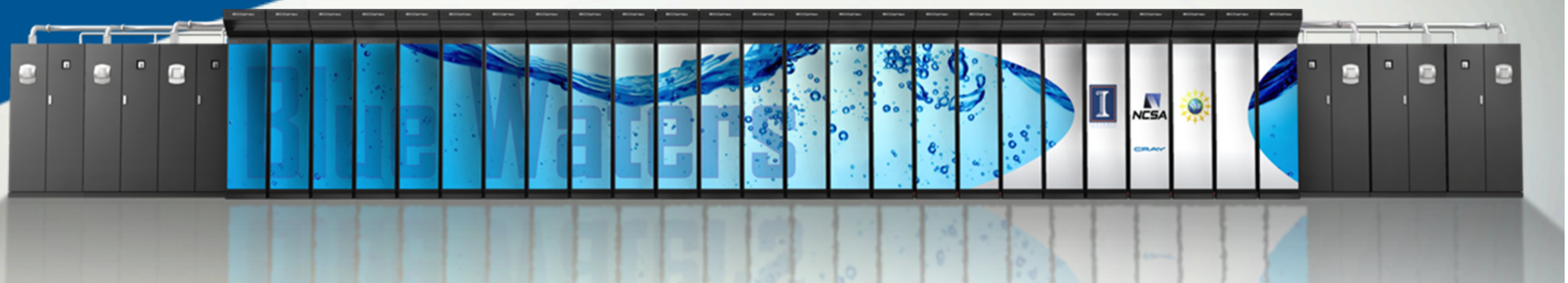


The Cray Gemini Network: Very Basic Architecture

Forest Godfrey
Senior Principal Engineer, Cray Inc.

November 12, 2013

Outline

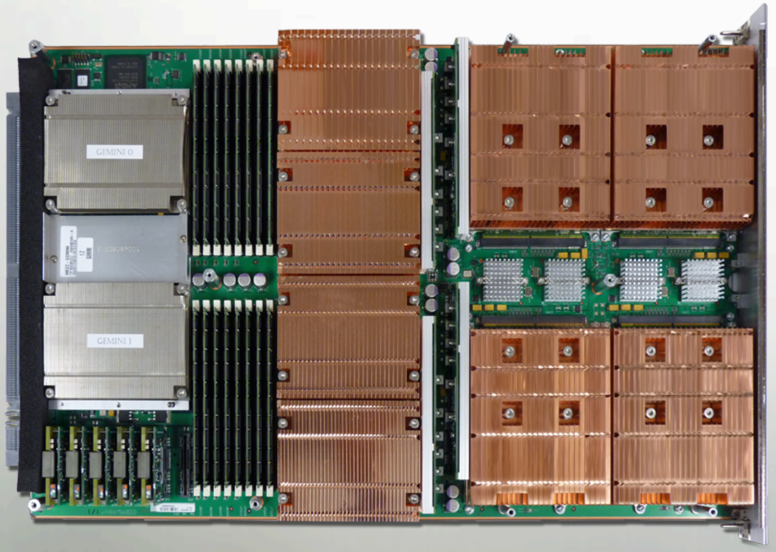
- About the Speaker
- Cray Gemini NIC Hardware Architecture Overview
- Cray XE/XK System Overview

