Productive Supercomputing with Cray Inc.

John M. Levesque

Director

Cray Supercomputing Center of Excellence

Cray's Family of Supercomputers

Cray X1E

- 1 – 150 TFLOPS
- 8 – 8,000+ processors
- Vector processing for uncompromised sustained performance

Cray XT3

- 1 – 100+ TFLOPS
- 256 – 30,000+ processors
- MPP Compute system designed for large-scale sustained performance
- Based on Sandia "Red Storm" collaboration

Cray XD1

- 50 GFLOPS - 2+ TFLOPS
- 12 – 2,000+ processors
- Mid-range system optimized for sustained performance
- With reconfigurable computing capability

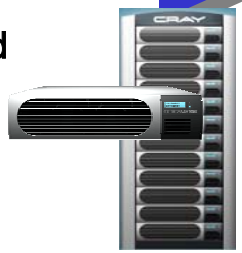
Purpose-Built High Performance Computers

The Cray Roadmap

Purpose-Built HPC Systems

Specialized HPC Systems

HPC Optimized Systems



BlackWidow 2007

Cascade
"Productive Computing"

2007 ***Eldorado***

Rainier

Cray X1E 2005

2008 ***Baker***

Cray XT3 Dual Core 2006

2007 ***Adams***

Cray XT3 2005

2005 **Cray XD1 Dual Core**

Performance
Programmability
Robustness

The Path to Productive Computing

The Cray X1E Supercomputer

Cray X1E

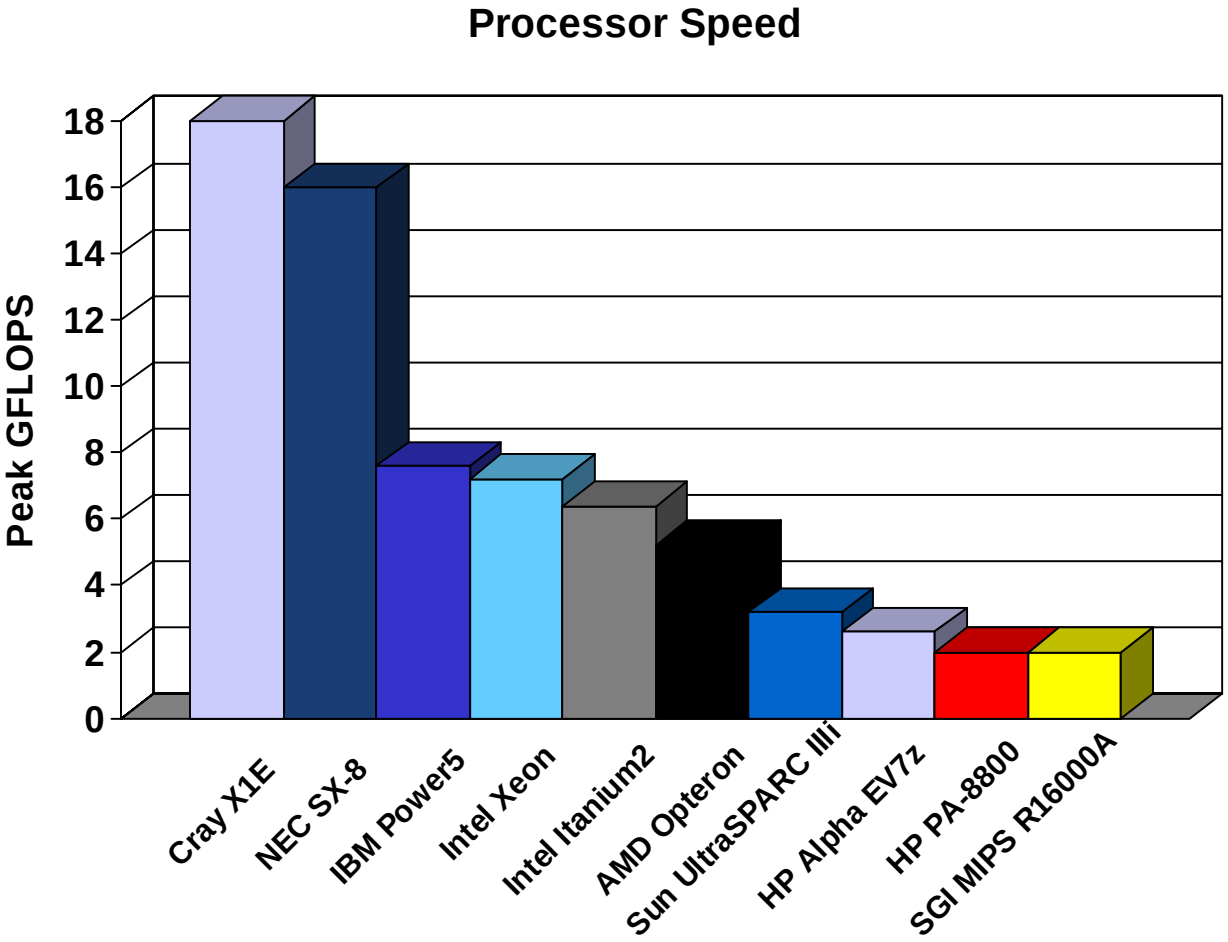
- **Designed for high sustained application performance**
- **Powerful vector processors**
  - 18+ GFLOPS peak
    - 34 GB/s memory BW per cpu
  - Automatic vectorization and streaming
- **Highly-scalable network interconnect**
  - Low latency, high bandwidth
    - 51 GB/s per compute module
  - Any processor can access any memory
  - System-wide, global I/O
  - Scales to 8192 cpu's & 150TF
- **True single system image**
  - Effective system management
  - Fast upgrades and reboots
  - Mature software environment

