INTERCONNECTION TECHNOLOGIES

Non-Uniform Memory Access Seminar

Elina Zarisheva

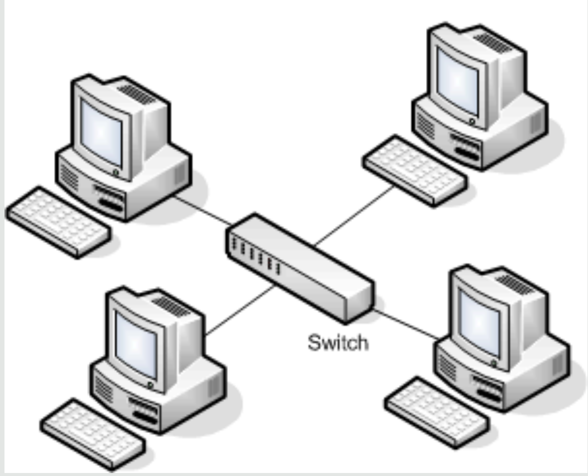
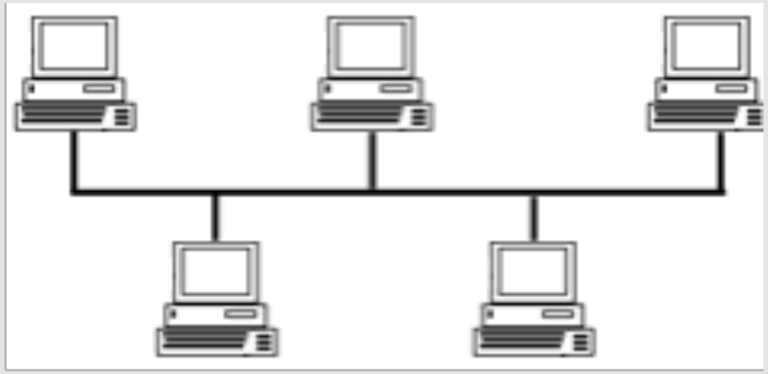
26.11.2014

Agenda

- Network topology
- Logical vs. physical topology
- Logical topologies
  - InfiniBand
  - Crossbar switch
- Interconnection technologies in NUMA system
  - AMD Hyper-Transport (HT)
  - Intel Quick-Path Interconnect (QPI)
  - NumaLink

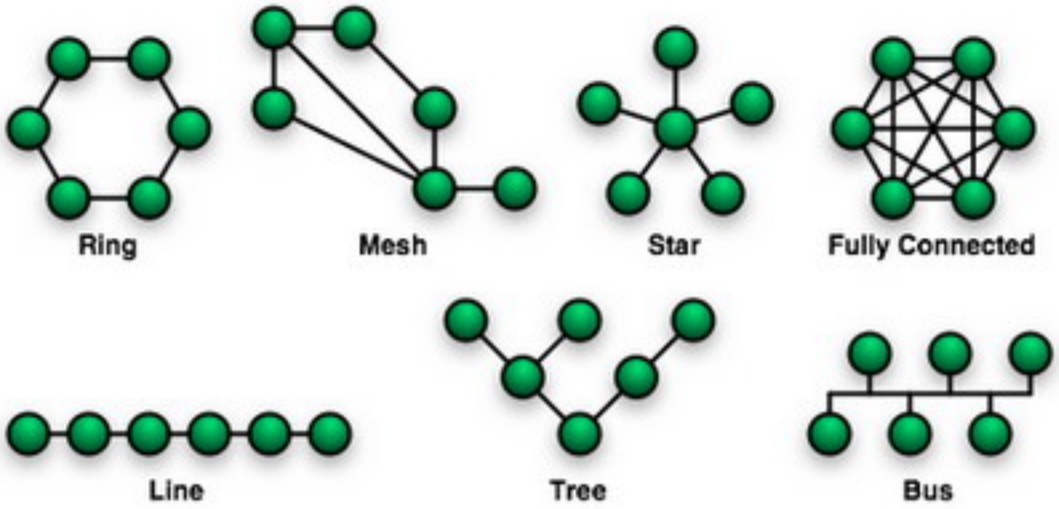
Network Topology

- Computer system or network equipment are connected to each other

Physical topologies	Logical topologies
	

Logical topologies

- Point-to-point
- Bus
- Daisy chain
- Ring
- Star
- Mesh
- Tree
- Hybrid
- Hypercube



Point-to-point topology

- Connects two nodes directly together
- Types:
  - simplex
  - half-duplex
  - full-duplex

+ simple

+ fast

+ medium is not shared

- support only two nodes

Bus topology

- Common medium (central bus) where the rest of nodes separately connected

  + more than one node

  + costs

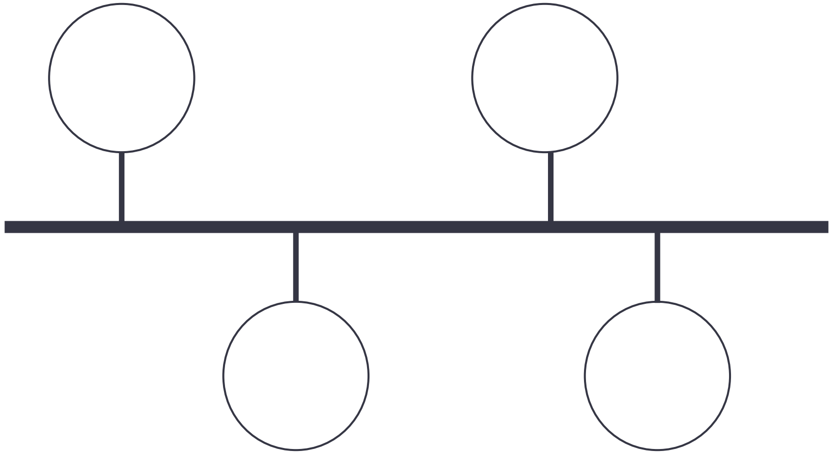
  - small networks

  - limitation of nodes

  - data collision

  - depended on central bus

  - security



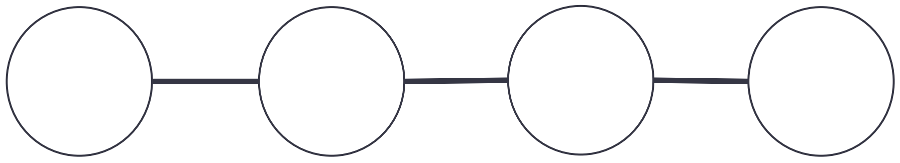
Daisy chain topology

- Node connects one after another

  + simple

  + scalability

  - slow for the opposite end of the chain



Ring topology

- Each node has two connection: to its nearest neighbors
- Data transmission happens indirectly
- Sending and receiving data with the help of the token
- Clockwise
- Double ring