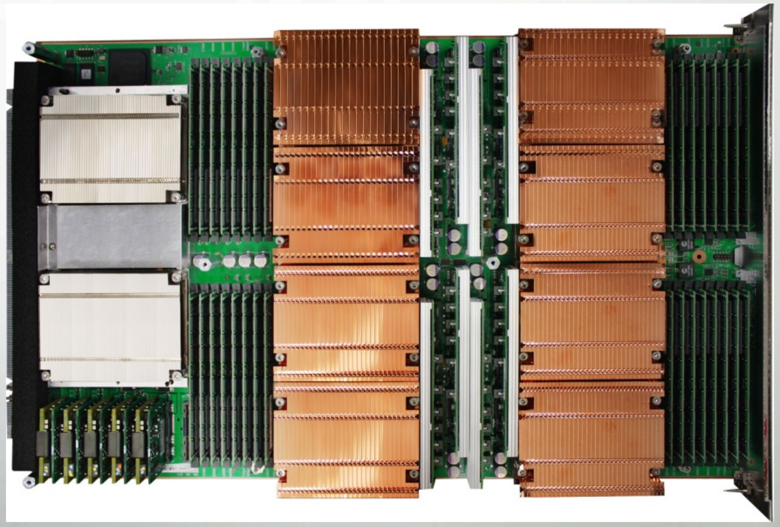
About Me

- Graduated from Carnegie Mellon in 1999 with Bachelor of Science in Computer Science
- Have been with Cray (or SGI when it owned Cray) ever since
- Started as a kernel programmer (Irix and Linux)
- Worked on SGI Origin and Altix systems as well as Cray X1, X1E, X2, XT series, XE series, and XK series (and two upcoming products). Served on architecture team for X2, XE and XK.
- Lead software architect for GPUs and future system control networks
- My brother, Brighten, is a professor in the CS Department at UIUC

Outline

- About the Speaker
- Cray Gemini NIC Hardware Architecture Overview
- Cray XE/XK System Overview





Gemini NIC Overview

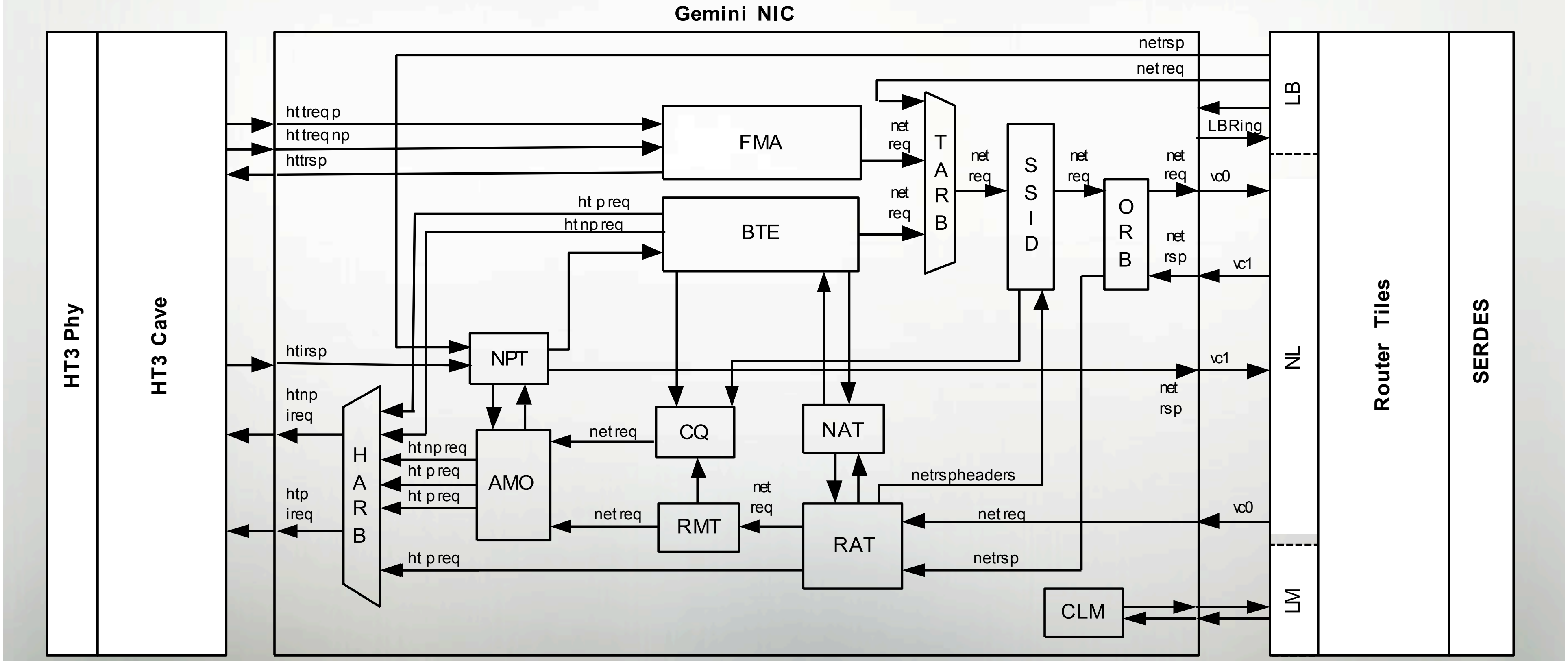
- Need to decouple network operations from the processor
  - Could just use a block transfer engine (BTE)
    - That only works at large enough transfer sizes
    - GAS languages need to deal with small transfers
- In addition to BTE, use a “Fast Memory Access (FMA) Window”
  - Goal: Allow processor directed network reads/ writes directly from user space without coupling processor instructions to network operations

