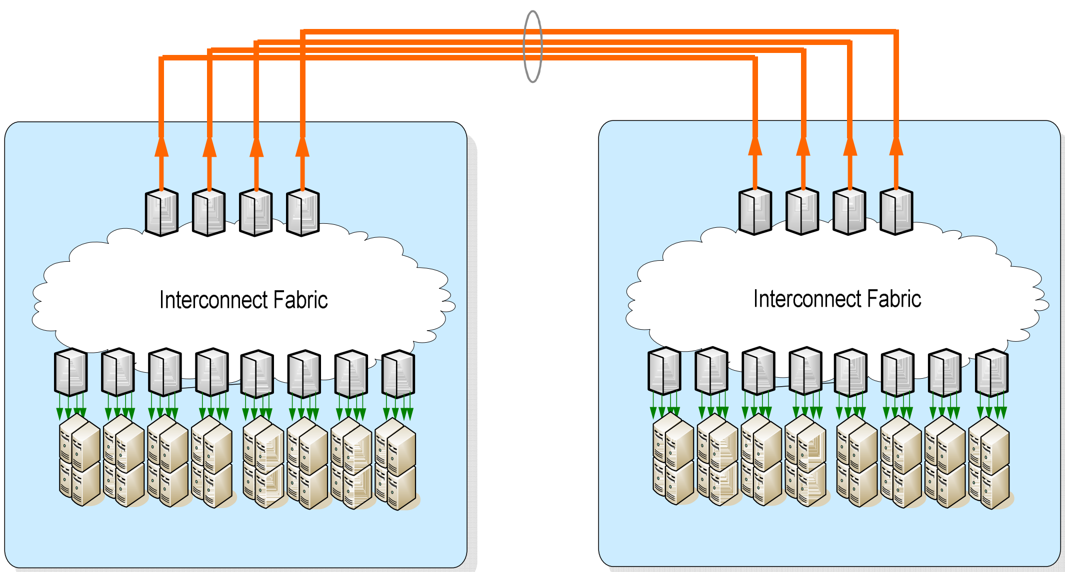
100GbE in Datacenter Interconnects: When, Where?

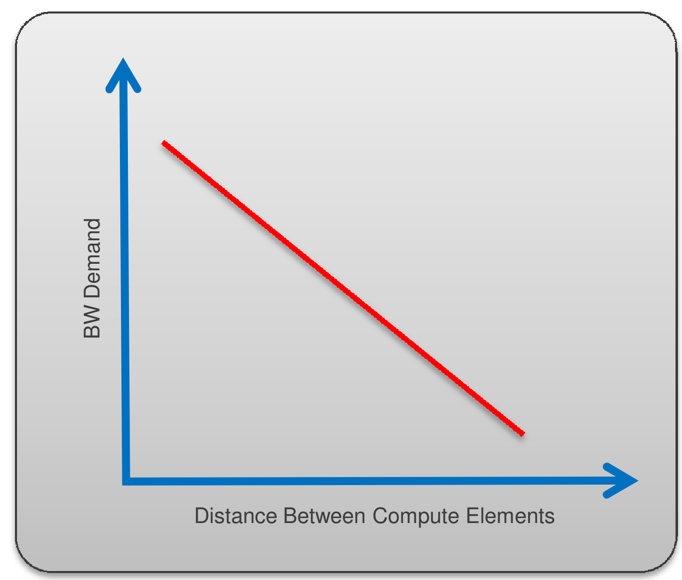
Bikash Koley
Network Architecture, Google

Sep 2009

Datacenter Interconnects

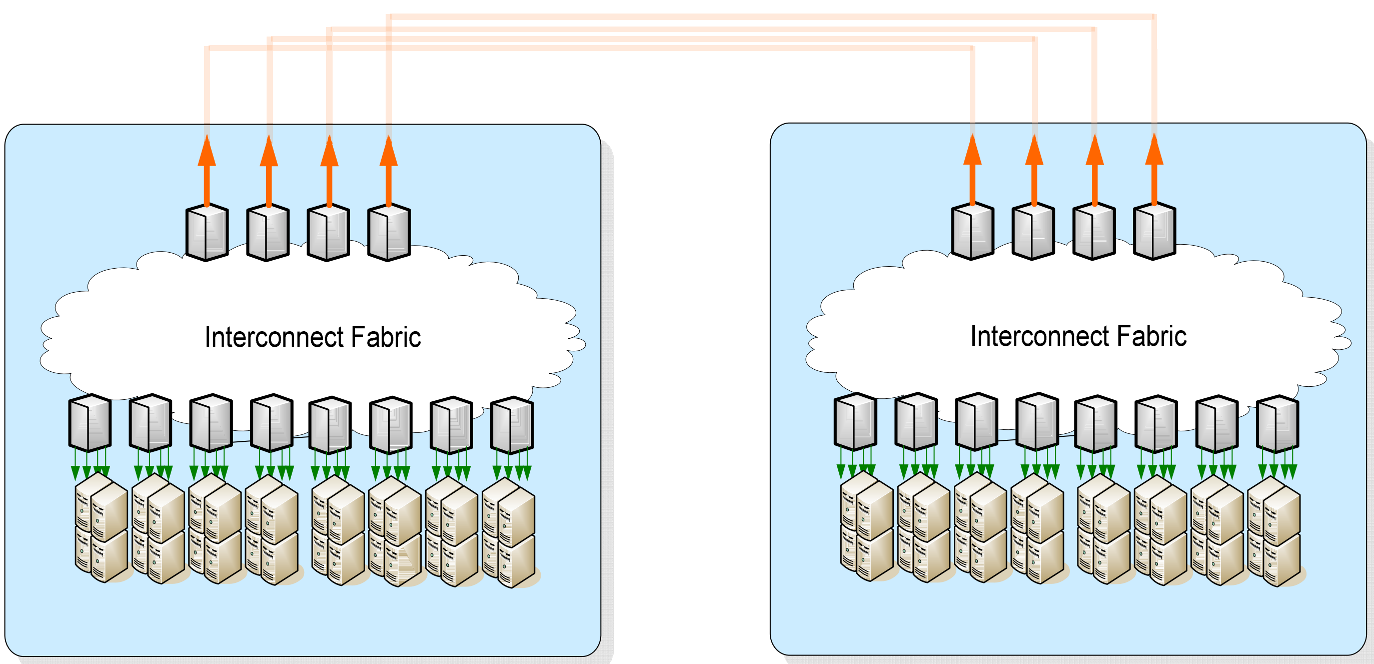
- Large number of identical compute systems
- Interconnected by a large number of identical switching gears
- Can be within single physical boundary or can span several physical boundaries
- Interconnect length varies between few meters to tens of kms
- Current best practice: rack switches with oversubscribed uplinks





