



IBM Research

Optical Technologies for Data Communication in Large Parallel Systems



Mark B. Ritter, Yurii Vlasov, Jeffrey A. Kash, and Alan Benner*

IBM T.J. Watson Research Center, *IBM Poughkeepsie

mritter@us.ibm.com

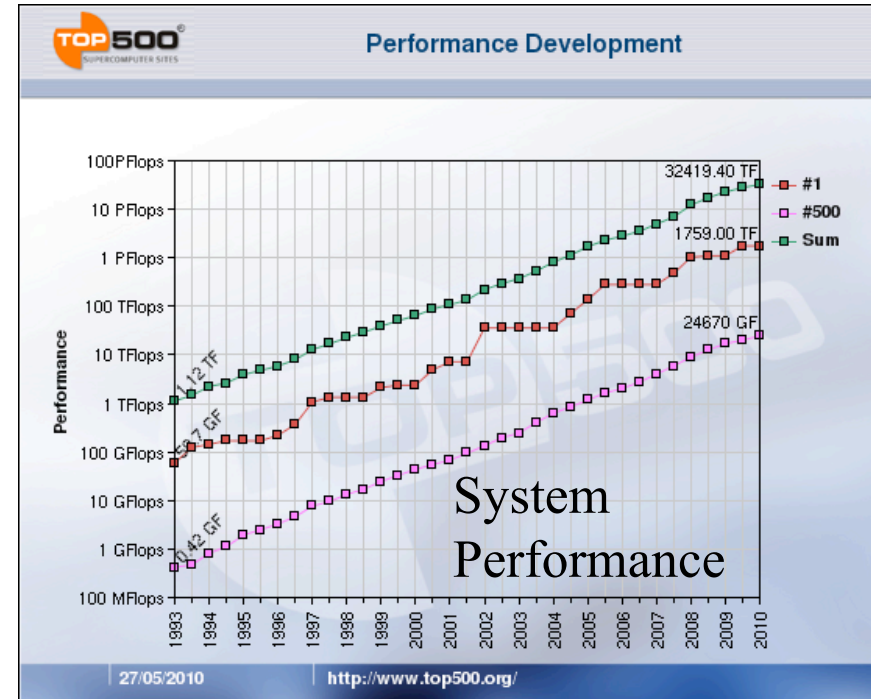
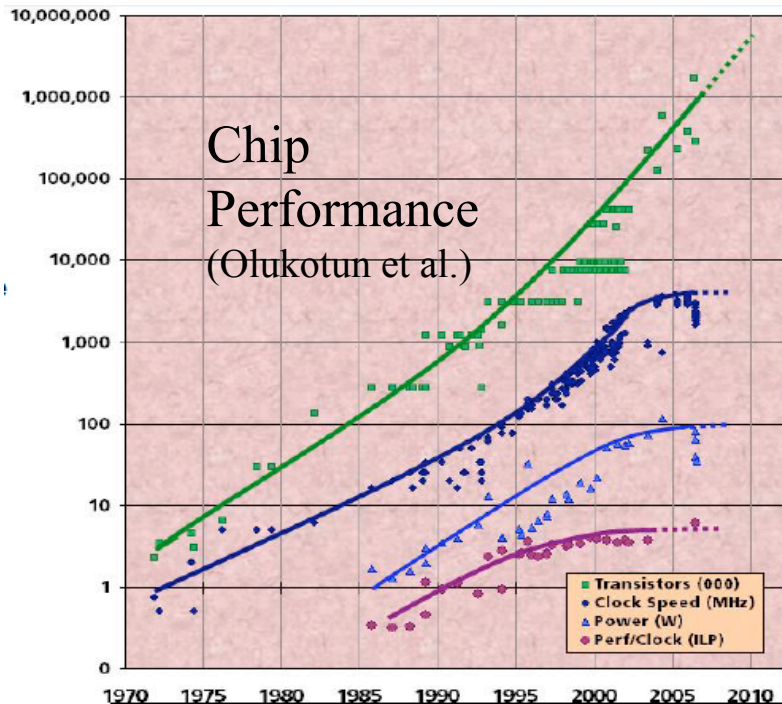


Outline

- HPC Performance Scaling and Bandwidth
- Anatomy of a Link
- Electrical and Optical Interconnect Limits
- Promise of Nanophotonic Technology
- Potential Insertion Points
- Summary

Copyright 2005. Barcelona Supercomputing Center - BSC

Performance Scaling Now Driven by Communication



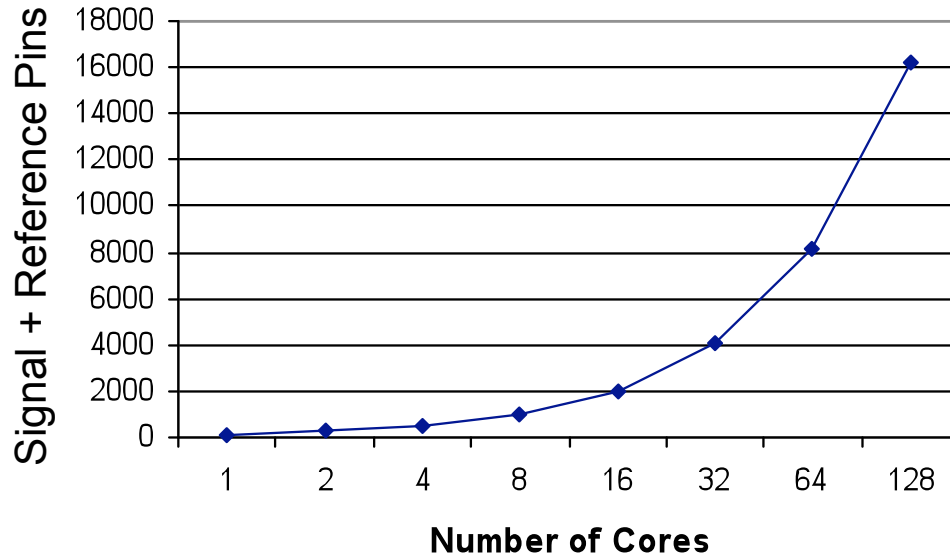
- System performance gains no longer principally from lithography-driven uniprocessor performance
- Performance gains now from parallelism exploited at chip, system level
- BW requirements must scale with System Performance, ~1B/FLOP (memory & network)
- Requires exponential increases in communication bandwidth at all levels of the system
 - Inter-rack, backplane, card, chip



Bandwidth: the Bane of the Multicore Paradigm:

- **Logic flops continue to scale faster than interconnect BW**
 - Constant Byte/Flop ratio with N cores (constant ν) means:

$$\text{Bandwidth}(N\text{-core}) = N \times \text{Bandwidth}(\text{single core})$$



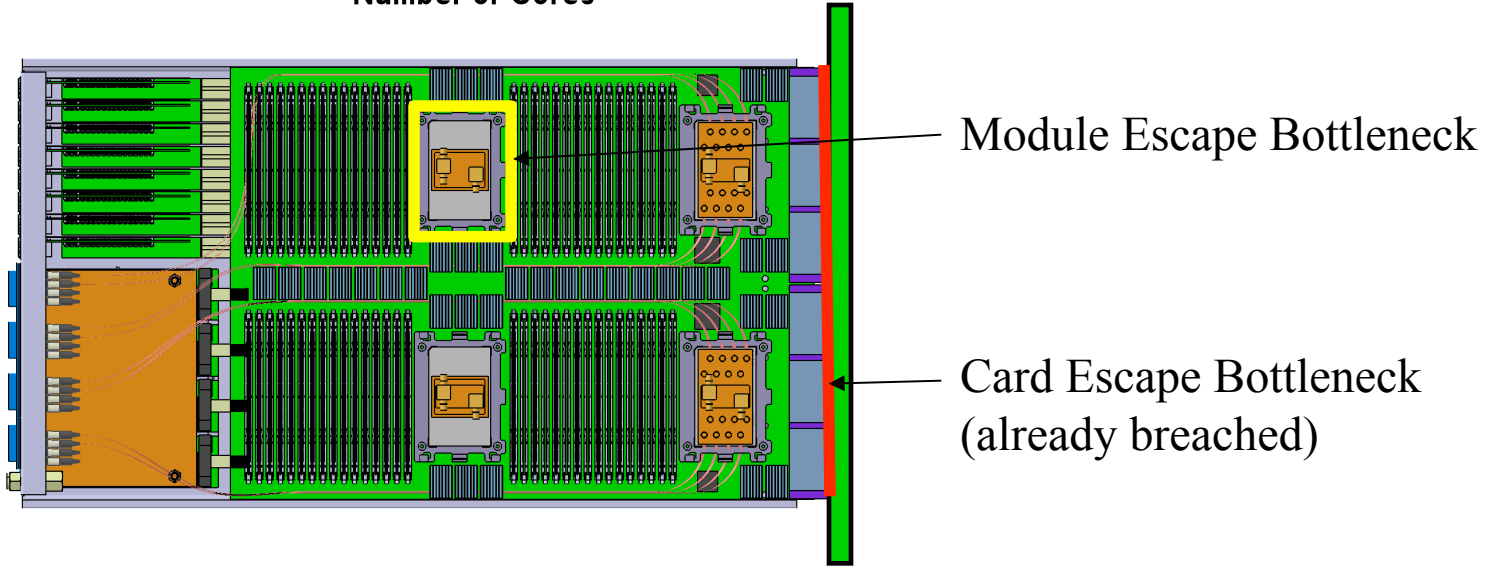
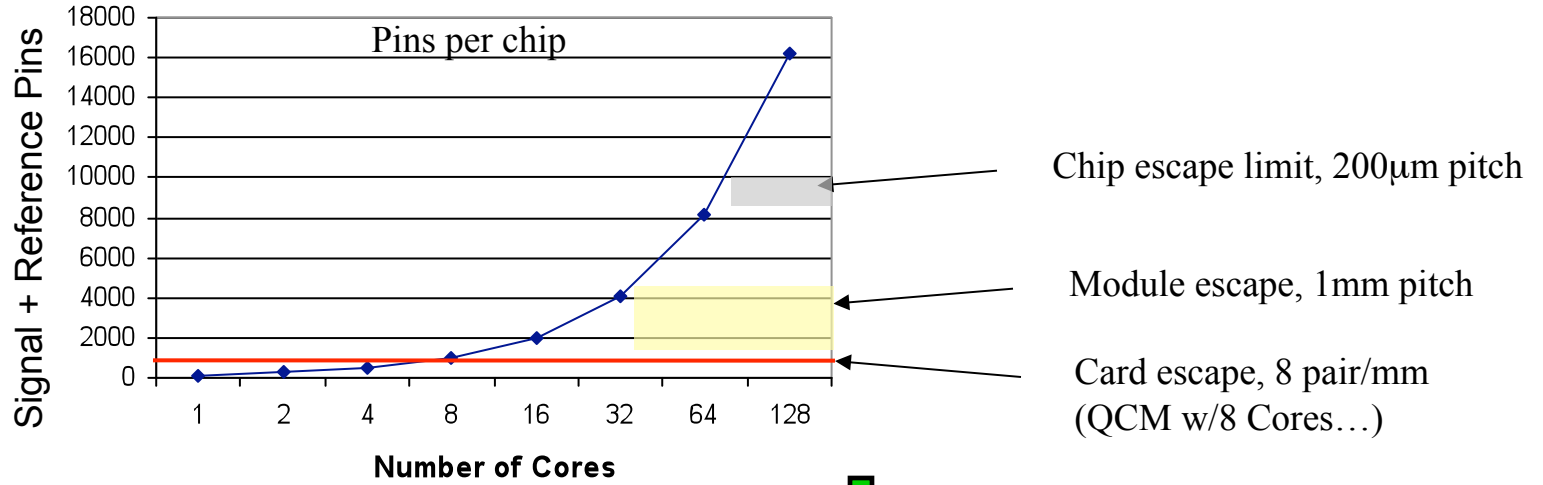
Assumptions:

