Cray XE6 Architecture

John Shalf

NERSC XE6 User Training

Feb 7, 2011

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

BERKELEY LAB Lawrence Berkeley National Laboratory

NERSC-6
Grace "Hopper"

Cray XE6

Performance

1.2 PF Peak

1.05 PF HPL (#5)

Processor

AMD MagnyCours

2.1 GHz 12-core

8.4 GFLOPs/core

24 cores/node

32-64 GB DDR3-1333 per node

System

Gemini Interconnect (3D torus)

6392 nodes

153,408 total cores

I/O

2PB disk space

70GB/s peak I/O Bandwidth

U.S. DEPA
ENER

BERKELEY LAB
Lawrence Berkeley
National Laboratory

Systems	2009	2015 +1/-0	2018 +1/-0
System peak	**2 Peta**	**100-300 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	2 TF or 10TF
Node memory BW	25 GB/s	0.2TB/s or 0.5TB/s	0.4TB/s or 1TB/s
Node concurrency	12	O(100)	O(1k) or 10k
Total Node Interconnect BW	3.5 GB/s	100-200 GB/s 10:1 vs memory bandwidth 2:1 alternative	200-400GB/s (1:4 or 1:8 from memory BW)
System size (nodes)	18,700	50,000 or 500,000	O(100,000) or O(1M)
Total concurrency	225,000	O(100,000,000) *O(10)-O(50) to hide latency	O(billion) * O(10) to O (100) for latency hiding
Storage	15 PB	150 PB	500-1000 PB (>10x system memory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(1 day)

Systems	2009	2015 +1/-0	2018 +1/-0
System peak	**2 Peta**	**100-300 Peta**	**1 Exa**
Power	**6 MW**	**~15 MW**	**~20 MW**
System memory	0.3 PB	5 PB	64 PB (+)
Node performance	125 GF	0.5 TF or 7 TF	2 TF or 10TF
Node mem BW	25 GB	0.2TB/s or 0	0.4TB/s or 1TB/s
Node concurren	12	(100)	k
Total Node Interconn			0GB/s from memory
System size (no			) or O(1M)
Total concurrency			for latency hiding
Storage		0	0 PB (>10x sys emory is min)
IO	0.2 TB	10 TB/s	60 TB/s (how long to drain the machine)
MTTI	days	O(1day)	O(1 day)

All the bad stuff they are warning you about exascale is already happening!
(its only going to get worse)

Things to Watch For in this Training

Preparing yourself for future hardware trends

- **CPU Clock rates are stalled (not getting faster)**
  - # nodes is about the same, but # processors is growing exponentially
  - Time to start thinking of parallelism from node level *(cores will drive you crazy)*
  - Go to Hybrid Parallelism to tackle intra-node parallelism so you can focus on # of nodes parallelism rather than # of cores
- **Memory capacity not growing as fast as FLOPs**
  - Memory per node is still growing, but per core is diminishing
  - Threading (OpenMP) on node can help conserve memory
- **Diminishing BW/flop makes locality essential**
  - *Vertical locality:* Careful cache-blocking and use of prefetch
  - *Horizontal locality:* NUMA effects (memory affinity: must always be sure to access data where it was first touched)

Cabinet Design

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC National Energy Research Scientific Computing Center

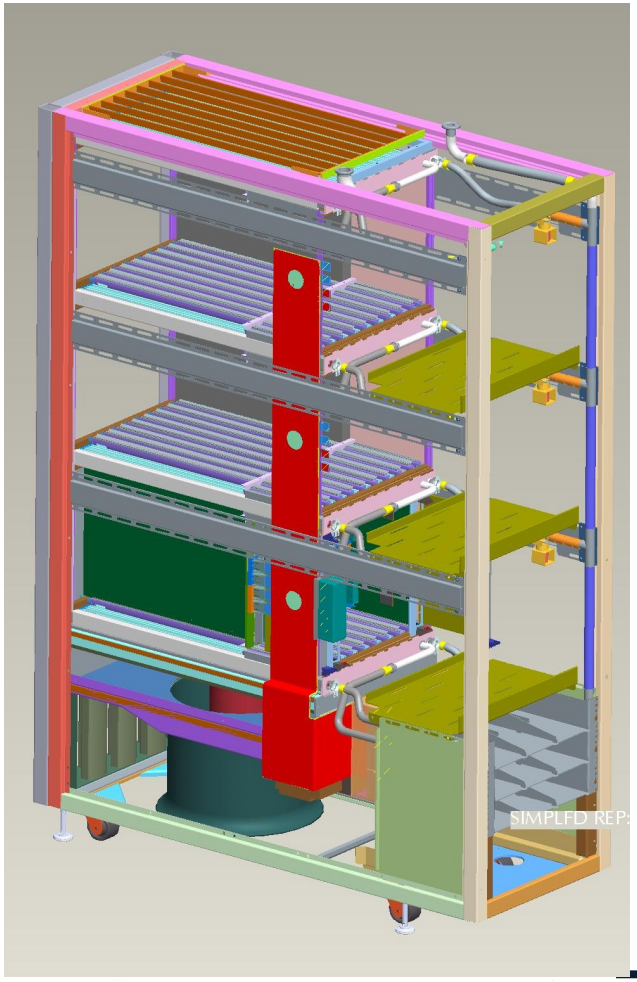
BERKELEY LAB Lawrence Berkeley National Laboratory

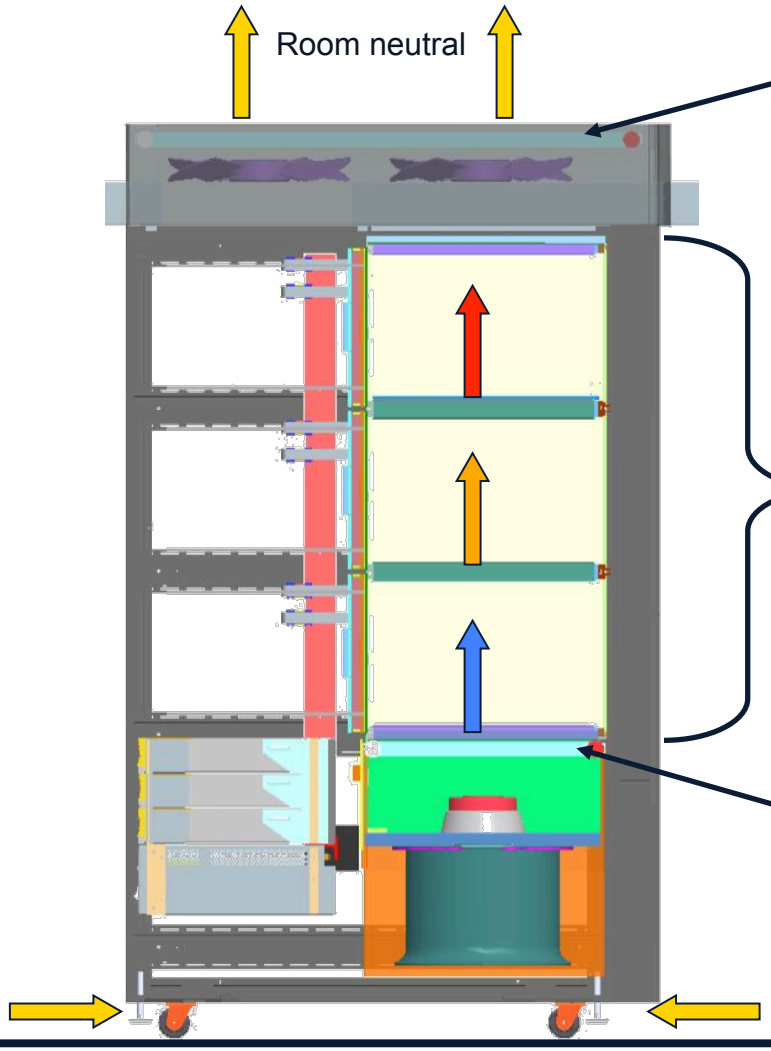
XE6 Cabinet Design

What's up with the hat??

Cray Cabinet Design

Energy Efficient Liquid Cooling





After-cooler assembly:
The extremely hot exhaust temperature of the HD air cooled chassis dramatically increases the capability of heat exchanger. This makes room neutral possible with single cooler assembly at exit.

HD Air Cooled Chassis:
Sandwich with R134a evaporators.

Pre-cooler assembly:
Required to operate in room environments over 20C.

Images Courtesy of Cray Inc.

Hopper Cooling Apparatus

XE6 Blade Design

U.S. DEPARTMENT OF ENERGY | Office of Science
NERSC National Energy Research Scientific Computing Center
BERKELEY LAB Lawrence Berkeley National Laboratory

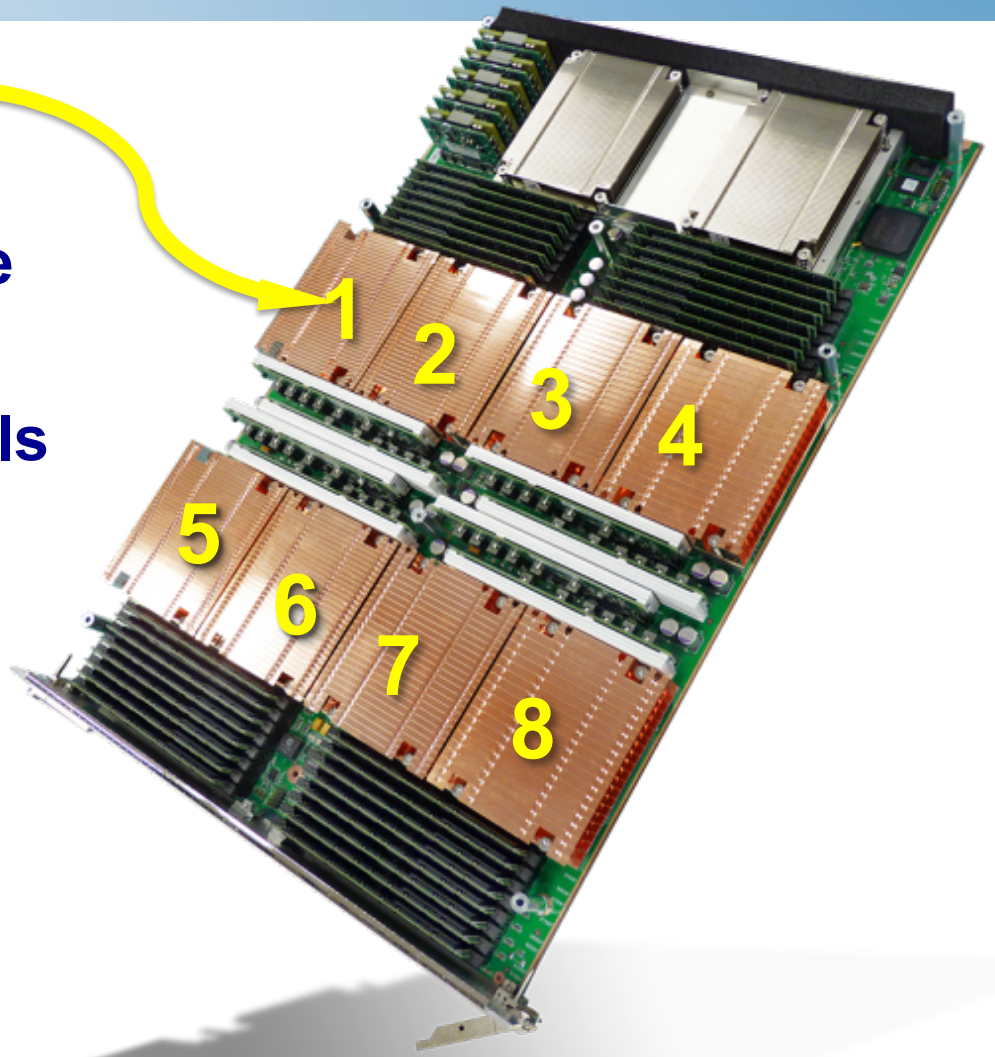
Cray XE6 Compute Blade

- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor

Images Courtesy of Cray Inc.

