The Impact of Optics on HPC System Interconnects

Mike Parker and **Steve Scott**

Hot Interconnects 2009
Manhattan, NYC

CRAY
THE SUPERCOMPUTER COMPANY

Will cost-effective optics fundamentally change the landscape of networking?

Yes.

- Changes the relationship between cost, cable length and signaling speed
- Opens the door to a new class of cost-effective topologies based on high-radix routers

It’s Been a Long Time Coming…

The *Massively Parallel Processing Using Optical Interconnects* conference series was started in 1994.

Some Conclusions from that 1997 Talk

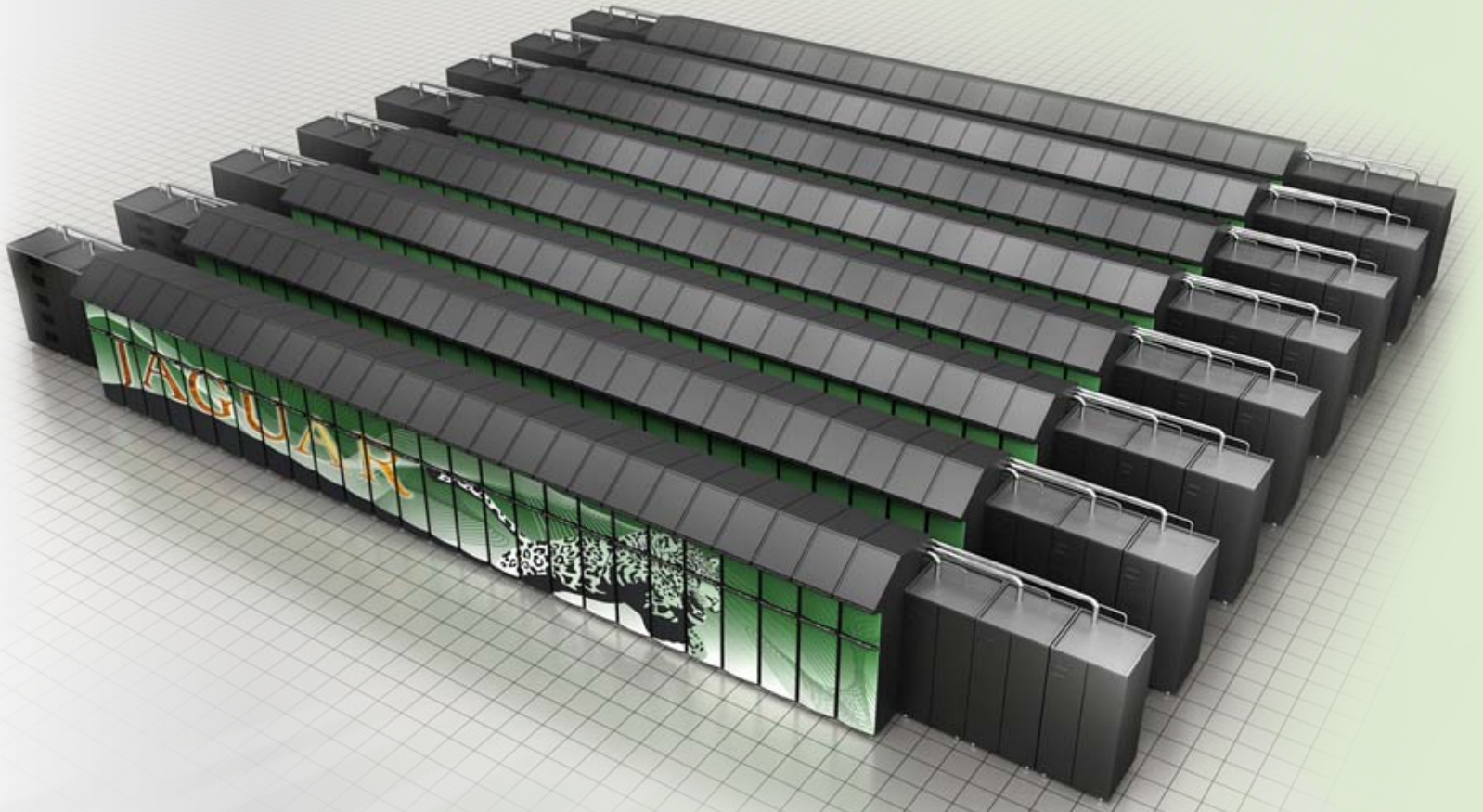
- Optics are really expensive compared to electrical signaling
- Copper's doing just fine for current MPPs and for the foreseeable future
- Optics are useful where you need distance (I/O and networking), and not really anyplace else
- The primary metrics of interest are
  - $/Gbps
  - Gbps per inch of board edge
  - Potential bandwidth off an ASIC

