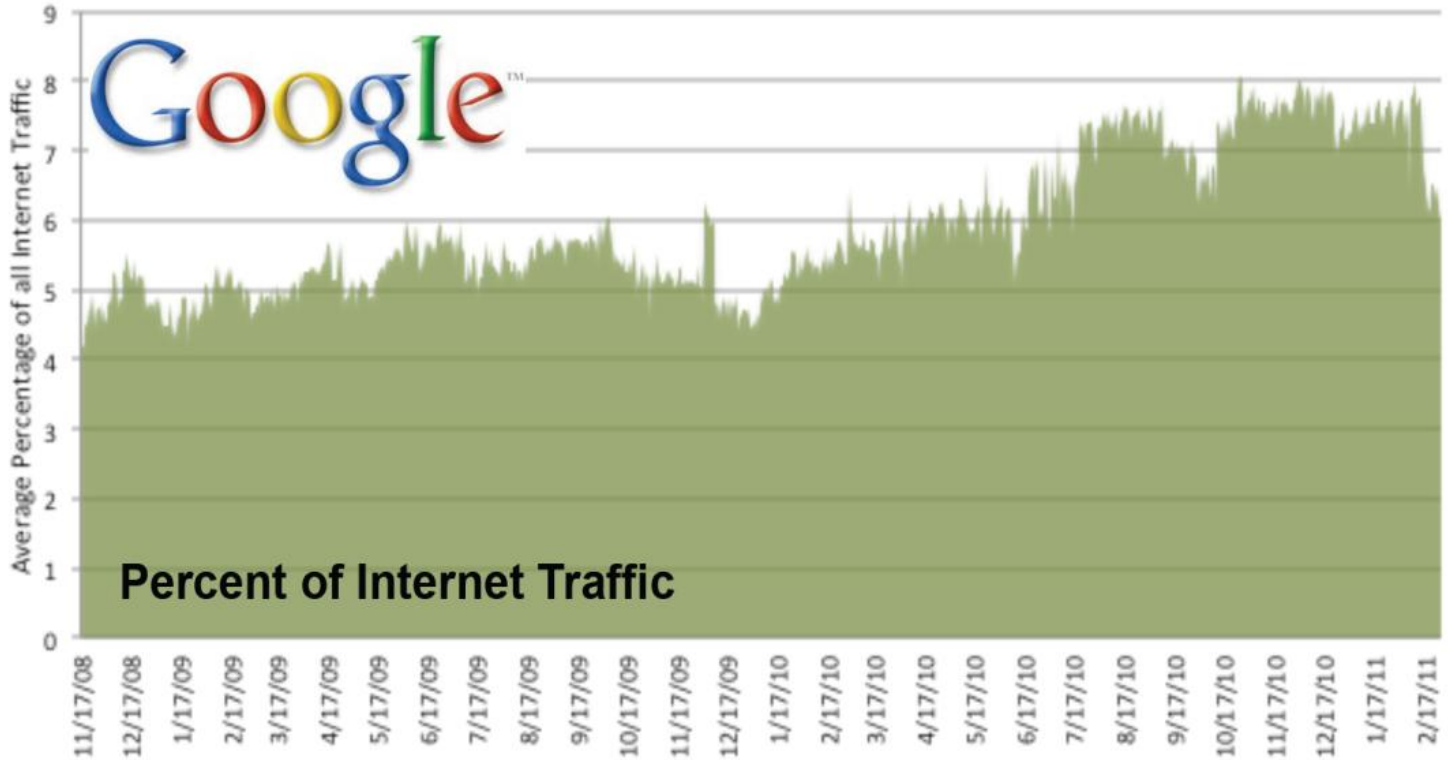
Ethernet Evolution and Warehouse Scale Computing

