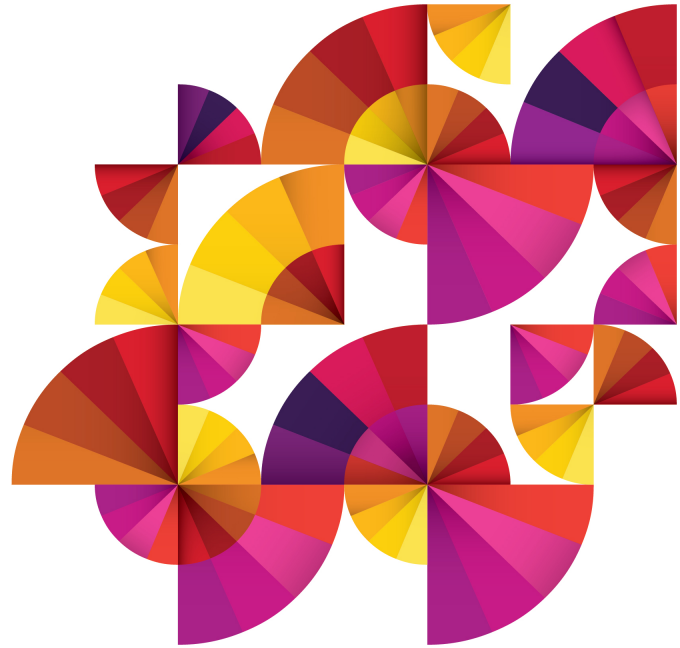


# Network challenges and directions for the Exascale era

ExaComm workshop, ISC '18  
28 June 2018



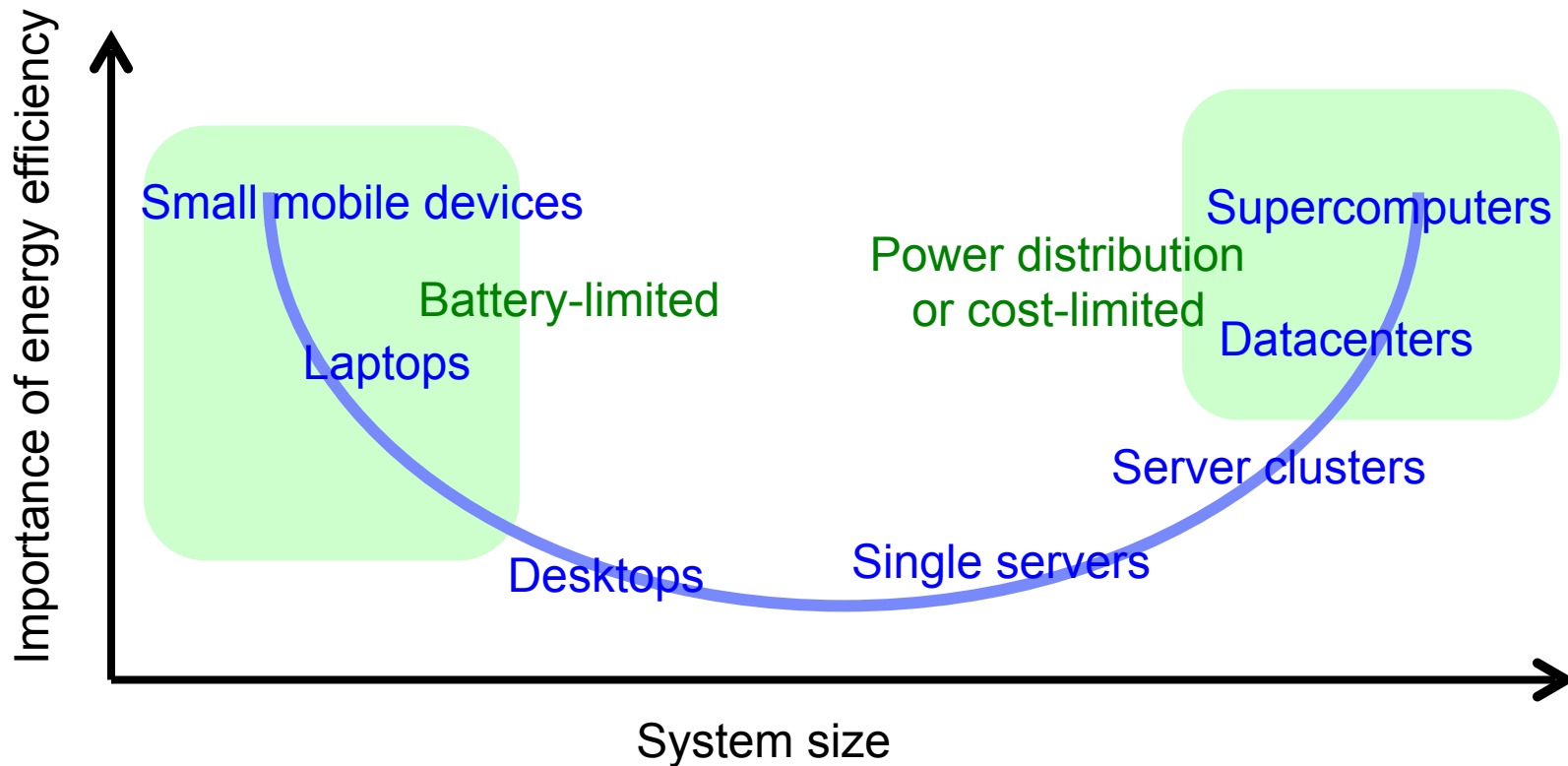
- Cost
- Energy
- Scalable performance
- Reliability

- Message rate, efficiency
- All-to-all performance
- Low latency
- Offload

- Large HPC customers (e.g., U.S. national labs) have said they desire 0.1 byte/flop or op
  - Total of ingress and egress bandwidth (0.05 + 0.05)
  - Although “fat” nodes make lower ratios acceptable
    - Surface to volume ratio decreases with memory size and computational capability
- For an exa-op system, this would be (50 + 50) PB/s total bandwidth to/from endpoints
- Is nirvana achievable?

