
Comparing Linux Clusters for the Community Climate System Model

Matthew Woitaszek

Michael Oberg

Henry Tufo



University of Colorado, Boulder
20 May 2004

Outline

- Introduction
 - Weather and climate on clusters
 - CCSM component models

- Design: Meet the machines

- Results
 - Platform Tests
 - Interconnect Tests

- Conclusions

Weather and Climate on Clusters

- ❑ Mesoscale meteorology uses clusters
 - ❑ NCAR/PSU MM5 ported to clusters in 1999 by Dorband and Michalakes
 - ❑ Clusters provide the best cost performance

- ❑ Community Climate System Model (CCSM)
 - ❑ Initial port to ANL Jazz by John Taylor in 2003
 - ❑ Cluster support depends on MPI implementation and requires significant environment editing
 - ❑ Sporadic reports of CCSM on clusters

The Community Climate System Model

- ❑ Four models and a flux coupler
 - ❑ Atmosphere model: Community Atmosphere Model (CAM)
 - ❑ Land model
 - ❑ Sea-ice model
 - ❑ Ocean model: Parallel Ocean Program (POP)

- ❑ We selected POP and CAM for our tests
 - ❑ Well-represented in the literature
 - ❑ Michael was able to get both to work

POP and CAM

- ❑ POP: The Parallel Ocean Program
 - ❑ From Los Alamos National Laboratory
 - ❑ Resolution 320x384x40: 1 degree, 40 levels
Currently used in CCSM
 - ❑ Resolution 640x768x40: ½ degree, 40 levels
Suggested for future high-resolution use

- ❑ CAM: The Community Atmospheric Model
 - ❑ Developed by NCAR and collaborators
 - ❑ T42L26: (128x64x26)

Platforms: Intel Xeon Clusters

	Xeon 2.4 Dual/Dolphin	Xeon 2.4 Single/Myrinet	Xeon 3.06 Dual/Infiniband
Nodes	64	350	48
CPU	Dual 2.4 GHz	Single 2.4 GHz	Dual 3.06 GHz
RAM/processor	1 GB	1 GB (selected)	1 GB
Interconnect	Dolphin	Myrinet PCIxD	Infiniband

- ❑ The first cluster is our cluster at the University of Colorado
- ❑ The second cluster is ANL Jazz
- ❑ The third cluster was provided by a vendor
- ❑ All have 512KB L2 cache, 3.2 GB/s memory bandwidth

Platforms: AMD Opteron Clusters

	Opteron 2.0 / Myrinet (A)	Opteron 2.0 / Myrinet (B)
Nodes	32	64+
CPU	Dual Opteron 2.0 GHz	Dual Opteron 2.0 GHz
RAM/processor	2 GB	2 GB
Interconnect	Myrinet PCIxD	Myrinet PCIxD
Mode	Native (64-bit)	Legacy (32-bit)

- ❑ Access to both clusters were provided by vendors
- ❑ Opterons have 1MB L2 cache

Platforms: Larger Computers

	IBM p690	SGI Altix 3700 (1300 MHz)	SGI Altix 3700 (1500 MHz)
CPUs	64 (used 8-way)	64	32
CPU	Power4 1.3 GHz	Itanium2 1.3 GHz	Itanium2 1.5 GHz
Interconnect	Colony: SP Switch2	SGI NUMALink3	SGI NUMALink3

- ❑ NCAR provided access to the p690 supercomputer
- ❑ SGI performed the benchmarking on the Altix machines using our software packages