- 3 GHz clock
- ~ 3 IPC
- 10 Gb/s I/O

- 1 B/Flop mem
- 0.1 B/Flop data
- 0.05 B/Flop I/O

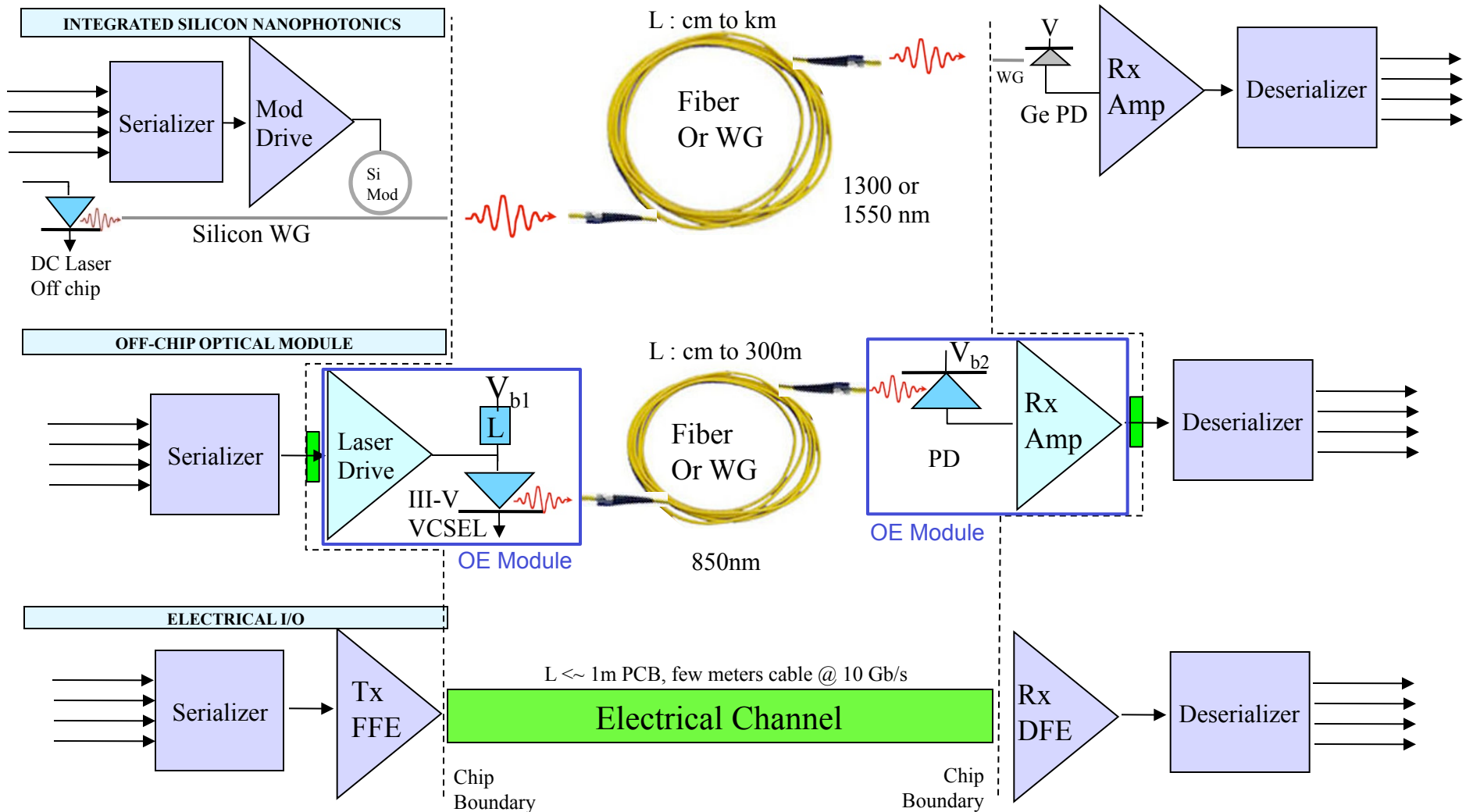
- 3Di (3D integration) will only exacerbate bottlenecks

Implications of BW Scaling:



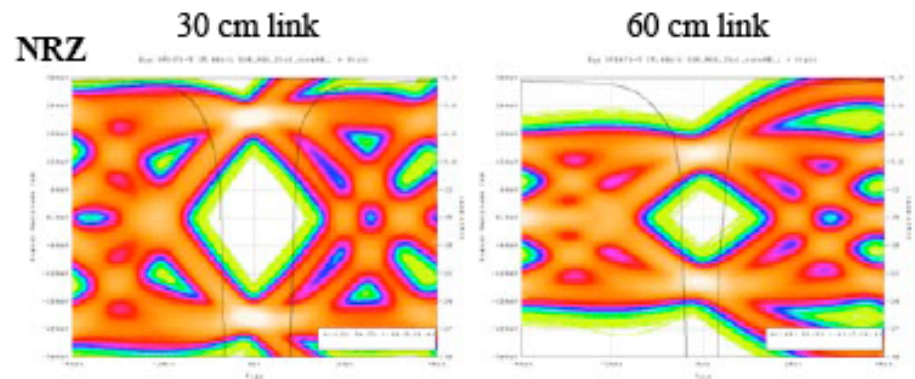
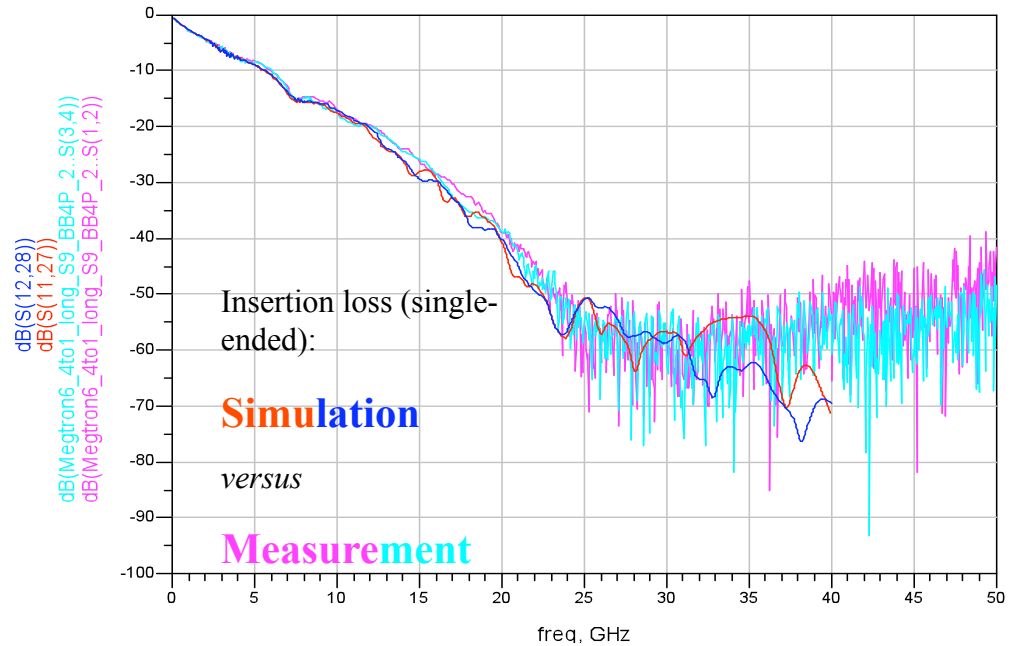
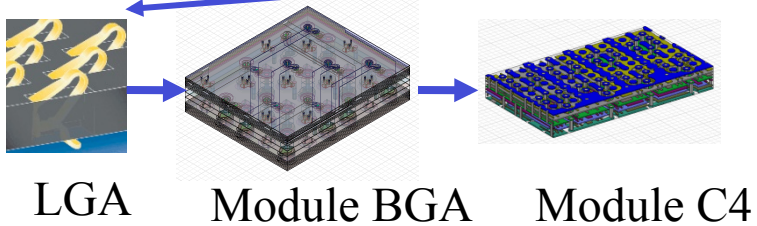
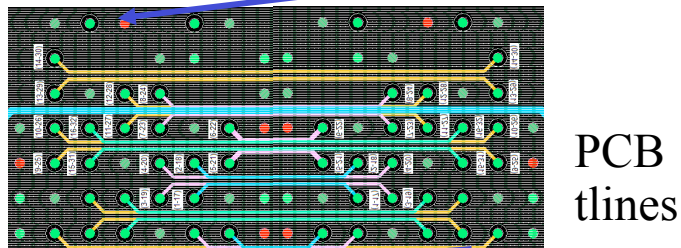
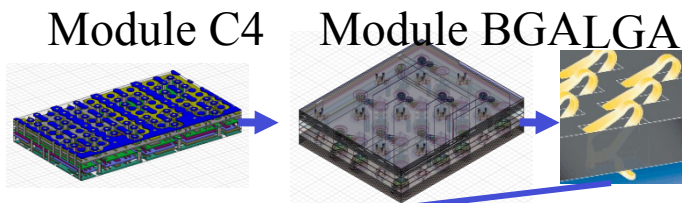
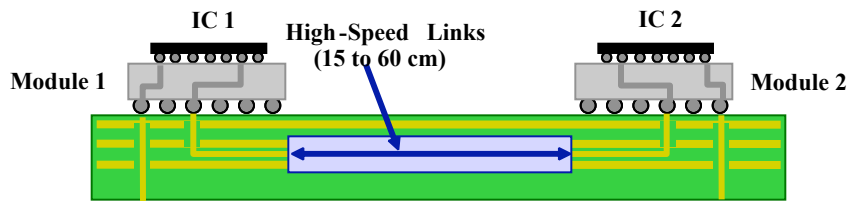
Only several generations left before module I/O limit is hit... what sets limits?