- **INTRA-DATACENTER CONNECTIONS**
- INTER-DATACENTER CONNECTIONS

Fiber-rich, Very large BW demand

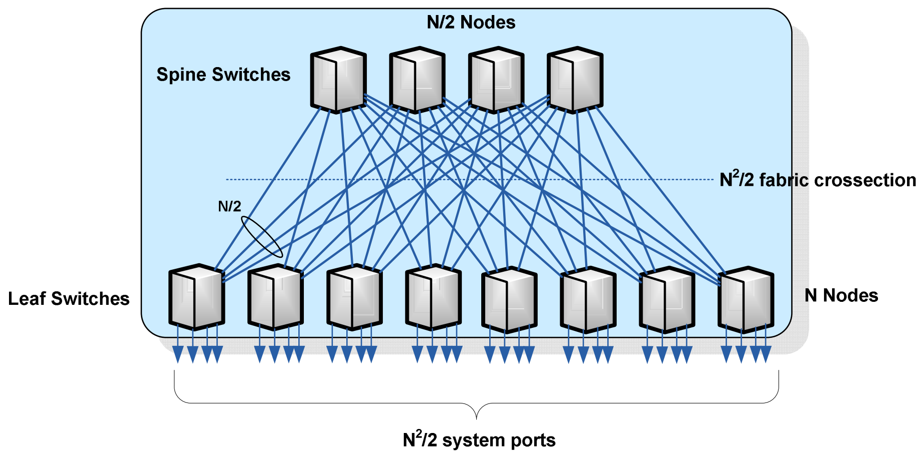


Datacenter Interconnect Fabrics

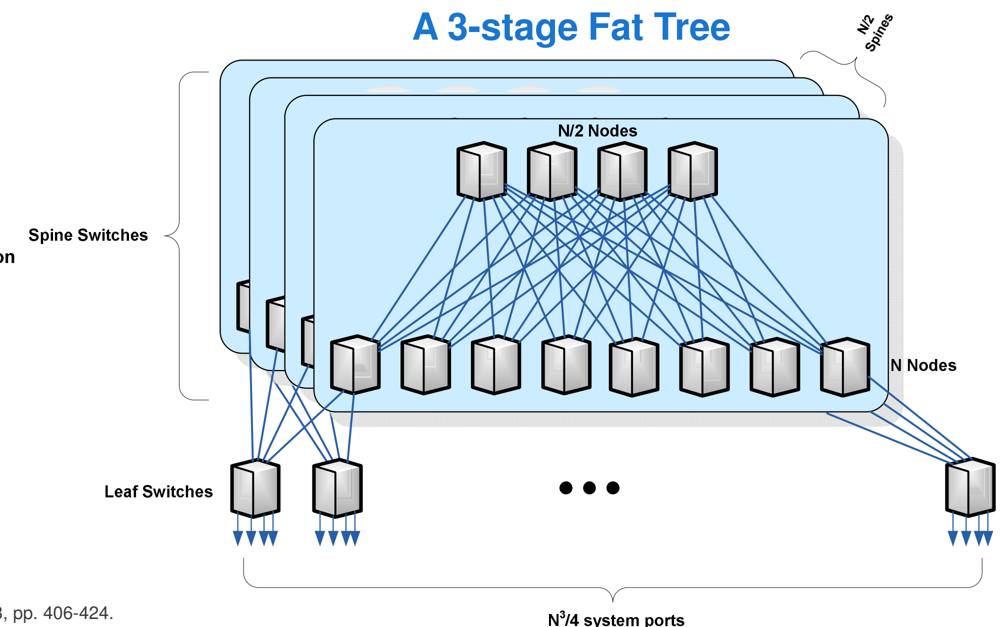
Google

- High performance computing/ super-computing architectures have often used various complex multi-stage fabric architectures such as Clos Fabric, Fat Tree or Torus [1, 2, 3, 4, 5]
- For this theoretical study, we picked the Fat Tree architecture described in [2, 3], and analyzed the impact of choice of interconnect speed and technology on overall interconnect cost
- As described in [2,3], Fat-tree fabrics are built with identical N-port switching elements
- Such a switch fabric architecture delivers a constant bisectional bandwidth (CBB)

A 2-stage Fat Tree



A 3-stage Fat Tree



1. C. Clos, *A study of non-blocking switching networks*, Bell System Technical Journal, Vol. 32, 1953, pp. 406-424.
2. Charles E. Leiserson: “Fat-Trees: Universal Networks for Hardware-Efficient Supercomputing.”, IEEE Transactions on Computers, Vol 34, October 1985, pp 892-901
3. S. R. Ohring, M. Ibel, S. K. Das, M. J. Kumar, “On Generalized Fat-tree,” IEEE IPPS 1995.
4 RUFT: Simplifying the Fat-Tree Topology, Gomez, C.; Gilabert, F.; Gomez, M.E.; Lopez, P.; Duato, J.; Parallel and Distributed Systems, 2008. ICPADS '08. 14th IEEE International Conference on, 8-10 Dec. 2008 Page(s):153 – 160
5. [Beowulf] torus versus (fat) tree topologies: http://www.beowulf.org/archive/2004-November/011114.html

