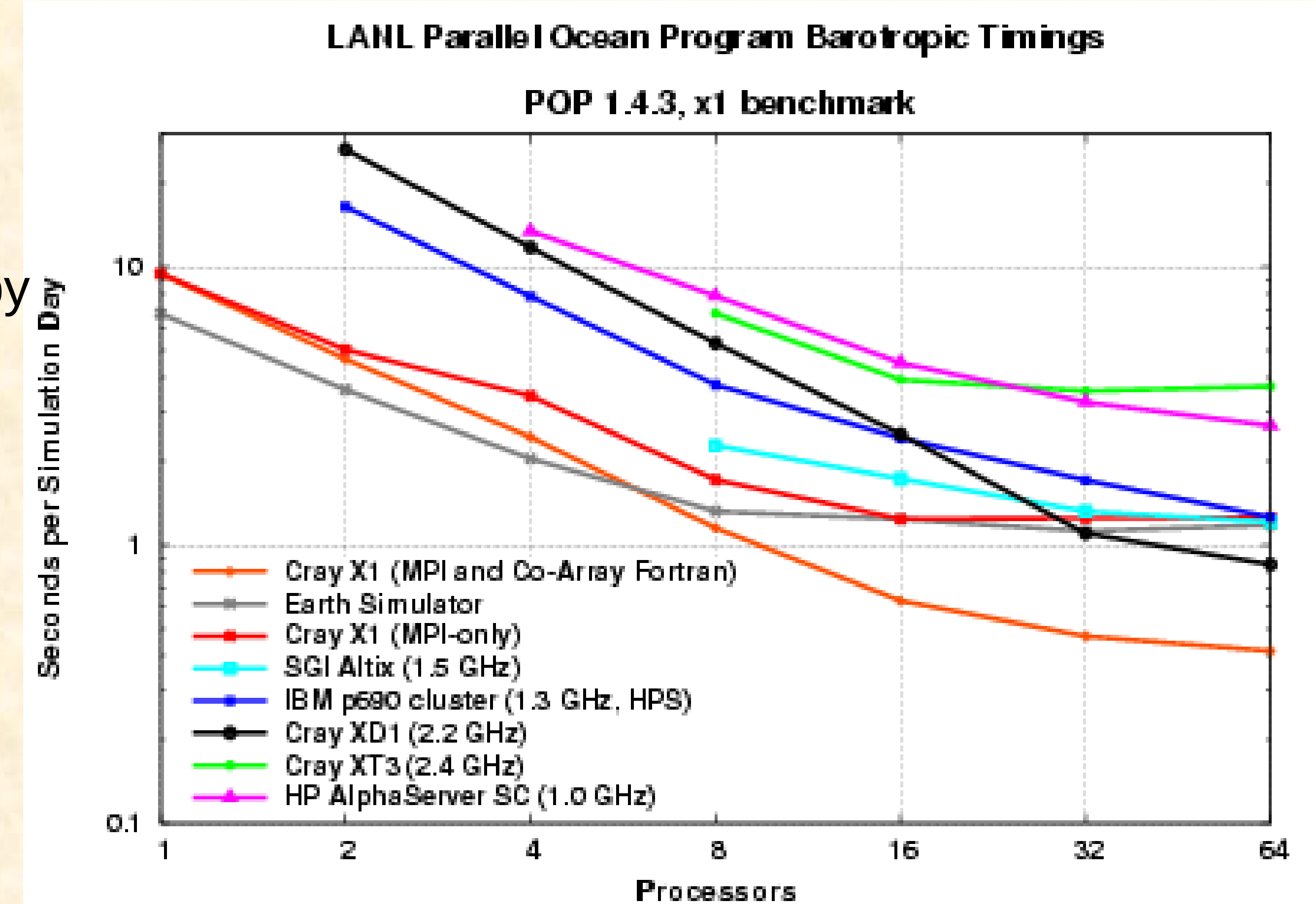
POP (cont.)

- XD1 performance just slightly below Altix
- Scaling well



POP: barotropic

- This portion of POP dominated by 2D implicit solve
- Known to scale poorly
- XD1 doing quite well