Anatomy of Communication Links:



All links have same basic features, the differences are in modulation and detection, and these differences determine power efficiency, distance x bandwidth, and density...

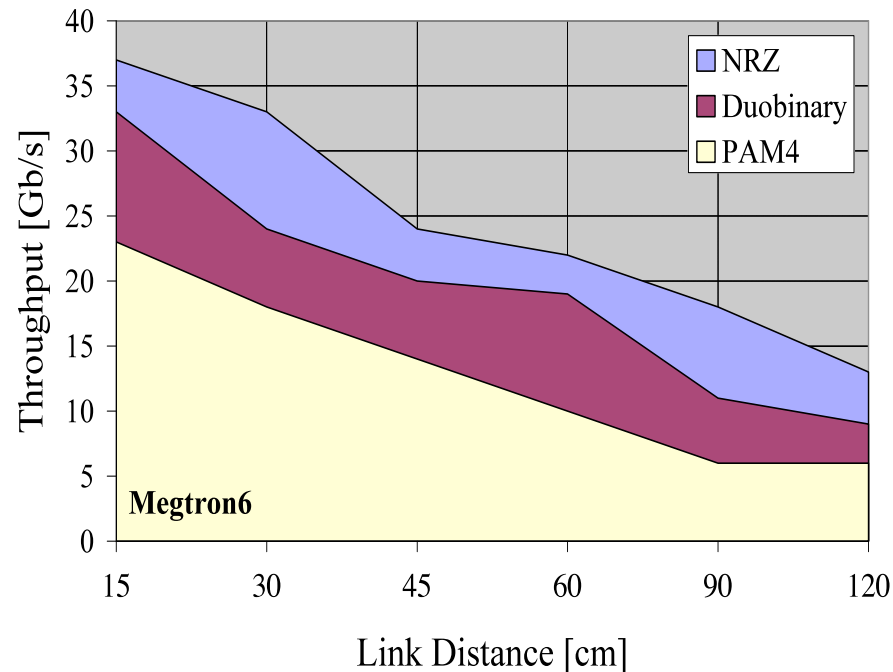
Electrical Interconnect Modeling



Modeling accuracy confirmed with measurement, model limits of electrical...

Electrical Interconnect Limits

- **Module-to-module on-board limits:**



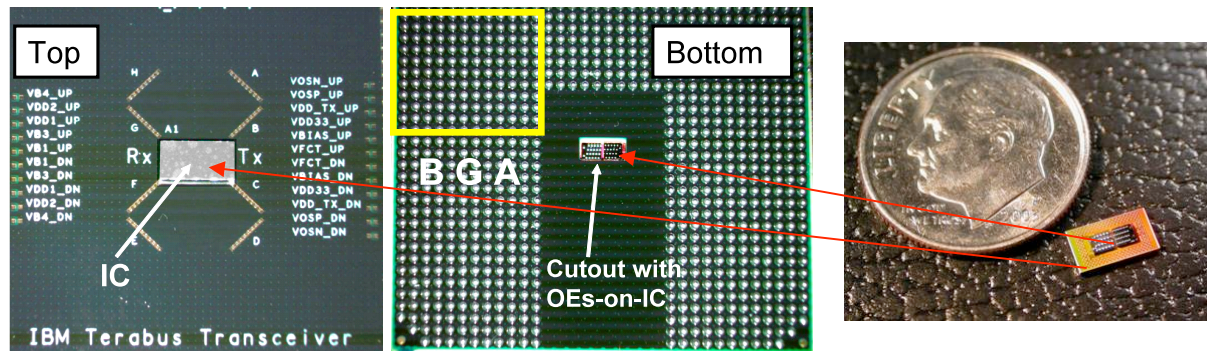
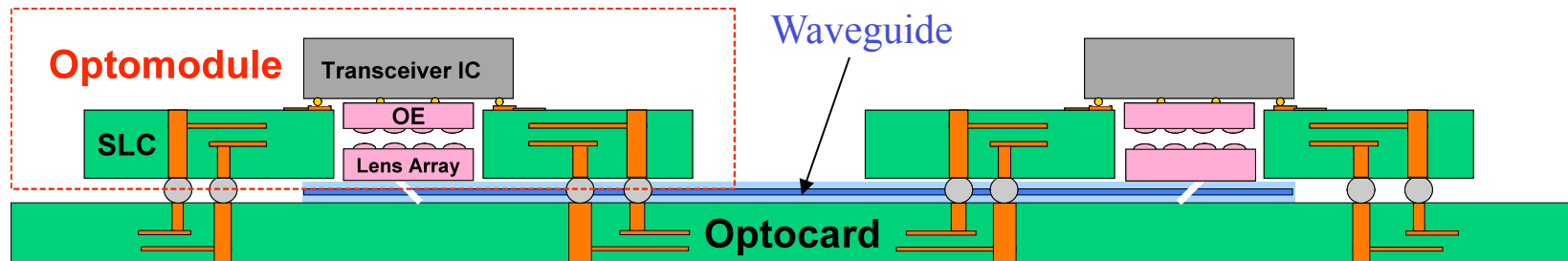
NRZ with FFE and DFE (and/or CTLE) best modulation for dense buses.

Achieve 25 Gb/s @ 45cm...

Costly dielectrics for > 25 Gb/s...

- **Off-board (backplane):**
 - Limits board-to-board bitrates to ~6.4 Gb/s for typical server configurations
- **Rack-to-rack** – already optical

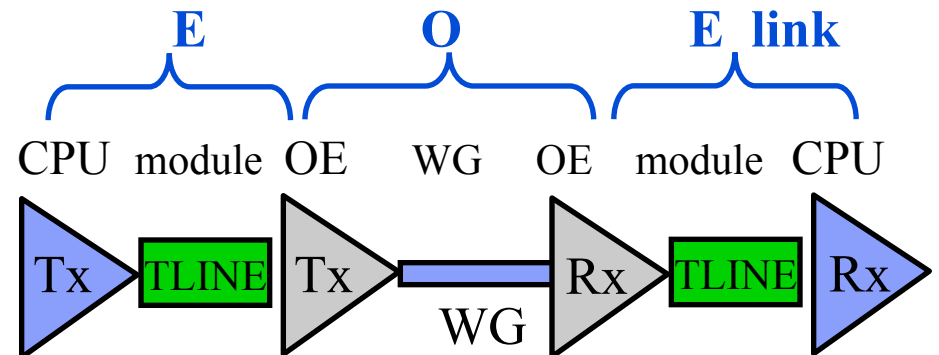
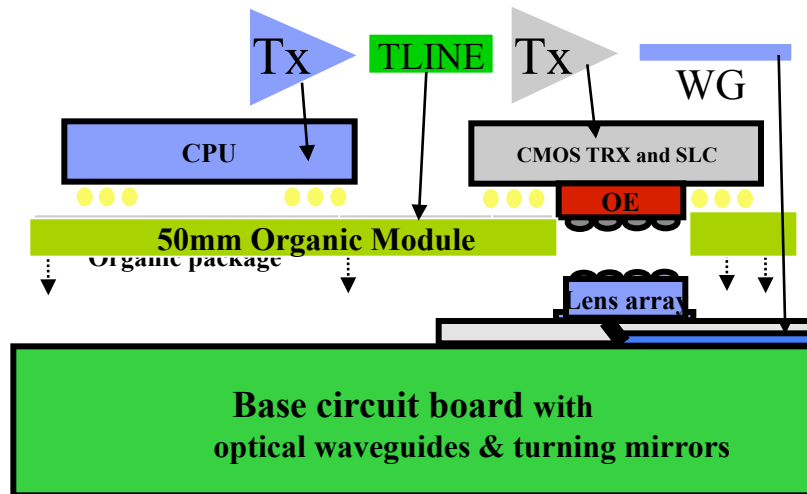
Optical Interconnect Modeling



Terabus
“Optomodule”
Kash et. al.