"I'll use optics when it can (without blowing some other metric)
— approach copper on cost (wins on cable bulk, distance, etc.)
— remove a bandwidth bottleneck based on connector density, ASIC I/O, electrical bandwidth ceiling, or (harder to quantify) mechanical feasibility"

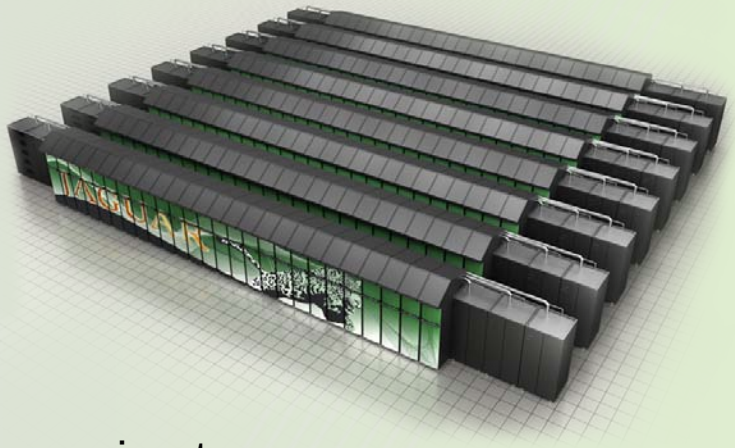
That time is now arriving…

HPC Systems at Cray

- Scalable multiprocessors for running capability scientific/technical apps
  - Thousands to tens of thousands of compute nodes
  - Tens to a few hundred cabinets (racks)

