



## The Cray XT3™ MPP Supercomputer

John M. Levesque

Director – Cray's Supercomputing Center  
of Excellence

Oct, 2005

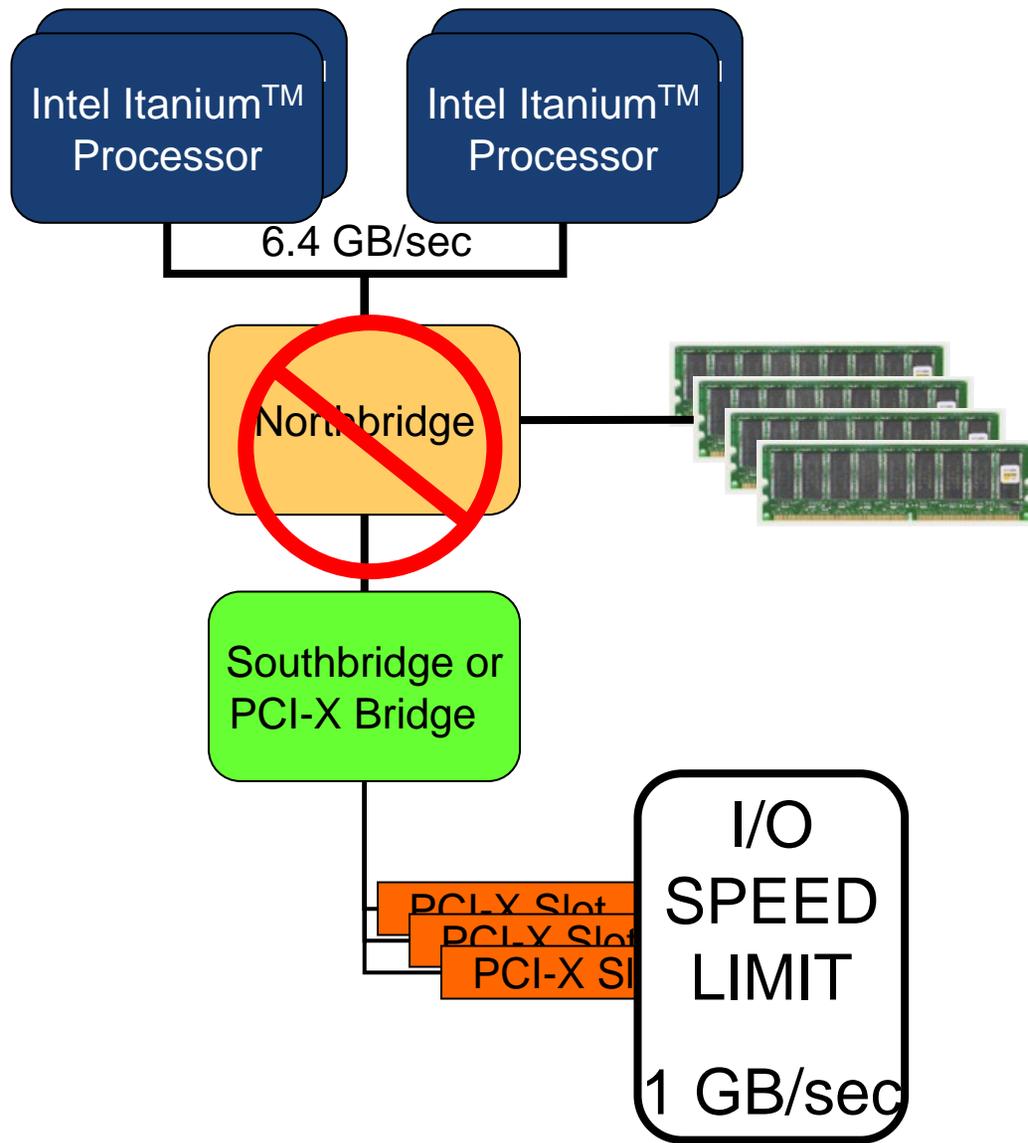


# Recipe for a good MPP

1. Select Best Microprocessor
2. Surround it with a balanced or “bandwidth rich” environment
3. Eliminate “barriers” to scalability
  - SMPs don’t help here
  - Eliminate Operating System Interference (OS Jitter)
  - Reliability must be designed in
  - Resiliency is key
  - System Management
  - I/O
  - System Service Life