- **Lowest-power links use optics as “analog repeaters” of signal with no clock recovery**
 - E-O-E modeling required: jitter adds up over two electrical links, one optical link

Optical Interconnect Modeling

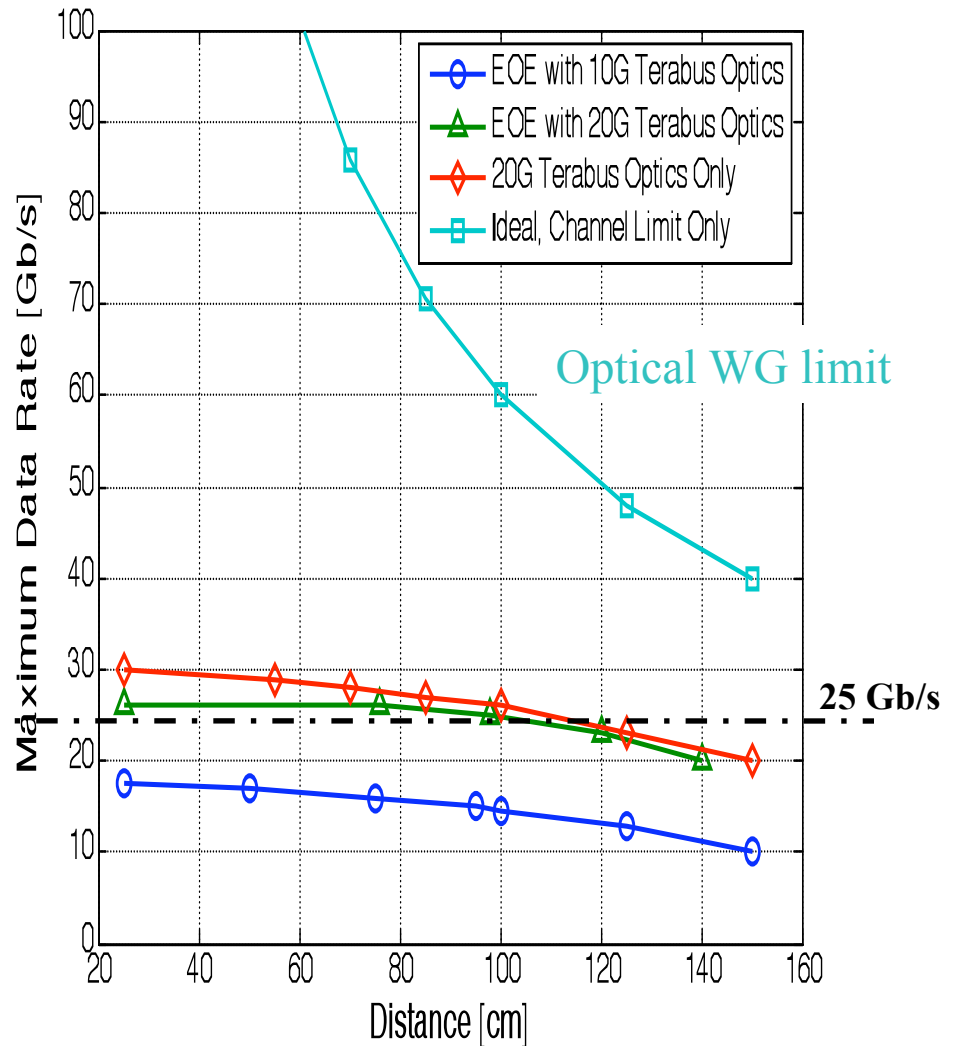
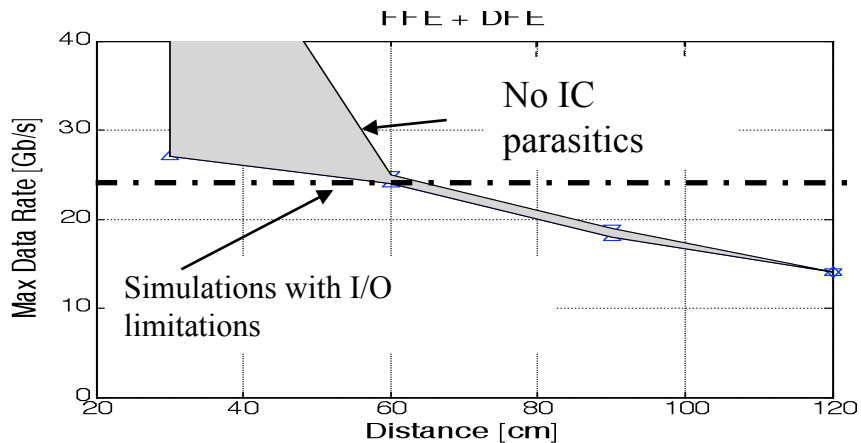


$$BW_{\text{oe}} = \left[\frac{1}{\sum_i \frac{1}{BW_i^2}} \right]^{1/2} \approx 26 \text{ GHz}$$

- Two electrical on-module links, each with ~ 30 GHz media BW
- One WG optical link with ~ 40 GHz media BW
- Assume electrical, optical I/O do not limit link BW, get 26 GHz BW
- Actual link includes electrical, optical I/O BW, drives system BW to < ~18 GHz
- **This limits overall EOE link bitrate to ~ 26 Gb/s @ 1 meter**
- Our models include EOE link using full dual-Dirac jitter convolution for end-to-end composite link

Electrical and Optical Link Reach

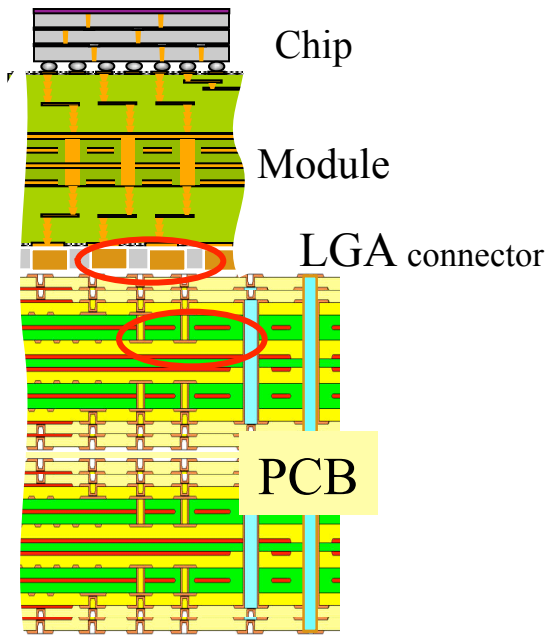
Predicted link reach @ 25 Gb/s:
 ~45 cm electrical links (Megtron 6)
 ~100 cm optical WG links



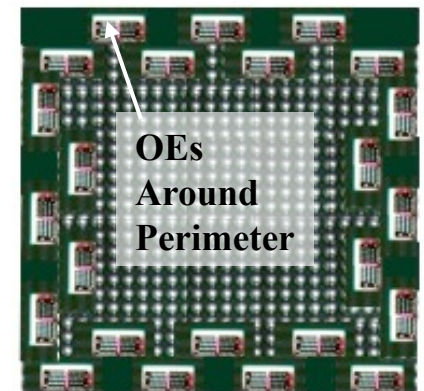
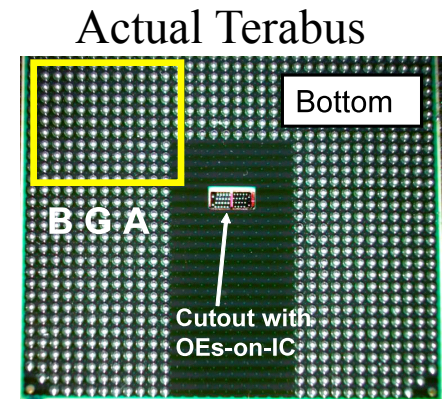
EOE links double the reach (current WG loss)

It's All About Bandwidth Escape:

Assume 60mm module, 1mm LGA pitch, 62.5 μ m WG pitch, what is escape BW?



Bandwidth of Elements (Tb/s)	
Electrical	Optical
C4 - 90 - 211	C4 - 90 - 211
Module 56 - 112	Module 56 - 112
LGA 12 - 23.5	OE Escape 46-100
LGA escape 17 - 29	Optical WG 64-166
Card 73 - 136	



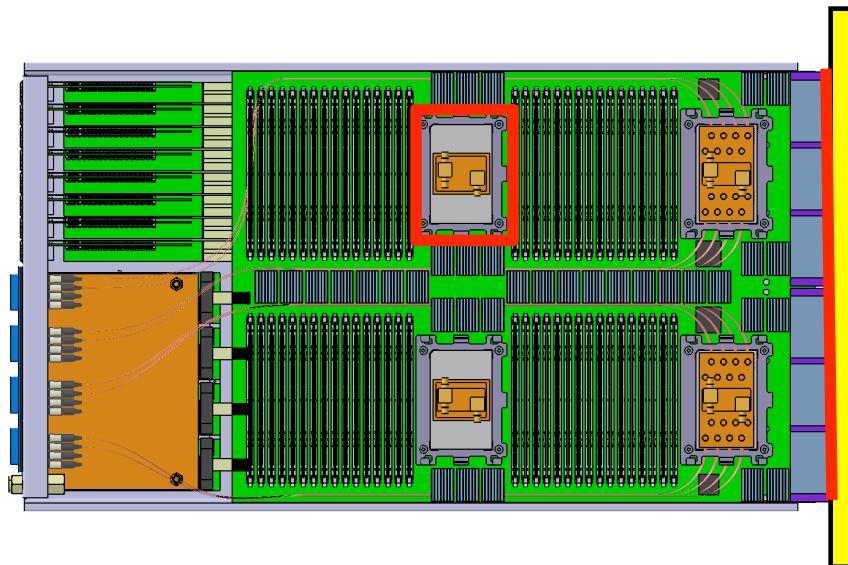
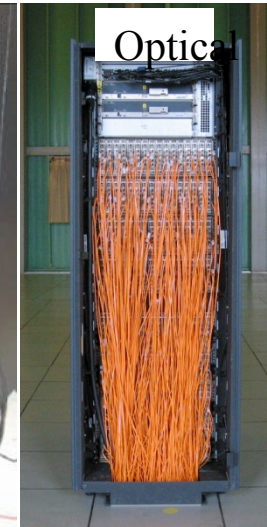
Notional Design