CAM

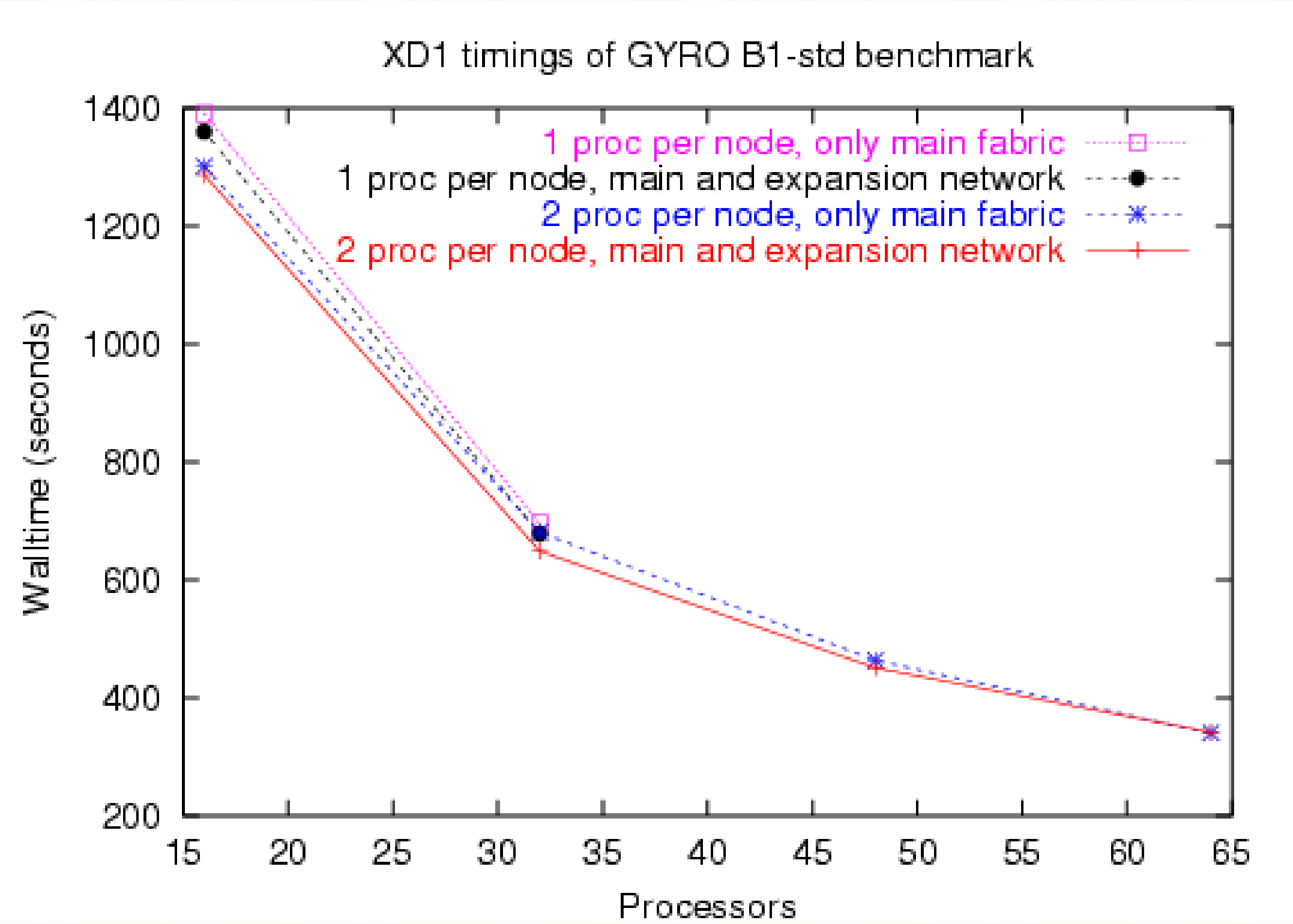
- **Community Atmospheric Model (CAM)**
  - **Atmospheric component of CCSM**
- **All is not rosy:**
  - **Results from CAM are not deterministic on the XD1 using the PGI 5.2-4 compilers**
  - **CAM *pergro* test known to fail on Opterons with PGI 5.2**
- **Positive outlook though**
  - **CAM has been shown to work on Opteron clusters with the Pathscale compilers**

GYRO

- **Simulates tokamak turbulence**
- **Solve time-dependent, nonlinear gyrokinetic-Maxwell equations**
- **Uses a 5-d grid**
- **Used the B1-std Benchmark problem**
  - **Flux-tube electrostatic simulation with kinetic electrons and collisions**
  - **140x8x8x16x20x2 grid**
  - **Uses multiples of 16 processors**

