Cray High Speed Networking

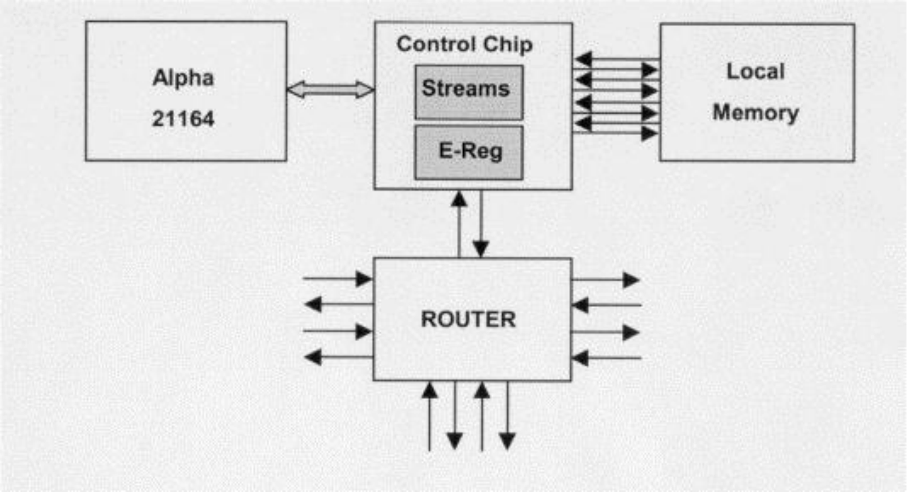
Robert Alverson

Overview

- **History of Cray networks**
- **The Cascade network**
  - Network interface
  - Topology
  - Routing
  - Implementation

History

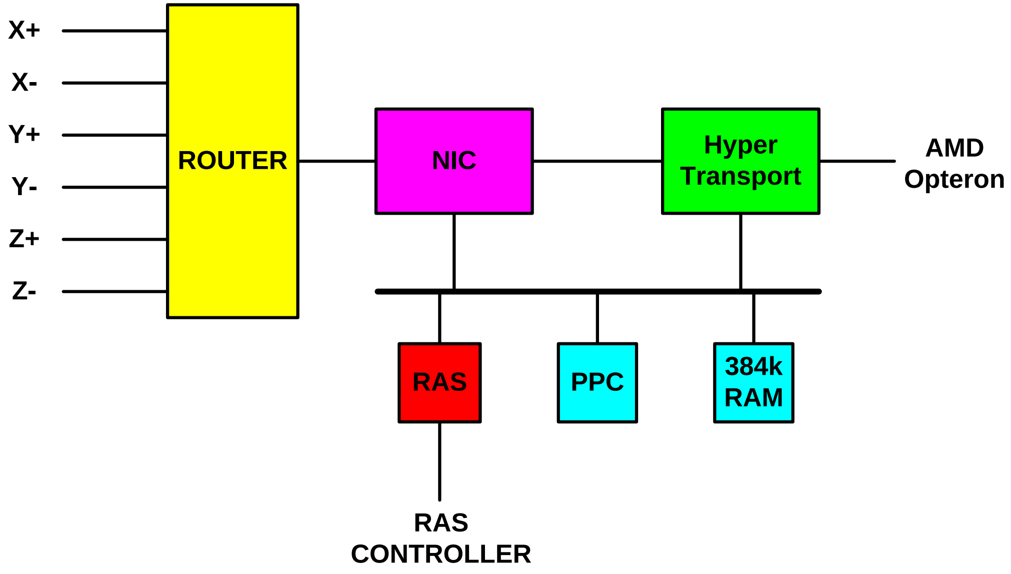
- **Cray Intel**
- **Pre-historic**
  - T3E torus
  - E-registers



  - Cray X2 “Black Widow”
  - Fat-tree using YARC 64 port switch

History

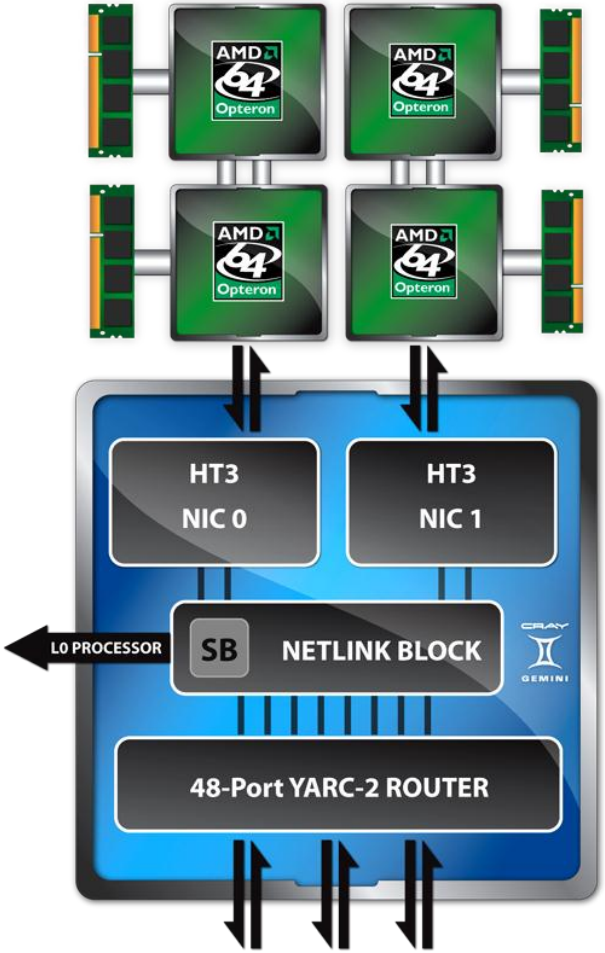
- **Seastar (Hot Interconnects 2003)**
  - NIC and 7 port switch integrated
  - HyperTransport 1.0
  - 4 virtual channels
  - Scalable messaging
  - Portals 3
  - Threadstorm interface
  - “Slow” PowerPC