Multimode optical transceivers could provide ~4x module escape BW

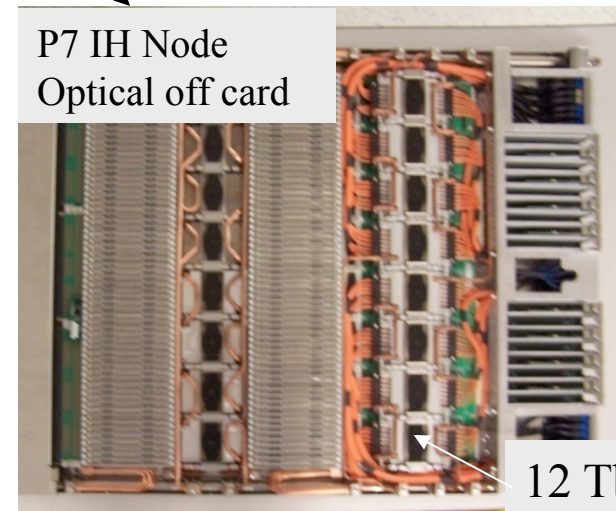
Escape Bandwidth Conclusions

For the high-end of HPC:

- Hit Rack electrical BW limit in early 2000's
- Hitting off-board BW now, P7 IH chose optics for off-board links
- Likely to hit electrical off-module BW limit soon (some packaging "fixes" - larger modules (S21 on module greater than board- tradeoff))



P7 IH Node
Optical off card



P7 IH System Hardware – Node Front View (~1000 Nodes in Blue Waters)

1m W x
1.8m D x
10cm H

Water
Connection

360VDC Input
Power Supplies

IBM's HPCS Program
partially supported by

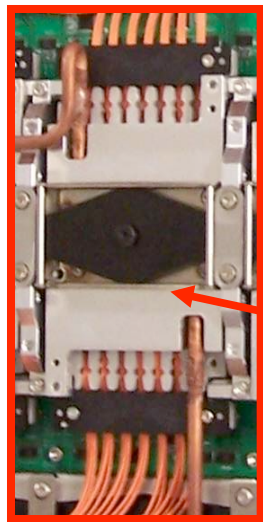


Memory
DIMM's (64x)

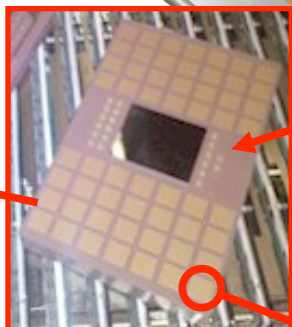
P7 QCM (8x)

Memory
DIMM's (64x)

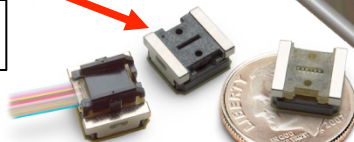
Hub
Module (8x)



Hub Assembly



MLC Module



Avago microPOD™

PCIe
Interconnect

PCIe
Interconnect

L-Link Optical Interface
Connects 4 Nodes to form Super Node

D-Link Optical Interface
Connects to other Super Nodes

D-Link Optical Interface
Connects to other Super Nodes

All off-node communication optical

PERCS/Power7-IH System - Data-Center-In-A-Rack

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 line cords. Up to 252 kW/rack max / 163 kW Typ.

Integrated Storage– 384 2.5” drives / drawer, 0-6 drawers / rack

230 TBytes/drawer (w/ 600GB 10K SAS disks), full RAID, 154 GB/s BW/drawer
Storage Drawers replace server drawers at 2-for-1 (up to 1.38 PetaBytes / 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

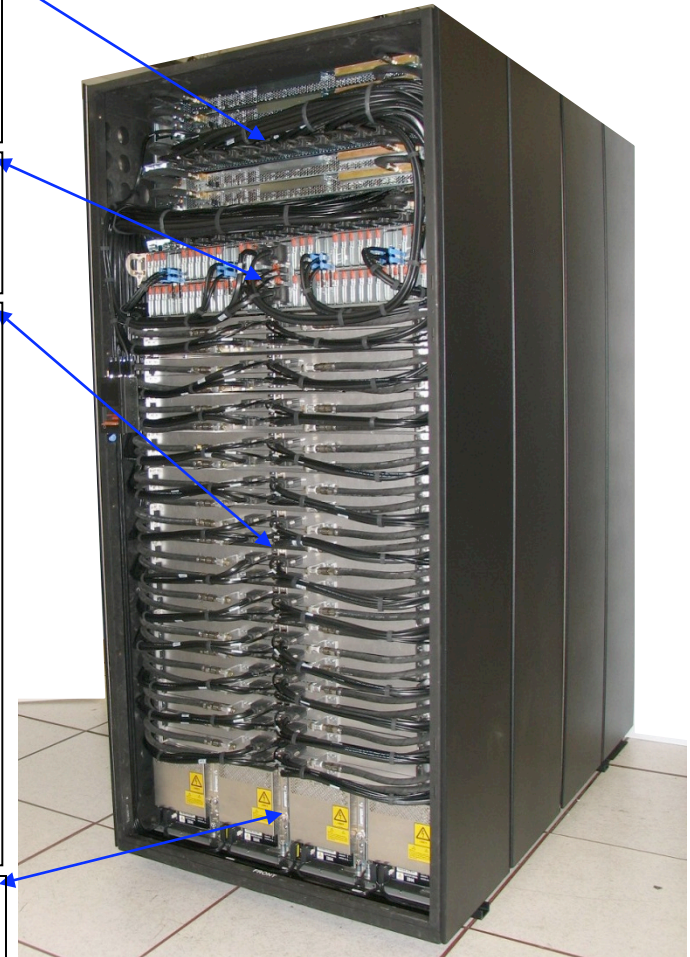
Local-remote optical links: 24 links to near hubs, (120+120) GB/s aggregate

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

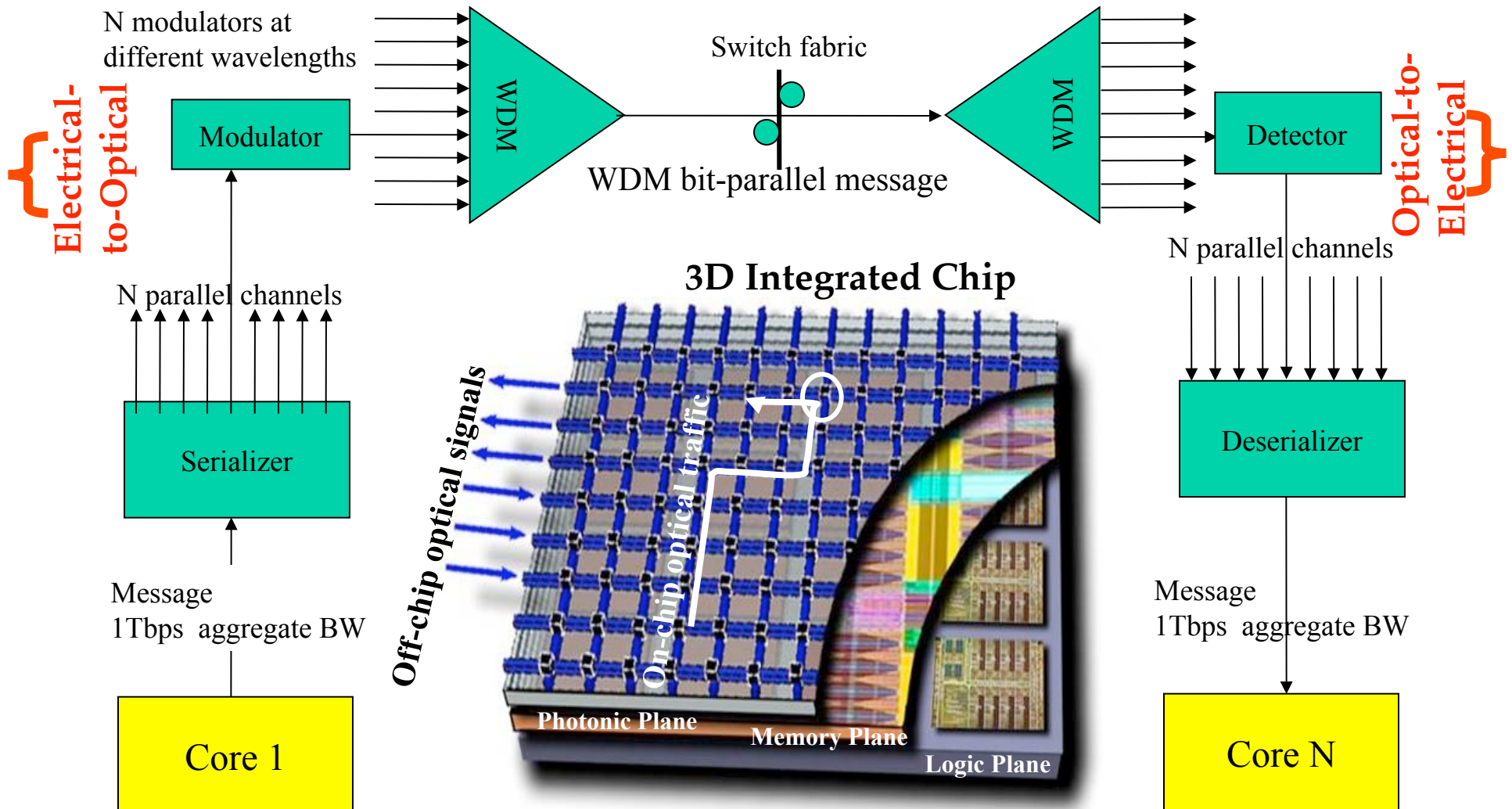
All thermal load transferred directly to building chilled water – no load on room