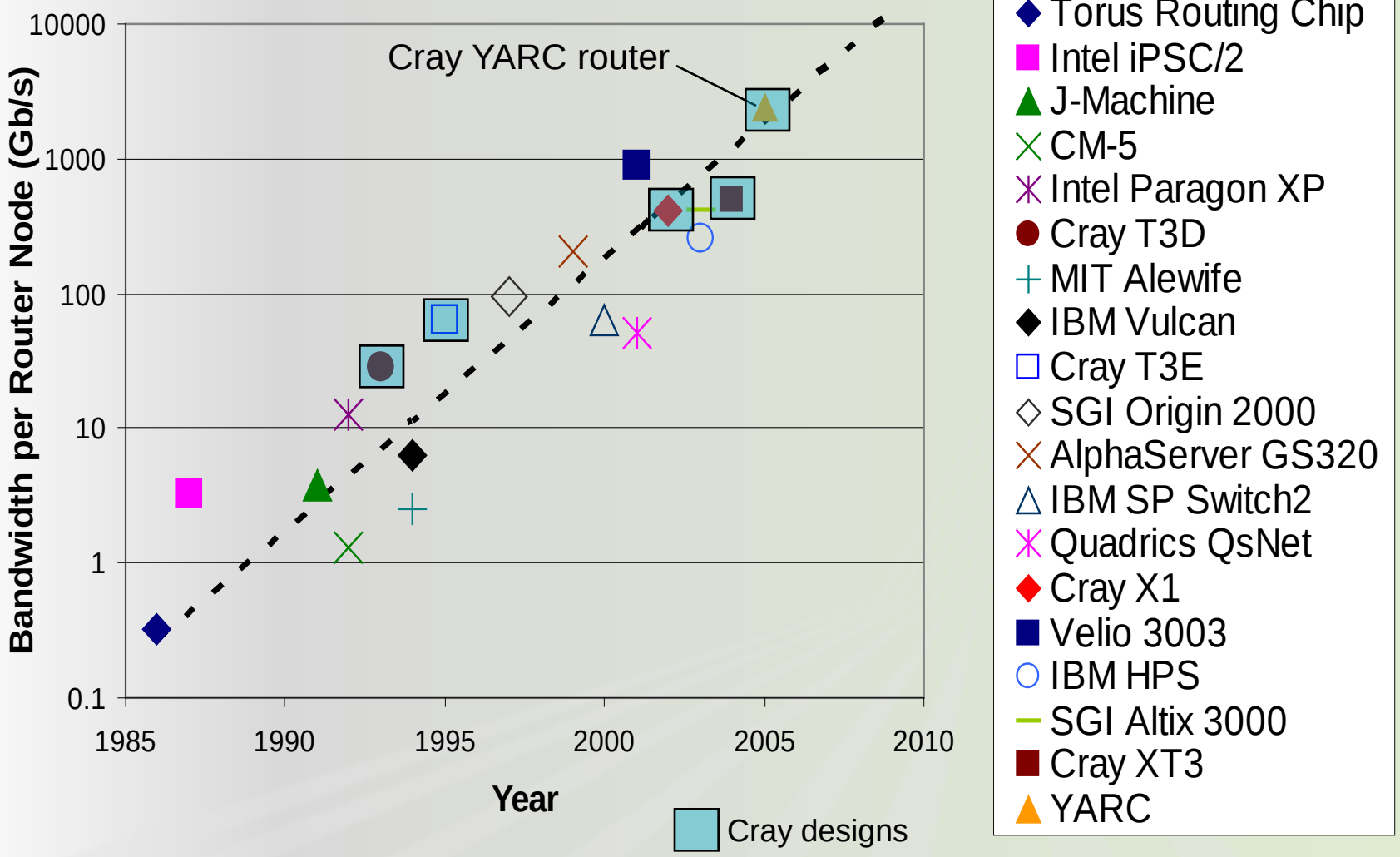


Cray's Interconnect Needs



- Driven by applications
- Communication characteristics
  - Point-to-point traffic
    - Broadcast used only very occasionally
    - Collectives can be performed with *virtual* spanning trees
  - Both message passing and global-address-space applications
    - ⇒ Both bulk data transfer and small packet performance are important
    - ⇒ We care about 53-byte packets
  - Mix of nearest (logical) neighbor, and "long-distance" communication
  - Logical→physical mapping means that communication is rarely really NN
    - ⇒ We focus a lot on *global bandwidth*
- Network performance
  - Per-node bandwidth of O(10 GBytes/s), scalable to large numbers of nodes
  - Latency matters, and is O(1us) across large networks
  - Both performance and price-performance matter