Bikash Koley
Google Network Architecture

Global Internet Traffic & Google



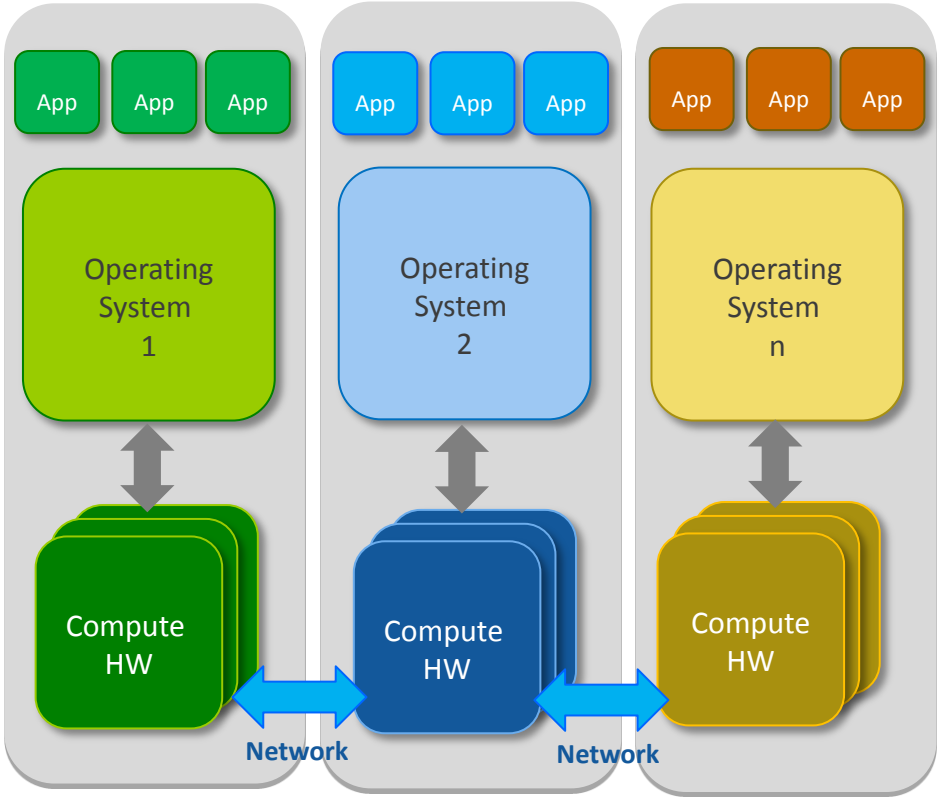
Internet Traffic Evolution 2007 - 2011
Presented by Craig Labovitz
April 6th, 2011 | Global Peering Forum

Google now represents an average 7% of all Internet traffic around the world. This number grows even larger (to as much as 8-12%) if include estimates of traffic offloaded by the increasingly common Google Global Cache (GGC) deployments and error in our data due to the extremely high degree of Google edge peering with consumer networks.

