Optical Interconnect Opportunities in Supercomputers and High End Computing

OFC 2012 Tutorial –
Category 14. Datacom, Computercom, and Short Range and Experimental Optical Networks (Tutorial)

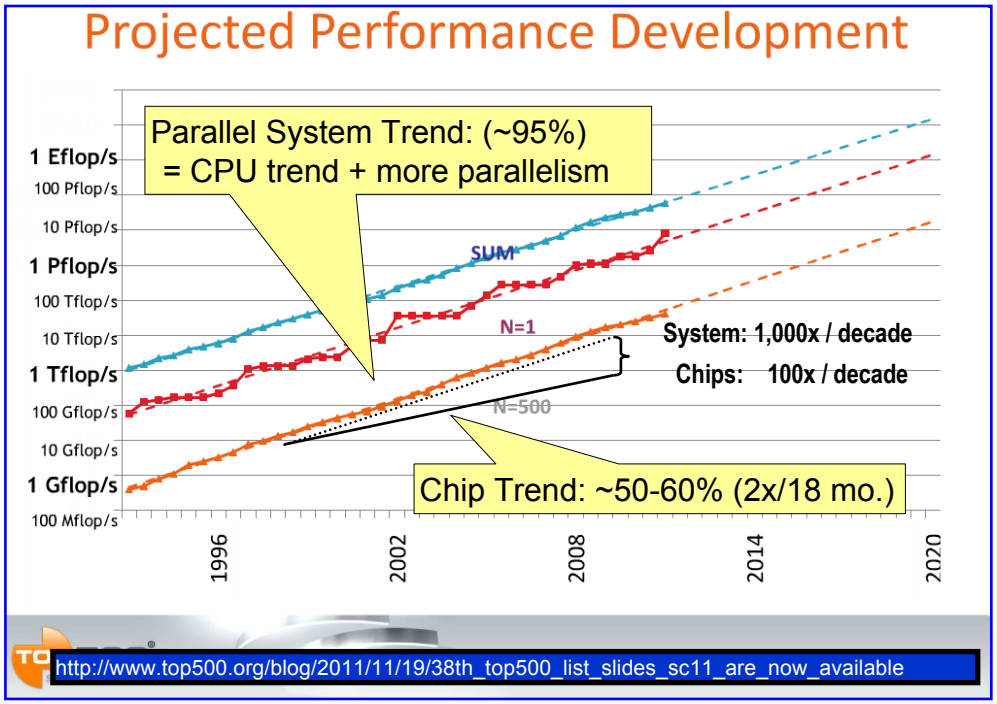
March 2012

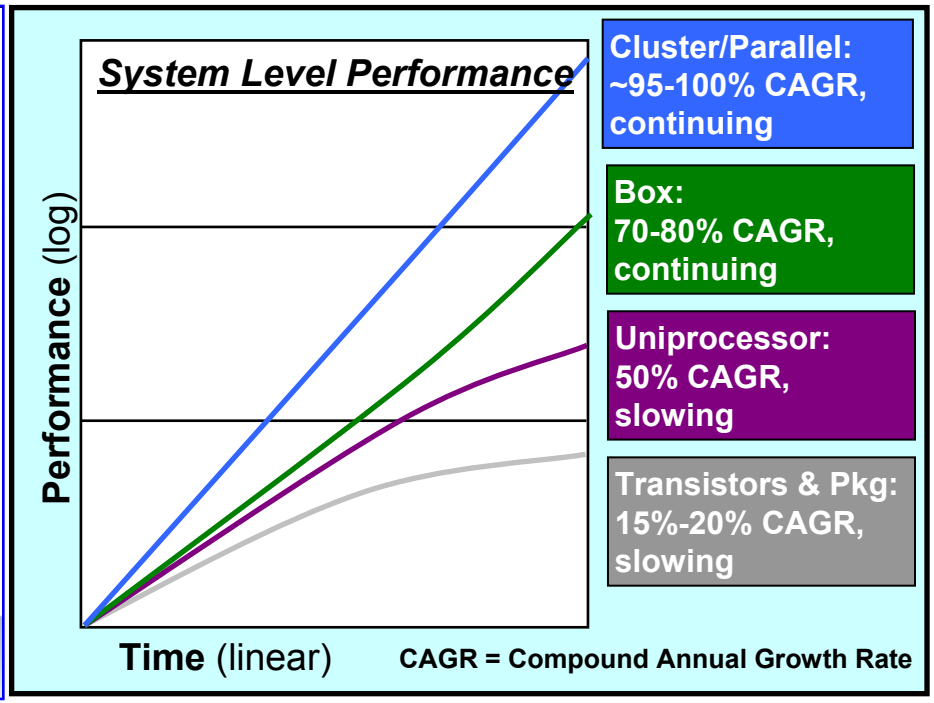
Alan Benner, bennera@us.ibm.com
IBM Corp. – Sr. Technical Staff Member, Systems & Technology Group
InfiniBand Trade Assoc. – Chair, ElectroMechanical Working Group

GOALS OF THIS TUTORIAL

- **Review optical interconnect from a systems architecture point of view**
  - Interconnect basics: What's important, what's not – future system needs
  - Data Centers: Infrastructure and Networking
  - HPC Systems / Supercomputer Systems
  - Review of some interesting research programs and progress
  - The rest of the decade – where are the challenges?

High-End computing systems: Steady Exponential Performance Growth





Note: Top500's Linpack needs moderate network performance
➔ Similar trends & growth rates apply to data centers.

- **System-level improvements will continue, at faster than Moore's-law rate**
  - System performance comes from aggregation of larger numbers of chips & boxes

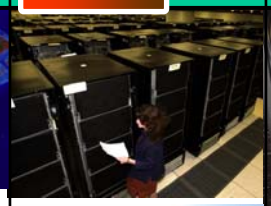
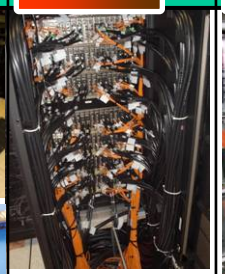
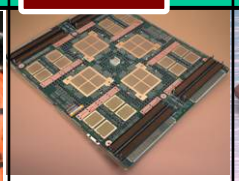
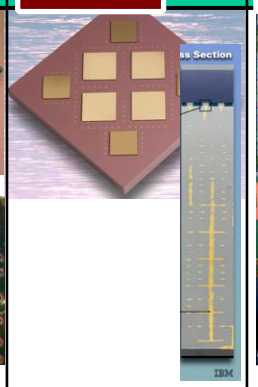
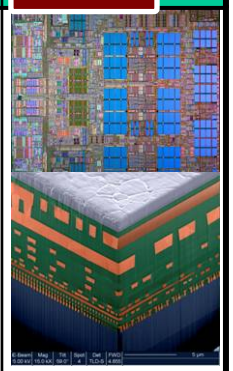
- **Bandwidth requirements must scale with system, roughly 0.5B/FLOP (memory + network)**
  - Receive an 8 Byte word, do ~32 ops with it, then transmit it onward ➔ 16B / 32 Operations
  - Actual BW requirements vary by application & algorithm by >10x : 0.5B/FLOP is an average

Optical Interconnect - Basics

The Landscape of Interconnect

PHYSICAL Link Types

Distinguished by Length & Packaging

	MAN & WAN	Cables – Long	Cables – Short	Backplane / Card-to-Card	Intra-Card	Intra-Module	Intra-chip
	 	 			 		
Length	Multi-km	10, - 300 m	1 m - 10 m	0.3 m - 1 m	0.1 m - 0.3 m	5 mm - 100 mm	0 mm - 20 mm
Typical # lanes per link	1	1 - 10s	1 - 10s	1 - 100s	1 - 100s	1 - 100s	1 - 100s
Use of optics	Since 80s	Since 90s	Since late 00's	Since 2010-2011	2012-2015	After 2015	Later

LOGICAL Link Types

Distinguished by Function & Link Protocol

	Internet	Local Area Network	Cluster / Data Center	Storage Area Network	Direct Attach Storage	I/O	Mezzanine Bus	SMP Coherency Bus	Memory Bus
	Traffic: IP Stds: Ethernet, ATM, SONET,	Traffic: HTML pages to laptops,.. Stds: 1G Ethernet, WiFi	Traffic: Intra-application, or intra-distributed-application Stds: InfiniBand, 1G Ethernet, 10/40/100Enet	Traffic: Read/Write to disk, shared Std: Fibre Channel	Traffic: Read/Write to disk, unshared Stds: SAS, SATA	Traffic: Load/store to I/O adapters Stds: PCI/PCIe	Traffic: Load/store to Hubs & bridges Stds: Hyper-Transport	Traffic: Load/store coherency ops to other CPUs' caches Stds: Hyper-transport	Traffic: Load/Store to DRAM or Memory Fanout chip Stds: DDR3/2/.
Key Characteristic	Inter-operability with "Everybody"	100-300m over RJ-45 / CAT5 cabling, or wireless	BW & latency to <60 meters	Dominated by FC	Shared tech. between servers & desktops	Shared tech between servers & desktops	Reliability	Reliability, massive BW, reliability	Reliability & cost vs. DRAM
Use of optics	Since 80s	Maybe Never? (Wireless, Building re-wiring, BW demand)	Since 2000s	Since 90s	Not yet	Scattered	Not yet	Coming	Coming later

Link Technology — Single-mode Optics — Mixed multi-mode optics & copper — Copper

The Landscape of Interconnect

PHYSICAL Link Types

Distinguished by Length & Packaging

	MAN & WAN	Cables – Long	Cables – Short	Backplane / Card-to-Card	Intra-Card	Intra-Module	Intra-chip
Length	Multi-km	10, - 300 m	1 m - 10 m	0.3 m - 1 m	0.1 m - 0.3 m	5 mm - 100 mm	0 mm - 20 mm
Typical # lanes per link	1	1 - 10s	1 - 10s	1 - 100s	1 - 100s	1 - 100s	1 - 100s
Use of optics	Since 80s	Since 90s	Since late 00's	Since 2010-2011	2012-2015	After 2015	Later

HPC- and Data Center-Specific Optical Interconnect

LOGICAL Link Types

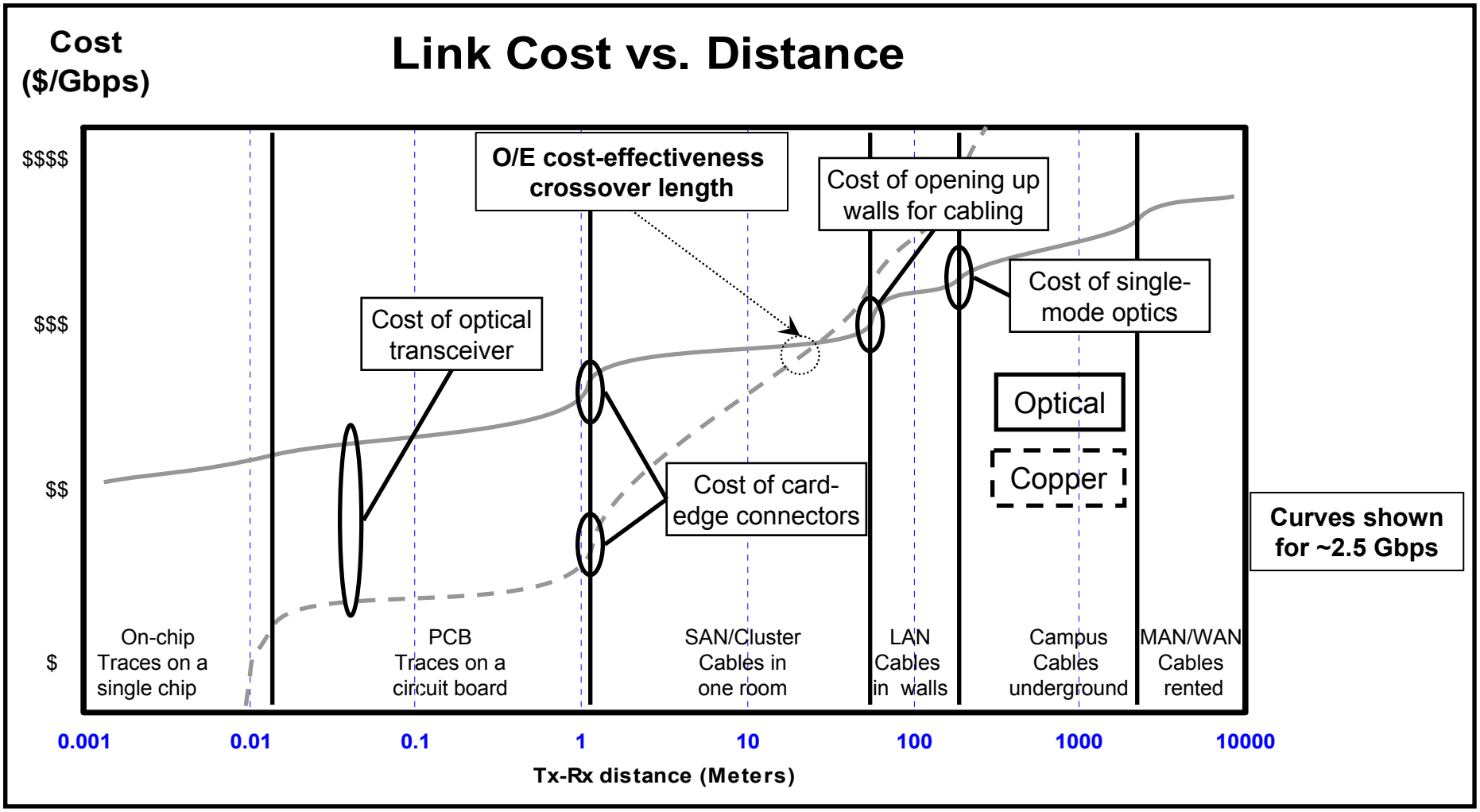
Distinguished by Function & Link Protocol

	Internet	Local Area Network	Cluster / Data Center	[illegible]	[illegible]	[illegible]	[illegible]	[illegible] cy Bus	Memory Bus
	Traffic: IP Stds: Ethernet, ATM, SONET,	Traffic: HTML pages to laptops,.. Stds: 1G Ethernet, WiFi	Traffic: Intra-application, or intra-distributed-application Stds: InfiniBand, 1G Ethernet, 10/40/100Enet	[illegible] Read/Write o disk, shared [illegible] Std: Fibre Channel	Read/Write to disk, unshared Stds: SAS, SATA	Load/store to I/O adapters Stds: PCI/PCIe	Load/store to Hubs & bridges Stds: Hyper-Transport	Load/store coherency ops to other CPUs' caches Stds: Hyper-transport	Traffic: Load/Store to DRAM or Memory Fanout chip Stds: DDR3/2/.
Key Characteristic	Inter-operability with "Everybody"	100-300m over RJ-45 / CAT5 cabling, or wireless	BW & latency to <60-250 meters	Dominated by FC	Shared tech. between servers & desktops	Shared tech between servers & desktops	Reliability	Reliability, massive BW, reliability	Reliability & cost vs. DRAM
Use of optics	Since 80s	Maybe Never? (Wireless, Building re-wiring, BW demand)	Since 2000s	Since 90s	Not yet	Scattered	Not yet	Coming	Coming later

Link Technology — Single-mode Optics — Mixed multi-mode optics & copper — Copper

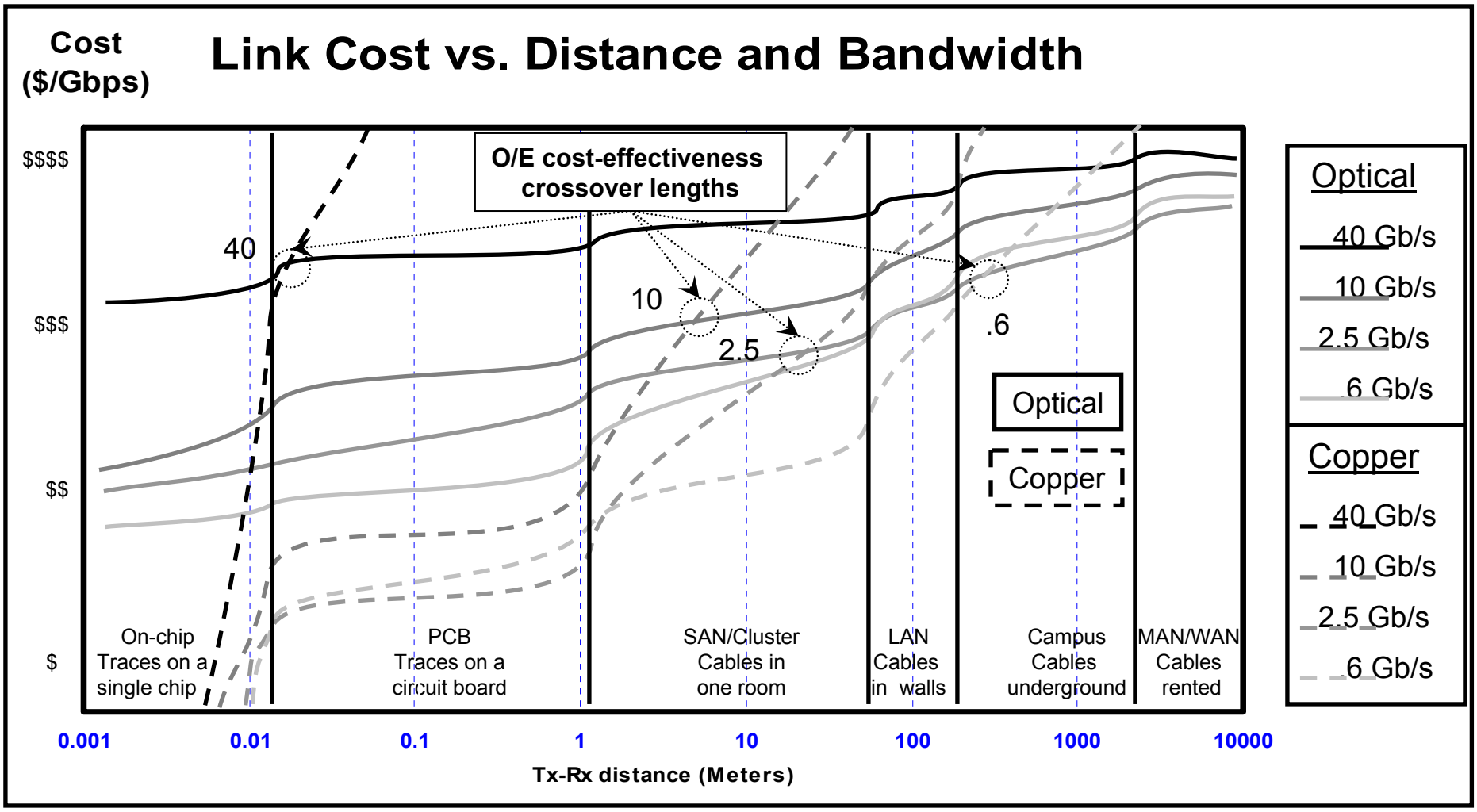
Optical vs. Electrical - Cost-Effectiveness Link Crossover Length

- **Qualitative Summary:**
  - At short distances, copper is less expensive. At longer distances, optics is cheaper
    - Expense is measured several ways: (parts cost, design complexity, Watts, BW density, etc.)
  - System design requires using optimal crossover length, using technology where appropriate



Cost-Effectiveness Link Crossover Length – Dependence on bit-rate

- **Over time, copper & optical get cheaper at pretty much the same rate**
  - ➔ The crossover length at a particular bit-rate have stayed pretty constant
- **As bit-rates have risen, a higher percentage of overall interconnect have moved to optics**
  - At 25 Gb/s, it appears that the crossover distance is ~2 - 3 M. Copper only works in-rack.