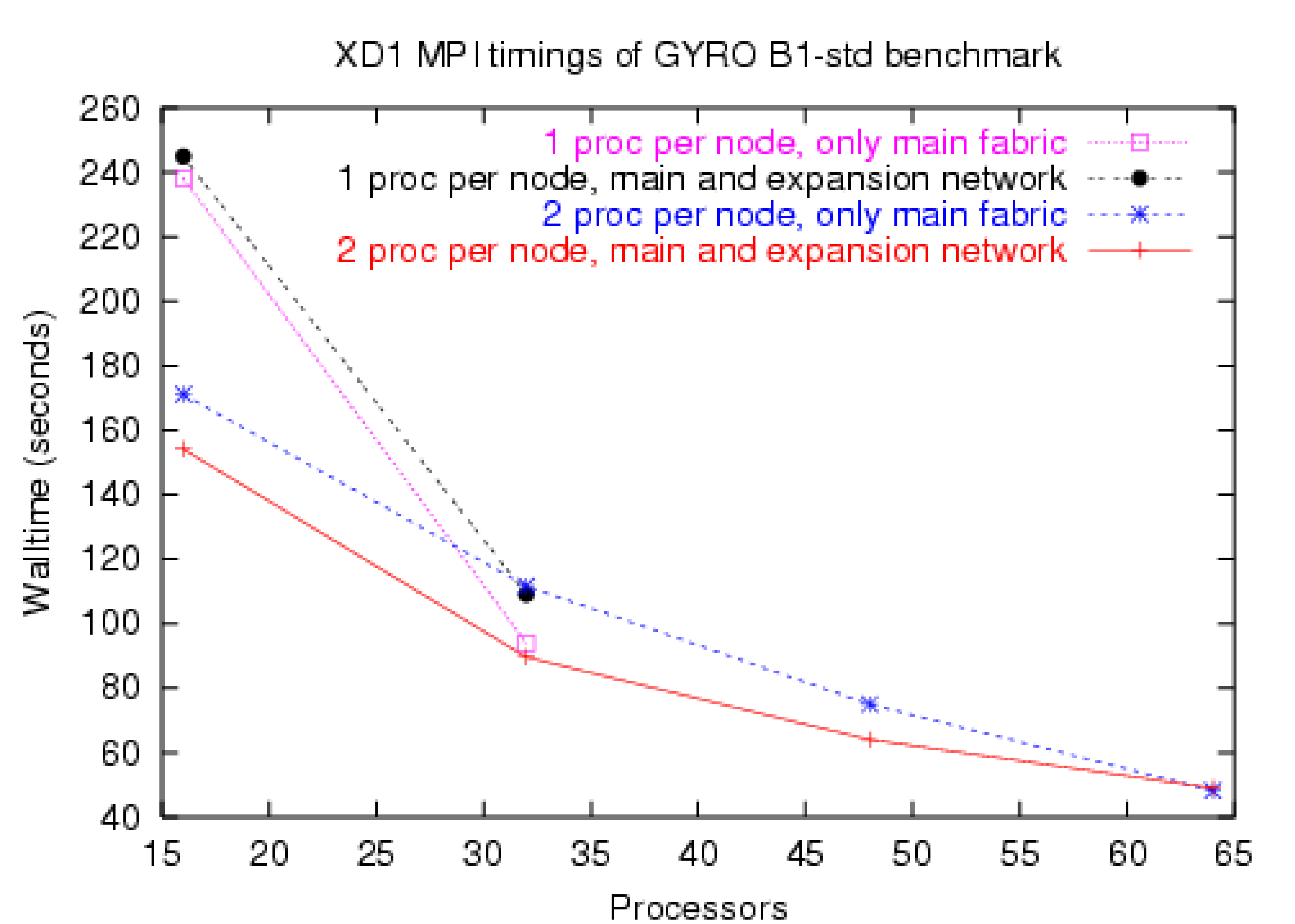
GYRO (cont.)

Indicates that 2p per node and both fabrics are more efficient



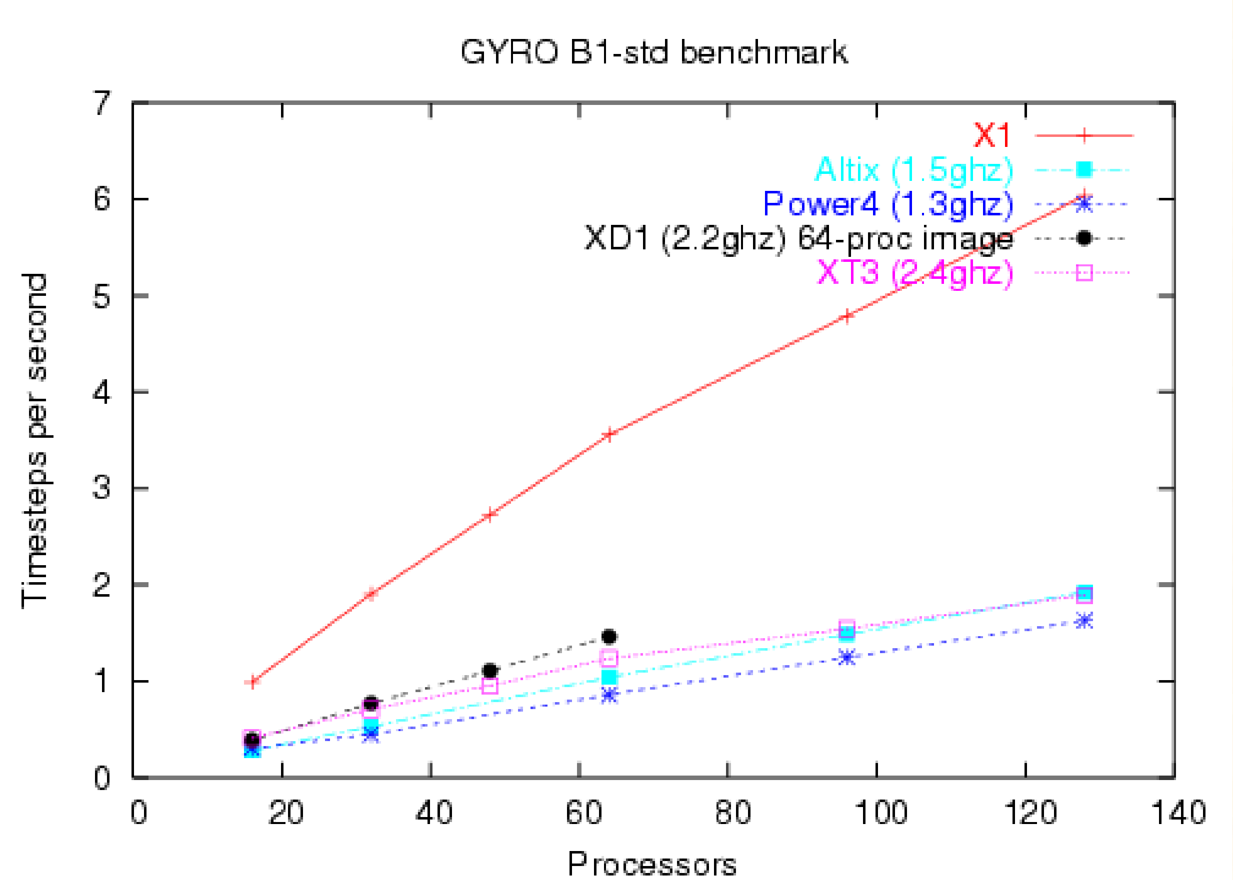
GYRO (cont.)