  + organized

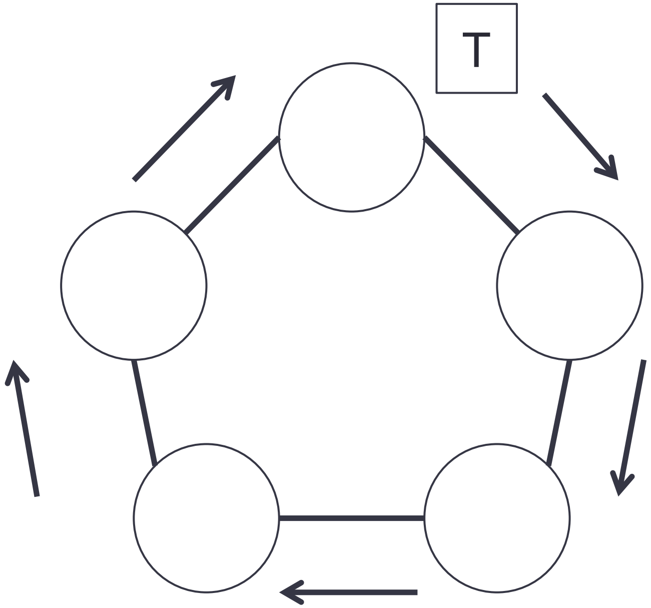
  + no date collision

  + no server

  + easy to add components

  - slow

  - dependent



Star topology

- Each node connects to a central point via a point-to-point link
- Central device:
  - hub
  - switch
- Independent access

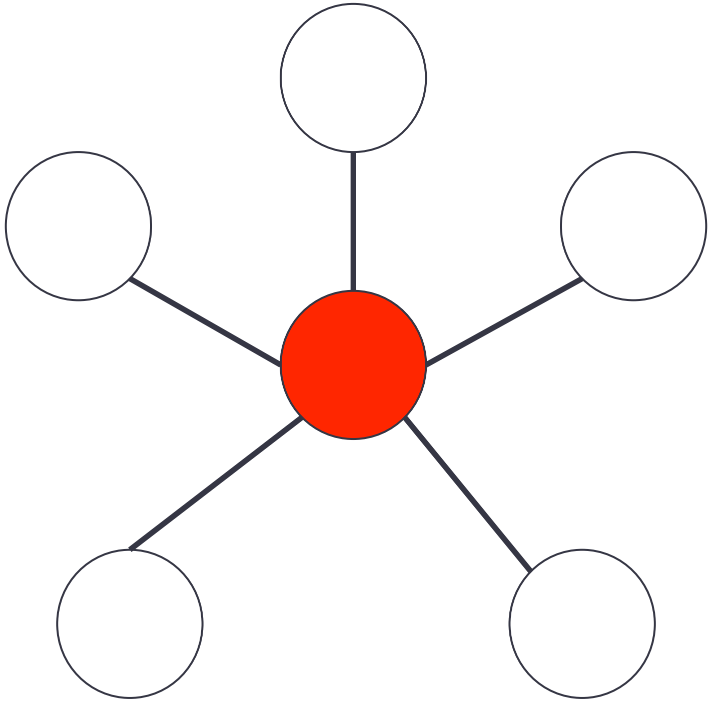
  + centralized management

  + failure of a node does not affect

  - central devise failure affect all the network

  - cost

  - number of nodes depends on capacity of central device



Mesh topology

- Nodes are connected directly to another nodes
- Types:
  - fully-connected
  - partly-connected

+ simultaneously

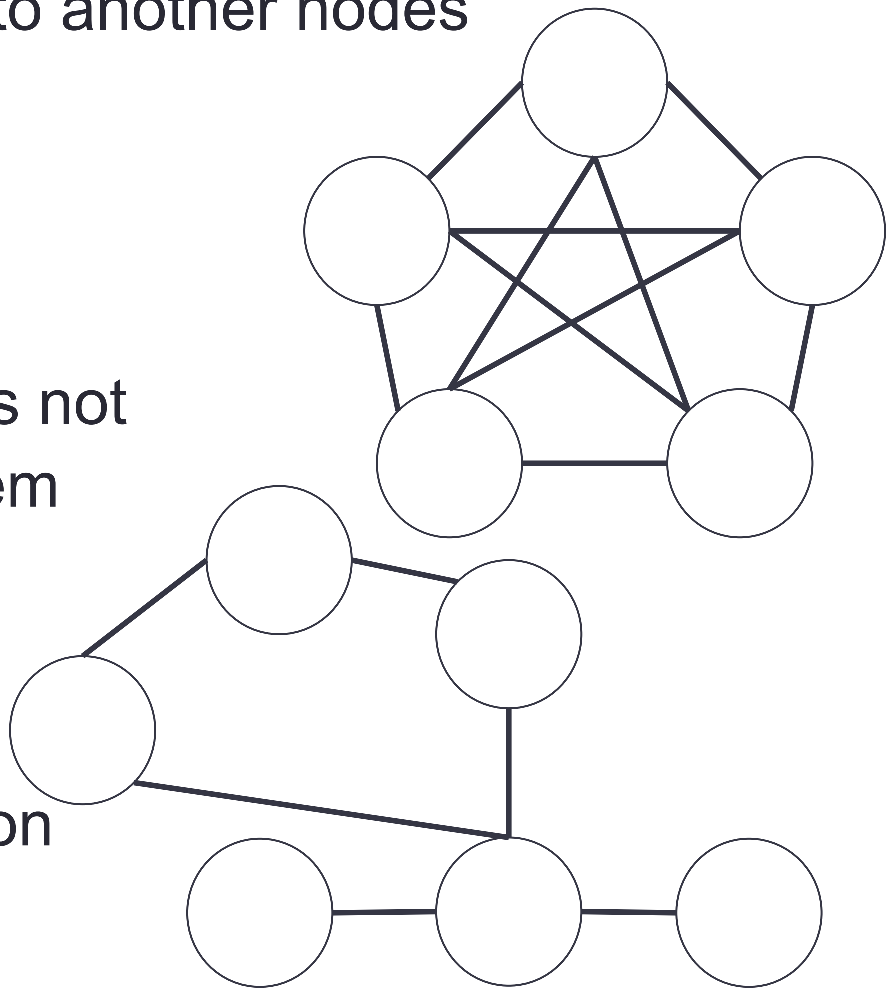
+failure of one node does not affect on the system

+ easy to modify

- high redundancy

- cost

- set-up and administration is difficult



Tree (hierarchy) topology

- Star topology are connected using Bus topology

  + expansion is easy

  + easy to manage and maintain

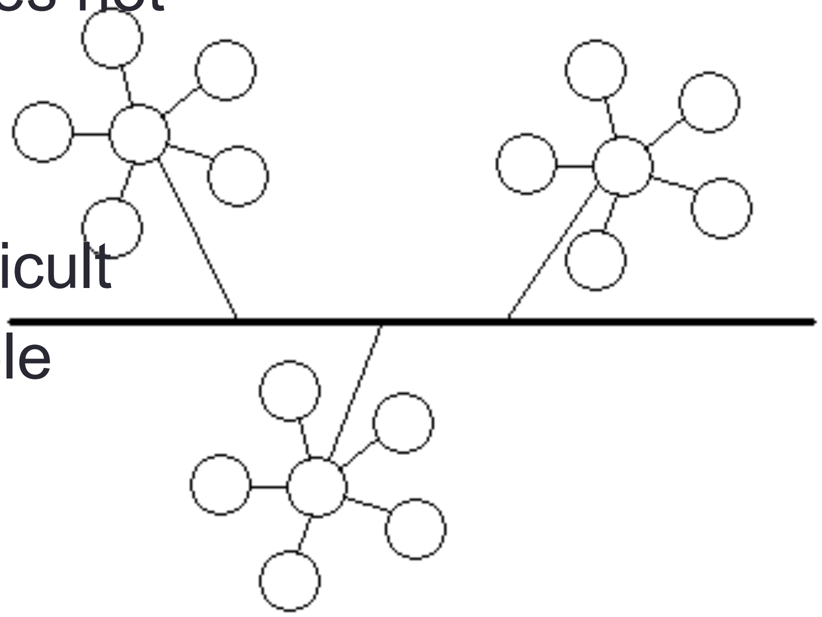
  + error detection and correction are easy