- **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 – extremely 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 extraneous hardware except optical fiber cables

Vision for 2020:	2020	1mW/Gb/s	\$0.025/Gb/s
------------------	------	----------	--------------

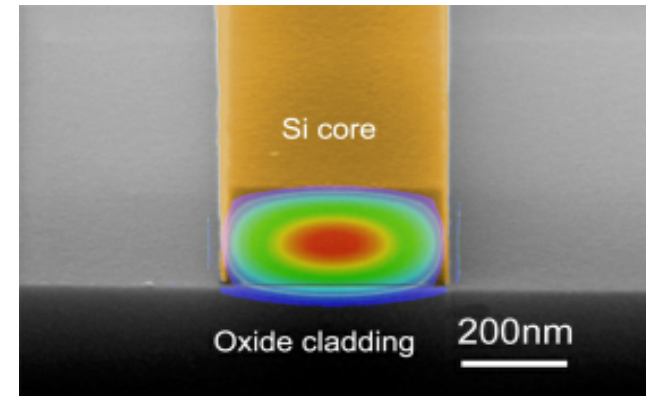
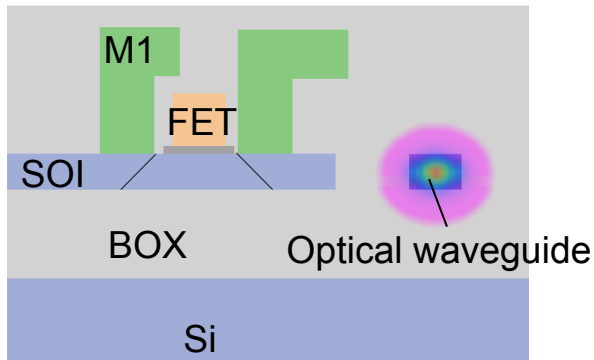
Silicon Nanophotonics for Optically connected 3-D Supercomputer Chip



Goal: Integrate Ultra-dense Photonic Circuits with electronics

- Increase off-chip BW
- Allow on-chip optical routing and interconnect

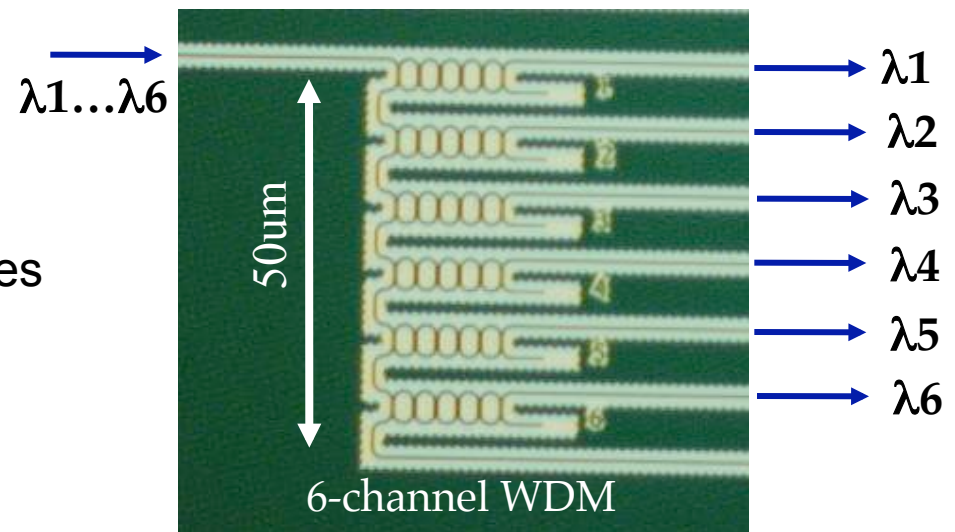
CMOS front end (FEOL) photonic integration for compatibility



→ Photonics sharing Si layer with FET body

Advantages:

- Deeply scaled Nanophotonics
- Best quality silicon and litho
 - Highest performance photonic devices
 - Lowest loss waveguides
 - Lowest power consumption
 - Most accurate passive λ control
- Most dense integration with CMOS
- Same mask set, standard processing
- Same design environment (e.g. Cadence)

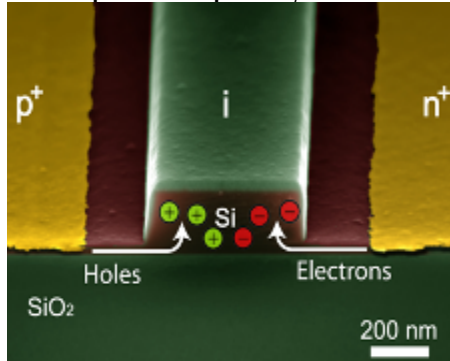


FEOL integrated Nanophotonic devices from IBM Si Photonics Group

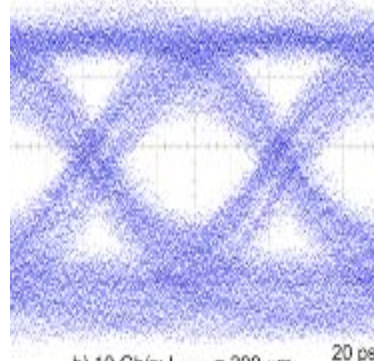
(Y. Vlasov, S. Assefa, W. Green, F. Xia, F. Horst, ... www.research.ibm.com/photronics)

1. Tx: Ultra-compact 10 Gbps modulator

Optics Express, Dec. 2007



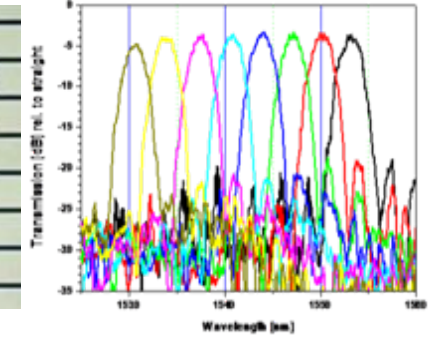
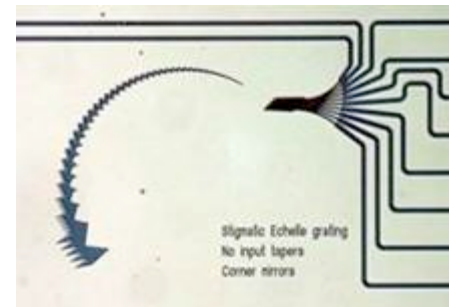
10 Gbps; $L_{MZM} = 200 \mu\text{m}$



b) 10 Gb/s; $L_{MZM} = 200 \mu\text{m}$ 20 ps

3. Ultra-compact WDM multiplexers

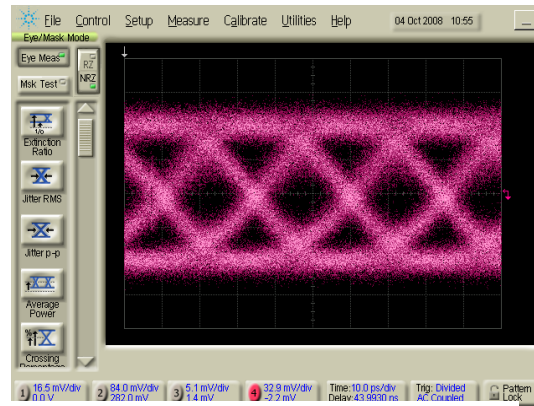
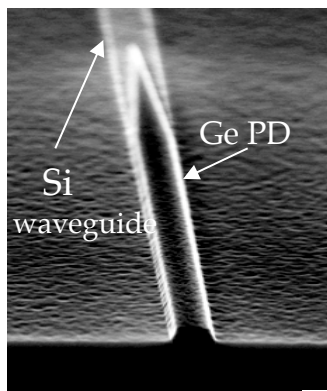
Optics Express May 2007, SPIE March 2008



Temperature insensitive; $30 \times 40 \mu\text{m}$

2. Rx: Ge waveguide photodetector

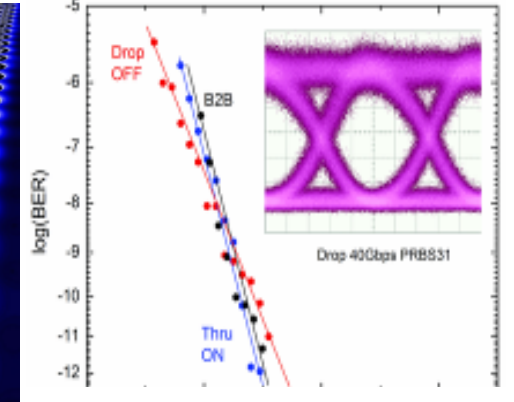
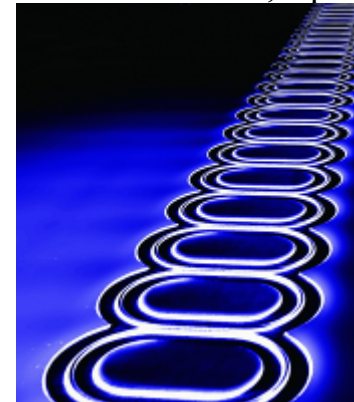
Optical Fiber Communications, March 2009