Cray XE6 Compute Blade

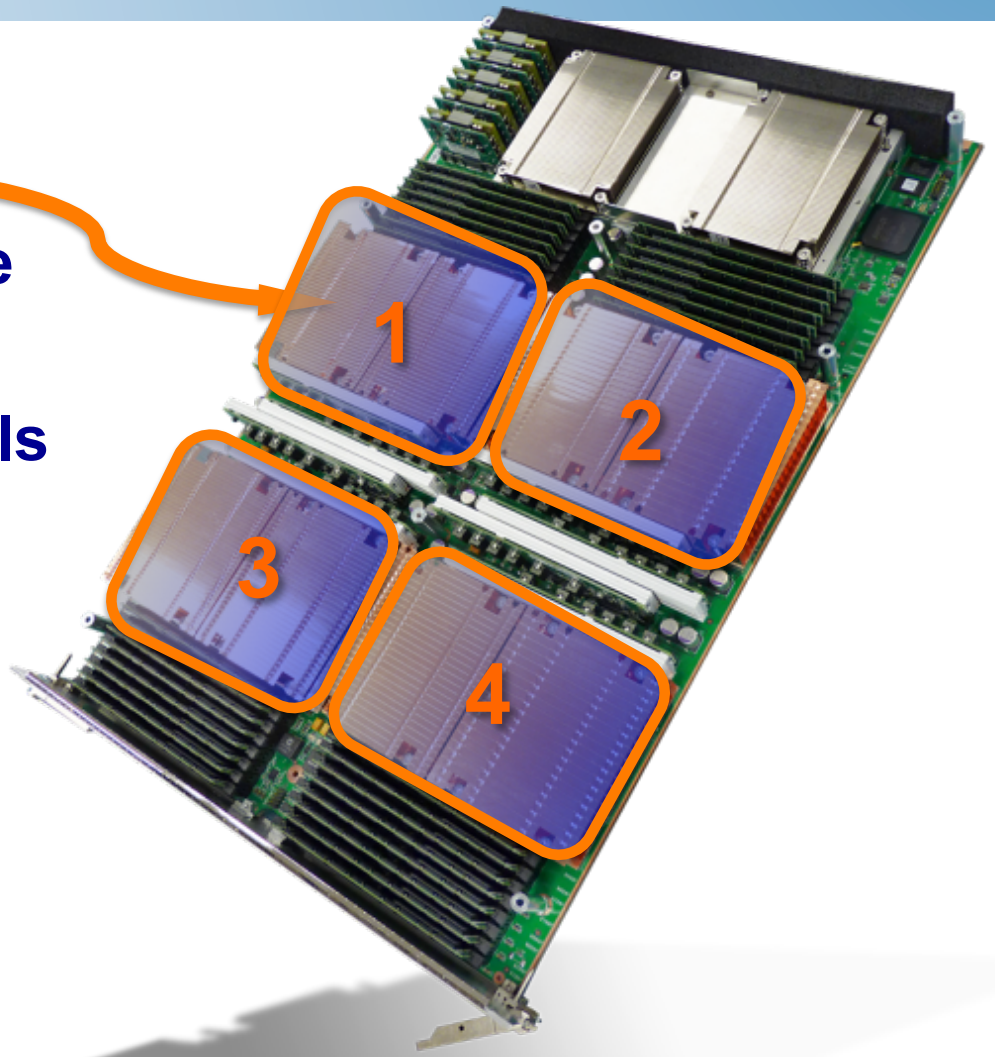
- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



Images Courtesy of Cray Inc.

Cray XE6 Compute Blade

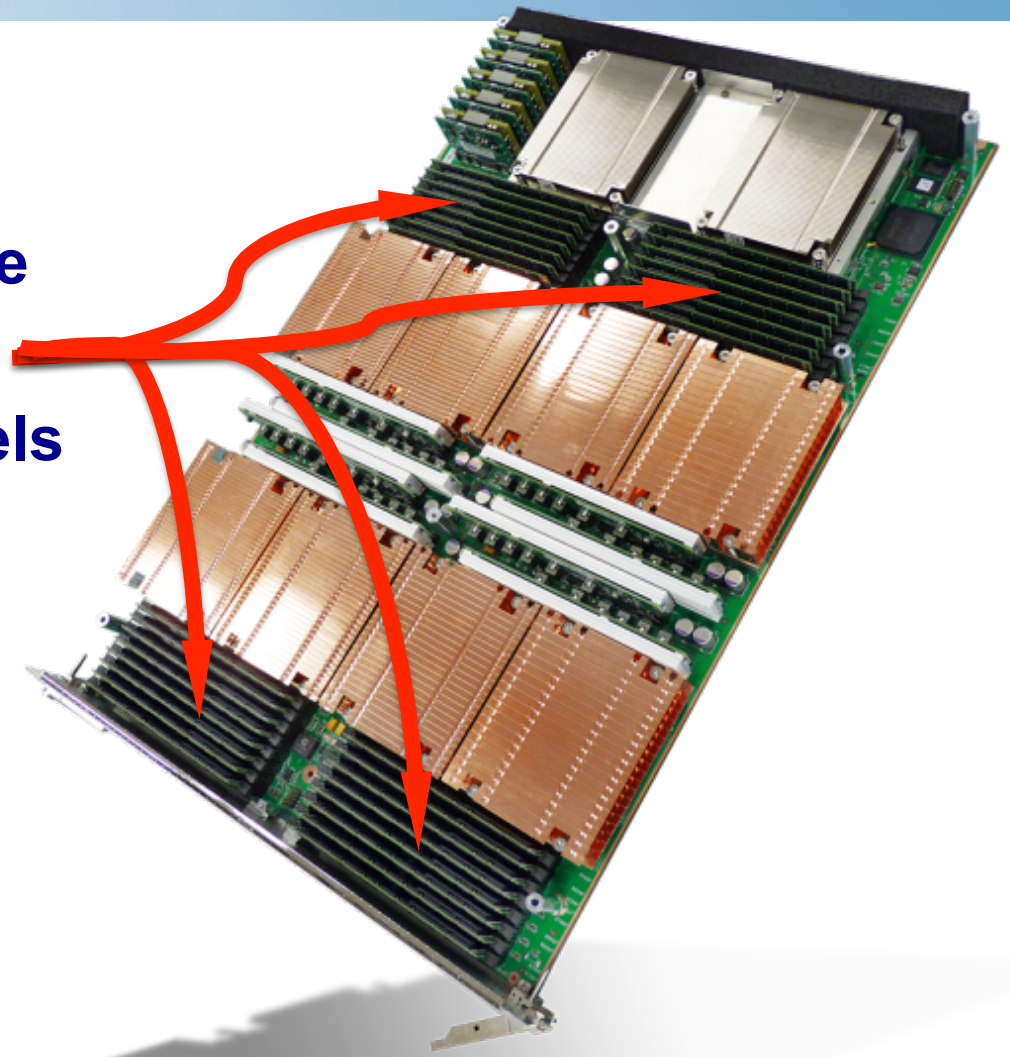
- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



Images Courtesy of Cray Inc.

Cray XE6 Compute Blade

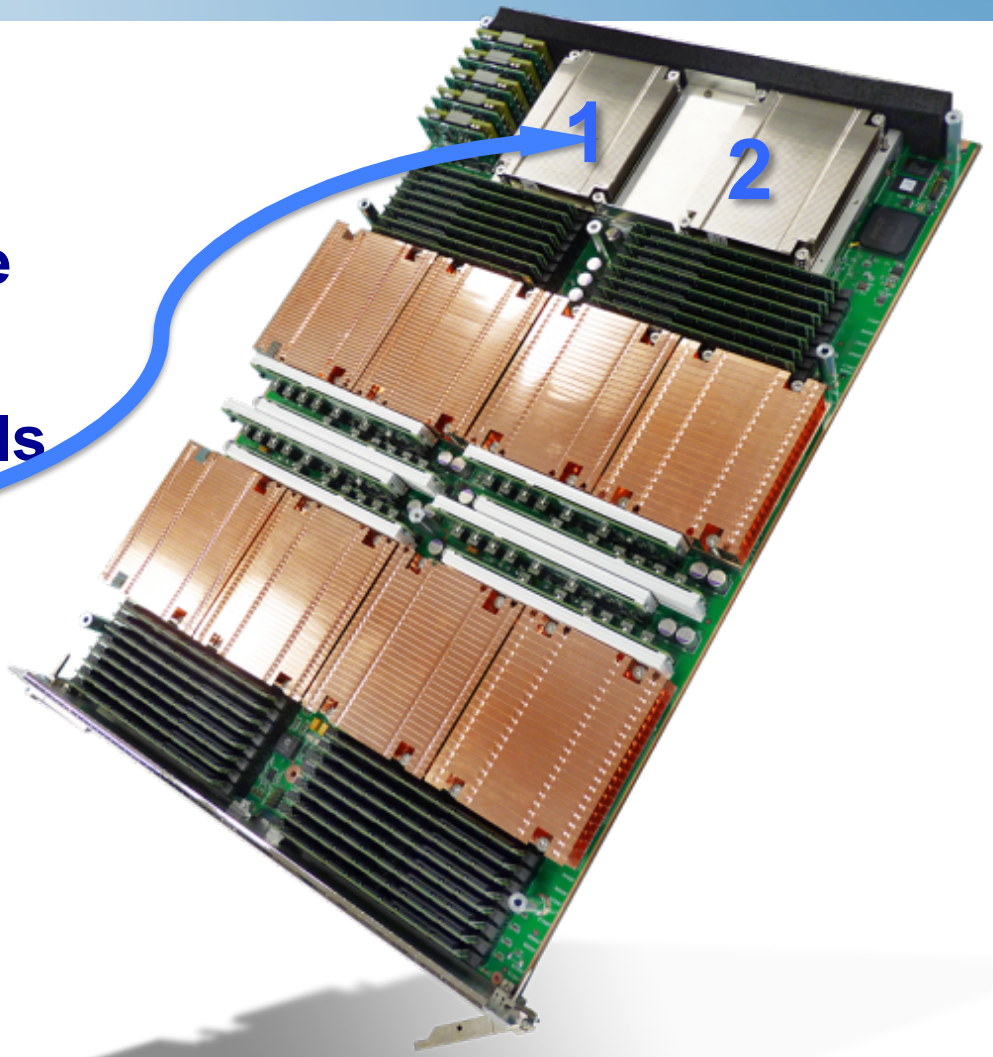
- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



Images Courtesy of Cray Inc.