History

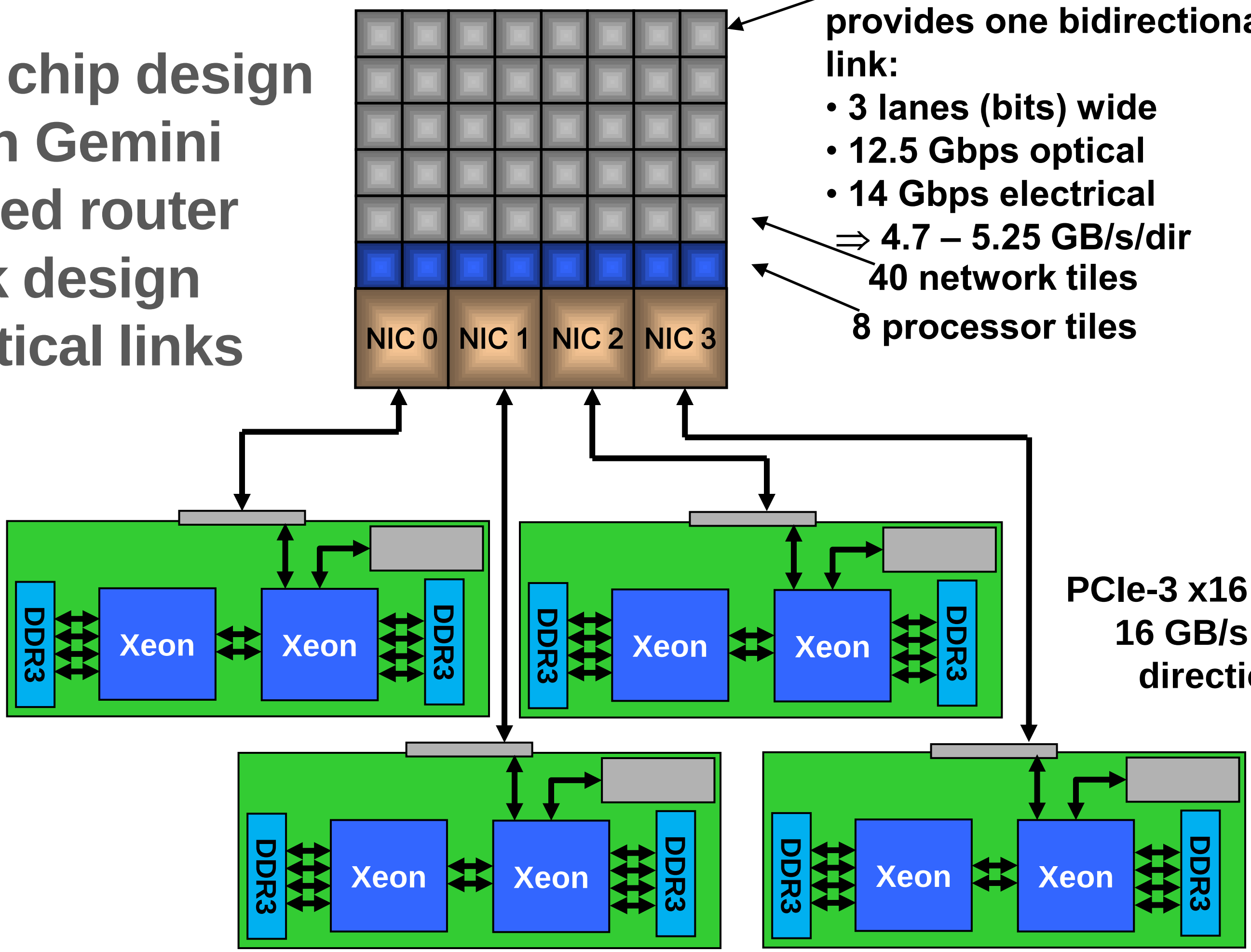
- **Gemini (Hot Interconnects 2010)**
  - 2 NICs and 48 port switch integrated
  - HyperTransport 3.0
  - 2 virtual channels
  - OctigaBay technology
  - Fine-grain remote PUT/GET
  - Support for more topologies including hypercube

  - Only ever used in torus



Cascade Network

- Aries ASIC
- System on a chip design
- NIC based on Gemini
- High radix tiled router
- New network design
- Electrical/optical links



48 router tiles, each provides one bidirectional link:

- 3 lanes (bits) wide
- 12.5 Gbps optical
- 14 Gbps electrical

⇒ 4.7 – 5.25 GB/s/dir

40 network tiles

8 processor tiles

PCIe-3 x16 bits at 16 GB/s per direction

PCI Express Interface

- Interoperate with Intel, AMD, …
- IOMMU provided (no reliance on AMD GART)
- Higher per packet overhead

Interface	Cray ASIC	Raw GB/s	Bytes for 8B Write
HyperTransport 1.0	Seastar	3.2	16
HyperTransport 3.0	Gemini	10.4	20
PCI Express 3.0	Aries	16.0	32

- More stringent PCIe deadlock requirements
  - Gemini can nearly deadlock, recovers
  - Aries drops PCIe writes when buffers would overflow
  - Provides for user level flow control to avoid back-pressure

New Aries NIC Features

- **Fast Memory Access (FMA)**
  - Minimize overhead for 8-64 byte operations
  - FMA launch: Fast path for single word put, get, and non-fetching AMO
- **User space Block Transfer Engine (RDMA)**
  - Reduces latency of issuing block transfer request
- **IOMMU in Aries**
  - Use of large pages is a performance optimization
  - Not dependent on address translation by PCI Express host