- At odds with earlier bandwidth data



GYRO (cont.)

- **Platform inter-comparison**
- **XD1 compares quite favorably**
- **Have data for 144p XD1 on later slide**

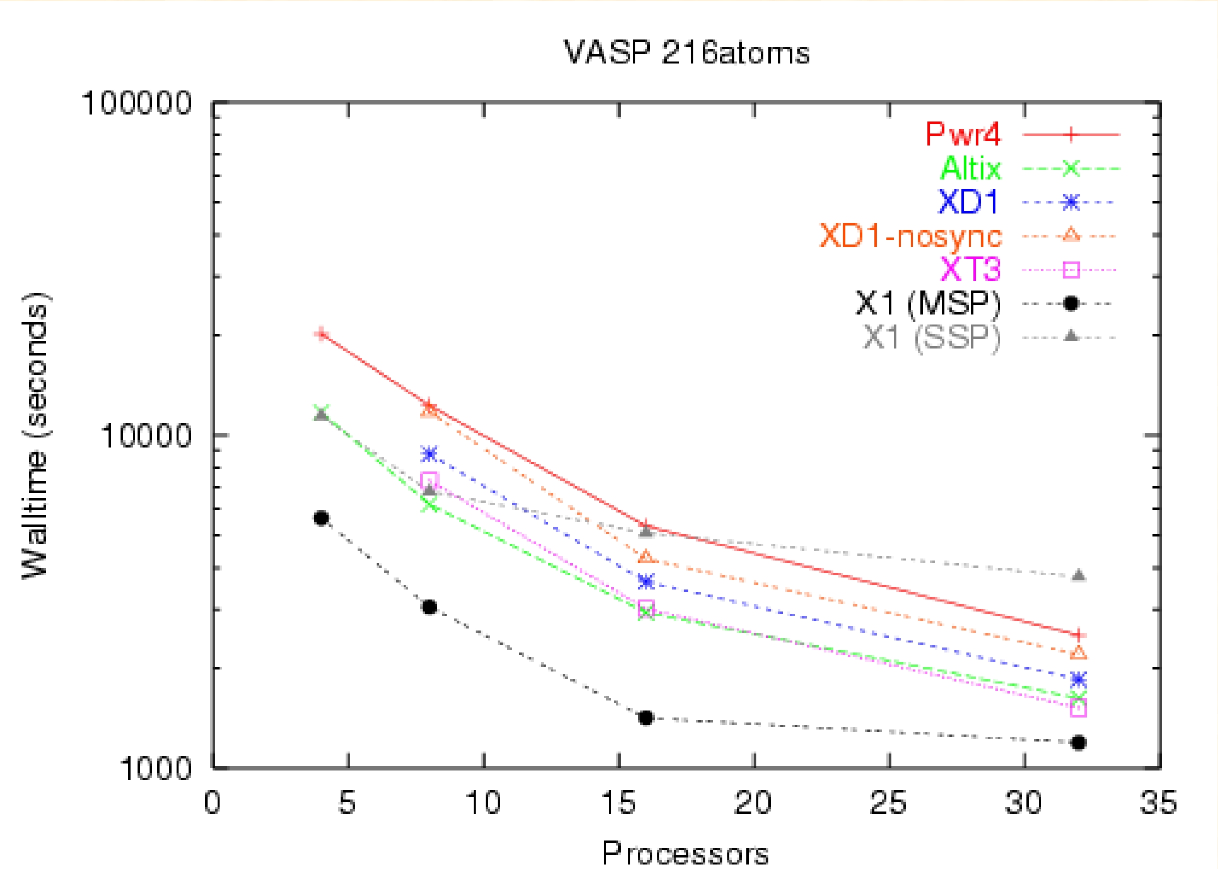


VASP

- **Vienna ab-initio Simulation Package for molecular dynamics**
  - **Uses pseudopotentials and plane wave basis sets**
  - **Approach based on finite-temperature local-density approximation and exact evaluation of instantaneous electronic ground state at each step**
- **Test case is a 216 atom benchmark**

VASP (cont.)