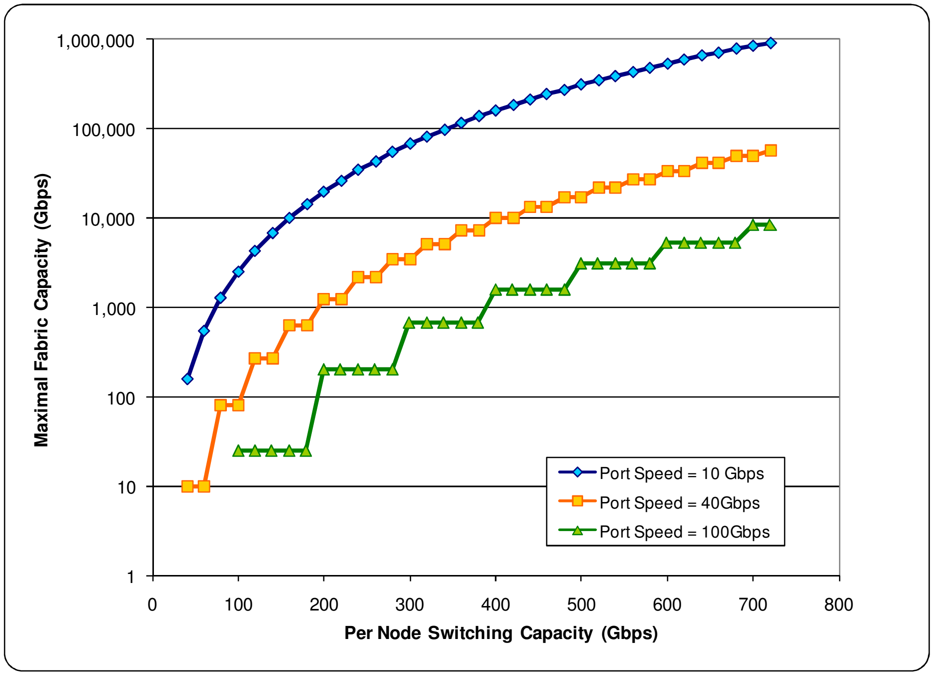
Interconnect at What Port Speed?

- A switching node has a fixed switching capacity (i.e. CMOS gate-count) within the same space and power envelope

- Per node switching capacity can be presented at different port-speed:
  - i.e. a 400Gbps node can be 40X10Gbps or 10X40Gbps or 4X100Gbps

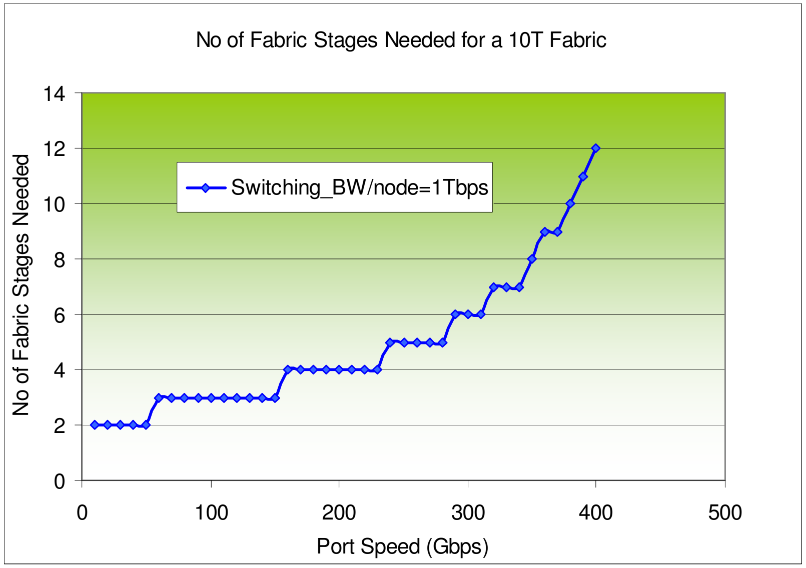
- Lower per-port speed allows building a much larger size maximal constant bisectional bandwidth fabric
- There are of course trade-offs with the number of fiber-connections needed to build the interconnect
- Higher port-speed may allow better utilization of the fabric capacity