Power Efficiency Study: Copper vs. Optical

Power Efficiency Design Example: 16 PF Scale Cabling Options

- **Thought Experiment:**
  - Imagine a 2014 Top-10 system – say 16 PF – Using POWER7-775 System Design

- **~16 PF System will require various lengths of links:**
  - <1m: Between 4 drawers of a SuperNode
  - 1-3m: Between 8 SuperNodes in 3-rack Building Blocks
  - 3-20m: Between "closely-spaced" Building Blocks (1/4 of other BBs in system)
  - 20-50m: Between "far-spaced" Building Blocks (3/4 of other BBs in system)

- **16 PF POWER 775 / PERCS system would need many many links**

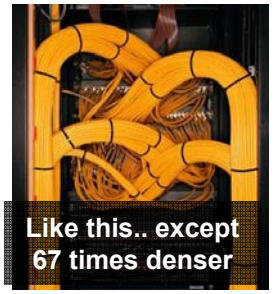
POWER7-775 / PERCS 16 PF System: # of Links				
	<1meter	**1-3 meter**	**3-20m**	**20-50m**
	In 4-drawer SuperNode	In 3-rack Building Block	of 1/4 system racks	of 3/4 system racks
Avg.# / drawer	96.00	1.75	30.00	96.00
# drawers	2,048	2,048	2,048	2,048
Total # of 120Gbps Transceivers	196,608	3,584	61,440	196,608

What if we interconnected with copper vs. optical?

16 PF-Scale Cabling Options: 10GBASE-T

- **Imagine we cabled this with "normal" 10G Ethernet (if it fit physically)**
  - Power utilization: ~3 Watts per 10G PHY transceiver (300mW/Gbps)
    - Inexpensive cables & connectors require high-power signal processing

	<1meter	1-3 meter	3-20m	20-50m	
Avg.# / drawer	96.00	1.75	30.00	96.00	
# drawers	2,048	2,048	2,048	2,048	
Total # of 120G XCVRs	196,608	3,584	61,440	196,608	Total Power, **MegaWatts**
Total # of 10GBase-T PHYs	2,359,296	43,008	737,280	2,359,296	
Power, Watts	7,077,888	129,024	2,211,840	7,077,888	16.5



➔ **At ~$1M per MWatt per year, with ~10-year machine life, 10GBase-T cabling would add >$165M in operating cost, above the machine cost**

16 PF-Scale Cabling Options: Optimized Copper

- **Imagine we cabled it with improved "Active copper cable", which allows lower power (75-150 mW/Gbps)**
  - Better twin-ax cables w/active circuits *inside* good connectors reduce the signal-processing required: 1.5W/20Gbps (<20m), or 5W (20-50m) (i.e., 75-250 mW/Gbps, length-dependent)
  - (…but it *still* won't fit – connectors & cables are too big..)

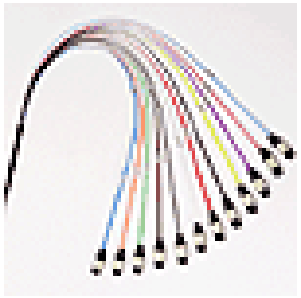
	<1meter	1-3 meter	3-20m	20-50m	
Avg.# / drawer	96.00	1.75	30.00	96.00	
# drawers	2,048	2,048	2,048	2,048	
Total # of 120G XCVRs	196,608	3,584	61,440	196,608	Total Power, **MegaWatts**
Total 20G Active cable ends	1,179,648	21,504	368,640	1,179,648	
Cable power, Watts	1,769,472	32,256	552,960	5,898,240	8.3

➔ **Active copper saves >$80M vs. passive copper in operating costs, over 10 years**

16 PF-Scale Cabling Options: Optical

- **Optical interconnect allows lower power (25 mW/Gbps)**
  - VCSEL/MMF requires <3W per 120Gbps (length-independent)

	<1meter	1-3 meter	3-20m	20-50m	
Avg.# / drawer	96.00	1.75	30.00	96.00	
# drawers	2,048	2,048	2,048	2,048	
Total # of 120G XCVRs	196,608	3,584	61,440	196,608	Total Power, **MegaWatts**
Cable power, Watts	589,824	10,752	184,320	589,824	1.4



➔**10-year cost of electrical power: <$15M**

➔**The message: In comparison to "cheap" 10GBASE-T, optical interconnect saves roughly $150M in machine operating costs over 10 years.**
- ➔*Plus* the connectors can actually fit in the system

➔**Better interconnect saves money in other ways, too**
- ➔Cables are much smaller/lighter/easy to install and manage
- ➔Signal integrity is more predictable across all lengths of cables
- ➔Efficient server utilization by moving jobs & data where most efficiently executed

Data Center Networking

Data Center Dynamics, 2011

- **Data Centers are growing in scale incredibly quickly:**
  - 1999 “Large” data center: 5,000 ft^2
  - 2004 “Large” data center: 50,000 ft^2
  - 2009 “Large” data Center: 500,000 ft^2
  - 2011 (started): IBM/Range Technology Data Center in China (near Beijing): ~624,000 ft^2

- **Power & Cooling Requirements growing nearly as fast**
  - 2001: 1-2 supercomputer centers in the world needed 10 MW of power
  - 2011: dozens of 10 MW data centers worldwide,
    US Gov’t planning 60 & 65 MW data centers

- **Power efficiency at all levels is critical**
  - Electrical power is the major ongoing cost for data centers.
  - Note: Moore’s law doesn’t apply to power and cooling – but there are efficiencies to be had

Facebook Data Center in the Oregon Desert

MIXING: Dampers let dry desert air into the facilities penthouse level. In the winter months, when the outside air is very cold, warm return air can be mixed in.

FILTERING: Air passes through filters to stop desert particles and insects from entering the system.

MISTING: Bacteria is killed and minerals removed in the facilities water treatment area. The treated water is then sprayed as a fine mist into the air. Evaporative cooling ensues, cooling the air to between 65°- 80°. A relative humidity of 35-65% is reached, eliminating problems of static electricity. Filters remove water particles from entering the system.

MOVING: Energy efficient 5 horsepower centrifugal fans move the cool air through air shafts down to the server floor where the air travels through the open servers that are stacked on racks. Each rack holds 90 servers."><

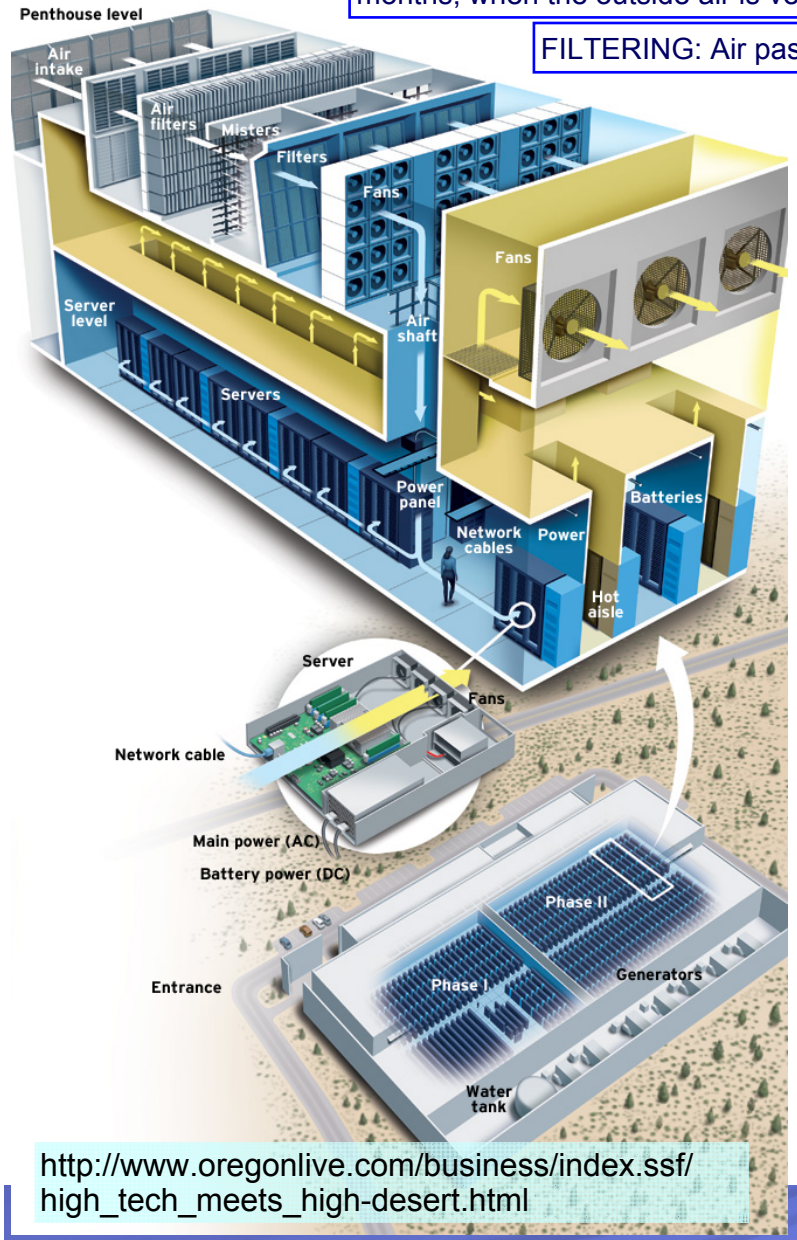
REMOVING: Exhaust fans remove the server return air (typically about 95°).

POWER CONVERSION: Conventional data centers convert power a number of times before it's used. Each conversion results in a loss of power. The custom servers run at a higher voltage and so can use power straight from the grid. First, power travels to a custom fabricated reactor power panel (where irregulatatiies are removed) and then to the servers themselves.

BATTERIES: The UPS system is a standby system. In the case of a power failure, batteries will provide 45 seconds of power to the servers until generators kick in.

OPEN CASING; Servers were designed without a cover to allow the air to freely pass through and cool the circuitry

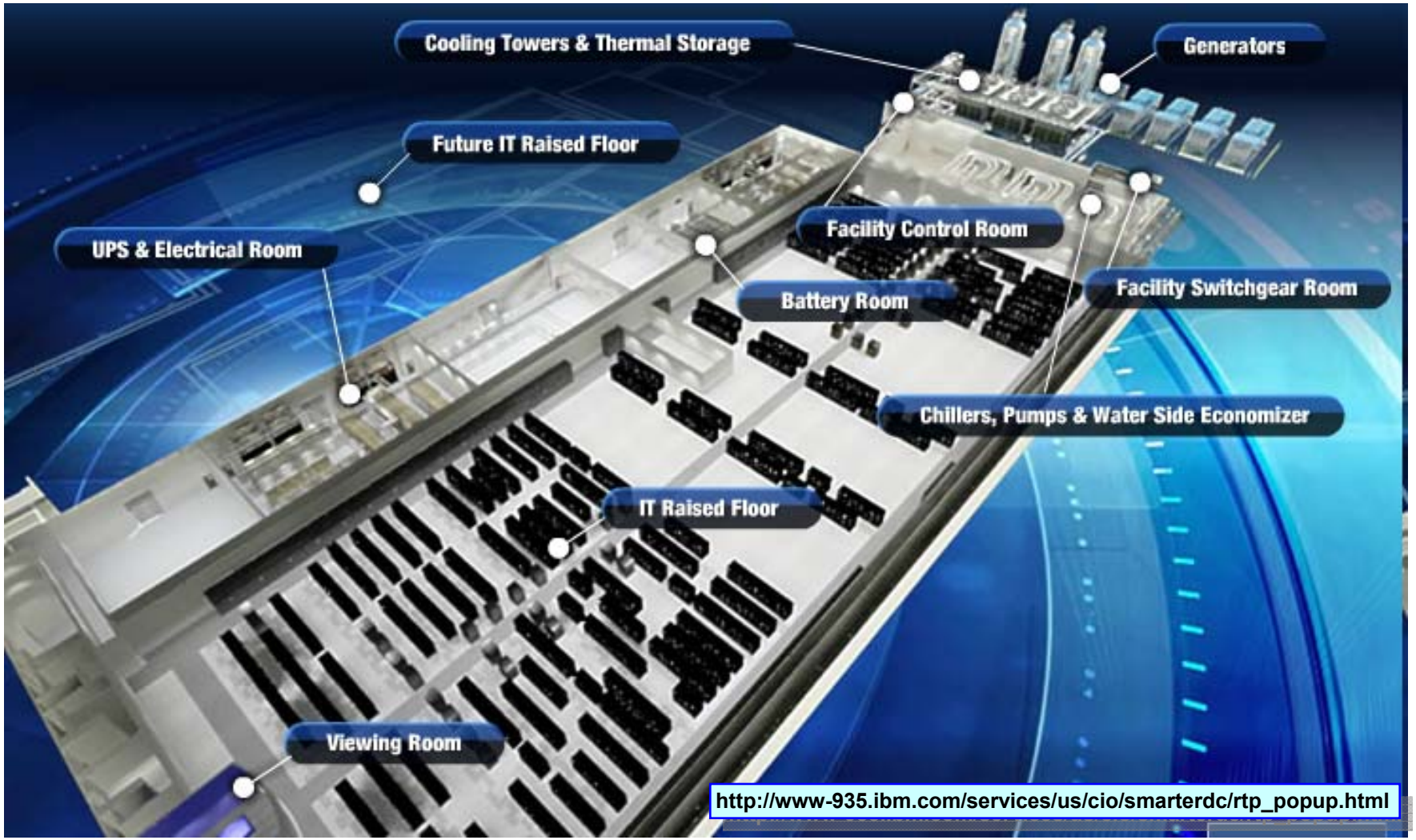
FANS: The servers were designed with bigger fans that use less energy.



http://www.oregonlive.com/business/index.ssf/high_tech_meets_high-desert.html

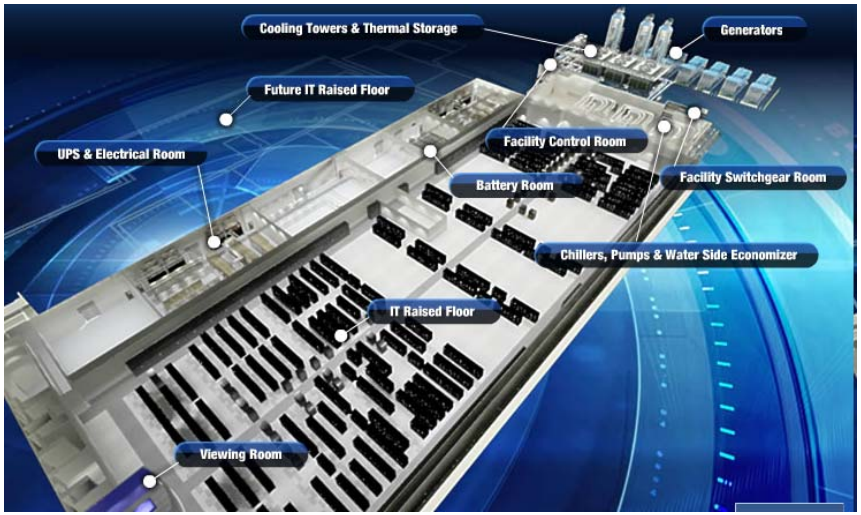
- Building-scale engineering required to support large-scale machines

Raleigh Leadership Data Center



- **Data center design reflects key strategies:**
  - Flexibility for growth for 20-30 years while IT equipment changes every 3-5 years.
  - Integrated management of IT and data center infrastructure.
  - Energy efficient power & cooling systems (LEED Gold) with full redundancy.

Raleigh Leadership Data Center – Equipment & photos




Two-cell 1300 ton cooling tower with variable speed fans


Six 2.5MW diesel generators for emergency power

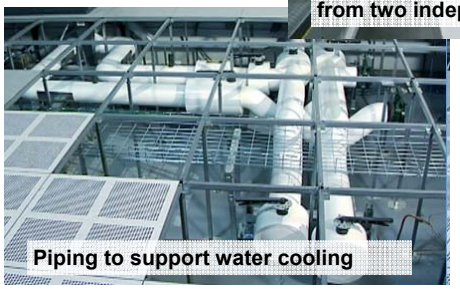

Three 50,000 gallon thermal storage tanks


1300-ton centrifugal chiller with variable speed drive


Water side economizer for 3900 hrs/yr of free cooling


3 MW electtrical switchgear from two independent sources


60,000 sq. ft. of IT raised floor space

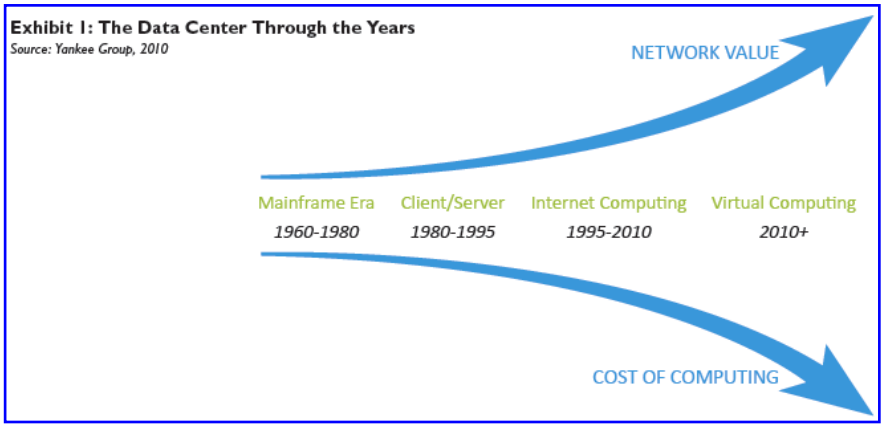

Piping to support water cooling

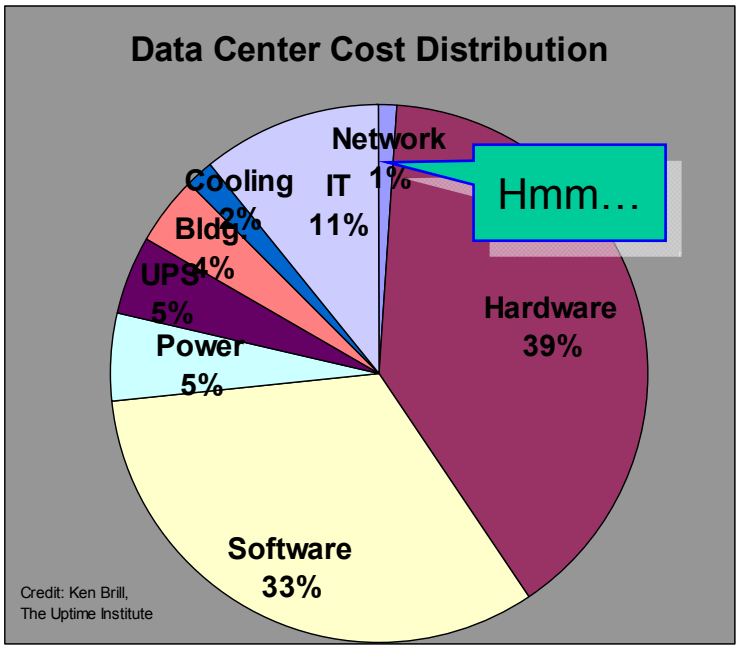

3 MW of wet cell battery capacity for 15 minutes of backup

- **Modern data center infrastructure is heavy-duty industrial-scale factory-style equipment**