  + failure of one segment does not affect to the system

  - central bus

  - maintenance becomes difficult

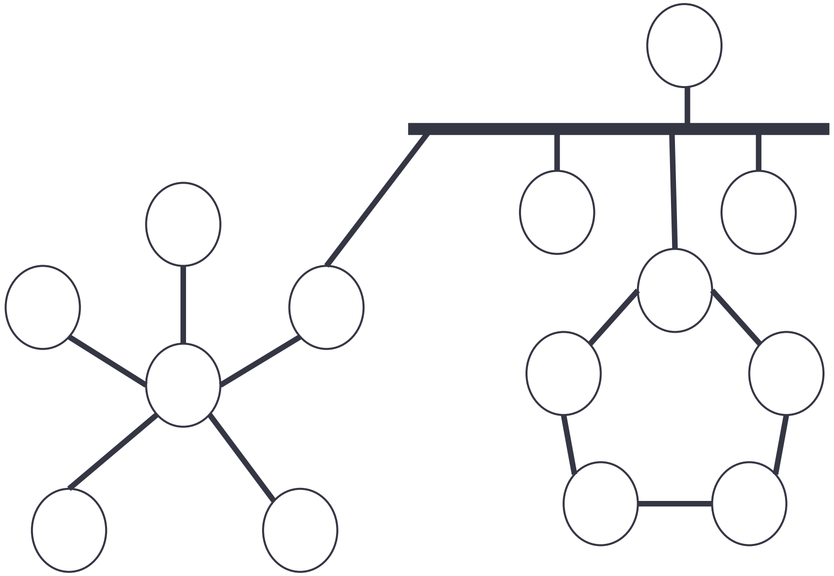
  - scalability depends on cable



Hybrid topology

- Combines two or more topologies
- Reap their advantages

+ scalable
+ flexible
+ effective
- complex
- cost



Hypercube topology

- The distance between any two nodes is at most log(p)
- Each node has log(p) neighbors

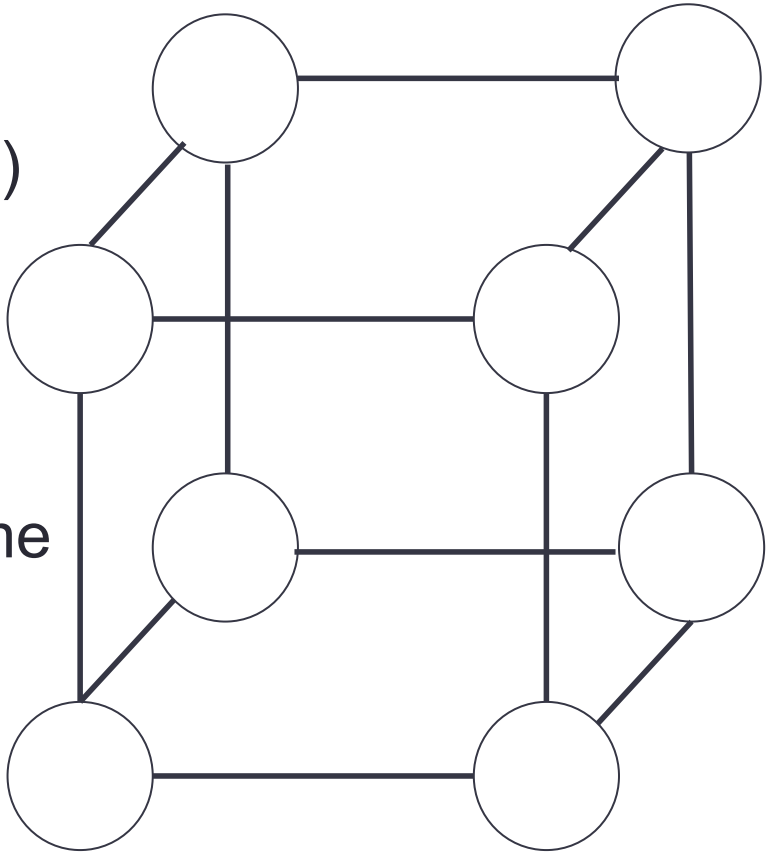
  + only log(p) neighbors

  + the longest route is log(p)

  + excellent connectivity

  - physical network infrastructure is ignored

  - building one node at a time

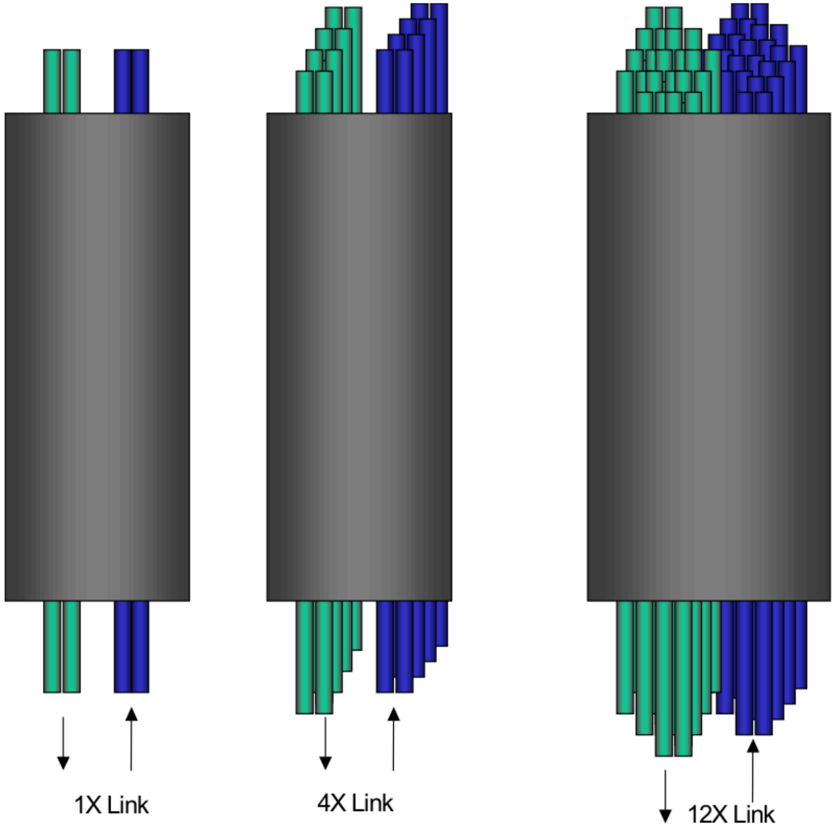


InfiniBand (IBA)