40 Gbps at 1V; 8fF capacitance


4. High-throughput nanophotonic switch

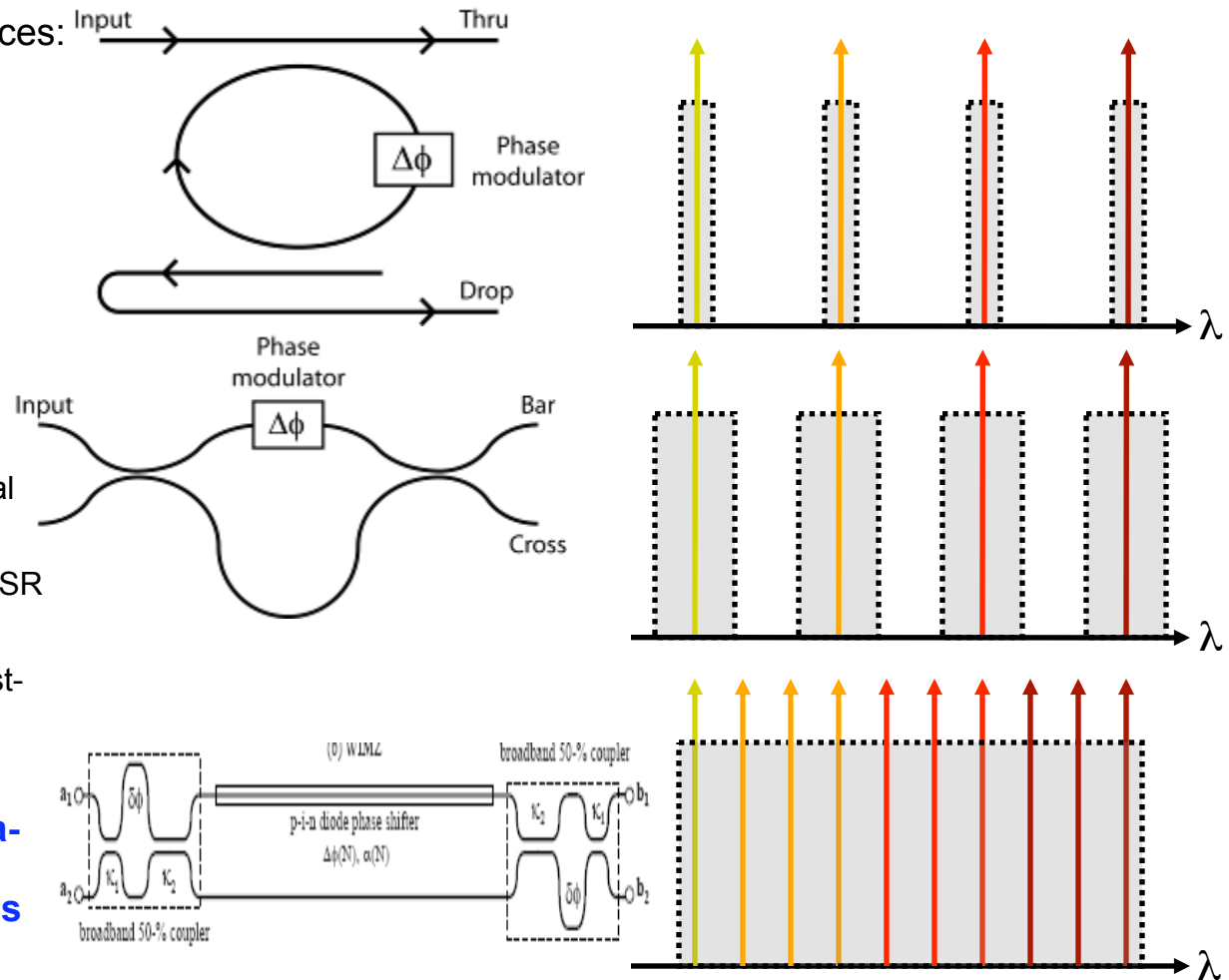
Nature Photonics, April 2008



Error free switching at 40 Gbps

Broad optical bandwidth for thermal and process stability and efficient use of optical spectrum

- Resonant or frequency sensitive devices: 
- **Ring resonator comb filter:**
 - Very narrow filter bands
 - Sensitive temperature variations
- **Conventional MZ device:**
 - Wavelength-sensitive coupler
 - Wider filter band
 - **Potential drawbacks:**
 - Unused regions between bands, spectral efficiency reduced
 - Wavelengths are restricted by varying FSR from dispersion
 - Fabrication tolerances may demand post-fabrication trimming
- **Broadband solution:**
 - **Design a filter with a single ultra-wide band, filled with multiple uniformly-spaced WDM channels**



B. G. Lee, A. Biberman, P. Dong, M. Lipson, K. Bergman, PTL **20**, 767-769 (2008).

P. Dong, S. F. Preble, M. Lipson, Opt. Express **15**, 9600-9605 (2007).

W. M. J. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov, Opt. Express **15**, 17106-17113 (2007).

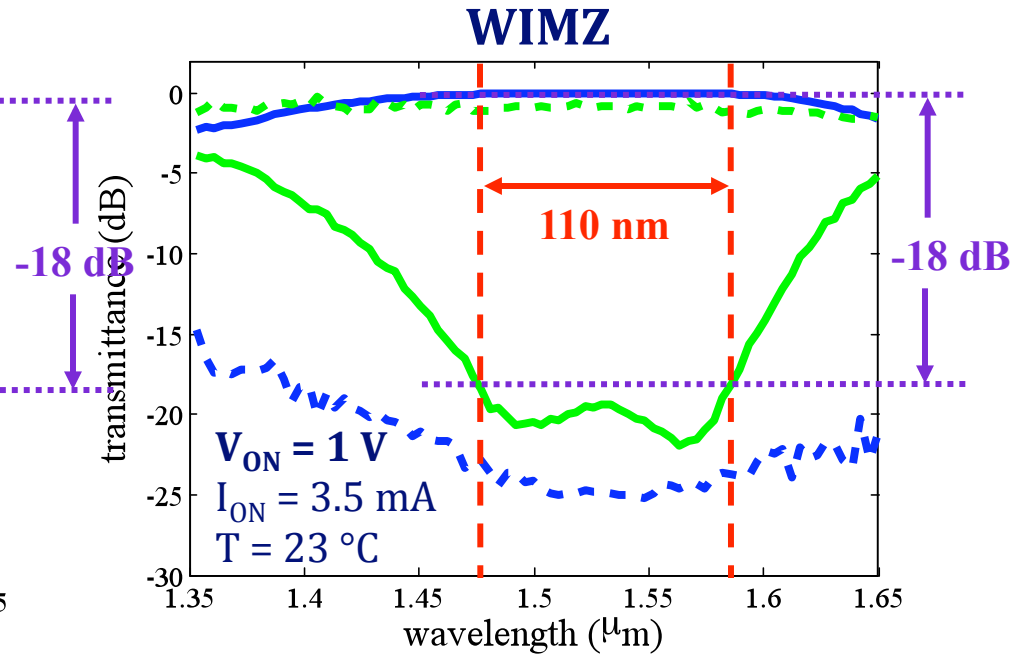
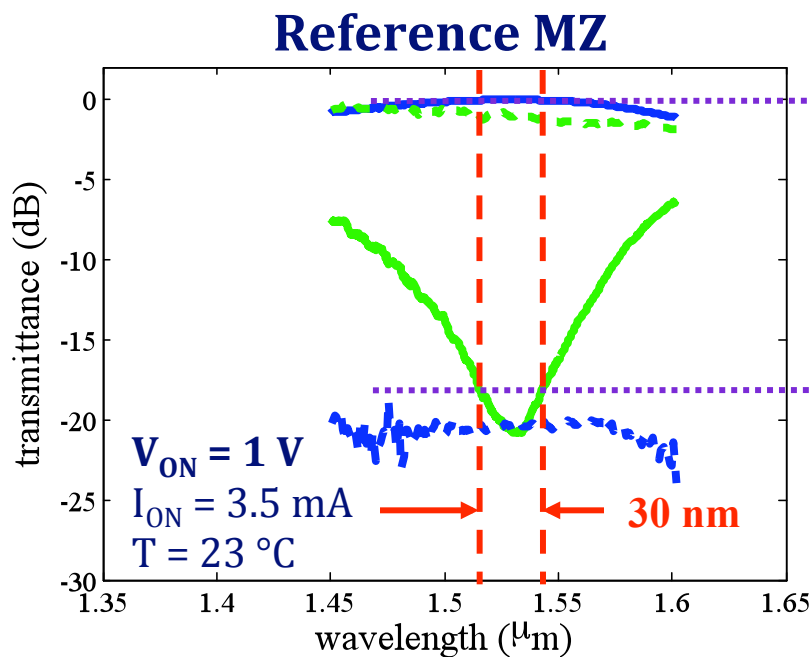
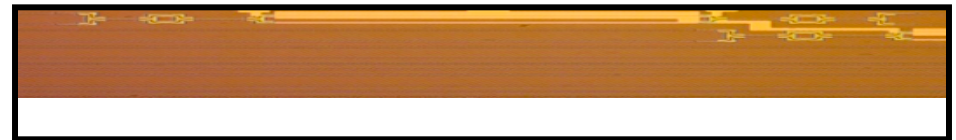
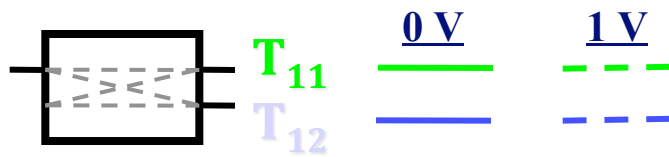
Y. Vlasov, W. M. J. Green, and F. Xia, Nature Photonics **2**, 242-246 (2008).