Purpose-built and optimized for high performance workloads

Cray X1E: The World's Fastest Processors

Cray X1E 4 dual-core processors per board: 144 GFLOPS

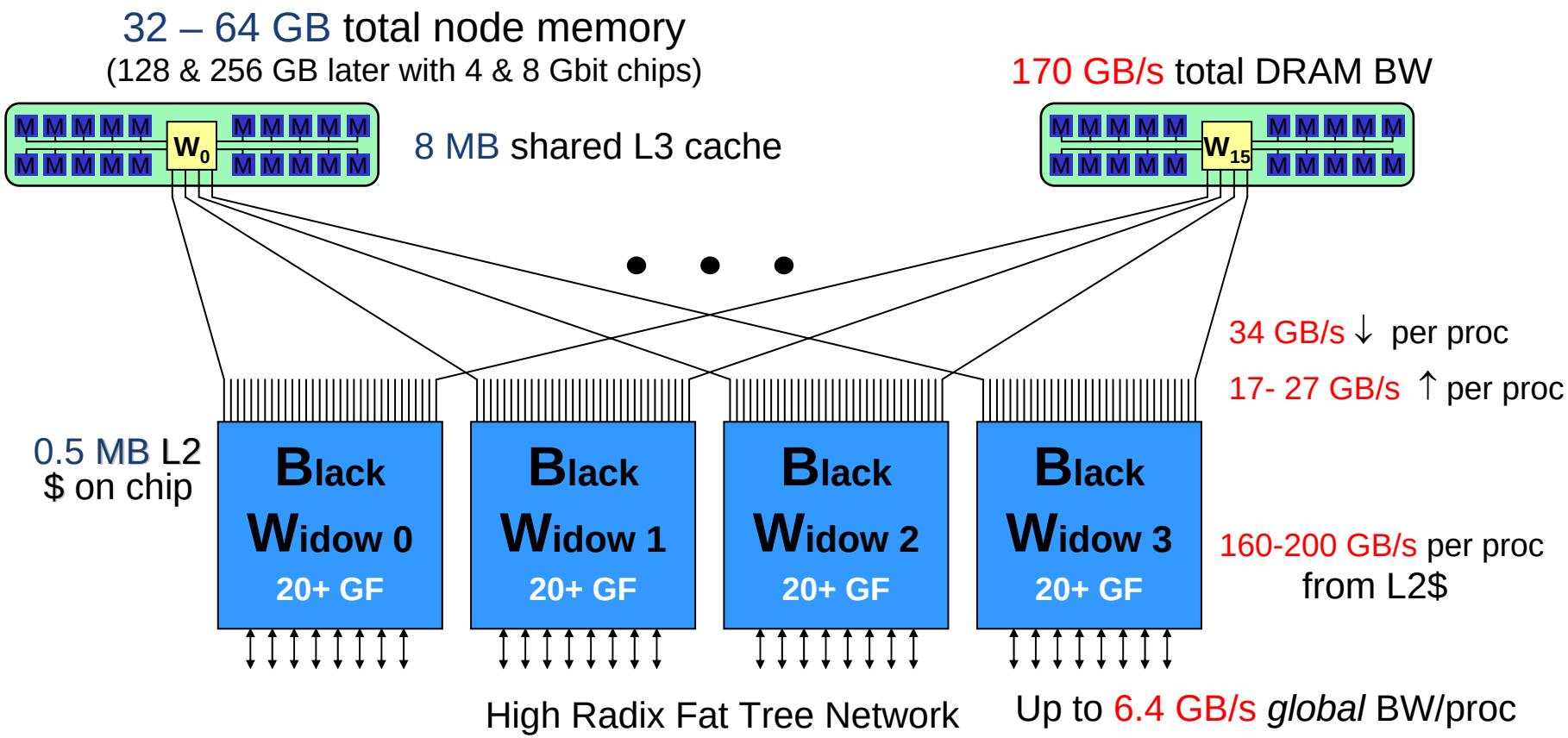
(billions of calculations/second)



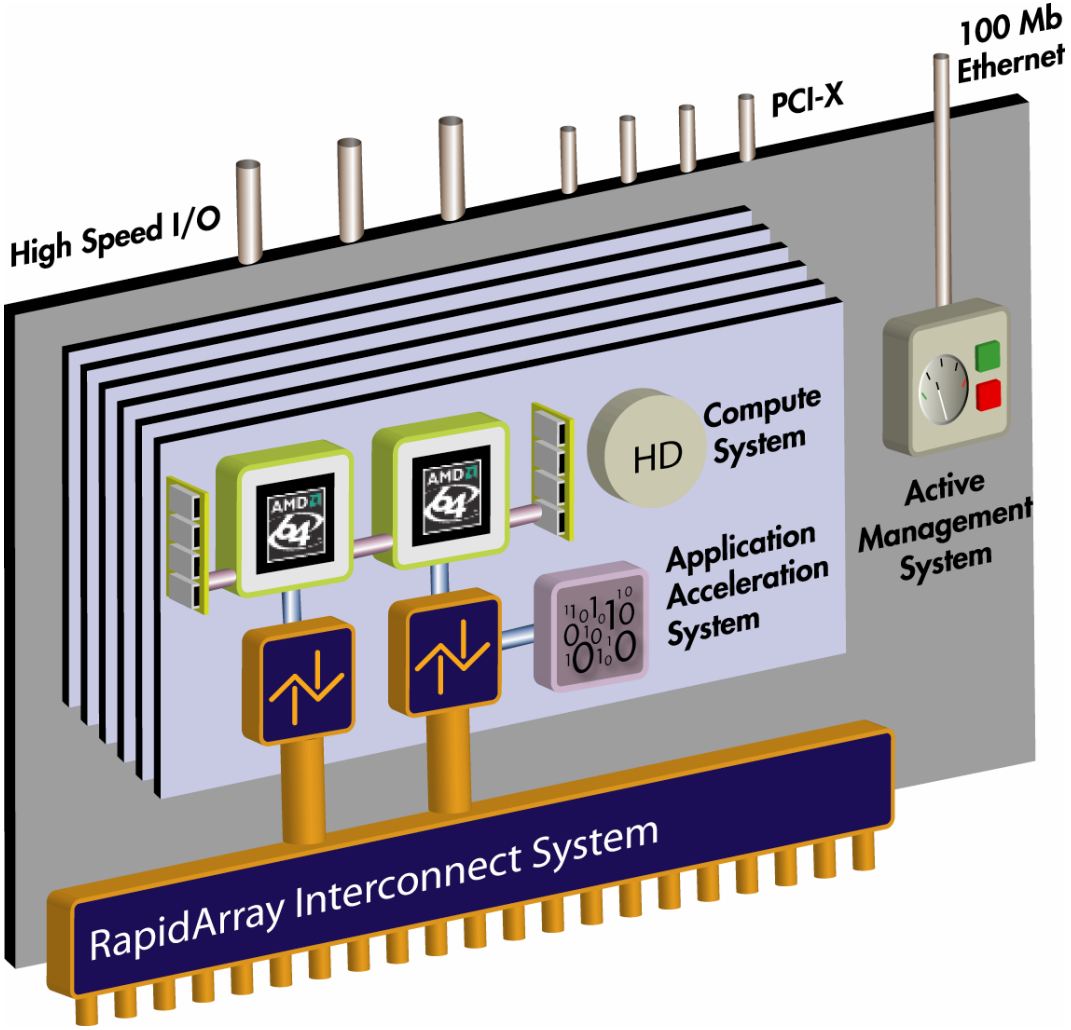
The World's Fastest Processor for the World's Toughest Problems

Cray BlackWidow Node

- **Second generation Vector MPP, shipping 1H 2007**
- **Major improvement in single thread scalar performance and price-performance**
- **Globally addressable memory with 4-way SMP nodes**
- **Two 4-way nodes per BlackWidow blade**



Cray XD1 System Architecture



Compute Blades

- 12 AMD Opteron 32/64 bit, x86 processors
- High Performance Linux
- Scaling to over 2,000 cpu's

