The Cray XE6 System

Ter@tec 2010 Workshop Ecole Polytechnique

June 16th

The Cray XE6

CRAY XE6

Designed to scale to over 1 million processor cores, every aspect of the Cray XE6 supercomputer – from its industry-leading resiliency features to its host of scalability-boosting technologies – has been engineered to meet science's ever-toughening demands for scalability, reliability and flexibility.

Cray Technology Innovations

System Interconnect	Custom interconnect and communications network
Systems Management & Performance	Software to productively manage and extract performance out of thousands of processors as a single system
Packaging	Very high density, upgradeability, liquid and air-cooling
Adaptive Supercomputing	Single integrated system

Building the Technologies and Infrastructure for Superior Scalability and Sustained Performance

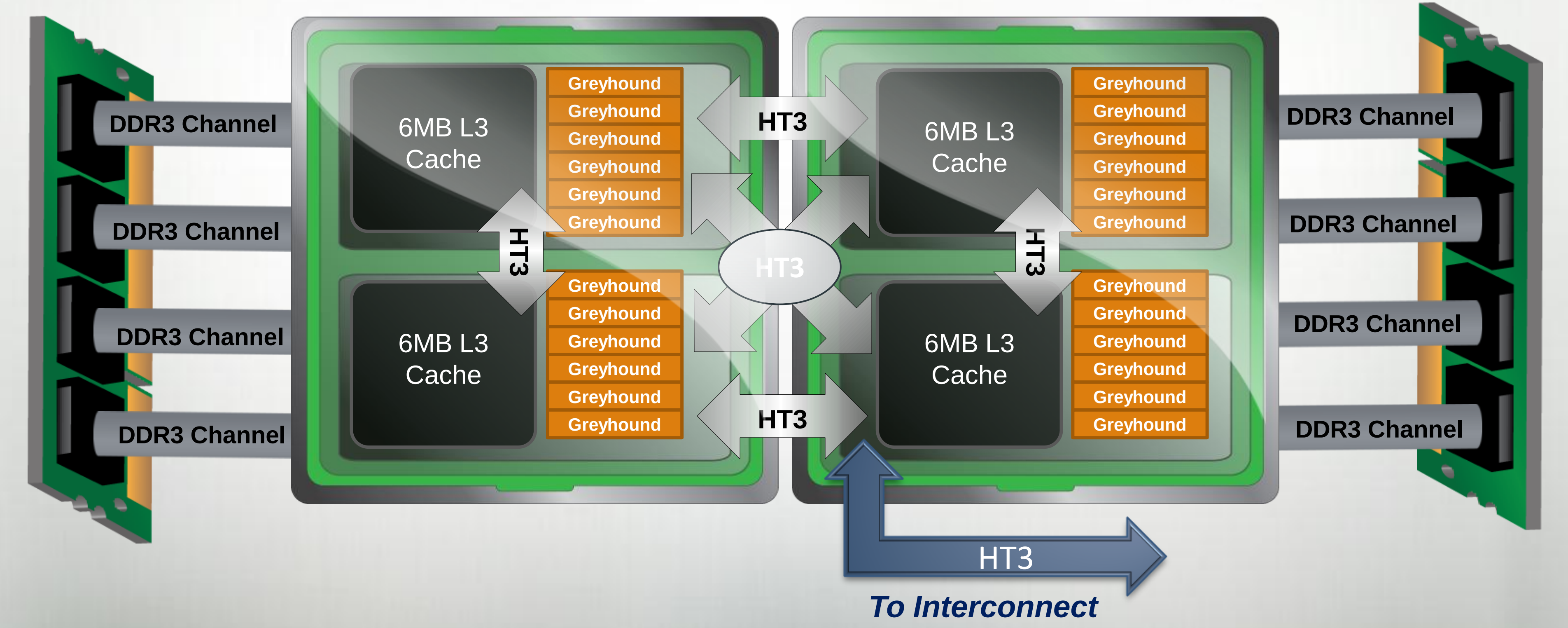
Cray XE6 System

- System announced June 2010 at CUG Edinburgh, Scotland
- First systems are shipping
- Key technologies
  - Series 6 blade to support AMD’s new 6100 series Opteron
  - Gemini interconnect
  - New XIO blade
  - CLE 3 Operating System

Cray XE6: Scalable Performance

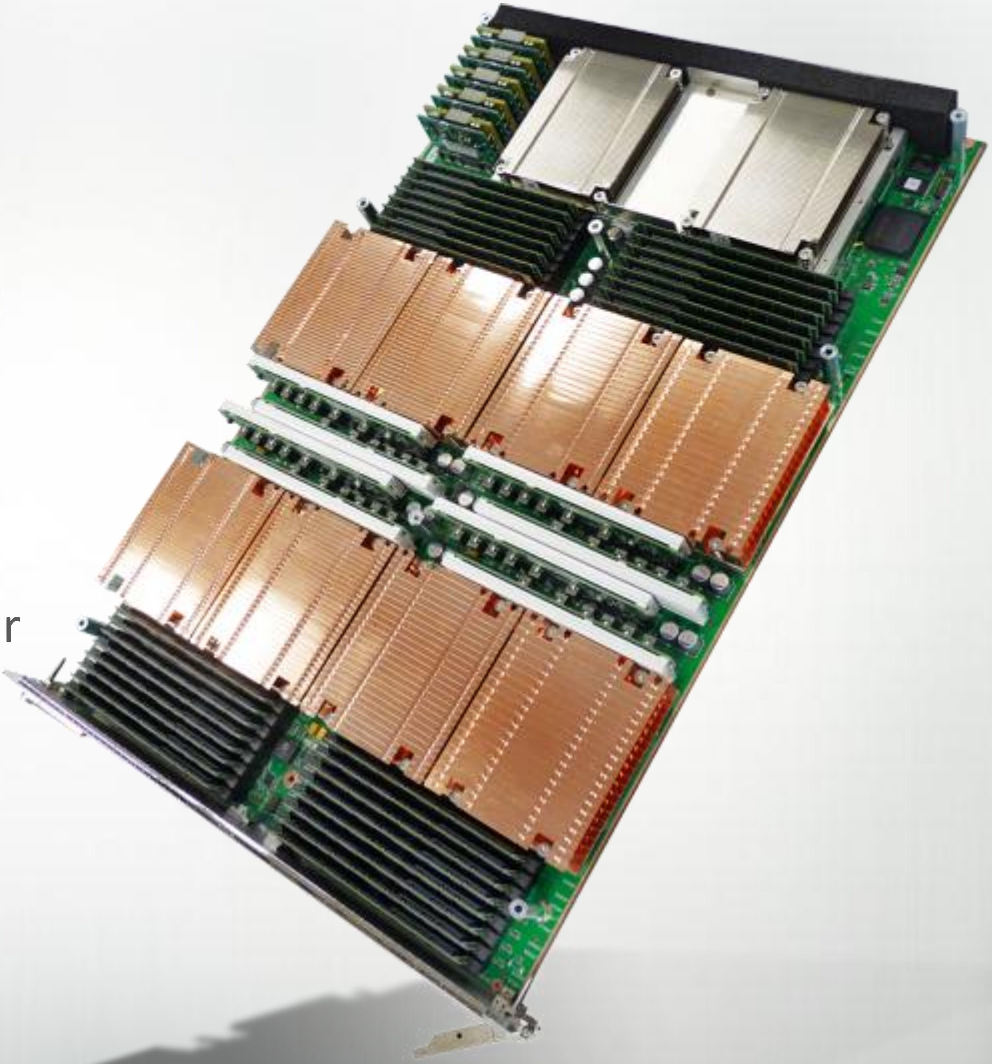
- AMD Opteron 6100 Series Processors
- Dual socket, 8 or 12 cores
  - Over 200 GFlops peak