Collective support

- **Latency optimization for our most important cases**
  - Integer and floating point add
  - Max/min, Compare & Swap, bit operations
- **NIC based**
  - No switch state to allocate or manage
  - No requirement to understand topology when constructing tree
  - Up to radix 32

Network Topology

- **Desire for more global bandwidth**
  - Largest Cray torus networks suffer on global traffic patterns
- **Application has non-local communication**
  - Unstructured traffic, communication load imbalance, many-to-many and all-to-all traffic, mismatch between system and job geometry
  - All of which increase the average hop count on a Torus
- **System benefits**
  - Reduced job-to-I/O interference
  - Reduced job-to-job interference

Dragonfly Network

Goals:

- Provide scalable global bandwidth
- Exploit low cost of short electrical links
- Reduce the required number of global optical hops
- Avoid the need for external router cabinets

Dragonfly concept

- Construct groups of locally-connected nodes
- Treat the group as single “super node” with very high radix
- Pool all the optical links coming out of the group into a single dimension
- Create a single all-to-all optical stage among the groups

Network Topology

- **Aries Dragonfly**
  - Two dimensions of all-all connected nodes comprise group
  - All-all connections between groups make dragonfly
- **Average hop count flat versus system size**
  - Direct route
    - Up to two hops within source group
    - One optical hop to destination group
    - Up to two hops within destination group
- **Bisection bandwidth per node fairly flat versus system size**
  - Asymptotically half of optical bandwidth
- **Heavy use of adaptive routing**
  - Select between direct route and Valiant route (random intermediate)
  - Adaptive feedback broadcast across chip

Network Links

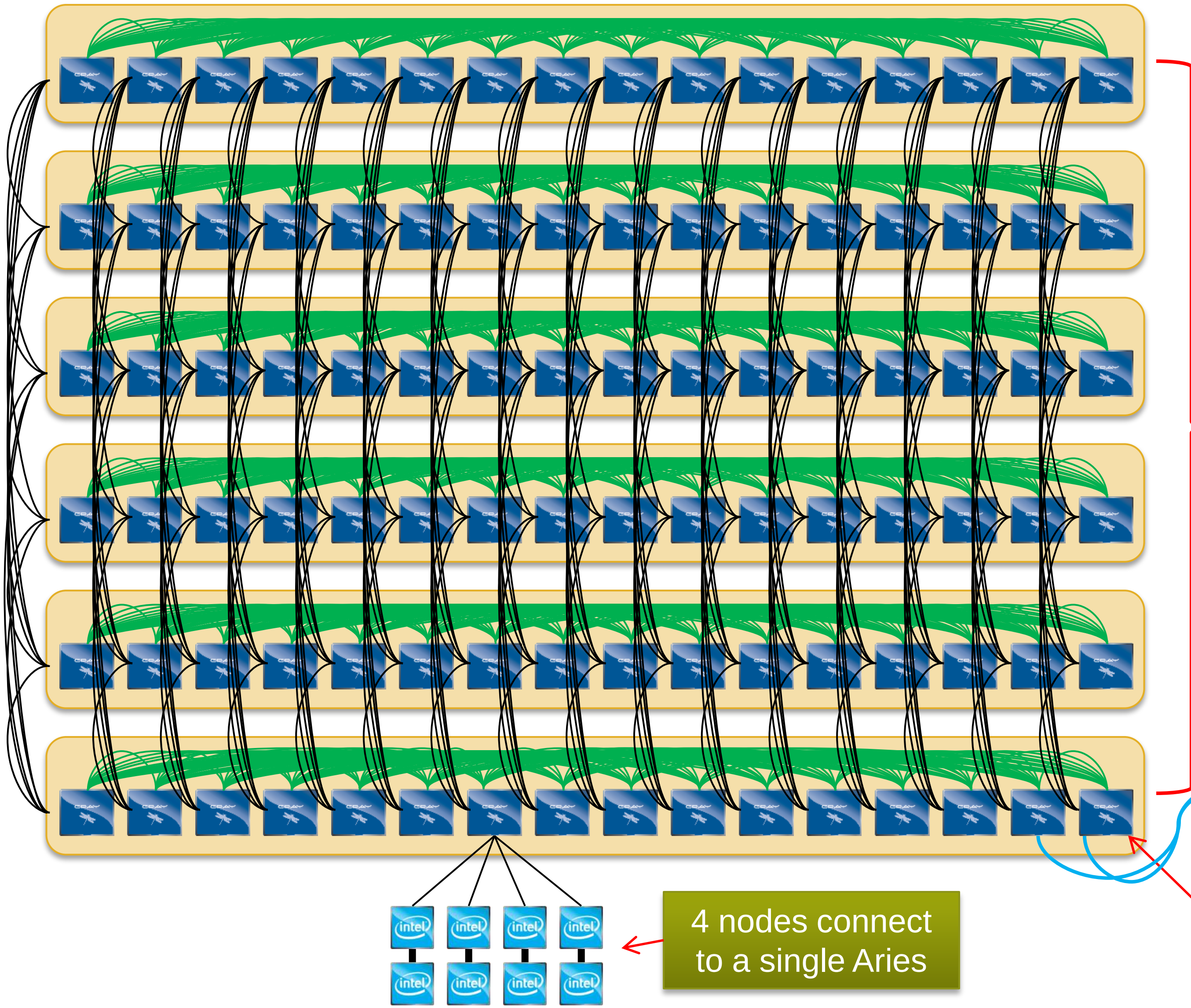
Electrical for short connections

- 14 Gbit/sec
- Within group

Optical for longer connections

- 12.5 Gbit/sec
- Group to group connections
- Expensive cables

Cascade – Local Electrical Network



backplanes connected with copper cables in a group:
“Black Network”