    - Meet your performance goals, at minimal cost, subject to various technology constraints

HPC Router Bandwidth Increasing Over Time



This Motivates High-Radix Routers

- During the past 20 years, the total *bandwidth per router* has increased by nearly four orders of magnitude, while *packet size* has remained roughly constant
  ⇒Changes the optimal router design

- Latency = (# hops)*(T_{hop}) + serialization_time
- Bandwidth/node = (# wires/router)*(signaling rate) / (# hops)
- Cost/bandwidth assuming constant link cost ∝ (# hops)

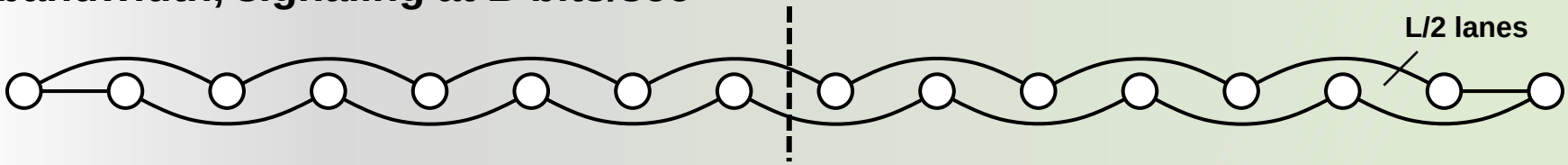
- By increasing the **radix** of the router, both the latency and the cost of bandwidth can be reduced

⇒ Utilize bandwidth by building networks with ***many narrow*** links rather than fewer fat links

