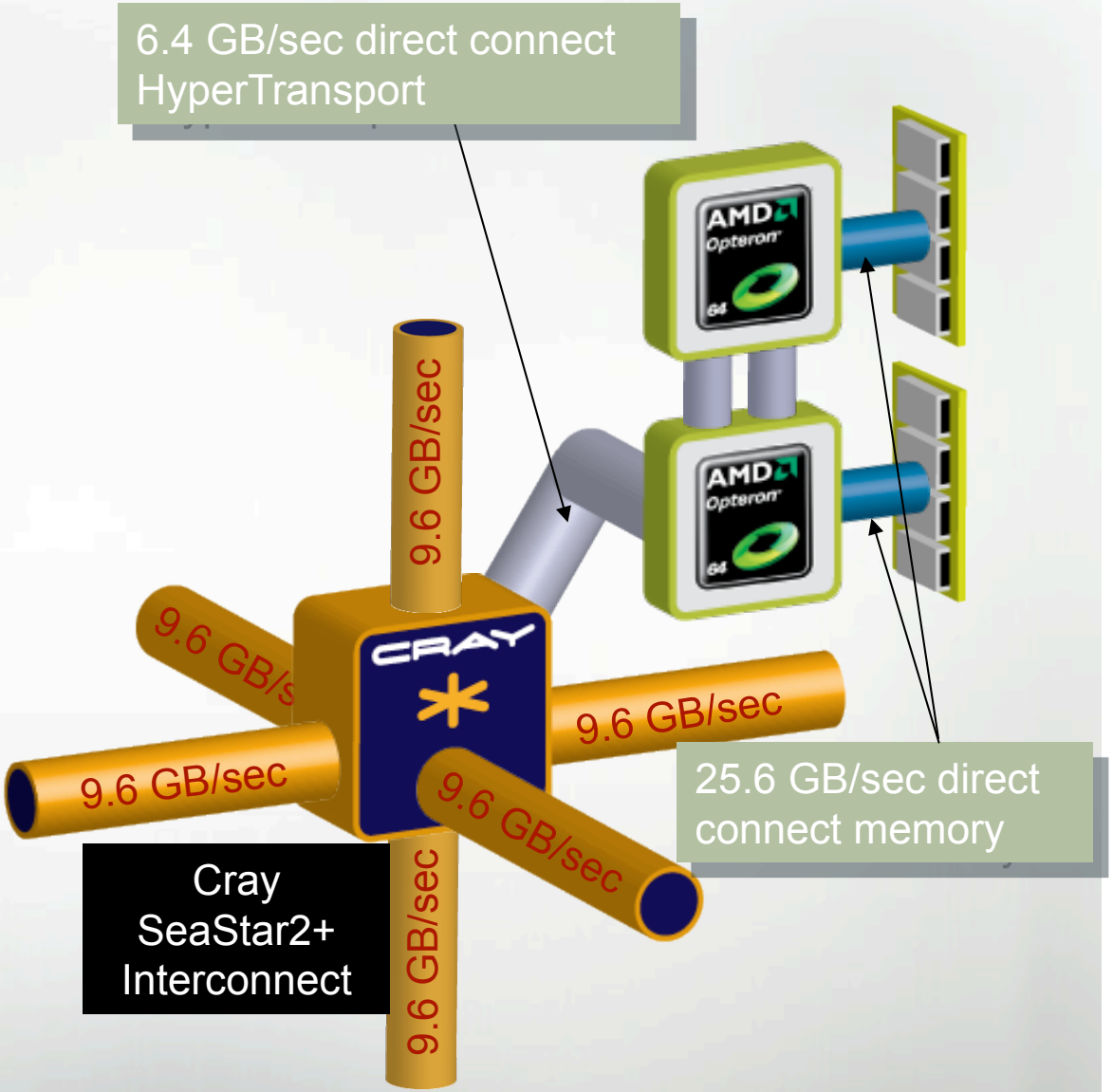
Interconnect

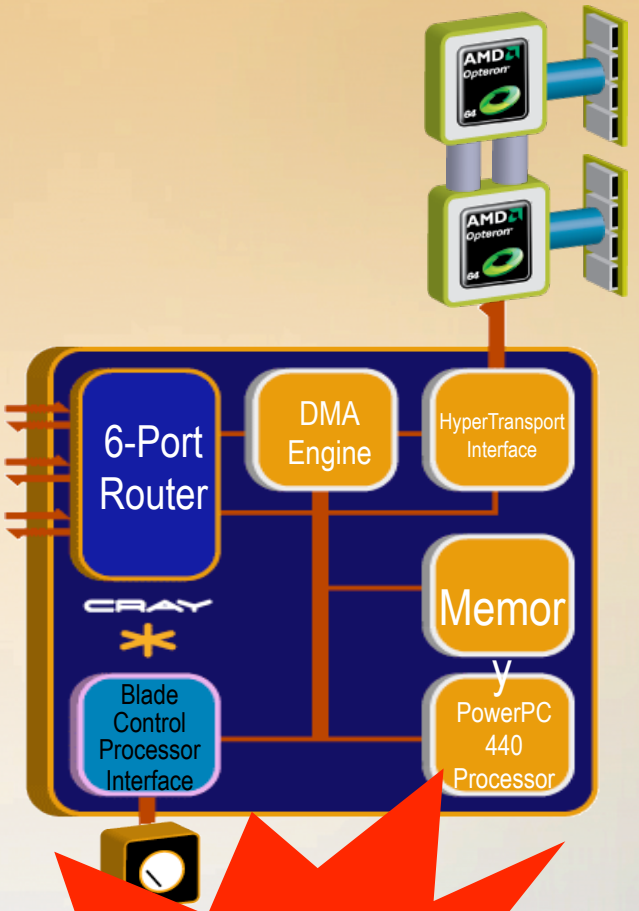
Cray XT5 Node

Cray XT5 Node Characteristics	
Number of Cores	8 or 12
Peak Performance Shanghai	76-86 Gflops/sec
Peak Performance Istanbul	125 Gflops/sec
Memory Size	8-32 GB per node
Memory Bandwidth	25.6 GB/sec



Cray SeaStar2+ Interconnect

- New firmware was released with "Amazon" in 2008 that will improved SeaStar performance
- Improvements:
  - Improved packet arbitration and aging algorithm lowers global latency
  - Using 4 virtual channels improves sustained global bandwidth