Data Center Networking – A few key observations

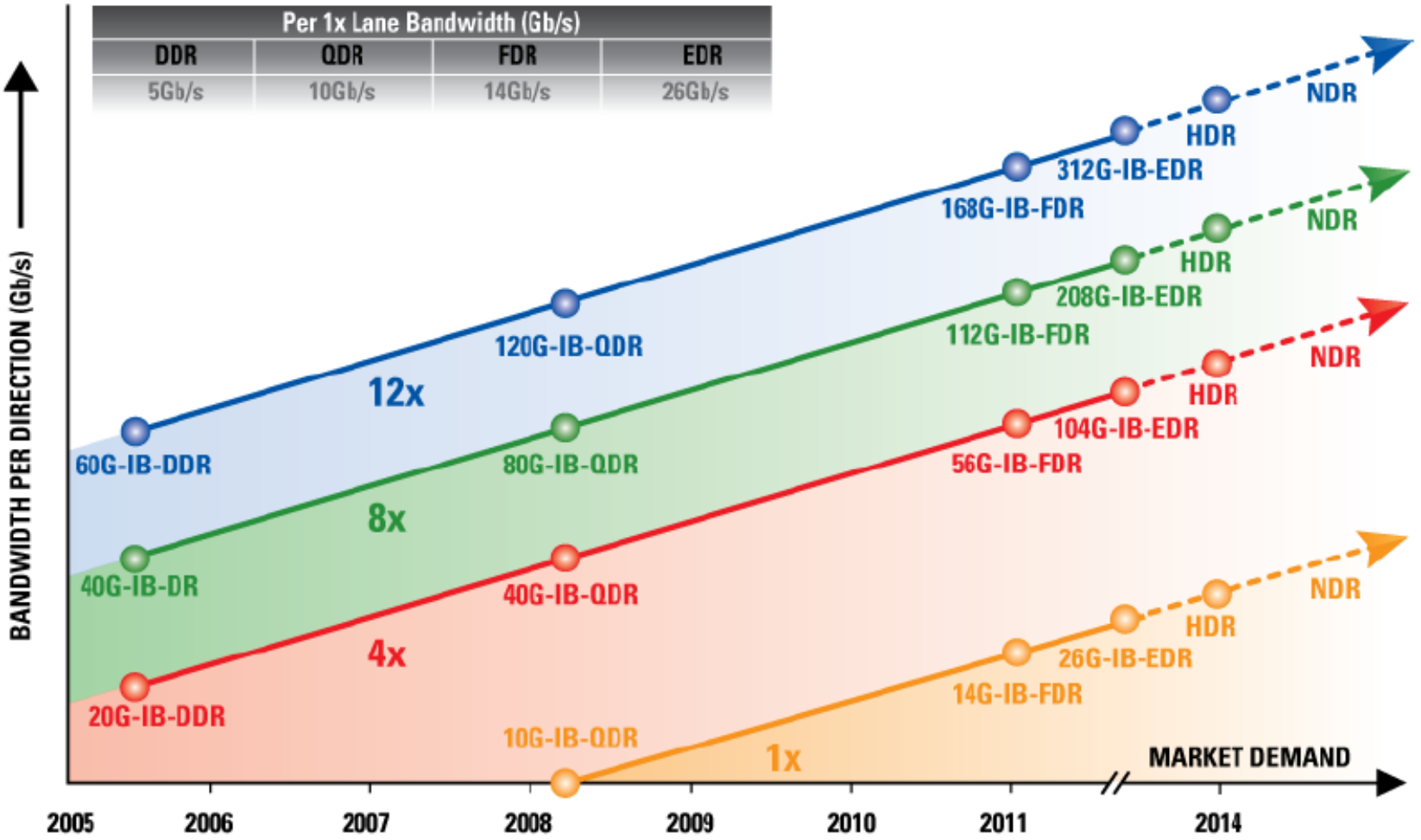
- **Improved DC networks are radically changing how data center apps run:**
  - Old style: "North / South" traffic: Each server handles 1 app for N desktop clients
    - Packets flowing into a data center go to specific servers, which sends packets back out.
  - New style: "East / West" traffic: N servers handle M apps as a virtualized pool for N clients
    - Packets flowing into a data center get flexibly directed to one of many servers, which generate *many* more server-to-server packets, and some packets go back out.
- **BW constraints (and *manageability* of traffic) still limit flexibility.**
- **Energy-efficient links are key – but higher-performance networks are more important**
  - High-BW links allow flexible placement of jobs & data ➔ high server utilization ➔ key benefit.





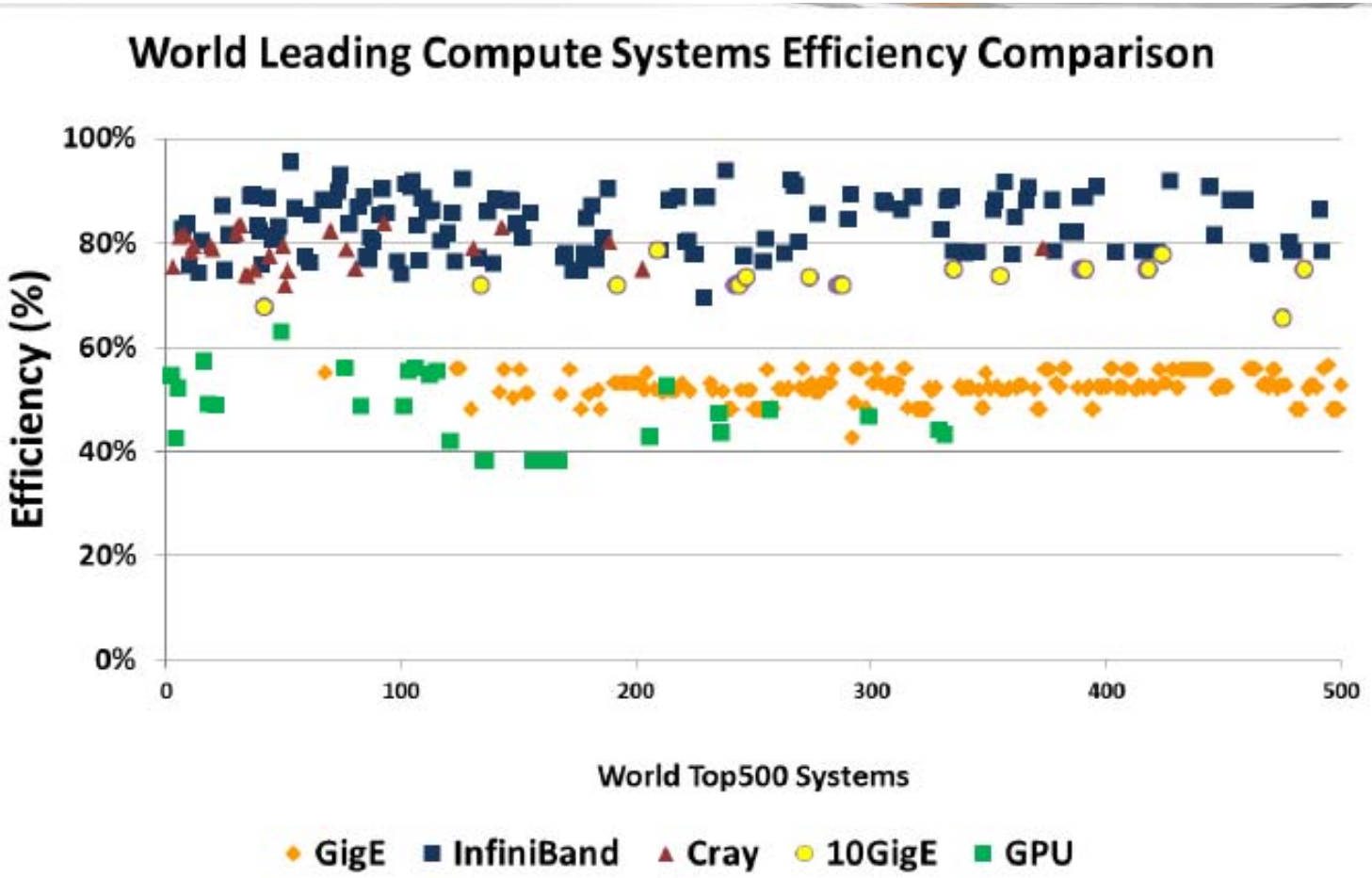
InfiniBand

InfiniBand Link Bandwidth Roadmap



- **56G-IB-FDR shipping now -- HCAs, switches, passive & active (copper & optical) cables**
  - Interoperability tested in Fall 2011 Plugfest
- **104G-IB-EDR expected in early 2013 – some cables demo'd already**

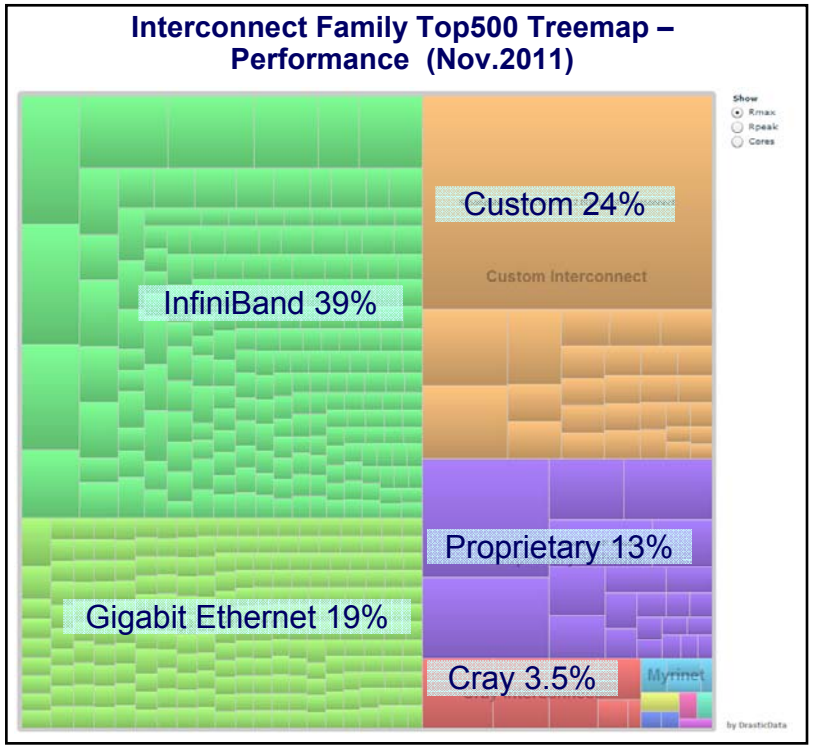
InfiniBand System Efficiency

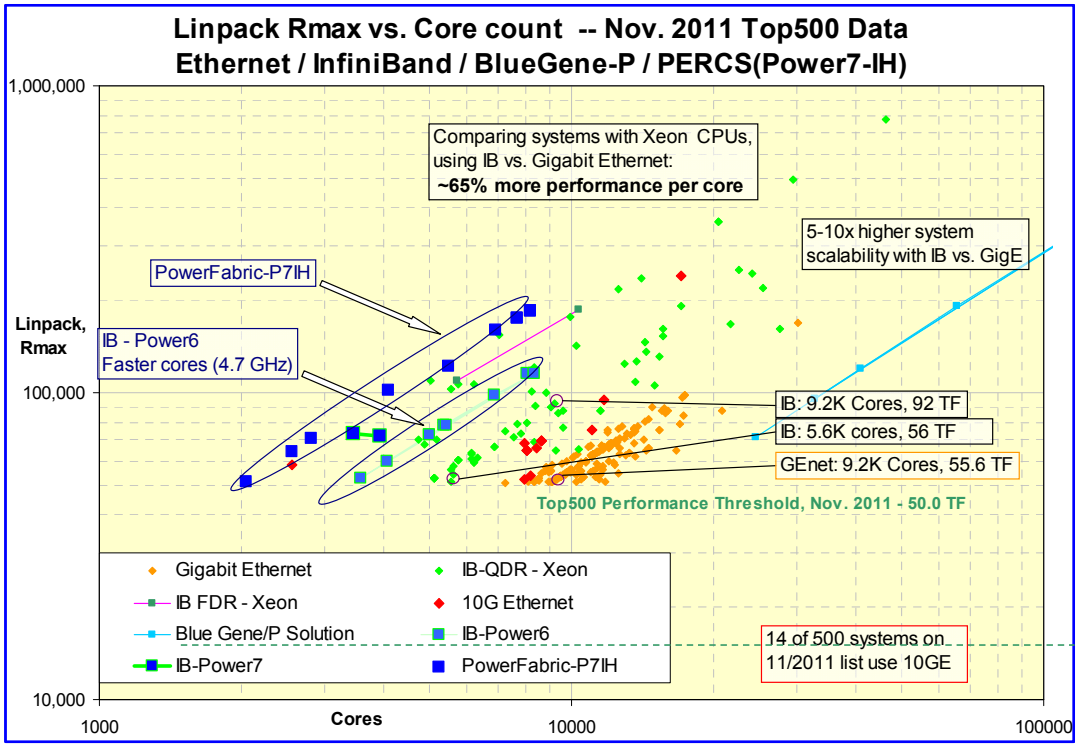


- TOP500 systems listed according to their efficiency
- InfiniBand is the key element responsible for the highest system efficiency
- Up to 96% efficiency

Brian Sparks *IBTA Marketing Working Group Co-Chair*

Top500: Impact of Interconnect on System Scaling





- **Left: Analysis of Top500 systems in terms of Interconnect Family.**
  - Majority of processing power is interconnected with InfiniBand interconnect
  - 2011: Custom & Proprietary Interconnects grew greatly – greater system-level requirements.

- **Right: Impact of Interconnect on System Cost/ Performance**
  - Switching from Gigabit Ethernet to InfiniBand allows either 65% fewer servers, or 65% better performance with same system size (on Linpack benchmark)

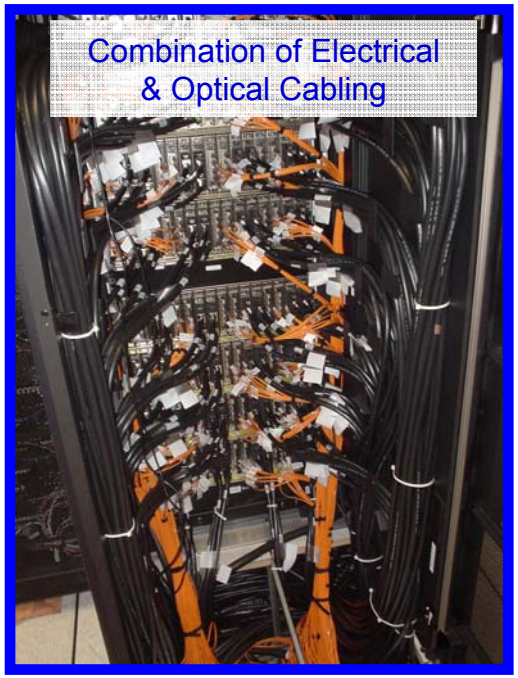
HPC Systems Networking

Rack-to-rack cabling: Recent history in HPC systems

2002: 40 TF/s

NEC Earth Simulator
- **all copper, ~1 Gb/s**

2005



IBM Federation Switch for ASCI Purple (LLNL)
- Copper for short-distance links (≤10 m)
- Optical for longer links (20-40m)

~3000 parallel links 12+12@2Gb/s/channel

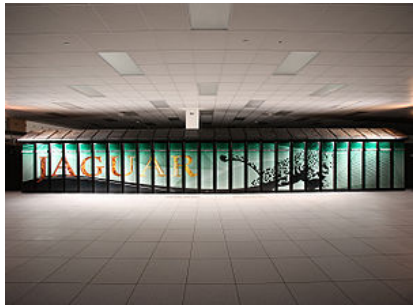
2008: 1 PF/s

IBM Roadrunner (LLNL)

*http://www.lanl.gov/roadrunner/

- 4X DDR InfiniBand (5Gb/s)
- 55 miles of Active Optical Cables

Cray Jaguar(ORNL)



*http://www.nccs.gov/jaguar/

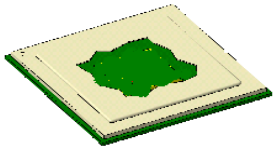
- InfiniBand
- 3 miles of optical cables, longest = 60m

- **Over time: higher bit-rates, similar lengths, more use of optics, denser connector packing**

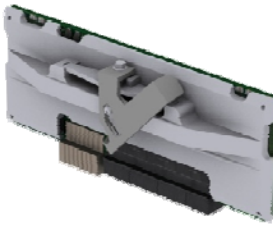
Blue Gene/Q

1. Chip:
16+2 μP
cores

2. Single Chip Module



3. Compute card:
One chip module,
16 GB DDR3 Memory,
Heat Spreader for H_2O Cooling



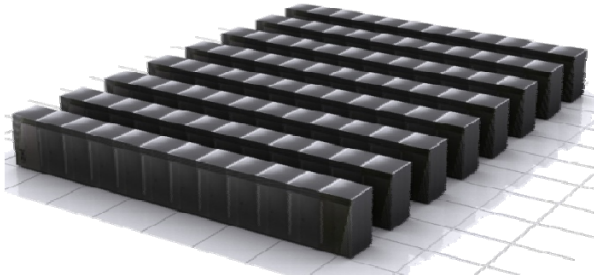
4. Node Card:
32 Compute Cards,
Optical Modules, Link Chips; 5D Torus