  - See Kim, et al, ISCA 2005 for details

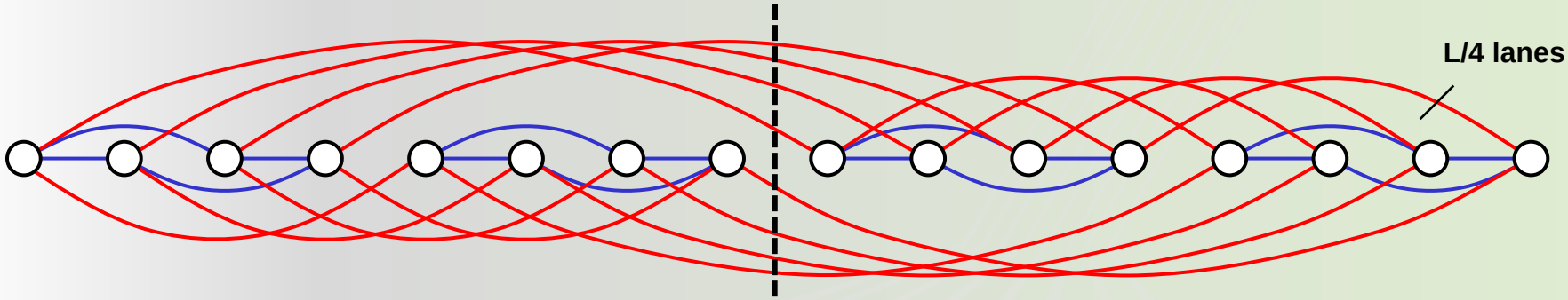
Advantage of Higher Radix

Consider a set of 16 nodes, each with L lanes (one signal in each dir) of pin bandwidth, signaling at B bits/sec



If wired as a radix-16 1D torus with link width L/2 lanes:

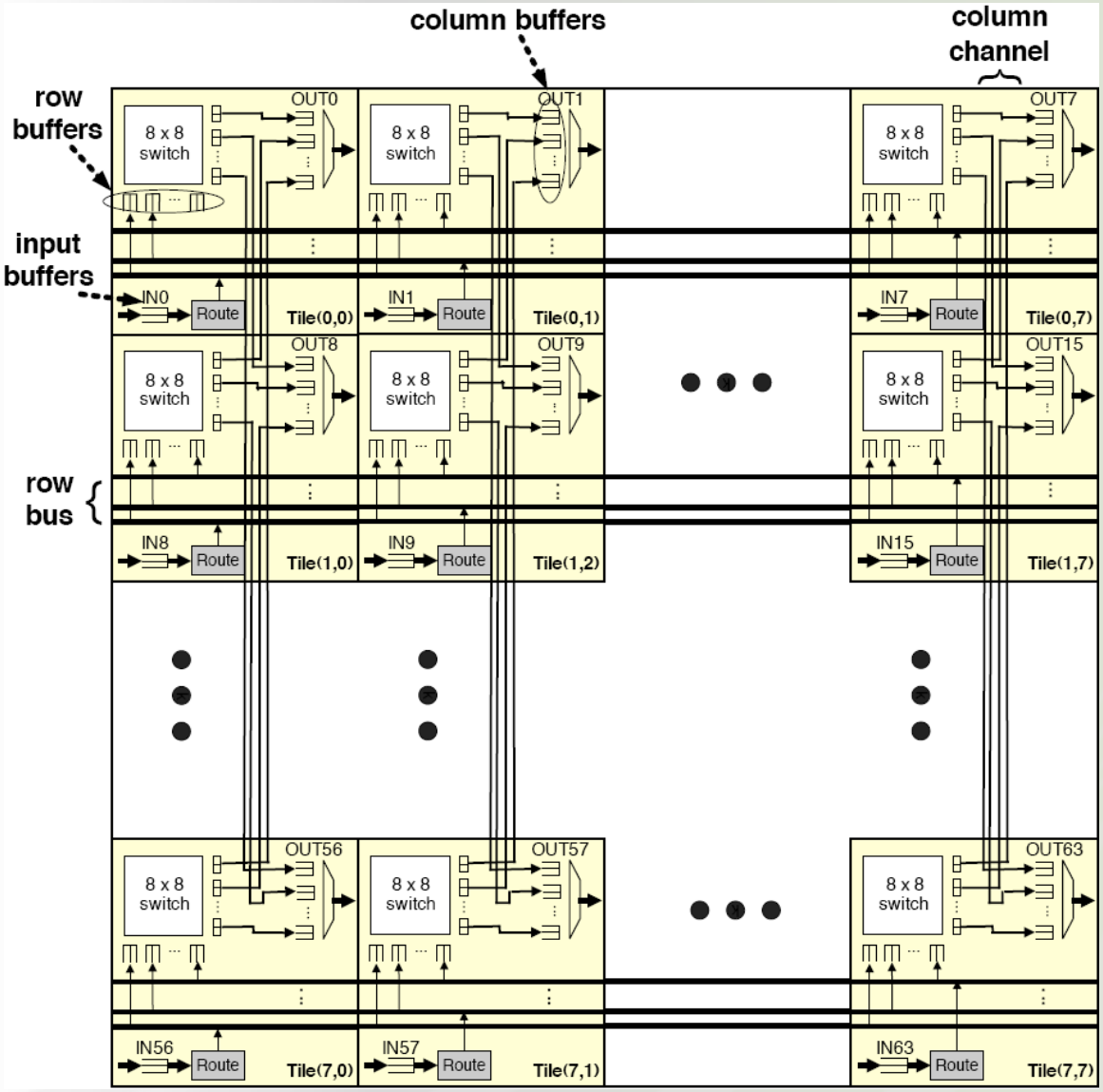
- Bisection bandwidth = 2L*B
- Average distance = 4 hops