RapidArray Interconnect

- 12 communications processors
- 1 Tb/s switch fabric

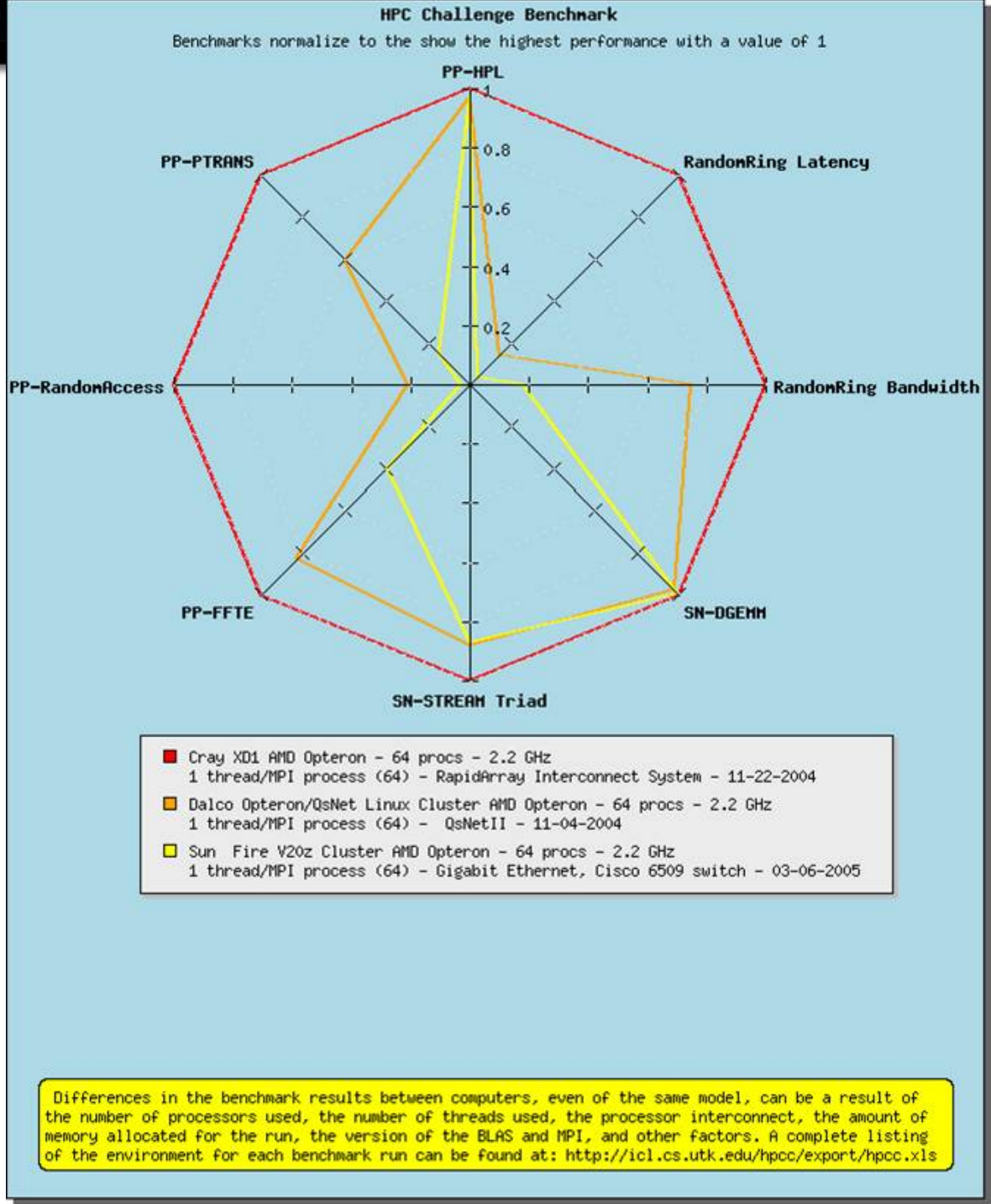
Active Management

- Dedicated processor

Application Acceleration

- 6 FPGA co-processors
- Reconfigurable computing
- Superlinear speed-up possibilities

Processors directly connected via integrated switch fabric



Source:
Jack Dongarra, University of Tennessee
March 2005

Rainier : The Move to Common Technologies

- **Common Technologies via XT3/Adams Front-end**
  - **Common service and I/O infrastructure**
  - **Common logon**
  - **Common job management**

- **One size doesn't fit all…**
  - **BlackWidow follow-on for vector**
  - **Adams follow-on for scalar**
  - **Tighter integration with Baker/Cascade**

- **Benefits**
  - **Leverage a few key technologies across products**
  - **Single solution for diverse workloads**
  - **Single user interface and environment**
  - **Improved performance and productivity by matching processor to the job**

Cascade Project

Cray Inc.
Stanford
Caltech/JPL
Notre Dame

Jet Propulsion Laboratory
California Institute of Technology

UNIVERSITY OF NOTRE DAME

Pathway to Petaflops – Cascade

- Advanced Research Program
  - Goal of a “trans-petaflops system”
  - Robust, easier to program, more broadly applicable
  - Focus on “time to solution”
- Phase I – Concept Development
  - Started in June 2002 for one year
  - Five total vendors
  - University partners
- Phase II – Concept Validation
  - Recent award of $49.9M to Cray
  - Two other vendors
  - Three-year contracts
- Phase III planned in 2006 – Product Development

"The cycle time from when engineers have an idea to when they have a program ready to run is one of the bottlenecks in high-end computers, and it will only become worse as we develop bigger and bigger machines,“