Cray XE6 Compute Blade

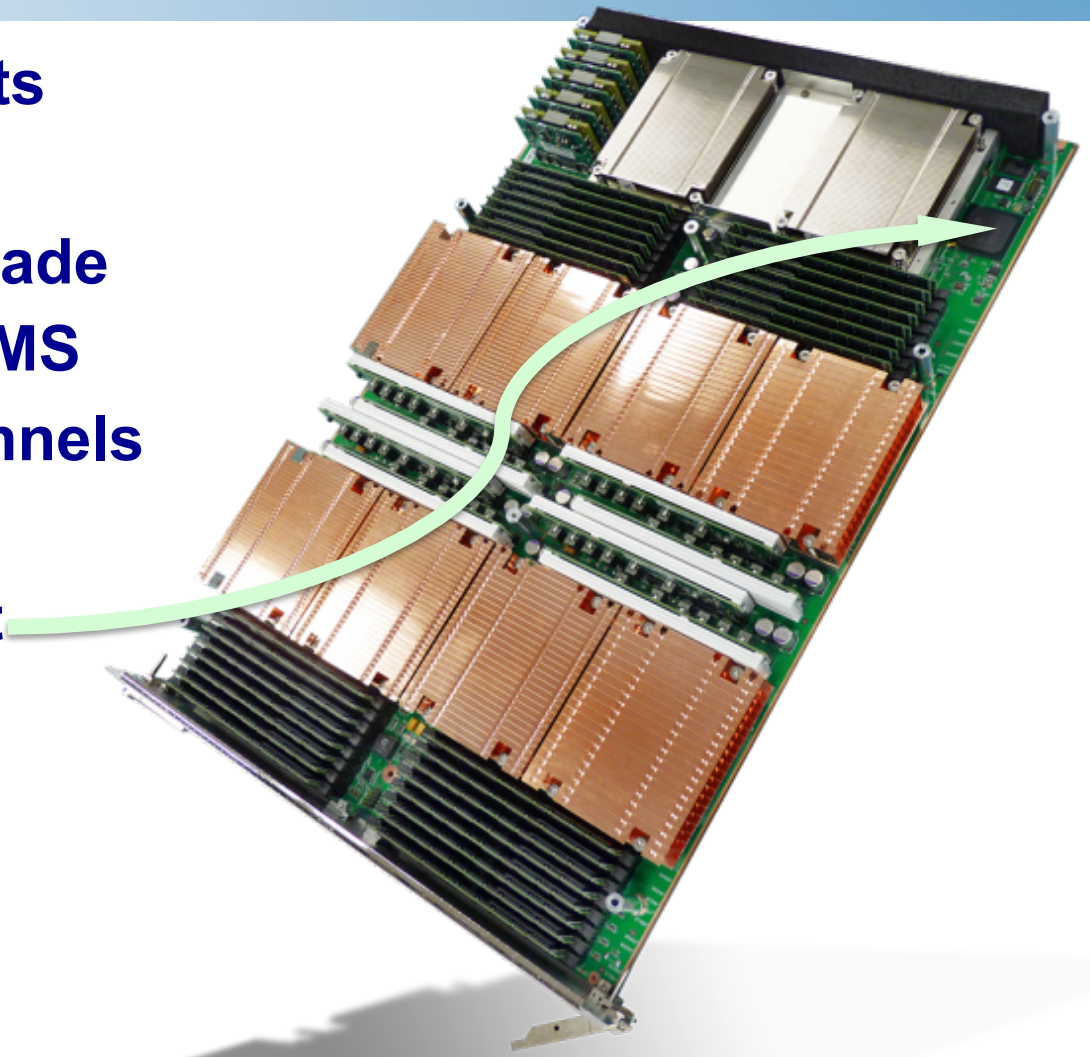
- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



Images Courtesy of Cray Inc.

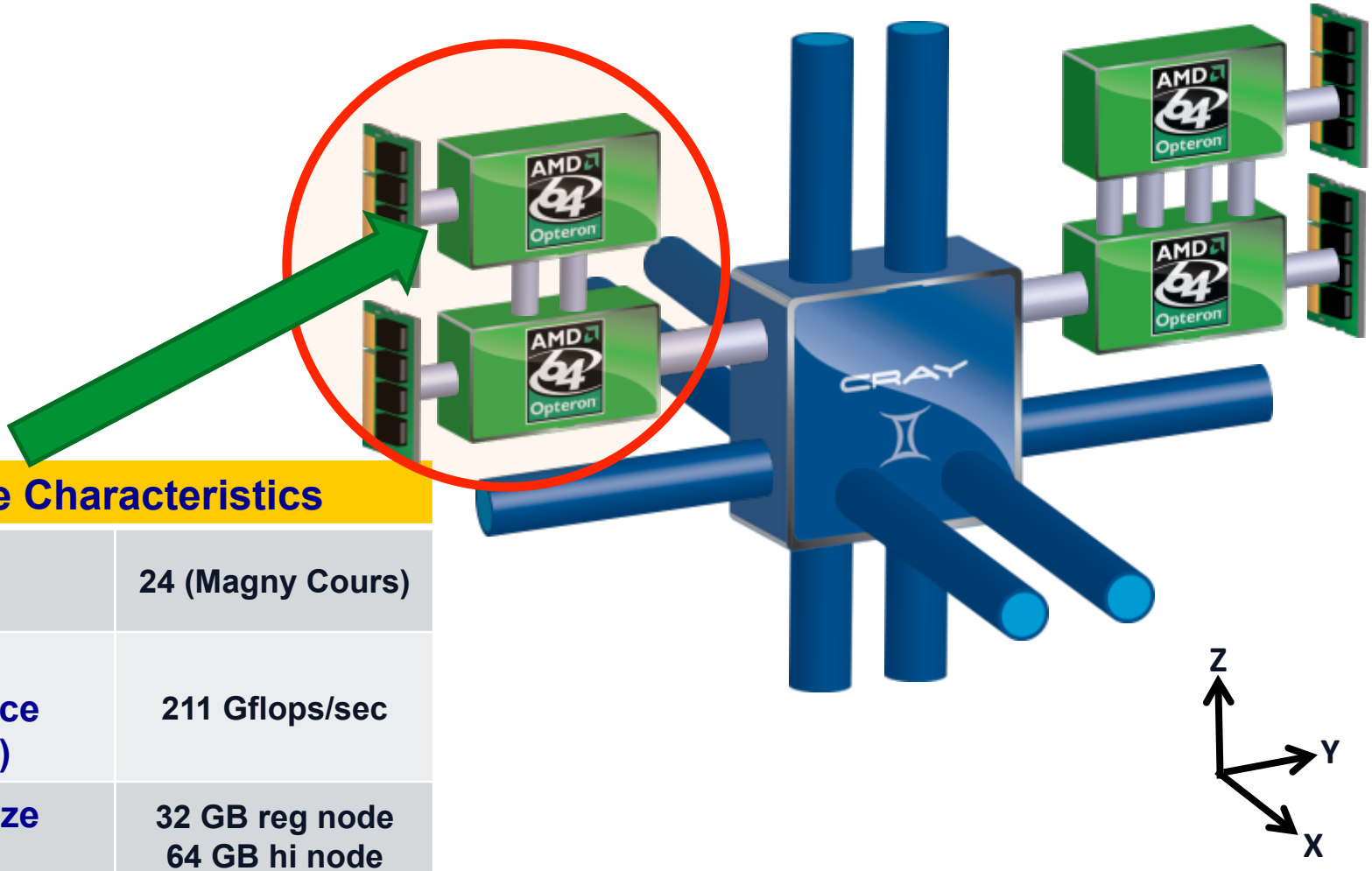
Cray XE6 Compute Blade

- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



Images Courtesy of Cray Inc.

Cray XE6 Compute Node

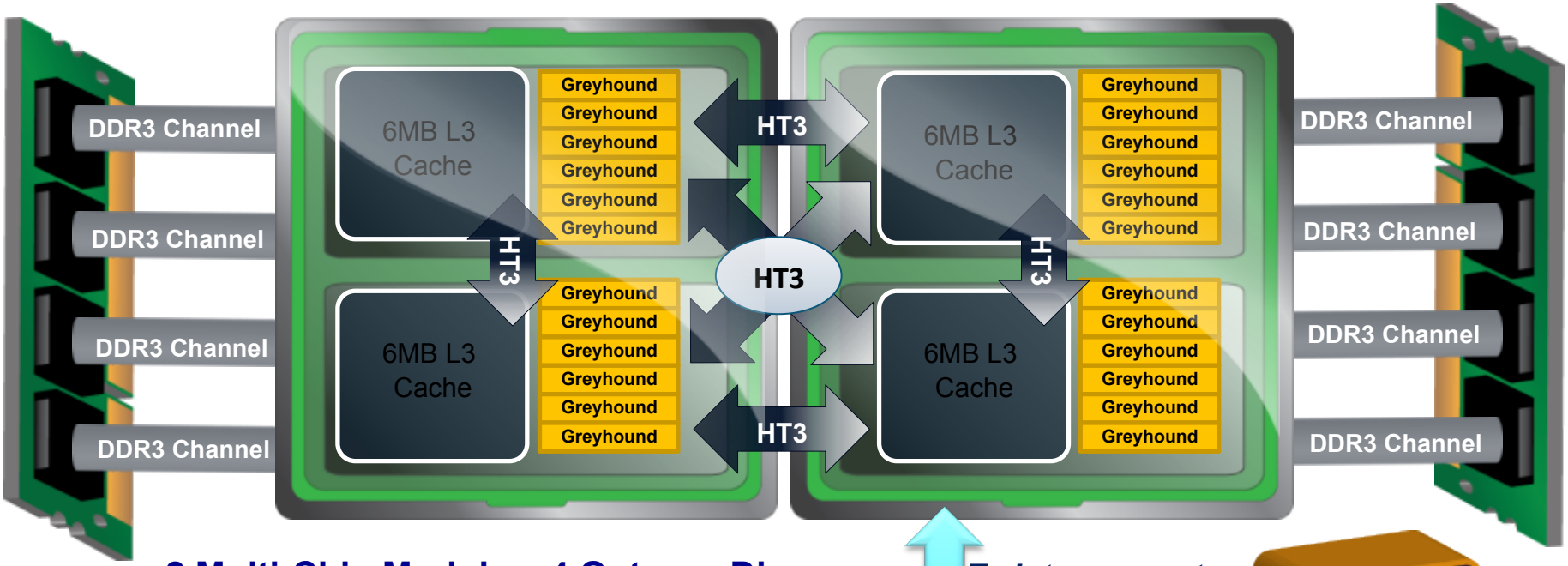


Node Characteristics	
Number of Cores	24 (Magny Cours)
Peak Performance MC-12 (2.2)	211 Gflops/sec
Memory Size	32 GB reg node 64 GB hi node
Memory Bandwidth (Peak)	83.5 GB/sec

Images Courtesy of Cray Inc.