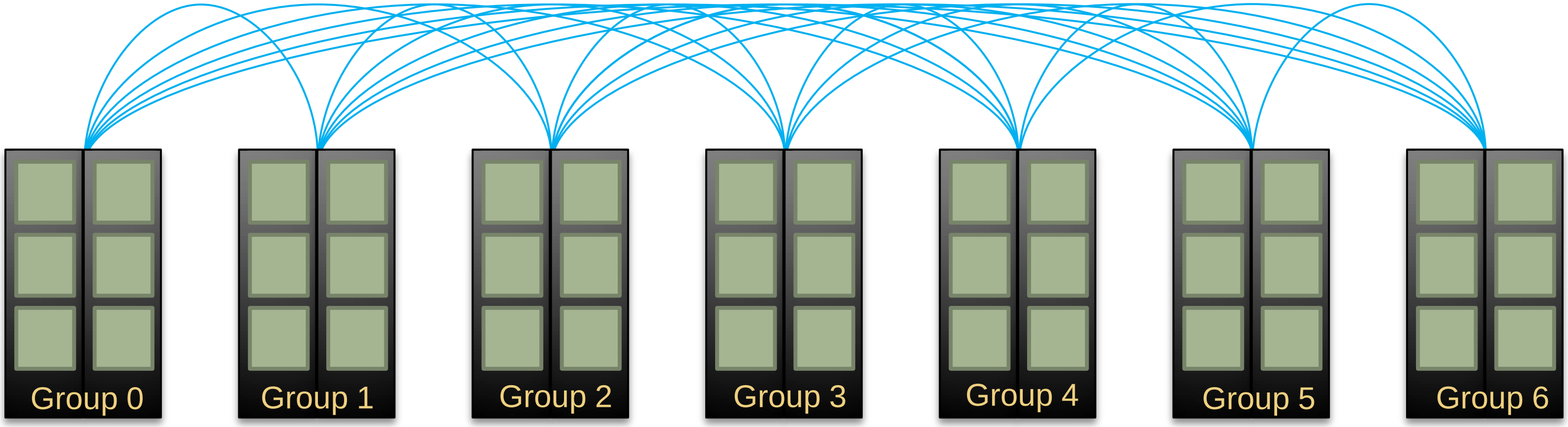
Optical cables interconnect groups
“Blue Network”

4 nodes connect to a single Aries

Aries connected by backplane
“Green Network”

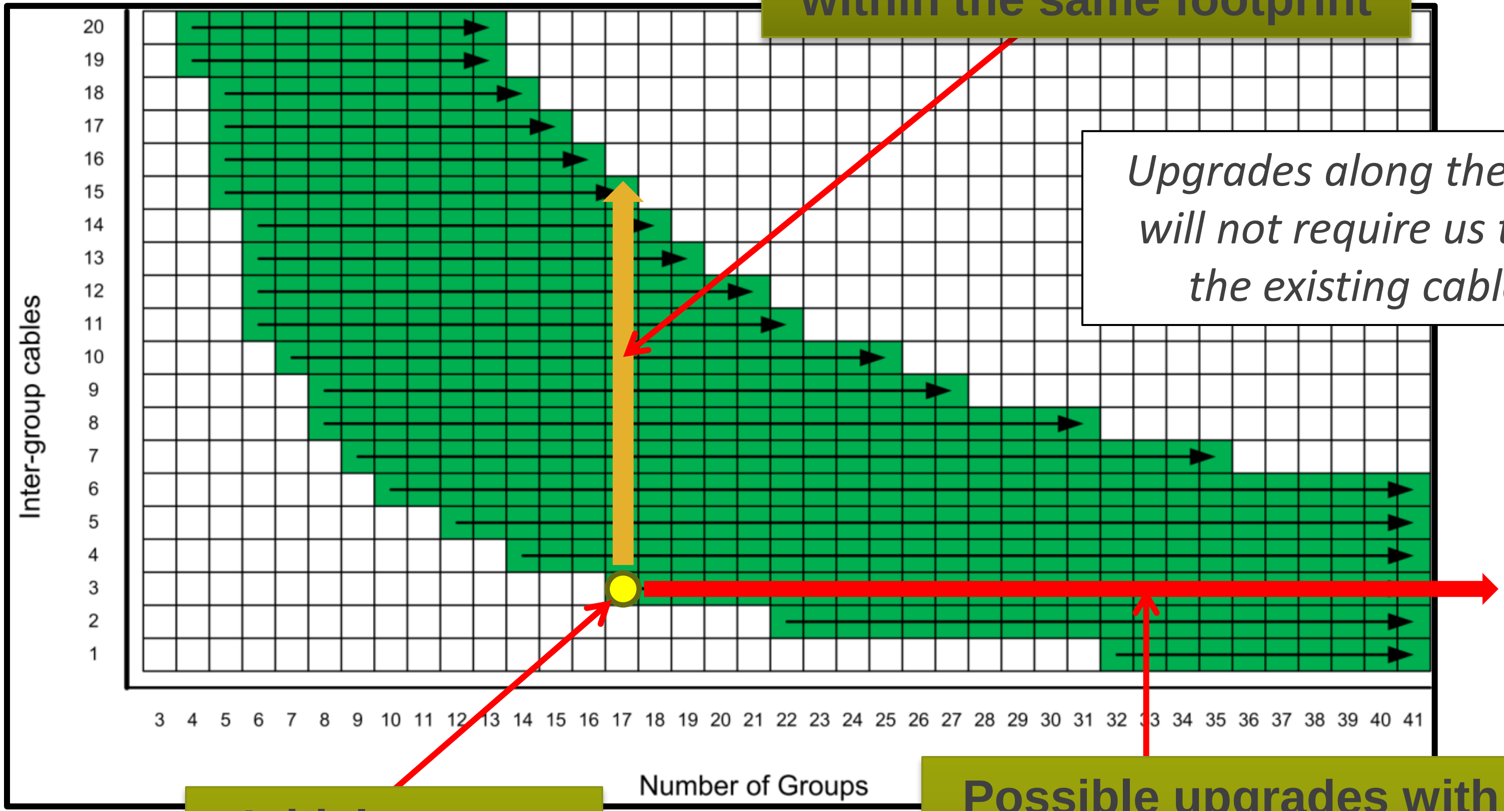
Cascade – Global Optical Network

- An all-to-all pattern is wired between the groups using optical cables (blue network)
- The global bandwidth can be tuned by varying the number of optical cables in the group-to-group connections



Example: A 7-group system is interconnected with 21 optical "bundles". The "bundles" can be configured between 2 or more cables wide, subject to the group limit.