- 15% rule
  - Supercomputer customers are typically willing to pay up to 15% of the system cost for the interconnection network
- Assume \$400M for an exaflop system: \$60M max for the interconnection network
- Assume highly-scalable, bandwidth- and cost-efficient topology:
  - Dragonfly - 2.5 bidirectional links/cables per endpoint
  - In the best case, only 0.5 links @ 50 Gb/s signaling per endpoint are optical
  - But this will increase to 1.5 links @ 100 Gb/s signaling
- Assume all network cost is from optical links (a very optimistic assumption)
  - Requires optical link cost  $\leq 15\text{¢/Gb/s}$  for 50 Gb/s signaling
  - An  $\leq 5\text{¢/Gb/s}$  for 100Gbps signaling
  - But today we are  $> \$1/\text{Gb/s}$ , resulting in \$400M for optics alone

# HPC network grand challenge #2: Energy



- Example: U.S. Dept. of Energy desires  $\leq 30$  MW for an exascale system
  - Assume 15% allocation of energy to the network: 4.5 MW
  
- Again assume a highly-scalable, bandwidth- and cost-efficient topology:
  - Dragonfly - 2.5 links per endpoint
  - In the best case, only 0.5 links @ 50 Gb/s signaling per endpoint are optical
  - But this will increase to 1.5 links @ 100 Gb/s signaling
  
- Assume switches and electrical links consume half of network power
  - It is actually much more than this today for cluster networks
  - Assume other half of energy from the optical links (2.25 MW)
  - Requires 5.6 pJ/bit for the optical links, with 50 Gb/s signaling
  - Requires 1.9 pJ/bit for optical links, with 100 Gb/s signaling
  - Difficult to achieve, but cost is a much bigger challenge

- Can the interconnect scale performance linearly?
  - Are there limits or inflection points to interconnect scaling?
- Can the system scale incrementally?
- Can the system be partitioned well?
  - Job isolation, QOS, jitter reduction
- Can messaging software scale to millions of endpoints?



- Ethernet commonality and in general, applicability to Cloud
- Higher switch radix and move to bandwidth-scalable topologies
- Increasing emphasis on offload
- Increasing level of optics integration

- Increasing commonality with Ethernet
- PHY and I/O macro convergence
  - Signaling rate
  - SERDES, training
  - Error detection and correction techniques
- Main motivation for interface convergence is cost savings
  - Signaling at 50 Gb/s and higher is a huge challenge
  - Shared design, verification, fabrication, and testing costs

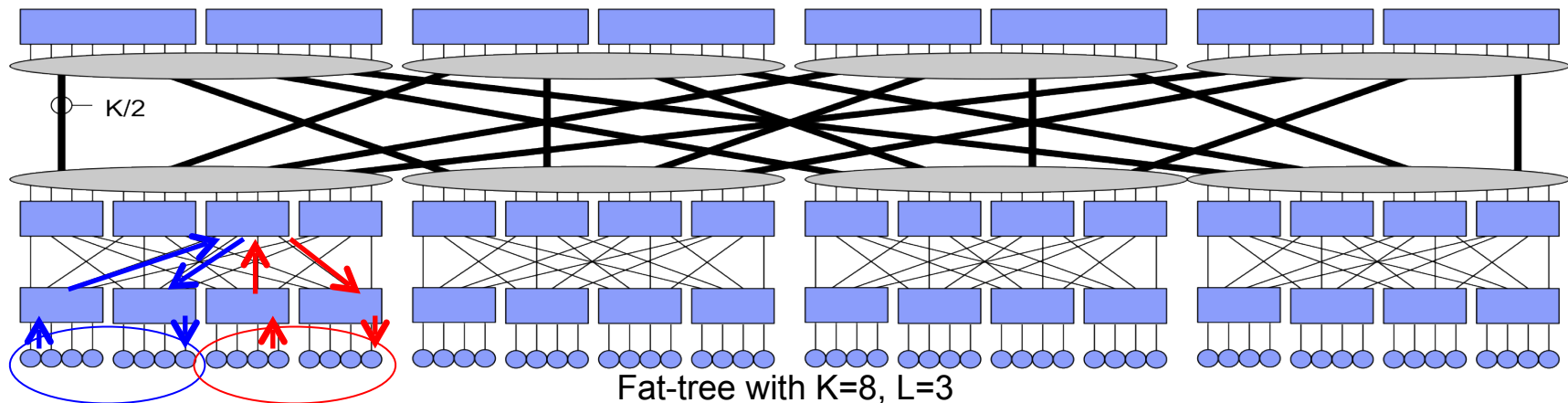
- The chosen Ethernet path for 50+ Gb/s signaling leverages PAM4 (Pulse Amplitude Modulation with 4 signal levels)
  - Doubles the signaling rate for the same Baud rate as NRZ (Non-Return to Zero)
- However, "eyes" are much narrower, and the impact of noise is relatively more pronounced
- Will typically require Forward Error Correction (FECs) to achieve an acceptable Bit Error Rate (BER)
  - Error detection and link-level retry are insufficient
- Increases latency on every hop, for checking and correction: ?? ns