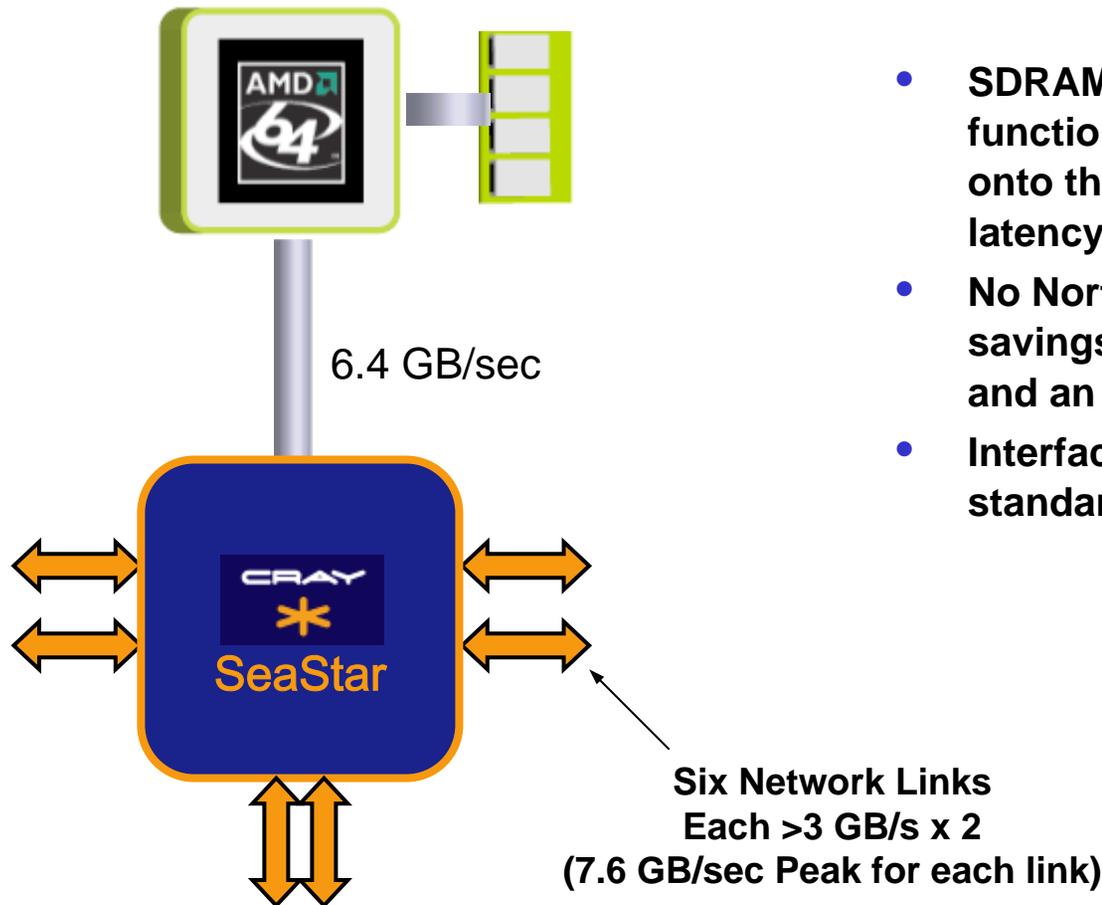
# Picking the best Processor: Why not Intel?



- Memory latency ~ 160 ns and *bandwidth is shared* between multiple processors
- Northbridge chip is 2<sup>nd</sup> most complex chip on the board. Typical chip uses about 11 Watts
- Any interconnect limited by speed of PCI-X since it's the fastest place to "plug in"
- Best place to tie in a high performance interconnect would be through the Northbridge, but this is difficult to do legally without an Intel bus license