5b. IO drawer:
8 IO cards w/16 GB
8 PCIe Gen2 x8 slots
3D I/O torus

5a. Midplane:
16 Node Cards

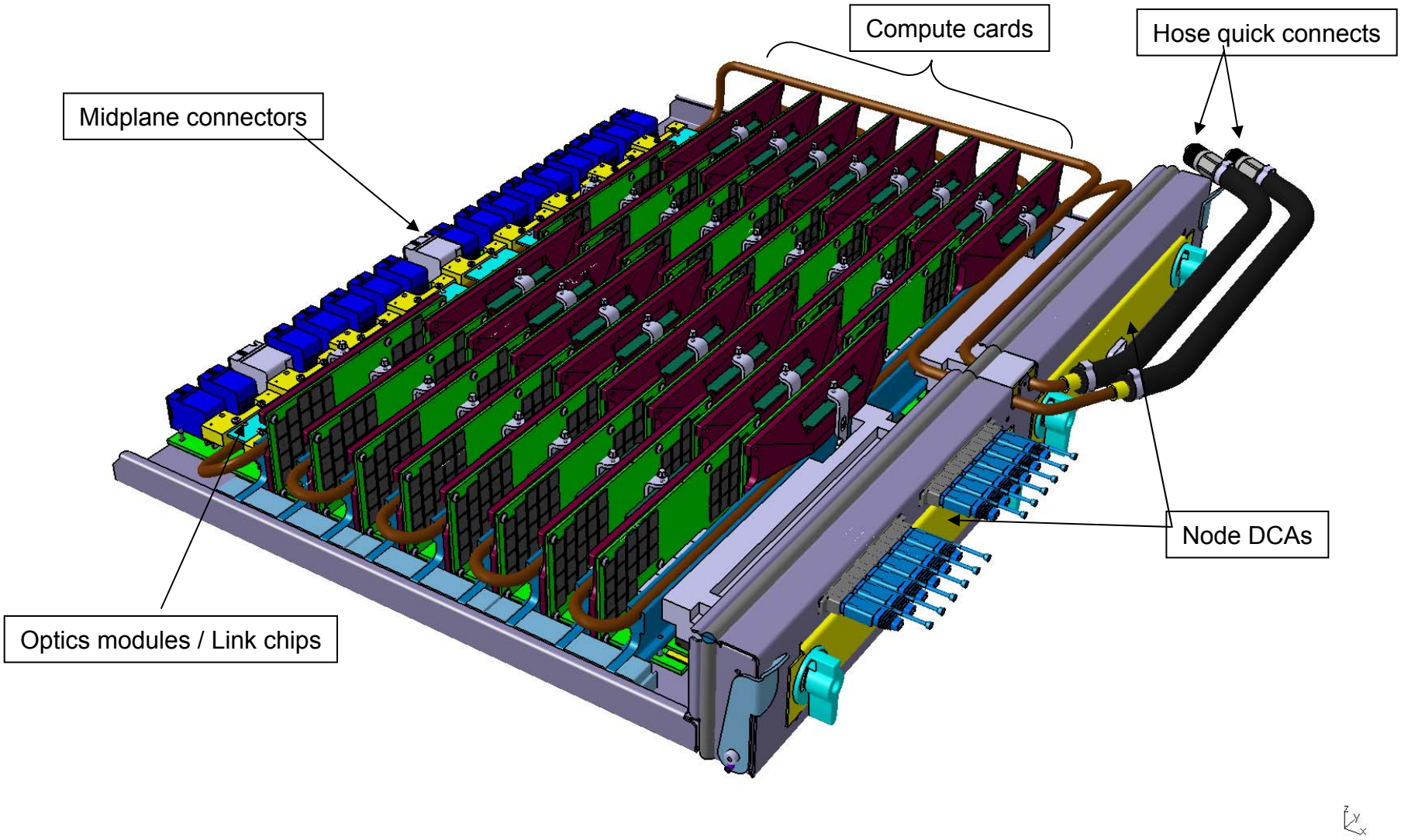
6. Rack: 2 Midplanes

7. System:
96 racks, 20PF/s

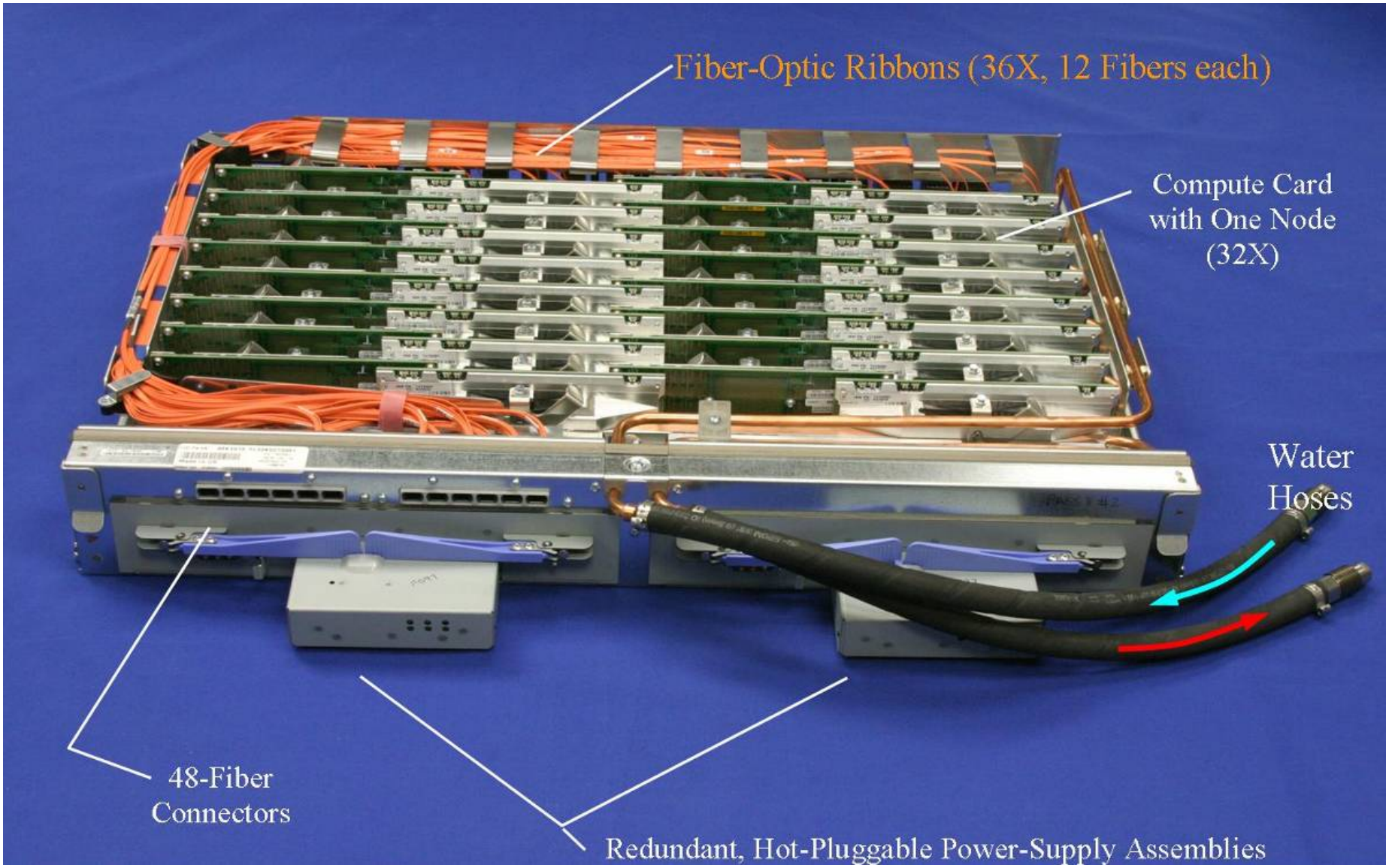


•Sustained single node perf: 10x P, 20x L
• MF/Watt: (6x) P, (10x) L (~2GF/W, Green 500 criteria)
• Software and hardware support for programming models for exploitation of node hardware concurrency

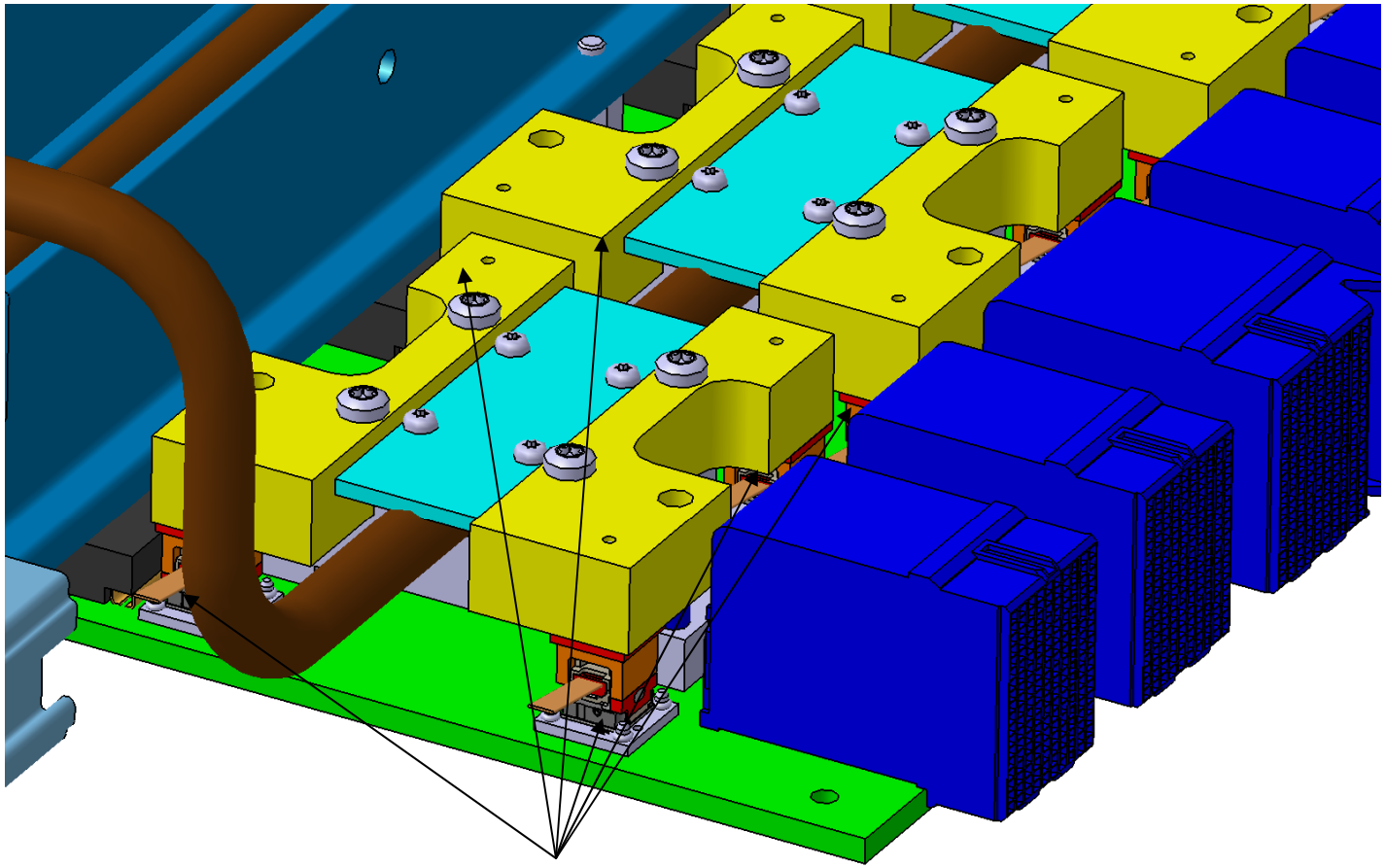
BG/Q Compute Drawer – Technical Drawing



BG/Q Compute Drawer

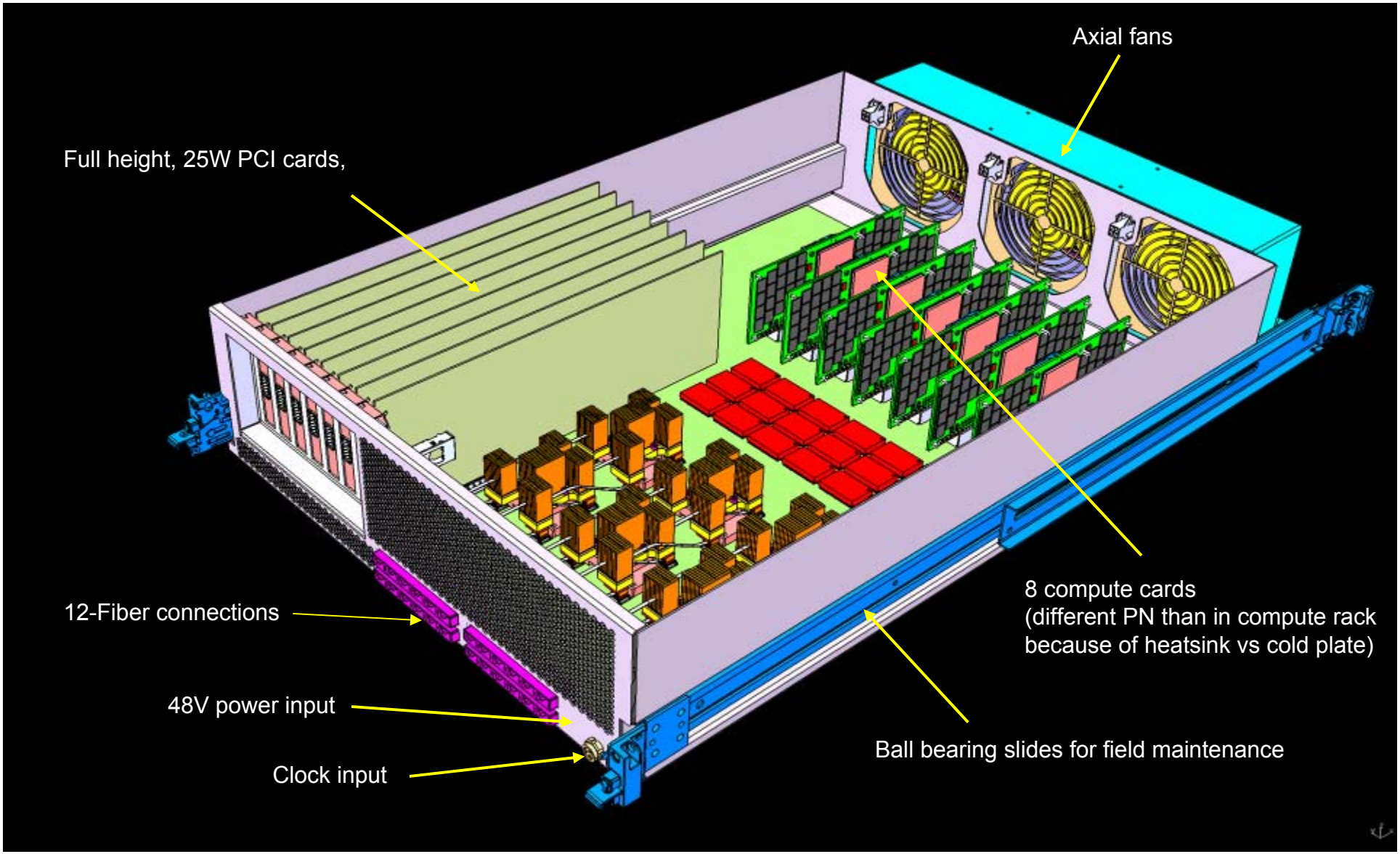


Compute Drawer – Rear Isometric View, showing optics modules

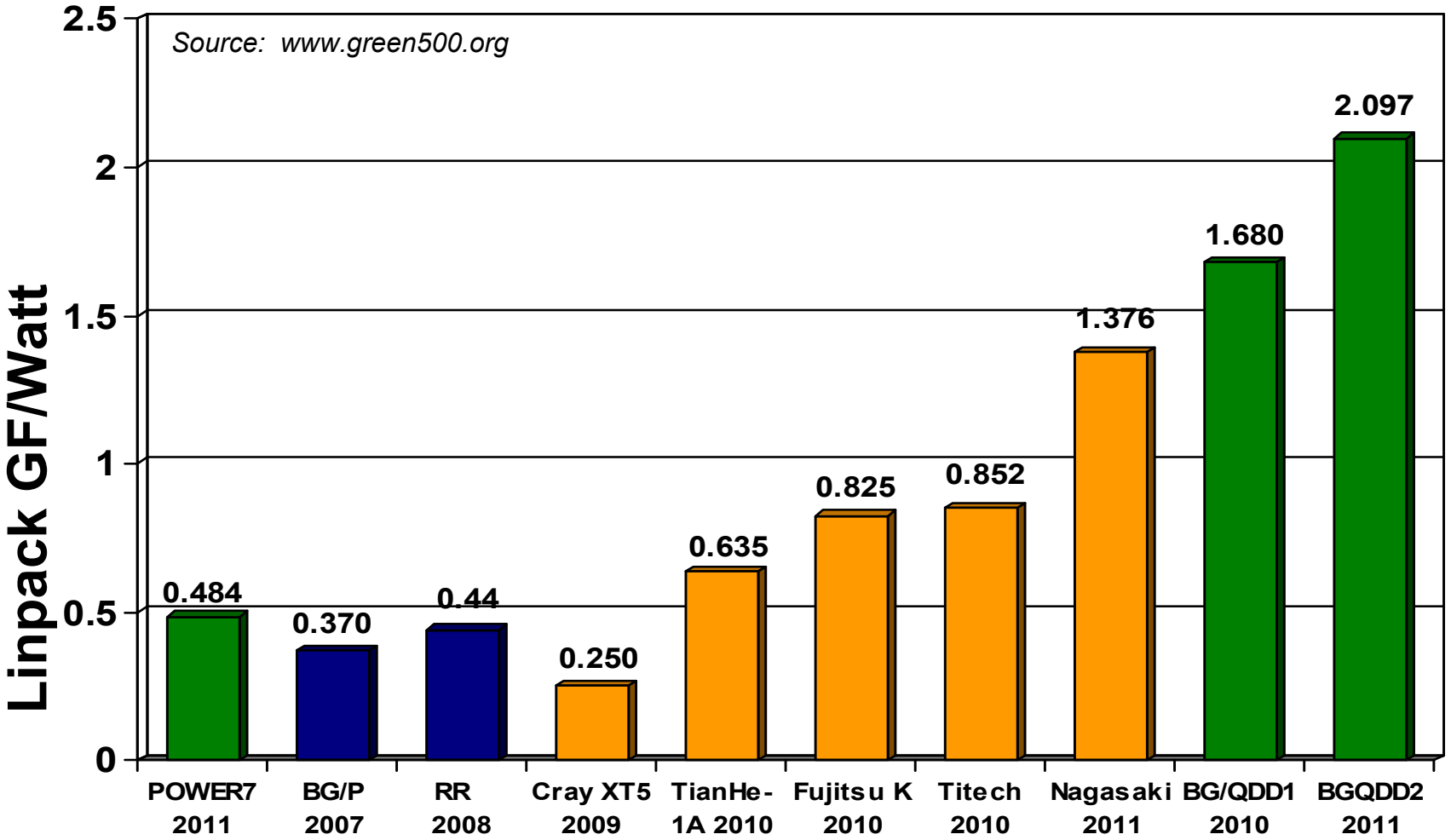


Optics modules placed in sockets
(mechanically retained by features in socket)

BG/Q Input/Output Drawer



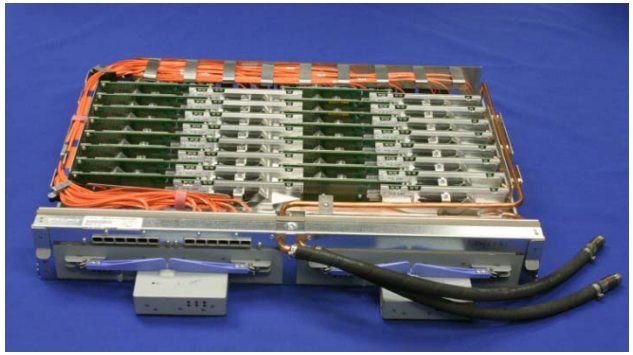
System Power Efficiency (Green500 06/2011)



At $.10/kWh => 1MW savings in power saves $1M/year. TCO saving is much more.
Low power is key to scaling to large systems

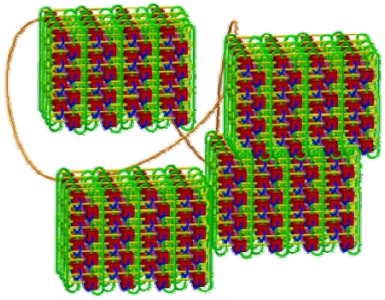
Blue Gene/Q

Industrial Design



32 Node Board

BQC DD2.0



5D torus

4-rack system

PERCS/Power 775 "Data-Center-In-A-Rack" System Architecture

- **All data center power & cooling infrastructure included in compute/storage/network rack**
  - No need for external power distribution or computer room air handling equipment.
  - All components correctly sized for max efficiency – very good 1.18 Power Utilization Efficiency
  - Integrated management for all compute, storage, network, power, & thermal resources.
  - Scales to 512K P7 cores (192 racks) – without any other hardware except optical fiber cables

Integrated Power Regulation, Control, & Distribution

Runs off any building voltage supply world-wide (200-480 VAC or 370-575VDC), converts to 360 VDC for in-rack distribution. **Full in-rack redundancy and automatic fail-over**, 4 power cords. Up to 252 kW/rack max / 163 kW Typ.

