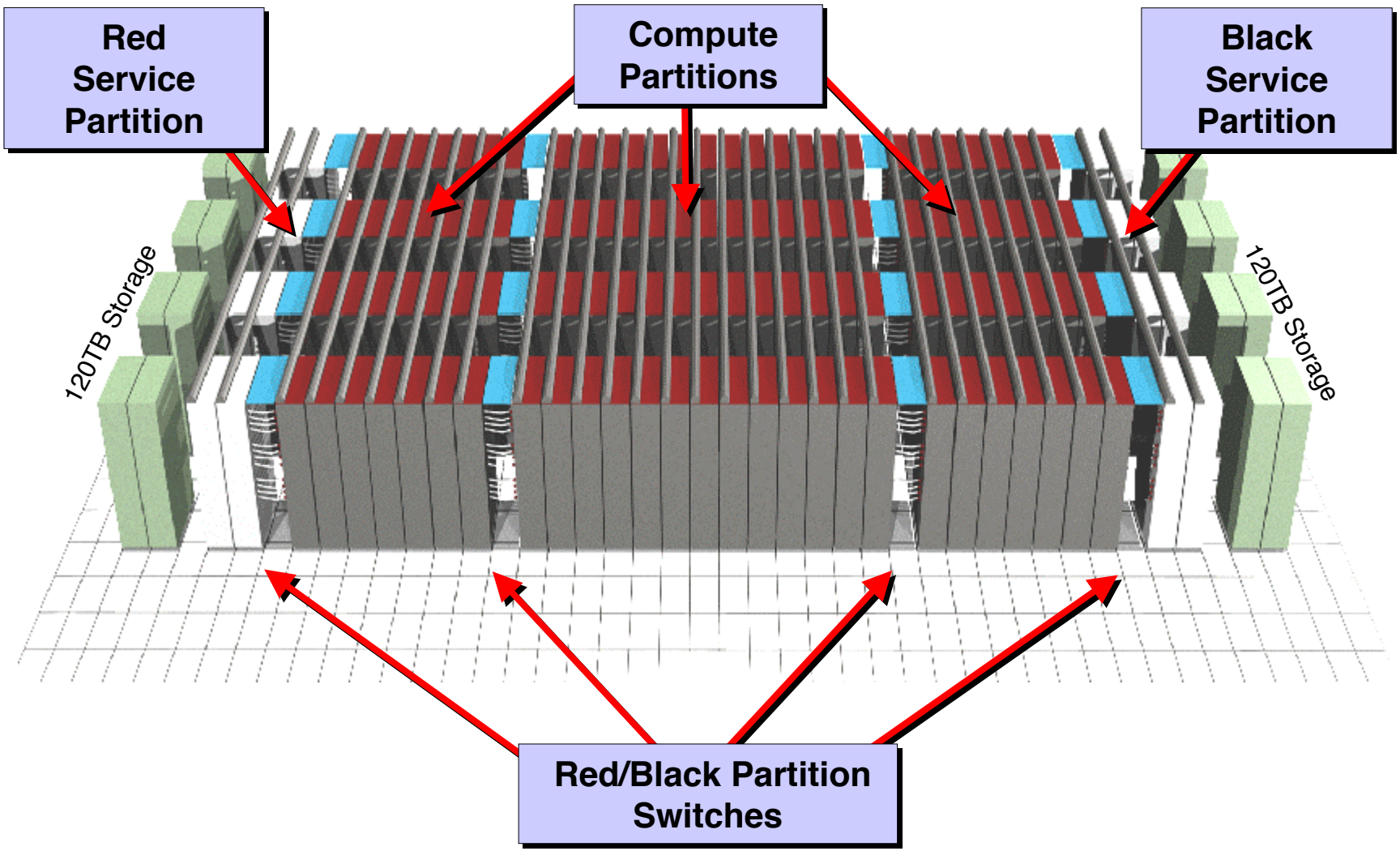
Red Storm

Robert Alverson

Red Storm Hardware Architect

Full System Configuration



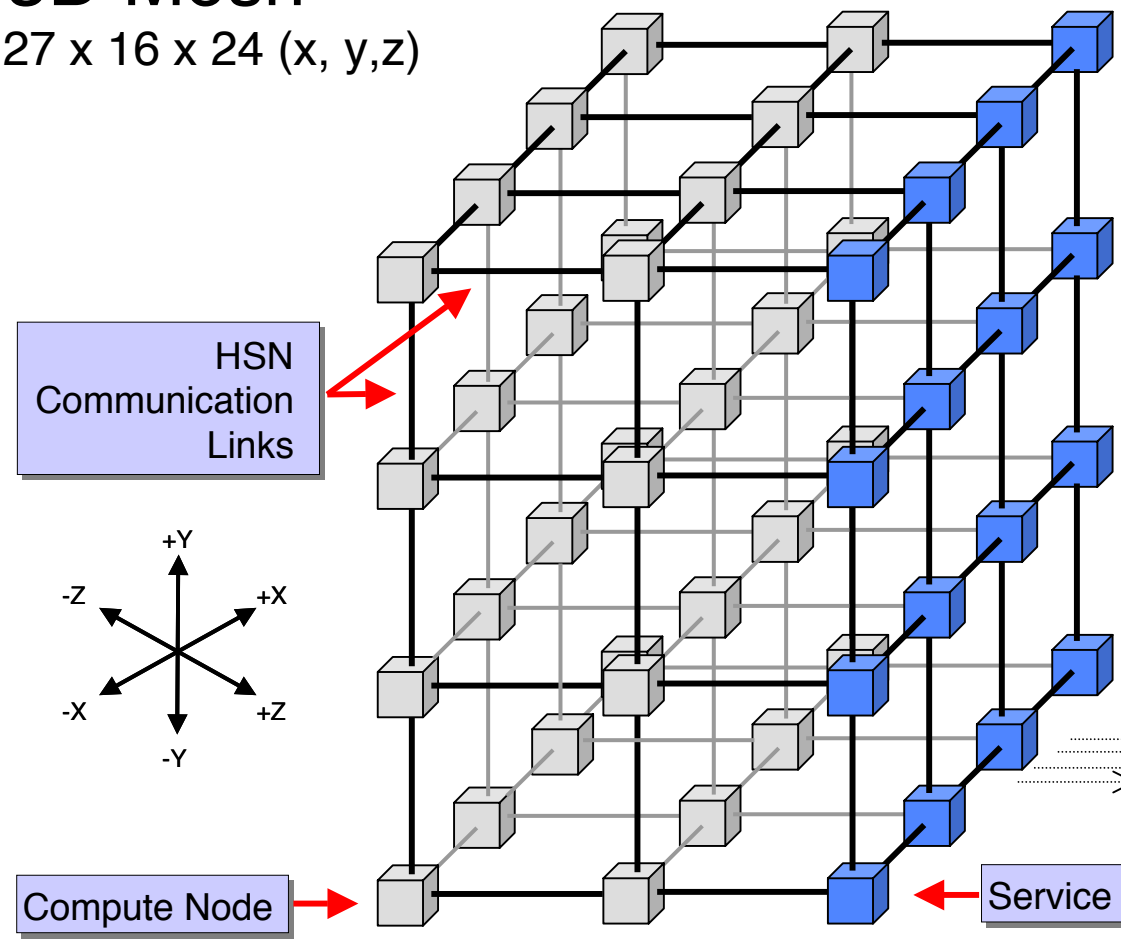
Red Storm System Overview

- 40TF peak performance
- 108 compute node cabinets, 16 service and I/O node cabinets, and 16 Red/Black switch cabinets
- 10,368 AMD Opteron™ compute processors
- 2 x 256 service and I/O processors(256P for red, 256P for black)
- 10 TB DDR memory
- 240 TB of disk storage(120TB for red, 120TB for black)
- Approximately 3000 ft² including disk systems
- <2.0 megawatts of power and cooling