Data-centers vs WSCs

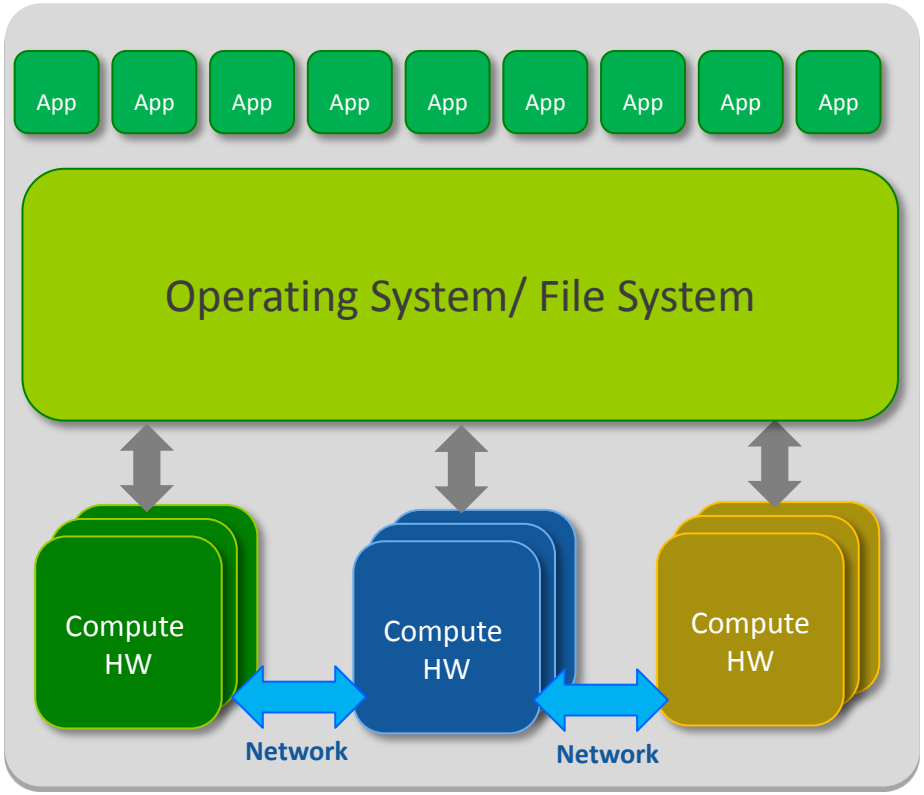
Google

Data-Center



- Heterogeneous hardware/ system software/ apps
- Partitioned resources managed separately
- Inefficient Compute