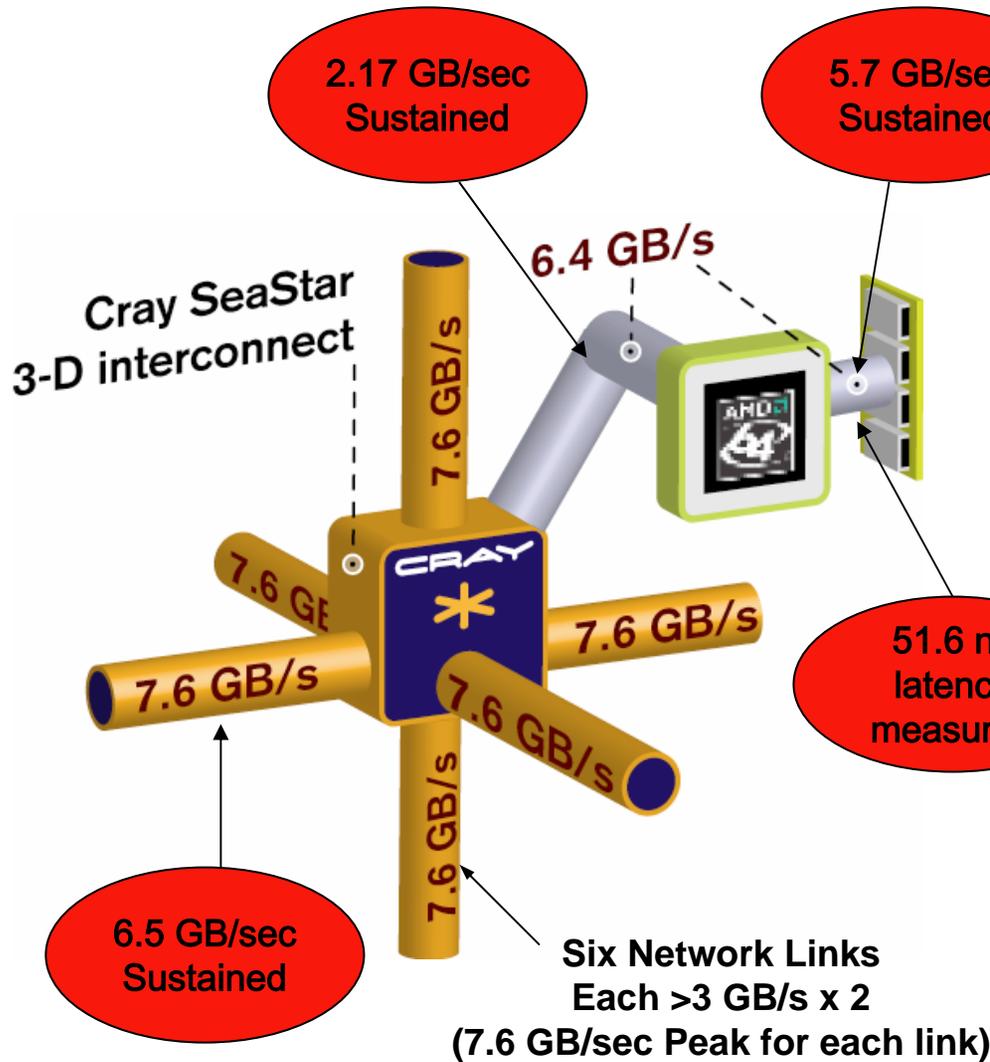
# AMD Opteron Generic System

## CRAY XT3 PE



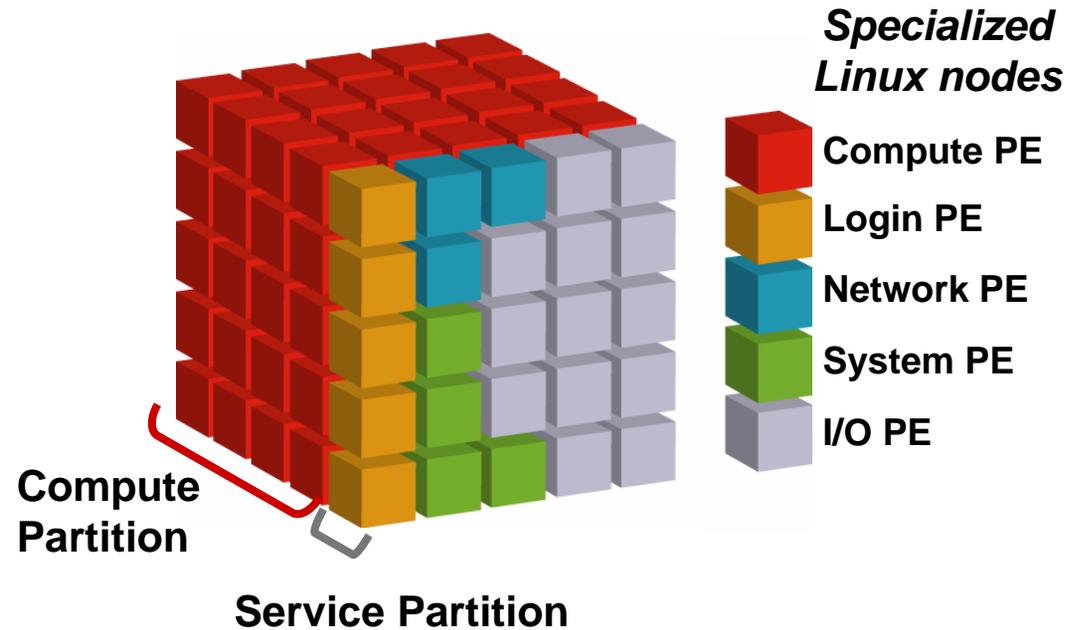
- SDRAM memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to 60-90 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

# Cray XT3 Processing Element: Measured Performance



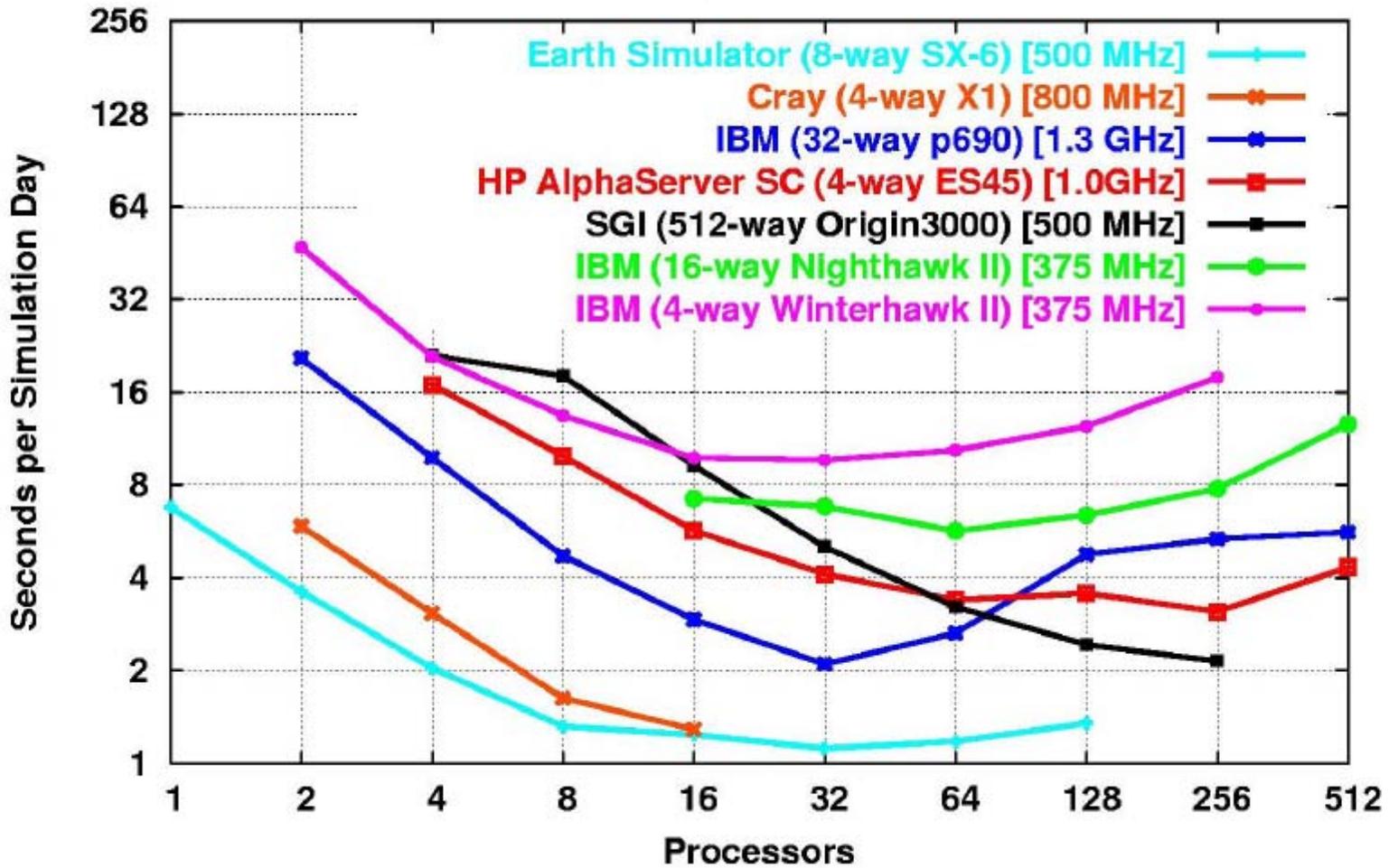
- SDRAM memory controller and function of Northbridge is pulled onto the Opteron die. Memory latency reduced to <60 ns
- No Northbridge chip results in savings in heat, power, complexity and an increase in performance
- Interface off the chip is an open standard (HyperTransport)