- Desire bandwidth-scalable topologies
- Take advantage of trend towards high-radix switches to flatten network
  - Fewer hops = reduced cost, energy, and latency
  - Disadvantages tori and similar nearest neighbor topologies
- Topology choice heavily influenced by costs and available technologies
  - Optics vs. electrical tradeoffs
  - Ratio of electrical links is decreasing with increasing signaling rate
  - For  $\geq 100$  Gb/s signaling, almost all links may be optical
    - Link and switch counts then become a good determiner of relative cost

# Topology options: Fat-tree



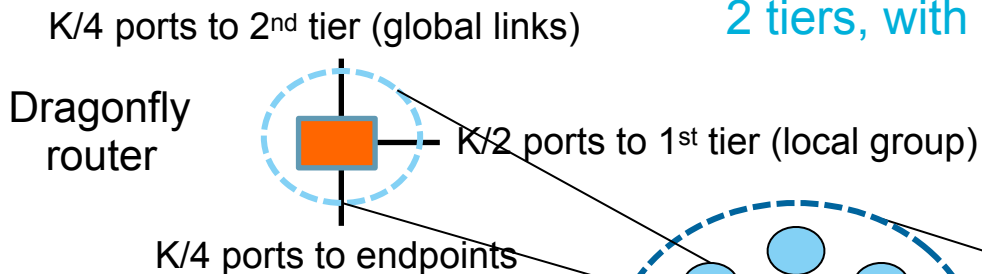
- Assume K-port switches, and an L-level fat-tree
- Scales to  $N = 2(K/2)^L$  endpoints =  $K^3/4$  for a 3-level Fat-tree
- Switches traversed =  $2L - 1$  (5 switches for a 3-level fat-tree)
- Links per endpoint =  $L$  (3 links for a 3-level fat-tree)
- Switches per endpoint (full tree) =  $(2L - 1)/K$
- Bisection bandwidth =  $BN$ , where  $B$  is the unidirectional link bandwidth
- Partitions: integer multiples of sub-trees with the same “parents”
- Easily accommodates tapering for reduced cost at reduced bandwidth



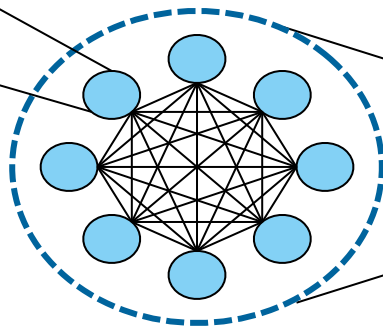
# Topology options: 2-tier Dragonfly



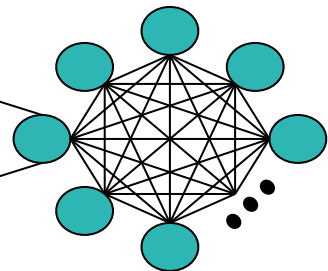
K-port Dragonfly router:  
2 tiers, with each tier fully connected



$K/4$  ports to endpoints



Local group  
(1<sup>st</sup>-tier connections)



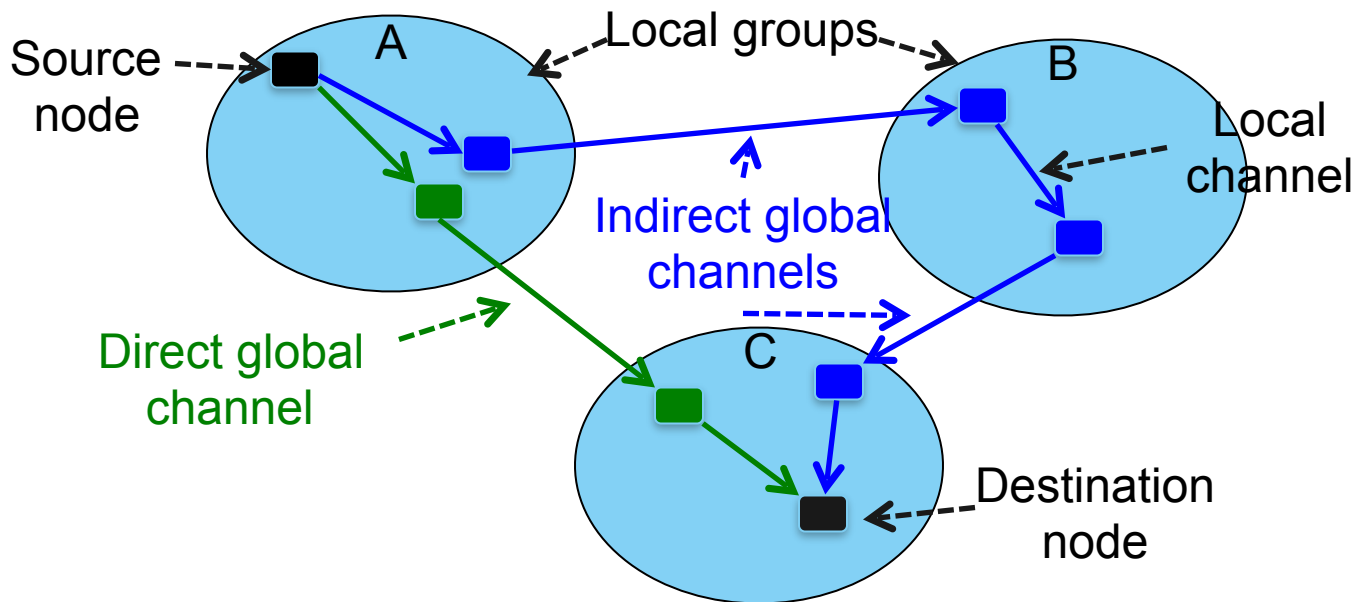
Full system (local groups  
connected via global links)