Red Storm Network

3D Mesh

27 x 16 x 24 (x, y,z)



HSN Communication Links

Compute Node

Service & IO Node

MPI Latency

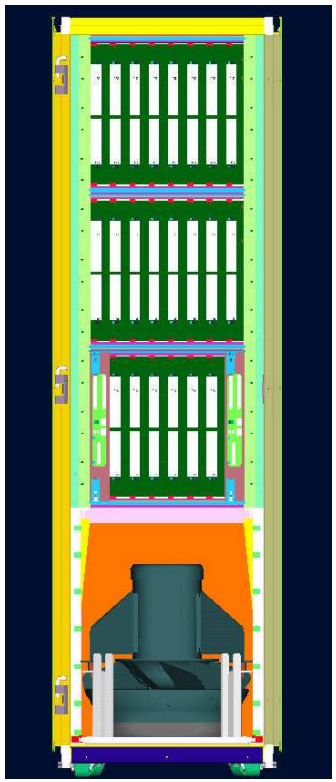
- 2 us Nearest neighbor
- 5 us Furthest point

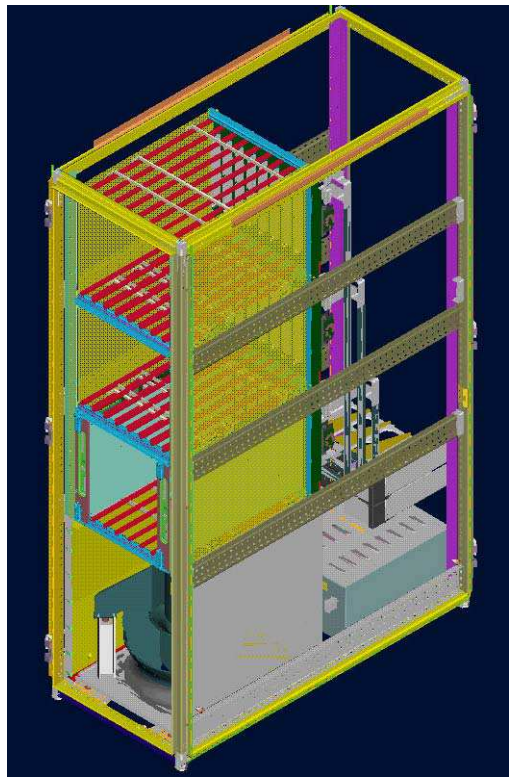
RAID Controllers

PCI - X

10 GigE Network

Red Storm Cabinet

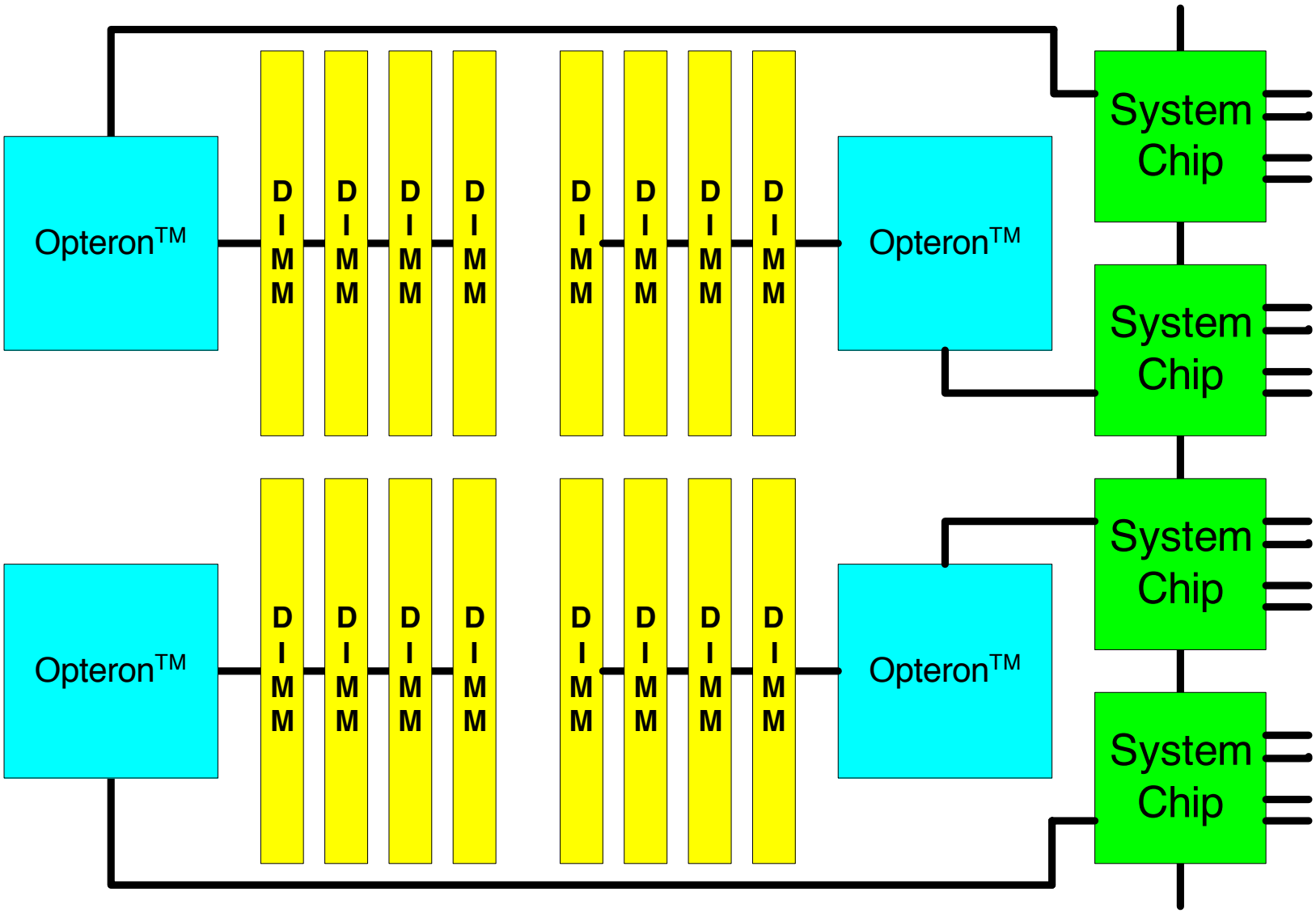