# Scalable Software Architecture: UNICOS/Ic

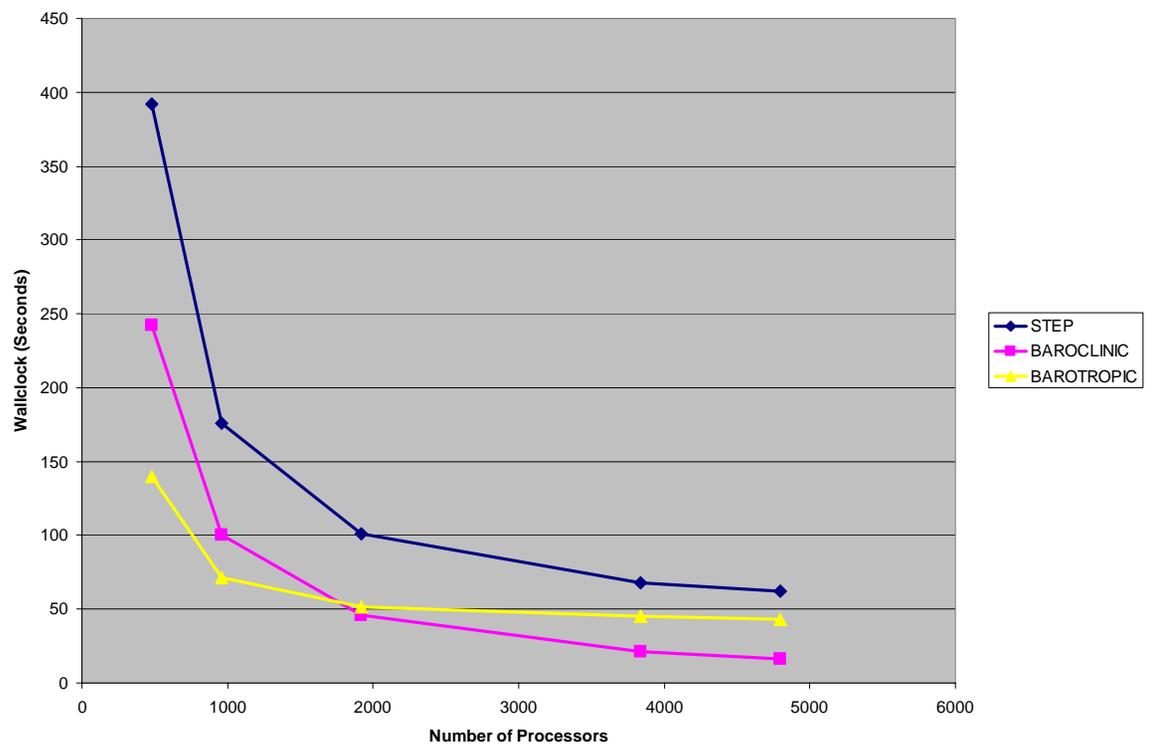


- Microkernel on Compute PEs, full featured Linux on Service PEs.
- Contiguous memory layout used on compute processors to streamline communications
- Service PEs specialize by function
  - 100 ms interrupt times
  - Will be synchronized if required
  - OS heartbeat checked once per second.
- Software Architecture eliminates OS “Jitter”
- Software Architecture enables reproducible run times