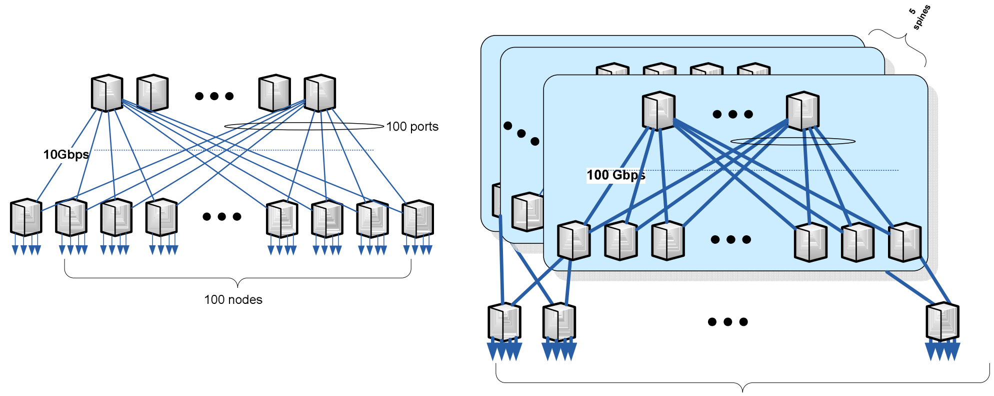
3-stage Fat-tree Fabric Capacity

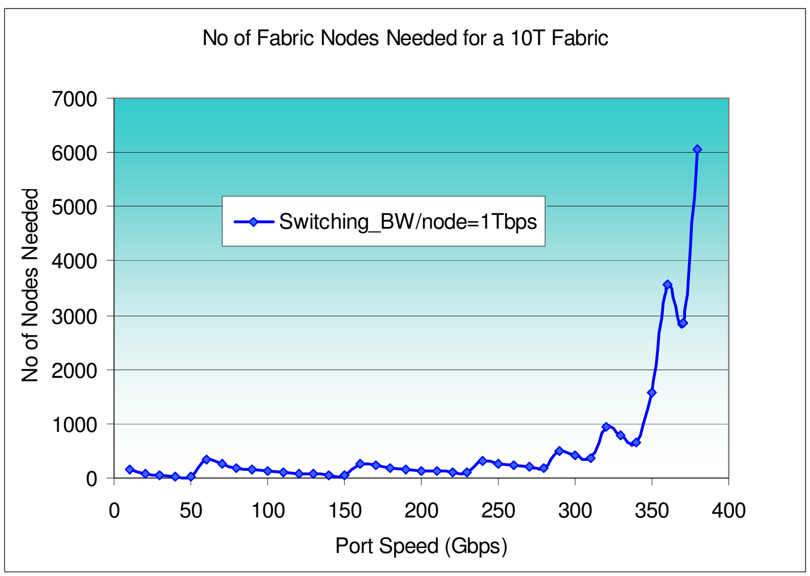


Fabric Size vs Port Speed

Constant switching BW/node of 1Tbps and constant fabric cross-section BW of 10Tbps Assumed

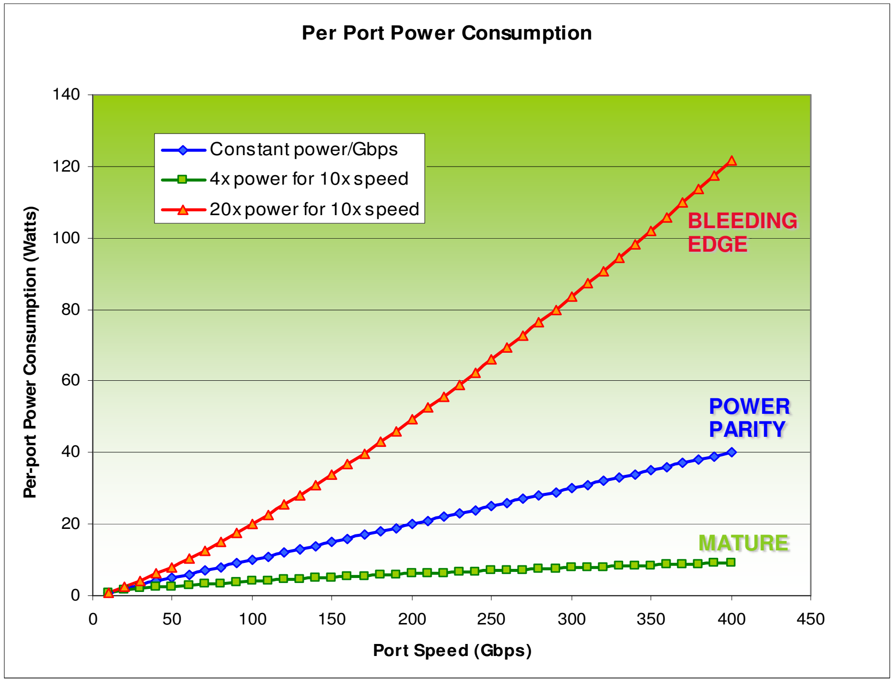


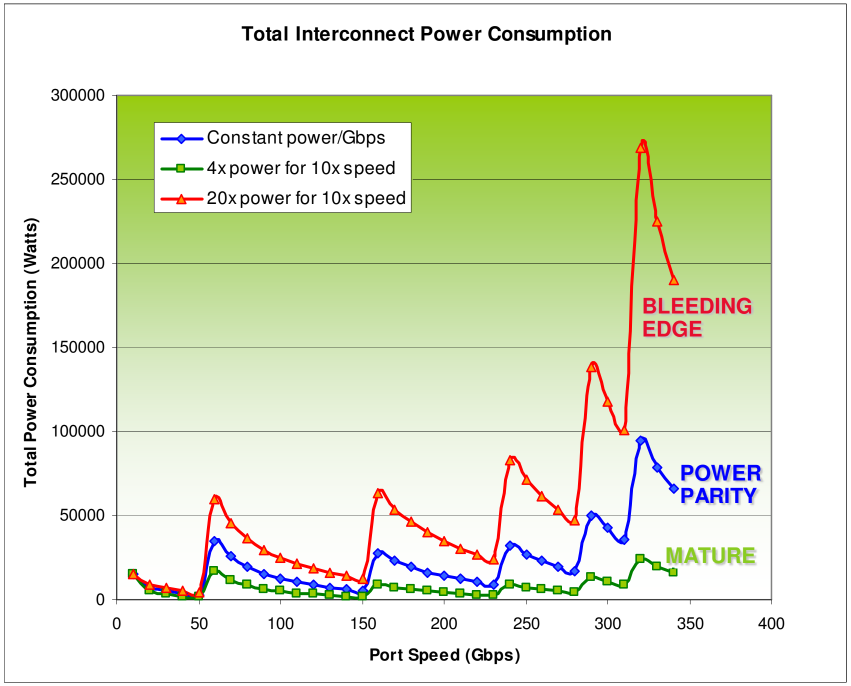




- Higher per port bandwidth reduces the number of available ports in a node with constant switching bandwidth
- In order to support same cross-sectional BW
    - more stages are needed in the fabric
    - More fabric nodes are needed