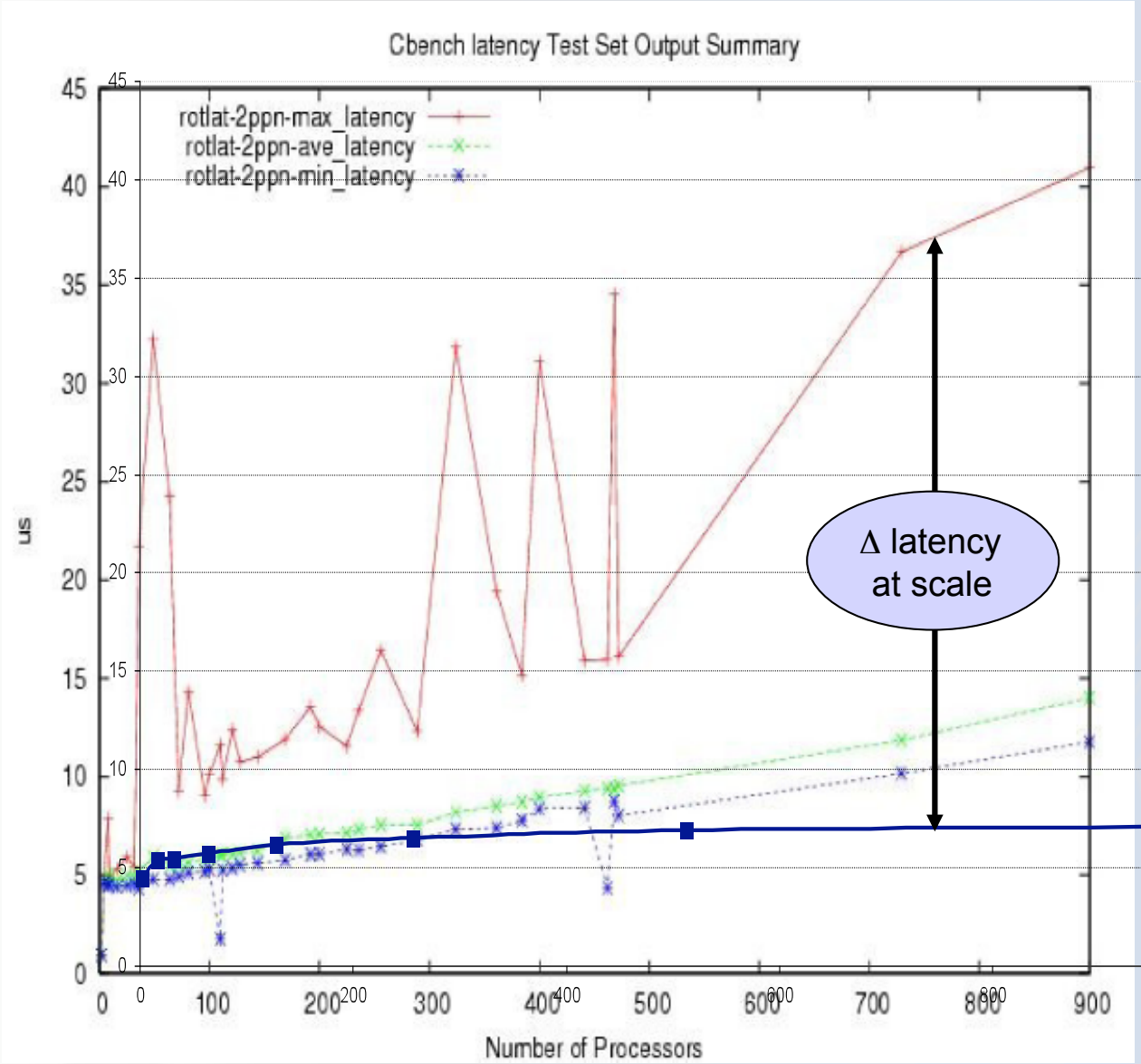


Packet arbitration and aging Improvement	
PTRANS	4.60%
MPIFFT	12.4%
AllReduce	12.4%
AllToAll	36.3%

Multiple virtual channels Improvement	
PTRANS	10–25%
MPIFFT	25%
RandomRing bandwidth	>40%