Experimental Design

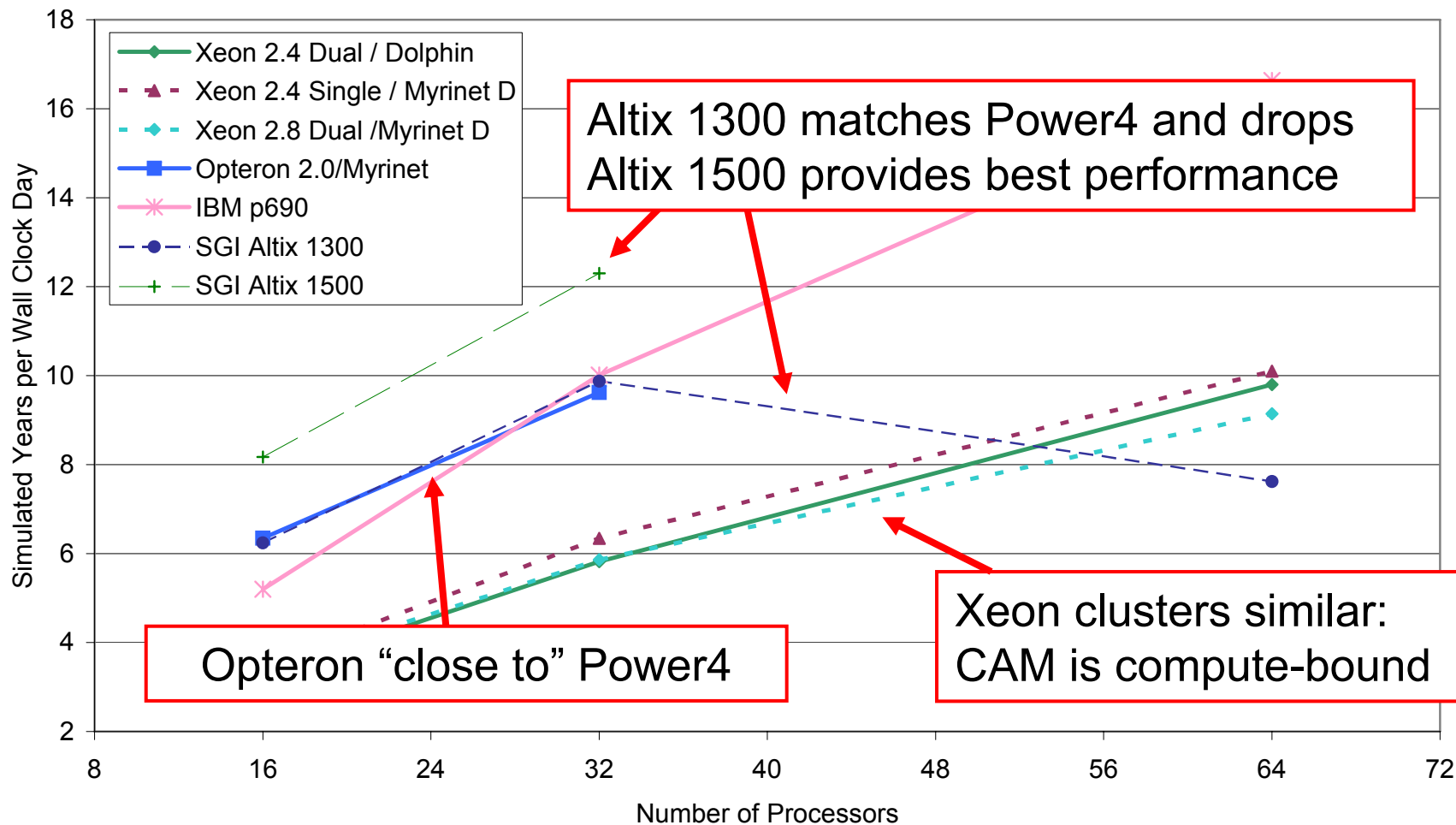
- ❑ Run on as many clusters as possible:
 - ❑ POP 320x384, POP 640x768, CAM T42L26
- ❑ Record times as reported by internal timers to avoid mpirun/poe startup overhead
- ❑ Repeat for averages and confidence intervals

Excessive Detail Orientation vs.
Grant-Funded and Courtesy Time Allocations

- ❑ Not all test variants were run on all machines
 - ❑ Consistent tests for platform comparisons
 - ❑ Additional tests on the CU Xeon cluster