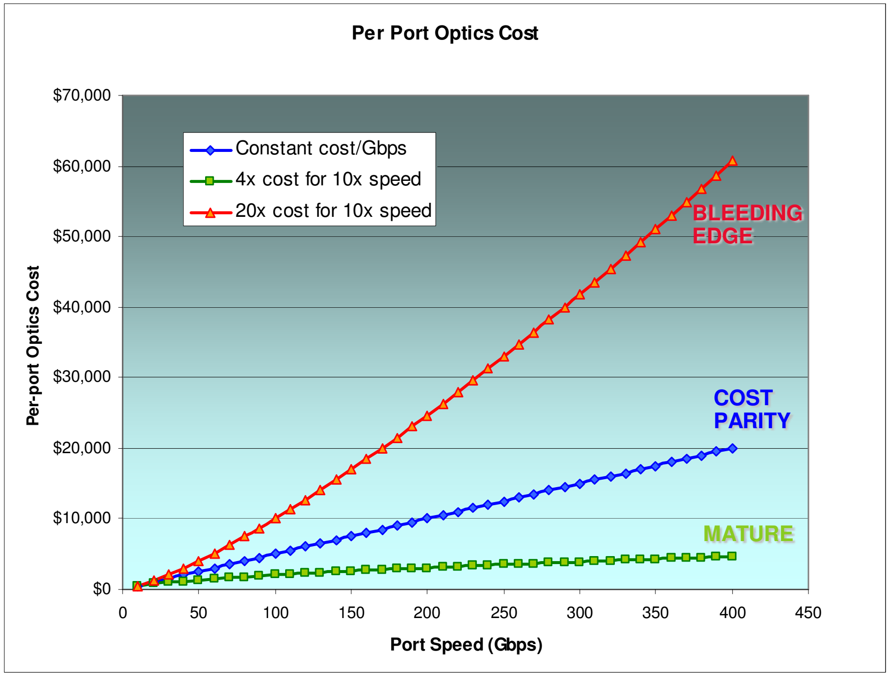
Power vs Port Speed

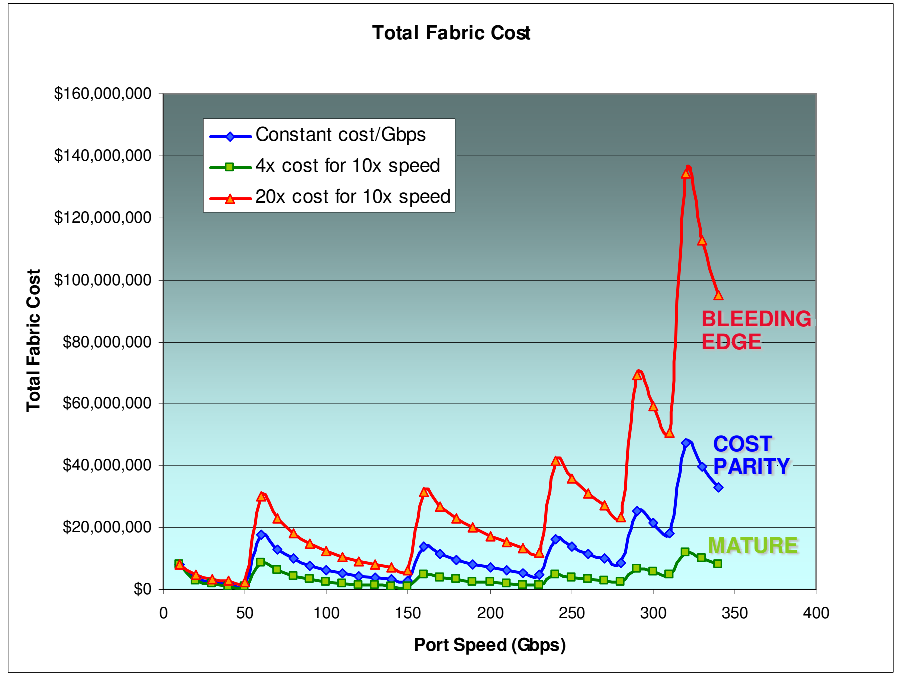




- **Three power consumption curves for interface optical modules:**
  - **Bleeding Edge: 20x power for 10x speed; e.g. if 10G is 1W/port, 100G is 20W/port**
  - **Power Parity: Power parity on per Gbps basis; e.g. if 10G is 1W/port, 100G is 10W/port**
  - **Mature: 4x power for 10x speed; e.g. if 10G is 1W/port, 100G is 4W/port**
- **Lower port speed provides lower power consumption**
- **For power consumption parity, power per optical module needs to follow the "mature" curve**