If wired as a radix-4 2D torus with link width L/4 lanes:

- Same number of pins per router
- Takes advantage of ability to use longer cables
- **Bisection bandwidth = 4L*B**
- **Average distance = 2 hops**

High Radix Router Microarchitecture

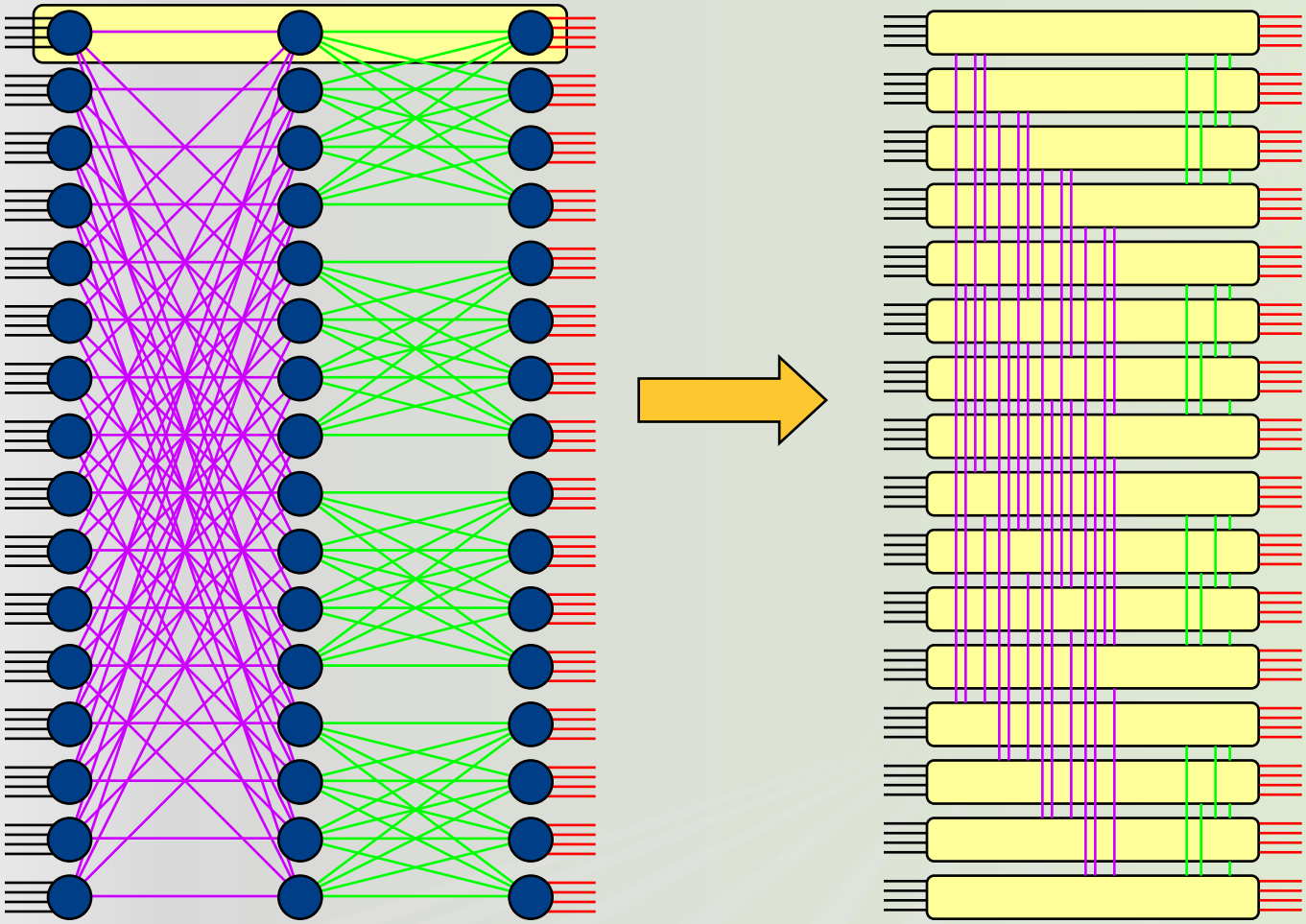


- Regular array of tiles
  - Easy to lay out chip
- No global arbitration
  - *All* decisions local
- Excellent performance
  - Non-blocking
  - Micro-pipelined
  - Internal speedup
- Simple routing
  - Small routing table per tile
  - High routing throughput for small packets
- See Scott, et al, ISCA 2006

Good Network Topologies for High Radix Routers

- Folded Clos (aka fat-tree)
  - Can scale global bandwidth linearly with processor count
  - Can load balance across network
    - Route any permutation conflict free; eliminate hot-spots
  - Low diameter compared to torus and hypercube
  - Many redundant paths (part of a balanced resiliency approach)
- Flattened butterfly
  - Like a butterfly, but all stages collapsed into single router