- 8 × DDR3 channels
  - 32 or 64 GB of memory
  - Bandwidth 85 GB/s



Cray XE6: Compute Blade

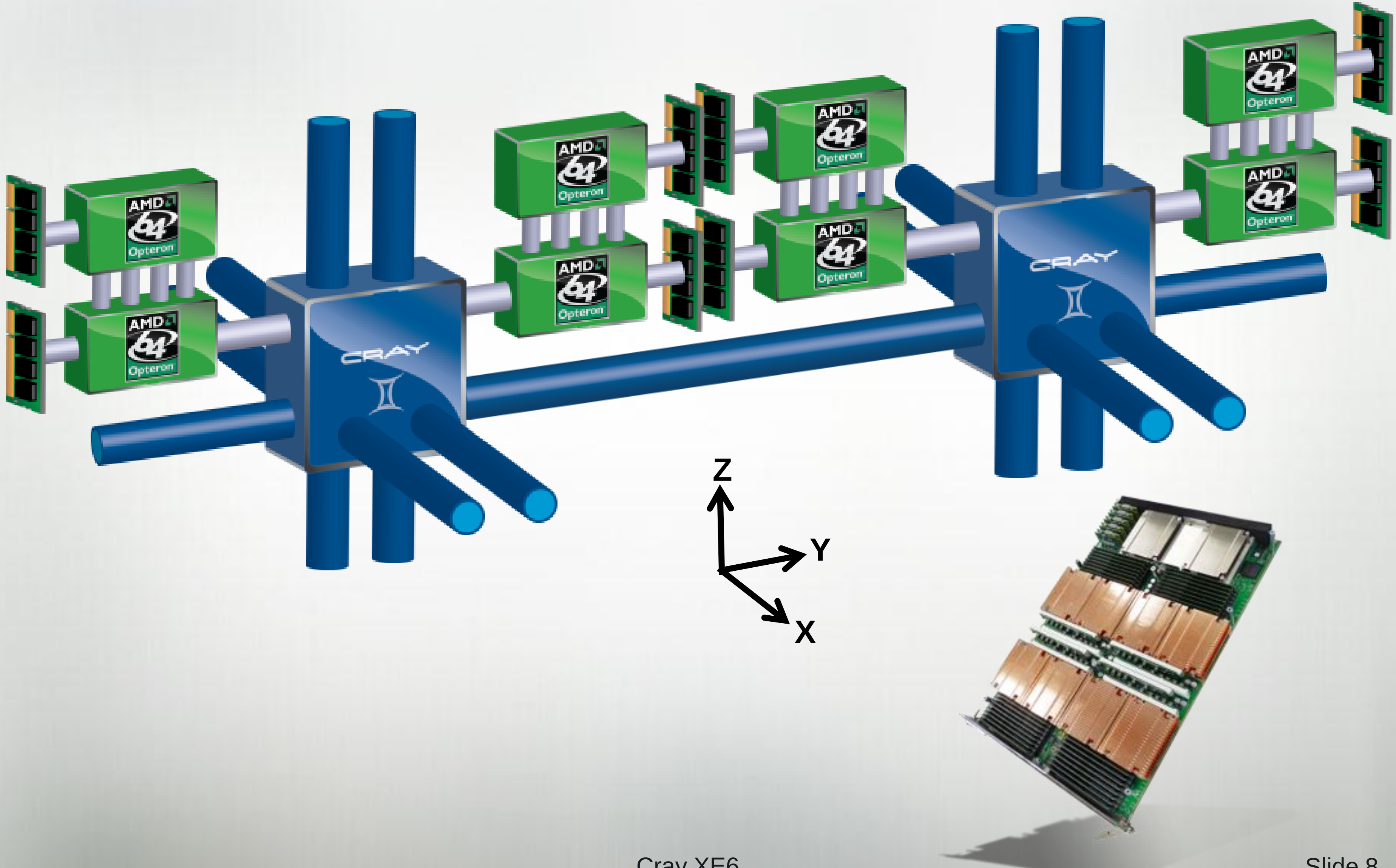
- 4 Compute nodes
- Each:
  - 2 Magny Cours Sockets
  - 16/24 Compute Cores
  - 8 DDR3 Memory channels
  - 8 DDR3 Memory DIMMS
- 2 Gemini ASICs
- L0 Blade management processor
- Fault tolerant power conversion



Cray XE6: I/O Blade

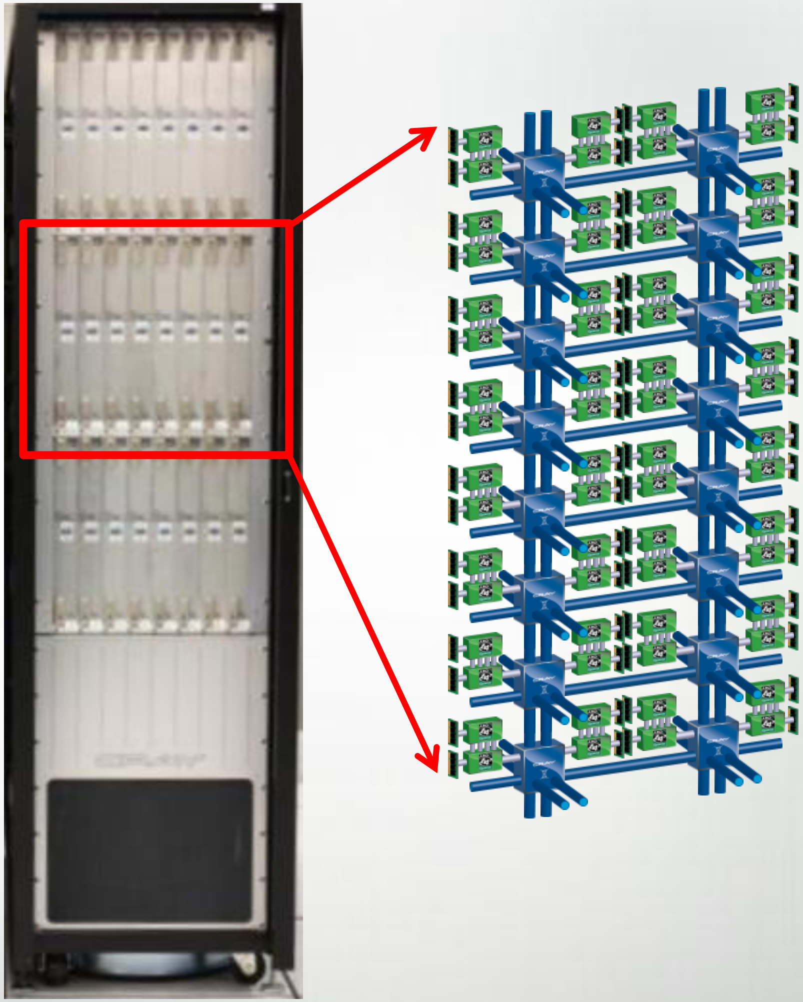
- Provides XE6 service nodes
- 4 single socket nodes
- Each:
  - 6 core CPU
  - 4 DDR2 DIMMs
  - 32 GB with 8GB DIMMs
  - AMD SR5670 Bridge
  - PCI-Express Gen2 x16
- Gemini Interconnect