Cascade Network Upgrade Options



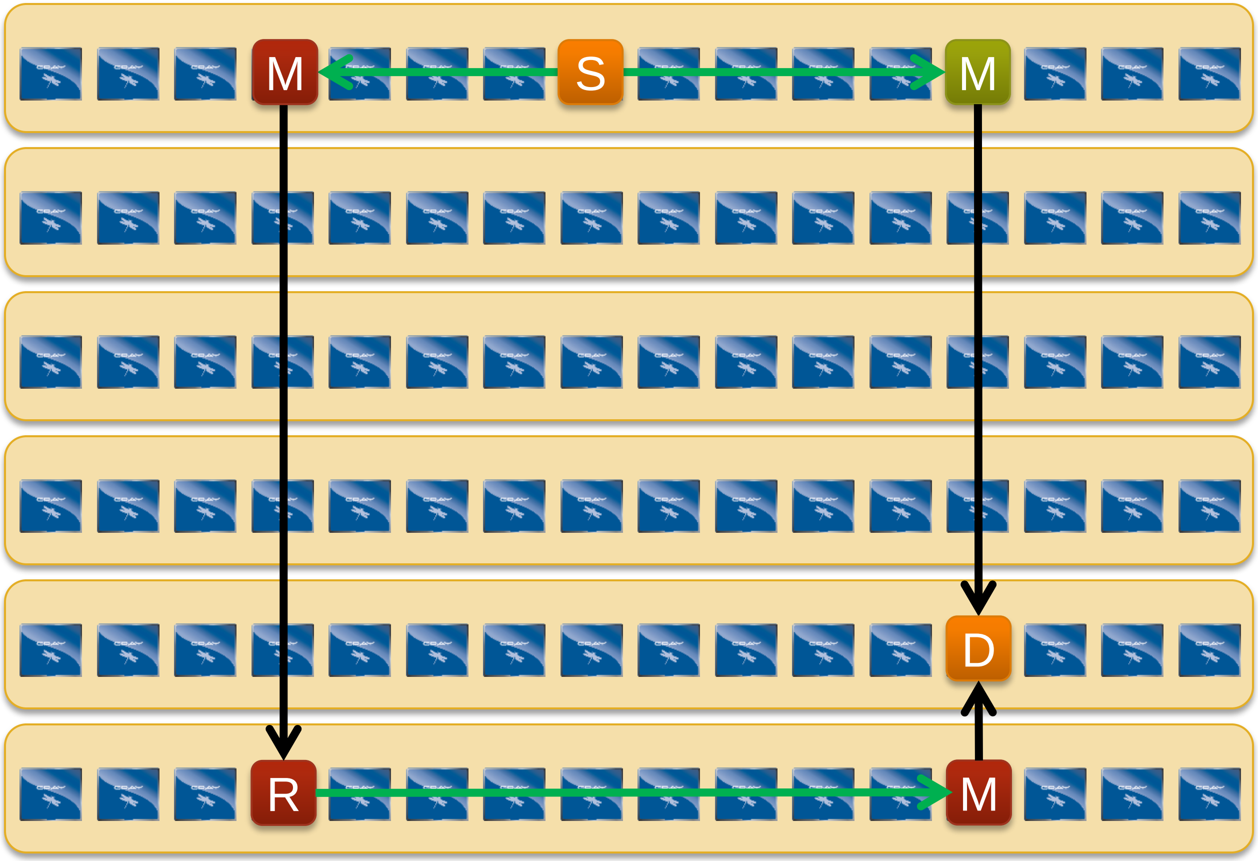
Upgrades to bandwidth within the same footprint

Upgrades along these arrows will not require us to disturb the existing cable mats

Initial system 17 groups, 3 wide bundles

Possible upgrades with 3-wide bundles (can carry this to 162 cabinets)

Cascade Routing – Intra Group



Minimal route between any two nodes in a group is just two hops

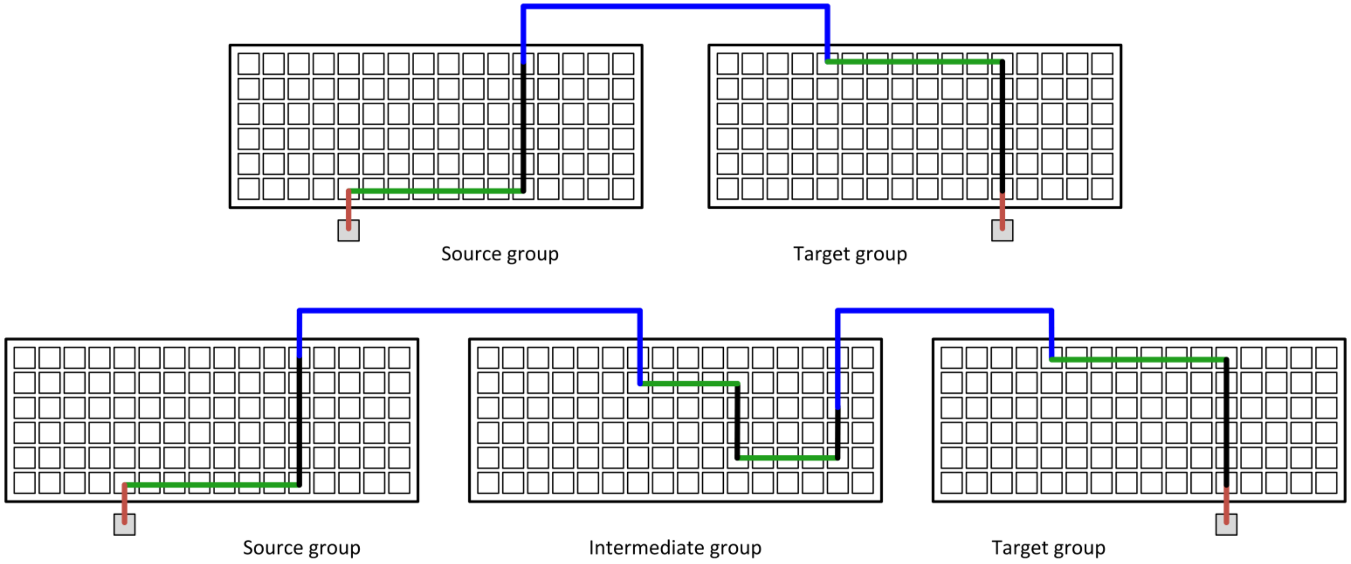
Non-minimal route requires up to four hops