Flattened Butterfly Topology

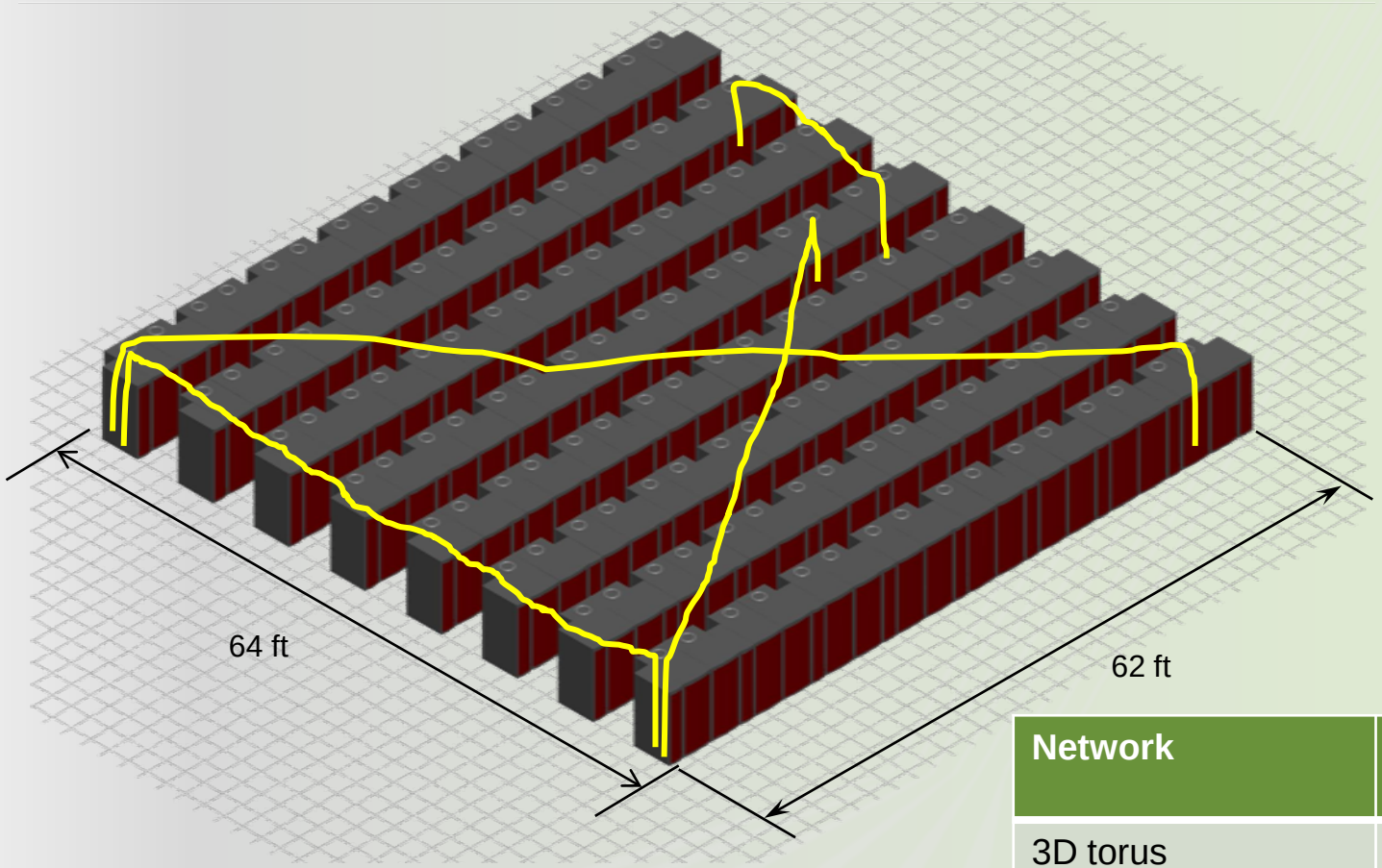


- Collapse multiple ranks of butterfly routers into a single rank of high-radix routers
- Links that had connected ranks of butterfly switches now connect routers within the single rank
- Each of these becomes a separate *dimension* in the flattened butterfly

Good Network Topologies for High Radix Routers

- Folded Clos (aka fat-tree)
  - Can scale global bandwidth linearly with processor count
  - Can load balance across network
    - Route any permutation conflict free; eliminate hot-spots
  - Low diameter compared to torus and hypercube
  - Many redundant paths (part of a balanced resiliency approach)
- Flattened butterfly
  - Like a butterfly, but all stages collapsed into single router
  - Half the global wire utilization of a fat-tree on uniform traffic, equal on worst-case traffic
  - *Requires* high radix router
  - Allows adaptive routing and load balancing (and really needs it)
  - See Dally, et al, ISCA 2007 for details
- Dragonfly
  - A variation on the flattened butterfly
  - Uses extra (inexpensive) local hops to reduce (expensive) global hops
  - Very large systems with only a single optical hop for well-balanced traffic
  - Still allows adaptive routing and global load balancing
  - See Kim, et al, ISCA 2008 for details