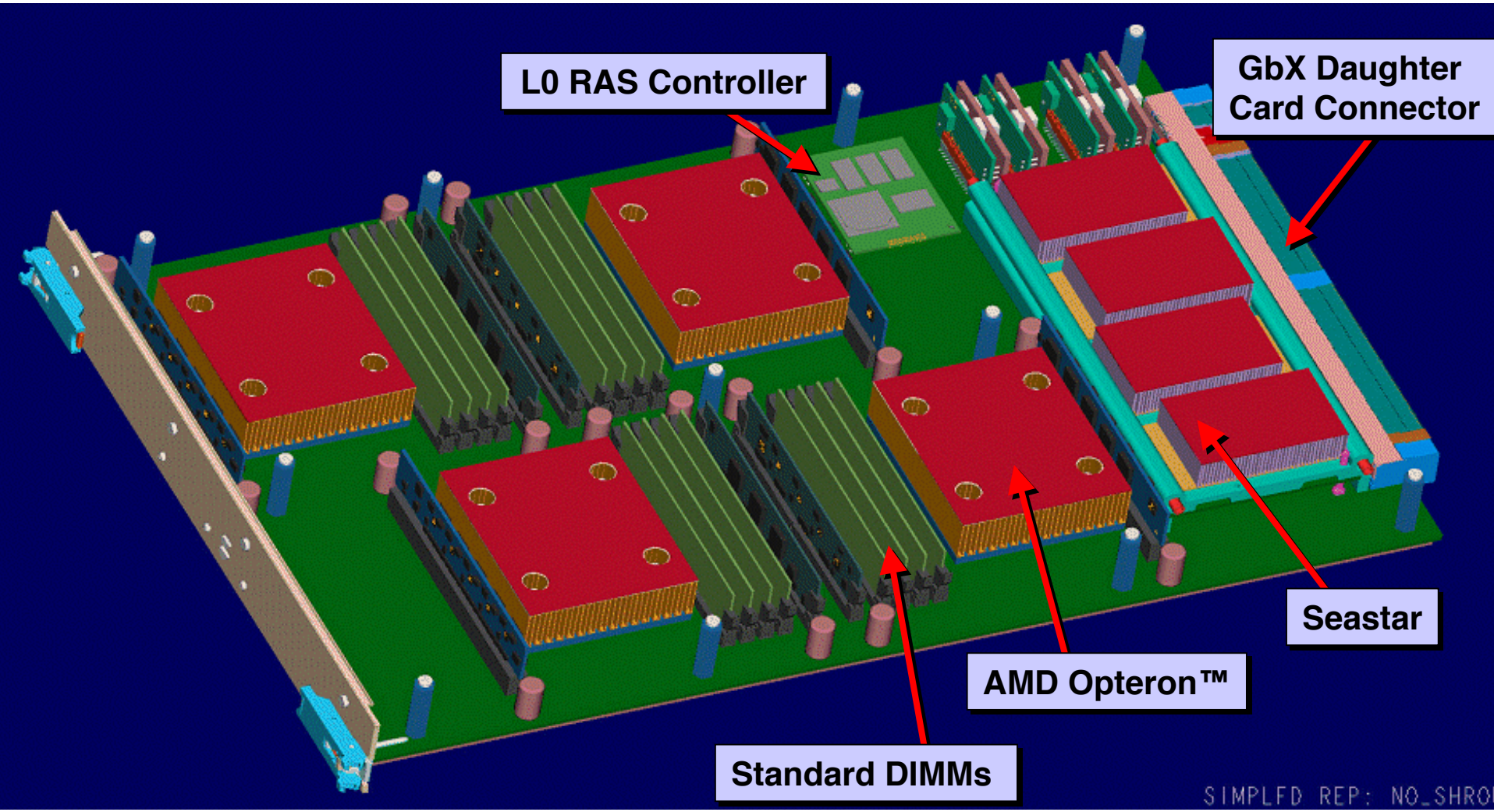
Compute Node

- Single (non-SMP) processor
- Integrated DDR memory controller
- Custom System Interface Chip
- No other support chips needed!
- Four nodes per board