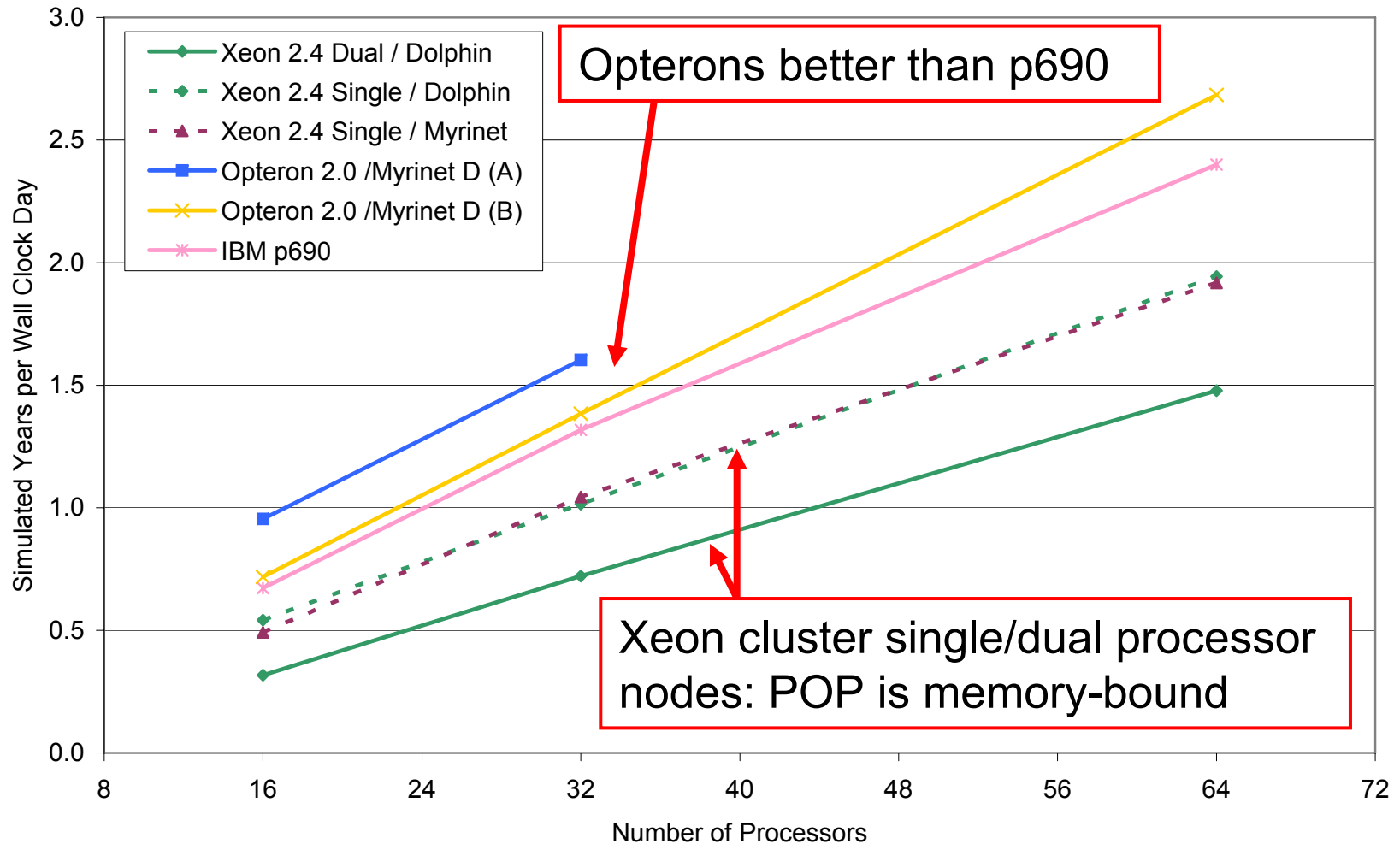
Platform Tests for CAM T42

CAM T42 Simulated Years per Wall Clock Day by Number of Processors



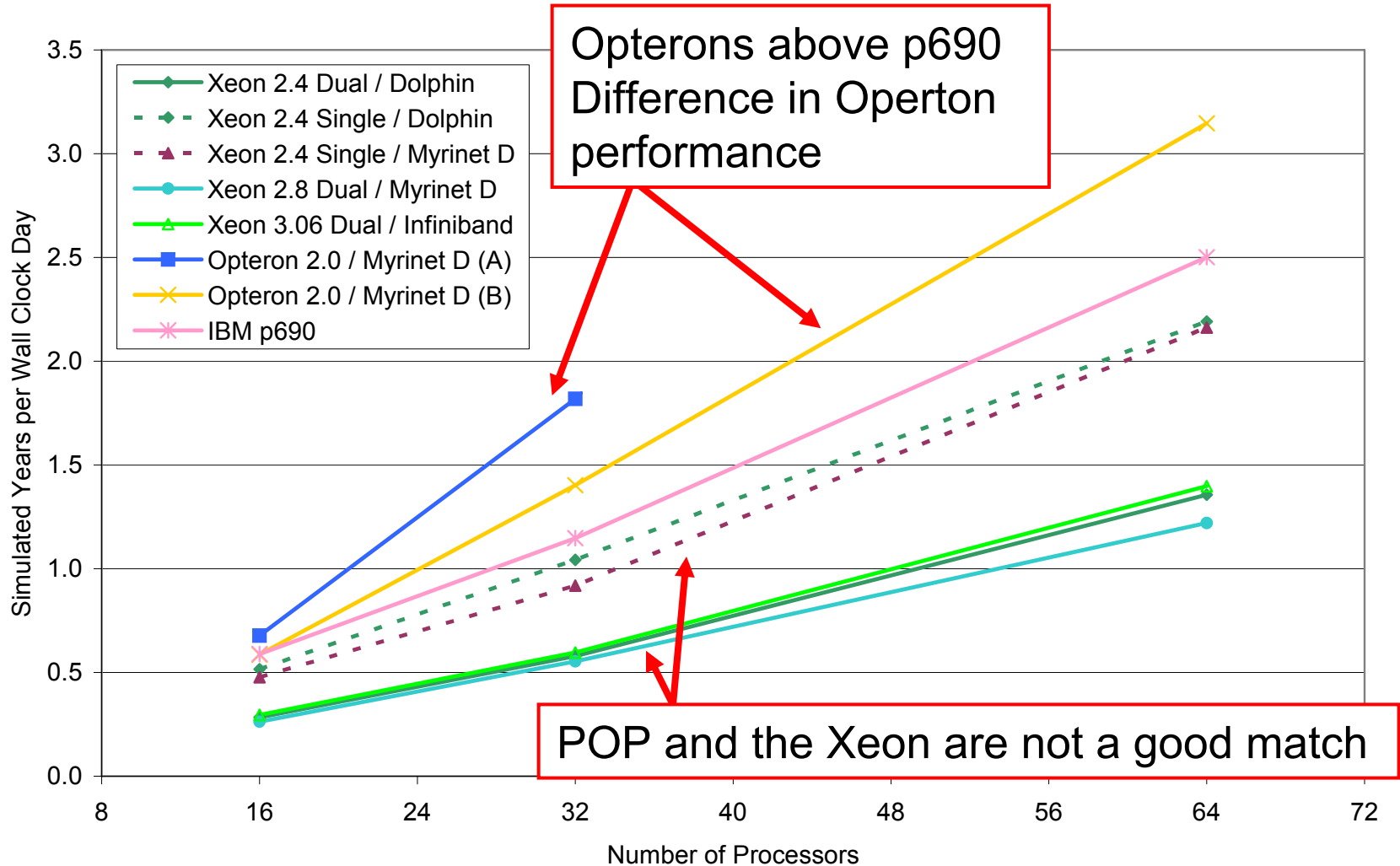
Platform Tests for POP 320x384

POP 320x384 Simulated Years per Day by Number of Processors