- Scalable to  $K^4/64$  endpoints
- 4 or 6 router/switch traversals
- 3 virtual channels per class
- 2.5 links per endpoint
- $4/K$  switches per endpoint
- Bisection bandwidth scales as  $BN/2$  (half of Fat-tree)
- Global bandwidth comparable to Fat-tree
- Non-interfering partition sizes only up to a full local group

# Dragonfly routing: direct and indirect paths



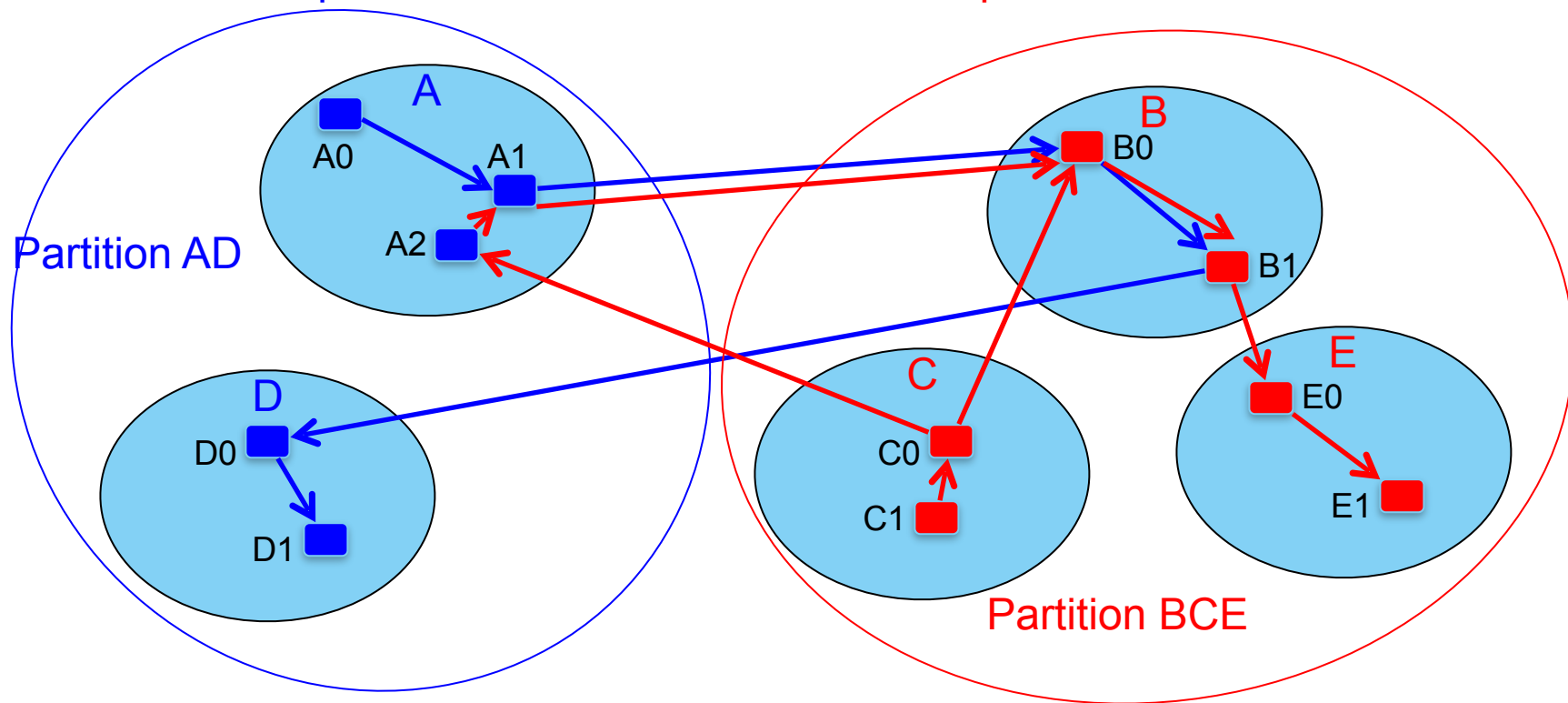
Direct path shown in green, indirect path shown in blue



# Dragonfly routing: partitions and indirect paths



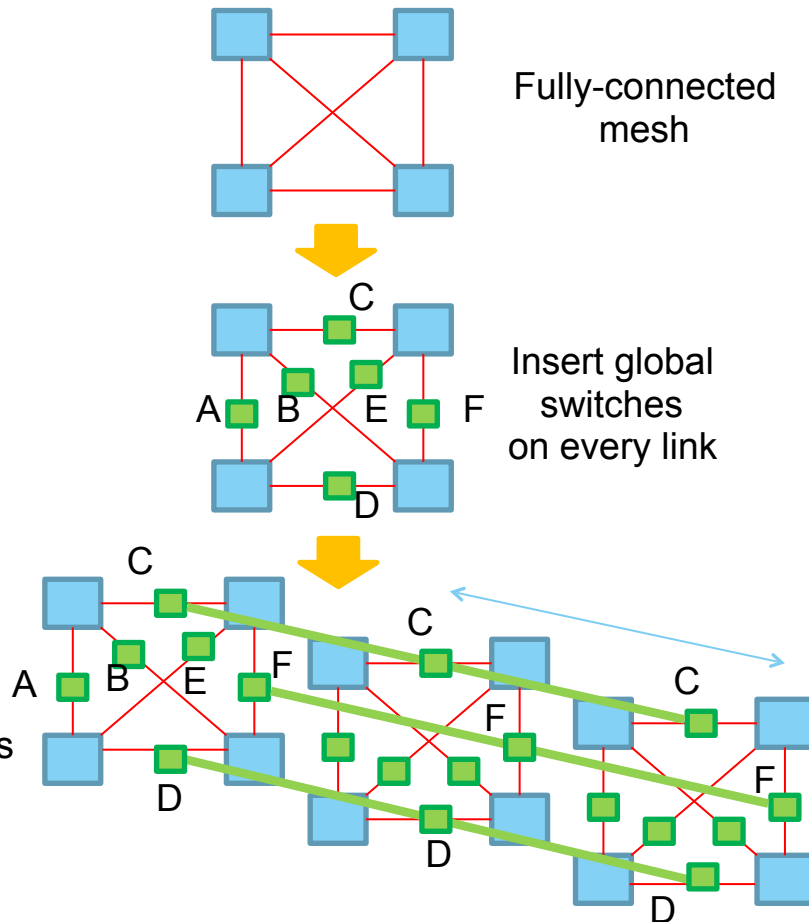
Partition AD path shown in blue, Partition BCE path shown in blue