Cost vs Port Speed

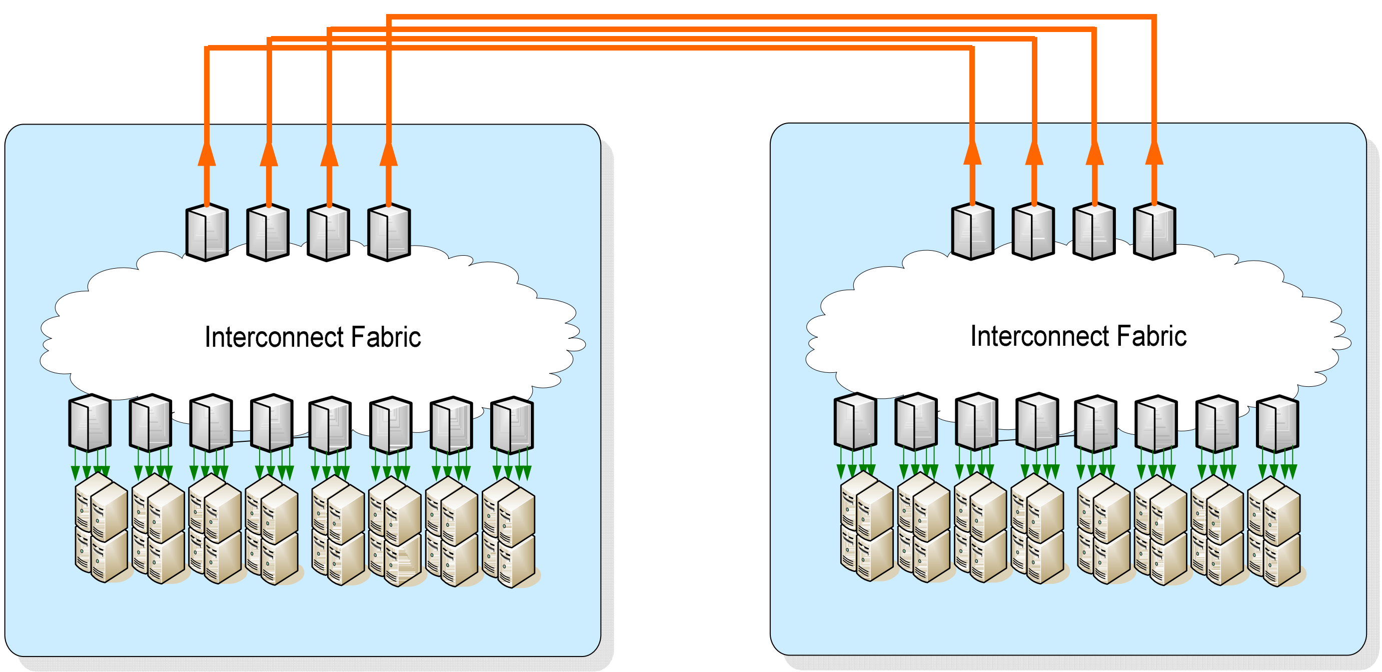




- **Three cost curves for optical interface modules:**
  - **Bleeding Edge: 20x cost for 10x speed**
  - **Cost Parity: Cost parity on per Gbps basis**
  - **Mature: 4x cost for 10x speed;**
- **Fiber cost is assumed to be constant per port (10% of 10G port cost)**
- **For fabric cost parity, cost of optical modules need to increase by < 4x for 10x increase in interface speed**

- INTRA-DATACENTER CONNECTIONS
- **INTER-DATACENTER CONNECTIONS**

Limited Fiber Availability, 2km+ reach

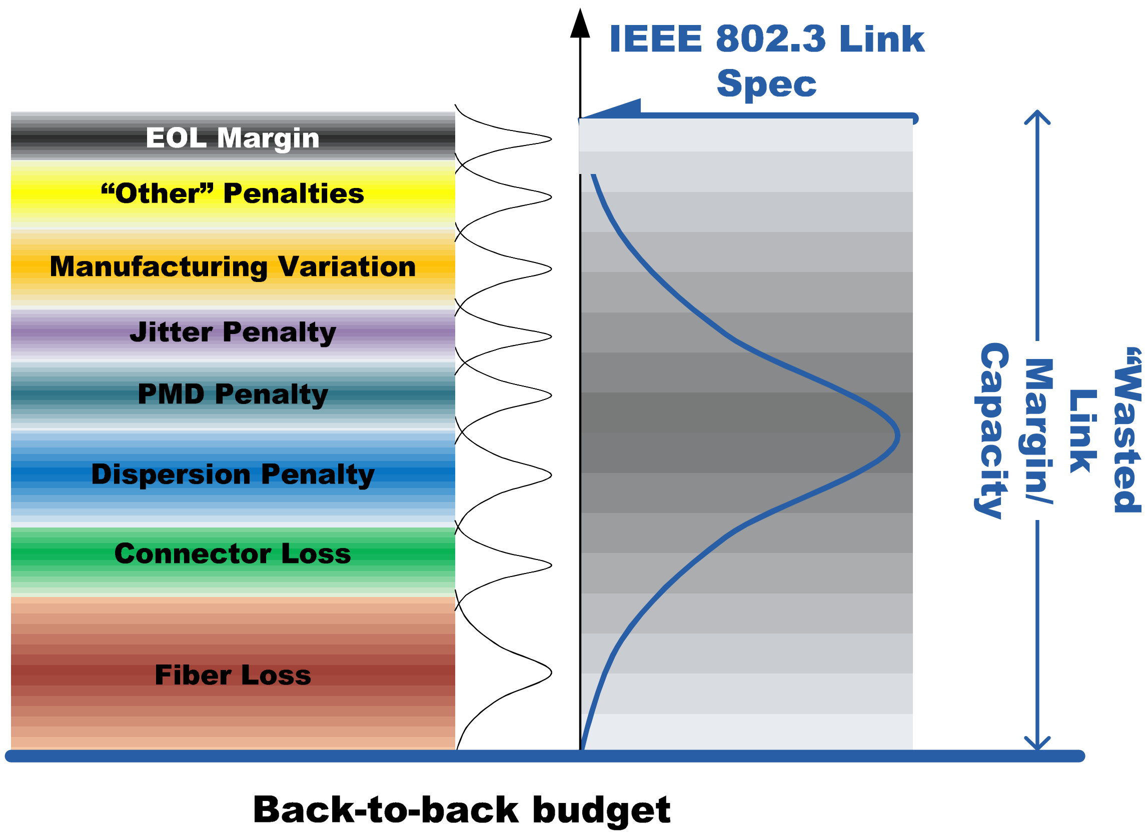


Beyond 100G: What data rate?