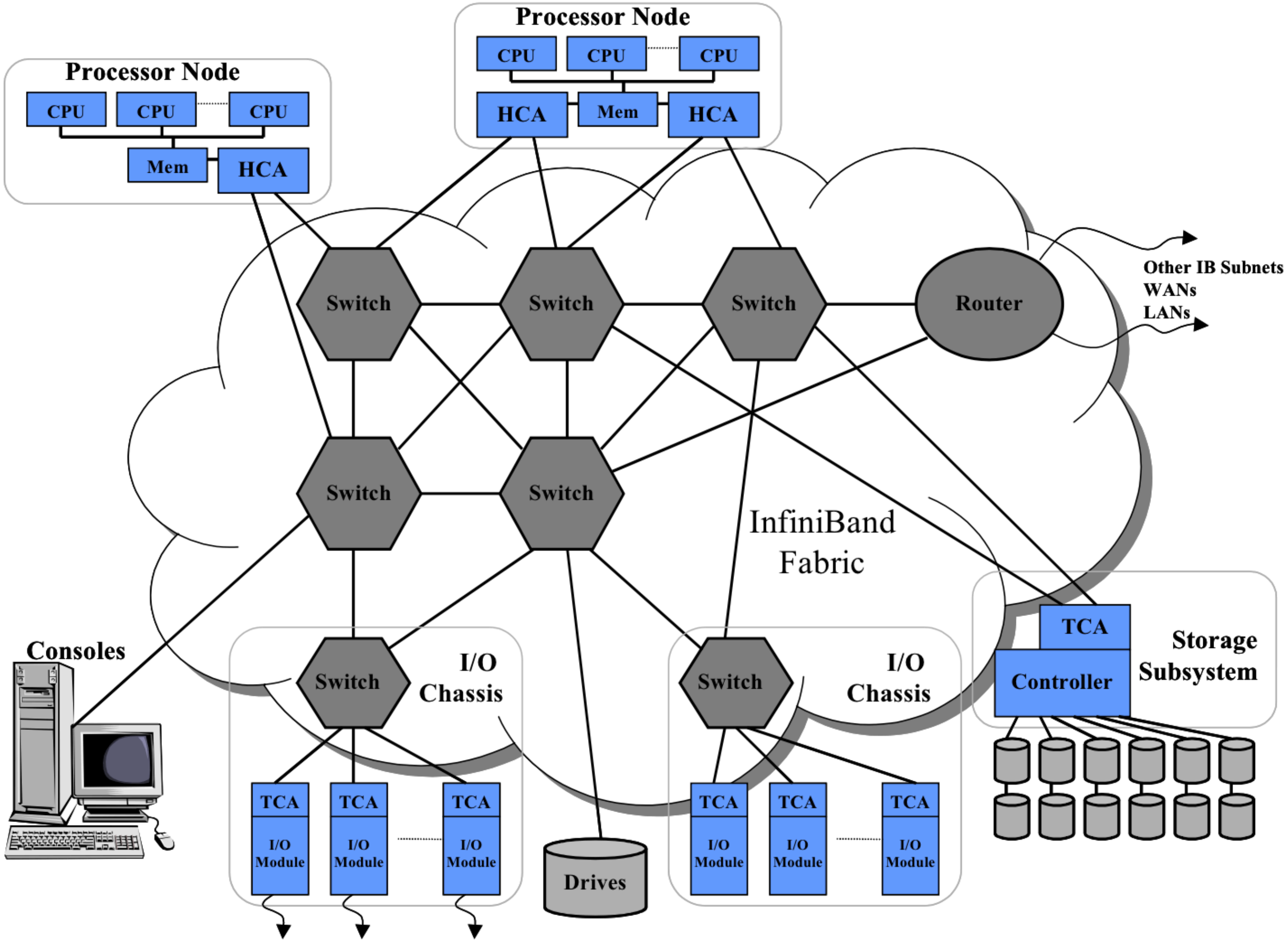
- High bandwidth
- Low latency
- Bi-directional
- “Shared-nothing” architecture
  - Not able to address directly
- Support for up to 64,000 addressable devices

IBA bandwidth

- 1X Link – 500MB/sec
- 4X Link – 2GB/sec
- 12X Link – 6 GB/sec



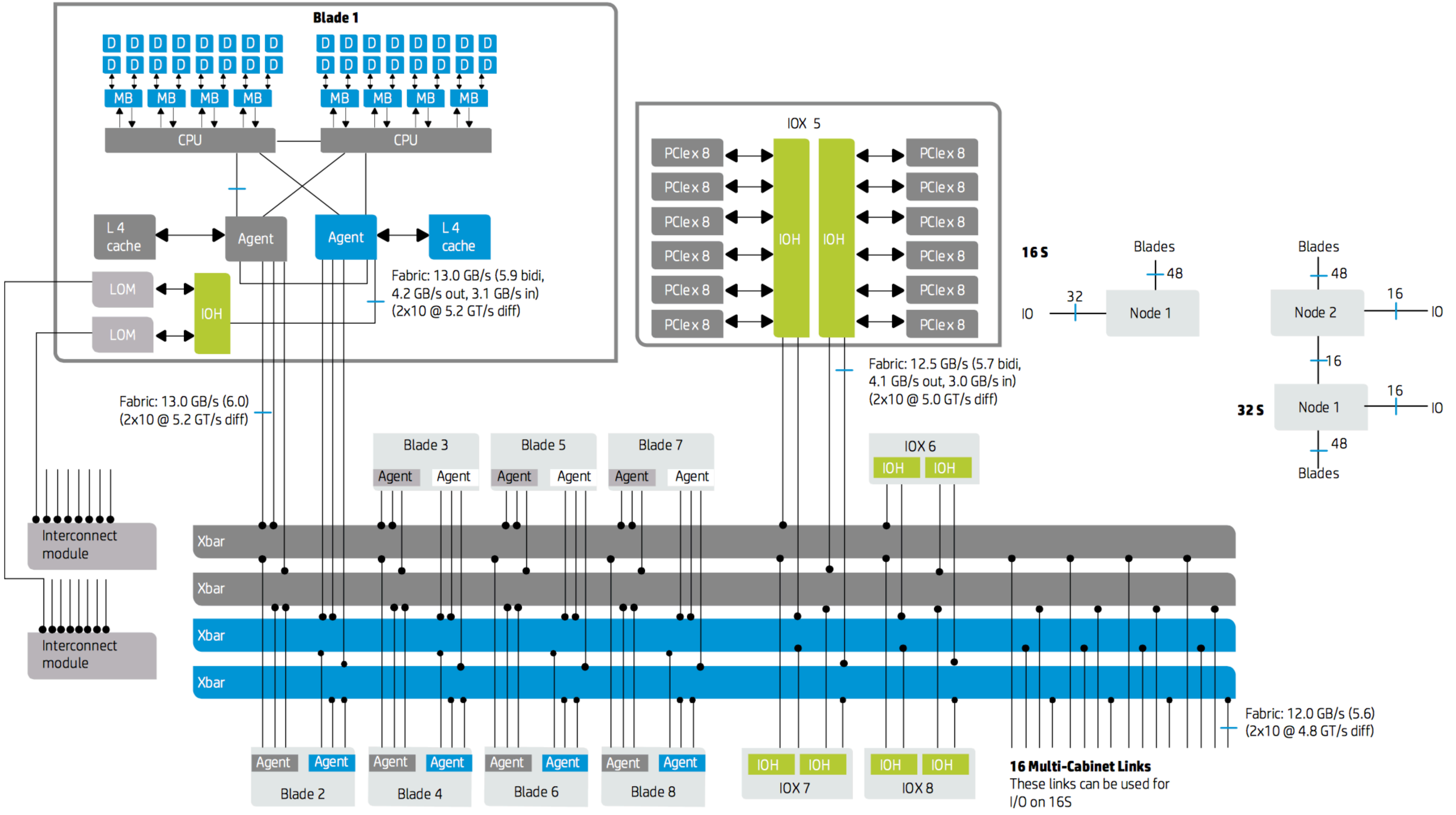
IBA System Fabric



IBA

- Used by 43% of the systems on the Top-500 list of supercomputers (Nov. 2010)
- Intel® Xeon Phi™
- Mellanox ConnectX IB