Cray XE6: Compute Blade Topology



Cray XE6 scaling up: cabinet

- Three chassis per cabinet
  - 8 blades per chassis
- 3D Torus network
  - No external switch cabinets
  - Each chassis provides a slice of the network
- Each cabinet
  - 96 dual socket nodes
  - 1536 or 2304 cores
  - 20 Tflop/s
  - 3 or 6 Terabytes of memory
  - Power consumption 25-50 KW



Cray XE6 scaling up: System

SPECIFICATIONS

Compute cabinets:	16
Nodes:	1500
Cores:	36,000
Peak:	320 Tflops
Memory:	96 TBytes
Power:	< 800kW
XDP cooling units:	4

Driving Sustained Performance & Energy Efficiency

Cray XT5 & XT5m → 250MF/W

Best performance per watt for x86 supercomputers

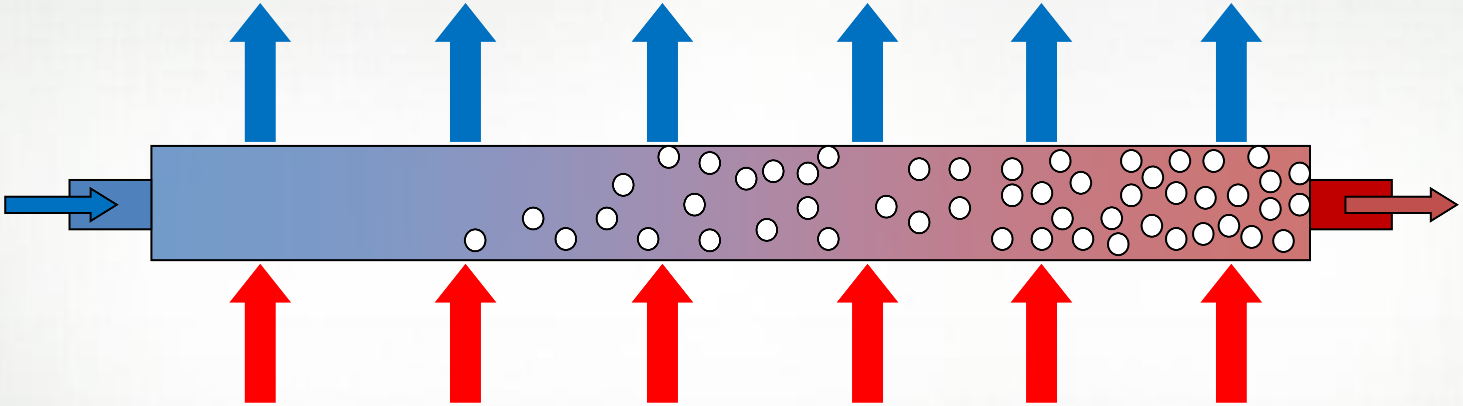
Cray XT6m & XE6 → 330+ MF/W

Setting standard for combining optimal performance with sustainability & upgradability

- 30-40% Improvement in "Peak" energy efficiency!
- ECOphlex cooling further reduces Total Cost of Ownership
- Industry leading "sustained performance" energy efficiency

Cray XE6: ECOphlex Cooling

Cool air is released into the computer room



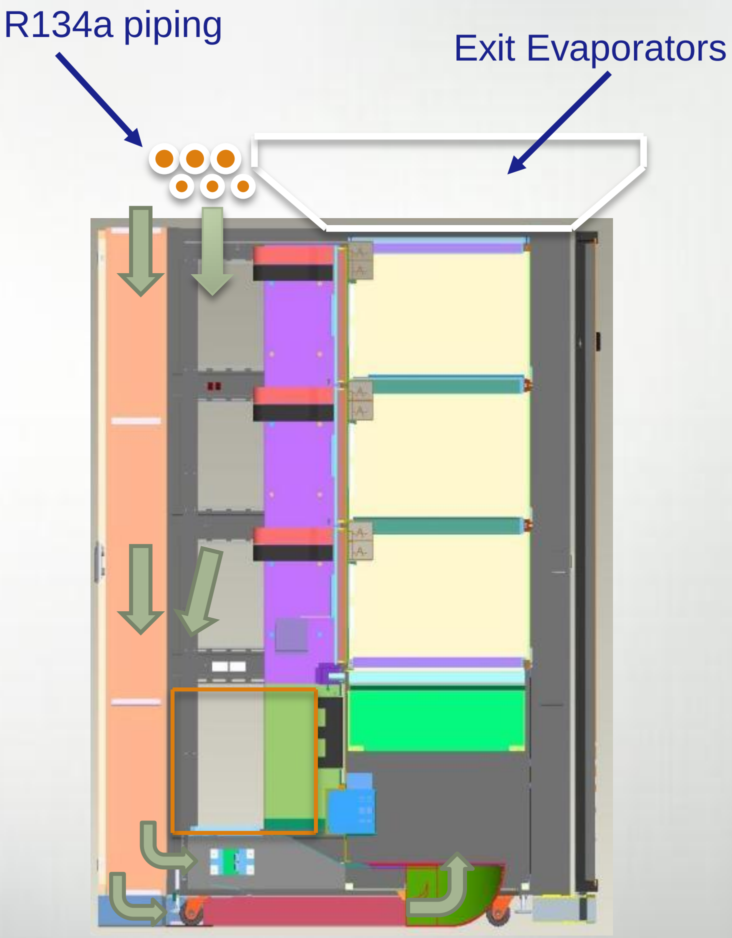
Liquid in

Liquid/Vapor mixture out

Hot air stream passes through evaporator, rejects heat to refrigerant via liquid-vapor phase change (evaporation).

Cray XE6: ECOphlex Cooling

- Increased efficiency under variable load
- Enhanced to support up to 130W per socket
- Enhanced sound kit to reduce noise
- One XDP for each 4 or 5 cabinets