Now Scaled to 150,000 cores

IB Cbench Latency



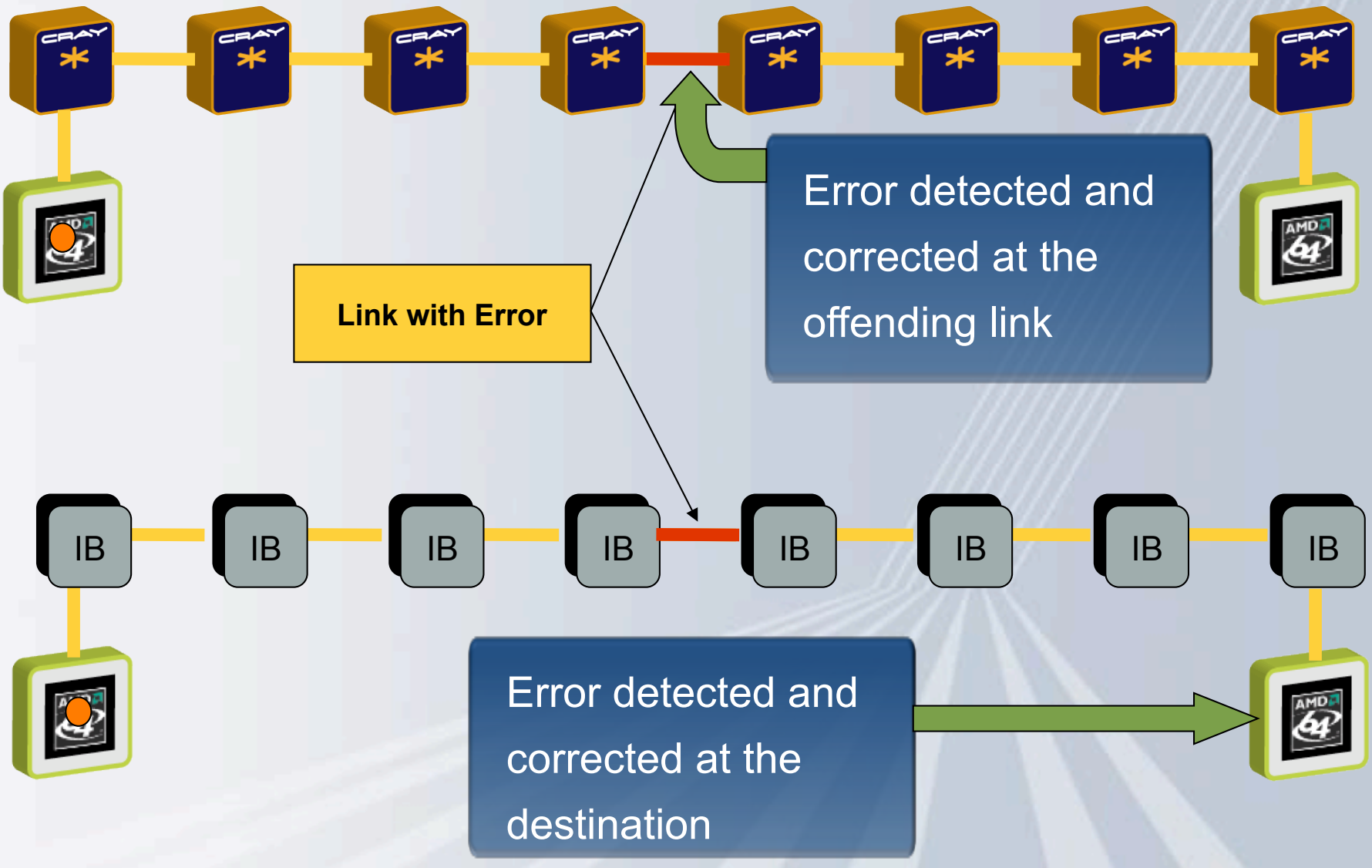
- IB shows a large spread best and worst case
- In MPP computing, we always wait for the slowest processor, so the *worst case figures are most important*
- Solutions include over-provisioning the interconnect and adaptive routing