Gemini Overview: Solutions: FMA Windows

- Split into two pieces – large (512MB) access window and small (4KB) control window
- Control window "aims" the access window at remote memory (sets target node, which memory registration, protection information, type of operation, etc.)
- Processor then writes directly into access window
- To do a remote read, command is set to read and the write into the local window causes remote memory to be written back to local memory
- In addition to reads/writes, various atomic memory operations are supported

Gemini Overview: NIC Architecture

- Need to add a variety of blocks to handle network operations/timeouts and completion of operations
- Final NIC



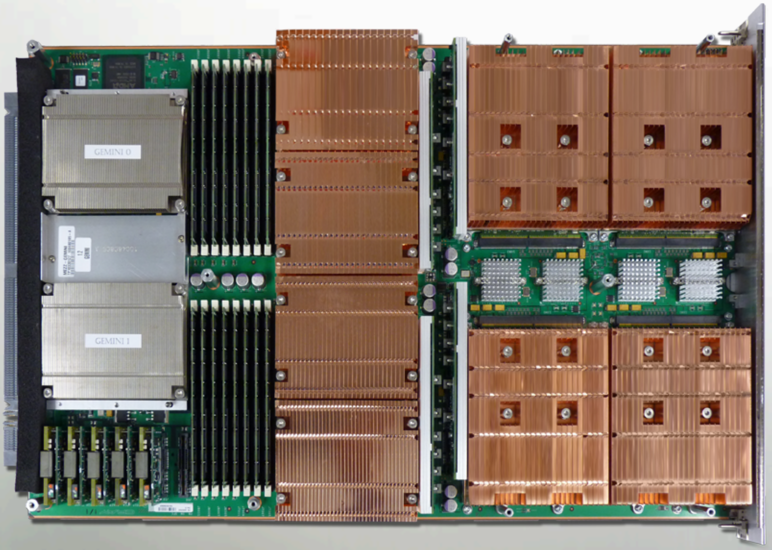
Gemini NIC

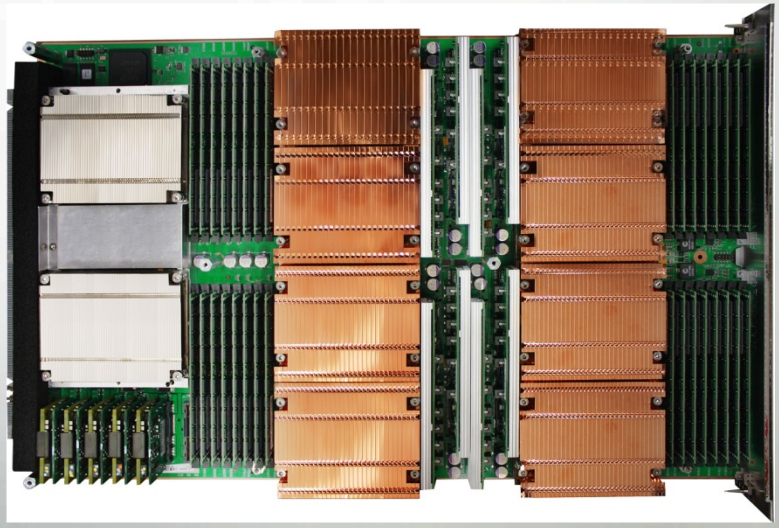
Difference from Datacenter

- Cray system looks a bit like a cluster
  - Lots of individual nodes each running Linux
- Unlike datacenters, tend to run small numbers of jobs at a time
  - 10's of thousands of nodes per job
- Need to be able to start and scale these large jobs
- In many ways, "cluster" is really a single system
  - Shared Filesystem
  - Share Job Launch
  - Shared Management