Crossbar switch (Integrity Superdome 2)



Fabric: 13.0 GB/s (5.9 bidi, 4.2 GB/s out, 3.1 GB/s in) (2x10 @ 5.2 GT/s diff)

Fabric: 12.5 GB/s (5.7 bidi, 4.1 GB/s out, 3.0 GB/s in) (2x10 @ 5.0 GT/s diff)

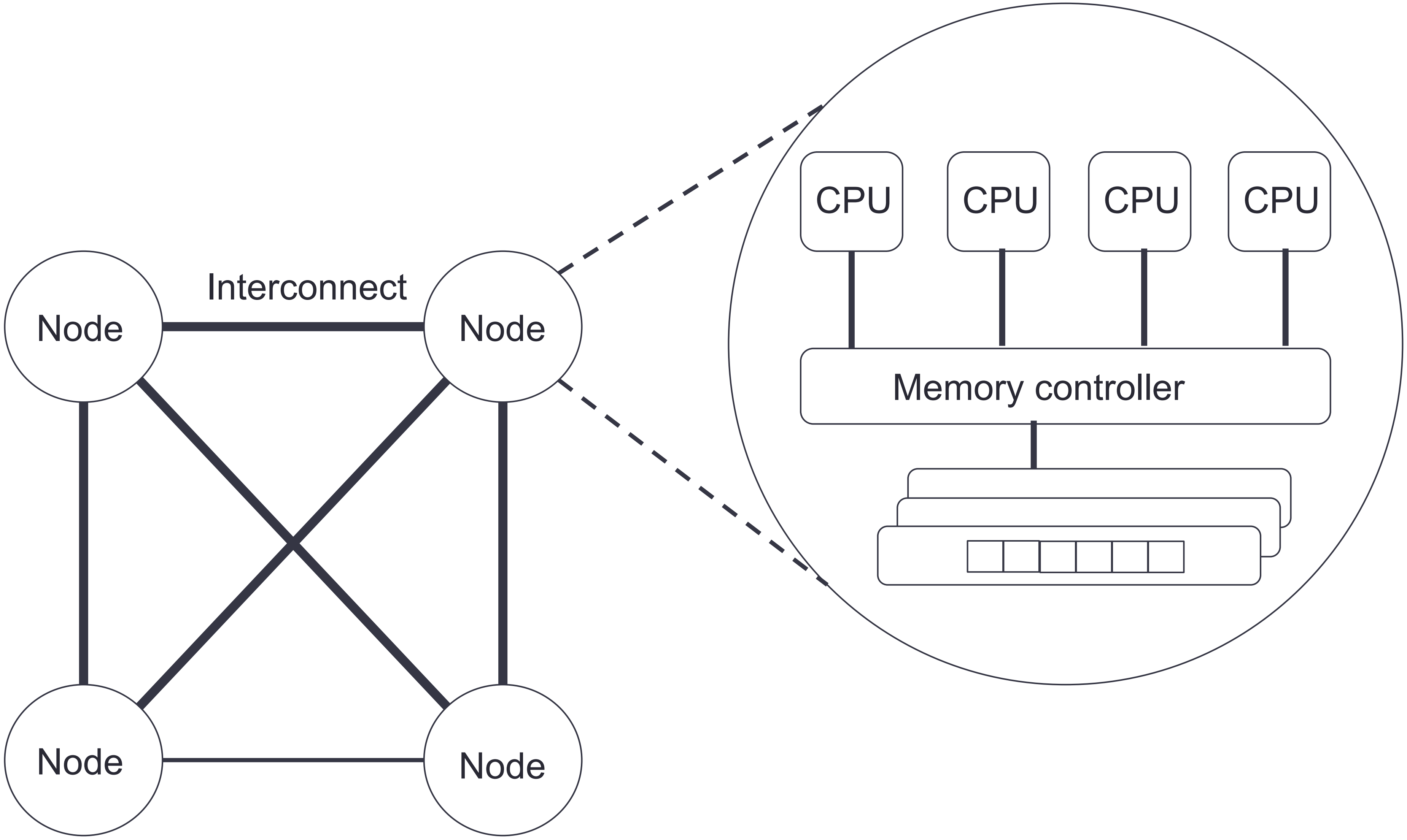
Fabric: 13.0 GB/s (6.0) (2x10 @ 5.2 GT/s diff)

Fabric: 12.0 GB/s (5.6) (2x10 @ 4.8 GT/s diff)