Images Courtesy of Cray Inc.

XE6 Node Details: 24-core Magny Cours



- **2 Multi-Chip Modules, 4 Opteron Dies**
- **8 Channels of DDR3 Bandwidth to 8 DIMMs**
- **24 (or 16) Computational Cores**
  - 64 KB L1 and 512 KB L2 caches for each core
  - 6 MB of shared L3 cache on each die
- **Dies are fully connected with HT3**
- **Snoop Filter Feature Allows 4 Die SMP to scale well**

U.S. DEPARTMENT OF ENERGY | Office of Science

Images Courtesy of Cray Inc.

BERKELEY LAB Lawrence Berkeley National Laboratory

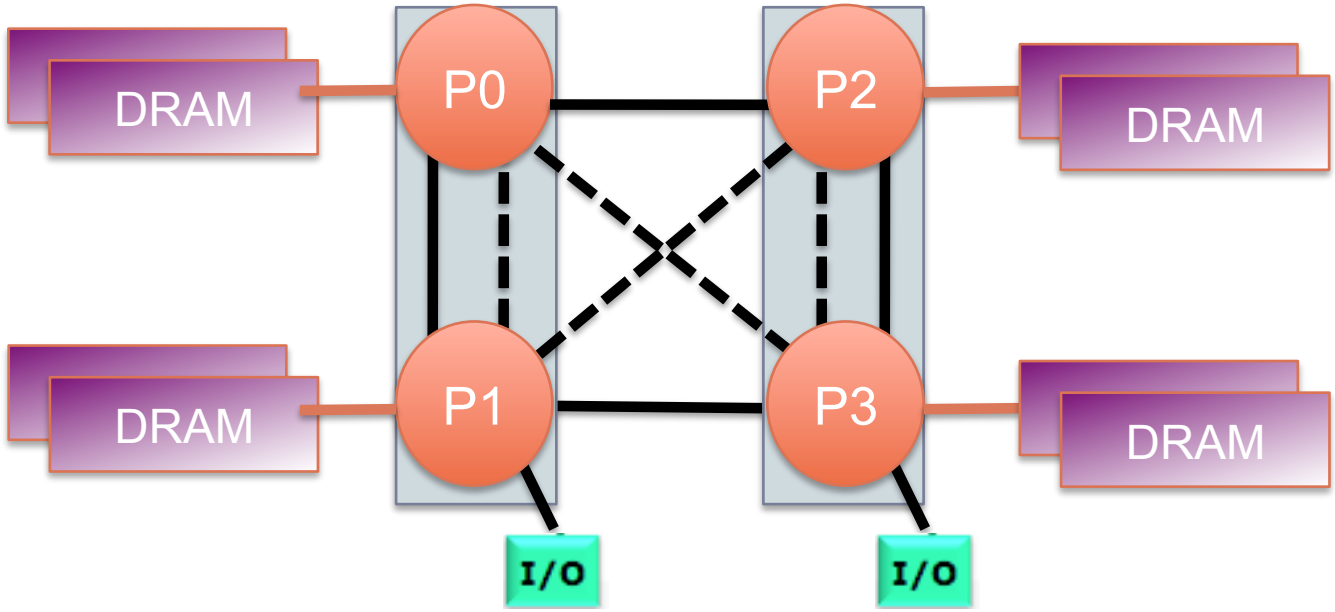
Hopper Node Topology
Understanding NUMA Effects

- **Heterogeneous Memory access between dies**
- **"First touch" assignment of pages to memory.**

2xDDR1333 channel
21.328 GB/s

3.2GHz x8 lane HT
6.4 GB/s bidirectional

3.2GHz x16 lane HT
12.8 GB/s bidirectional



- **Locality is key** *(just as per Exascale Report)*
- **Only *indirect* locality control with OpenMP**

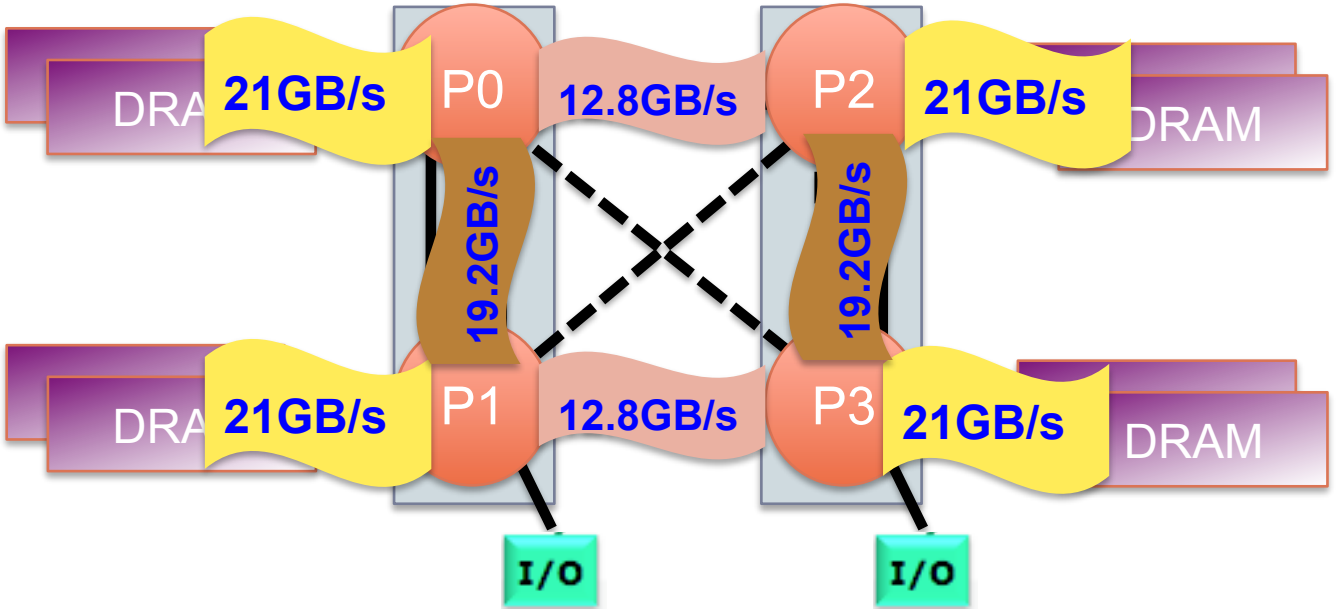
Hopper Node Topology
Understanding NUMA Effects

- **Heterogeneous Memory access between dies**
- **"First touch" assignment of pages to memory.**

2xDDR1333 channel
21.328 GB/s

3.2GHz x8 lane HT
6.4 GB/s bidirectional

3.2GHz x16 lane HT
12.8 GB/s bidirectional



- **Locality is key** *(just as per Exascale Report)*
- **Only *indirect* locality control with OpenMP**

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

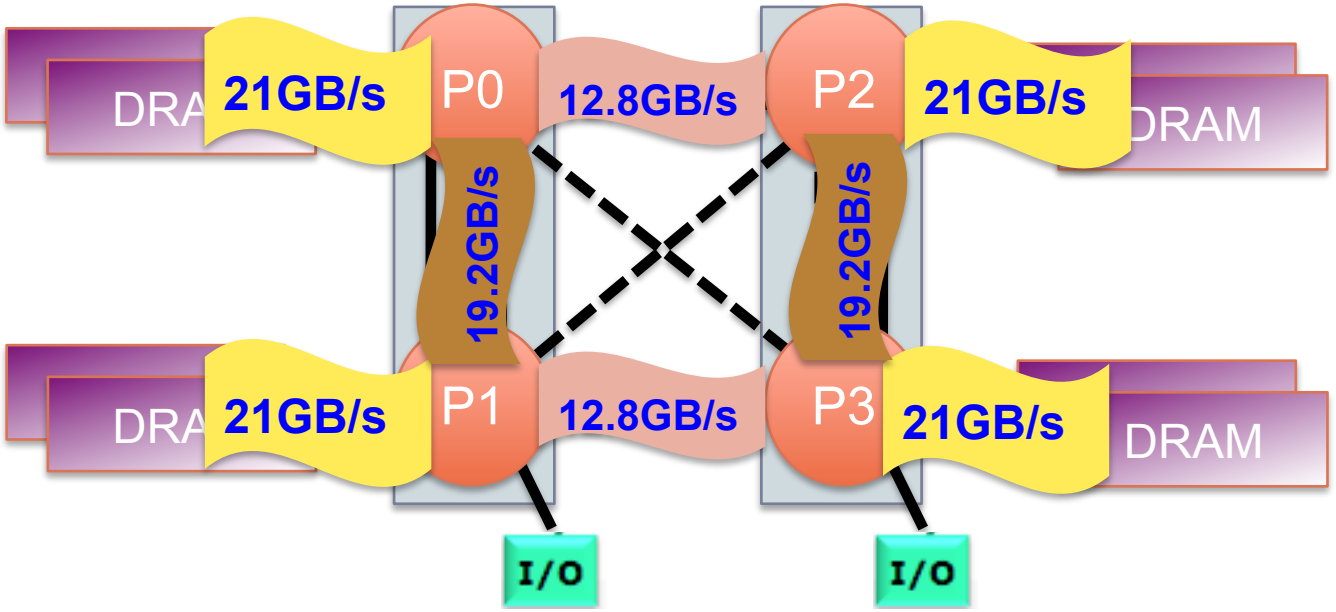
Hopper Node Topology
Understanding NUMA Effects

- **Heterogeneous Memory access between dies**
- **"First touch" assignment of pages to memory.**

2xDDR1333 channel
21.328 GB/s

3.2GHz x8 lane HT
6.4 GB/s bidirectional

3.2GHz x16 lane HT
12.8 GB/s bidirectional



- **Locality is key** *(just as per Exascale Report)*

Launch threads on "NUMA Nodes" (see COE talk)

Interconnect

Evolution of Cray Interconnects

Seastar (Franklin)

- **Designed for efficient RDMA for MPI/2-sided messaging**
- **Performance**
  - Latency 8-9usec
  - Injection Bandwidth 1.2GB/s
  - Link Bandwidth 4GB/s per link
- **Non-adaptive routing**
- **Built for scalability to 250K+ cores**