# Topology options: stacked full mesh

- Simultaneously discovered by IBM and Fujitsu
- Names “multi-layer full mesh” by Fujitsu
- Scalable to  $K^3/8$  endpoints
- 3 or 5 switch traversals
- 2 virtual channels per class
- 2 links per endpoint
- $\sim 3/K$  switches per endpoint
- Bisection bandwidth scales as  $\sim BN/2$  (half of Fat-tree)
- Global bandwidth comparable to Fat-tree
- Many isolated partition sizes possible



 = Global switch  
 = local (TOR) switch connecting endpoints

# Topology options: stacked full mesh

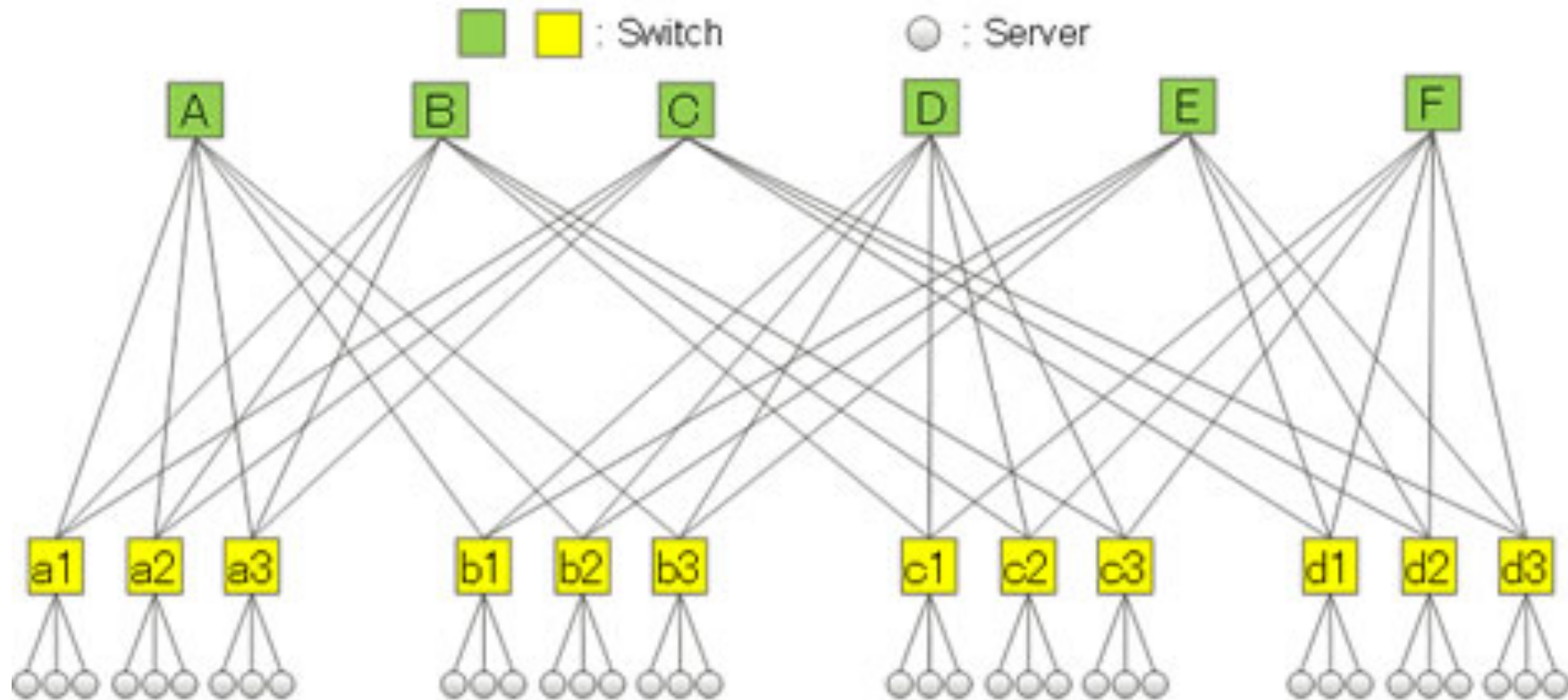


Figure from Fujitsu: <http://www.fujitsu.com/global/about/resources/news/press-releases/2014/0715-02.html>