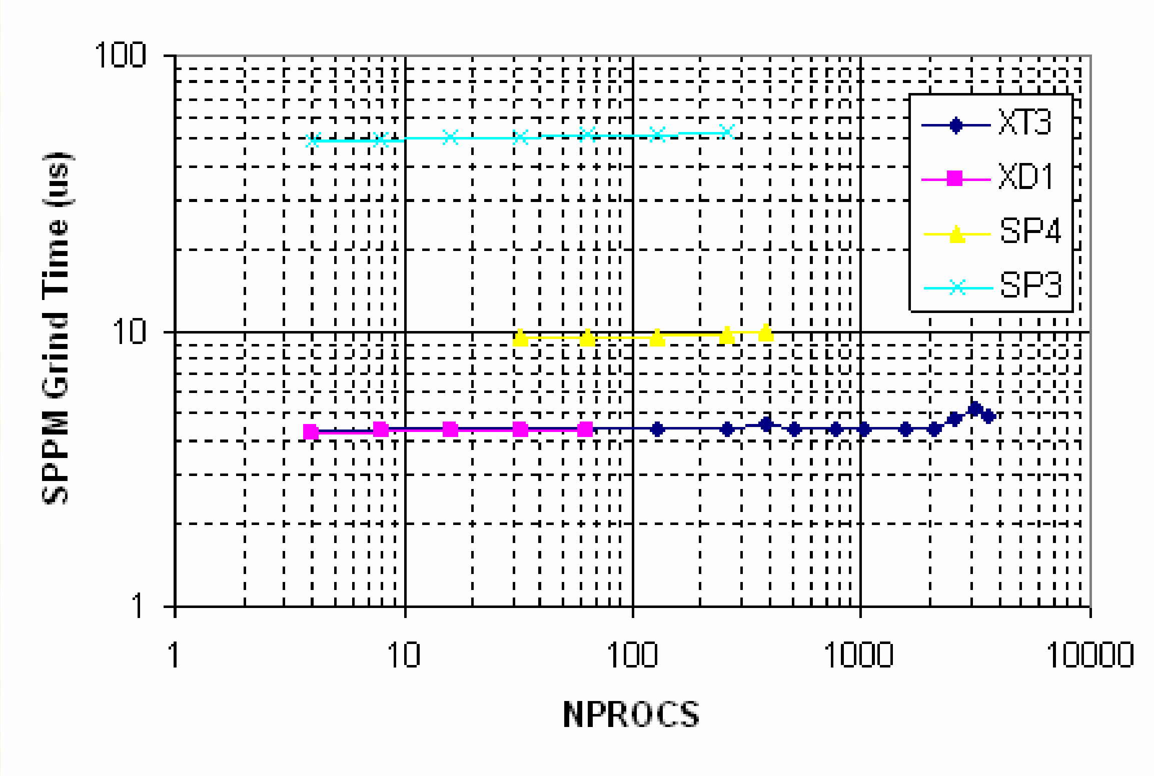
- **XD1 compares well**
- **LSS makes a noticeable difference**



sPPM

- **sPPM solves a 3-D gas dynamics problem on Cartesian mesh, using simplified version of piecewise parabolic method**
- **Makes use of a split scheme of X, Y, and Z Lagrangian and remap steps**
  - **Computed as 3 sweeps through the mesh**
- **Test case is a shock passing through a gas benchmark**
  - **24 billion zones**

sPPM (cont.)



SPPM Grind Time (us)
100
10
1
1
10
100
1000
10000
NPROCS
XT3
XD1
SP4
SP3

Recent upgrade

- Went from 2 6-chassis systems to 1 12-chassis system (1st in US, 2nd worldwide by 2 days)
- 12 main and 12 expansion fabric links on each chassis
  - 6-chassis system:
    - two RA cables connect each pair of chassis
  - 12-chassis configuration:
    - one RA cable connects each pair of chassis
- So we essentially cut our interconnect bandwidth in half

Note:

- Fat-tree topologies, which use switch chassis, offer a key advantage over the direct-connect topologies: the bisection bandwidth is better. Fat-tree topologies work well in large systems (or small systems that will likely grow) on which applications frequently exchange data between processes that run in different Cray XD1 chassis, such as applications that perform frequent all-to-all communication.
  - Taken from the Cray XD1 RapidArray Interconnect Topologies manual

Revisit interconnect setup

Note that for a 12 chassis system, there are pairs of chassis that share two links

**My best guess as to the actual mapping

From	Port	To	Port
1	1	2	1
1	12	2	12
1	11	3	12
1	10	4	12
1	9	5	12
1	8	6	12
1	7	7	12
1	6	8	12
1	5	9	12
1	4	10	12
1	3	11	12
1	2	12	12