- Robert Graybill, HPCS Program

UNIVERSITY OF NOTRE DAME

JPL Jet Propulsion Laboratory California Institute of Technology

LELAND STANFORD JUNIOR UNIVERSITY ORGANIZED 1891

The Cascade Approach

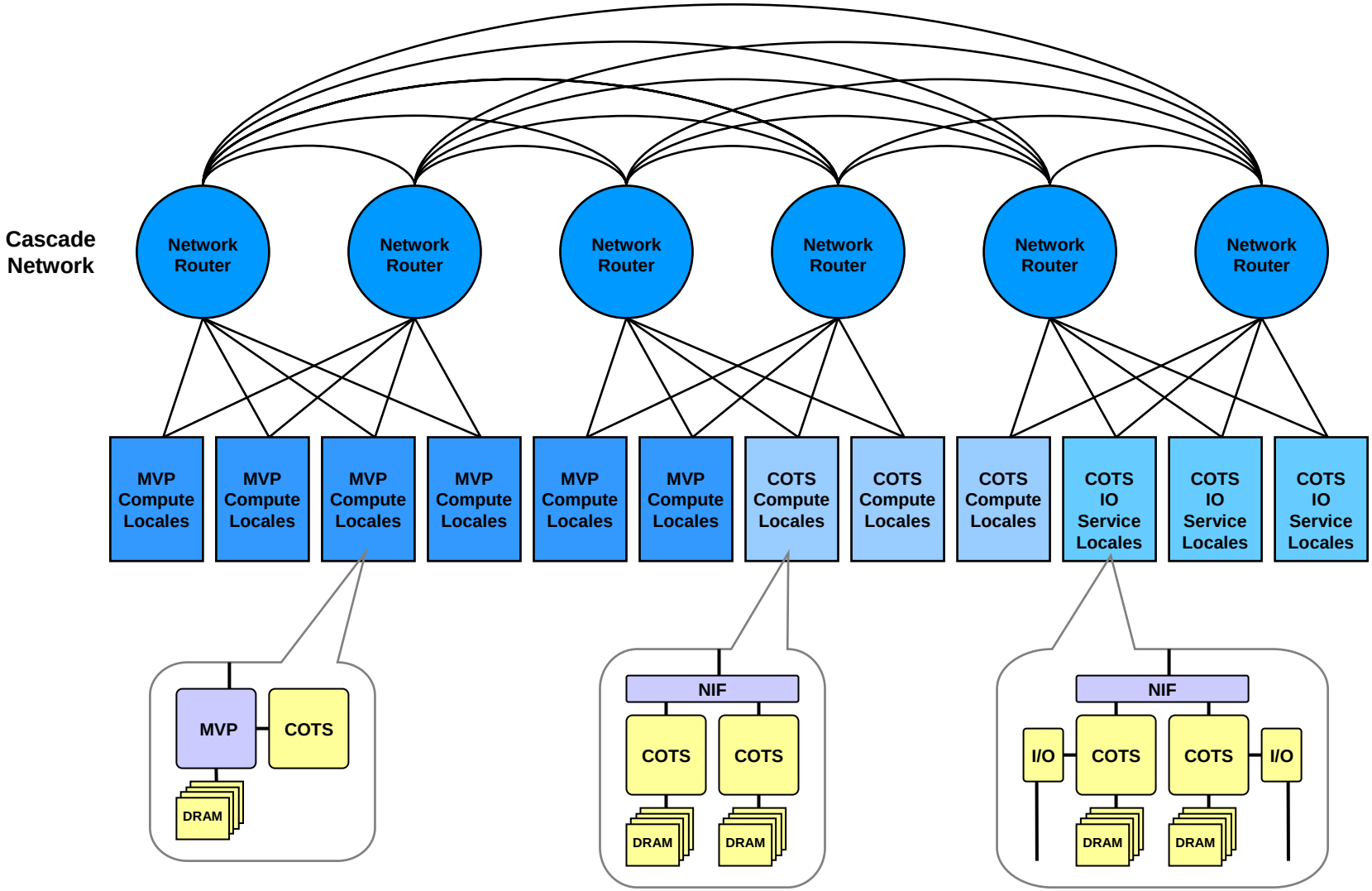
- **Ease the *development* of parallel codes**
  - **Support shmem(), UPC and CAF**
  - **Develop the Chapel high productivity language**
- **Provide programming tools to ease *debugging* and *tuning***
  - **Automatic performance analysis (expert system)**
- **Design a heterogeneous, tailorable machine that can match the attributes of a wide variety of applications:**
    - **Thread level parallelism**
    - **Data-level parallelism**
    - **Regular and sparse bandwidth of varying intensities**
    - **Spatial and temporal locality**

  ⇒ **Increases performance**

  ⇒ **Significantly eases programming**

- **Provide robustness at scale**
  - **Hardware features for fault tolerance**
  - **Omak transactional operating system for resiliency**

Conceptual HW Configuration



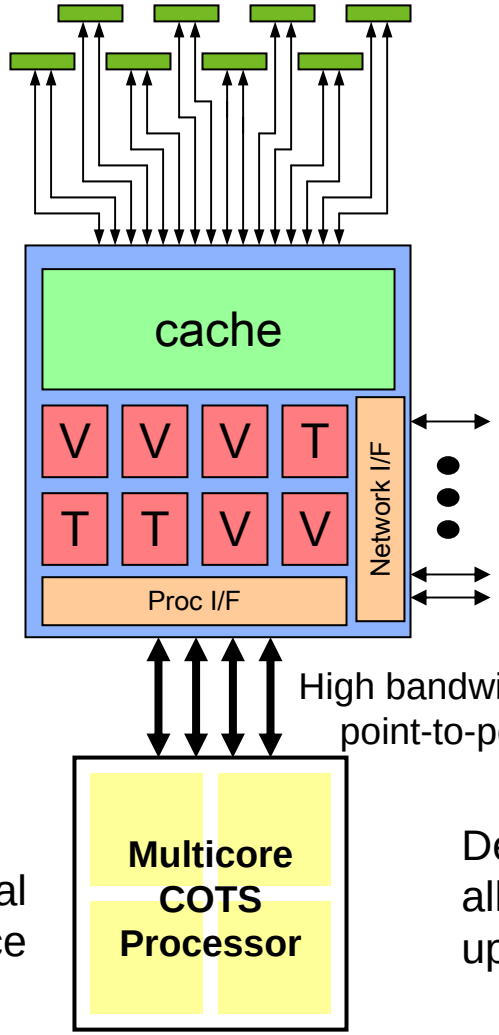
Heterogeneous Computing

Support both commodity and custom DIMMs

Memory and network provide strong support for sparse accesses

Multiple cores, can dynamically switch between *vector* or *multithreaded* mode, providing acceleration for DLP or TLP workloads.

First generation vector only.



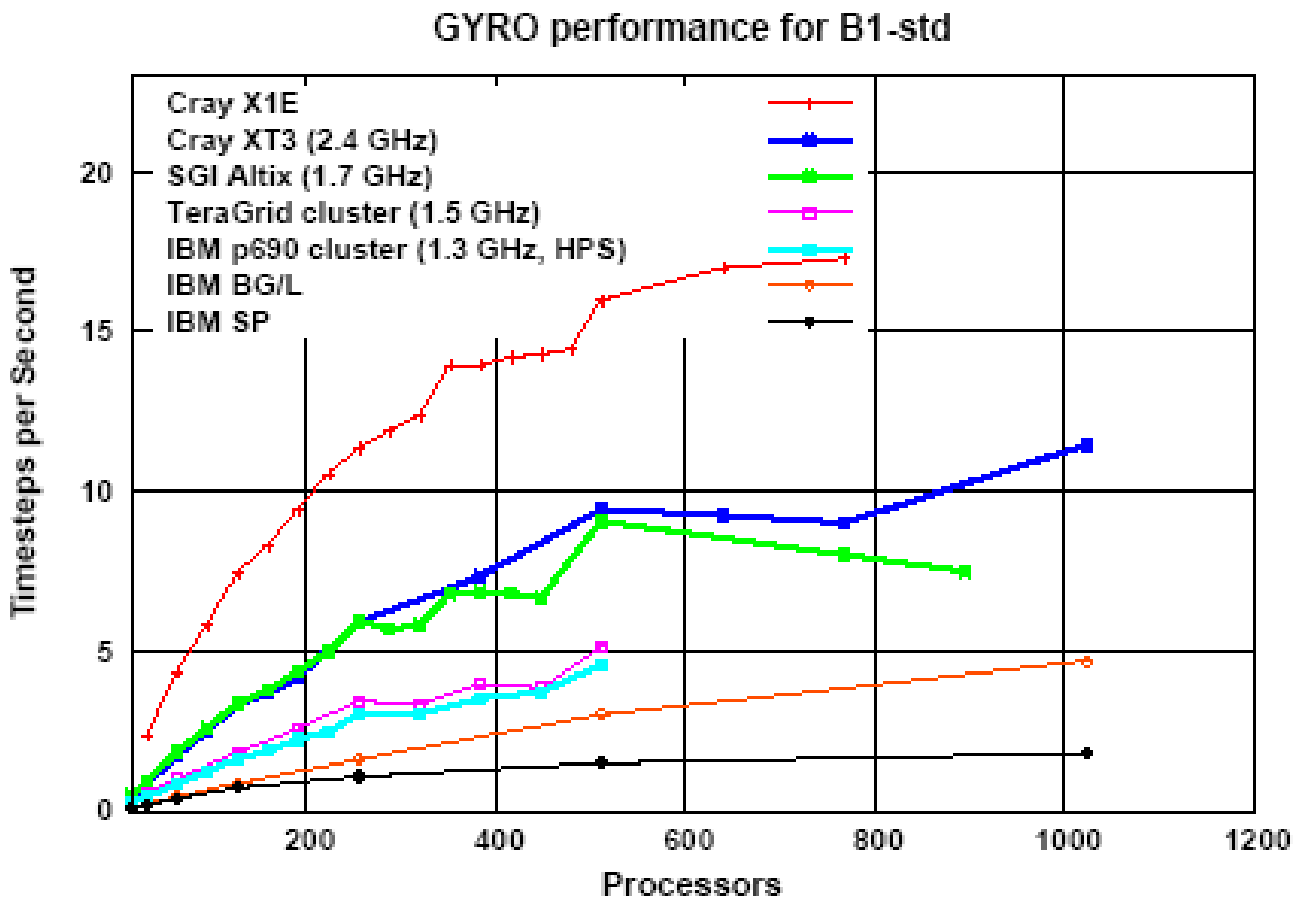
Flexible interconnect allows *very* high global bandwidth