The group has sufficient bandwidth to support full injection rate for all nodes

Adaptive routing selects between minimal and non-minimal paths based on load

Cascade Routing – Inter Group

Packets start on a minimal path, may switch to a non-minimal path if the load is lower

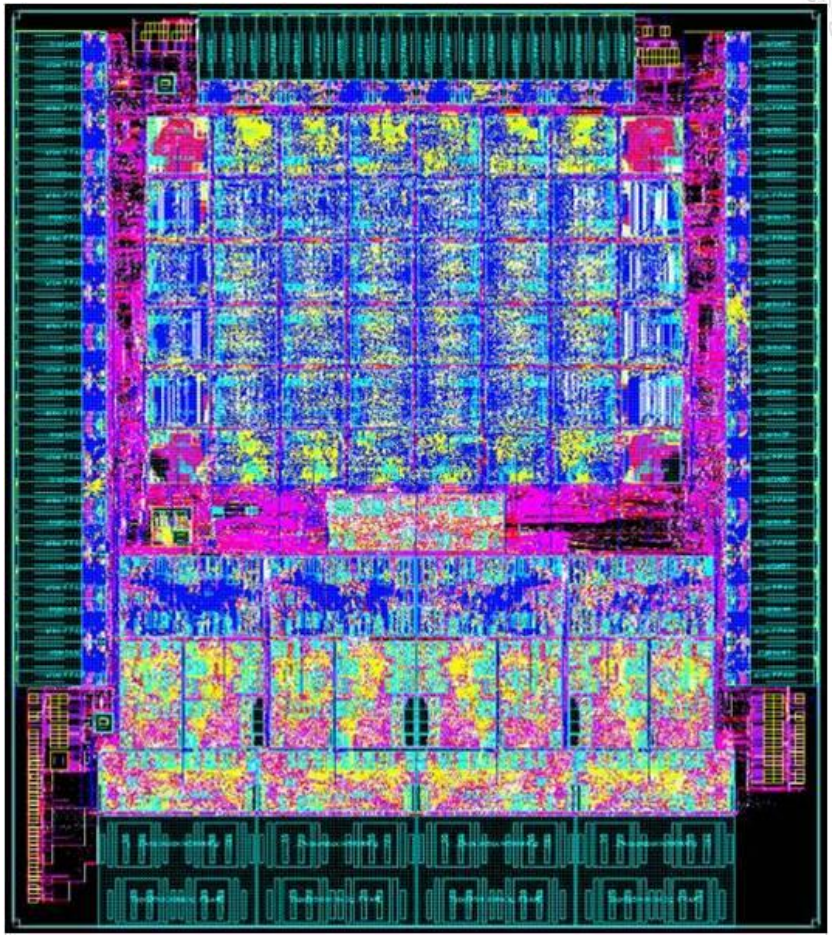


Comparison with Fat-Tree

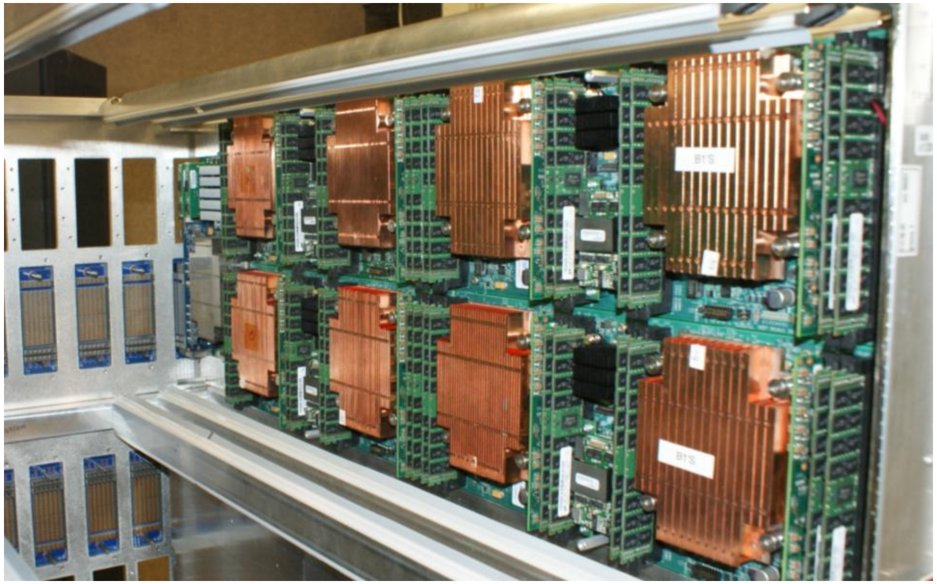
- **A fat-tree can also provide high global bandwidth**
- **But:**
  - Cost per node increases with system size. In particular the cost of the external router cabinets increases
  - Fat-tree requires twice as many optical links for a given global bandwidth
  - Two optical hops per path for fat-tree vs. one for dragonfly
- **The higher the proportion of traffic we can route minimally the bigger the advantage of Dragonfly**
  - Precisely the purpose of adaptive routing in Aries.
- **Traffic patterns in use by our most important customers**
  - Global traffic (all-to-all, uniform random) tends to be self load balancing, minimal routing works well
  - Traffic requiring non-minimal routing is more local, load on the global links is low. We have plenty of headroom for those that need two optical hops