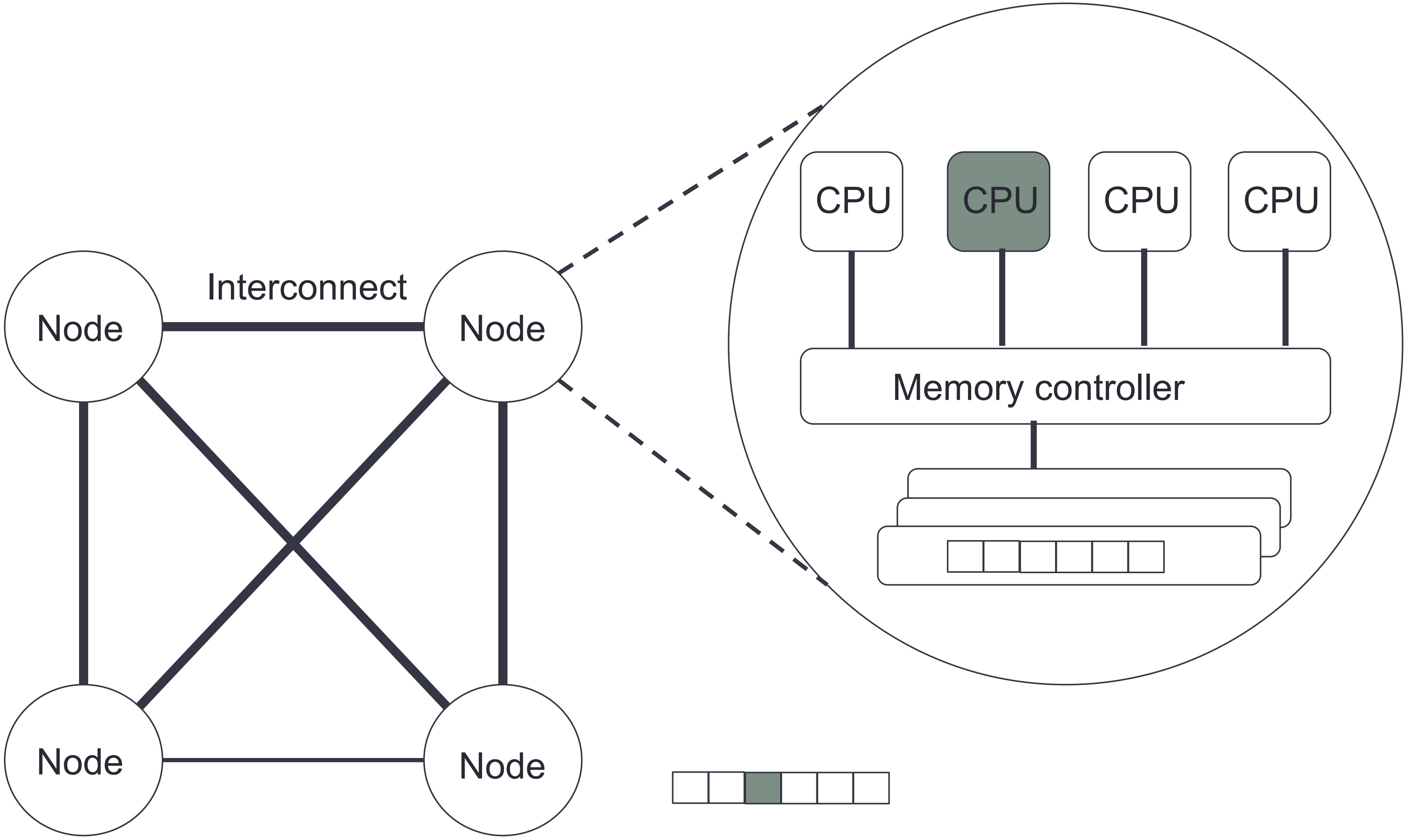
16 Multi-Cabinet Links
These links can be used for I/O on 16S

INTERCONNECTION TECHNOLOGIES IN NUMA

Non-Uniform Memory Access



Interconnectors



NUMA Interconnectors

- AMD Hyper-Transport (HT)
- Intel Quick-Path Interconnect (QPI)
- NumaLink

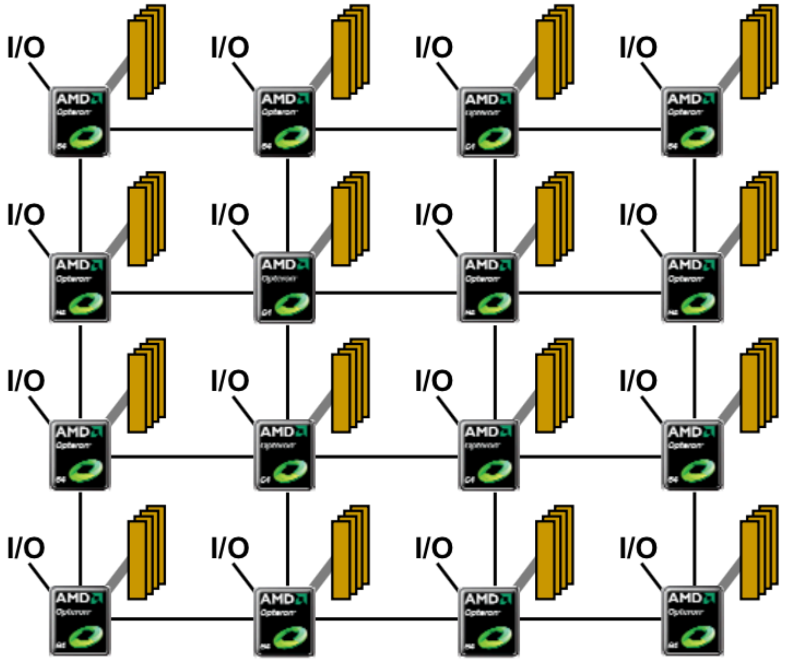
AMD Hyper-Transport (HT)

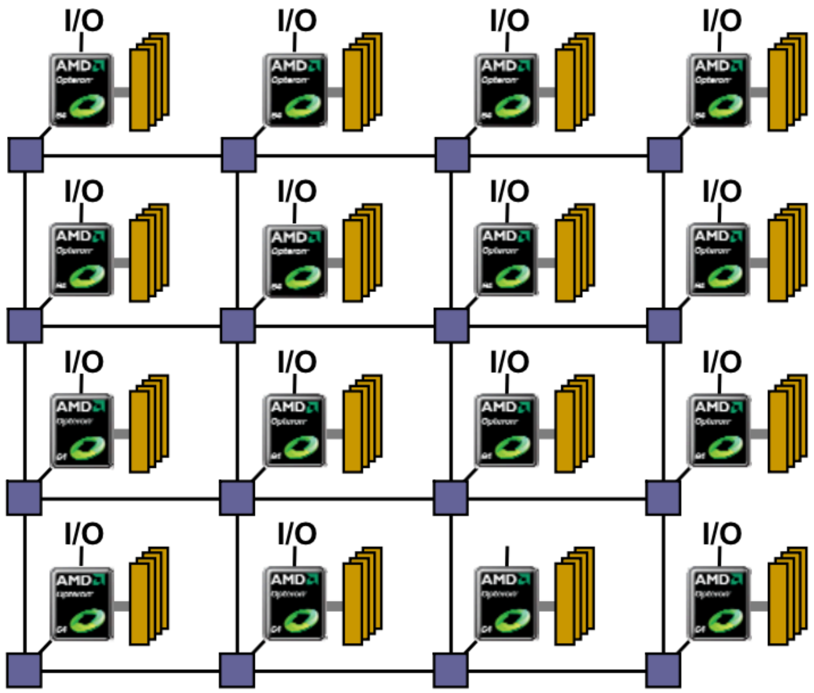
- High bandwidth
- Low latency
- Bi-directionally
- Multiple configuration
  - Daisy chain
  - Star
  - Mesh
- HyperTransport 3.10

AMD Hyper-Transport (HT)

- Minimal chain: two nodes
- Maximal chain: 32 nodes
- Up to 3.2 GB/sec over 8-bit I/O link
- Up to 12.8 GB/sec over 16-bit I/O link
- Up to 25.6 GB/sec over 32-bit I/O link

HT Implementation





AMD Hyper-Transport (HT)