Δ latency at scale

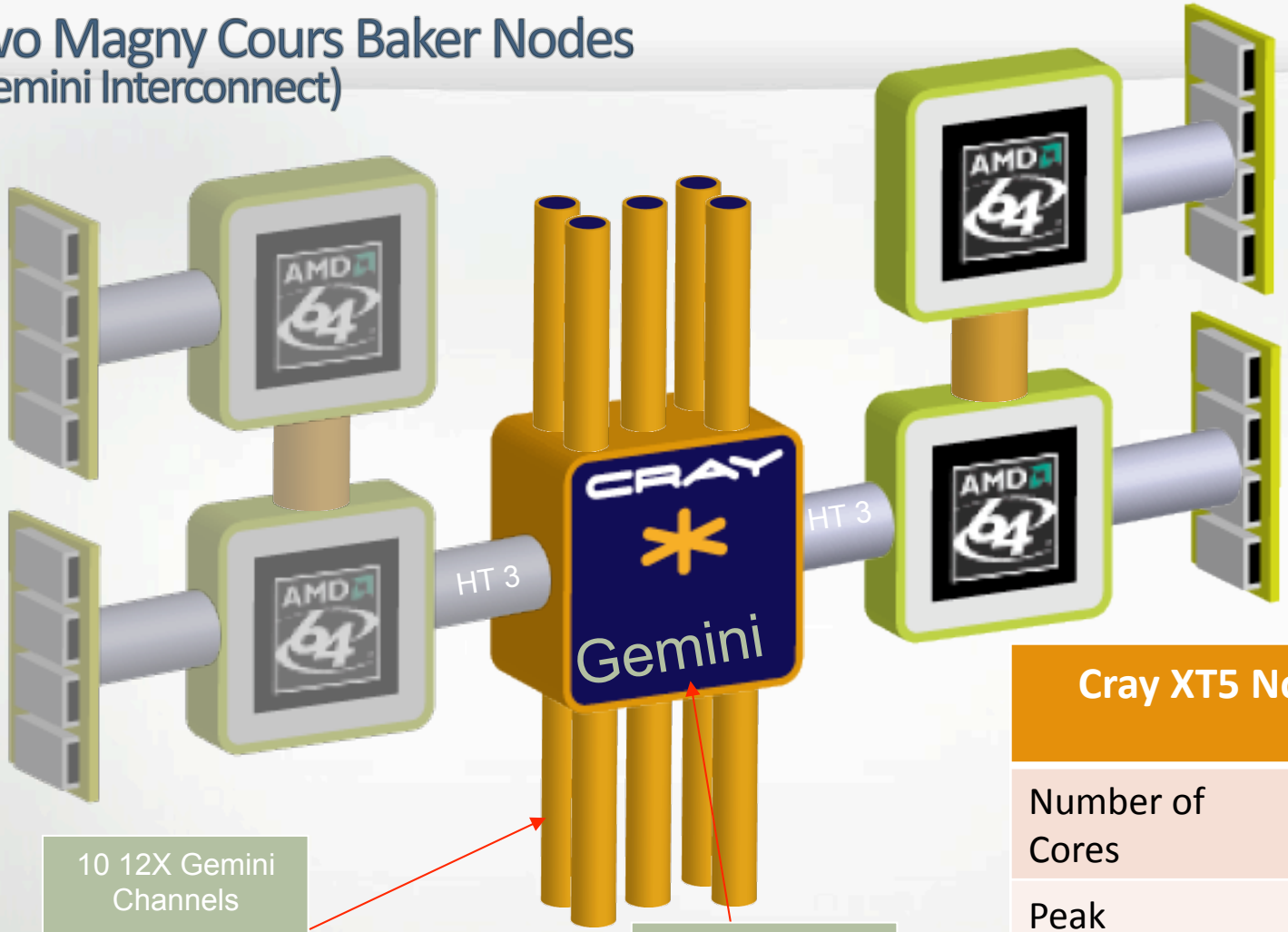
Cray SeaStar2 Maximum Latency

Source: Presentation by Matt Leininger & Mark Seager, OpenFabrics Developers Workshop, Sonoma, CA, April 30th, 2007

The Importance of Link Level Reliability



Two Magny Cours Baker Nodes
(Gemini Interconnect)