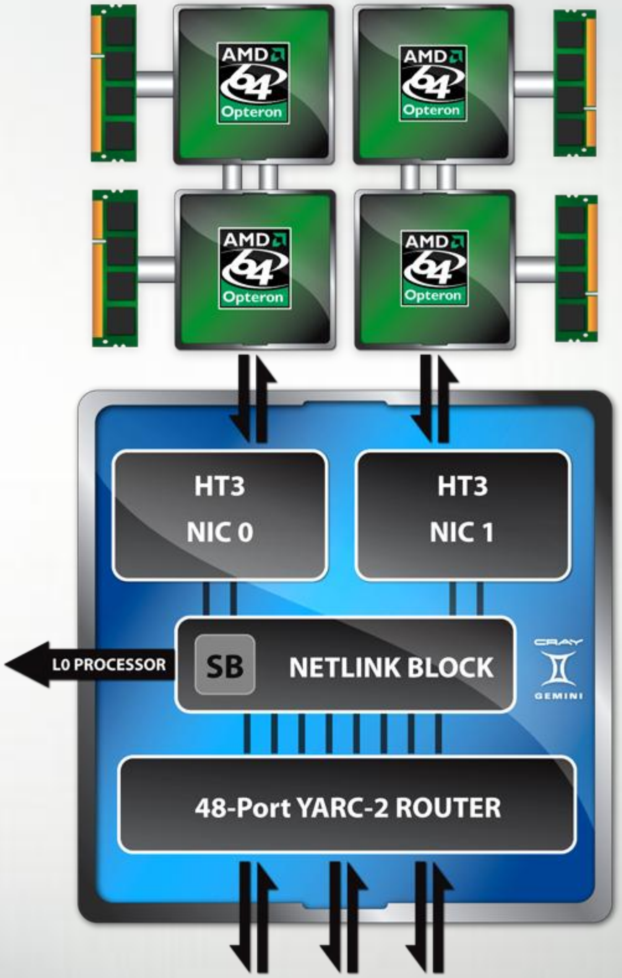


R134a piping

Exit Evaporators

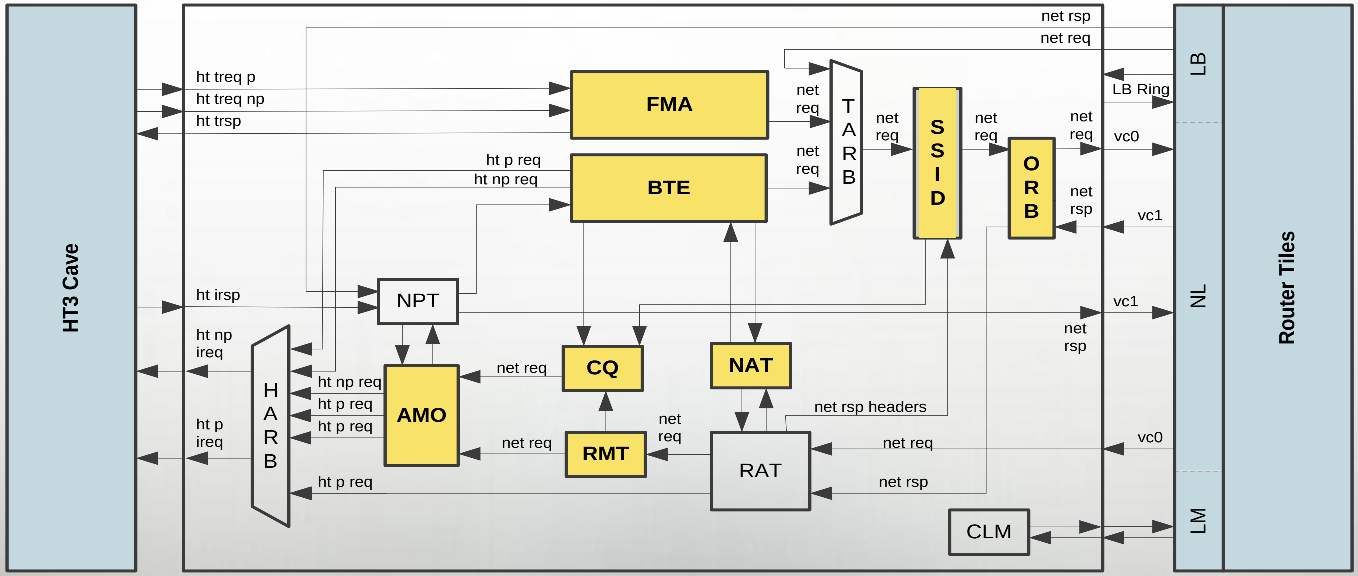
Cray XE6: Gemini Network

- System on a Chip design
  - 2 HyperTransport NICs
  - Embedded high performance router
- 3D Torus network
  - XT5/XT6 systems field upgradable
- Advanced features
  - Globally memory access
  - High rate of small messages
  - Hardware support for PGAS languages: put/get/amo



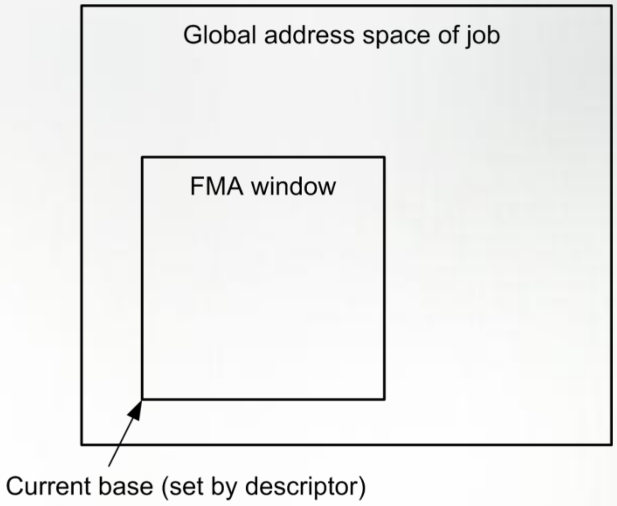
Cray XE6: Gemini NIC Design

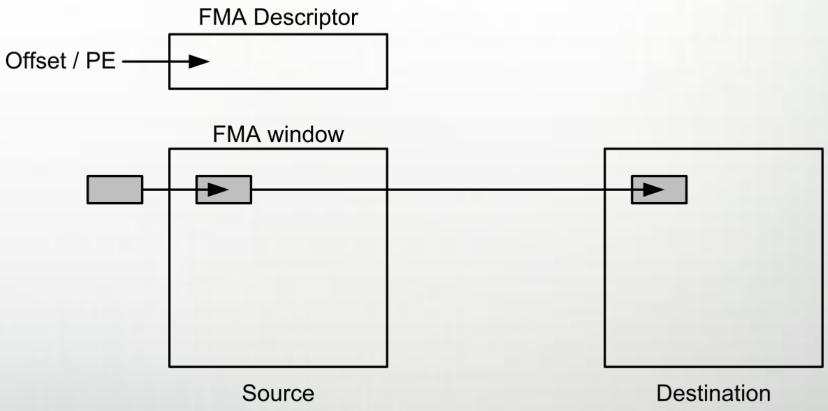
- HyperTransport 3 host interface
- Hardware pipeline maximizes issue rate
- Fast memory access (FMA)
- Block transfer engine (BTE)
- Hardware translation of node ids and user virtual addresses
- AMO cache
- Network bandwidth dynamically shared between NICs



Cray XE6: Fast Memory Access

- FMA provides a local window into the global address space
- Initialise FMA descriptors and associated windows
- Writes to windows generate put/get/amo
- Writes to the descriptor to modify destination and base address of window (not always needed)
- Can also modify how the address bits map to PEs





Cray XE6: Gemini MPI Features

- FMA provides low-overhead OS-bypass.
    - Lightweight issue of small transfers
- DMA offload engine
    - Allows large transfers to proceed asynchronously of the application
- Designed to minimize memory usage on large jobs
    - Typically 20MB/process including 4MB buffer for unexpected messages
- Adaptive routing:
    - Reduces network contention
    - Automatically routes around link failures
- AMOs provide a fast synchronization method for collectives

Cray XE6: Gemini PGAS Features

- Globally addressable memory provides efficient support for
  - UPC, Co-array FORTRAN, SHMEM
- Pipelined global loads and stores
  - Allows for fast execution of irregular communication patterns
- Atomic memory operations
  - Provides fast synchronization method for one-sided communication
- Cray DMAPP application interface
  - Cray Programming Environment targets this directly
  - Available for 3rd party tools