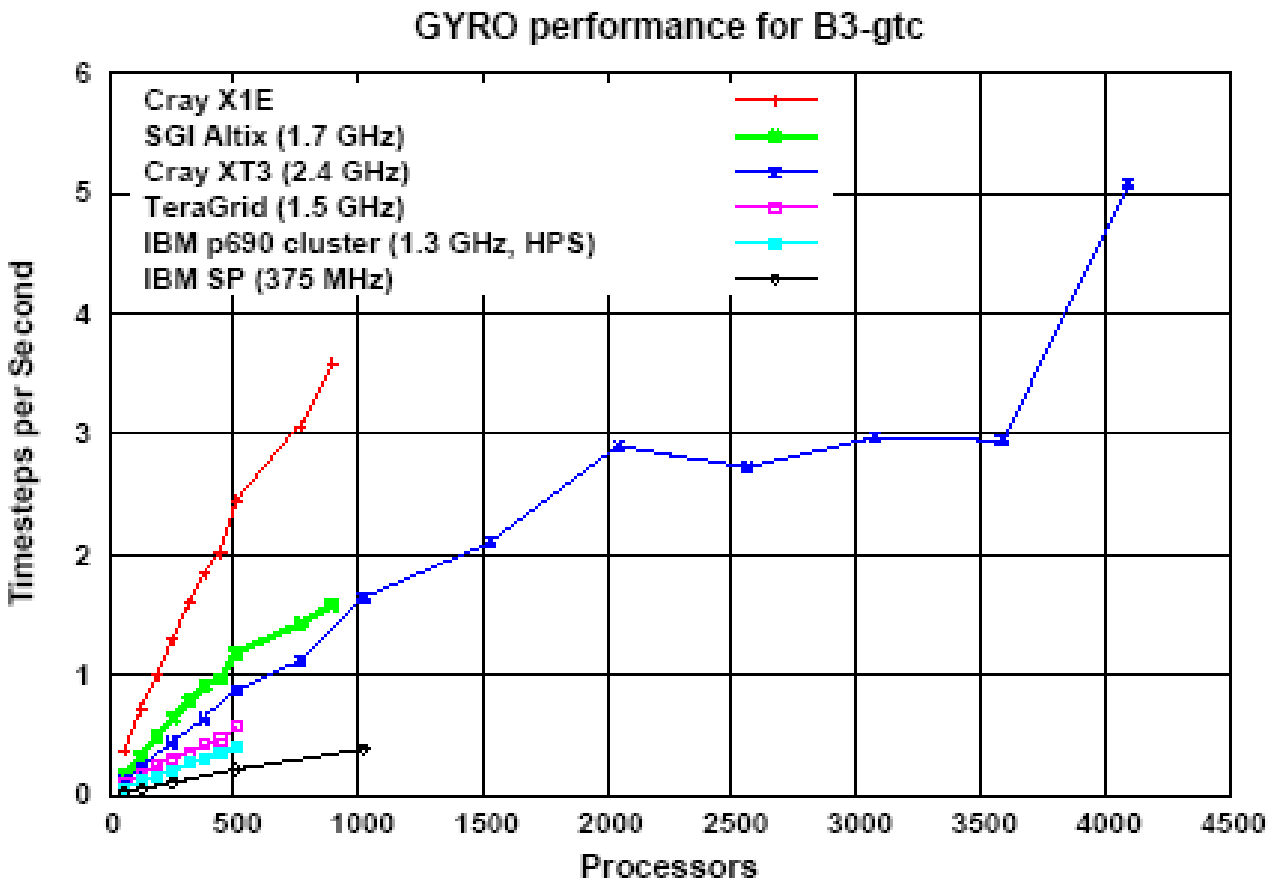
High bandwidth, point-to-point links

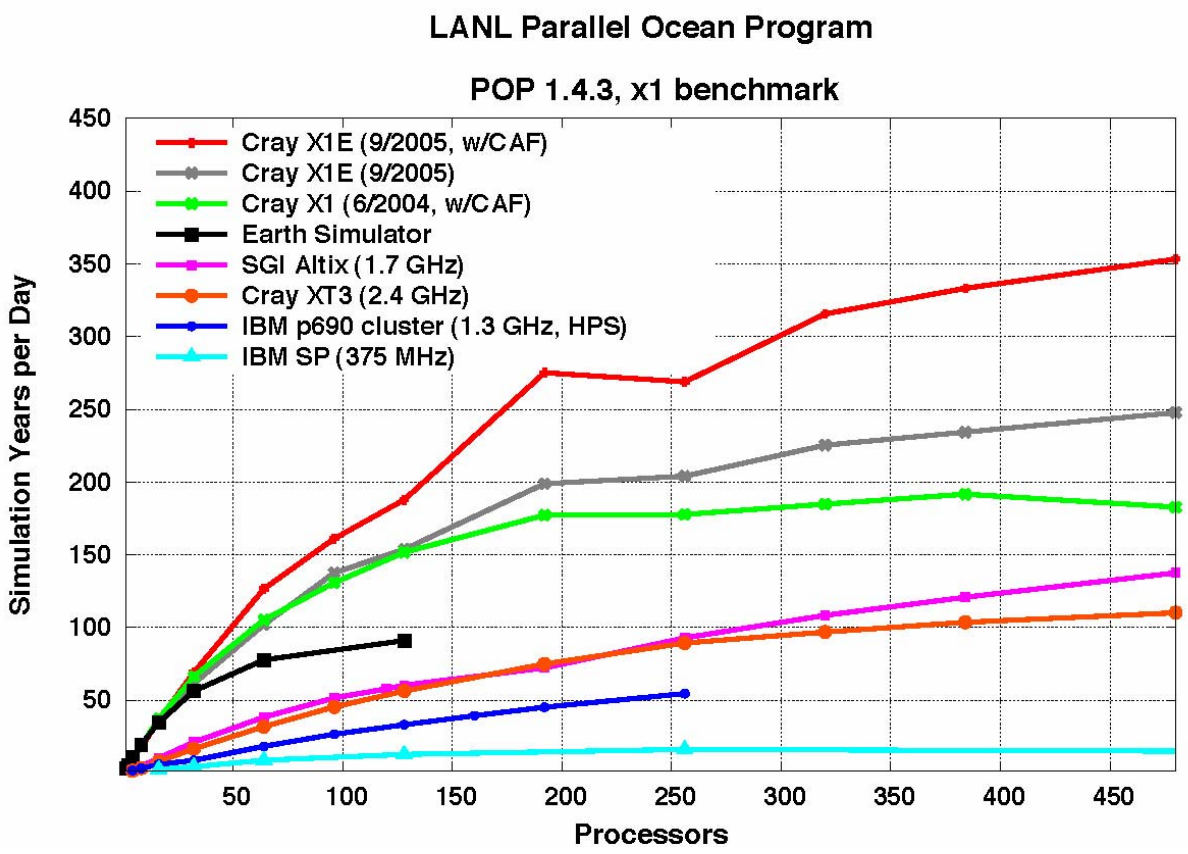
Best of class serial scalar performance

Decoupled computation model allows COTS processor to queue up work to be consumed later.