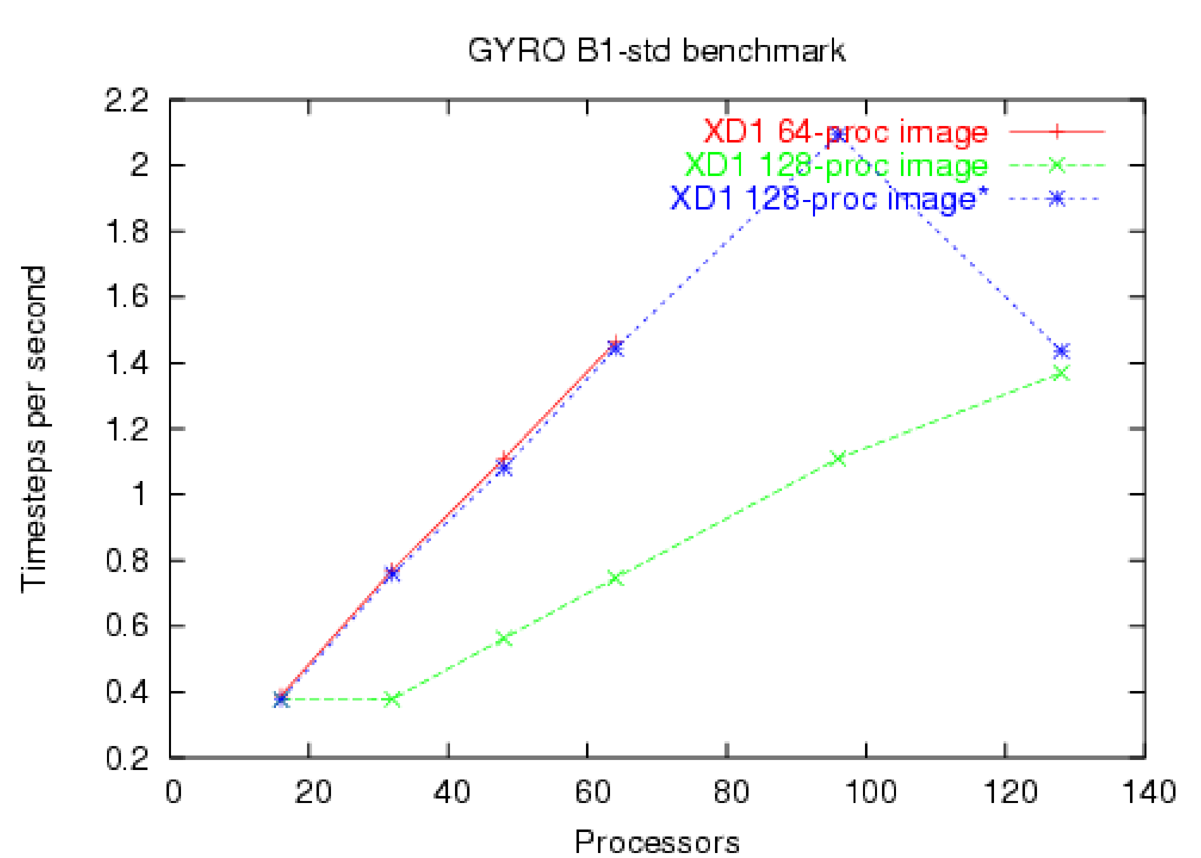
GYRO again

Same XD1 data plot as earlier, but with data from upgrade

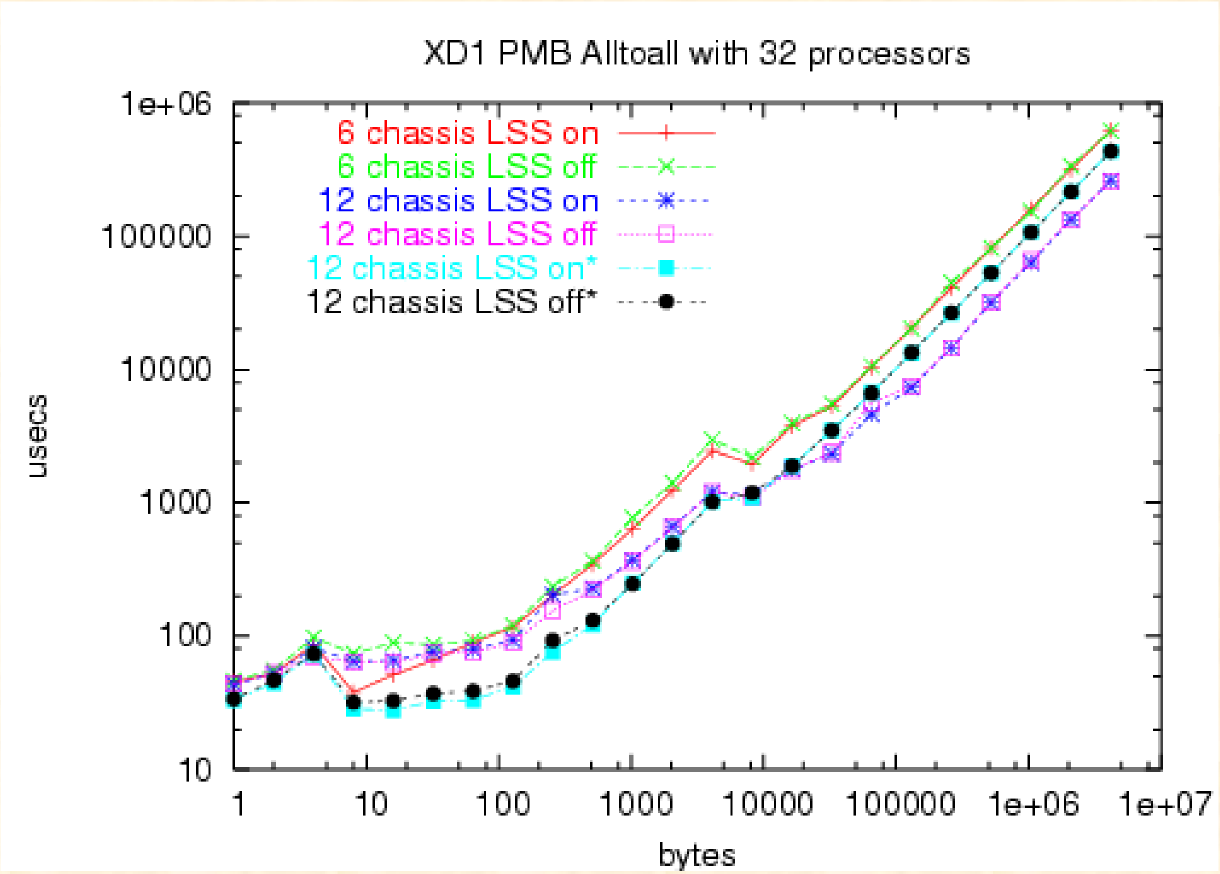
Other platforms left out

* Carefully chose nodes that were in pairs of chassis with extra links

Result of bandwidth reduction going to 12 chassis system



PMB alltoall again



Summary

- **Performs comparably well**
- **LSS is a good thing**
- **Two fabrics are better than one**
  - **When the expansion fabric is reliable**
- **Bandwidth reduction when upgrading 2 6-chassis system to 1 12-chassis system**
  - **Bandwidth limited codes suffer greatly**
  - **6 and 12 chassis systems have extra links**