An Adaptive Linux OS designed specifically for HPC

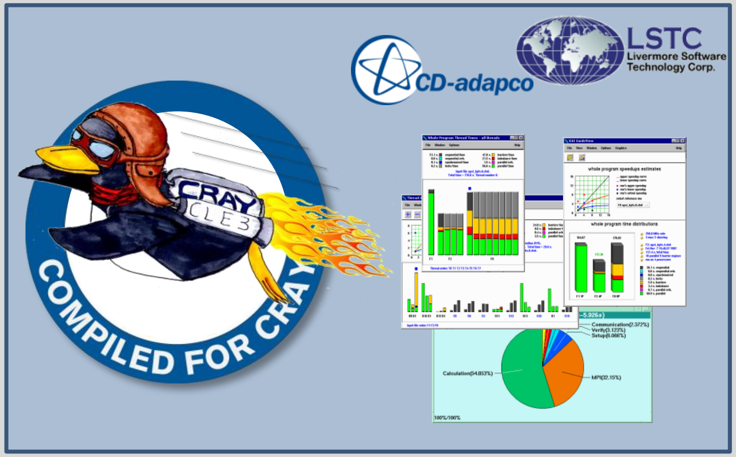
CRAY LINUX ENVIRONMENT CLE3

ESM – *Extreme Scalability Mode*

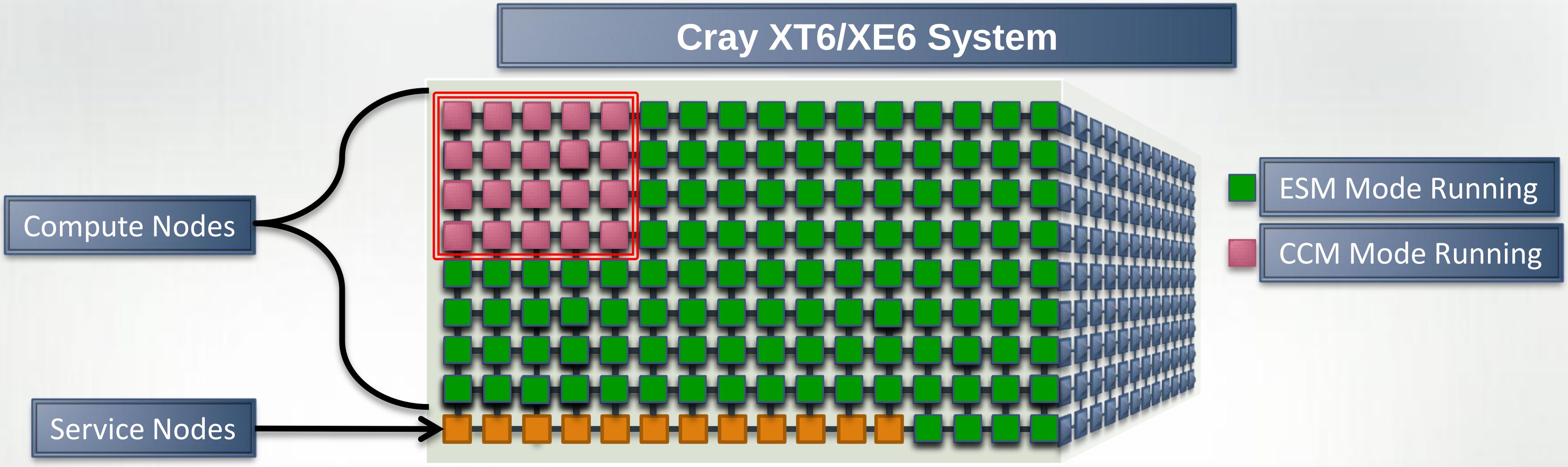
- No compromise *scalability*
- Low-Noise Kernel for scalability
- Native Comm. & Optimized MPI
- Application-specific performance tuning and scaling

CCM –*Cluster Compatibility Mode*

- No compromise *compatibility*
- Fully standard x86/Linux
- Standardized Communication Layer
- Out-of-the-box ISV Installation
- ISV applications simply install and run



CLE3 simultaneous CCM and ESM Modes



- Submit CCM application through batch scheduler

  qsub -q ccm_queue AppScript
- Node scheduled then configured for CCM
- Executes the batch script and application
- After CCM job completes, CCM nodes cleared
- CCM nodes available for ESM or CCM mode Applications

Cray XE6: Programming Environment

Every XT6 Cray System Includes

Cray Integrated Tools

- Cray Compilation Environment
  - Fortran/C/UPC/CAF/C++
- Optimized OpenMP/MPI Libraries
- CrayPat, Cray Apprentice2
- Optimized Math Libraries
  - Iterative Refinement Toolkit
  - Cray PETSc, CASK

Customer-selected Options

Compilers

- PGI, PathScale

Debuggers

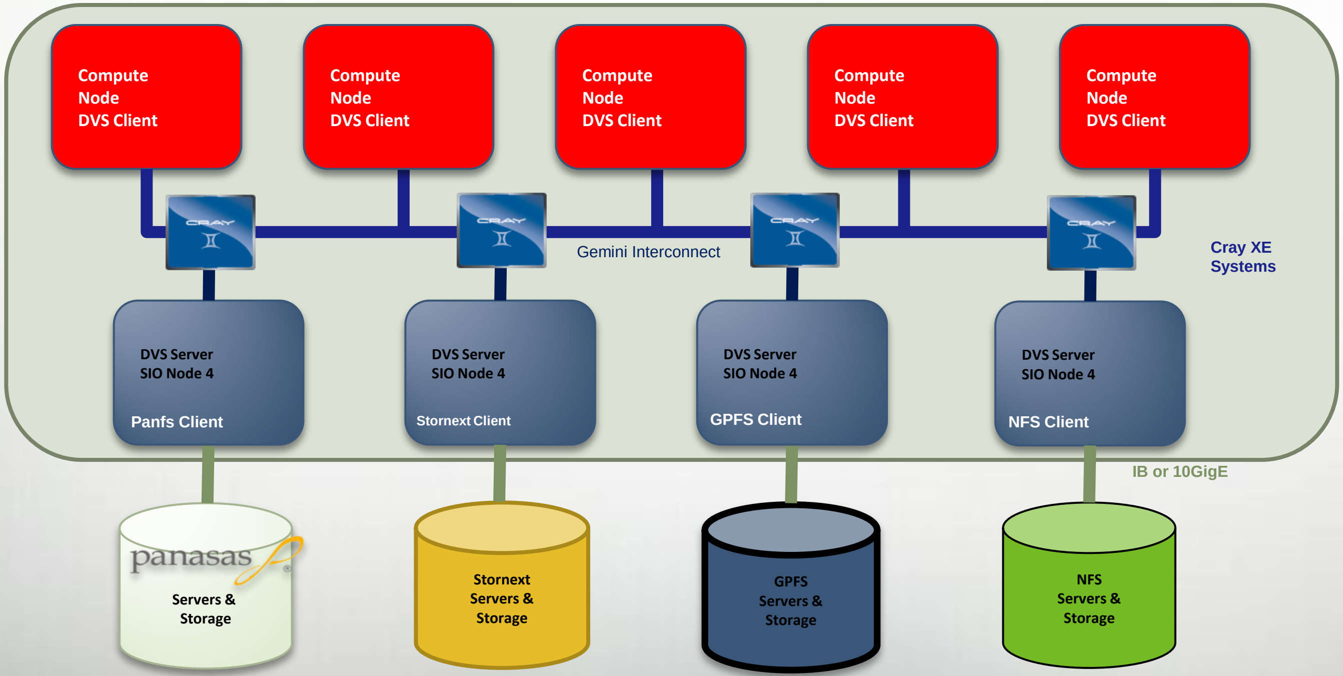
- TotalView, Allinea DDT

Schedulers

- Moab, PBS Professional, LSF

Cray XE6: Filesystems

- Lustre parallel filesystem
  - DDN or LSI storage connected via Fibre Channel or Infiniband
- Data Virtualization Service (DVS)
  - Support for multiple filesystems (e.g. Panasas and GPFS)