Gemini (Hopper)

- **Designed for global address space and PGAS**
- **Performance**
  - 100x improvement in message throughput
  - 3x improvement in latency
  - Link Bandwidth 9.3GB/s
- **Adaptive Routing**
  - improved fault tolerance
- **Scalability to 1M+ cores**

Cray Gemini

- **3D Torus network**
- **Supports 2 Nodes per ASIC**
- **168 GB/sec routing capacity**
- **Scales to over 100,000 network endpoints**

–Link Level Reliability and Adaptive Routing

–Advanced Resiliency Features

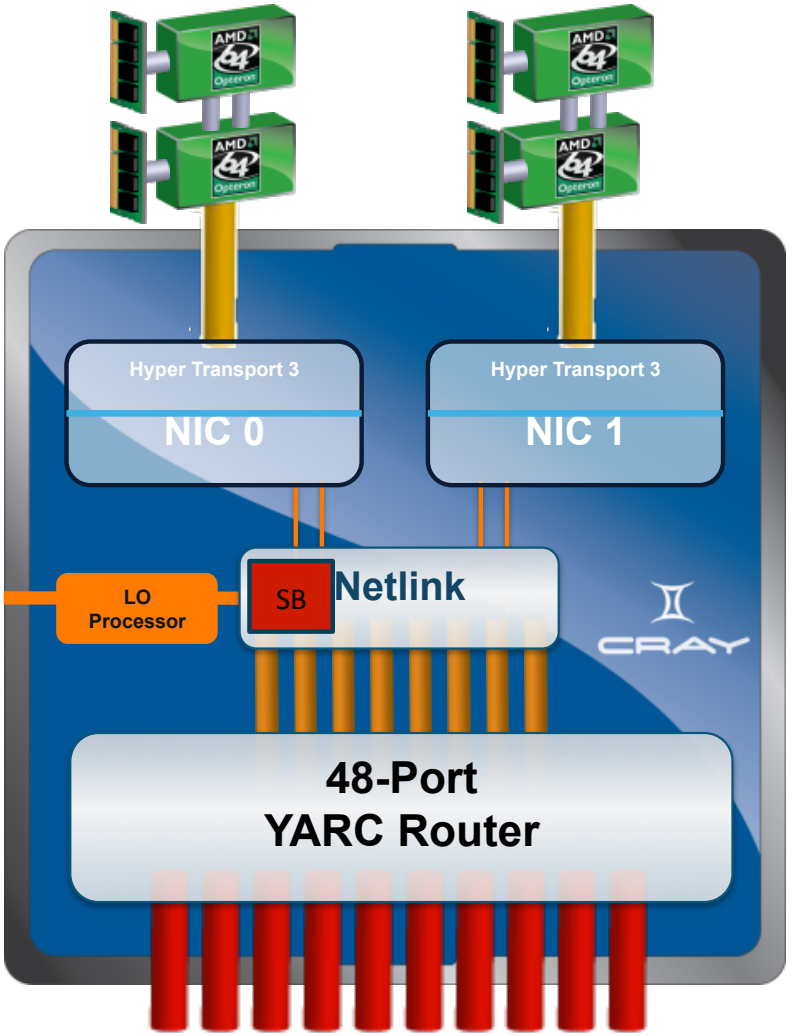
- **Provides global address space**
- **Advanced NIC designed to efficiently support**

–MPI*: Millions of messages/second*

–One-sided MPI

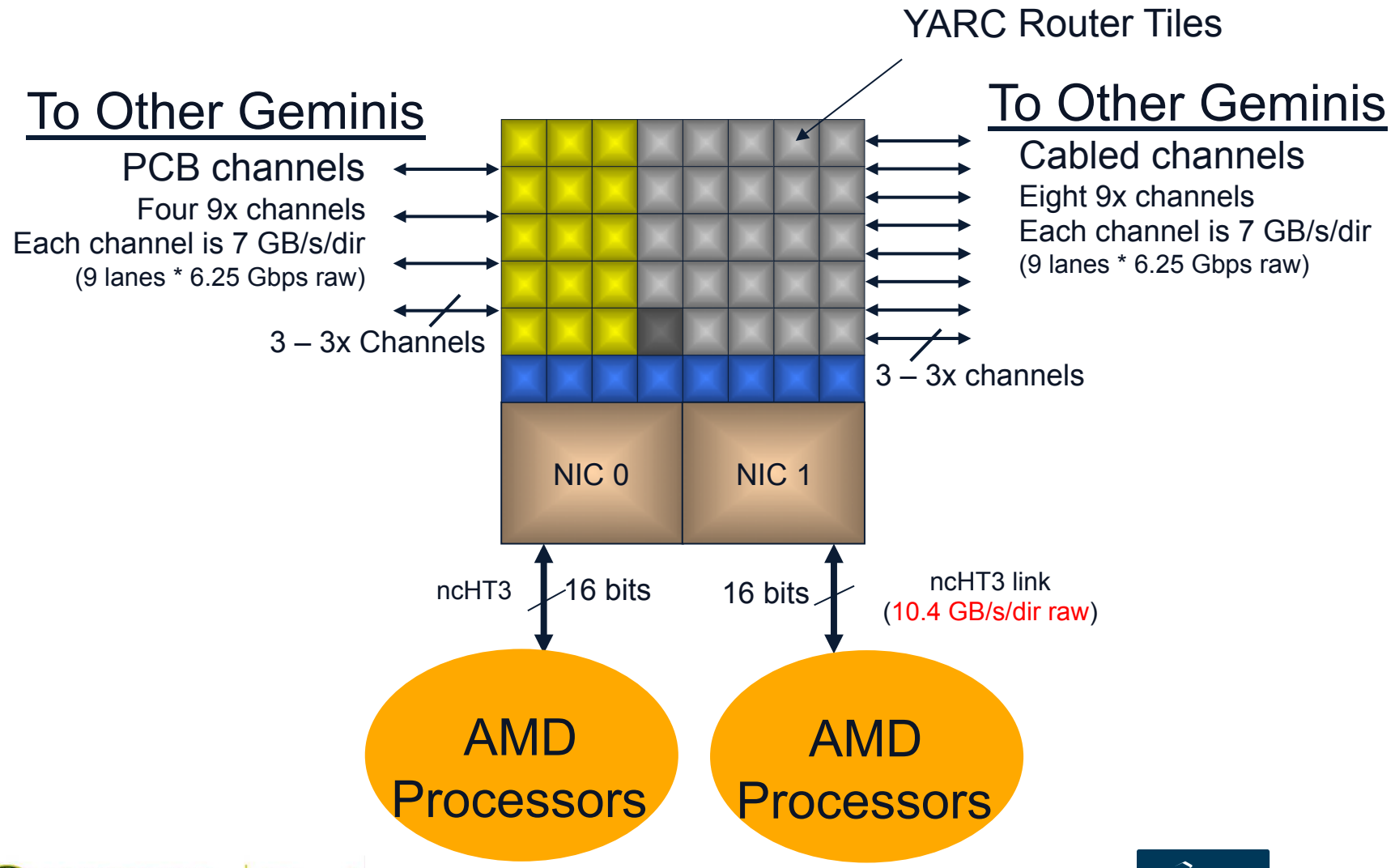
–UPC, FORTRAN 2008 with coarrays, shmem

–Global Atomics



Images Courtesy of Cray Inc.