10 12X Gemini Channels

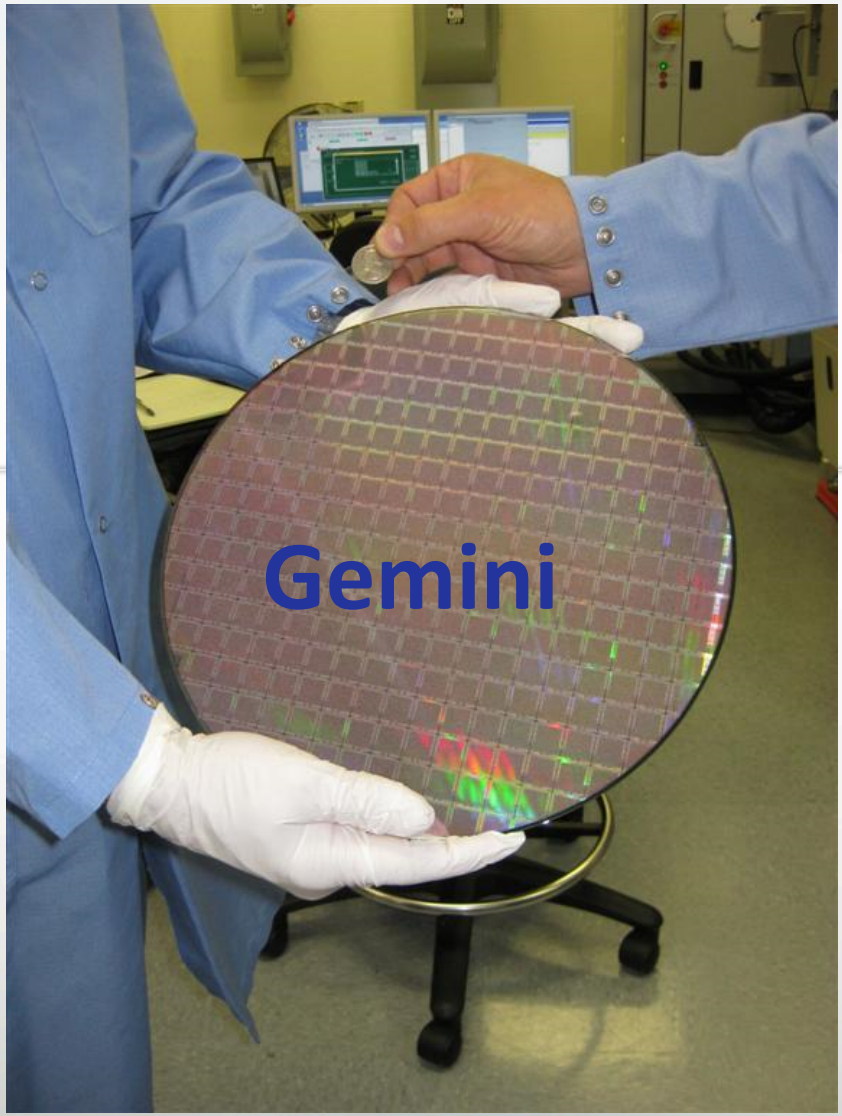
(Each Gemini acts like two nodes on the 3D Torus)