- Opteron
- Athlon 64
- Phenom

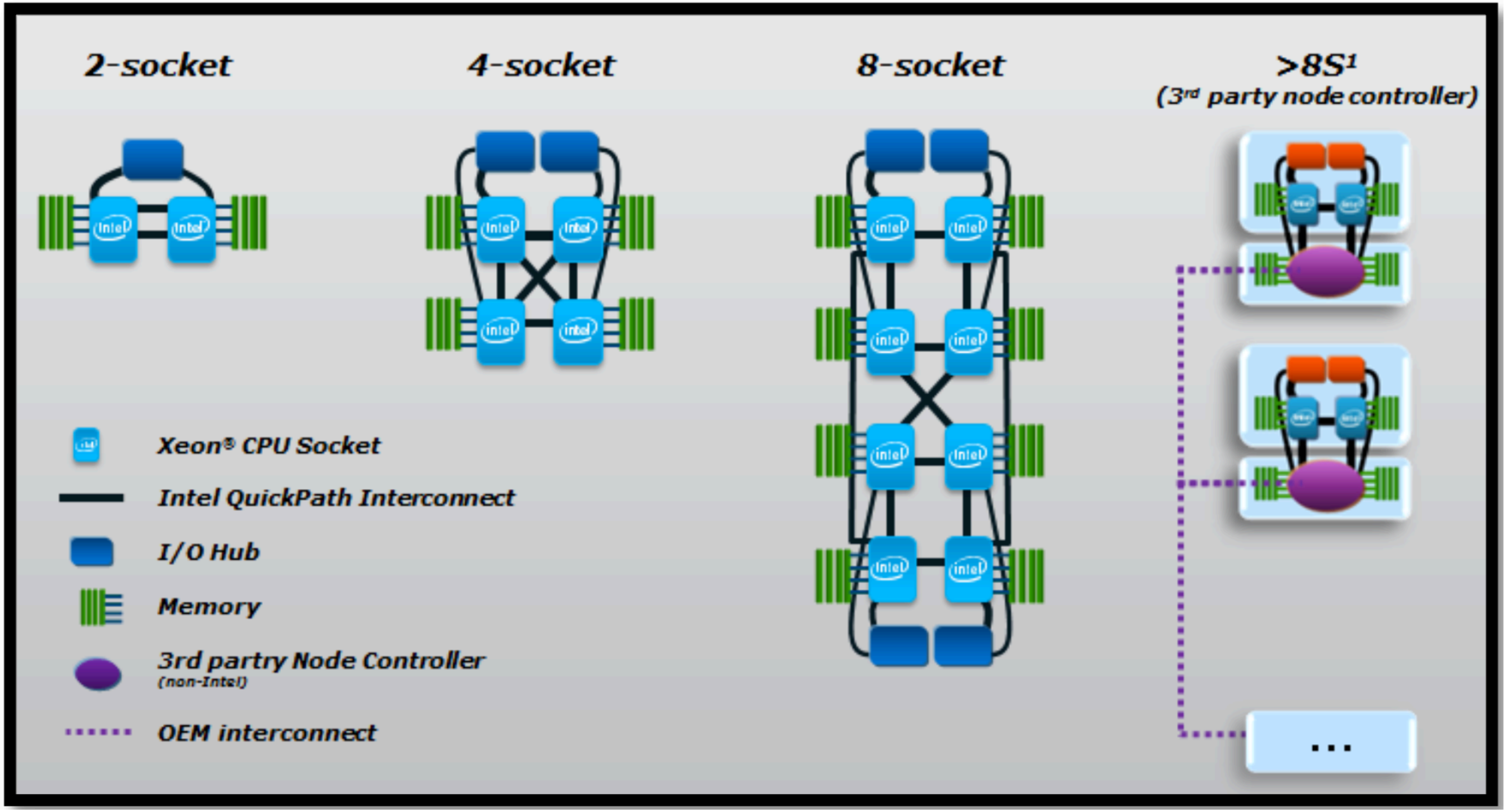
Intel Quick-Path Interconnect (QPI)

- High bandwidth
- Low latency
- Bi-directional
- Up to 25.6. GB/sec per link
- High scalable

QPI Physical Layer

- Two 20-lane point-to-point data links
- Full duplex with a separate clock pair in each direction
- 42 signals
- Total number of pins is 84
- 20 data lanes are divided onto four "quadrants" of 5 lanes each

Intel Quick-Path Interconnect (QPI)



Intel Quick-Path Interconnect (QPI)