Integrated Storage – 384 2.5" HDD or SSD drives /drawer

230 TBytes\drawer (w/600 GB 10K SAS disks), 154 GB/s BW/drawer, software-controlled RAID, up to 6/rack (replacing server drawers) (up to **1.38 PBytes / rack**)

Servers – 256 Power7 cores / drawer, 1-12 drawers / rack

Compute: 8-core Power7 CPU chip, 3.7 GHz, 12s technology, 32 MB L3 eDRAM/chip, 4-way SMT, 4 FPUs/core, Quad-Chip Module; **>90 TF / rack**

No accelerators: normal CPU instruction set, robust cache/memory hierarchy

Easy programmability, predictable performance, mature compilers & libraries

Memory: 512 Mbytes/sec per QCM (0.5 Byte/FLOP), **12 Terabytes / rack**

External IO: **16 PCIe Gen2 x16 slots / drawer**; SAS or external connections

Network: Integrated Hub (HCA/NIC & Switch) per each QCM (8 / drawer), with 54-port switch, including total of 12 Tbits/s (1.1 TByte/s net BW) per Hub:

Host connection: 4 links, (96+96) GB/s aggregate (0.2 Byte/FLOP)

On-card electrical links: 7 links to other hubs, (168+168) GB/s aggregate

Local-remote optical links: 24 links to near hubs, (120+120) GB/s aggregate

Distant optical links: 16 links to far hubs (to 100M), (160+160) GB/s aggregate

PCI-Express: 2-3 per hub, (16+16) to (20+20) GB/s aggregate

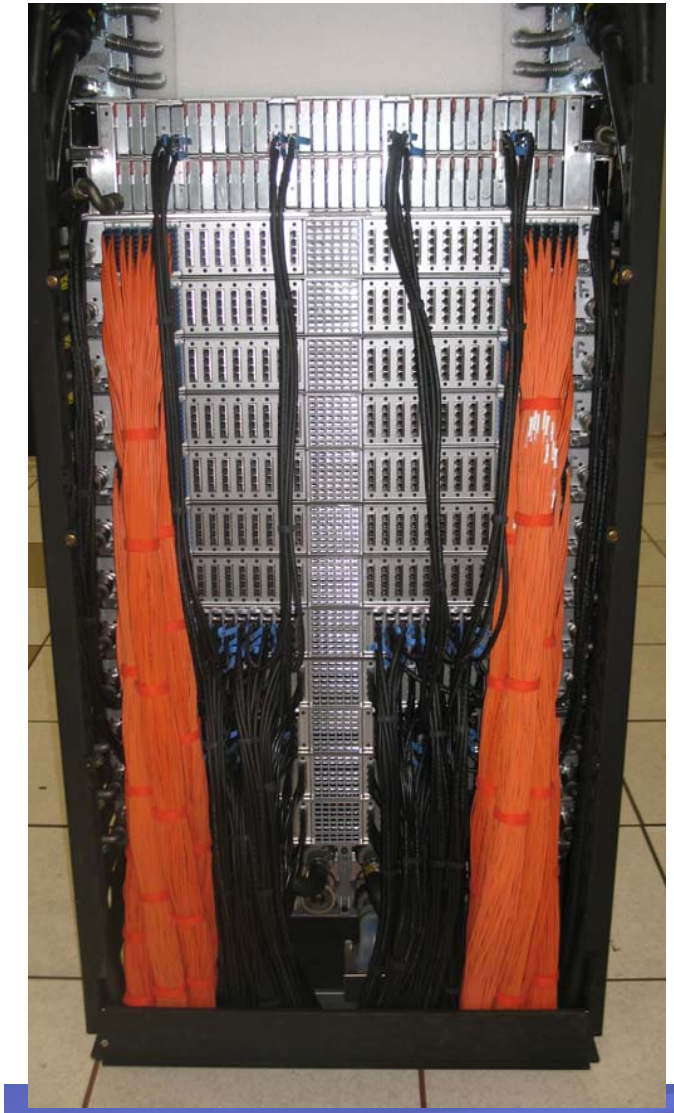
Integrated Cooling – Water pumps and heat exchangers

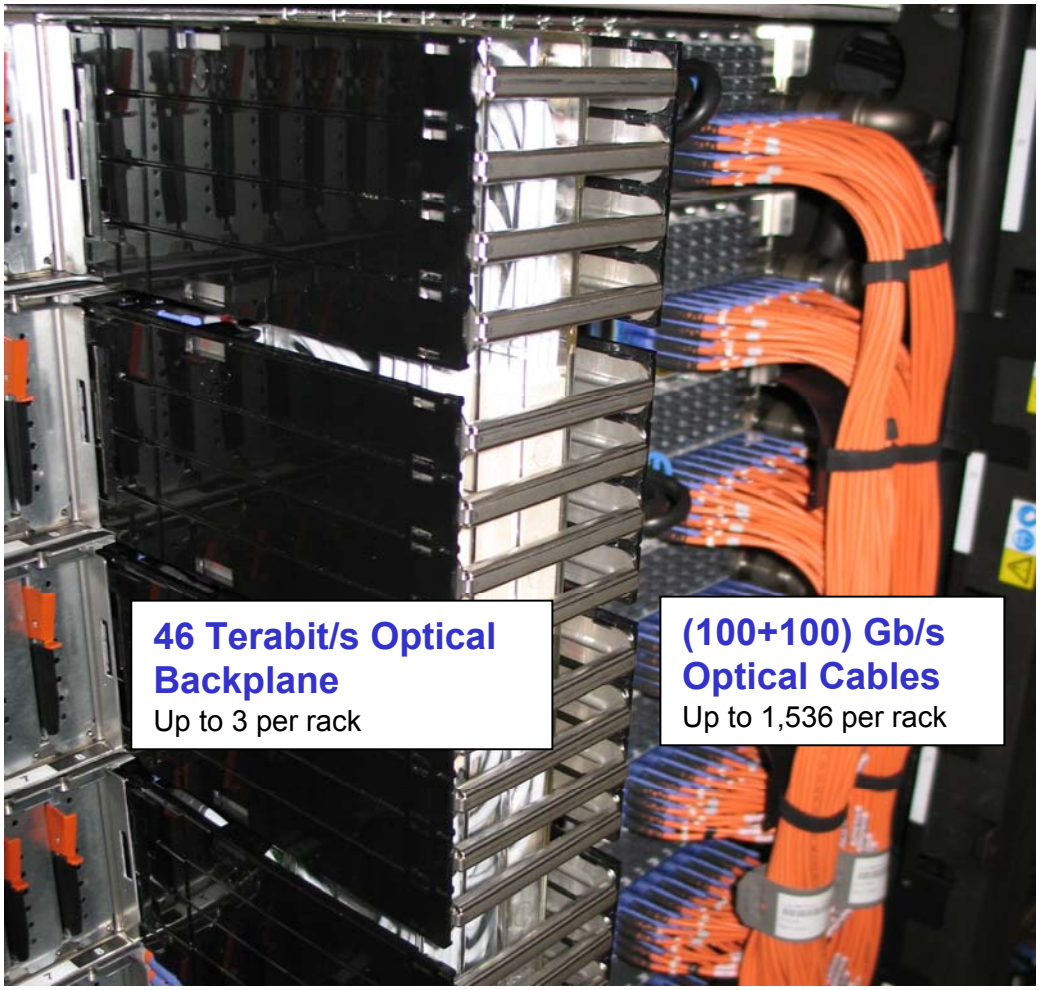
All heat transferred directly to building chilled water – **no thermal load on room**

33

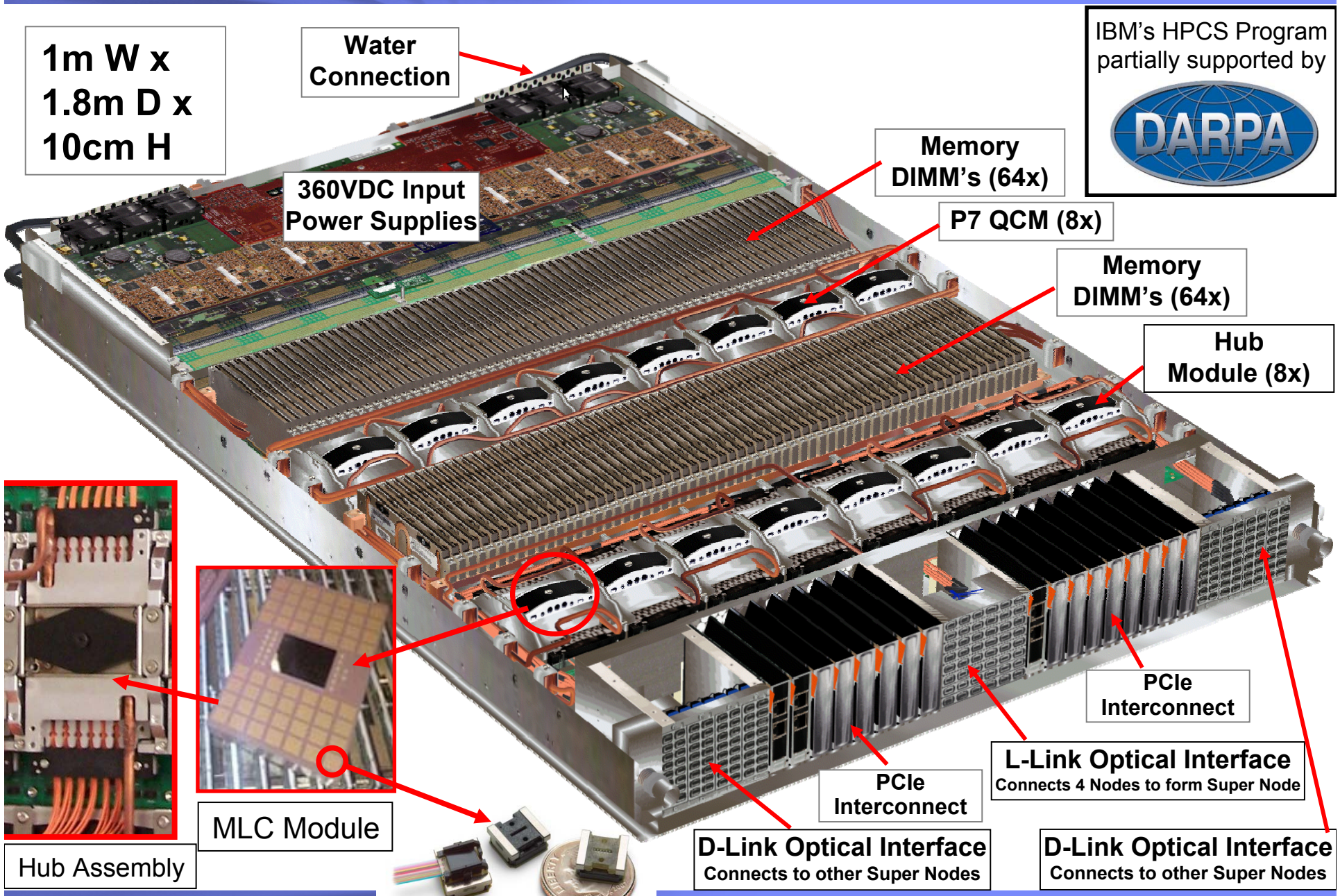
P7-IH – Cable Density

- **Many many optical fibers**
  - Each of these cables is a 24-fiber multimode cable, carrying (10+10) GBytes/sec of traffic





P7 IH System Hardware – Node Front View (Blue Waters: ~1200 Node drawers)

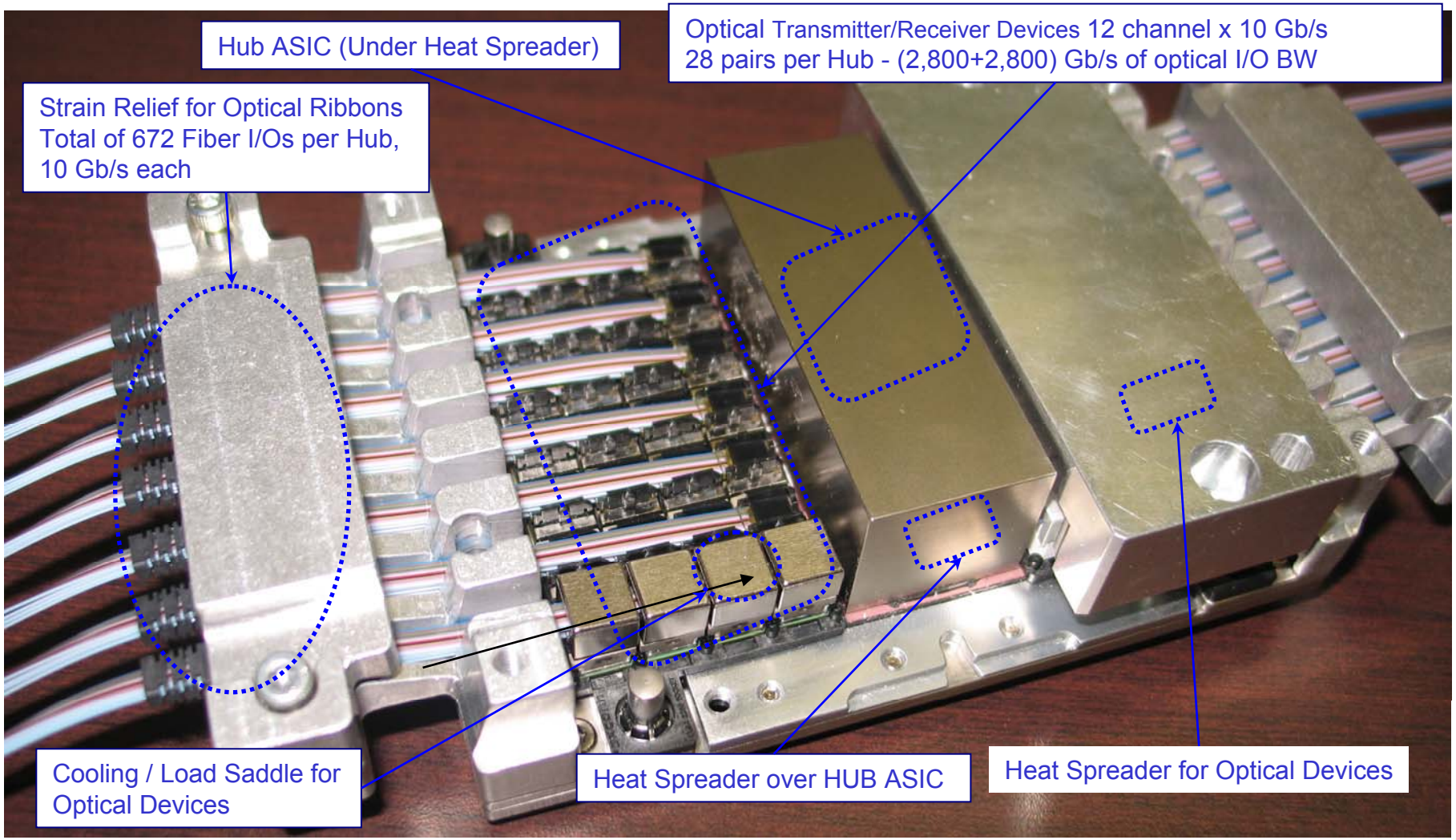


All off-node communication optical

35

Hub Module – MCM with Optical I/Os

- **This shows the Hub module with full complement of Optical I/Os.**
  - Module in photo is partially assembled, to show construction – full module HW is symmetric



Overview: Recent strategic directions in IBM Research

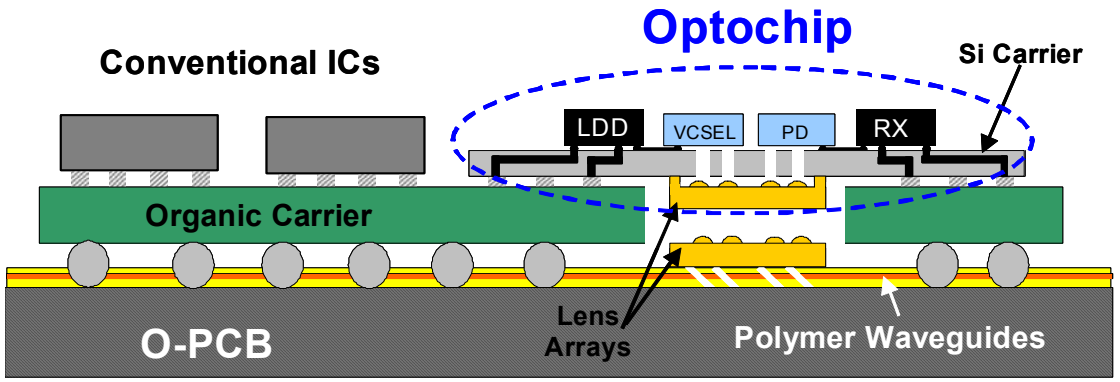
IBM Optical Interconnect Research: Meeting Key Challenges for Optical Links

- **Increasing aggregate system performance will demands more optical links**
  - Bandwidth demands steadily increasing → higher channel rates, more parallel channels
  - Optical link budgets substantially more challenging at higher data rates
  - Density requirements becoming increasingly important as number of links in systems grows

- **IBM Research has active programs in a variety of areas of optical interconnect**
  - Transceiver Opto-Mechanical Design – Advanced Packaging, 3D Chip-Stacking and silicon carriers, Through silicon optical vias.
    - Example: 24 + 24 channel highly integrated transceivers
  - Optical PCBs – Polymer Optical Waveguides, both above and in PCBs
  - Advanced Circuit Design in SiGe & CMOS Drivers & Receivers
    - Example: >30Gb/s SiGe links, 25 Gb/s CMOS links
    - Optical Transmitter Equalization for better link margin, jitter, power efficiency
  - Silicon Photonics

24-channel 850-nm transceivers packaged on Si carriers

- **850-nm is the datacom industry standard wavelength**
  - Multiple suppliers, low-cost, optimized MMF fiber bandwidth
- **Retain the highly integrated packaging approach: dense Optomodules that "look" like surface-mount electrical chip carriers**
- **Si carrier platform: high level of integration of the electrical and optical components with high density interconnection, requires through-silicon-vias (both optical and electrical)**

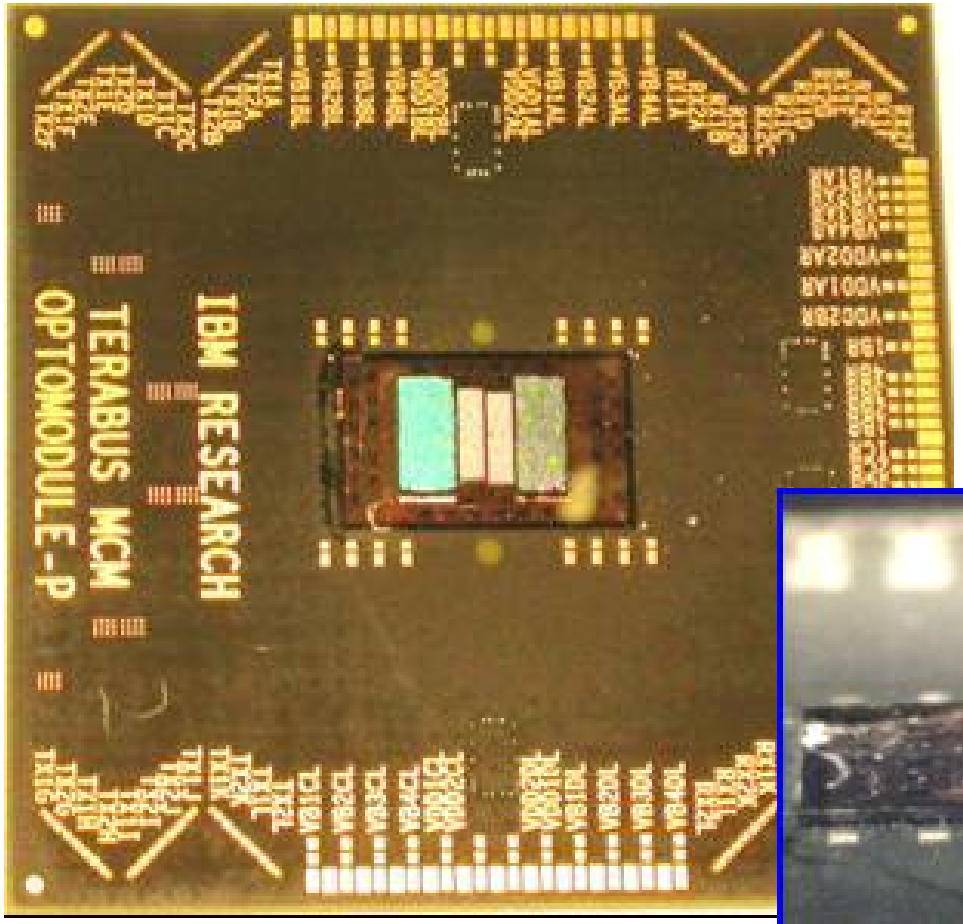


Optically enabled MCM (OE-MCM)