High Radix YARC Router with adaptive Routing

168 GB/sec capacity

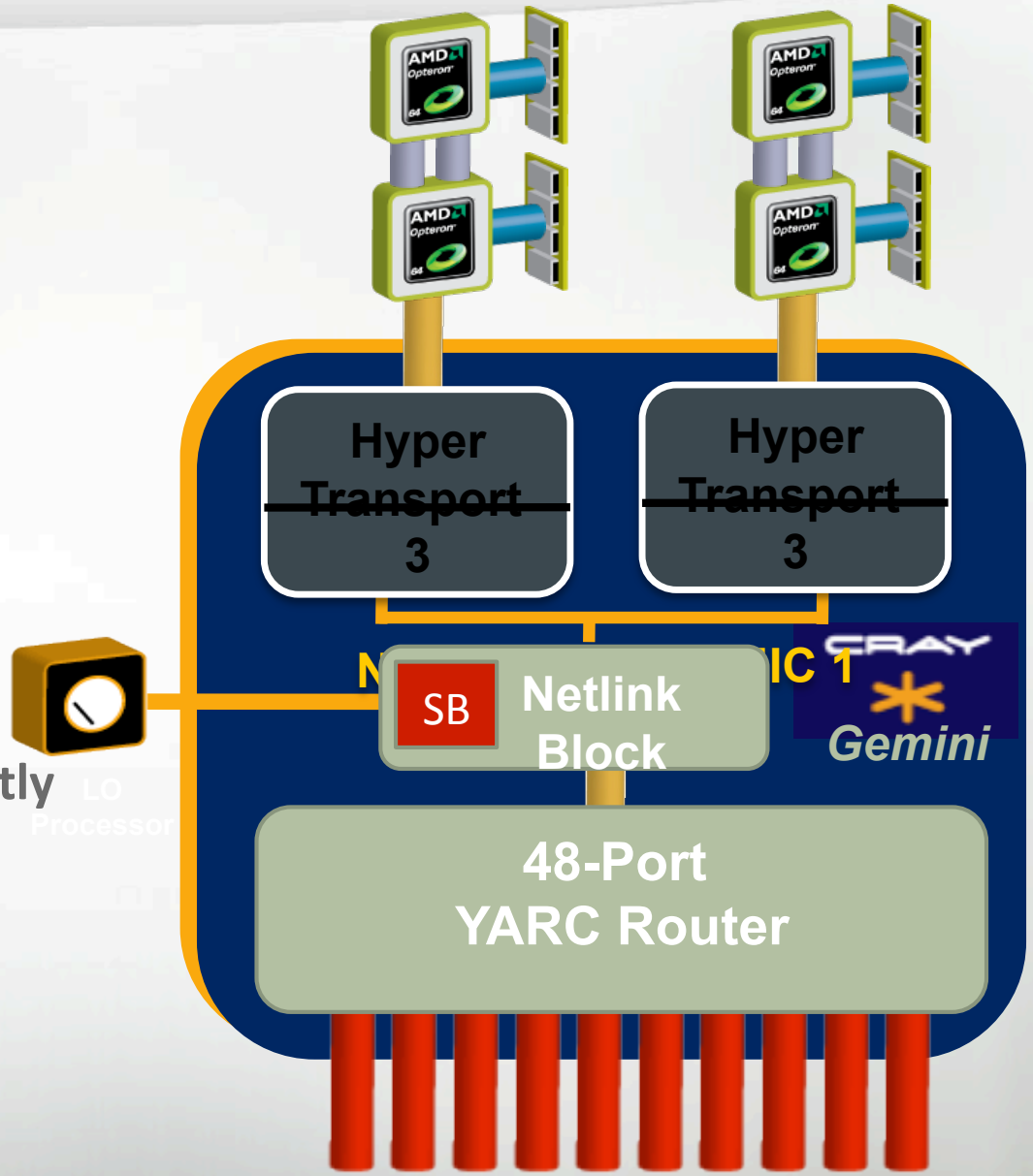
Cray XT5 Node Characteristics	
Number of Cores	24
Peak Performance	182 Gflops/s
Memory Size	32 or 64 GB per node
Memory Bandwidth	85 GB/sec