WSC

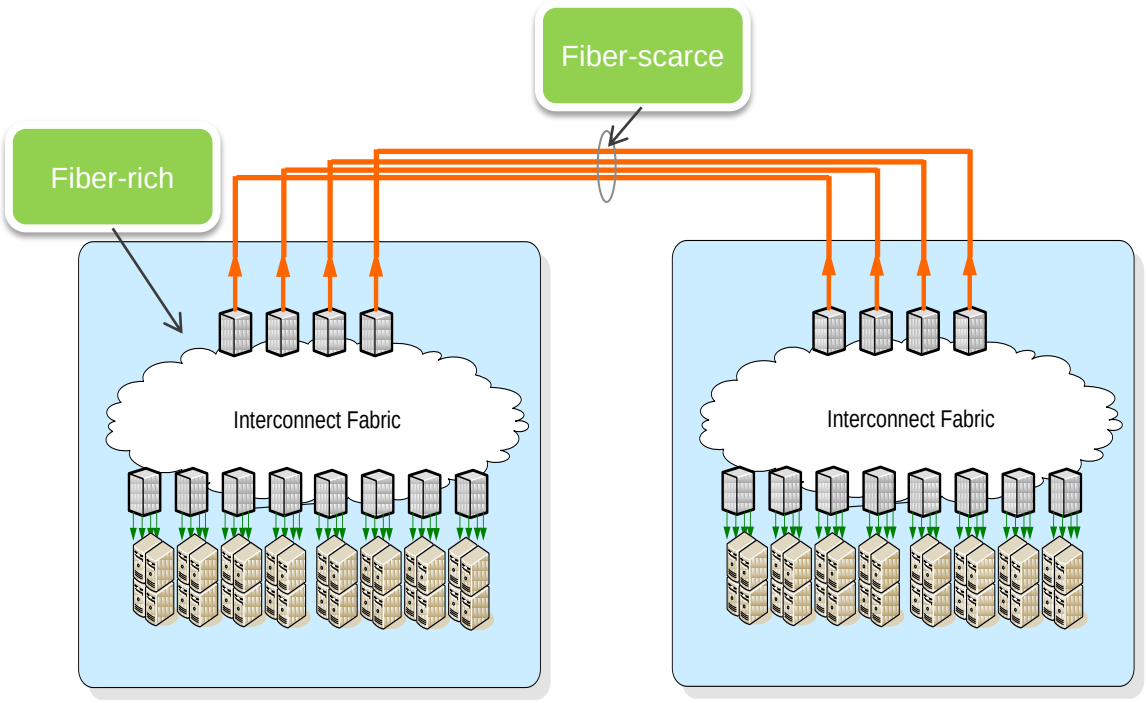


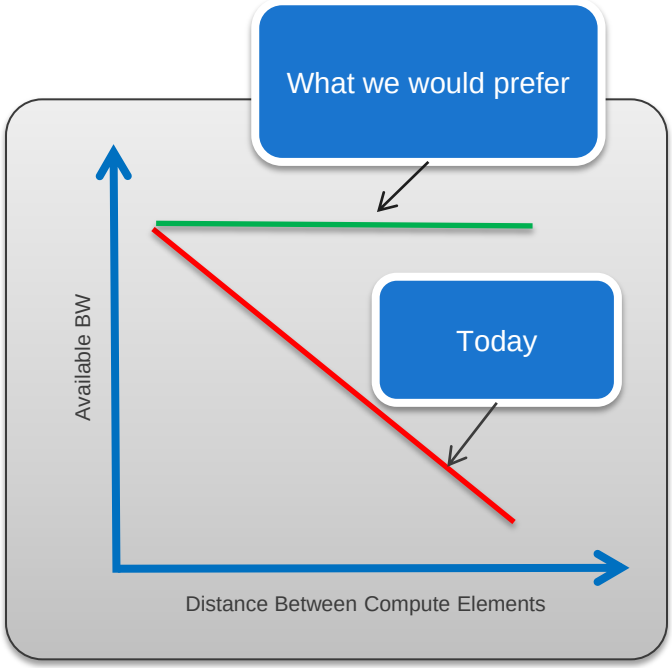
- Homogeneous hardware/ system software
- Common pool of resources managed centrally
- Very Efficient Compute