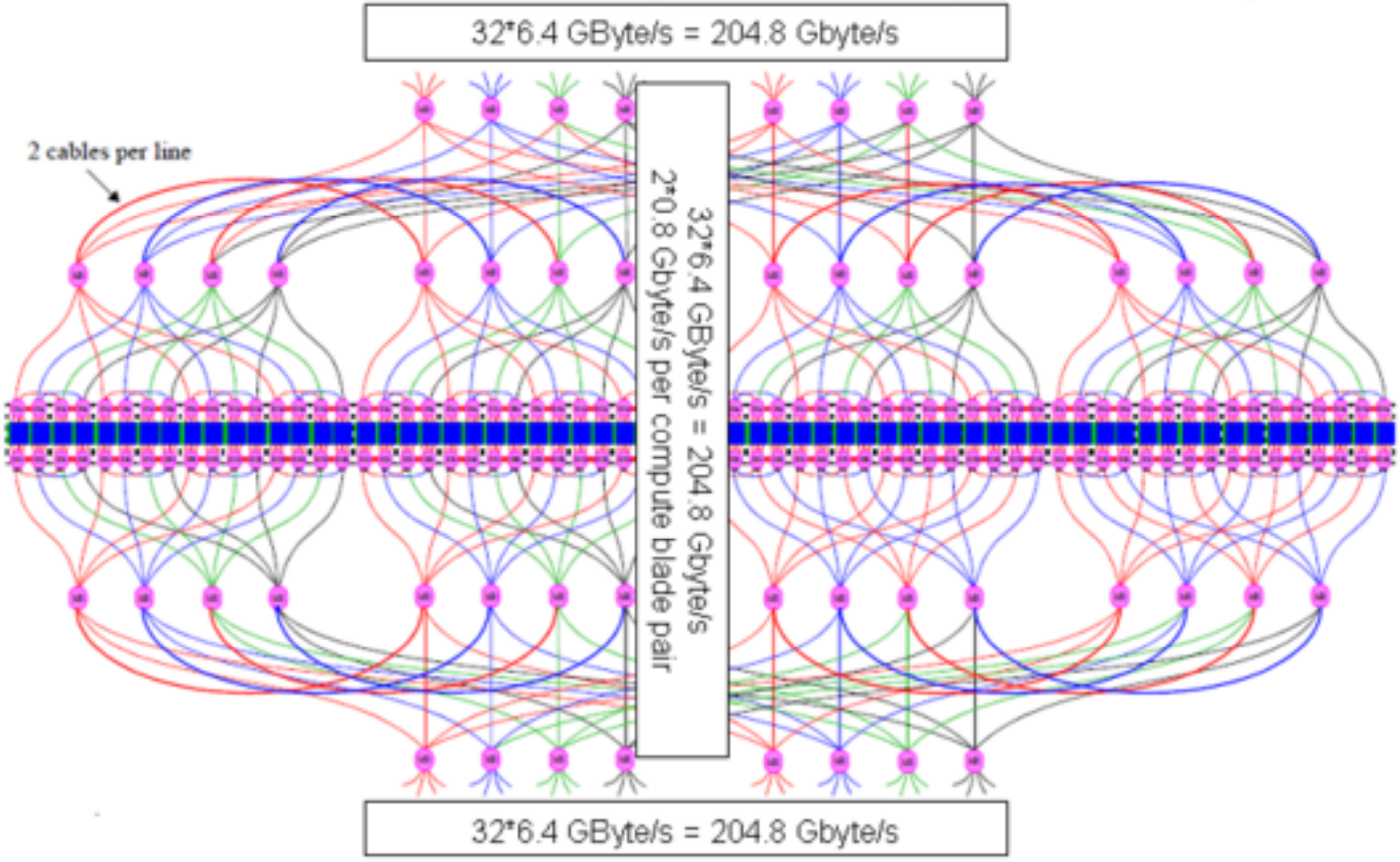
- Xeon
- Core i7
- Itanium

SGI Numalink

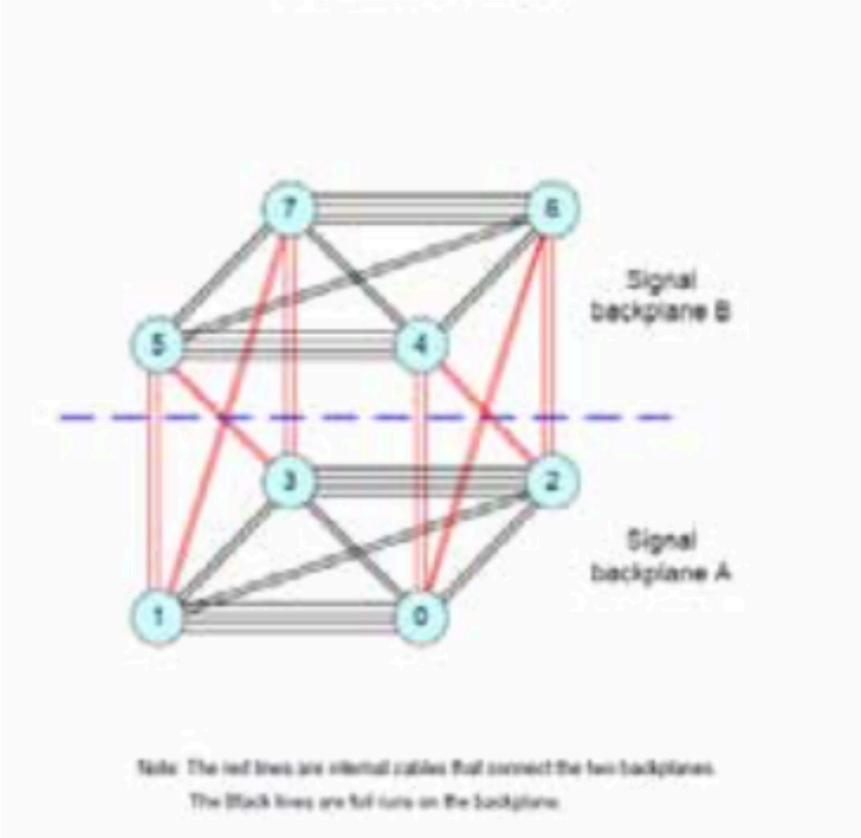
- High bandwidth
- Low latency
- Bi-directional
- Connected into the memory infrastructure of the system
- Reduce one-way request to memory
- Numalink3 3.2 GB/sec
- Numalink4 6.4 GB/sec
- Different topology
- Max 2048 nodes

SGI Numalink. Tree configuration

Intra-Partition
NUMAlink configuration (Fat Tree)

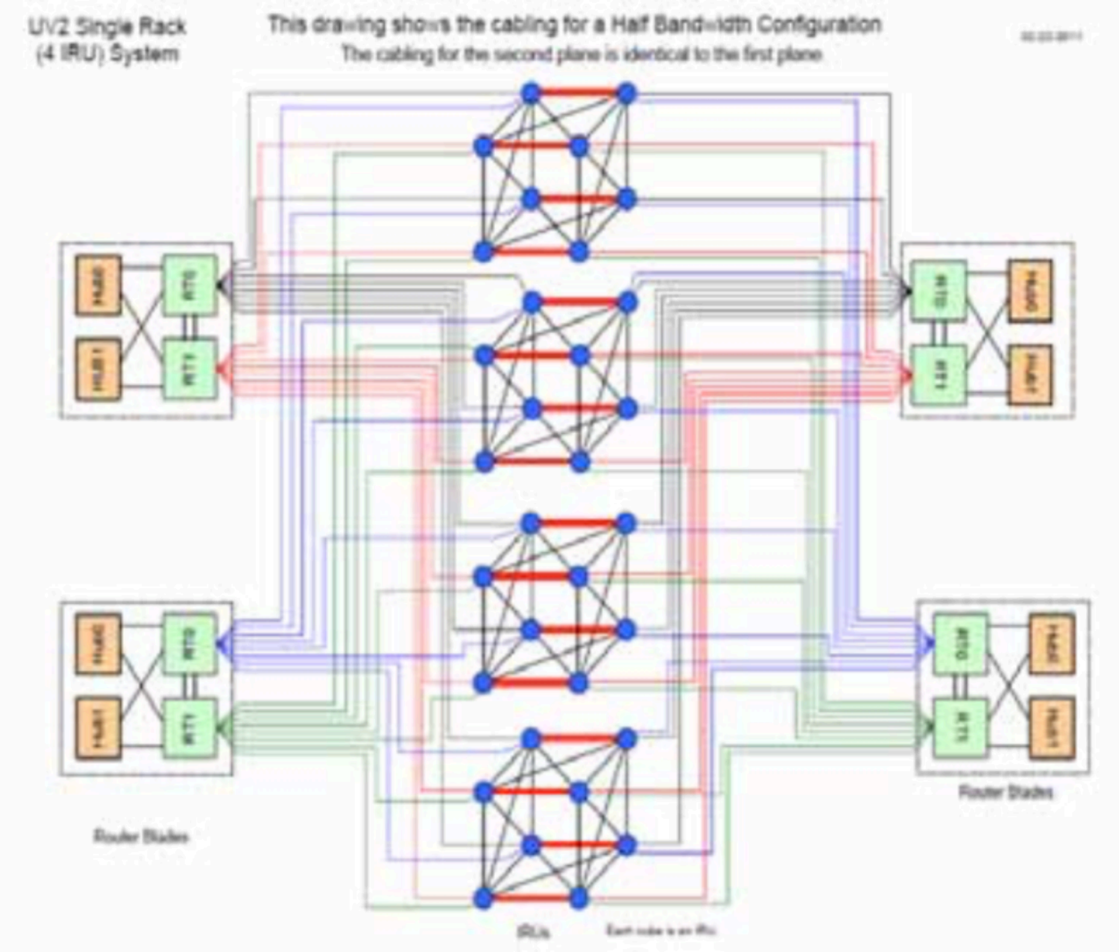

32*6.4 GByte/s = 204.8 Gbyte/s
2 cables per line
32*6.4 GByte/s = 204.8 Gbyte/s
2*0.8 Gbyte/s per compute blade pair
32*6.4 GByte/s = 204.8 Gbyte/s

SGI UV 2000. Hypercube configuration



Note: The red lines are internal cables that connect the two backplanes.
The Black lines are full runs on the backplane.

SGI Numalink. Hypercube configuration


UV2 Single Rack
(4 IRU) System
This drawing shows the cabling for a Half Bandwidth Configuration
The cabling for the second plane is identical to the first plane
HUB 0
HUB 1
RT 0
RT 1
Router Blades
Router Blades
IRUs
Each cube is an IRU

SGI Numalink

- Altix servers and supercomputers

Questions?