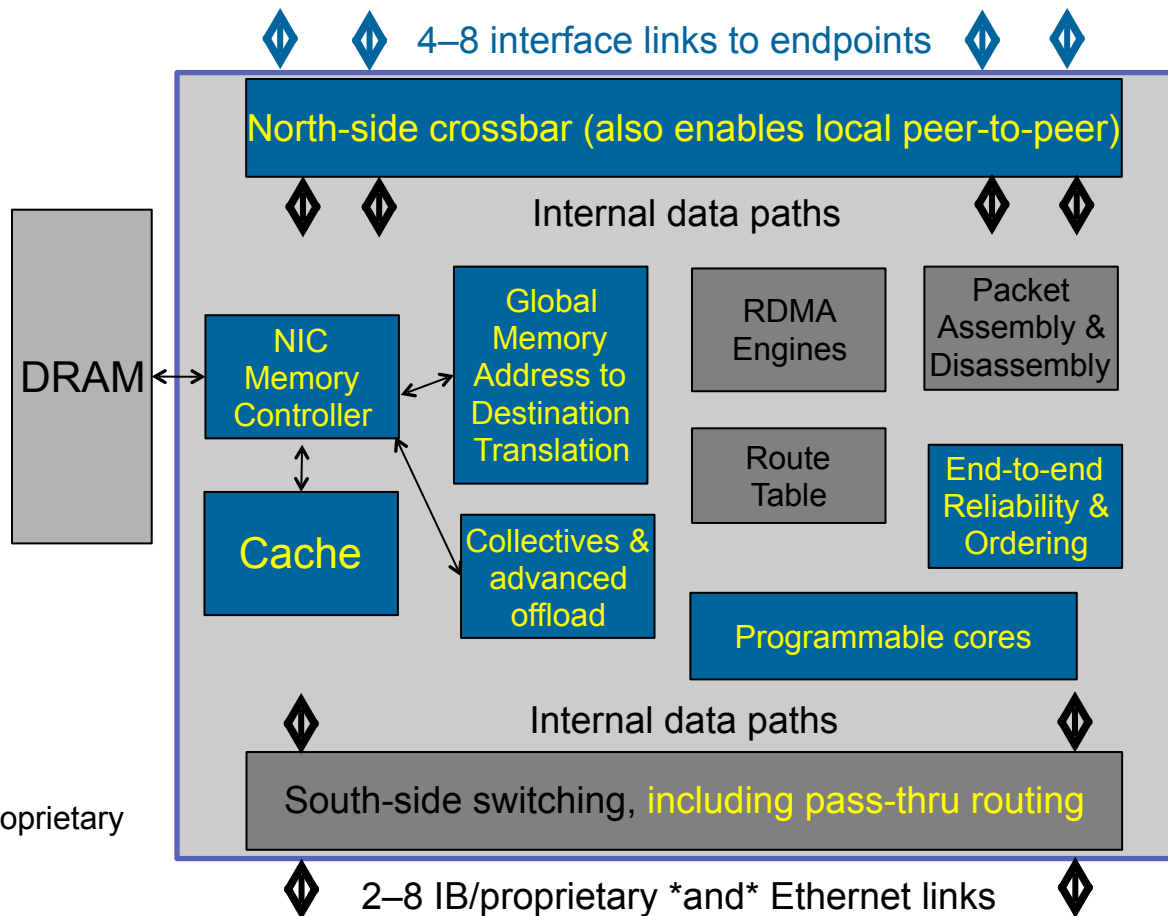
# Topology comparison table



Architecture	Approximate Max Scale K=36, 48, 64	Links Per Endpoint	Switch Ports Per Endpoint	Switch Traversals Direct, Worst Case Indirect	Range of partition sizes	Virtual Channels per Traffic Class
2-level Fat-tree	648, 1152, 2048	2	3	3, NA	Good	1
3-level Fat-tree	11664, 27648, 64K	3	5	5, NA	Good	1
4-level Fat-tree	205K, 648K, 2M	4	7	7, NA	Good	1
Stacked full mesh	5184, 12288, 29127	2	3	3, 5	Medium	2
2-tier Dragonfly	26244, 82994, 256K	2.5	4	4, 6	Only within local group	3

- For fat-trees, 3 levels is the sweet spot balancing scale and complexity
- Stacked full mesh attractive within its scale (about half that of a 3-level fat-tree)

# Offload: “SuperNIC” architecture



- Dual protocol:
1. InfiniBand or proprietary
  2. Ethernet

- Many interesting opportunities for processor offload
  - Active messages/transactions, including remote atomic operations
  - Complex collectives
  - Efficient gather/scatter
  - Message completion handling
  - Message aggregation
  - Send/receive messages without host processor involvement
    - Direct protocol hand-off to GPUs, other accelerators, smart storage, etc.