Gemini Interconnect

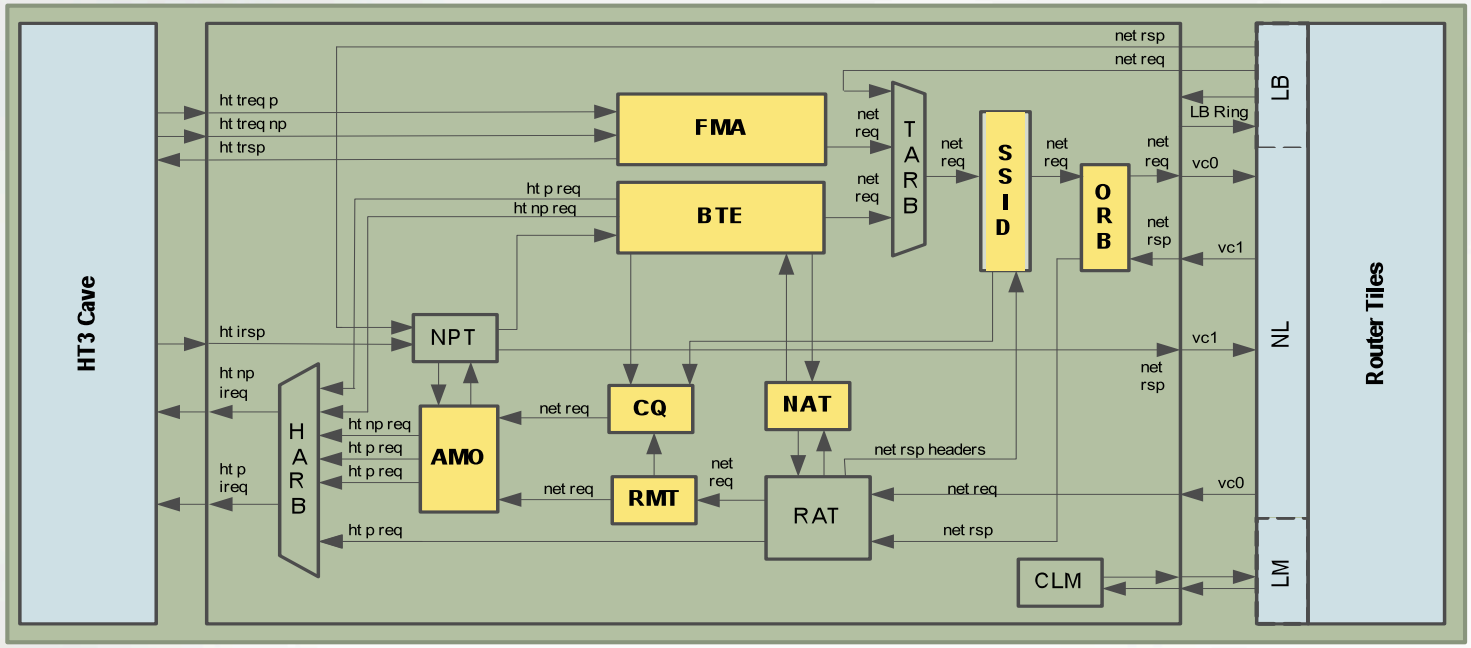


Cray Gemini ASIC

- 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
  - One-sided MPI
  - Shmem
  - UPC, Coarray FORTRAN, Titanium, Global Arrays



Gemini NIC block diagram



- **FMA (Fast Memory Access)**
  - Mechanism for most MPI transfers
  - Supports tens of millions of MPI requests per second
- **BTE (Block Transfer Engine)**
  - Supports *asynchronous* block transfers between local and remote memory, in either direction
  - For use for large MPI transfers that happen in the background

Gemini Reliability Features

- Will support warm-swap of blades
- Can map around bad links without rebooting
- Adaptive Routing – multiple paths to the same destination
- Packet level CRC carried from start to finish
- Network channels can automatically degrade
- Large blocks of memory protected by ECC
- Can better handle failures on the HT-link, discards packets instead of putting backpressure into the network
- Improved error reporting and handling
- Performance counters allowing tracking of app specific packets
- The “send/receive” channel protocol supports end-to-end reliable communication. (used by MPICH2 and OpenMPI)
- The RDMA protocol supports low overhead verification of success or failure. The low overhead error reporting allows the programming model to replay failed transactions

Gemini – Status

- Cray approved the netlist release 8/22/08
- First Wafers out of fab on 10/25/08
- Software infrastructure in place
- First Gemini mezzanine assemblies powered up 11/17/08
- First bugs in parts found and characterized, fibbed parts returned
- First MPI message traffic on 2/10/09
  - Un-optimized, zero-byte latency between two nodes was less than 2 microseconds