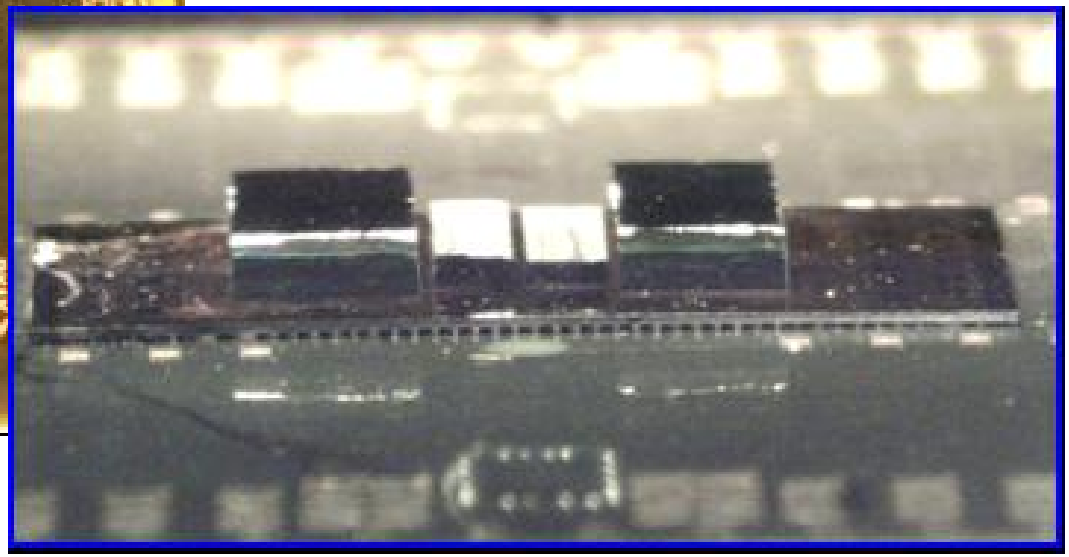
Terabus 850 nm

- 24TX + 24 RX Transceiver
  - 2x12 VCSEL and PD arrays
  - 2 130nm CMOS ICs
- **TSV Si carrier**
  - Optical vias in Si carrier
  - Side-by-side flip chip assembly

Assembled 24-channel 850-nm modules for optical PCB links

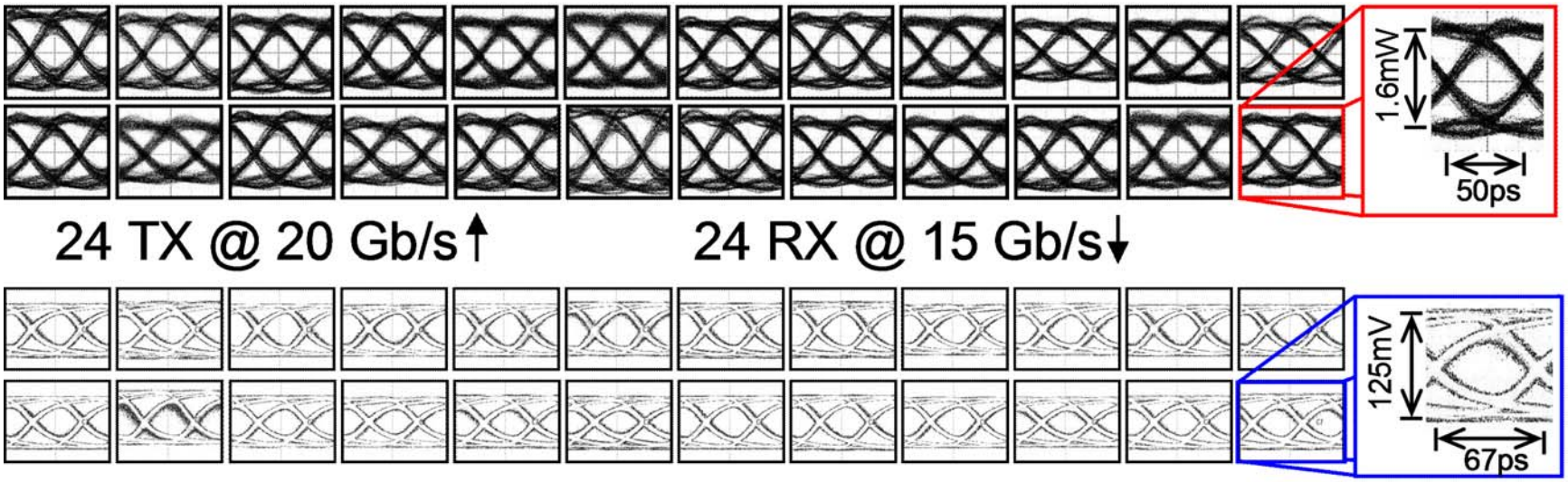


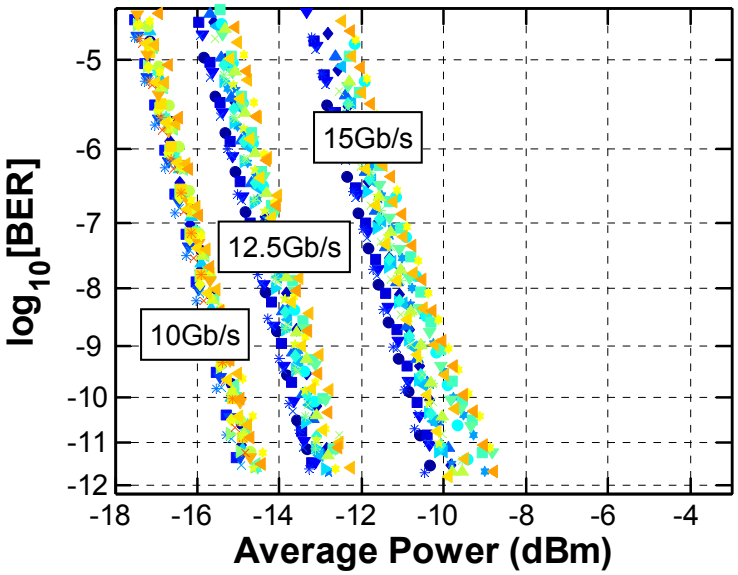
- Flip-chip assembly of OE and CMOS chips to Si-carrier using AuSn solder “micro bumps”
- Flip-chip attachment of Si-carrier Optochip to organic carrier using PbSn solder transfer process



First row of solder joins visible beneath the Optochip

360Gb/s, 24-channel, 850-nm Transceiver Modules Demonstrated

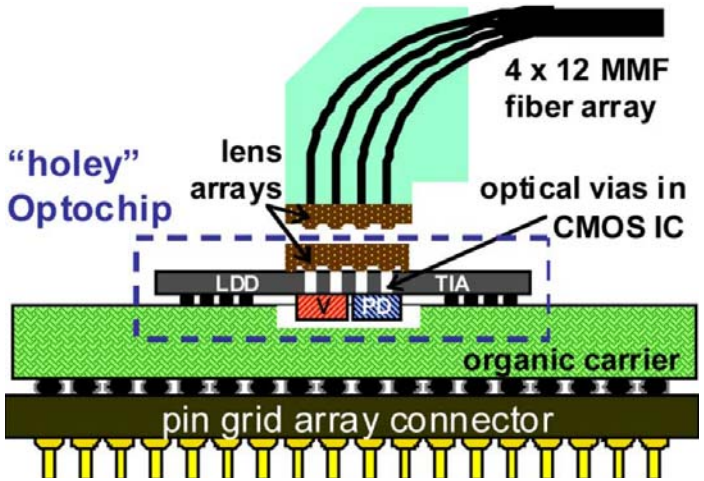




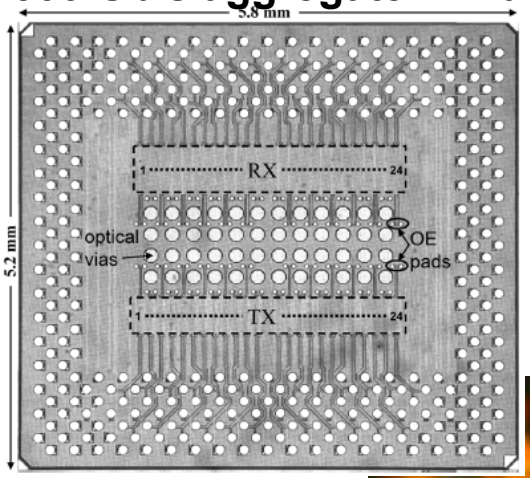
- **Highest aggregate bandwidth for any 850-nm parallel optical module: 360 Gb/s bi-directional**
- **Power efficiency < 10 pJ/bit**

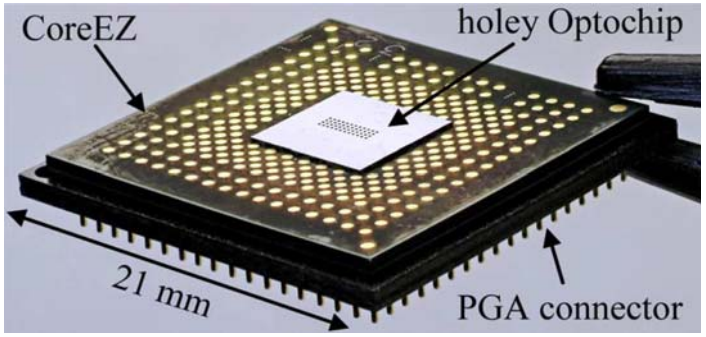
- F. E. Doany *et al.*,"Terabit/s-Class 24-Channel Bidirectional Optical Transceiver Module Based on TSV Si Carrier for Board-Level Interconnects," ECTC 2010, June 2010.

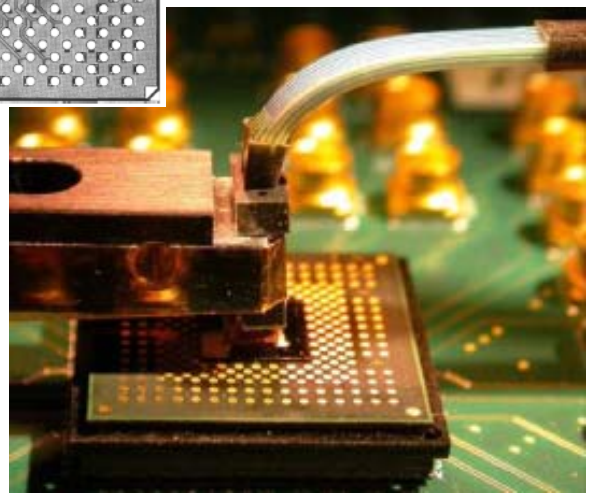
"Holey" Optochip –CMOS IC with optical though-silicon-vias



- **(24+24)x12.5 Gbps single-chip transceiver**
- **Flip-chip mounting of VCSELs & PDs directly on driver/receiver circuits**
- **300 Gb/s aggregate BW at 8.2 pJ/bit,**

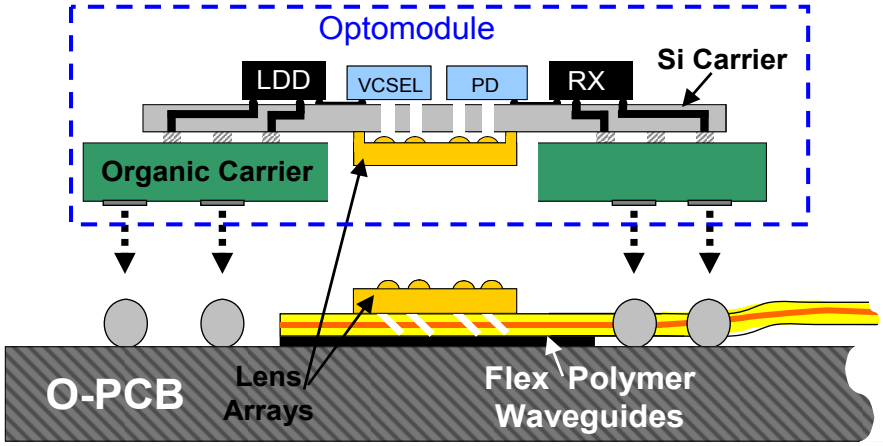


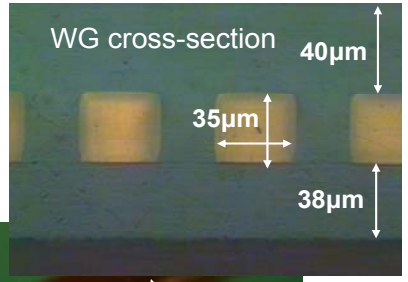


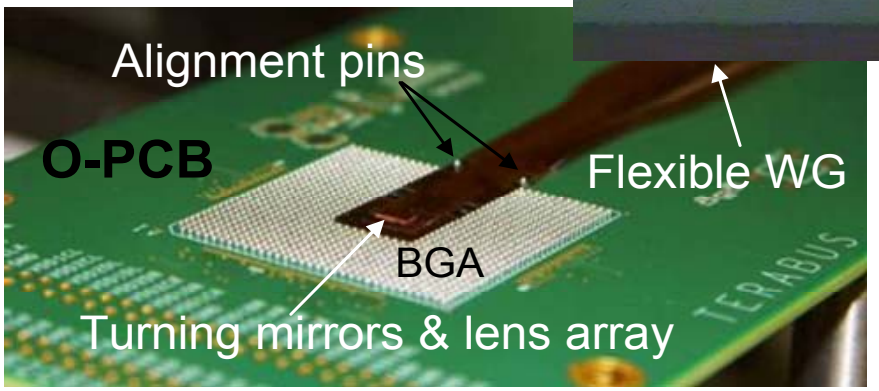


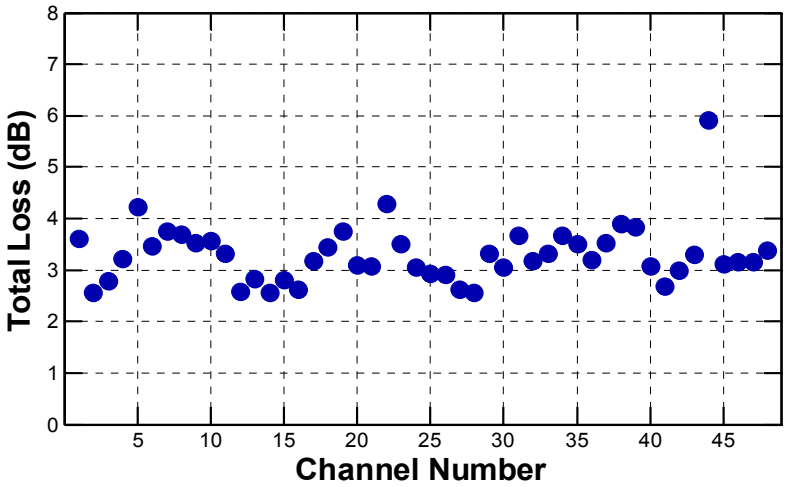
- C. L. Schow, *et al.*,"A 24-Channel, 300 Gb/s, 8.2 pJ/bit, Full-Duplex Fiber-Coupled Optical Transceiver Module Based on a Single "Holey" CMOS IC," J. Lightwave Tech., Vol. 29, No. 4, Feb. 2011.

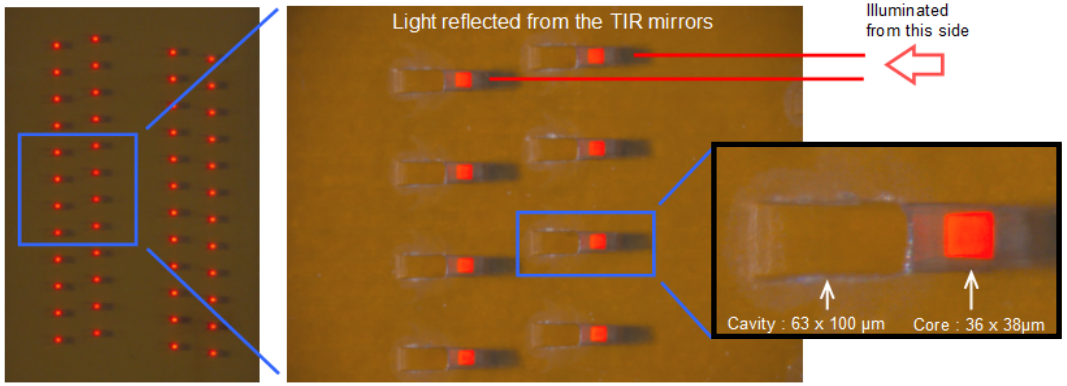
o-PCB preparation and assembly







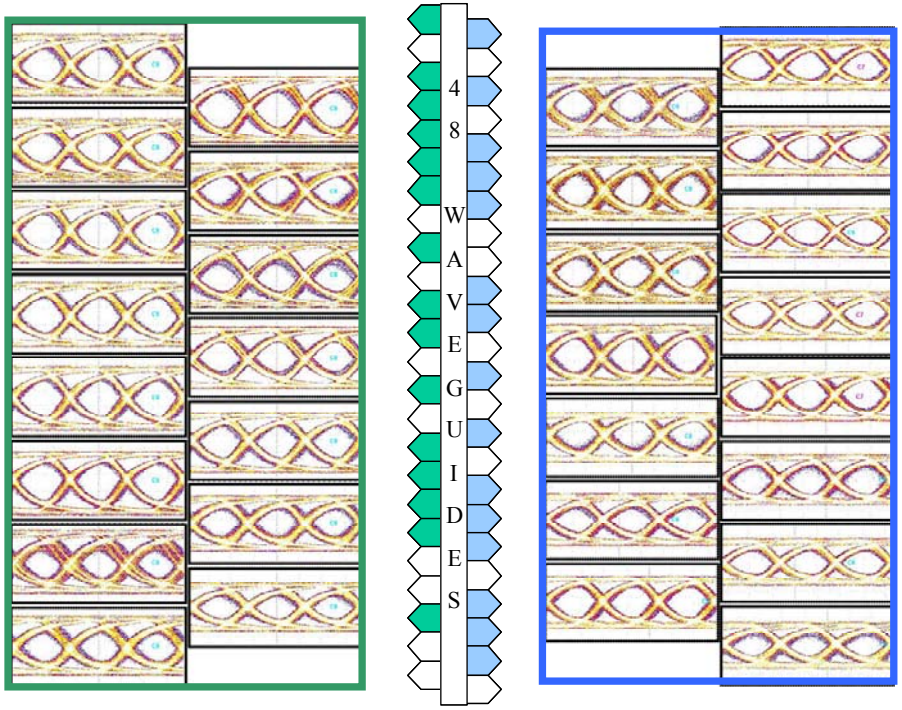




- **45° turning mirrors formed by laser ablating air cavities in the WGs**
  - Total internal reflection (TIR) mirrors, 0.5-0.7 dB loss
- **48 element WG lens arrays aligned to the flex WG**
- **WG flex attached to PCB with pre-deposited BGA solder balls**