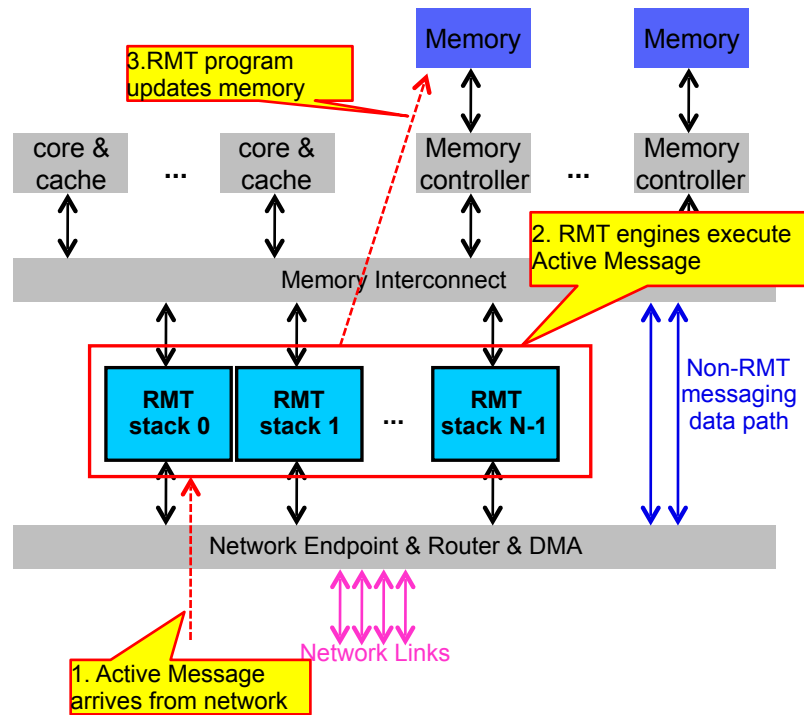
- Programmable/configurable engine advantages:
  - More flexible function
  - More robust if there are design errors or oversights
  - Can support many functions with one unit
- Hardwired:
  - For the given function: more efficient, higher performance
- And FPGAs are in the middle of this spectrum

- In (or attached to) the NIC?
  - Minimizes dependence on host architecture
  - Can upgrade network independently
- Closer to the node memory?
  - Low-latency, high-bandwidth access to host memory
  - Efficient packed gather/scatter packet transfer over NIC-host bus
- Support in switches
  - Collective support provides clear advantages in latency
  - And can provide bandwidth advantages, depending on the implementation

# Remote Memory Transaction (RMT) investigation



- RMT request = Active Message
  - Initiates program execution on receiver node
  - Updates remote memory user data
- RMT engines in/near network interface
  - Many programmable engines
  - Tiny and power efficient
  - Optimized for data movement
- Network topology agnostic
- Near memory with low-latency, high-bandwidth access to entire address space





# Co-packaging: Changing Approach for Building Switches



## Today

Switch Chip



- Bandwidth limited by pin density at package / board interface
- Large energy costs for driving > 10 cm transmission lines
- Switch chip area constraints

## Next Step



- Bandwidth limited by pin density at chip / package interface
- Some energy still spent on few cm long electrical transmission lines
- More silicon area for switching function = more ports

## Future Solution



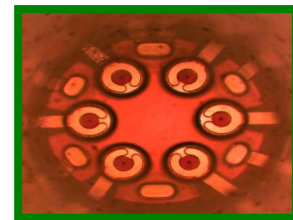
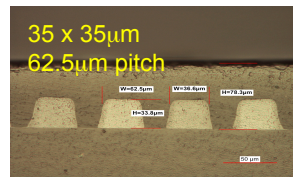
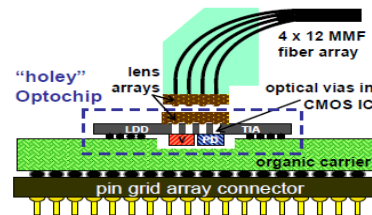
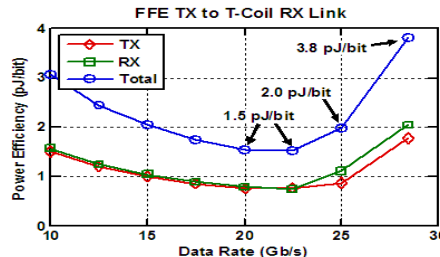
- Bandwidth limited by spectral efficiency
- Energy spent on steering pipes rather than processing/transmitting bits

- Avoid distortion, power, & cost of ASIC-interfacing electrical links
- Move beyond chip & module pin-count limits