Outline

- About the Speaker
- Cray Gemini NIC Hardware Architecture Overview
- Cray XE/XK System Overview



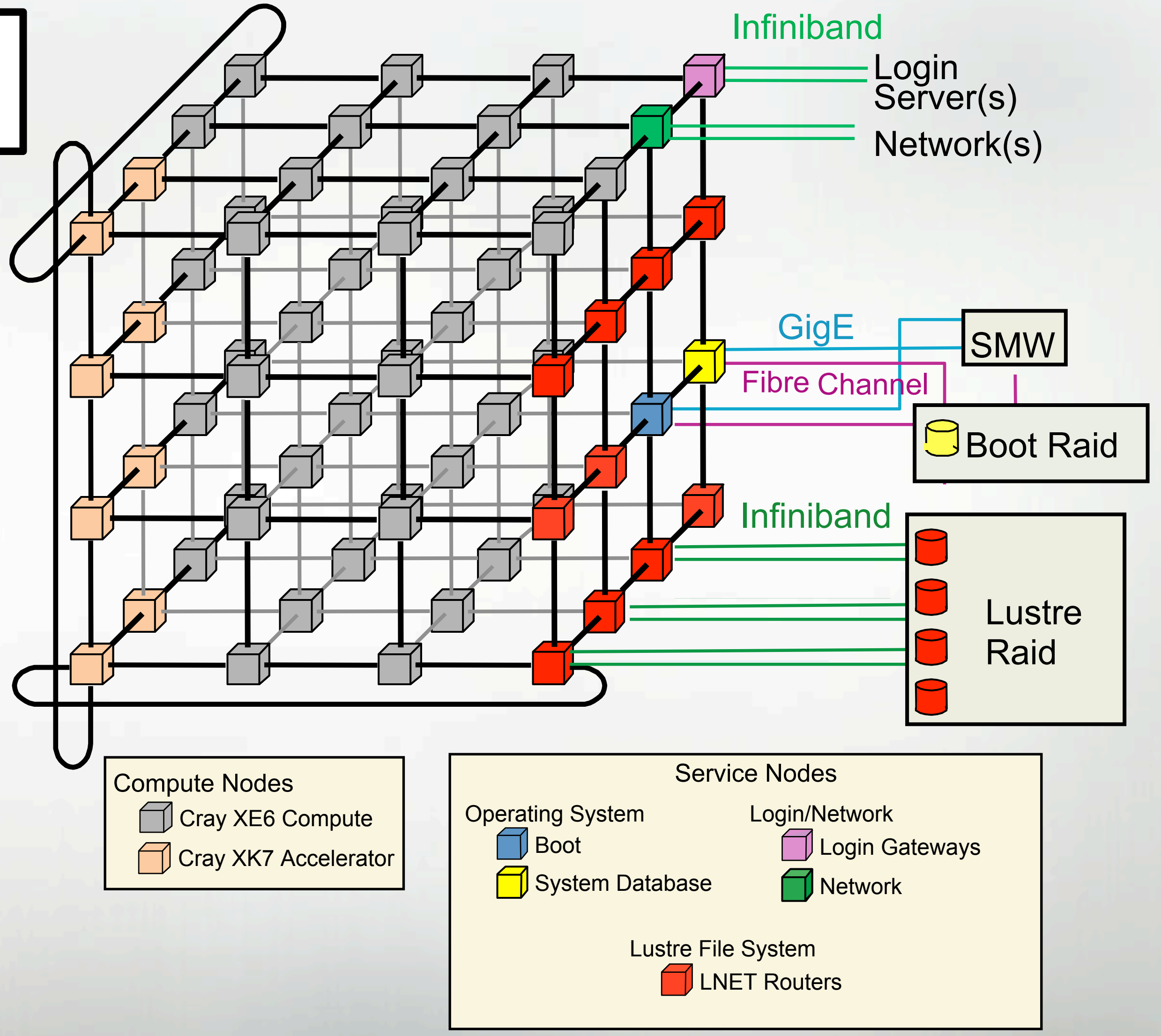


Gemini Overview: Router Architecture

- Integral 48 port router
- 8 ports are internal only
- 40 external ports are arranged to form 3 dimensional torus
  - 6 links: X+, X-, Y+, Y-, Z+, Z-