J. Van Campenhout, W. Green, S. Assefa, and Y. A. Vlasov, Opt. Express **17**, 24020-24029, 2009.

WIMZ Provides Large Bandwidth, Low Crosstalk and a CMOS-Compatible Drive Voltage

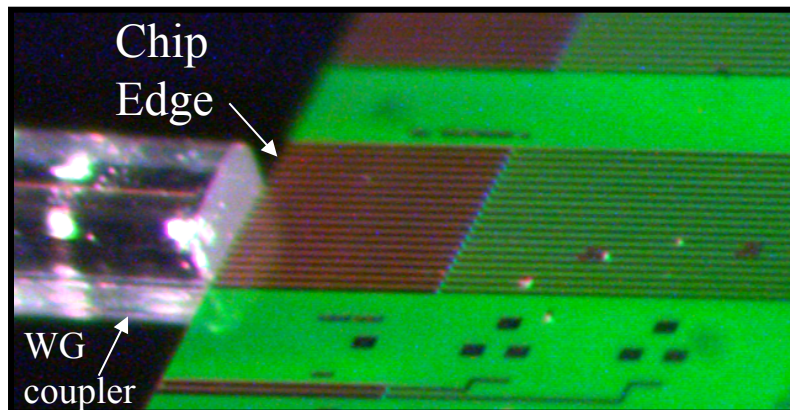
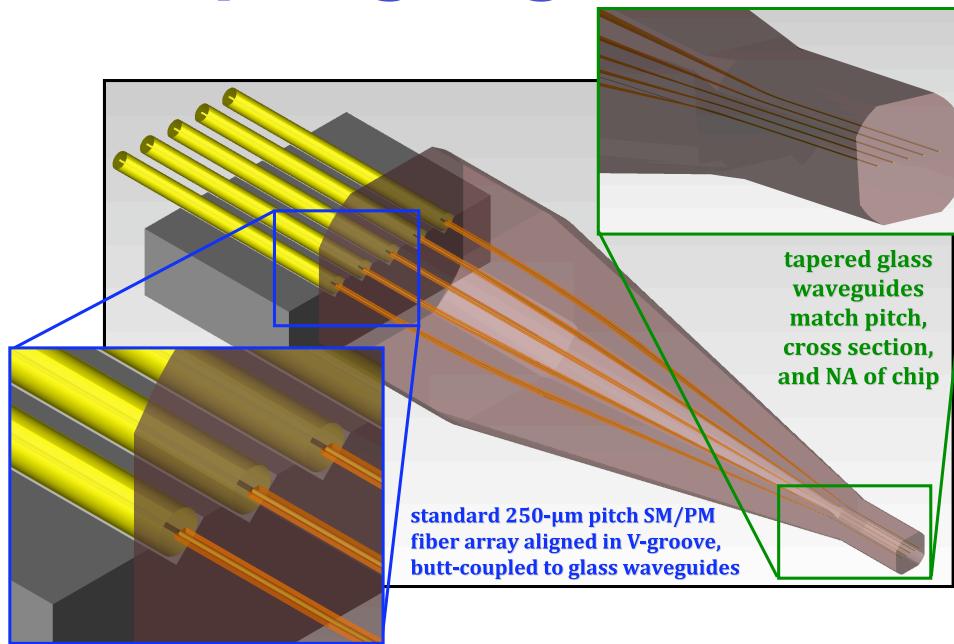


- Measured with TE-polarized broadband LED and OSA
- Normalized to total OFF-state power in both outputs



[J. Van Campenhout *et al.*, Optics Express **17** (26) 2009]

Coupling Light to Si Photonics



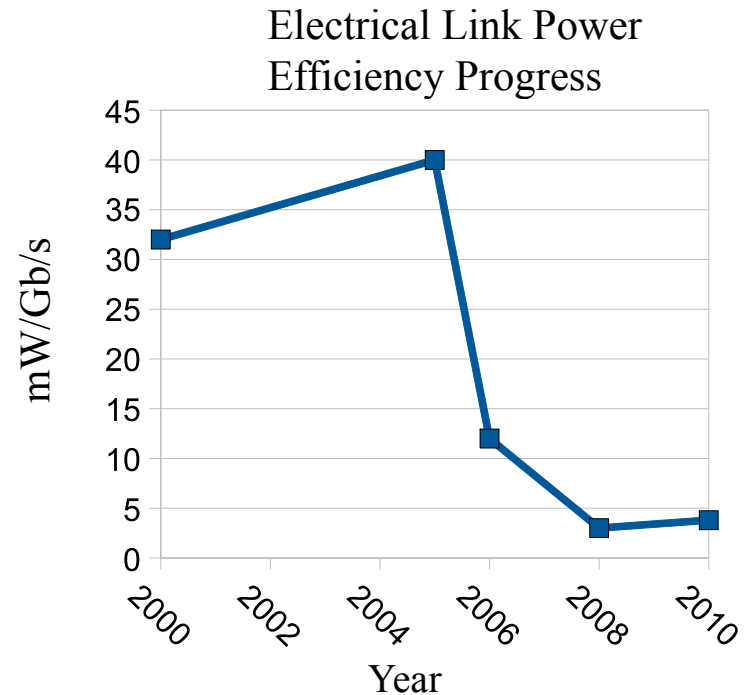
- Multichannel tapered coupler allows interfacing 250- μm -pitch PM fiber array with 20- μm -pitch silicon waveguide array
- 8-channel coupling demonstrated
- < 1 dB optical loss at interface
- Uniform insertion loss and crosstalk

B. G. Lee, F. E. Doany, S. Assefa, W. M. J. Green, M. Yang, C. L. Schow, C. V. Jahnes, S. Zhang, J. Singer, V. I. Kopp, J. A. Kash, and Y. A. Vlasov, *Proceedings of OFC 2010, paper PDP44*.

Power Issues

Exascale system example

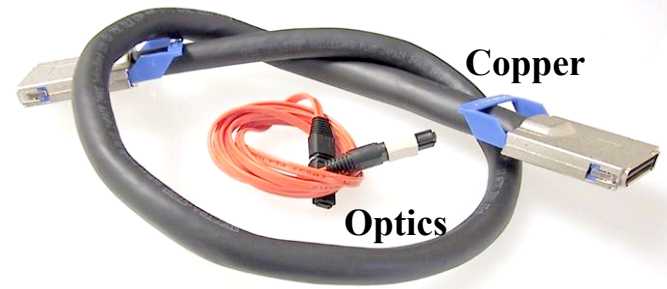
- 0.2 B/Flop comm BW,
1 B/Flop memory BW
- $1 \text{ mW/Gb/s} \times 2 \times 10^9 \text{ GF} =$
0.2 MW I/O Power
- Typical I/O 10+ mW/Gb/s...
2 MW I/O Power !! Ouch !



Link Type	Power Efficiency (pJ/bit)	Distance	50 Tb/s Off-module I/O Power (W)
Electrical	3	< 1 m	150
EOE	7	< 100 m	350
Silicon Nanophotonics	2 - 3	km	100 - 150

**Power will ultimately limit module escape BW to $\sim 25 - 50 \text{ Tb/s}$
Reducing power of Si nanophotonic devices is key...**

Cost Analysis



- **Cost comparisons need system approach**
 - Design + BOM + Assembly + Test
 - Same system optimized for electrical or optical technology – difficult
- **HPC system power and cost needs are clear:**

Year	Peak Performance	(Bidi) Optical Bandwidth	Optics Power Efficiency (mW/Gb/s)	Optics Power Consumption	Cost (\$/Gb/s, aggressive)	Optics Cost
2008	1PF	0.012PB/s (1.2×10^5 Gb/s)	100	0.010 MW	10	\$1 M
2012	10PF	1PB/s (10^7 Gb/s)	60	0.50 MW	1.0	\$9 M
2016	100PF	20PB/sec (2×10^8 Gb/s)	10	2 MW	0.2	\$30 M
2020	1000PF (1EF)	400PB/sec (4×10^9 Gb/s)	3	10 MW	0.025	\$80 M

Exascale: I/O consume power and \$\$ of system...

(Table after Benner, 2009)

Summary

- **Optical links provide escape bandwidth advantage**
 - Rack-to-rack links have been optical
 - Off-card links just hitting BW wall (P7iH)
 - Off-module links will hit wall this decade
 - Costly to exceed 25 Gb/s in dense electrical buses
- **Power issues will dominate I/O bandwidth growth**
 - Electrical I/O power efficiency improving greatly
 - EOE solutions provide distance advantage, but higher power
 - Si nanophotonics solutions need focus to reduce I/O power
- **Optics will see increasing use as electrical BW limits approached**
 - Electrical solutions growing more costly
 - Optical technology cost takedowns must continue for Exascale and general use
 - Integrated Silicon Nanophotonics offers greatest promise to push BW

References

- 1) Barcelona Supercomputing Center (BCS), Barcelona, Spain; IBM Mare Nostrum system installed in 2004
- 2) T.J. Beukema, "Link Modeling Tool," Challenges in Serial Electrical Interconnects, IEEE SSCS Seminar Fort Collins, CO, March 2007.
- 3) S. Rylov, S. Raynolds, D. Storaska, B. Floyd, M. Kapur, T. Zwick, S. Gowda, and M. Sorna, "10+ Gb/s 90-nm CMOS serial link demo in CBGA package," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 9, pp. 1987-1991, Sept. 2005.
- 4) L. Shan, Y. Kwark, P. Pepeljugoski, M. Meghelli, T. Beukema, J. Trehwella, M. Ritter, "Design, analysis and experimental verification of an equalized 10 Gbps link," *DesignCon 2006*.
- 5) J.F. Bulzacchelli, M. Meghelli, S.V. Rylov, W. Rhee, A.V. Rylyakov, H.A. Ainspan, B.D. Parker, M.P. Beakes, A. Chung, T.J. Beukema, P.K. Pepeljugoski, L. Shan, Y.H. Kwark, S. Gowda, and D.J. Friedman, "A 10-Gb/s 5-tap DFE/4-tap FFE transceiver in 90-nm CMOS technology," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 12, pp. 2885-2900, Dec. 2006.
- 6) D.G. Kam, D. R. Stauffer, T.J. Beukema and M. B. Ritter, "Performance comparison of CEI-25 signaling options and sensitivity analysis," OIF Physical Link Layer (PLL) working group presentation, November 2007.
- 7) M.B. Ritter et. al. "The Viability of 25 Gb/s On-board Signaling" 28th ECTC, May 2008.
- 8) F. Doany et al.: "Measurement of optical dispersion in multimode polymer waveguides", LEOS Summer Topical Meetings, June 2004.
- 9) Alan Benner, "Optics in Servers – HPC Interconnect and Other Opportunities," IEEE Photonics Society - Winter Topicals 2010, Photonics for Routing and Interconnects, January 11, 2010

This work was supported in part by Defense Advanced Research Projects Agency under the contract numbers HR0011-06-C-0074, HR0011-07-9-002 and MDA972-03-0004.