POP Barotropic Timings  
POP 1.4.3, x1 benchmark

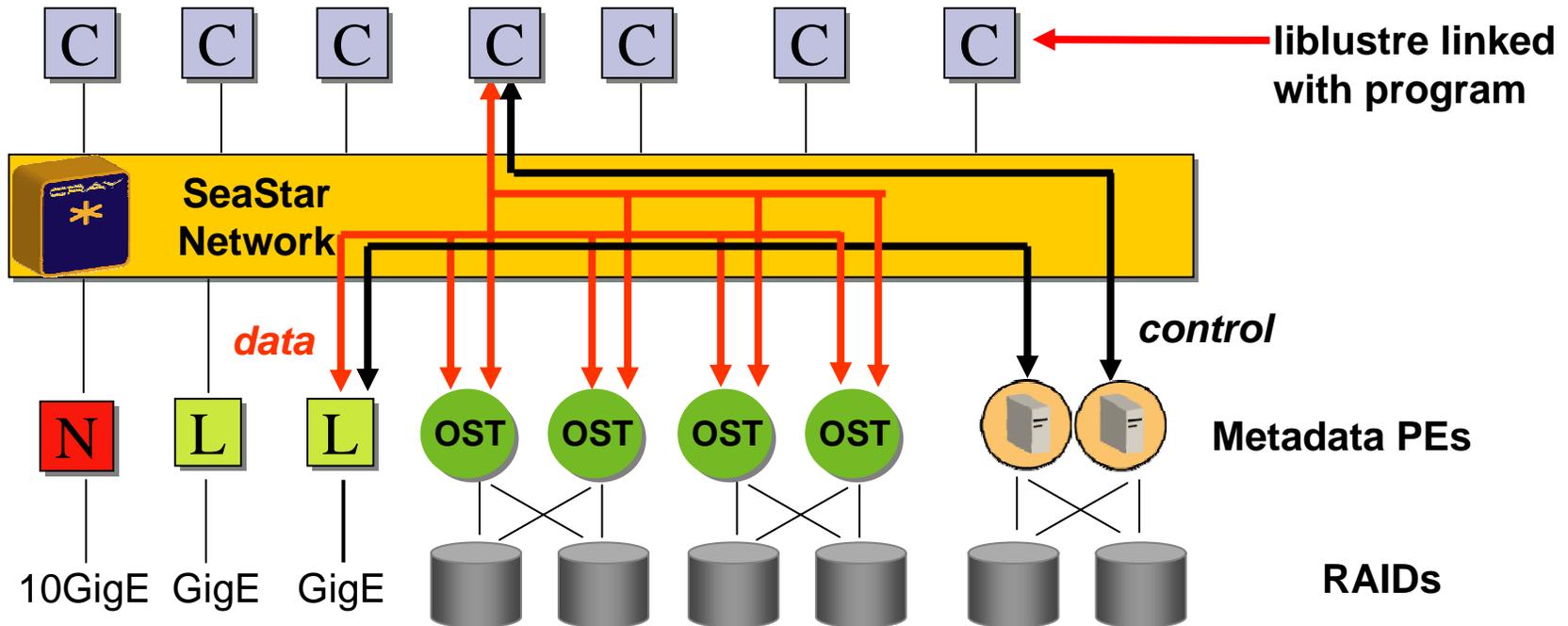
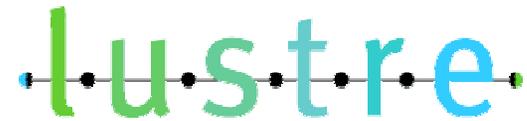