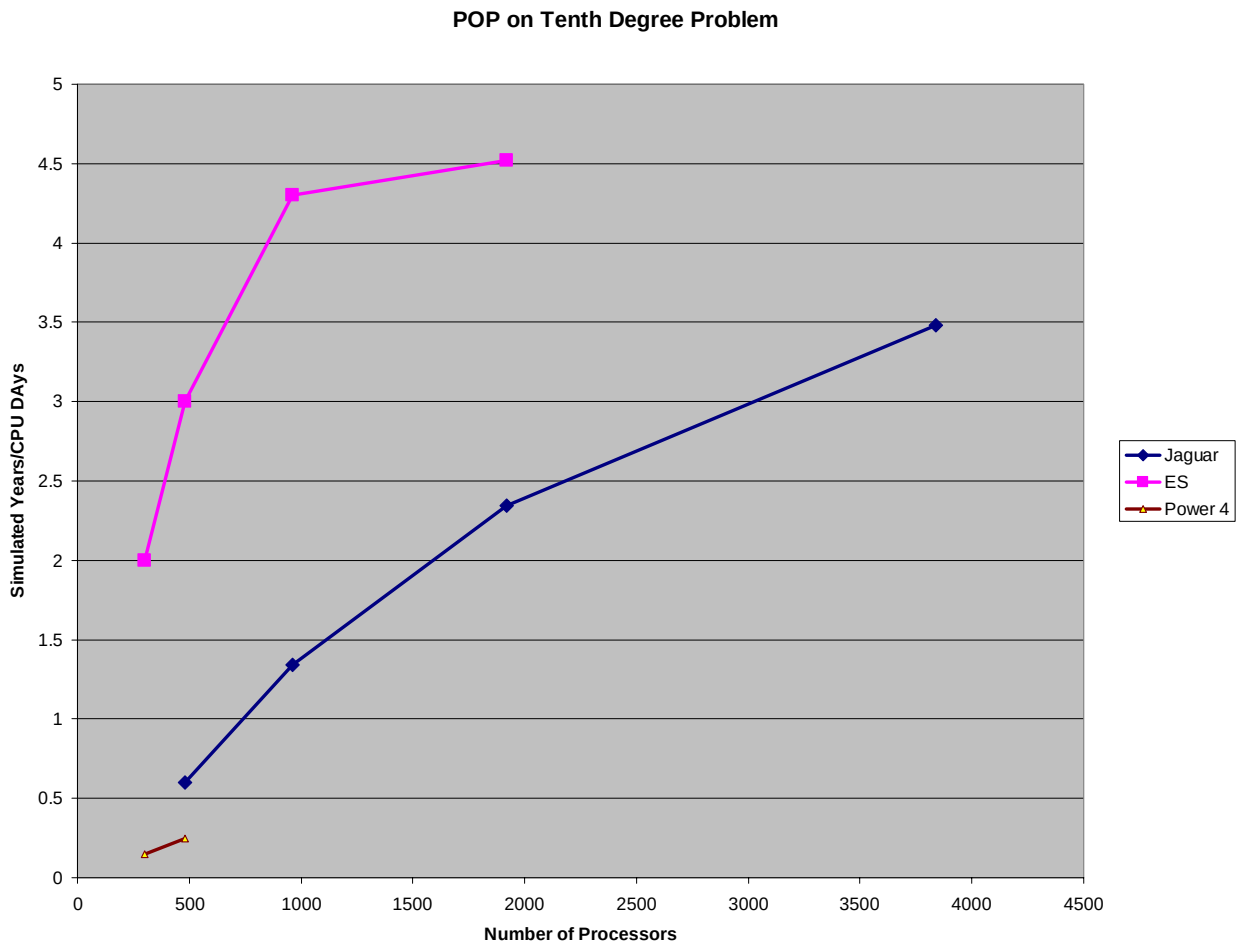
Custom compute nodes fit into common system infrastructure with commodity-based nodes.