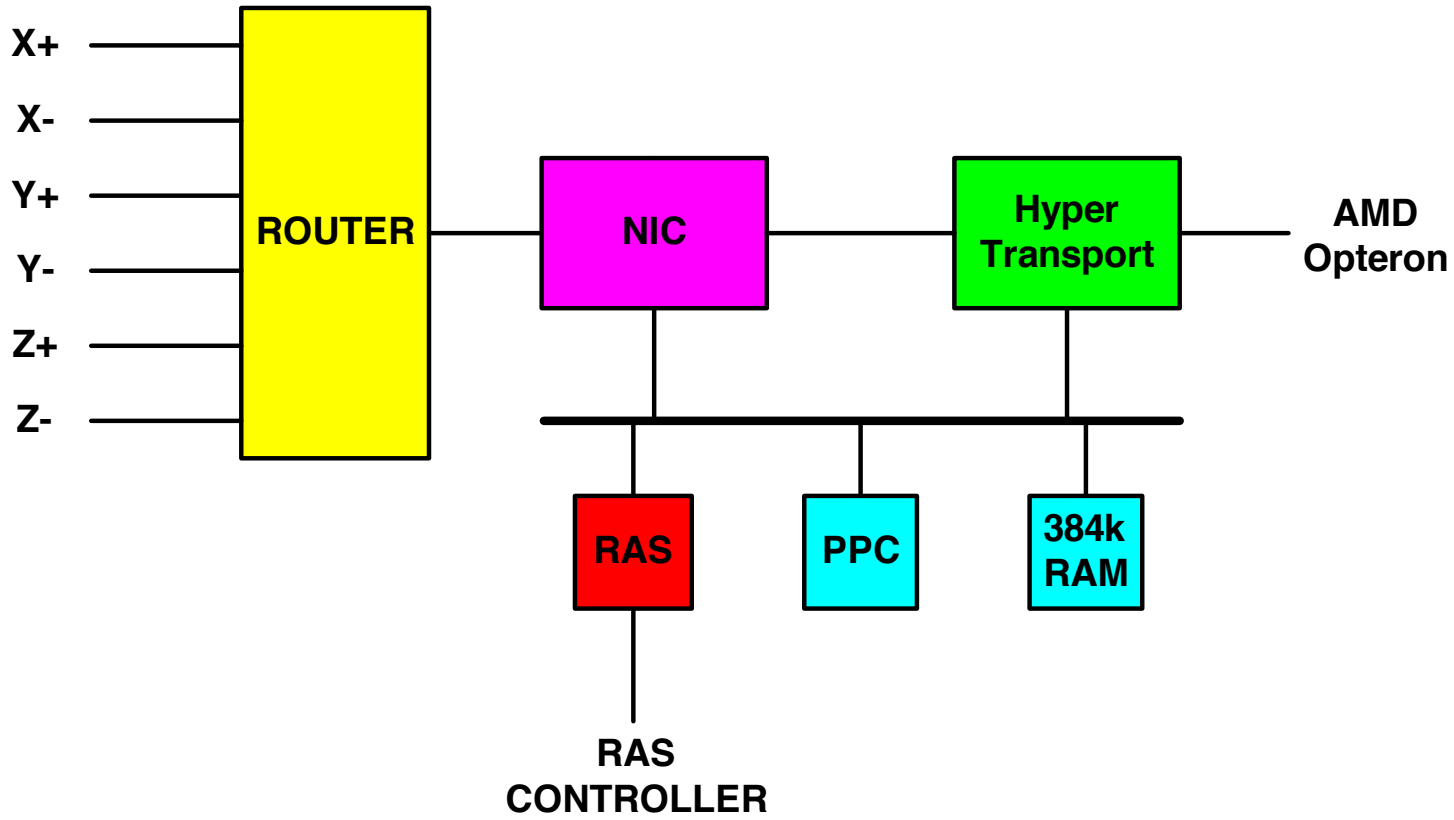
Compute Board



4 Node Compute Board



System Chip


X+
X-
Y+
Y-
Z+
Z-
ROUTER
NIC
Hyper Transport
AMD Opteron
RAS
PPC
384k RAM
RAS CONTROLLER

System Interface Chip

- 800 MHz DDR HyperTransport™ to Opteron™ processor
- Supplies boot code to processor
- 384 kB on-board scratch RAM
- Message passing network interface with 1.5 Gbyte/sec user bandwidth, each direction
- 7 port router

System Chip Features

- IBM 0.13u ASIC process
- 500 MHz embedded PowerPC™
- 16 bit 1.6 Gbit/sec HyperTransport™
- Six 12 channel 3.2 Gbit/sec High Speed Serial links
- GDA Technologies Inc. HyperTransport Cave

HyperTransport™

- High bandwidth link to processor,at least 3 Gbyte/sec required
- PCI-X BW insufficient, latency sometimes high
- HyperTransport is open standard

Network Interface

- Message based
- DMA between Opteron™ memory and network for high bandwidth (cache coherent)
- Minimal host overhead
- Supports reception of multiple simultaneous messages

Integrated Router

- 6 high speed network links per ASIC
- More than 4 Gbyte/sec per link
- Reliable link protocol with CRC-16 and automatic retry
- Support for up to 32k nodes in 3D toroidal mesh