Warehouse-Scale Computer Interconnects

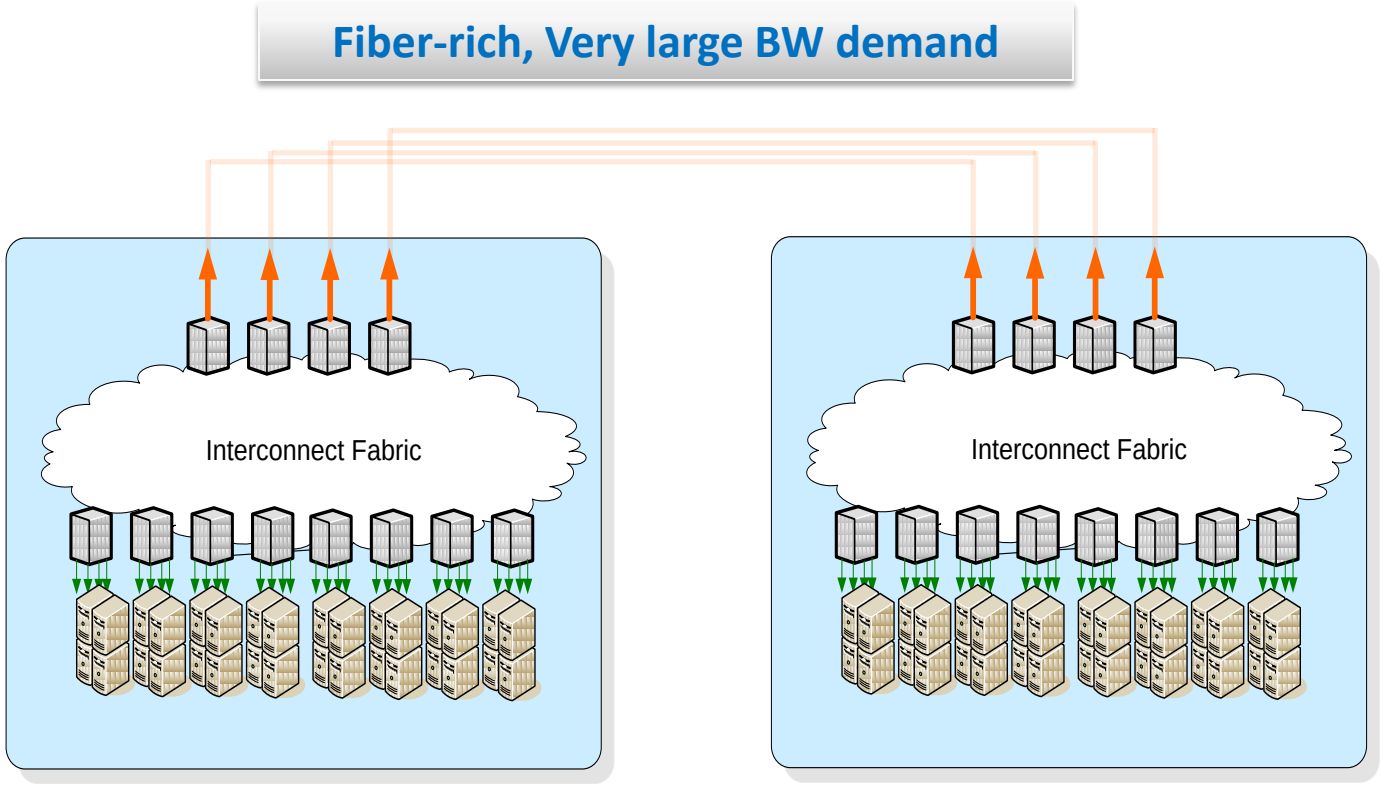
Google

- 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 thousands of kms

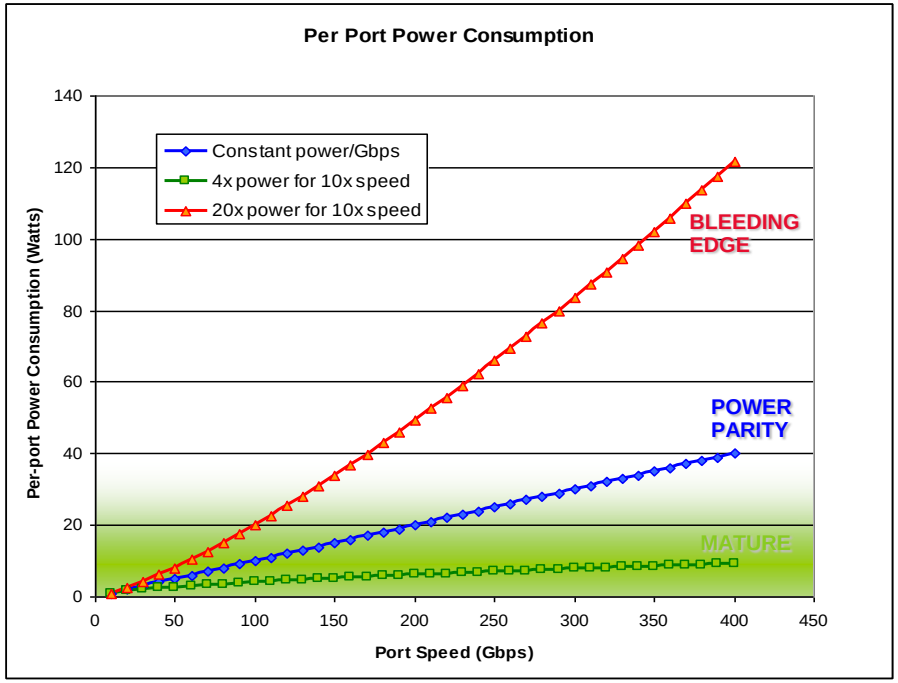


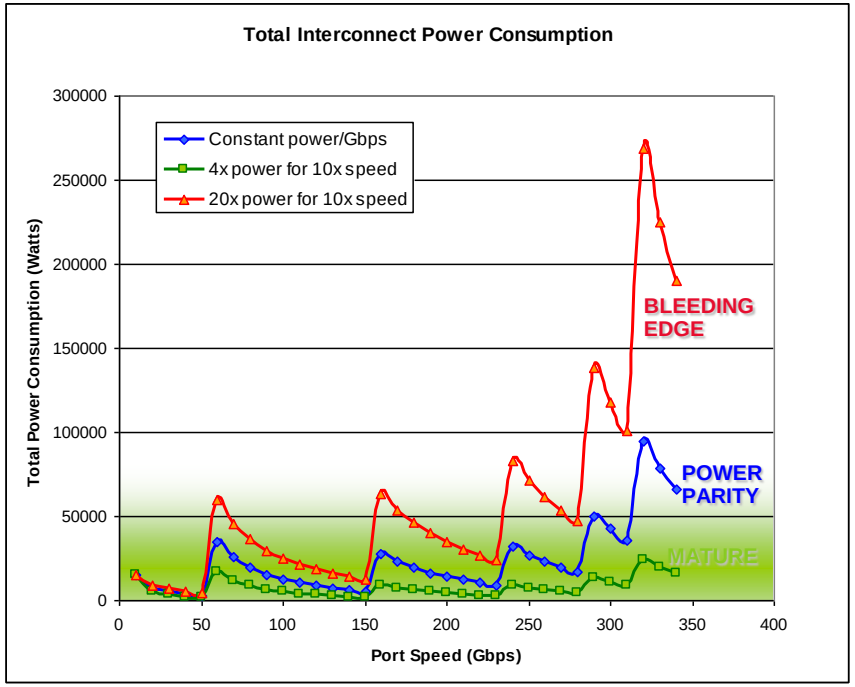


- **INTRA-DATACENTER CONNECTIONS**
- **INTER-DATACENTER CONNECTIONS**



Power vs Port Speed

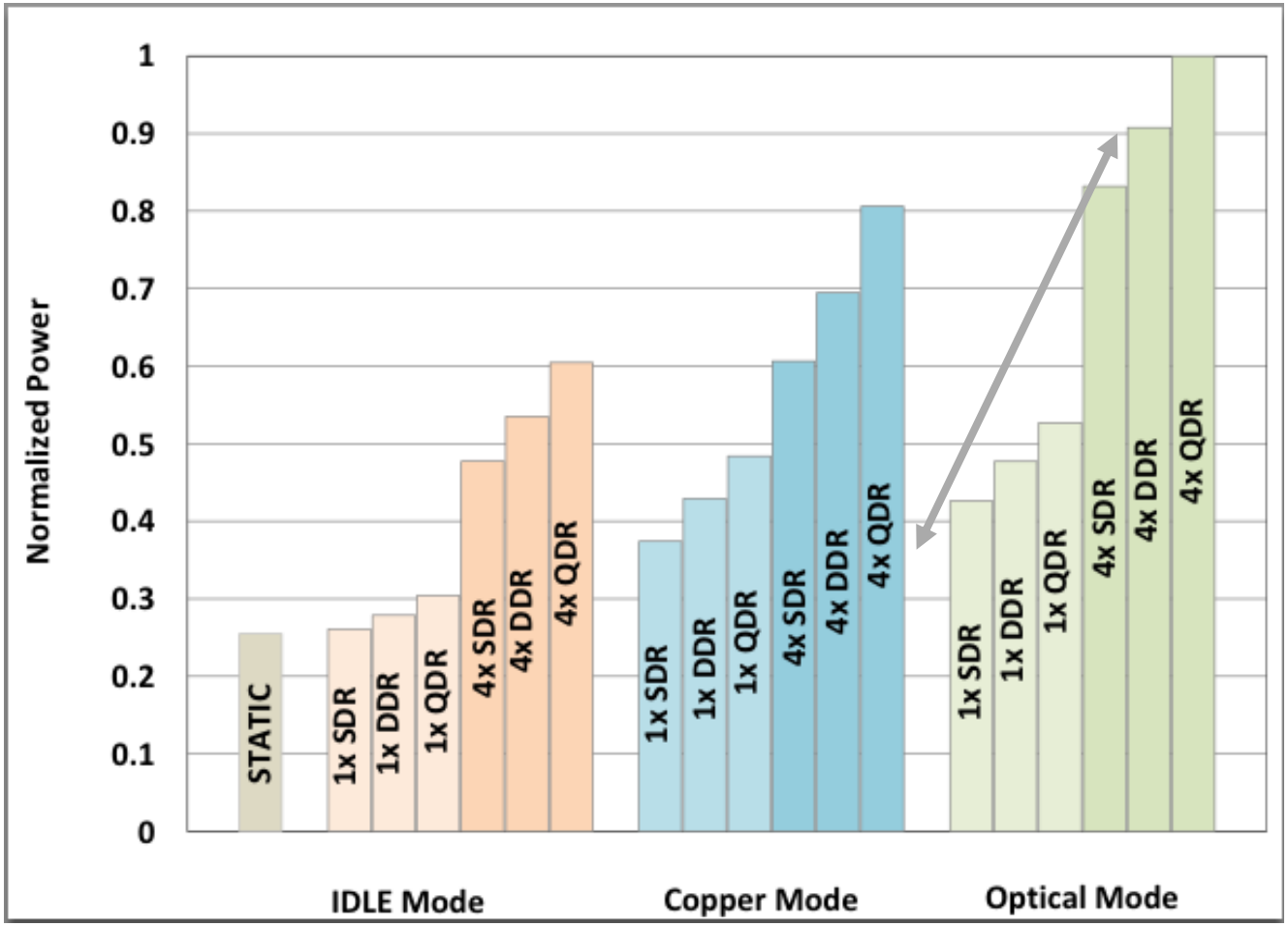




- Three power consumption curves for interface optical modules:
  - Bleeding Edge: 20x power for 10x speed; e.g. if 10G is 1W/port, 100G is 20W/port
  - Power Parity: Power parity on per Gbps basis; e.g. if 10G is 1W/port, 100G is 10W/port
  - Mature: 4x power for 10x speed; e.g. if 10G is 1W/port, 100G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

Energy Proportional Networks

"**Energy Proportional Datacenter Networks**", Dennis Abts, Mike Marty, Philip Wells, Peter Klausler, Hong Liu, Proceedings of the International Symposium on Computer Architecture, 2010, pp. 338-347.