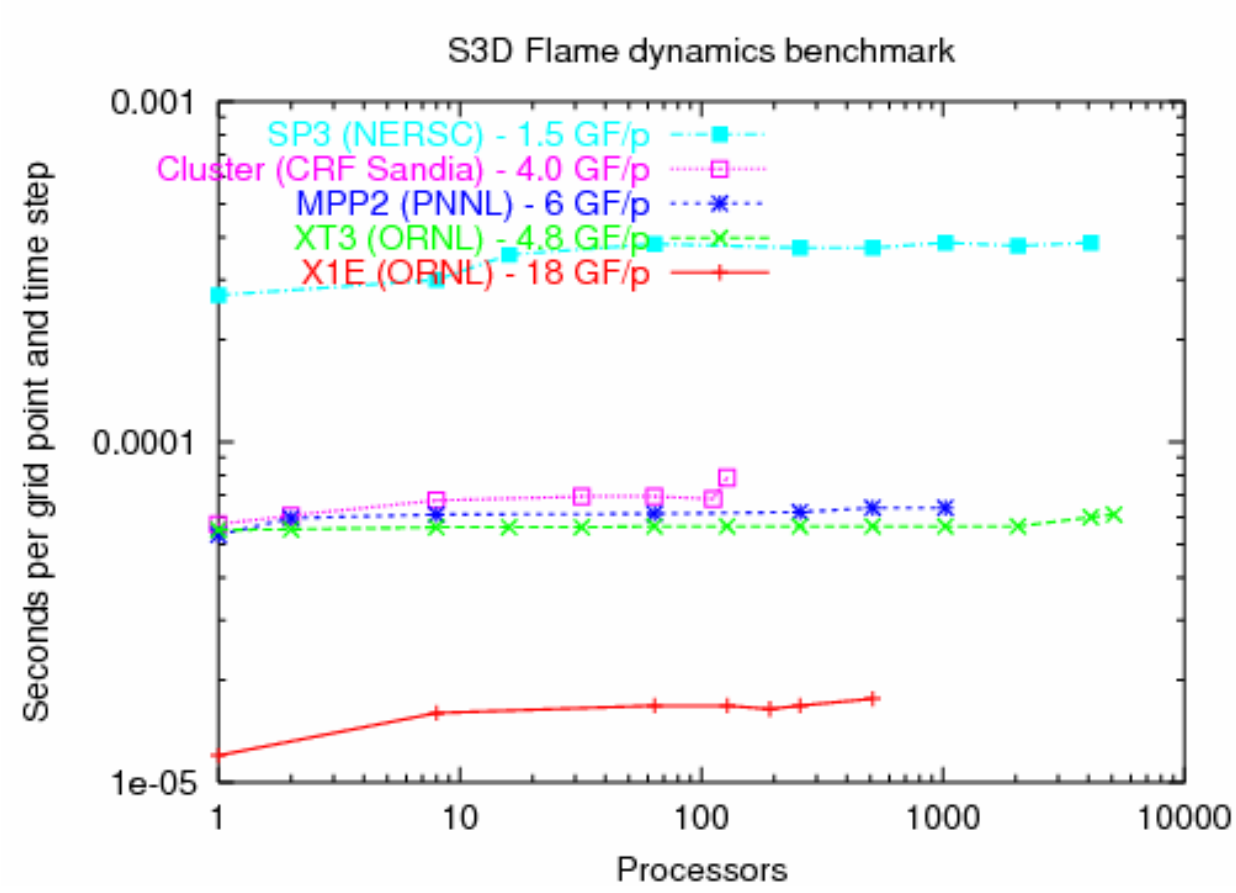

GYRO performance for B1-std
Cray X1E
Cray XT3 (2.4 GHz)
SGI Altix (1.7 GHz)
TeraGrid cluster (1.5 GHz)
IBM p690 cluster (1.3 GHz, HPS)
IBM BG/L
IBM SP
Timesteps per Second
0
5
10
15
20
200
400
600
800
1000
1200
Processors

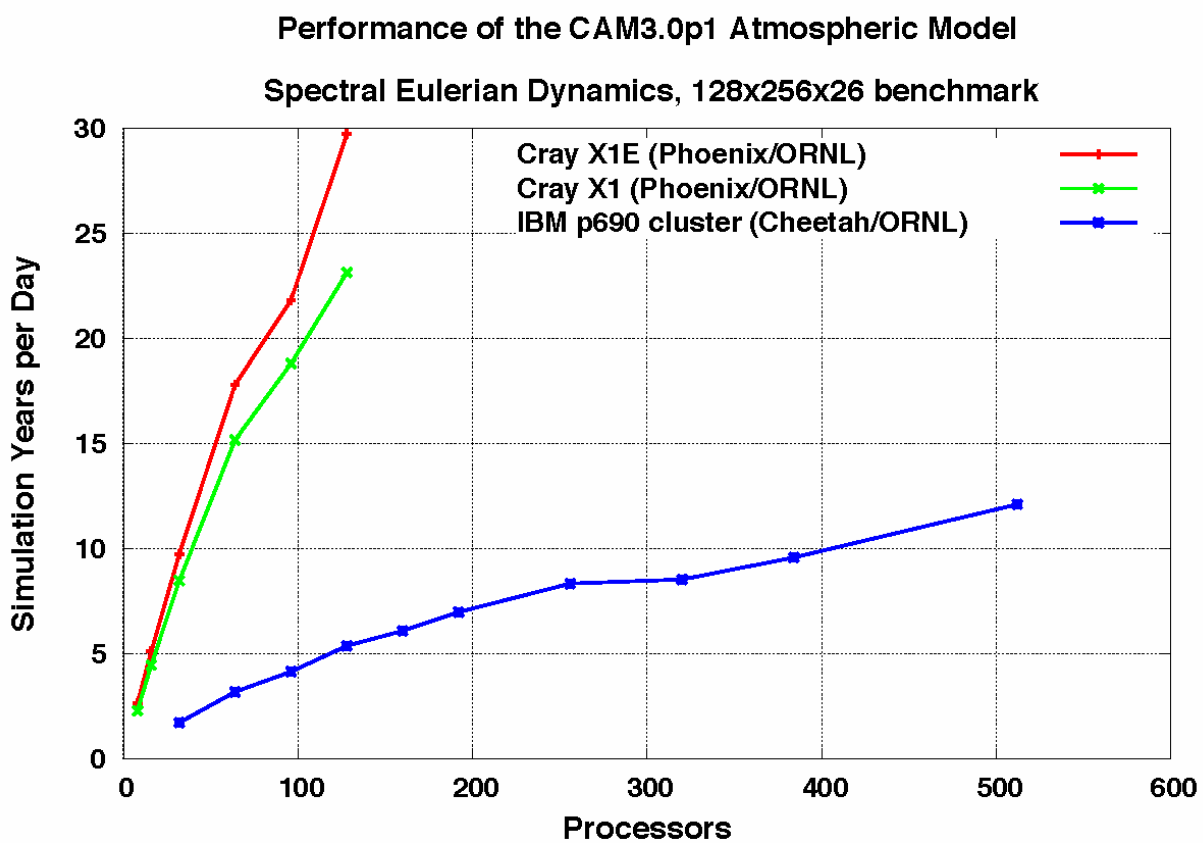

GYRO performance for B3-gtc
Cray X1E
SGI Altix (1.7 GHz)
Cray XT3 (2.4 GHz)
TeraGrid (1.5 GHz)
IBM p690 cluster (1.3 GHz, HPS)
IBM SP (375 MHz)
Timesteps per Second
Processors
0
1
2
3
4
5
6
0
500
1000
1500
2000
2500
3000
3500
4000
4500