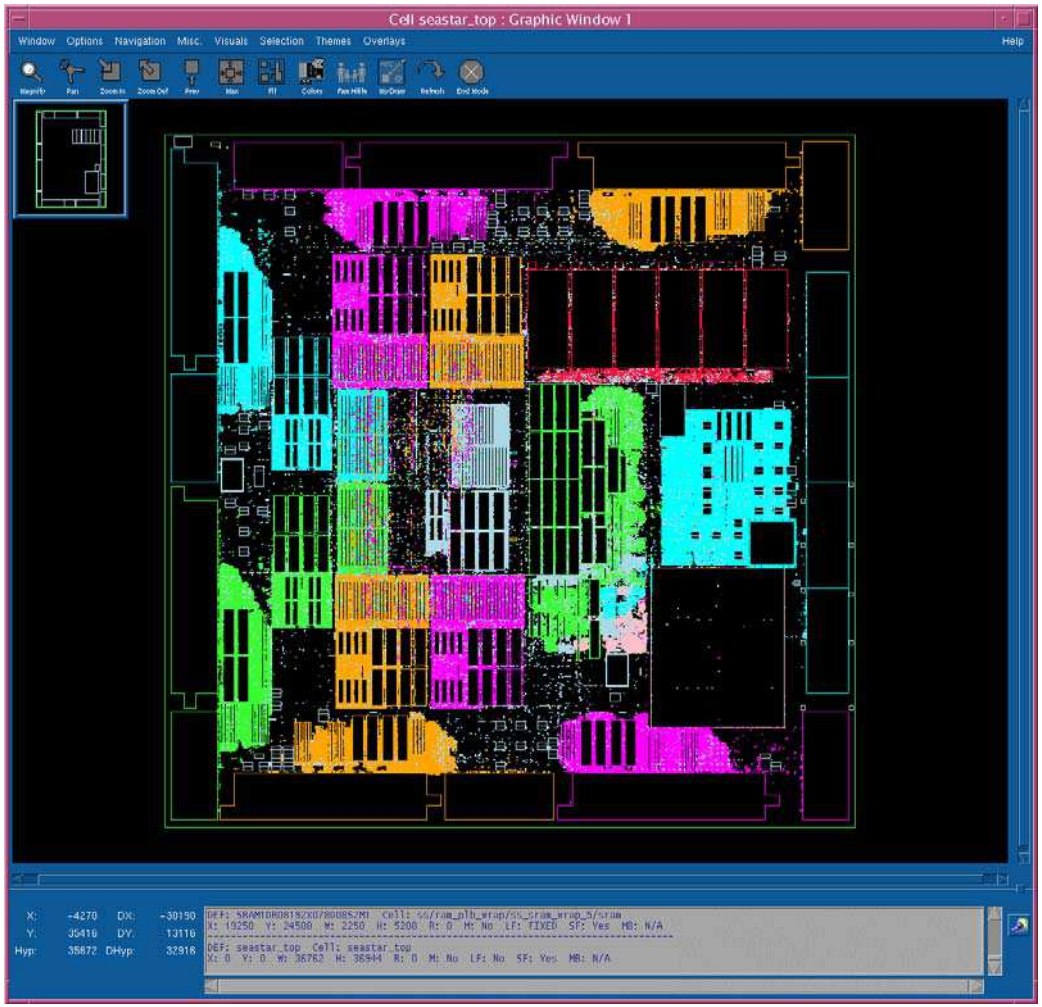
Embedded PowerPC

- Message preparation
- Message demultiplexing (MPI matching)
- System monitoring

Reliability

- Reliable link protocol
- SECDED on scratch RAM with scrubbing
- SEC on routing lookup tables
- Parity protection on DMA tables
- Monitor port accesses PowerPC™ and Opteron™ state

Die Layout


Cell seastar_top : Graphic Window 1
Window Options Navigation Misc. Visuals Selection Themes Overlays Help

Wrap up

- High bandwidth, low latency MPI
- Scalable to 32k nodes
- Reliable at large scale
- Questions?