Aries Data

- **40nm process**
- **Die size: 16.6 x 18.9mm**
- **Gate count: 217M**
- **184 lanes of high speed SerDes**
  - 30 optical network lanes
  - 90 electrical network lanes
  - 64 PCI Express lanes



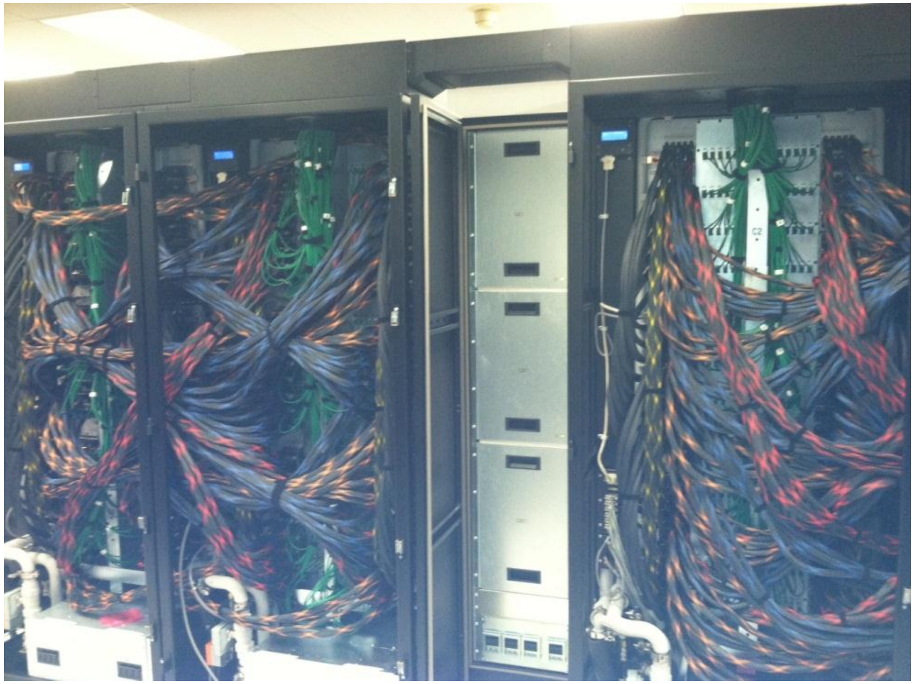
Compute Blade



- 4 Nodes
- Intel Xeon (Sandybridge) CPUs
- PCI-Express Gen3 x16 host interface

Network Wiring

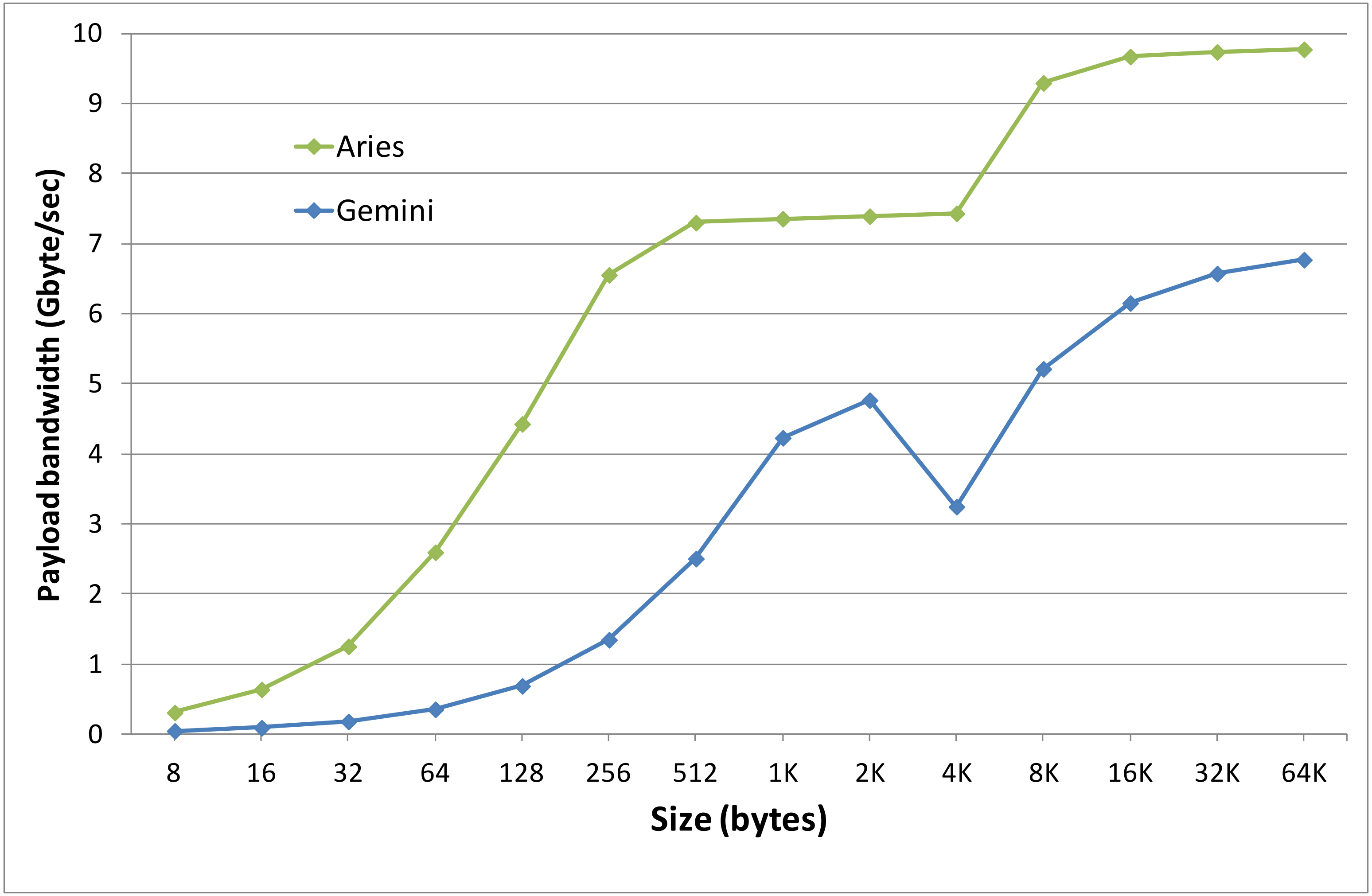
- Optical links
  - Green
  - Exit to top of cabinet
- Electrical links
  - Multi-colored to help organize
  - All-all by chassis evident