Cray XE6: Production Reliability

- Building on the XT5 base:
- Mature, reliable Linux-based OS
  - Cray Linux Environment 3.x
- Resilient system services
  - Meta data
  - Boot servers
  - System management workstation
- Over provisioning of critical resources to meet a given service level
- SeaStar provides link-level retry

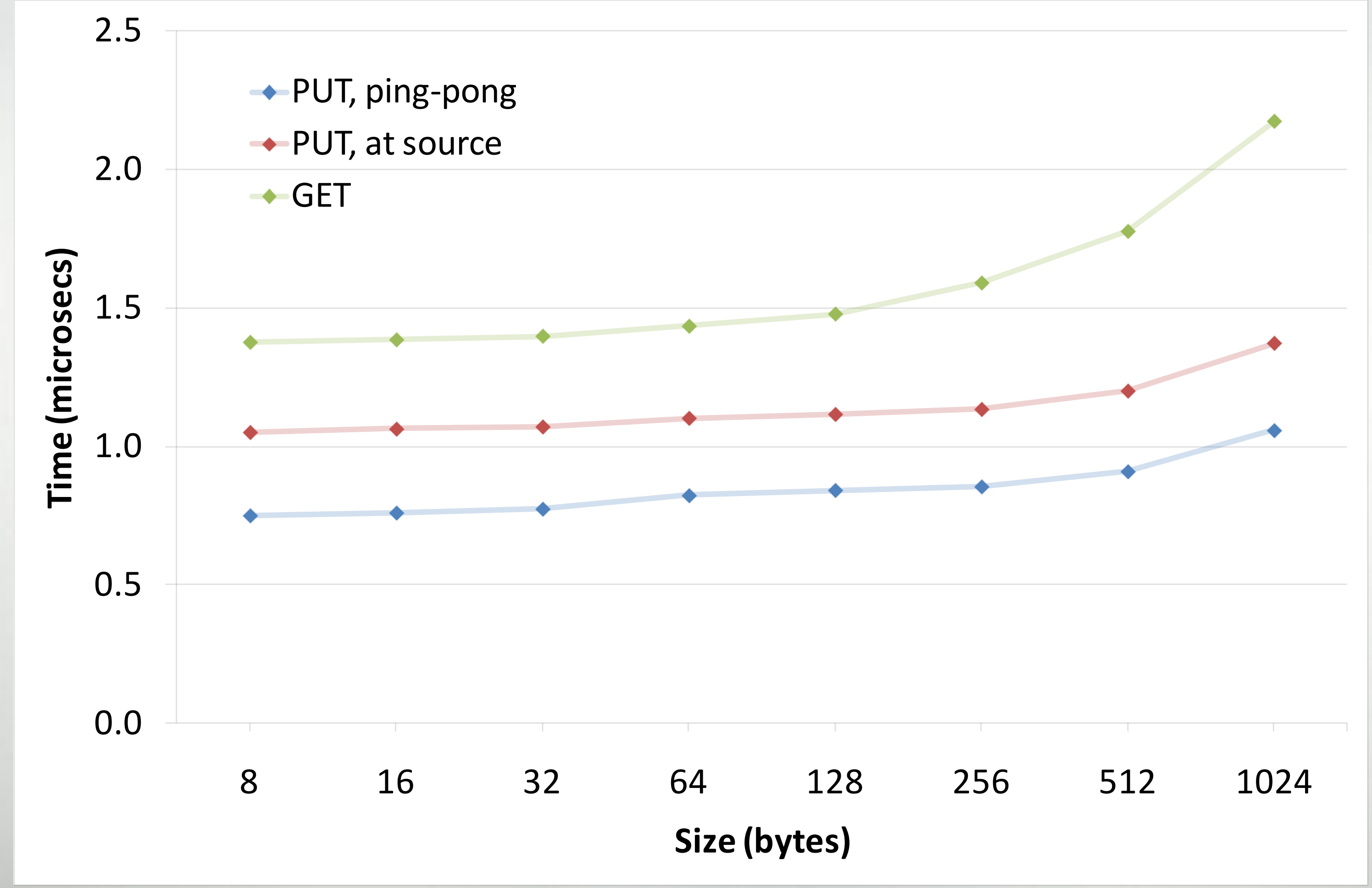
- Gemini brings new features:
  - Adaptive routing around errors
  - Warm swap blades
  - End-to-end reliable MPI

- NodeKARE: Node Knowledge And Recognition
  - Per-node tests run if a job fails
  - Checks more possible sources of error:
    File system checks, memory usage, application termination, site-specific check