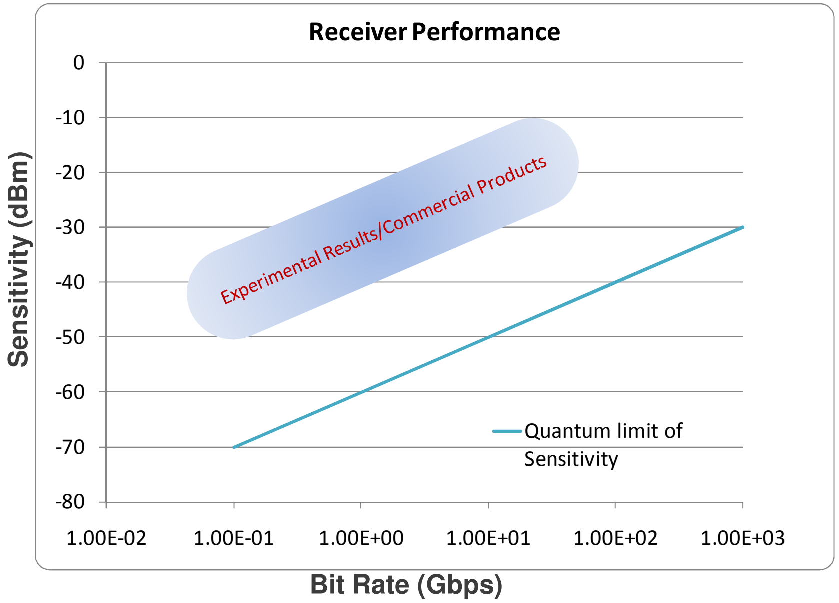
• 400Gbps? 1Tbps? Something “in-between”? How about all of the above?

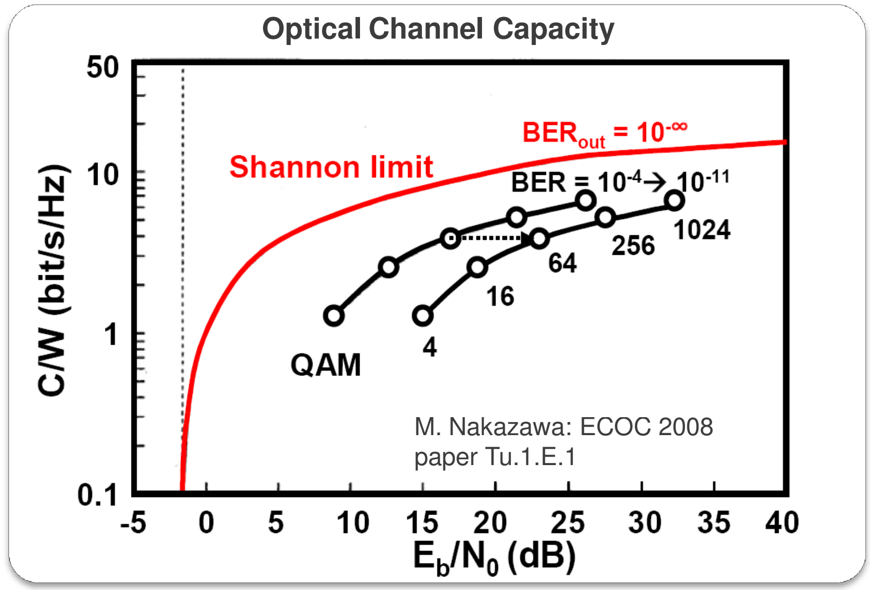
•Current optical PMD specs are designed for absolute worst-case penalties

•Significant capacity is untapped within the statistical variation of various penalties



Where is the Untapped Capacity?



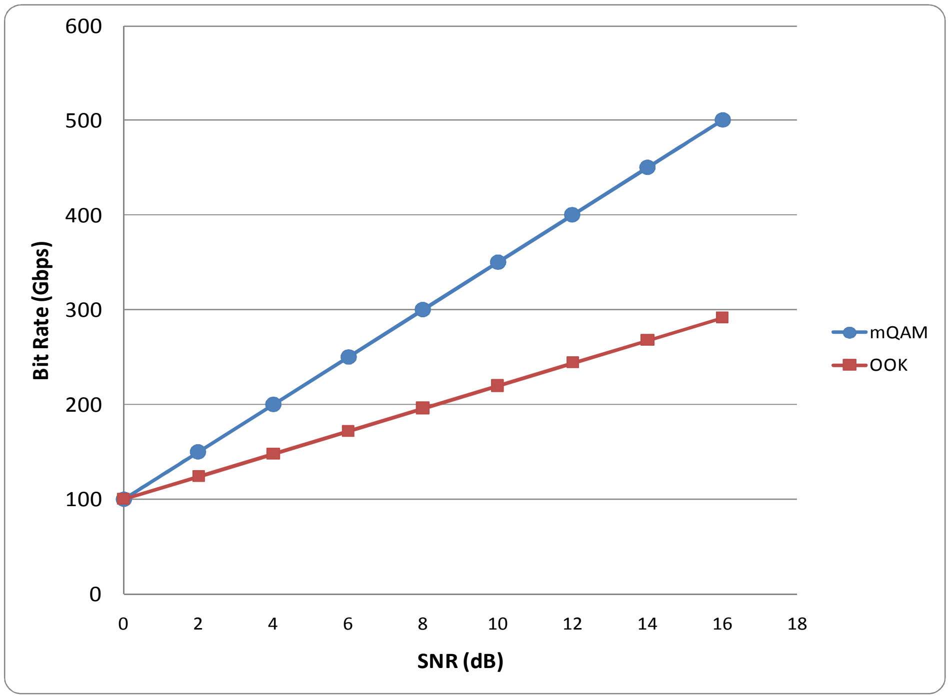


- **Unused Link Margin ≡ Untapped SNR ≡ Untapped Capacity**
- **In ideal world, 3dB of link margin will allow link capacity to be doubled**
- **Need the ability to use additional capacity (speed up the link) when available (temporal or statistical) and scale-back to the base-line capacity (40G/100G?) when not**

Rate Adaptive 100G+ Ethernet?