Platform Tests for POP 640x768

POP 640x768 Simulated Years per Day by Number of Processors



Interconnect Comparison

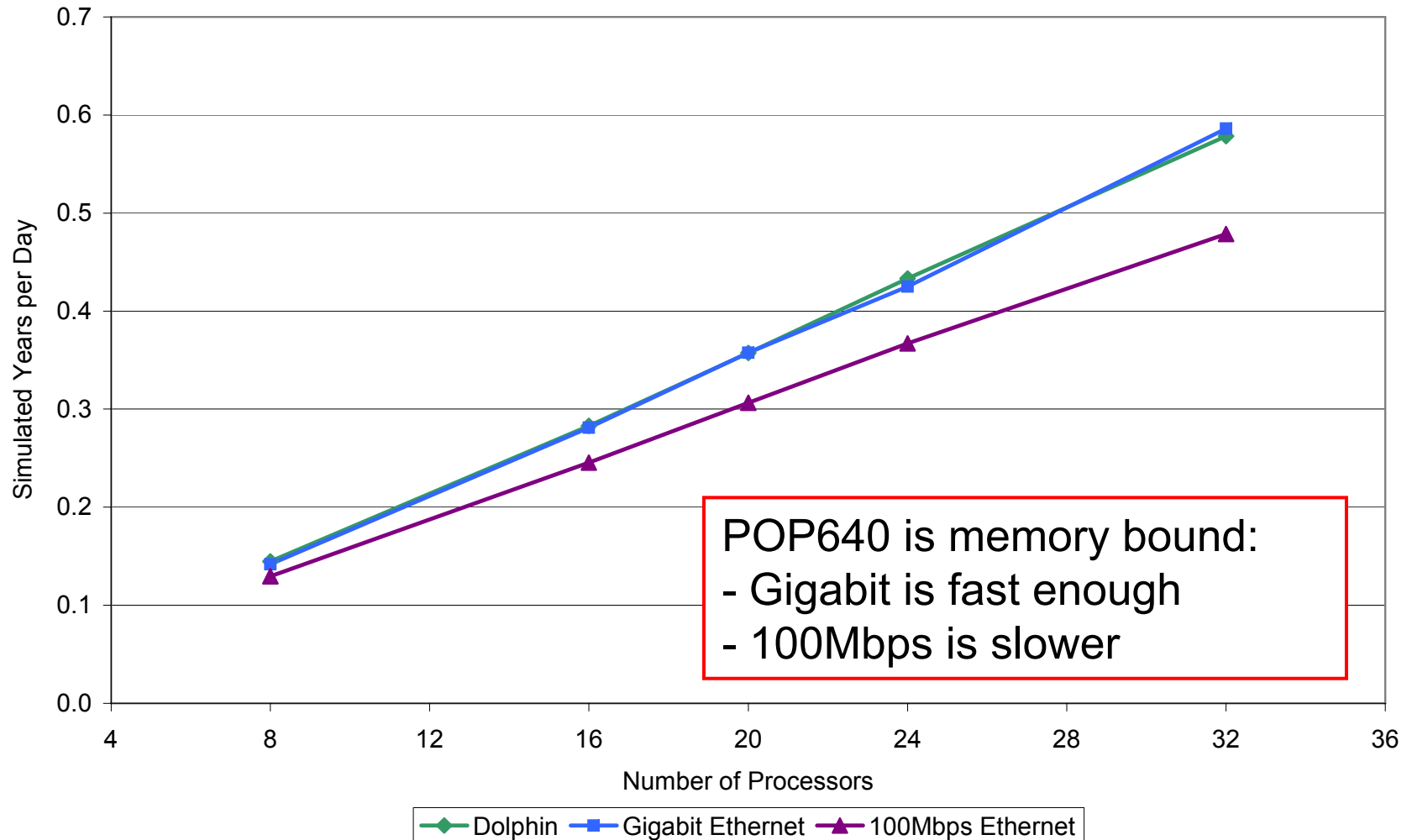
- ❑ A favorite question when building a cluster:

Do we really need Myrinet, Dolphin, or Infiniband, or will gigabit Ethernet work for our applications?

- ❑ Vital statistics for Myrinet, Dolphin, and Infiniband are frequently cited and similar
 - ❑ 250 MB/s, 4-6 μ s
 - ❑ But will gigabit Ethernet work just fine?
- ❑ We ran POP and CAM tests on a Xeon cluster:
 - ❑ Dolphin
 - ❑ Gigabit Ethernet and 100MBps Ethernet