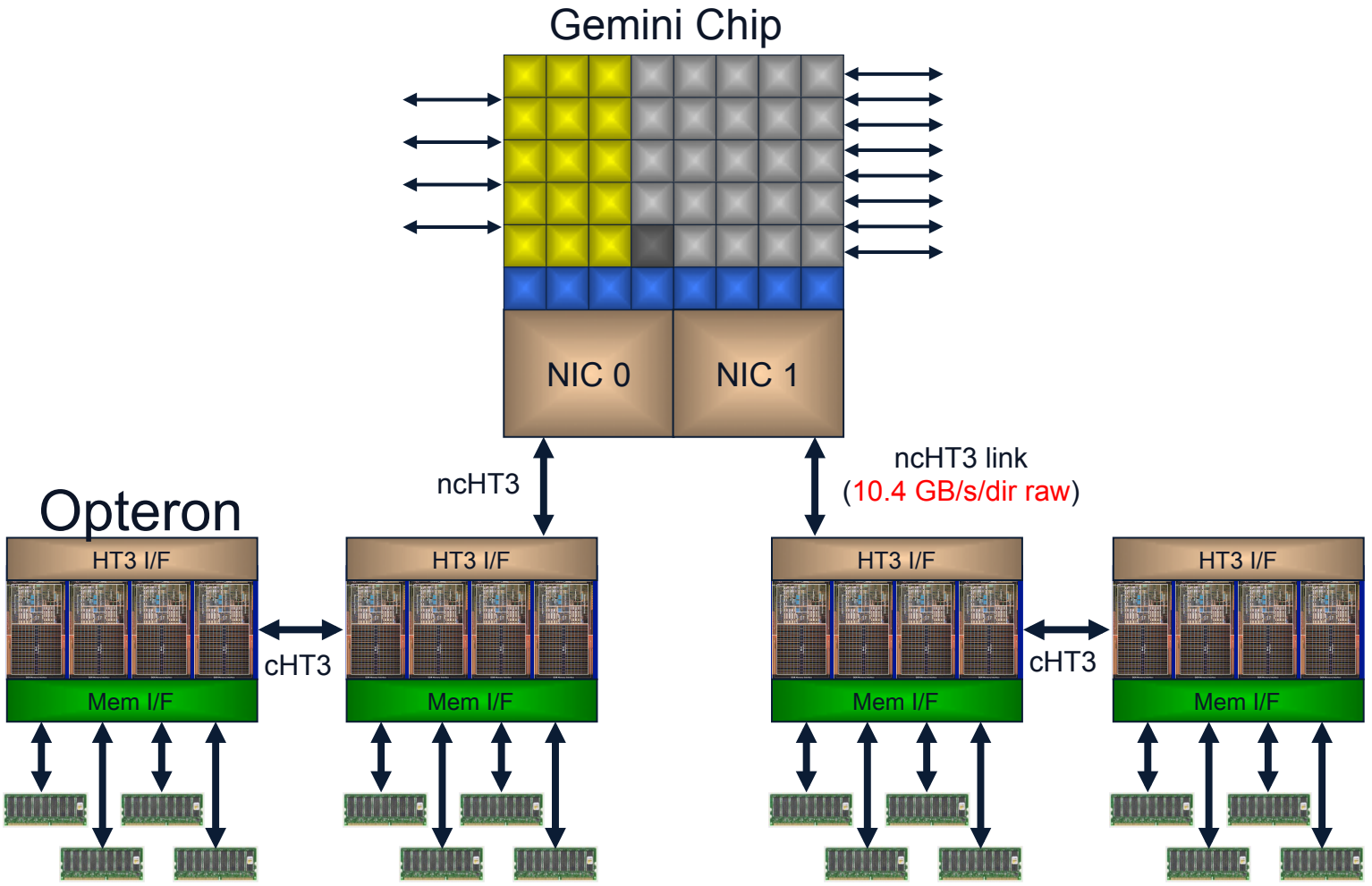
Gemini / Baker NIC



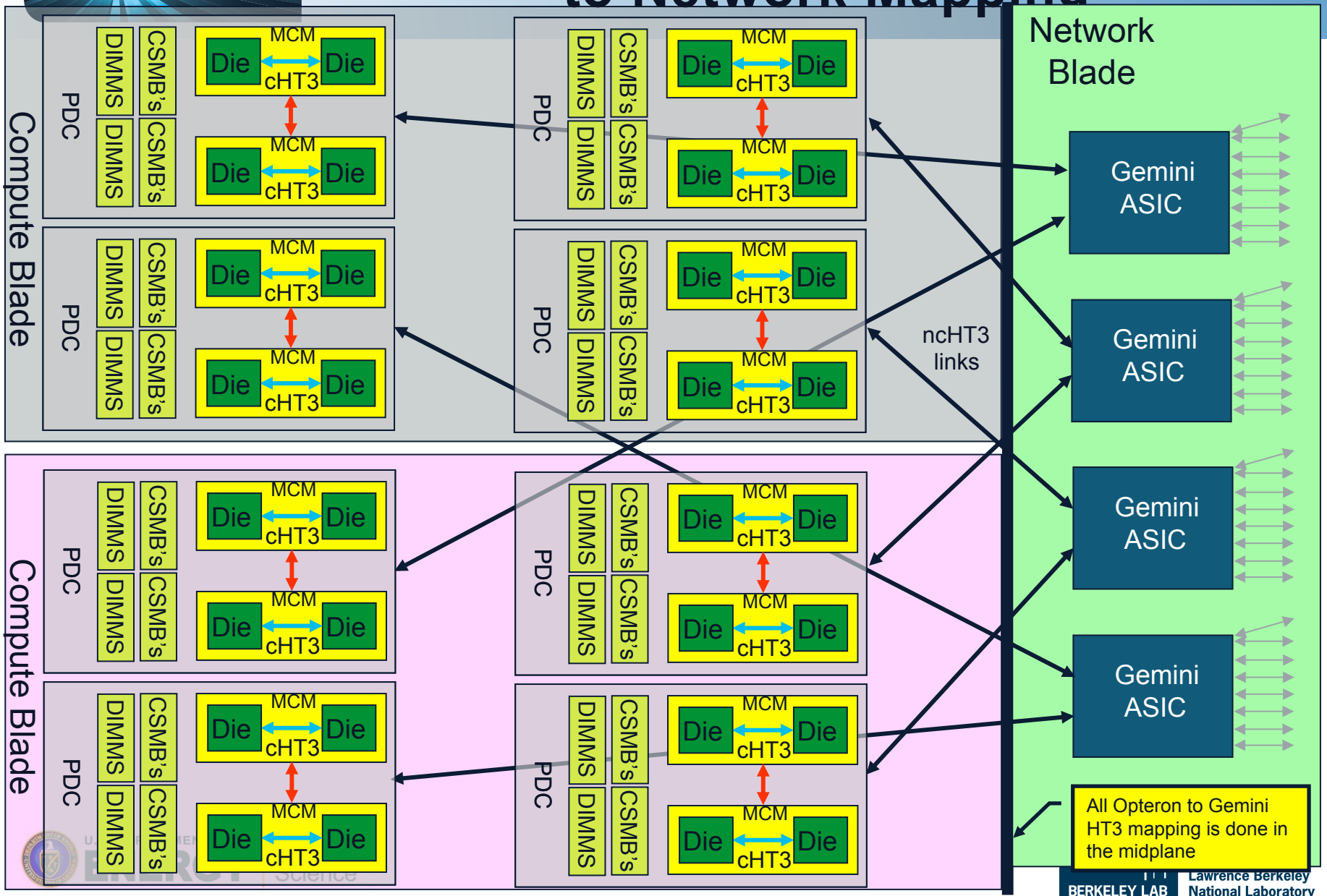
Images Courtesy of Cray Inc.