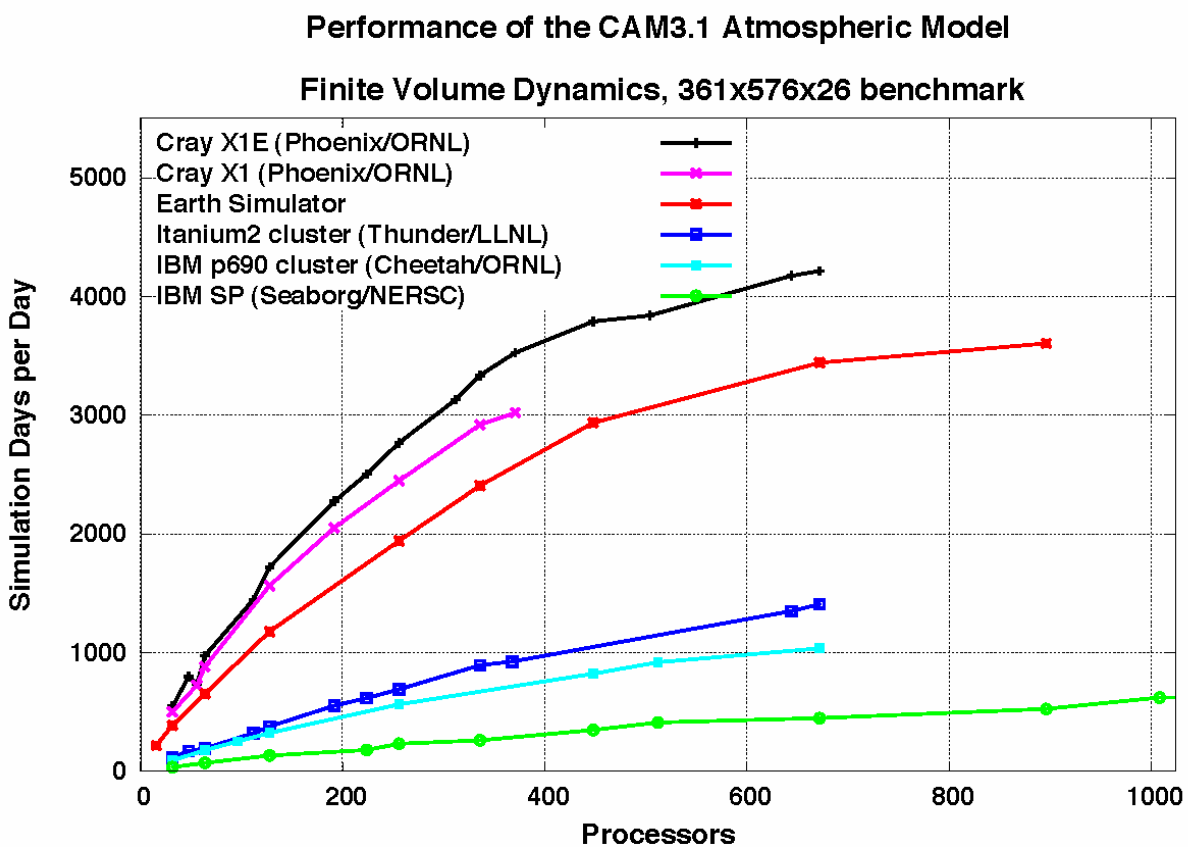

LANL Parallel Ocean Program
POP 1.4.3, x1 benchmark
Cray X1E (9/2005, w/CAF)
Cray X1E (9/2005)
Cray X1 (6/2004, w/CAF)
Earth Simulator
SGI Altix (1.7 GHz)
Cray XT3 (2.4 GHz)
IBM p690 cluster (1.3 GHz, HPS)
IBM SP (375 MHz)
Simulation Years per Day
Processors
450
400
350
300
250
200
150
100
50
50
100
150
200
250
300
350
400
450

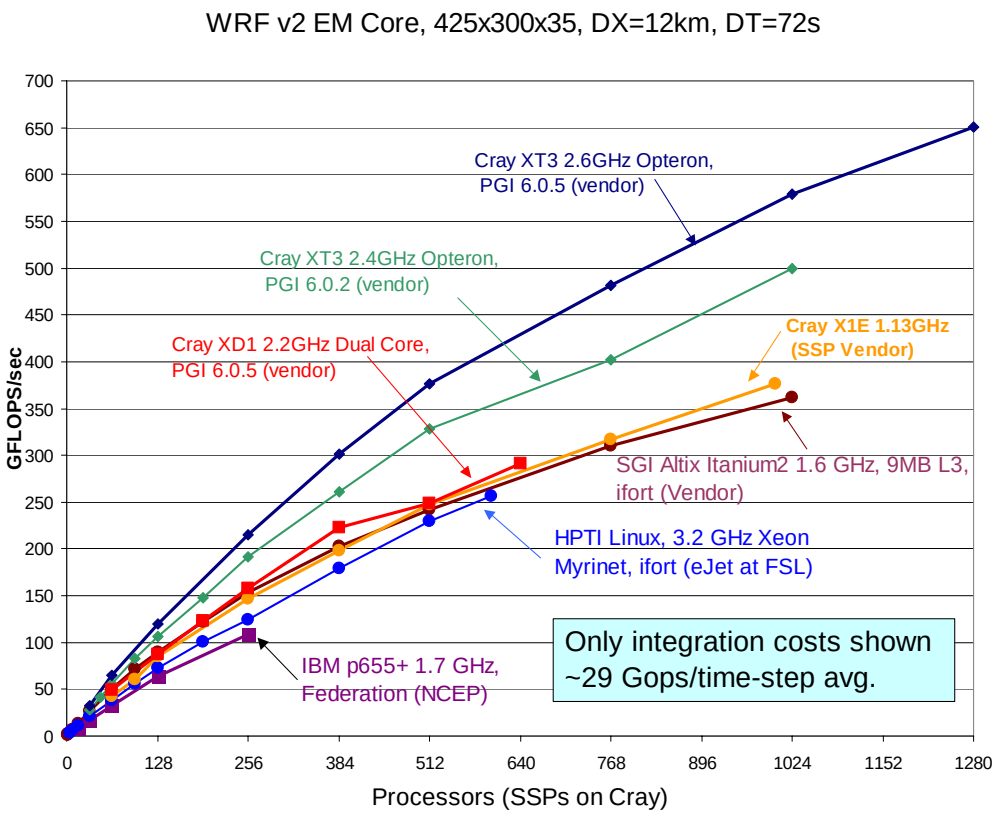
1/10 degree POP problem




S3D Flame dynamics benchmark
SP3 (NERSC) - 1.5 GF/p
Cluster (CRF Sandia) - 4.0 GF/p
MPP2 (PNNL) - 6 GF/p
XT3 (ORNL) - 4.8 GF/p
X1E (ORNL) - 18 GF/p
Seconds per grid point and time step
0.001
0.0001
1e-05
1
10
100
1000
10000
Processors


Performance of the CAM3.0p1 Atmospheric Model
Spectral Eulerian Dynamics, 128x256x26 benchmark
Cray X1E (Phoenix/ORNL)
Cray X1 (Phoenix/ORNL)
IBM p690 cluster (Cheetah/ORNL)
Simulation Years per Day
Processors
0
5
10
15
20
25
30
0
100
200
300
400
500
600


Performance of the CAM3.1 Atmospheric Model
Finite Volume Dynamics, 361x576x26 benchmark
Cray X1E (Phoenix/ORNL)
Cray X1 (Phoenix/ORNL)
Earth Simulator
Itanium2 cluster (Thunder/LLNL)
IBM p690 cluster (Cheetah/ORNL)
IBM SP (Seaborg/NERSC)
Simulation Days per Day
0
1000
2000
3000
4000
5000
0
200
400
600
800
1000
Processors


WRF v2 EM Core, 425x300x35, DX=12km, DT=72s
GFLOPS/sec
Processors (SSPs on Cray)
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700
0
128
256
384
512
640
768
896
1024
1152
1280
Cray XT3 2.6GHz Opteron, PGI 6.0.5 (vendor)
Cray XT3 2.4GHz Opteron, PGI 6.0.2 (vendor)
Cray XD1 2.2GHz Dual Core, PGI 6.0.5 (vendor)
Cray X1E 1.13GHz (SSP Vendor)
SGI Altix Itanium2 1.6 GHz, 9MB L3, ifort (Vendor)
HPTI Linux, 3.2 GHz Xeon Myrinet, ifort (eJet at FSL)
IBM p655+ 1.7 GHz, Federation (NCEP)
Only integration costs shown
~29 Gops/time-step avg.

Cray….

- The *only* company dedicated solely to Supercomputing
  - **Deep experience base**
  - **Uniquely responsive to our customers' needs**
- Our approach:
  - **Purpose-built machines for HPC**
  - **Leading system balance (processor, memory, network, storage)**
  - **Designed for efficiency and reliability at scale**
  - **Focus on end-user productivity and time-to-solution**
- Today's solutions:
  - **X1E ←→ XT3 ←→ XD1**
- Sustained innovation and vision:
  - ***Rainier:*** **Integrated computing capability in 2007**
  - ***Cascade:*** **Aggressive research program for 2010**