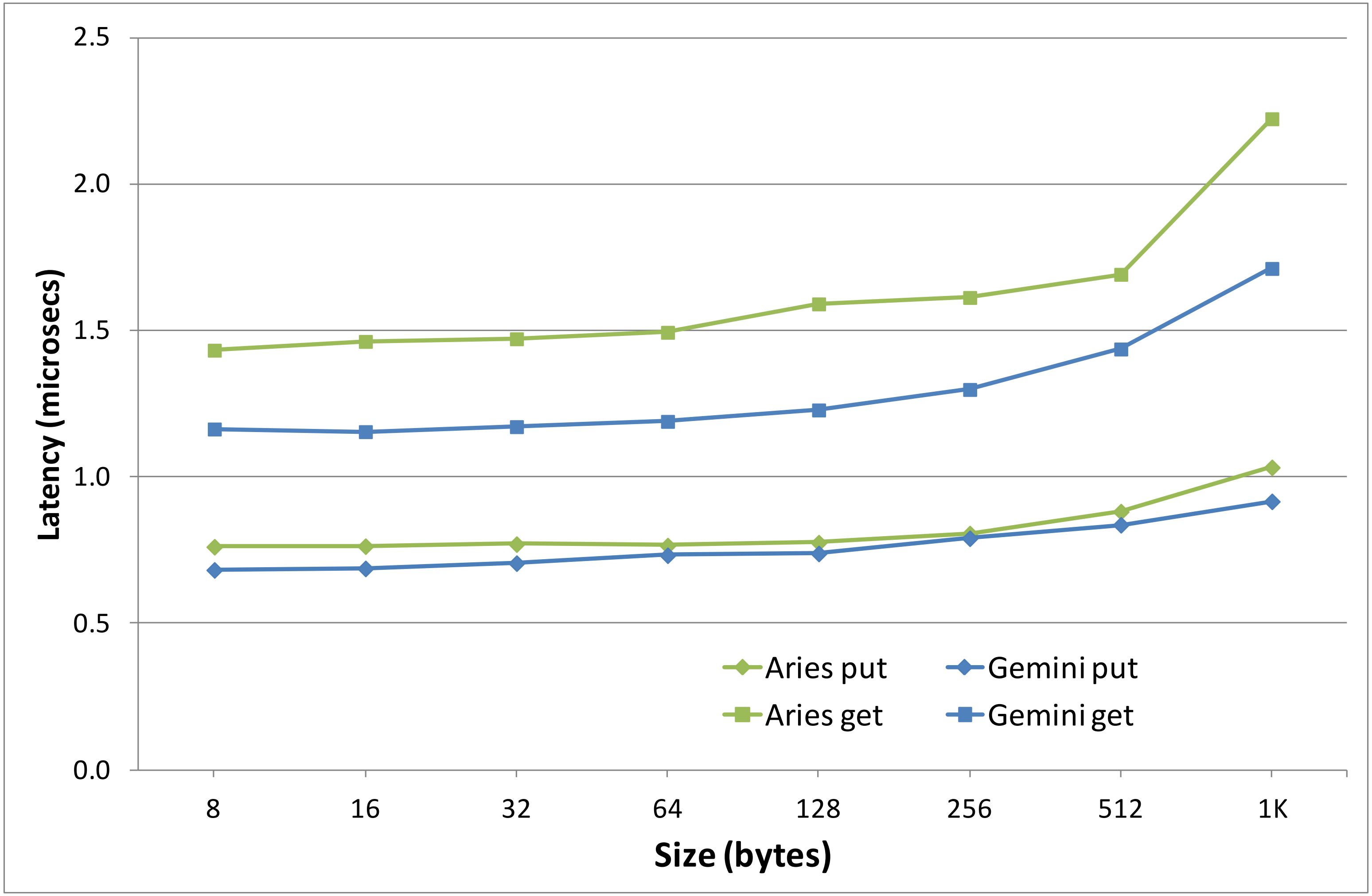
Performance

- **Tests use**
  - DMAPP (as close to the hardware as possible)
- **1-16 processes per node**
- **Early software, not tuned**

Bandwidth Comparison with Gemini



Latency Comparison with Gemini



Summary

- **Aries improves on the successful Gemini ASIC**
  - Improved injection and global bandwidth
  - Improved scaling up to 90k nodes
- **Dragonfly topology has good combination of low latency, scalable bandwidth**
  - Wider class of applications run efficiently for a fixed budget

Acknowledgement

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

Thank you!