Gemini / Baker NIC

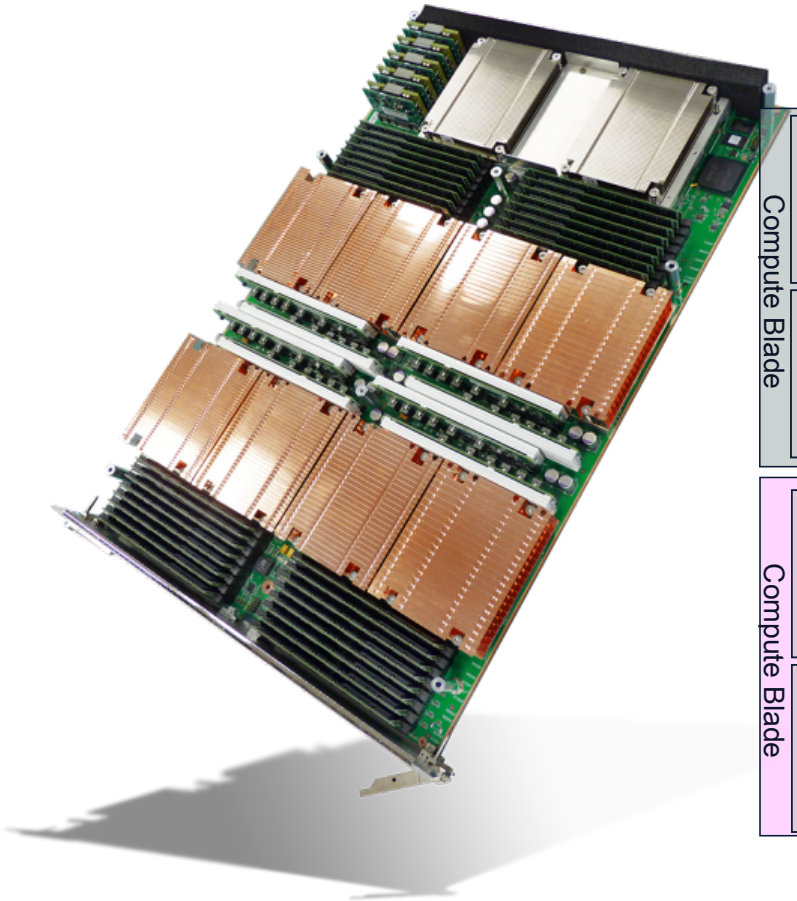
two compute nodes per NIC

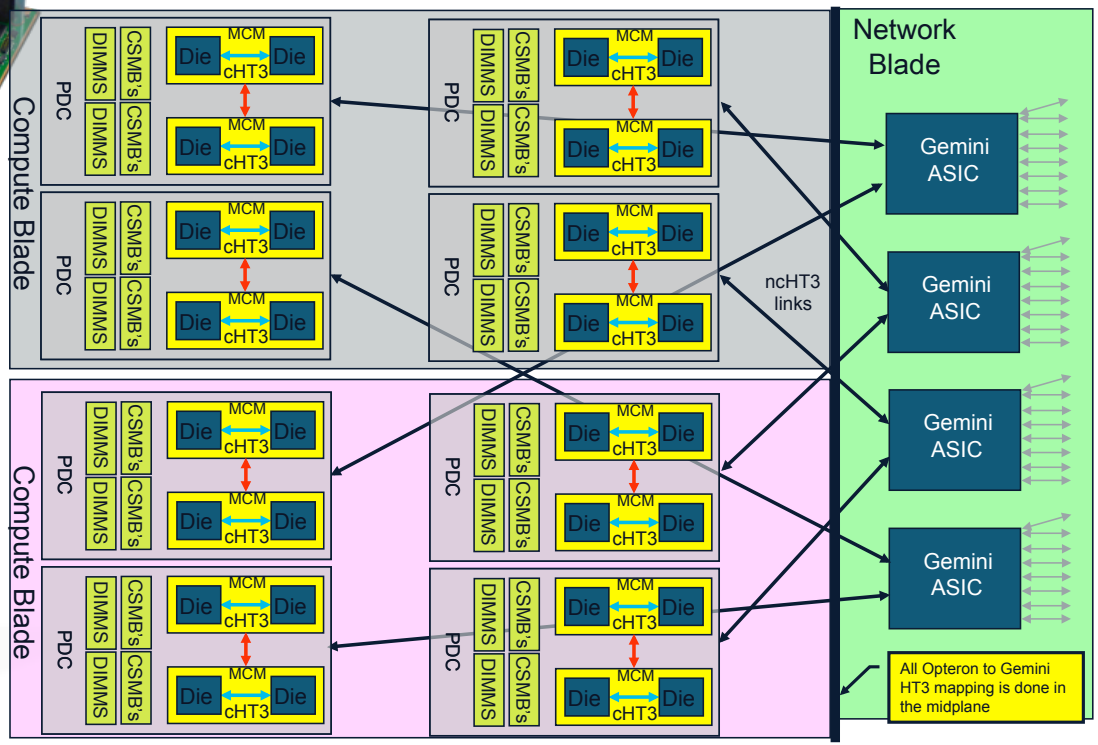


2 Baker Compute Blades to Network Mapping



2 Baker Compute Blades to Network Mapping

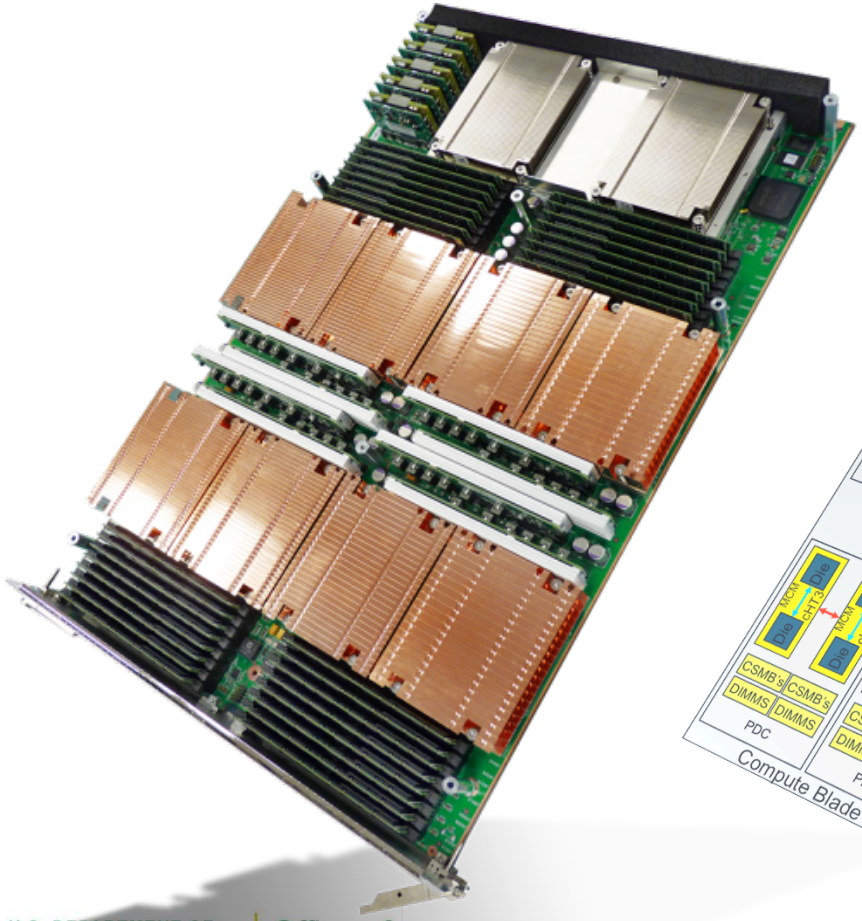


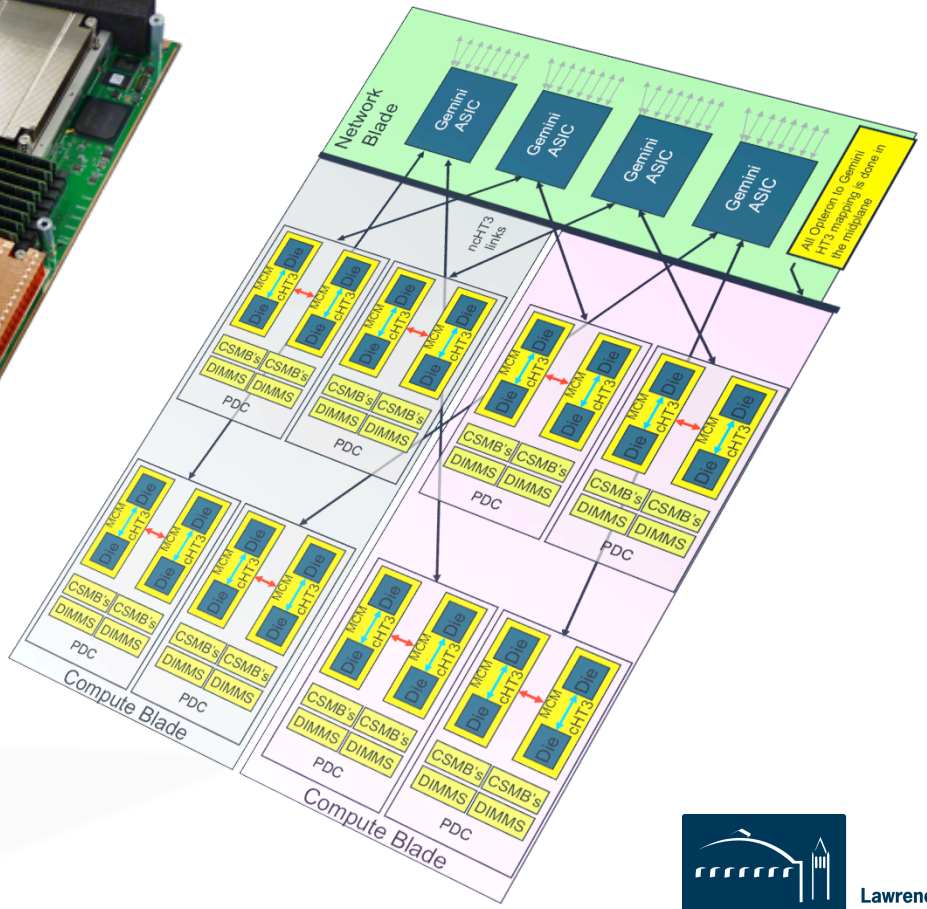


U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

2 Baker Compute Blades to Network Mapping

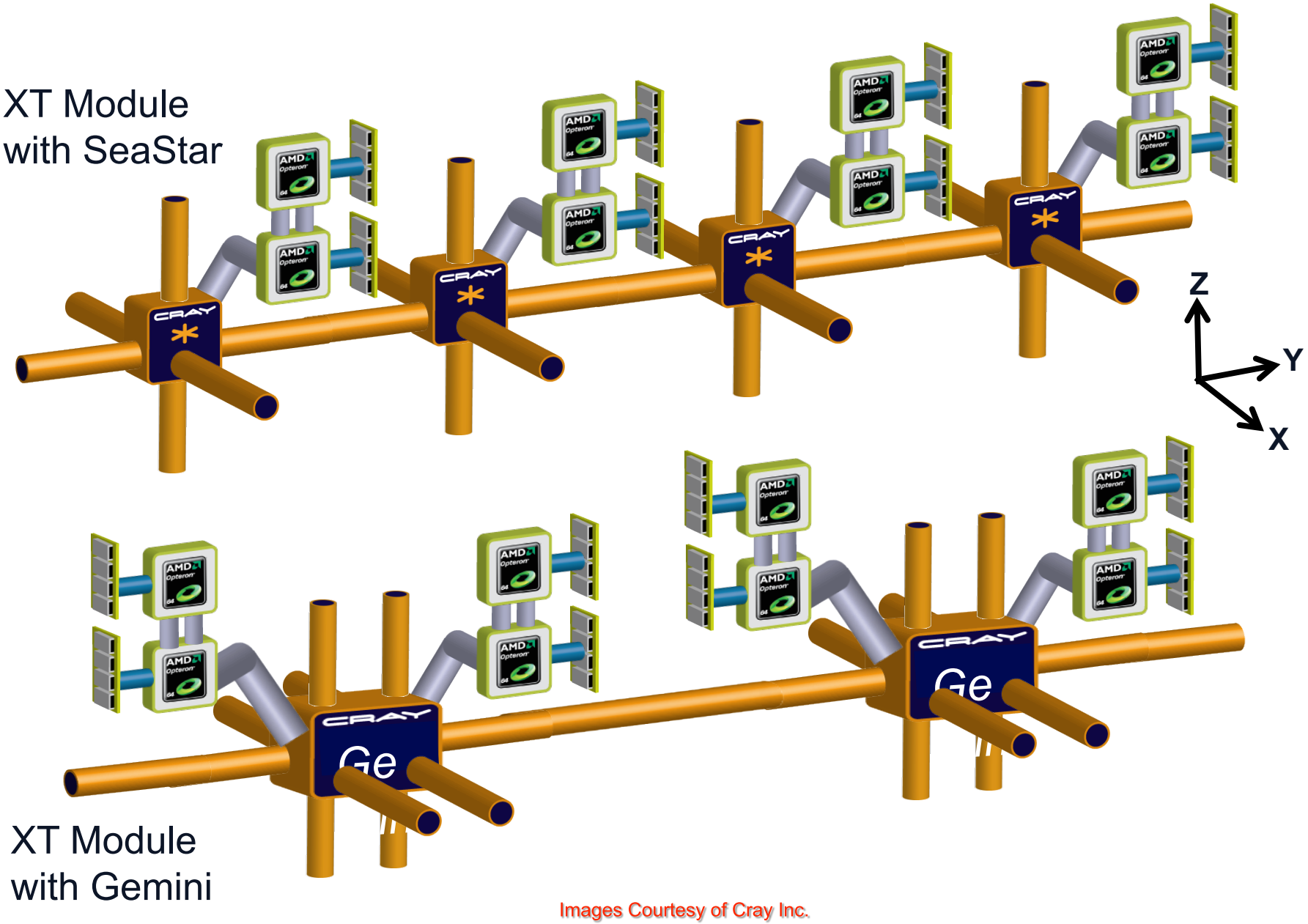




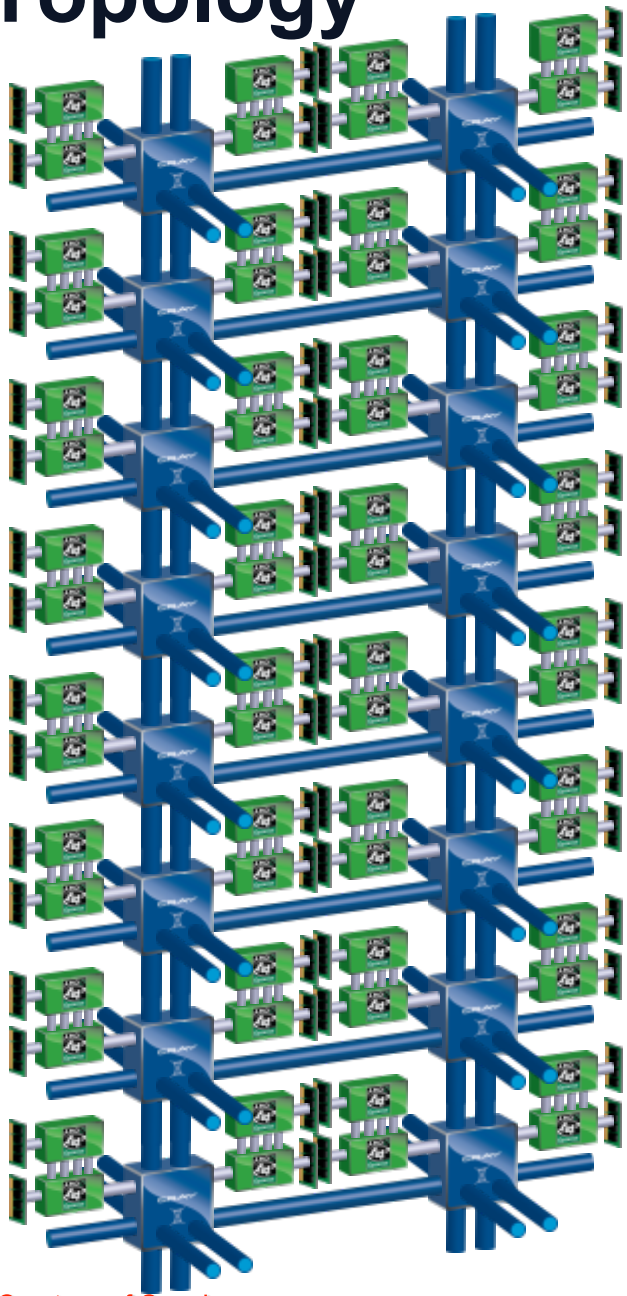
U.S. DEPARTMENT OF ENERGY | Office of Science

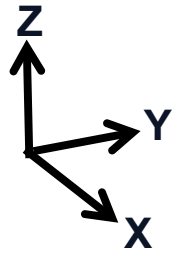
BERKELEY LAB Lawrence Berkeley National Laboratory

Gemini vs SeaStar – Topology



Cray XE6 Chassis Topology





Images Courtesy of Cray Inc.

Wiring up the Cabinets

Images Courtesy of Cray Inc.