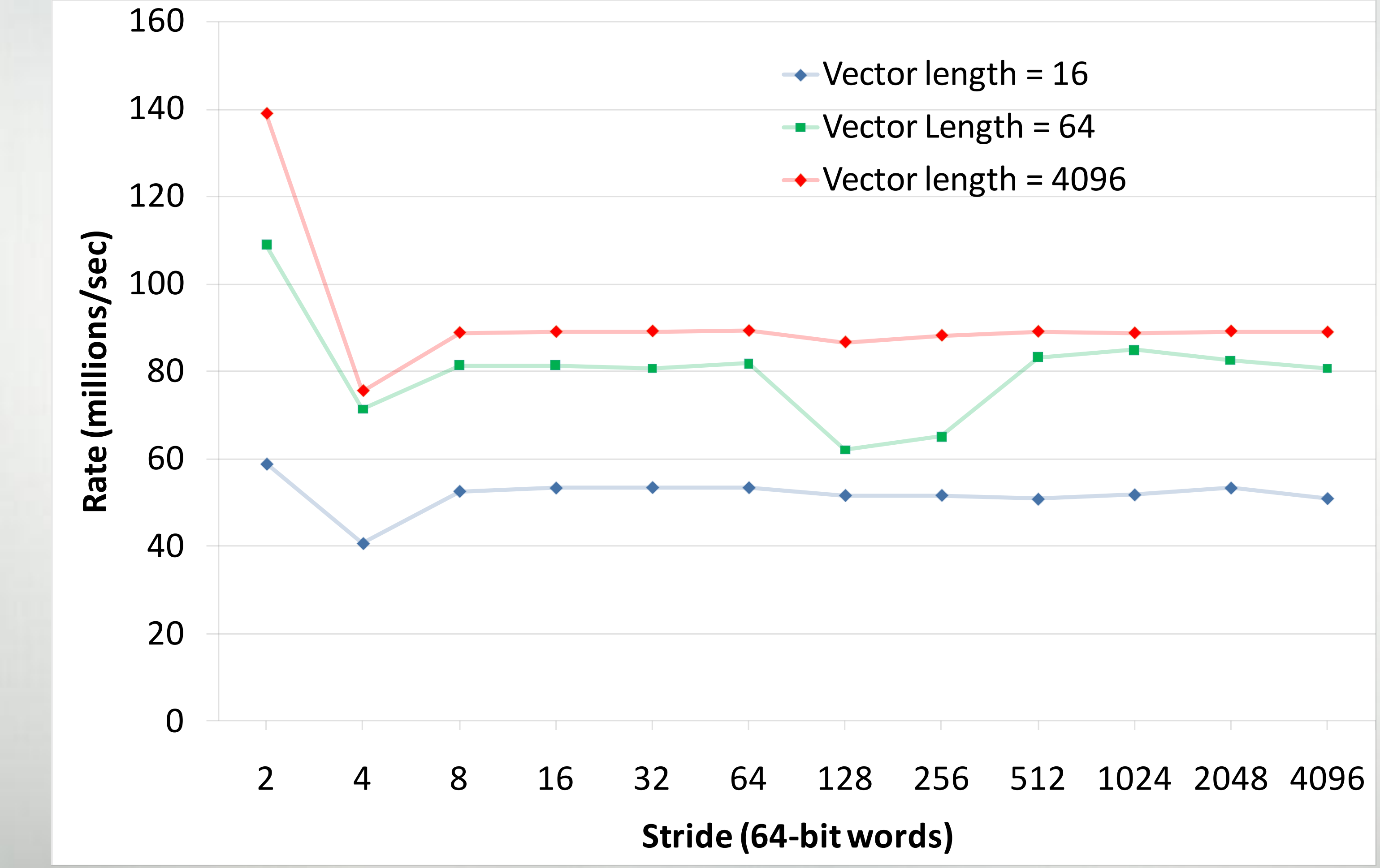
Cray XE6: Early Performance Data

- Preliminary performance data from Gemini software release
- MPI latency: < 1.5 µs
- MPI message rates: 20 times that of XT5
- Injection bandwidth: > 6 GB/s per node
- PGAS latencies: put < 1 µs, get < 1.5 µs