- **More tests needed to understand 12 chassis system performance**

Questions

- **Questions? Comments?**
- **For more information on these studies, see**
  - **http://www.csm.ornl.gov/evaluation**
- **faheymr@ornl.gov**

Extra Slides

- **PSTSWM kernel**

PSTSWM

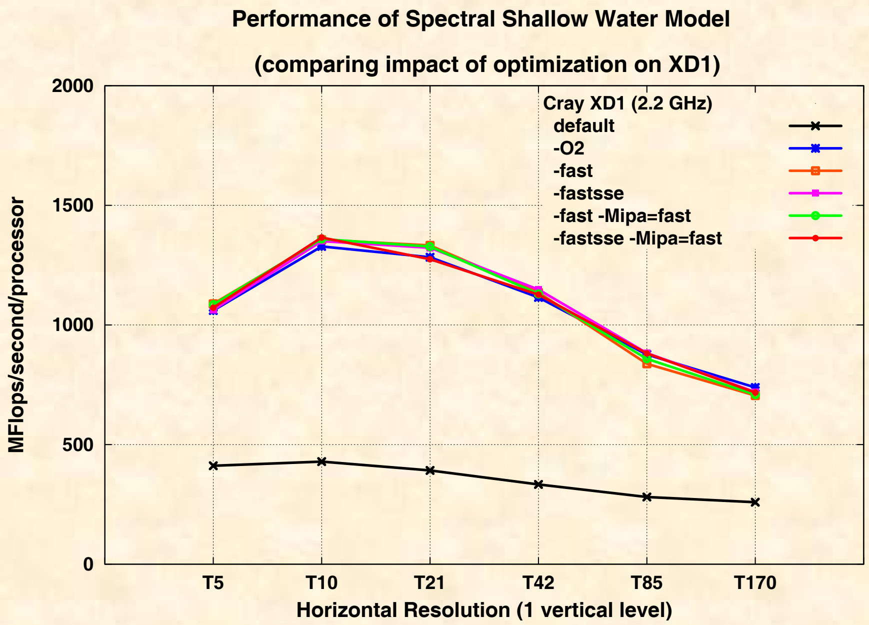
- **Parallel Spectral Transform Shallow Water Model**
  - **Important computational kernel in spectral global atmospheric models**
  - **99% of fp operations are multiply or add**
  - **Exhibits little resuse of operands**
    - **Exercises memory subsystem as problem size is scaled**
    - **Can be used to evaluate impact of memory contention**

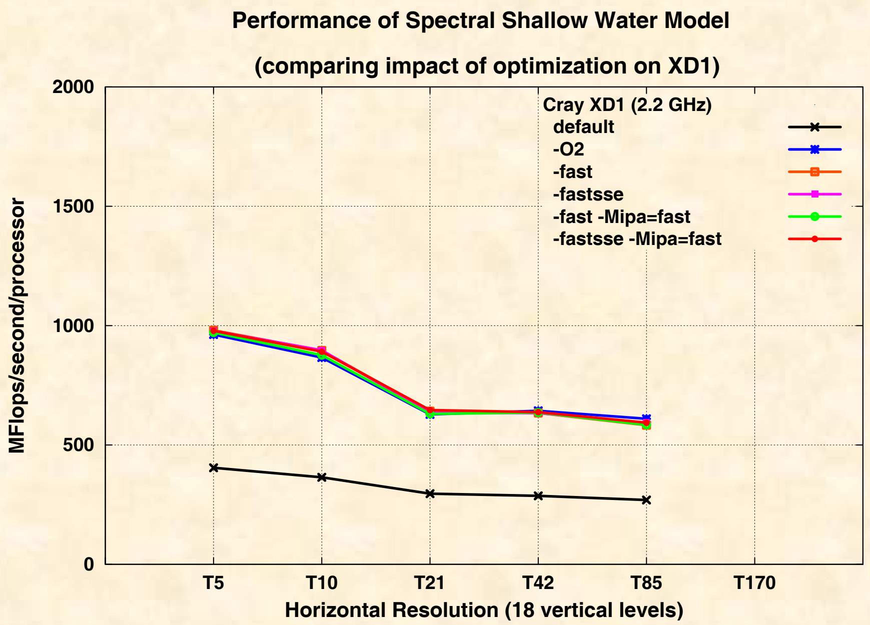
PSTSWM (cont.)