I/O and External Login Nodes

U.S. DEPARTMENT OF ENERGY | Office of Science
NERSC National Energy Research Scientific Computing Center
BERKELEY LAB Lawrence Berkeley National Laboratory

DVS and External I/O

- **DVS nodes provide a gateway to I/O and login services (breaking out from the torus)**

- **External login nodes are "outside" of the torus, so you can still get work done while the main system is down**

- **Externalized I/O provides a gateway to the NERSC Global Filesystem (GPFS), and eventually will enable cleaner separation of scratch filesystems**
  - New environment variables to control striping across DVS servers in addition to Lustre striping
  - Ongoing work to control GPFS tunable parameters via DVS

Observations / Conclusions

What you should try to get out of the rest of this training!

(things to watch for)

U.S. DEPARTMENT OF ENERGY | Office of Science
NERSC National Energy Research Scientific Computing Center
BERKELEY LAB Lawrence Berkeley National Laboratory

Things to Watch For in this Training

Preparing yourself for future hardware trends

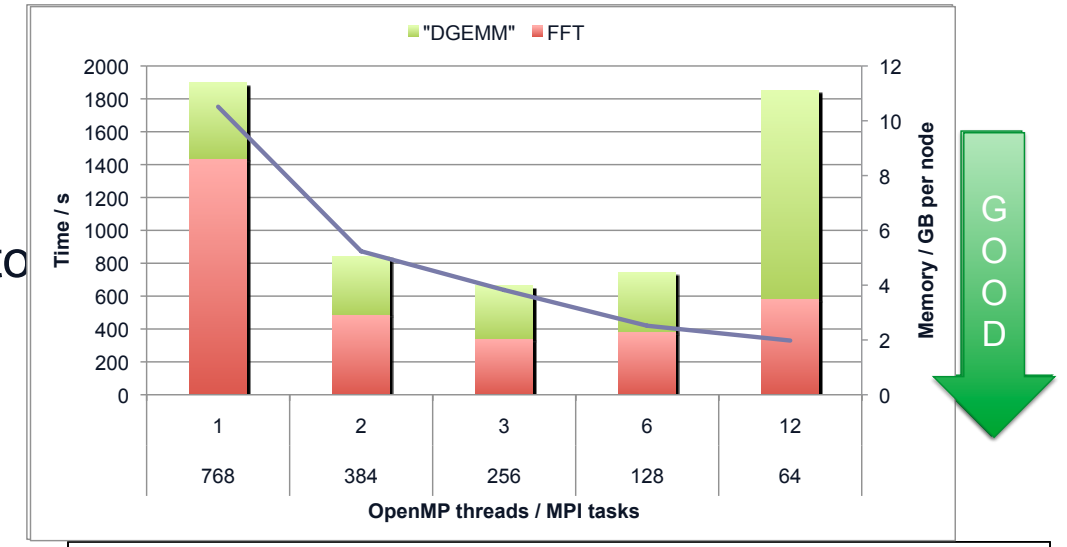
- **CPU Clock rates are stalled (not getting faster)**
  - **# nodes is about the same, but # processors is growing exponentially**
  - **Time to start thinking of parallelism from node level** *(thinking about the exponential growth rate in cores will drive you crazy)*
  - **Go to Hybrid Parallelism to tackle intra-node parallelism so you can focus on # of nodes parallelism rather than # of cores**

- **Memory capacity not growing as fast as FLOPs**
  - **Memory per node is still growing, but per core is diminishing**
  - **Threading (OpenMP) on node can help conserve memory**

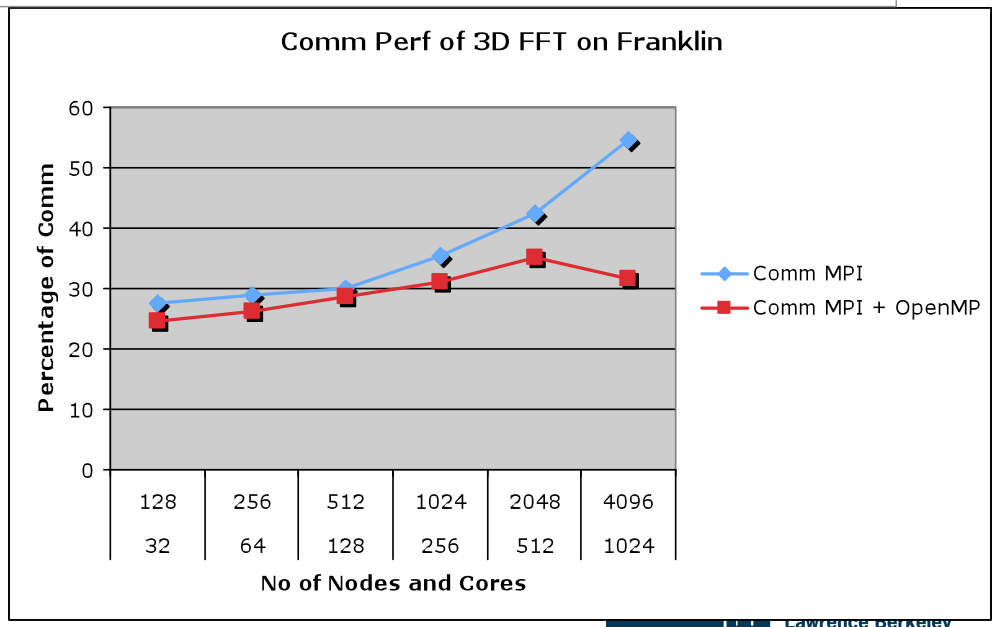
- **Diminishing BW/flop makes locality essential**
  - ***Vertical locality:*** **Careful cache-blocking and use of prefetch**
  - ***Horizontal locality:*** **NUMA effects (memory affinity: must always be sure to access data where it was first touched)**

Example OpenMP for PARATEC

- **MPI+OMP Hybrid**
  - Reduces memory footprint
  - Increases performance up to NUMA-node limit

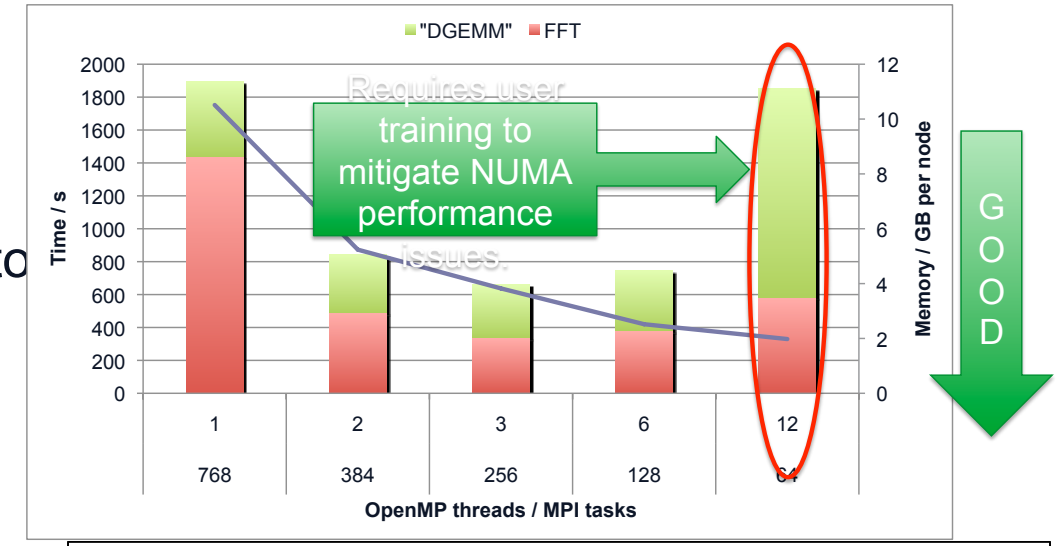


- **Hybrid Model improves 3D FFT communication performance**
  - Enables node to send larger messages
  - Substantial improvements in communications efficiency

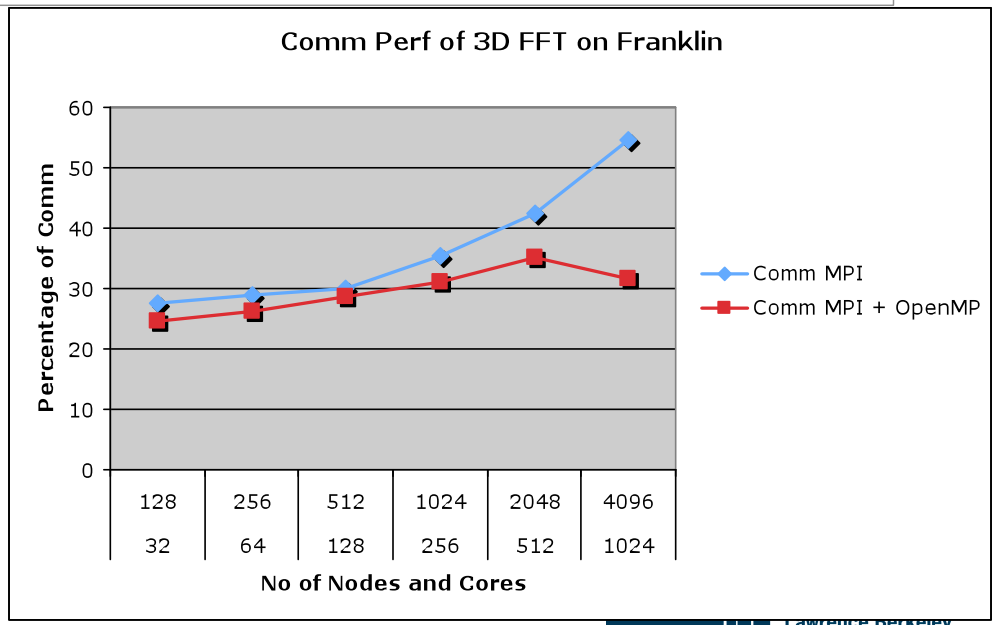


Example OpenMP for PARATEC

- **MPI+OMP Hybrid**
  - Reduces memory footprint
  - Increases performance up to NUMA-node limit



- **Hybrid Model improves 3D FFT communication performance**
  - Enables node to send larger messages
  - Substantial improvements in communications efficiency



Watch Carefully in this Training for Solutions to the Following Challenges

(prepare your codes for the future!)

- **Parallelism:** ***Focus on scaling on # of nodes*** *(tractable)* ***rather than # cores in system*** *(which is intractable)*
- **Memory Capacity:** ***Share Data that is common to all processors in node to reduce memory footprint***
- **Memory Bandwidth:** ***Respect memory locality***
  - *Vertical locality with blocking & prefetch*
  - *Horizontal locality using NUMA nodes*
- **Interconnect Performance:** ***Threading on-node can enable improved surface-to-volume ratio for domain decomposition in some cases***
  - *Larger/aggregated messages can mean better MPI performance*