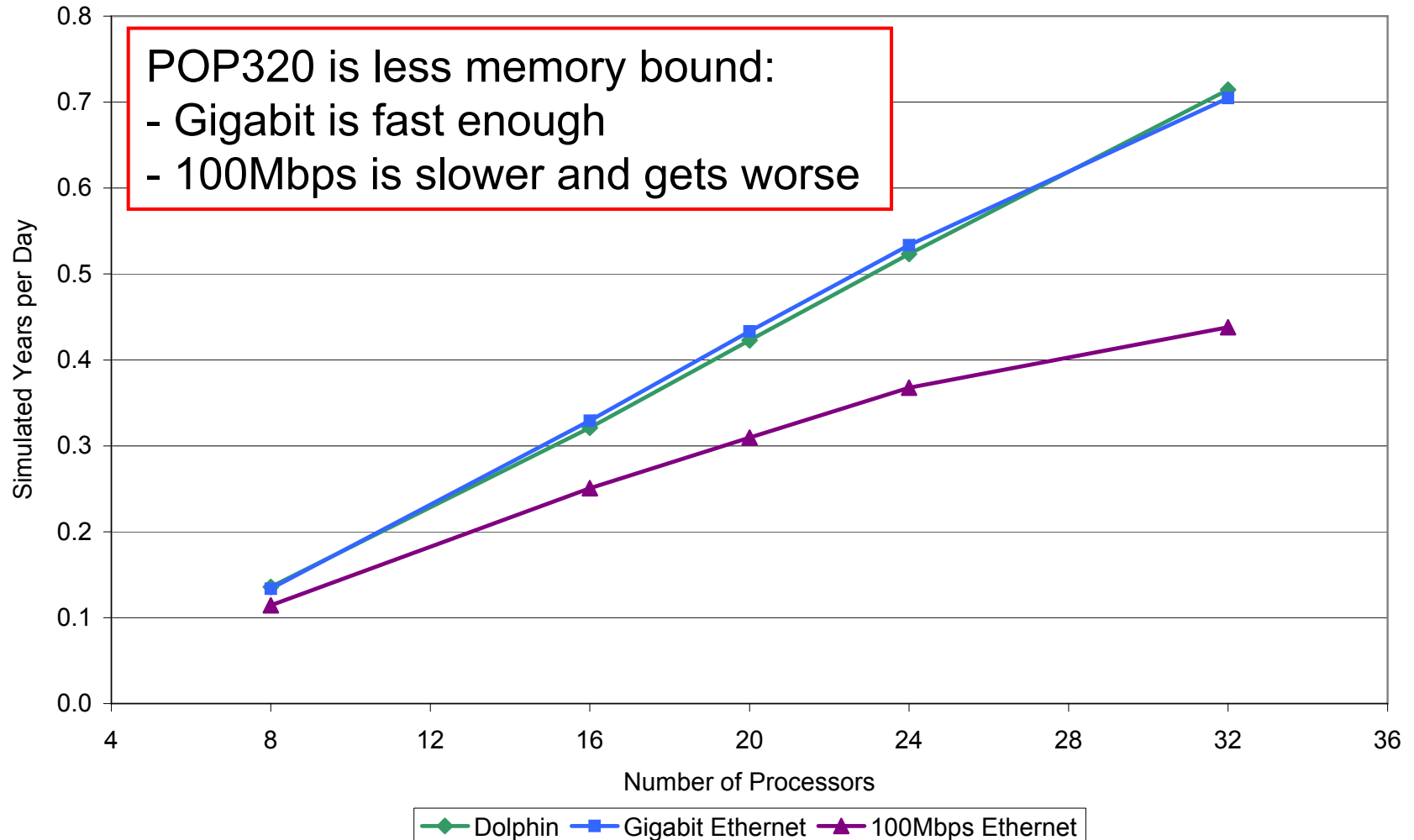
Interconnect Comparison for POP 640x768

POP 640x768 Simulated Years per Wall Clock Day by Number of Processors



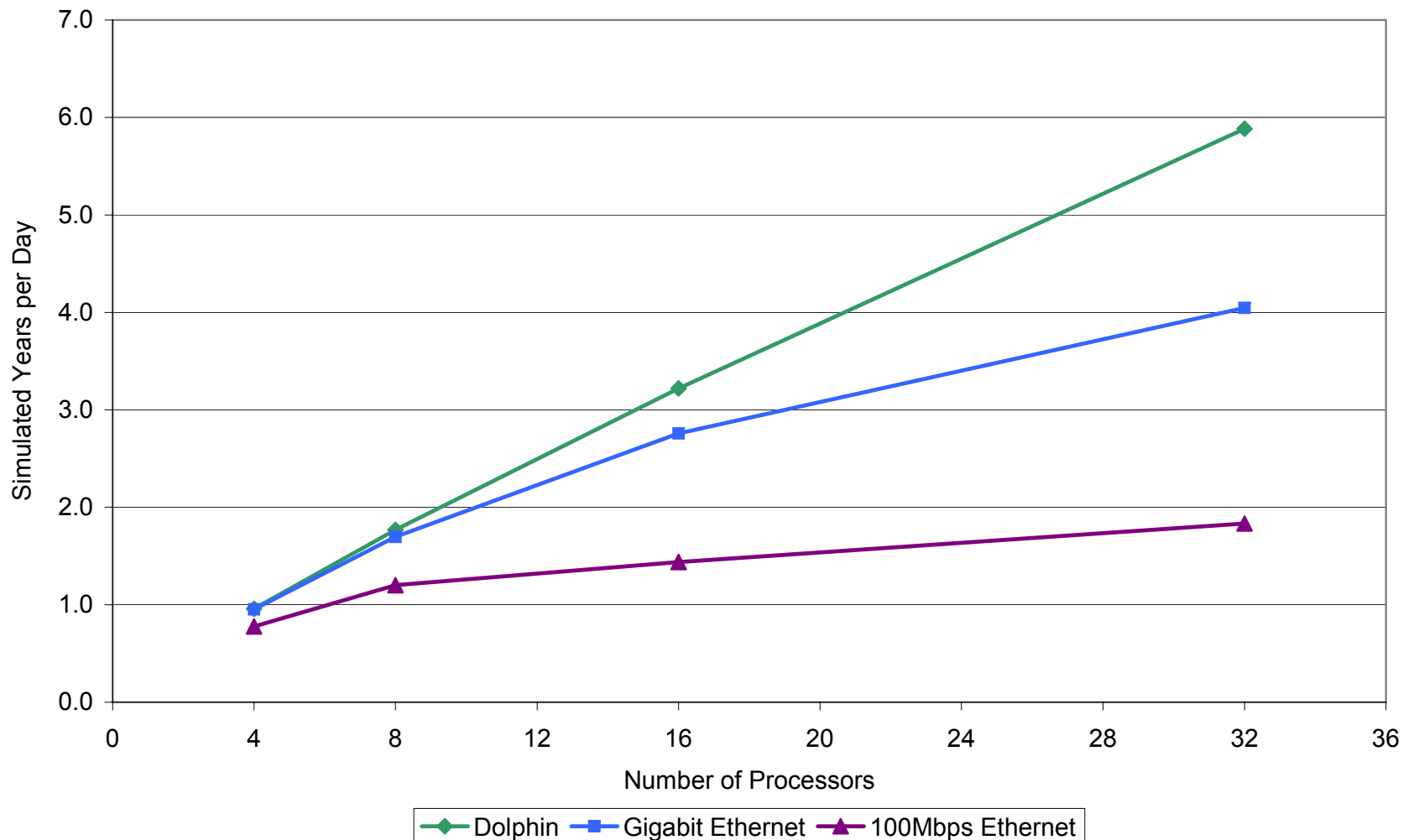
Interconnect Comparison for POP 320x384

POP 320x384 Simulated Years per Wall Clock Day by Number of Processors



Interconnect Comparison for CAM T42

CAM T42 Simulated Years per Wall Clock Day by Number of Processors



Conclusions

□ Platforms

- The AMD Opteron clusters provide performance that rivals the IBM p690 supercomputer
- The Xeon clusters appear to be limited by the memory architecture

□ Interconnect

- A high performance interconnect is probably best

□ Integrated models are rate limited and the components will interact

Future Work

- ❑ Enhanced CCSM support for clusters
 - ❑ Streamlined build and batch system integration
 - ❑ Internal model tuning for processor architecture and cluster interconnects

- ❑ This future work may actually happen!
 - ❑ NCAR is purchasing several large Opteron clusters
 - ❑ Michael Oberg recently interviewed for a job with NCAR's the High Performance Systems Section

Thank you!

Questions?

matthew.woitaszek@colorado.edu

michael.oberg@colorado.edu

tufu@cs.colorado.edu



<http://hemisphere.cs.colorado.edu>

Emergency Slide: Opteron Memory Benefits

- ❑ Integrated memory controller

- ❑ Three HyperTransport Links
 - ❑ 3.2 GB/s in each direction

- ❑ SMPs connected using HyperTransport
 - ❑ Point-to-point connection
 - ❑ Up to 8-way SMP
 - ❑ Memory access is essentially uniform: “DRAM hit” vs. “DRAM conflict” according to AMD