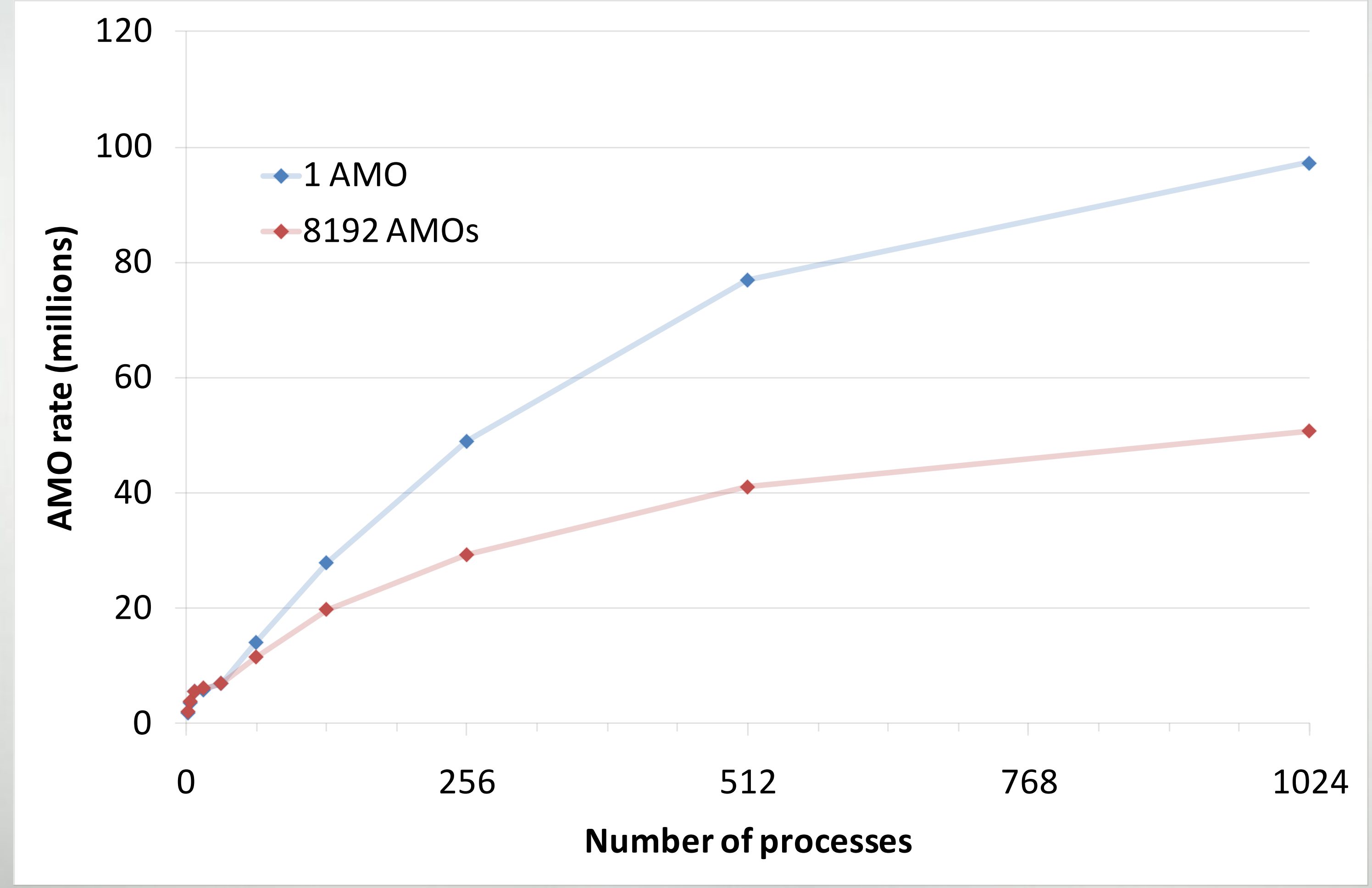
Cray XE6 Performance: put/get latency



Cray XE6 Performance: Strided put rate



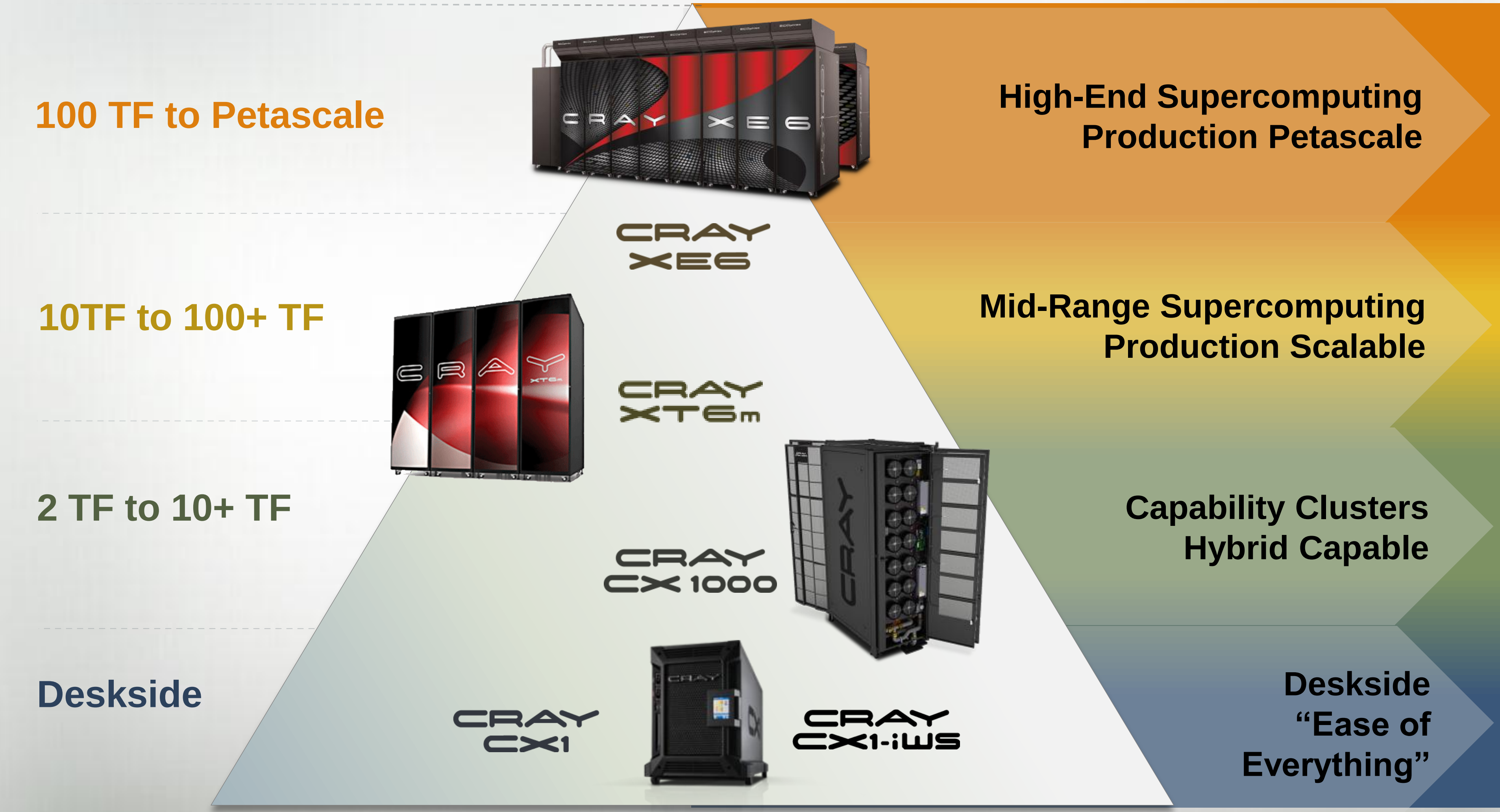
Cray XE6 Performance: AMO Rate



Cray XE6: Early Performance Data

- How will this translate to application performance ?
- Increased performance on strong scaling applications.
  - Gemini enables more nodes to be used more efficiently.
- Increased injection bandwidth and message rates
  - Improve performance on multi-core nodes
- Hardware put/get provides performance to PGAS applications
  - For the first time on mainstream systems

Cray XE6: Pinnacle of Cray’s product range



100 TF to Petascale

High-End Supercomputing
Production Petascale

CRAY XE6

10TF to 100+ TF

Mid-Range Supercomputing
Production Scalable

CRAY XT6m

2 TF to 10+ TF

Capability Clusters
Hybrid Capable

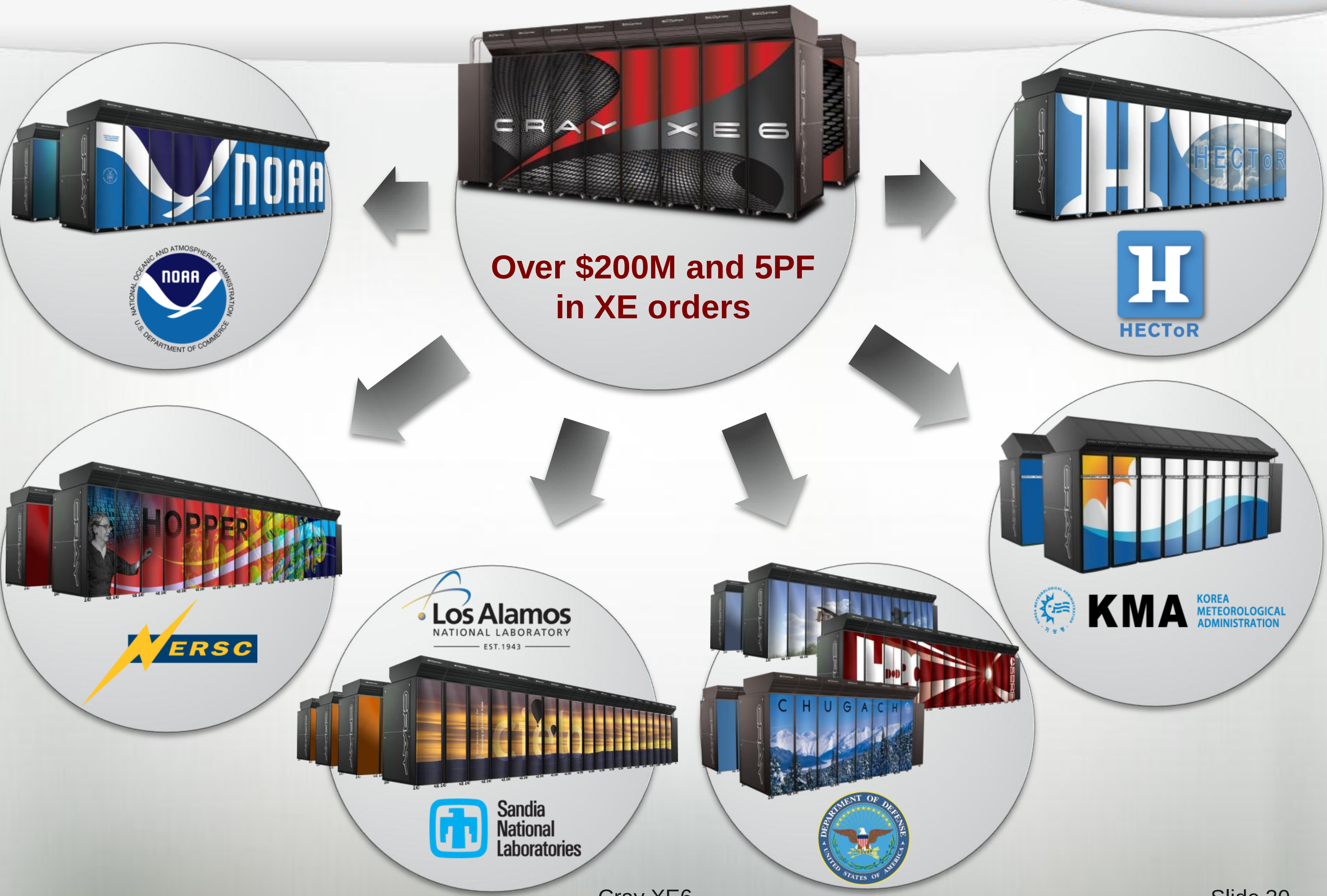
CRAY CX 1000

Deskside

Deskside
“Ease of
Everything”

CRAY CX1

CRAY CX1-iWS



Over $200M and 5PF in XE orders

- Thank you

- Further information: www.cray.com