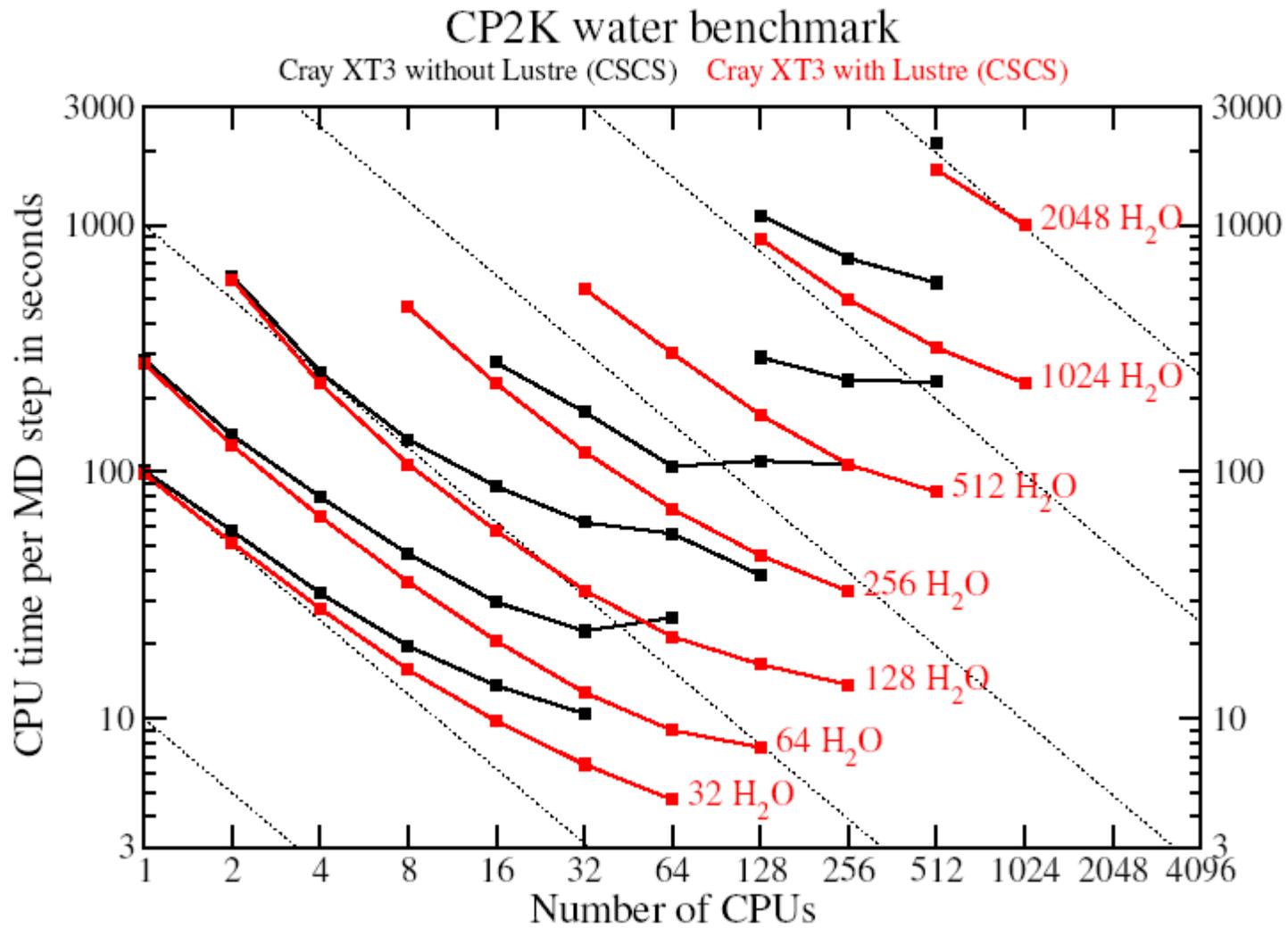
POP Tenth Degree on XT3



# Scalable I/O

- Global Parallel File System: Lustre
  - Open Source, Vendor Neutral
  - Highly Scalable, block allocation NOT serialized
  - Liblustre for MPPs
  - OST Software Failover, Dual Path controllers







CRAY XT3  
SCALABLE BY DESIGN

© 2005 CRAY INC. ALL RIGHTS RESERVED.  
ANTOLAP, DE GRUIR