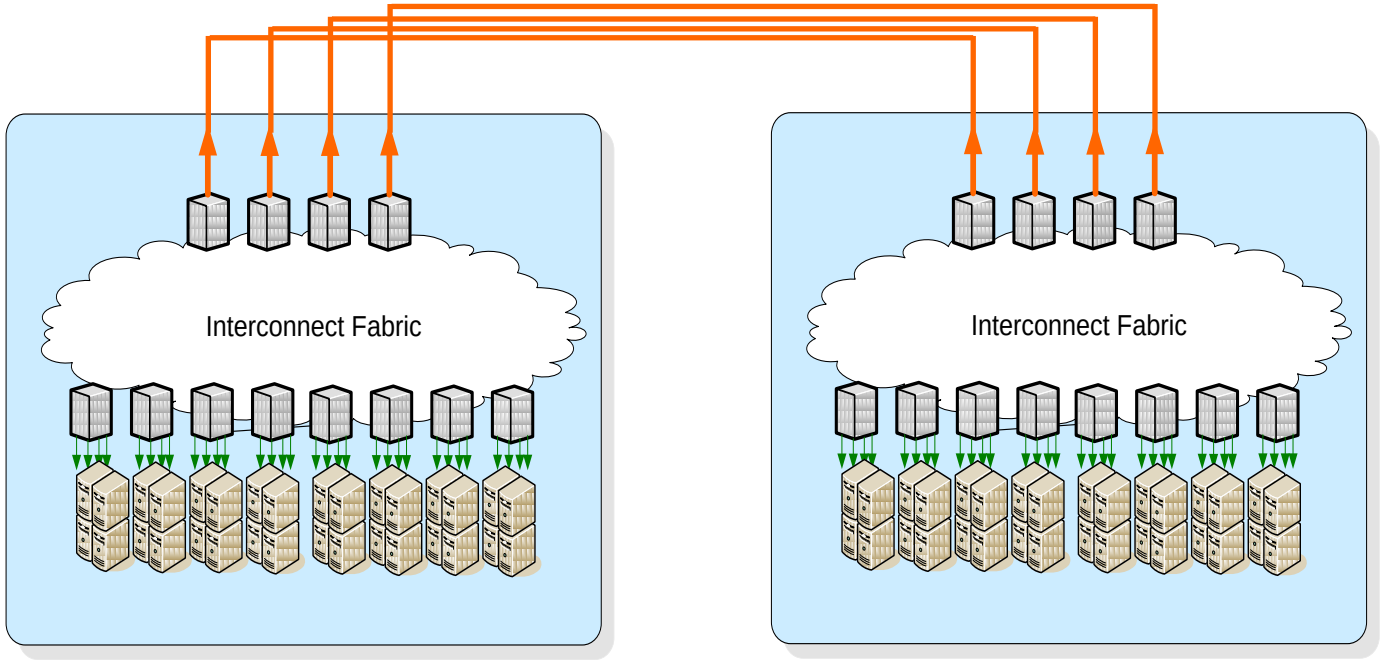


Infiniband
16x range of speed
and
60% dynamic power

Ethernet
40G sometimes is 4x10, could we get 1,2,3,4x10 support with energy savings?
Same for 100G/400G/1TbE (nx10) or (mx25).

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

Limited Fiber Availability, 10km-nx1000km reach



Inter-datacenter Network

Google

FIGURE OF MERIT	Cost Power	Cost Power	Cost	Spectral efficiency Reach		
TECHNOLOGY	MMF	SMF WDM	WDM	10G / 25G DWDM	100G /400G/Tbit?	
REACH	Intra-Datacenter	Campus	Metro	Regional	LH/ULH	Trans-Oceanic