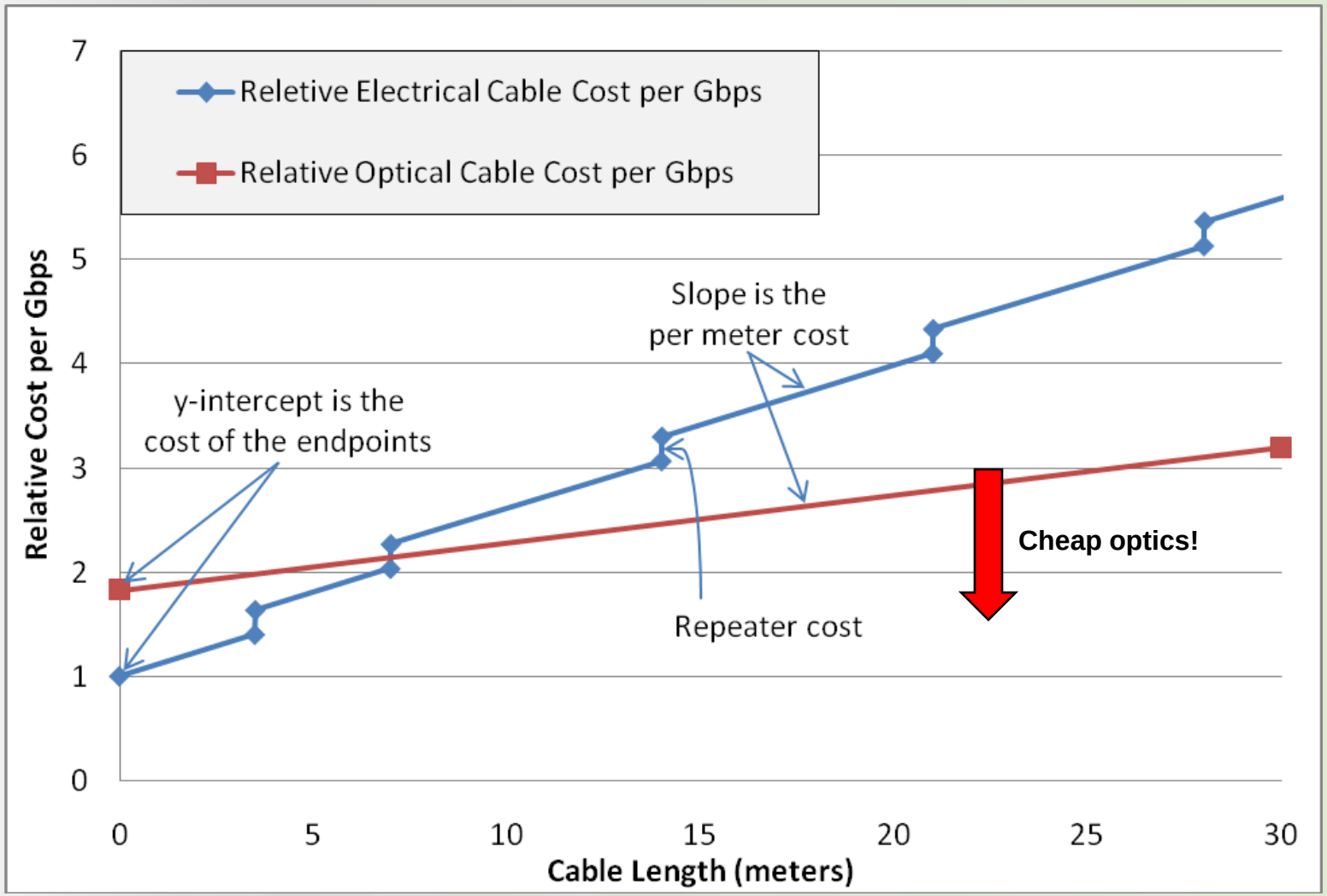
High Radix Networks Require Longer Cables



- Example 128-cabinet system
- Assume 10% slack in cables, and 2m drops inside cabinets

Network	Longest cable (m)
3D torus	7
Flattened Butterfly	25
Folded Clos	25
Dragonfly	34

Cost per Gbps for Optics and Copper



Feedback to Optical Vendors/Researchers

From the perspective of an HPC system vendor

- What really matters:
  - Cost per Gbps (over some physical path)
  - Gbps per inch of board edge
  - Potential bandwidth off an ASIC
    - Integrated silicon photonics could be a big win!
  - Watts per Gbps (this one is new since 1997)
- Matters a bit:
  - Cable bulk and bend radius (already very good here)
  - Hard error rate (component reliability)
- Doesn't matter much:
  - Bandwidth per fiber
  - Bandwidth per cable
  - Transient bit error rate (below 1e-9 or so)
- And really not interested in:
  - Broadcast-based networks
    - Have the expense of listening to everyone, but most traffic is not for you!
  - Anything that requires tuning a receiver (and thus is slow)
  - Optical switching (electrical switching is just fine, thanks)
  - Free-space optics

Summary and a Look to the Future

- Optical links are finally making sense for HPC system interconnects
  - Cray is now designing our first hybrid electrical/optical network
- Performance and price-performance of optical networks is relatively insensitive to distance
  - This enables a new class of high-radix networks with low network diameter
    $\Rightarrow$ *lower latency*
    $\Rightarrow$ *more cost-effective global bandwidth*
- Driving down the cost of optical links will further strengthen this argument

- Expect incremental improvement over the next several years
  - Optical/electrical price-performance crossover distance will continue to drop
  - Shorter cables…. backplane…. on-board….. on module… on chip (?)
- Next big advancement will be integrated optics directly off package (die?)
  - Has the potential to provide a major increase in off-chip bendwidth, and significantly reduce signaling power

Thank You.

Questions?