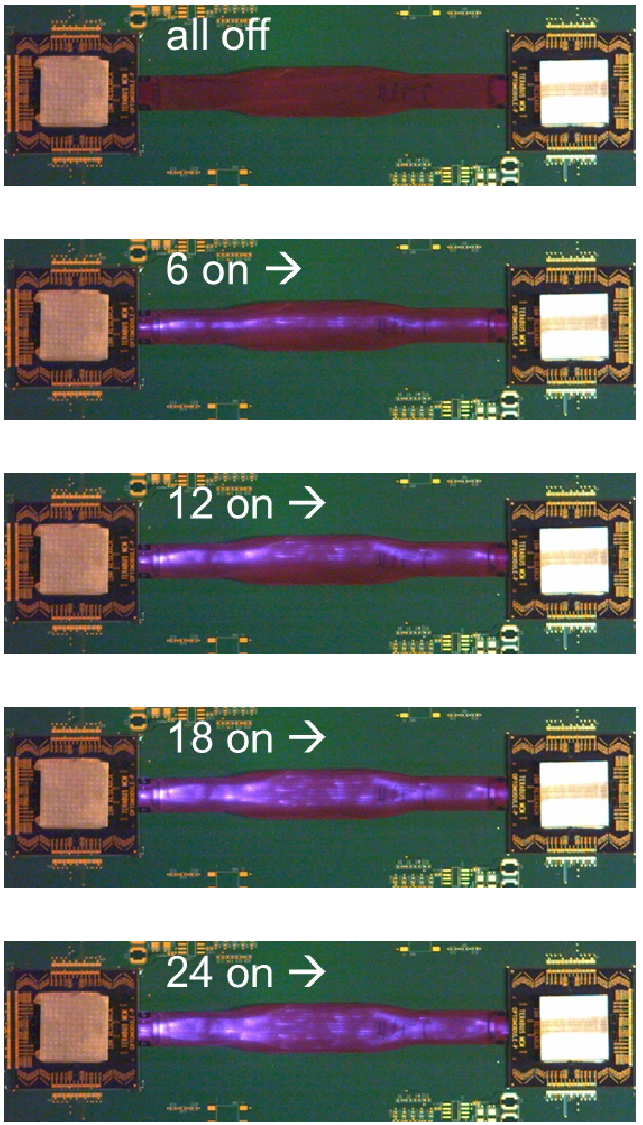
850-nm Optical PCB in Operation

15 Gb/s



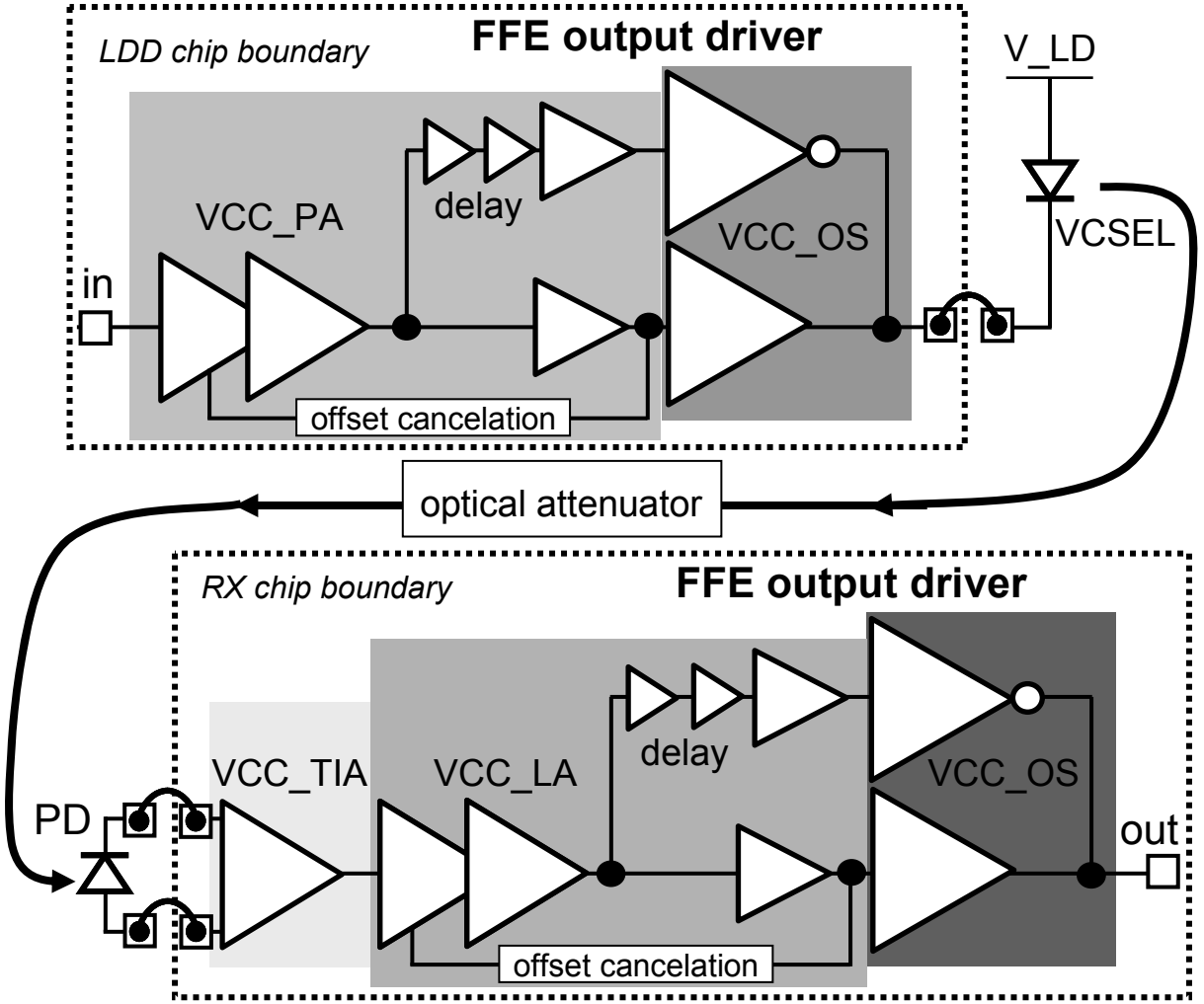
15 + 15 channels

- **15 channels each direction at 15 Gb/s, BER < 10^{-12}**
- **225 Gb/s bi-directional aggregate**
- **145 mW/link = 9.7 pJ/bit**



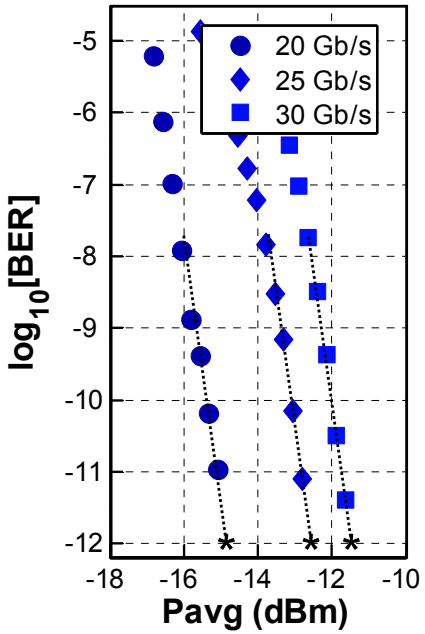
- F. E. Doany *et al.*,"Terabit/s-class board-level optical interconnects through polymer waveguides using 24-channel bidirectional transceiver modules," ECTC 2011 June 2011.
- C. L. Schow *et al.*, "225 Gb/s bi-directional integrated optical PCB link," OFC 2011, post-deadline paper, Mar. 2011.

SiGe 8HP: Pushing the Speed Limits of VCSEL Links

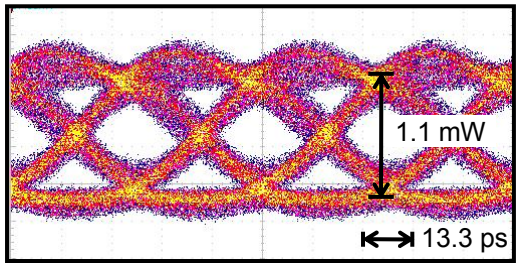


- **FFE circuit included in TX output for VCSEL pre-distortion/pre-emphasis and in RX output to drive through packages and boards**

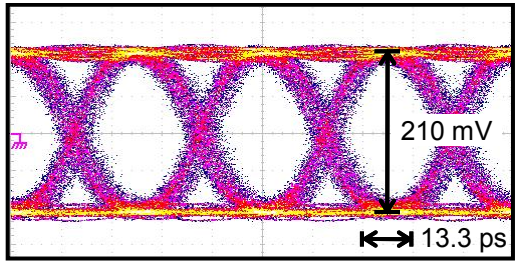
Record SiGe 8HP full-link: 30 Gb/s using 10Gb/s OEs

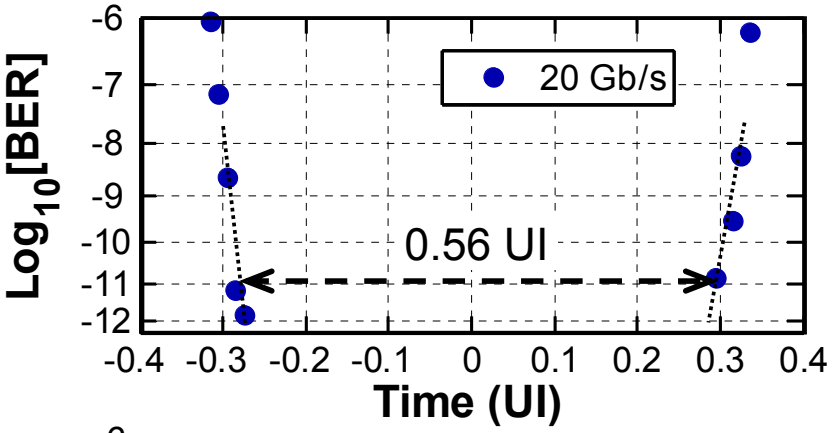


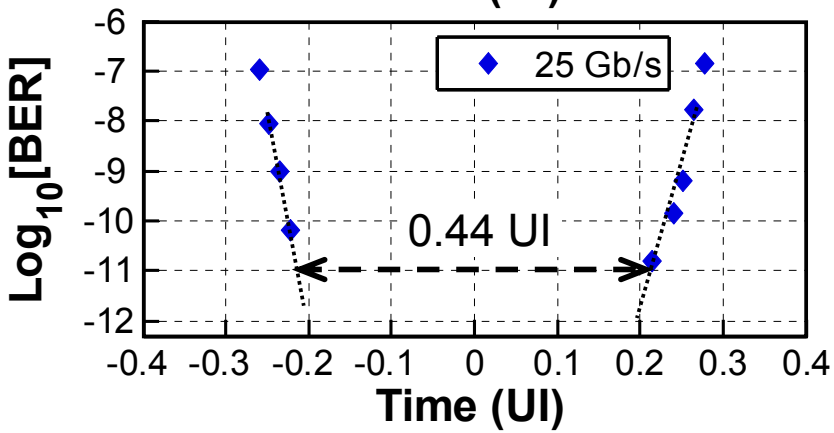
VCSEL output, 30 Gb/s



RX output, 30 Gb/s



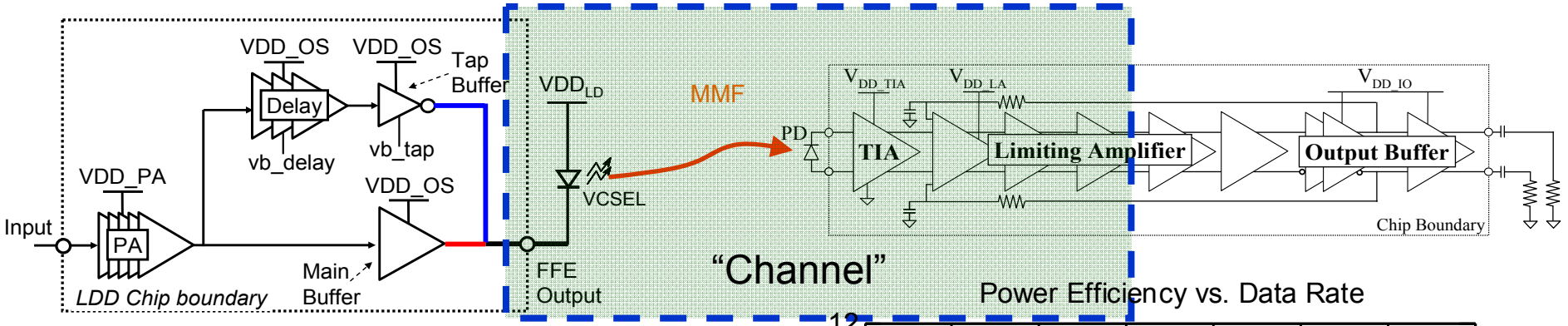




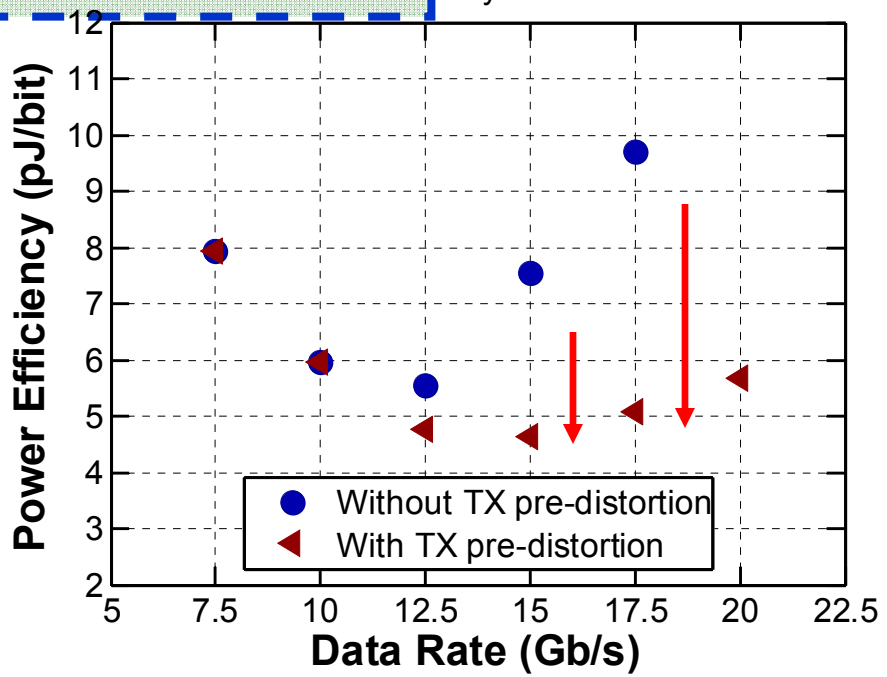
- **First 30Gb/s VCSEL based link**
- **10 Gb/s VCSELs**
- **Applications for multimode reference receiver**
- **Novel TIA design**
- **Operates with margin at 30G**
- **100m transmission with minimal penalty verified at 25 Gb/s**

- C. L. Schow and A. V. Rylyakov, "30 Gbit/s, 850 nm, VCSEL-based optical link," *Electron. Lett.*, September 1, 2011.

Applying Signal Processing to Low Power Optical Links



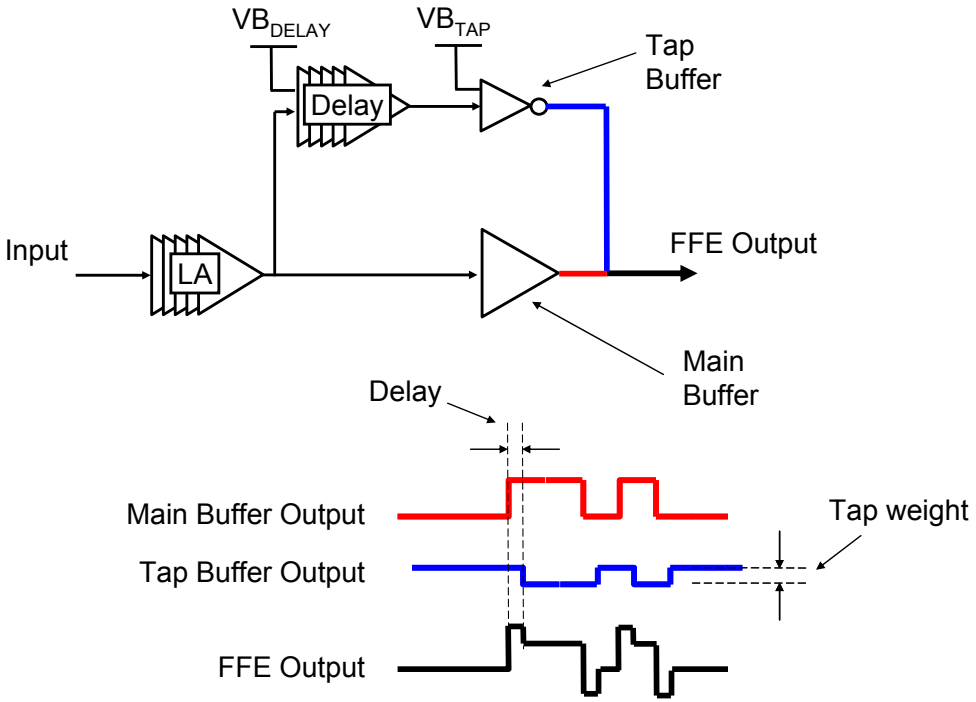
- Electrical links have increasingly used signal processing to improve performance…
  - optics can do this too!
- Pre-distortion compensation for combined VCSEL/TIA and LA:
  - Increases obtainable link speed to 20Gb/s
  - 5.7pJ/bit total link power consumption while maintaining BER < 10^{-12} and >200mV$_{ppd}$ at RX outputs

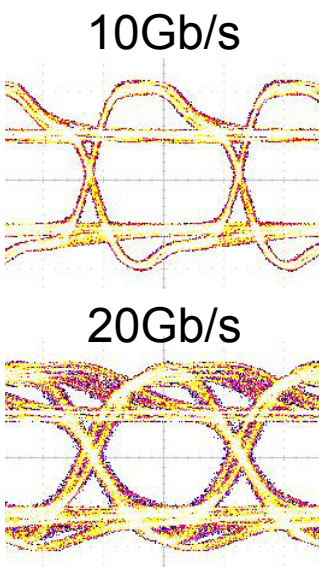


- C. L. Schow *et al.* "Transmitter pre-distortion for simultaneous improvements in bit-rate, sensitivity, jitter, and power efficiency in 20 Gb/s CMOS-driven VCSEL links," *OFC 2011*, post deadline paper, Mar. 2011.