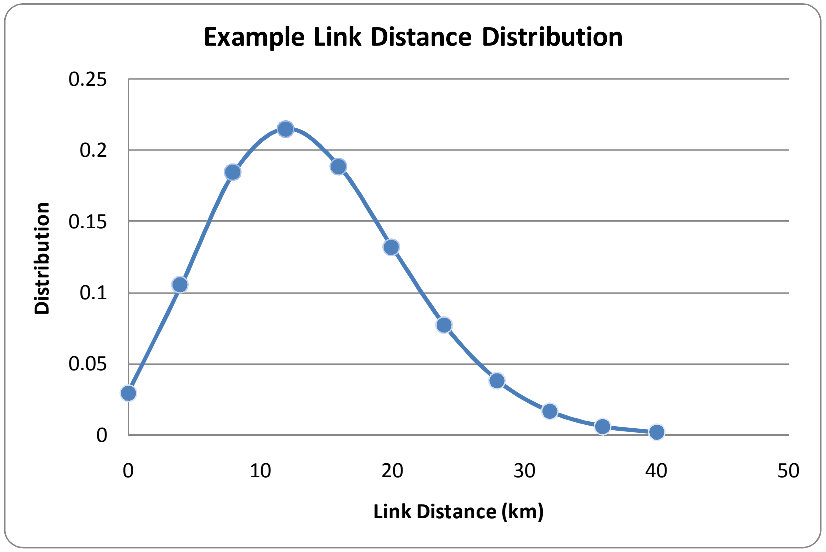
Google

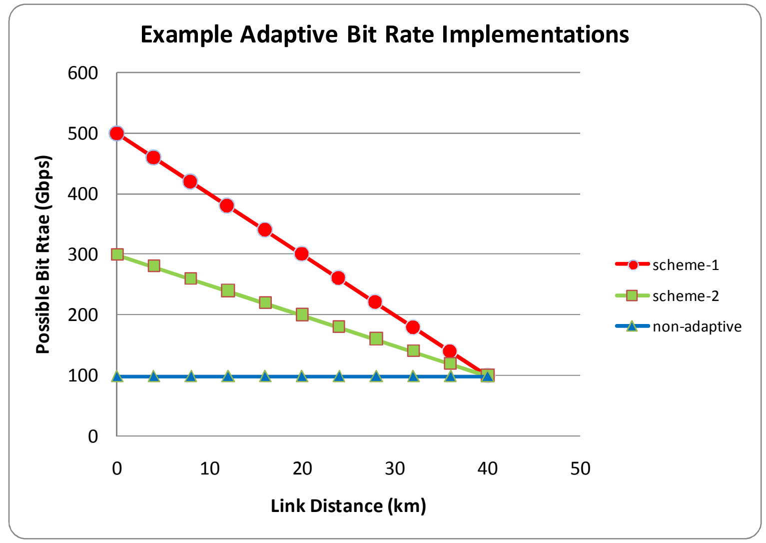
- There are existing standards within the IEEE802.3 family:
  - IEEE 802.3ah 10PASS-TS: based on MCM-VDSL standard
  - IEEE 802.3ah 2BASE-TL: based on SHDSL standard
- Needed when channels are close to physics-limit : We are getting there with 100Gbps+ Ethernet
- **Shorter links ≡ Higher capacity (matches perfectly with datacenter bandwidth demand distribution, see slide # 3)**

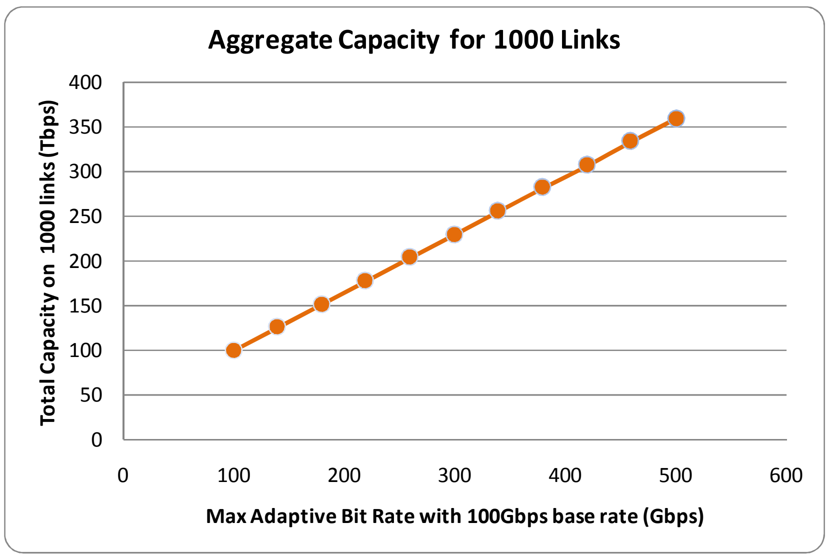


- **How to get there?**
  - High-order modulation
  - Multi-carrier-Modulation/OFDM
  - Ultra-dense WDM
  - **Combination of all the above**

Is There a Business Case?







- An example link-length distribution between datacenters is shown
  - Can be supported by a 40km capable PMD
- Various rate-adaptive 100GbE+ options are considered
  - Base rate is 100Gbps
  - Max adaptive bit-rate varies from 100G to 500G
- Aggregate capacity for 1000 such links is computed

Q&A

Google™