Gemini Overview: Router Architecture

Blue Waters 3D Torus
23 x 24 x 24 Gemini



Compute Nodes
- ⬜ Cray XE6 Compute
- 🟧 Cray XK7 Accelerator

Service Nodes

Operating System
- 🟦 Boot
- 🟨 System Database

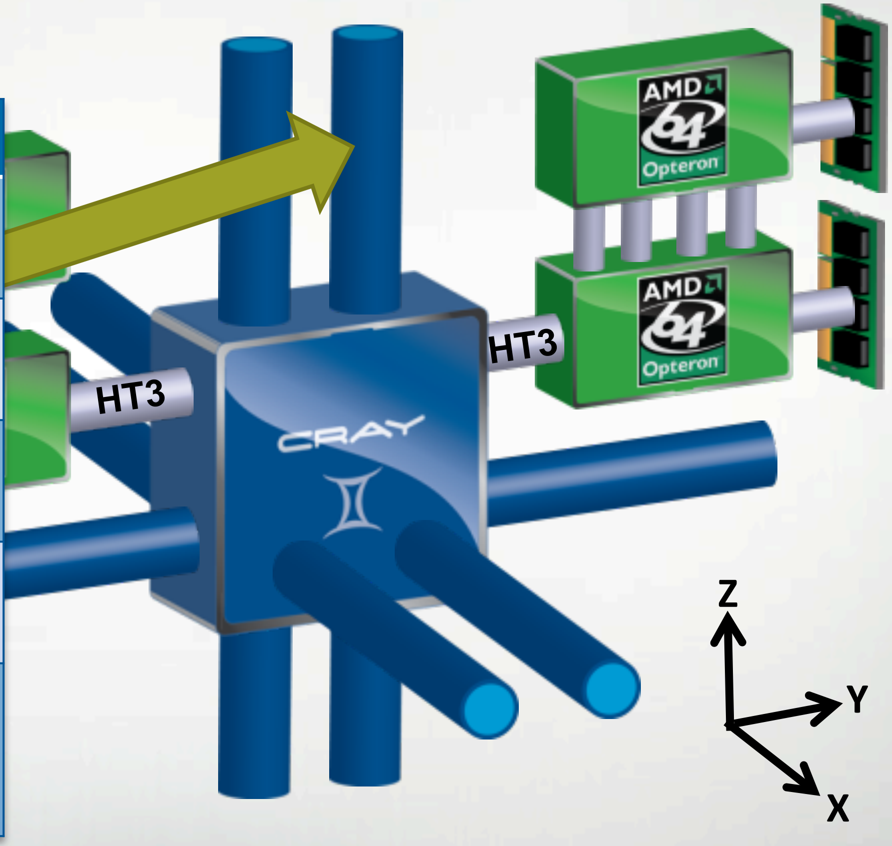
Login/Network
- 🟪 Login Gateways
- 🟩 Network

Lustre File System
- 🟥 LNET Routers

Cray XE6 Node Block Diagram

Blue Waters is currently scheduled to contain 22,752 XE6 compute nodes

Node Characteristics	
Number of Cores*	16
Peak* Performance	REDACTED
Memory Size	64 GB per node
Memory Bandwidth (Peak)	REDACTED
Interconnect Injection Bandwidth (Peak)	9.6 GB/sec per direction