# Path to increased network bandwidth per link

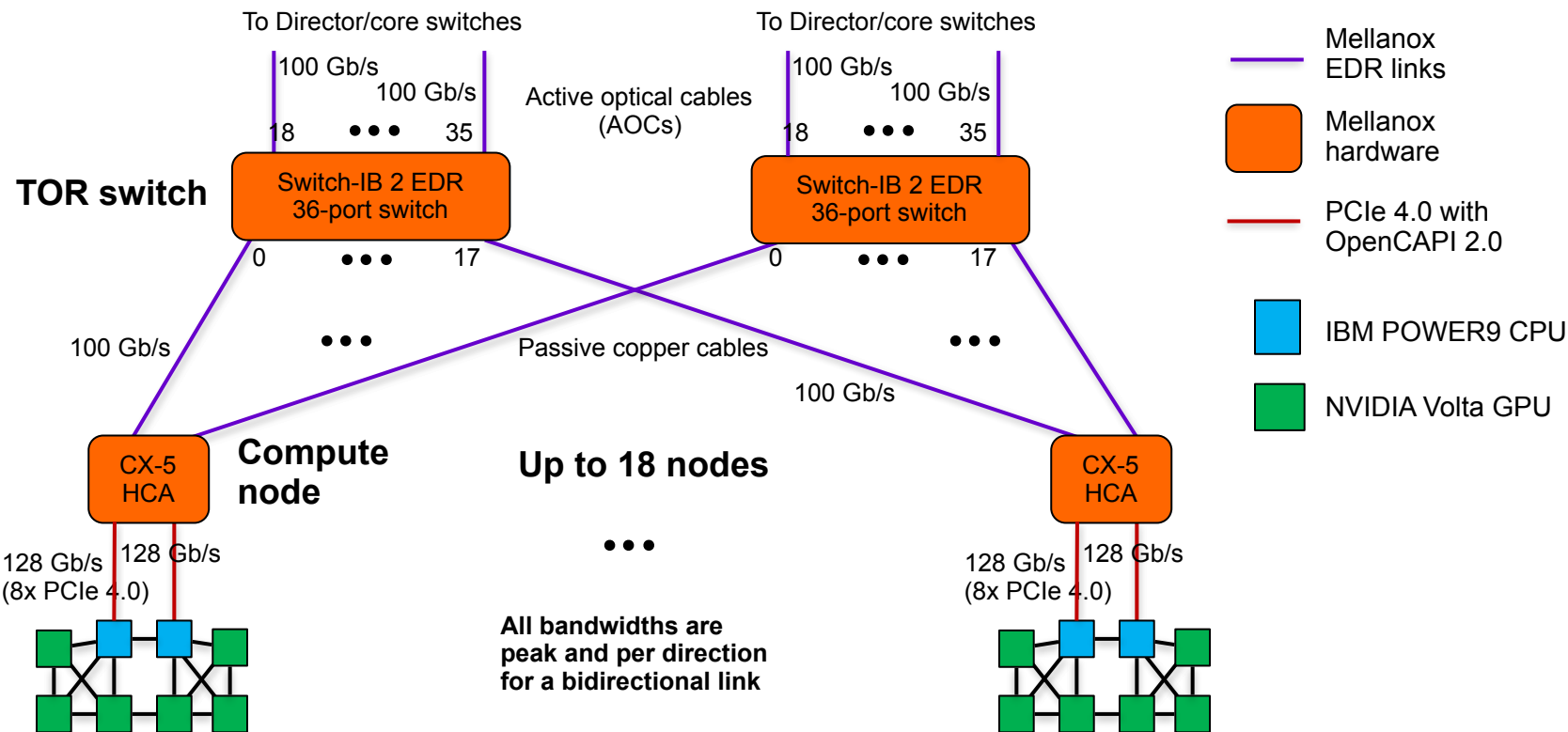


- Data Rate
  - Ultimately limited by power
  - Can be mitigated by tight packaging
- VCSELs vs. Silicon Photonics
- Number of physical lanes
  - Increase number of fibers
  - Closely packed optical waveguides increases density
  - Multicore fiber can reduce fiber count by 4x or more
- PAM4
  - 4 signaling levels
  - Doubles signaling rate compared to NRZ
- WDM
  - CWDM with multimode possible for ~2-4 wavelengths
  - Si Photonics for >4 wavelengths (also multi-km distance)

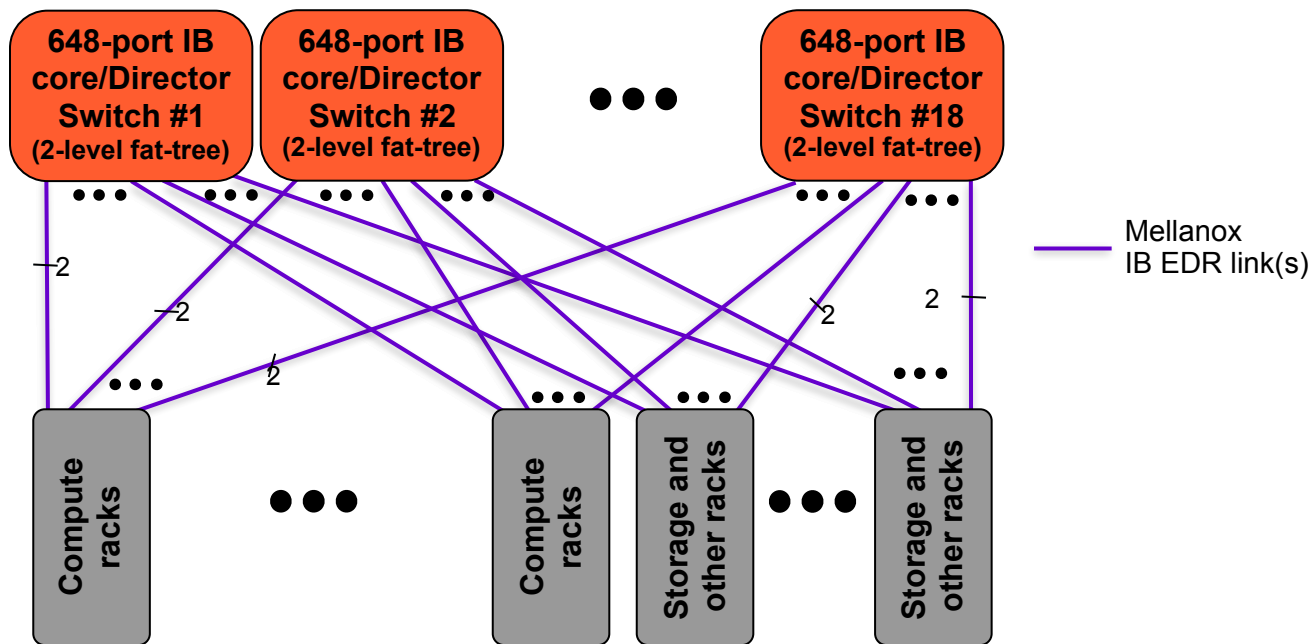


- Networks are aggressively targeting future HPC and Analytics challenges
- Scalable performance
  - Technology, topologies, messaging software
- Cost: public enemy number one
  - Technology & topologies
  - Leverage commodity when possible, exploit commonality with ethernet
- Low latency, high messaging rate, offload, collectives
  - Holistic, end-to-end design philosophy with codesigned messaging stack
  - Overlap communication with computation
  - Hardwired and programmable support in NICs, switches, and near-memory
  - Location matters: compute near data to minimize data movement

# CORAL: Summit compute rack InfiniBand components



# CORAL: complete InfiniBand EDR network



Thanks!

Danke schön!