These experiments examine serial performance, both using one processor and running the serial benchmark on multiple processors simultaneously. Performance is measured for a range of horizontal problems resolutions for 1 and 18 vertical levels

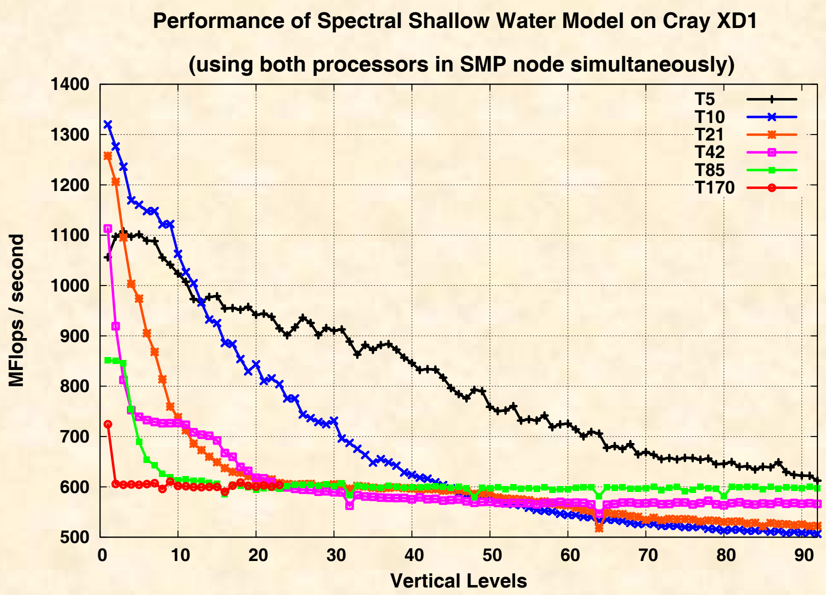
Problem	Horizontal Resolutions
T5	8x16
T10	16x32
T21	32x64
T42	64x128
T85	128x256
T170	256x512

PSTSWM (cont.)

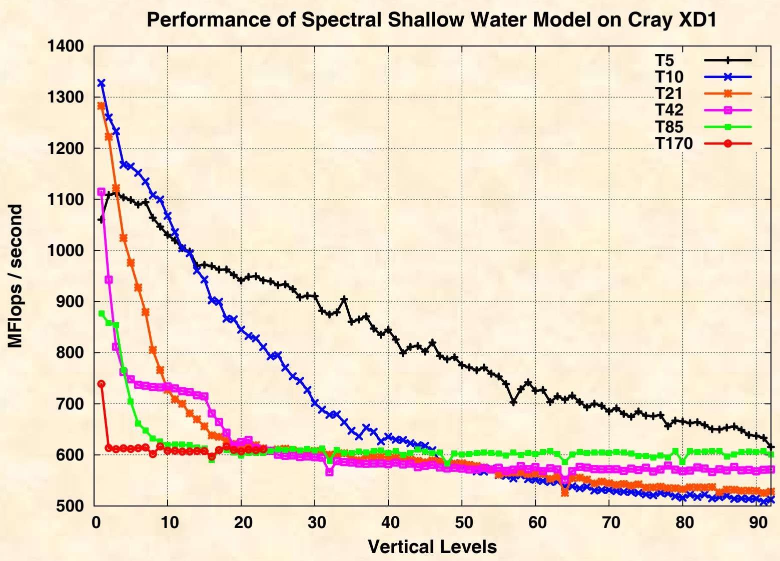

Performance of Spectral Shallow Water Model
(comparing impact of optimization on XD1)
Cray XD1 (2.2 GHz)
default
-O2
-fast
-fastsse
-fast -Mipa=fast
-fastsse -Mipa=fast
MFlops/second/processor
2000
1500
1000
500
0
T5
T10
T21
T42
T85
T170
Horizontal Resolution (1 vertical level)


Performance of Spectral Shallow Water Model
(comparing impact of optimization on XD1)
Cray XD1 (2.2 GHz)
default
-O2
-fast
-fastsse
-fast -Mipa=fast
-fastsse -Mipa=fast
MFlops/second/processor
2000
1500
1000
500
0
T5
T10
T21
T42
T85
T170
Horizontal Resolution (18 vertical levels)

PSTSWM (cont.)



Performance of Spectral Shallow Water Model on Cray XD1
(using both processors in SMP node simultaneously)
MFlops / second
Vertical Levels
T5
T10
T21
T42
T85
T170



Performance of Spectral Shallow Water Model on Cray XD1
MFlops / second
Vertical Levels
T5
T10
T21
T42
T85
T170