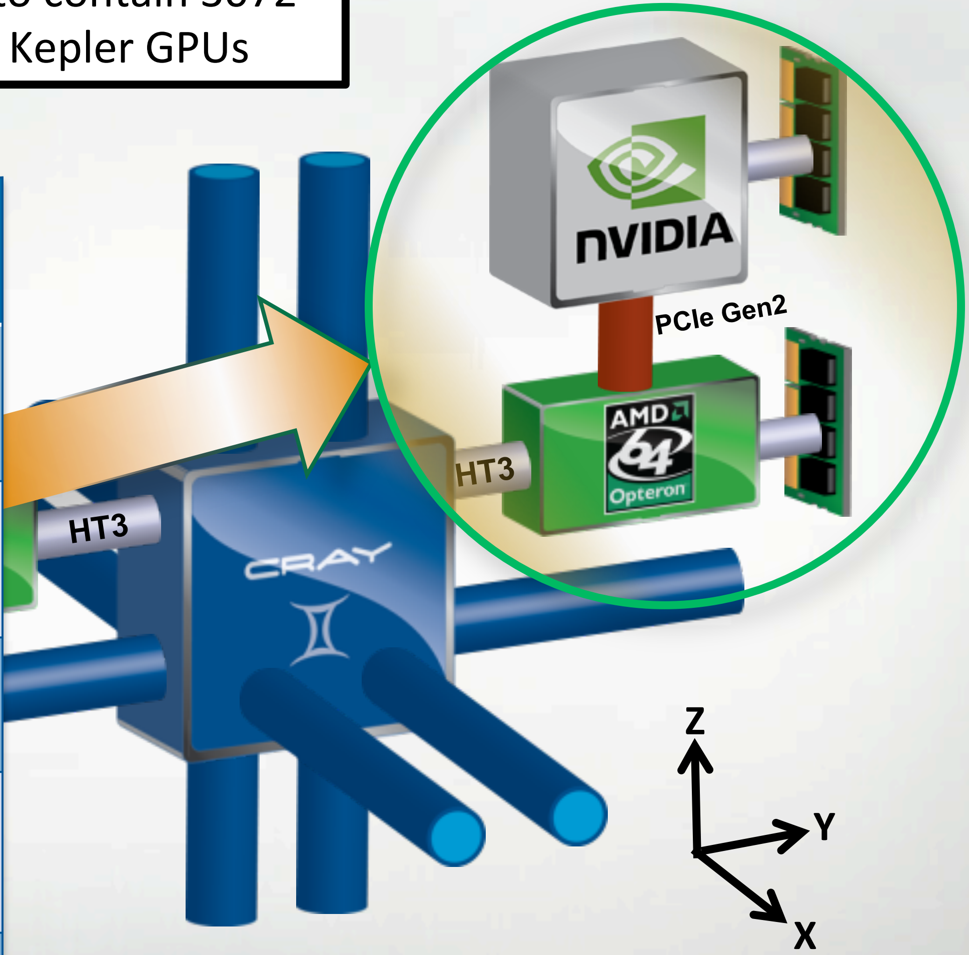


**Exact calculation of these numbers is beyond the scope of this talk*

Cray XK6 Block Diagram

Blue Waters is currently scheduled to contain 3072 XK compute nodes with NVIDIA™ Kepler GPUs

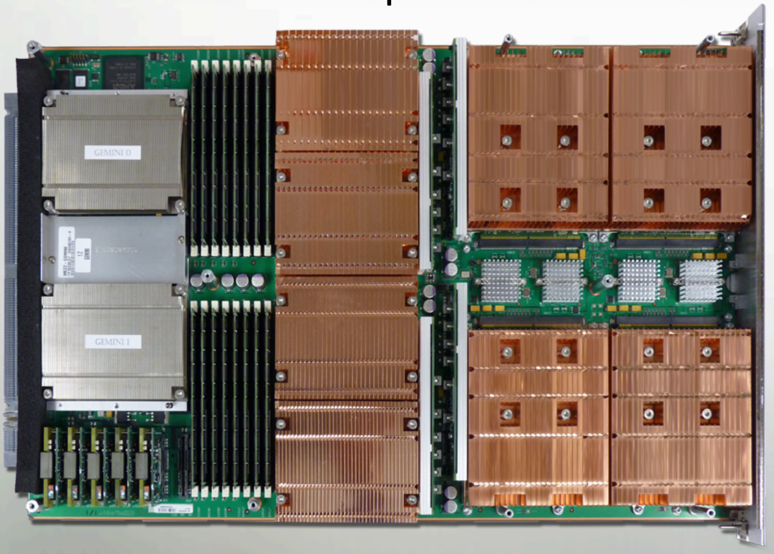
XK7 Compute Node Characteristics	
Host Processor	AMD Series 6200 (Interlagos)
Host Processor Performance	REDACTED
Kepler Peak (DP floating point)	REDACTED
Host Memory (peak)	REDACTED
Kepler Memory	6GB GDDR5 capacity (180 GB/sec)



Questions?

- Forest Godfrey can be reached at fgodfrey@cray.com.

GPU Compute Blade



Opteron Compute Blade