FFE Equalizers for Both TX and RX Outputs

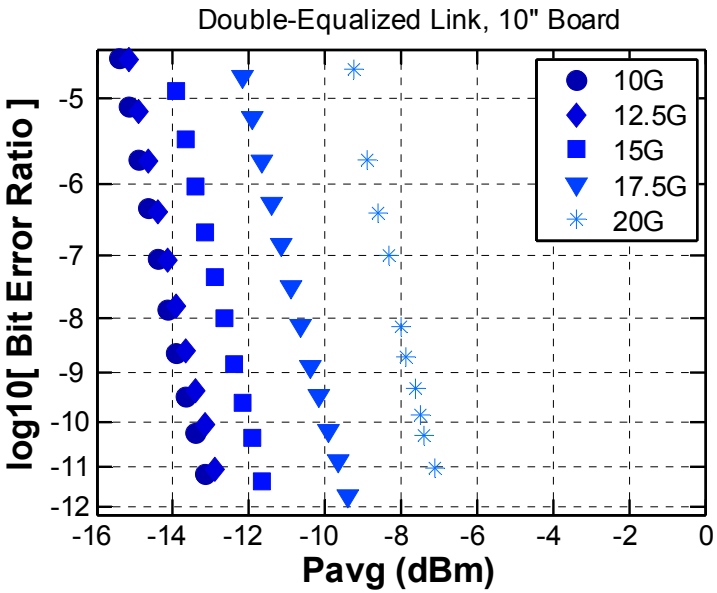
Feed-Forward Equalizer (FFE) circuit for adjustable output pre-emphasis

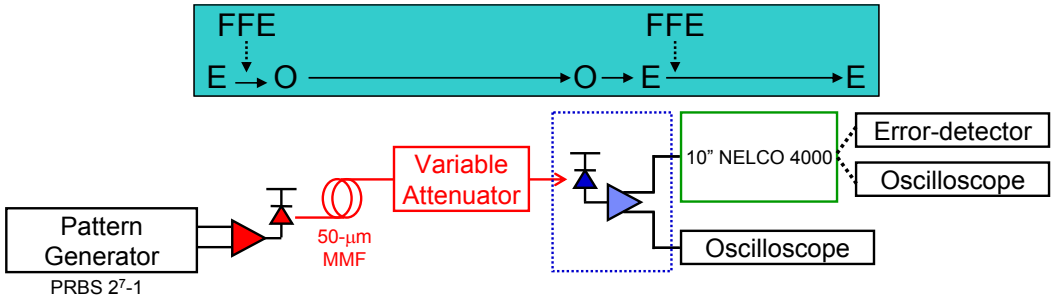




- **Feed-Forward Equalizer (FFE) design leveraging extensive electrical serial link design**
- **Equalization heavily applied to VCSEL outputs for improved link performance → first demonstration**

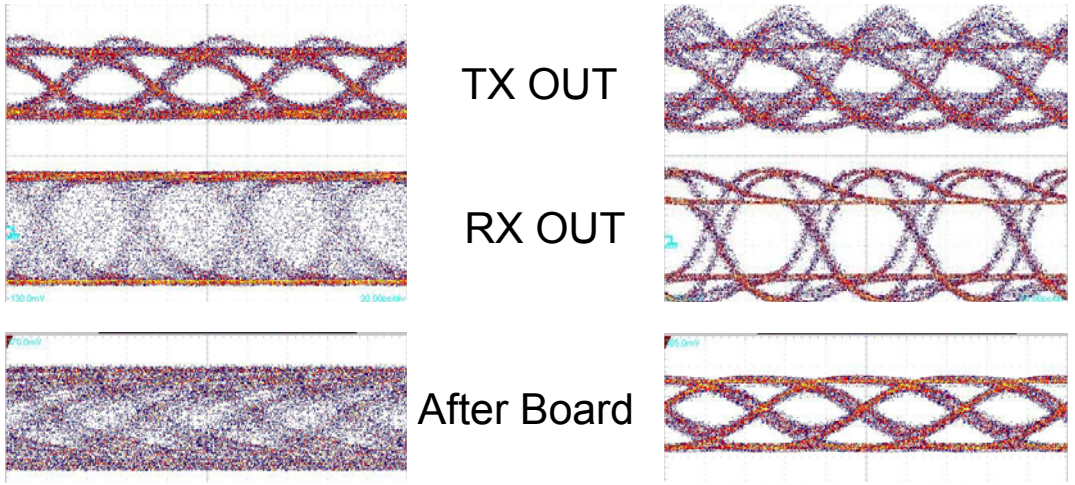
Double Equalized Links: 20 Gb/s





	Power (mW)
TX_PA	49
TX_OS	23
VCSEL	10.7
TX Total	**82.7**
TX Equalizer (included in TX total)	5.4
RX_TIA	27.3
RX_LA	65.1
RX_IO	31.2
RX Total	**123.6**
RX Equalizer (included in RX total)	3.9
Link Total	**206.3**

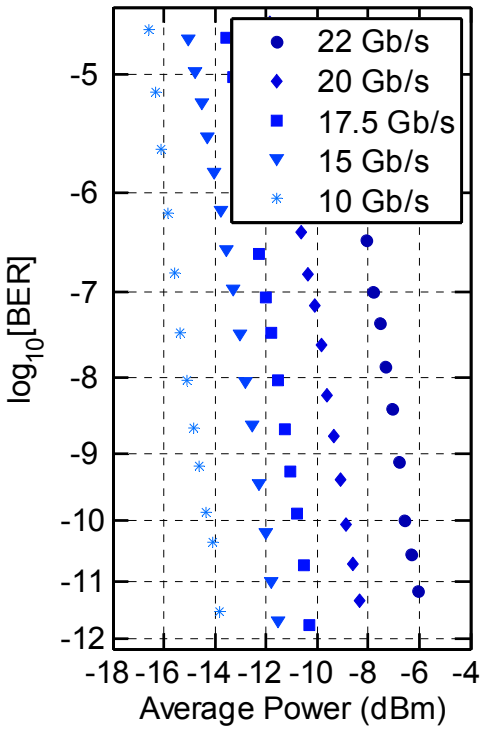
TX ER = 2.0
TX output power:
OMA = -1.4 dBm
P_{avg} = +0.3 dBm

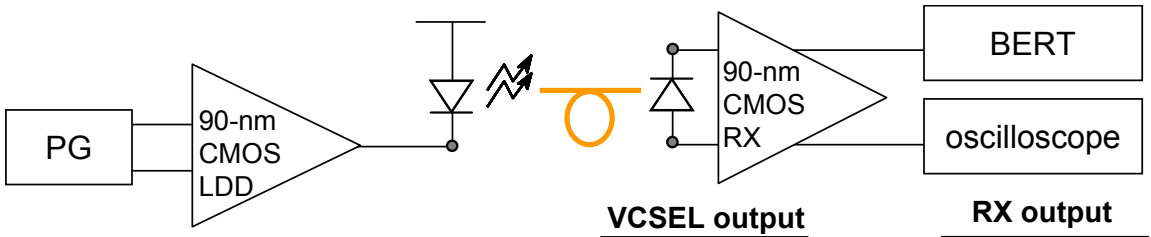


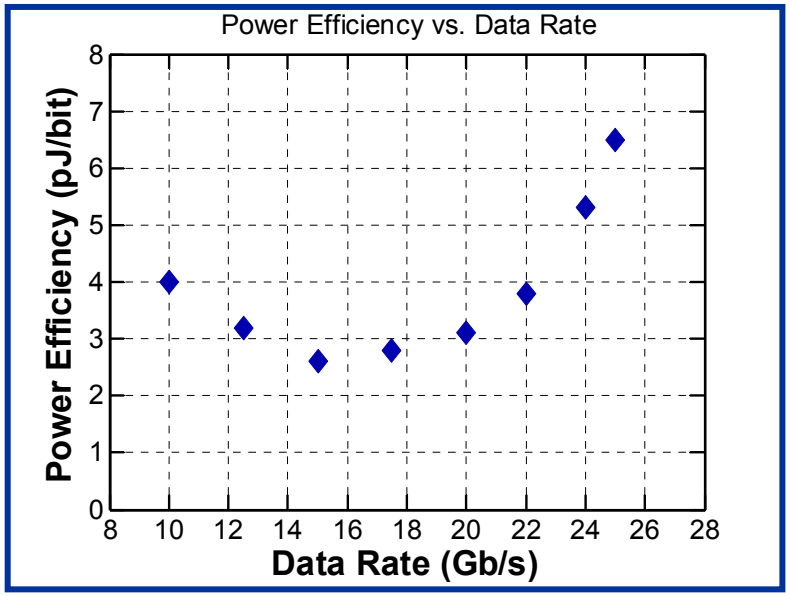
- **Equalizers enable 20Gb/s operation**
- **Dramatic improvements in eye opening**
  - Additional 0.22 UI (22 ps) eye opening, even at 10 Gb/s

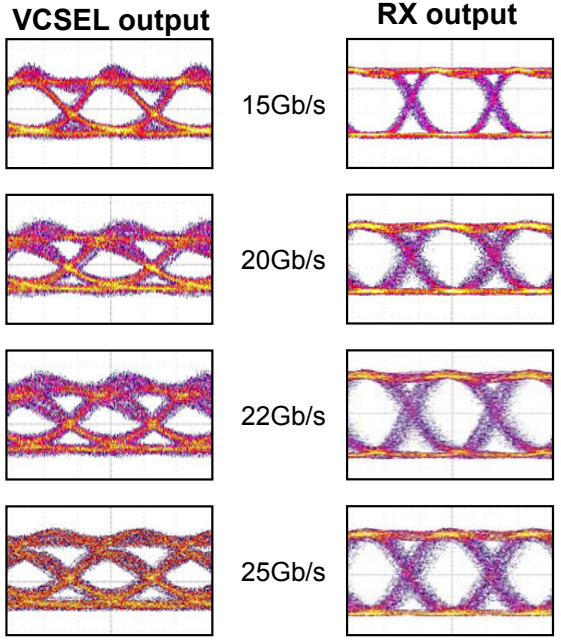
- A. V. Rylyakov *et al.*, "Transmitter Pre-Distortion for Simultaneous Improvements in Bit-Rate, Sensitivity, Jitter, and Power Efficiency in 20 Gb/s CMOS-driven VCSEL Links," *J. of Lightwave Technol.*, 2012.

Extending CMOS links to 25 Gb/s







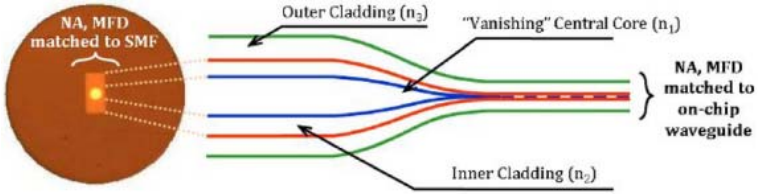


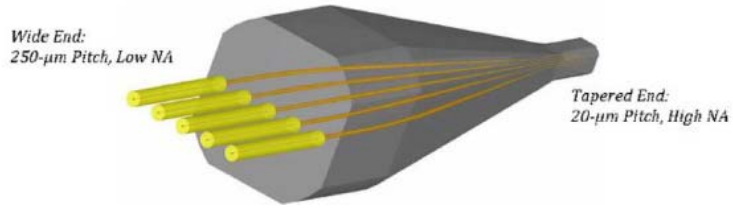
- **Links operate up to 25 Gb/s: a first for CMOS**
- **Record power efficiencies: 2.6pJ/bit @ 15 Gb/s, 3.1 pJ/bit @ 20 Gb/s**
- **Transmitter equalization will likely yield further improvement**

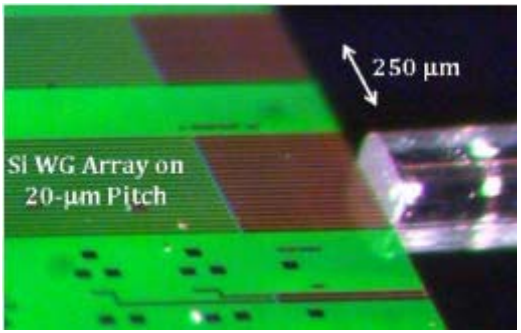
- C. L. Schow et al., “A 25 Gb/s, 6.5 pJ/bit, 90-nm CMOS Based Multimode Optical Link” Submitted to IEEE Photonics Technol. Lett., 2011.

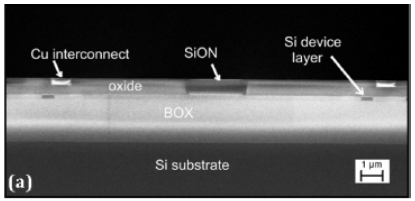
Silicon Photonics-Related: Coupling to on-chip waveguides

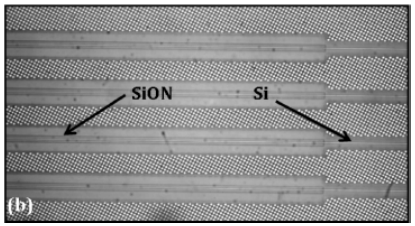
- **Edge-coupling of optical waveguides in silicon photonics chip matches well with standard IC packaging practice & power/cooling requirements.**
- **Key problem: low-loss coupling to standard optical fiber**

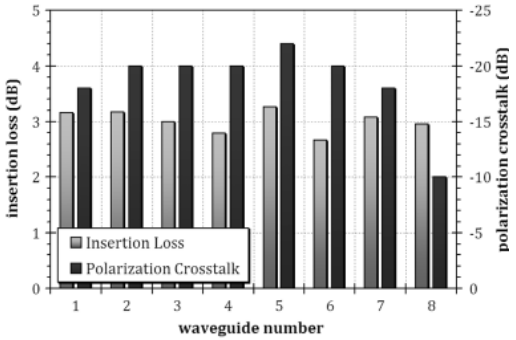


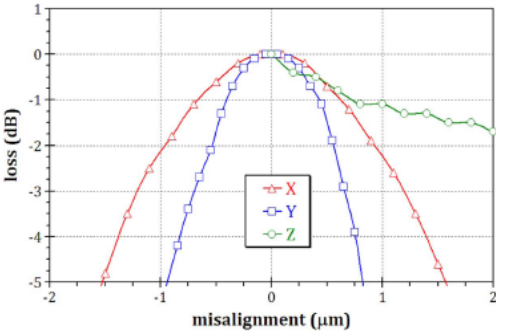












• F. E. Doany et al., "Multichannel High-Bandwidth Coupling of Ultradense Silicon Photonic Waveguide Array to Standard-Pitch Fiber Array", JLT, Vol. 29, No. 4, Feb.2011

Looking Forward: Exascale Systems

Evolution of Supercomputer-scale systems – 1980s-2020s

Supercomputing - 1980s

1-8 processors in 1 rack

Supercomputing 2000s:

10,000s of CPUs in 100s of racks

Supercomputing 2020s:

10M to >100M CPU cores, >500 racks?

??

- **In 2018-2020, we'll be building Exascale systems – 10^{18} ops/sec – with 10s of millions of processing cores, near billion-way parallelism**
  - Yes, there are apps that can use this processing power:
    - Molecular-level cell simulations,
    - Modeling brain dynamics at level of individual neurons,
    - Multi-scale & multi-rate fluid dynamics, …
- **<u>Massive</u> interconnection (BW & channel count) will be needed - within & between racks.**

2015-2020 – Exascale Computing Systems

- **We're expecting to need to build balanced ExaFLOP/s scale systems in ~2018**
  - 100-Million to 1 Billion-way parallelism

← Yes, 100 Million to Billion-way systems

- **Roadmaps to Exascale: well explored in DARPA/IPTO industry-wide study**
  - "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems", by Peter Kogge et. al., http://www.nd.edu/~kogge/reports.html
    - (Peter Kogge is a former IBM Fellow, now at Notre Dame)

- **Key points regarding interconnect / networking:**
  - "The single most difficult and pervasive challenge perceived by the study group dealt with energy, namely,...energy per operation"
  - "[The] ***energy in data transport*** will dwarf the traditional computational component in future Exascale systems....particularly so for the largest data center class." ***[italics added]***

- **➔Exaggerating a bit: Energy for data transport is *the* problem for exascale systems**
  - ~ 200x more energy needed to transport a bit from a nearest-neighbor chip than to operate on it.
    - Energy needed for a floating-point operation (~'13-'16): 0.1-0.05 pJ/bit
    - Energy needed for data transport on-card, ~3-10 inches: 2-10 pJ/bit ,← up to 200x higher
    - Energy needed for data transport across a big system: ~20-100 pJ/bit ← up to 2,000x higher
      - Assume: 3-7-hop network diam., 3-8 pJ/bit per link for transmission, 2 pJ/bit routing in ASIC

Yes, I know the software people will disagree, -- software is another critical problem for exascale.

The Road to Exascale

Year	Peak Performance	Machine Cost	Total Power Consumption
2008	1PF	$150M	2.5MW
2012	10PF	$225M	5MW
2016	100PF	$340M	10MW
2020	1000PF (1EF)	$500M	20MW

- **Assumptions: Based on typical historical trends (see, e.g., top500.org and green500.org):**
  - 10X performance, 4 years later, costs 1.5X more dollars
  - 10X performance, 4 years later, consumes 2X more power

Acknowledgment: J. Kash

How much optics, and at what cost?

Year	Peak Performance	(Bidi) Optical Bandwidth	Optics Power Consumption	Optics Cost
2008	1PF	0.012PB/s (1.2×10^5 Gb/s)	0.012MW	$2.4M
2012	10PF	1PB/s (10^7 Gb/s)	0.5MW	$22M
2016	100PF	20PB/sec (2×10^8 Gb/s)	2MW	$68M
2020	1000PF (1EF)	400PB/sec (4×10^9 Gb/s)	8MW	$200M

- **Target >0.2Byte/FLOP I/O bandwidth plus >0.2Byte/FLOP memory bandwidth**
  - 2008 optics replaces electrical cables (0.012Byte/FLOP, 40mW/Gb/s)
  - 2012 optics replaces electrical backplane (0.1Byte/FLOP, 10% of system power/cost)
  - 2016 optics replaces electrical PCB (0.2Byte/FLOP, 20% of system power/cost)
  - 2020 optics on-chip (or to memory) (0.4Byte/FLOP, 40% of system power/cost)

Acknowledgment: J. Kash

Cost and Power per bit (unidirectional)

Year	Peak Performance	number of optical channels	Optics Power Consumption	Optics Cost
2008	1PF	48,000 (@ 5Gb/s)	50mW/Gb/s (50pJ/bit)	$10,000 per Tb/s
2012	10PF	$2x10^6$ (@ 10Gb/s)	25mW/Gb/s	$1,100 per Tb/s
2016	100PF	$4x10^7$ (@ 14-25 Gb/s)	5mW/Gb/s	$170 per Tb/s
2020	1000PF (1EF)	$8x10^8$ (@ ~25 Gb/s?)	1mW/Gb/s	$25 per Tb/s

- **Future directions for optical cables:**
  - Lower cost (reducing >60%/year)
  - Much more BW (increasing >210%/year)
  - Much lower power (improving >45%/year)
- **Variety of methods for reaching these targets**
  - Higher bitrates: 10-20-20 Gb/s per channel
  - Smaller footprint for O/E modules
  - Move optics closer to logic
  - New technologies

Acknowledgment: J. Kash

Summary

Summary Remarks

- **The future is bright.**
- **Optics will play a steadily-increasing role in systems – Must feed the transistors**
  - Bandwidth-density, power-efficient data transport, reliable signal integrity
- **Parallel optical interconnects are fast replacing copper cables today**
- **Lots of interesting systems-level challenges, lots of technologies to choose from**
- **Optical interconnect for supercomputers and other high-end compute systems will likely grow at >200% CAGR (deployed Gb/s), assuming cost can be improved at 60% CAGR ($/Gb/s) and power can be improved at 45% CAGR (mw/Gb/s) at the same time.**

We're banking on this happening – the question is (/ questions are):
How?

- **For Exascale systems in 2015-2020, interconnect is *the* interesting technical problem.**
  - CPUs/GPUs/SPUs/APUs get the glory, and are interesting business-wise, but technically, FLOPs are easy. Storage capacity is harder, but technically requires no breakthroughs.
  - Data transfer – chip/chip, card/card, rack/rack – is *hard*.
    - Will account for >80% of the system power, & 50-90% (app-dependent) of performance

Thank you kindly