Power-Efficient, High-Bandwidth Optical Interconnects for High Performance Computing

Fuad Doany, IBM T. J. Watson Research Center

Acknowledgements & Disclaimer

- IBM Colleagues
  - C.Schow, F. Libsch, A. Rylyakov, B. Lee, D. Kuchta, P. Pepeljugoski, C. Baks, C. Jahnes, R. Budd, J. Proesel, J. Kash, Y. Kwark, C. Tsang, J. Knickerbocker, Y. Vlasov, S. Assefa, W. Green, B. Offrein, R. Dangel, F. Horst, S. Nakagawa, Y. Taira, Y. Katayama, A. Benner, D. Stigliani, C. DeCusatis, H. Bagheri, K. Akasofu, B. Offrein, R. Dangel, S. Nakagawa, M. Taubenblatt, M. Soyuer and many others…
- Emcore Corporation
  - N.Y. Li, K. Jackson
- Endicott Interconnects
  - B. Chan, H. Lin, C. Carver

- IBM's Terabus project partially supported by DARPA under the Chip to Chip Optical Interconnect (C2OI) Program

The opinions expressed in this presentation are those of the Author and do not necessarily represent the position of any Government funding agencies.

Evolution of Optical interconnects

Time of Commercial Deployment (Copper Displacement):

1980's | 1990's | 2000's | > 2012

WAN, MAN
metro,long-haul

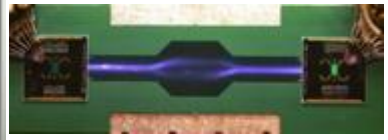
Telecom

LAN
campus, enterprise

Datacom

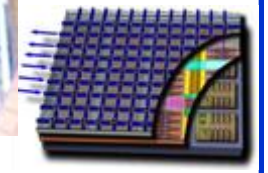
System
intra/inter-rack

Board
module-module



Module
chip-chip

IC
on-chip



Computercom

- BW*distance advantage of optics compared to copper leading to widespread deployment at ever-shorter distances
- As distances go down the number of links goes up putting pressure on power efficiency, density and cost

Increasing integration of Optics with
decreasing cost, decreasing power, increasing density

Outline

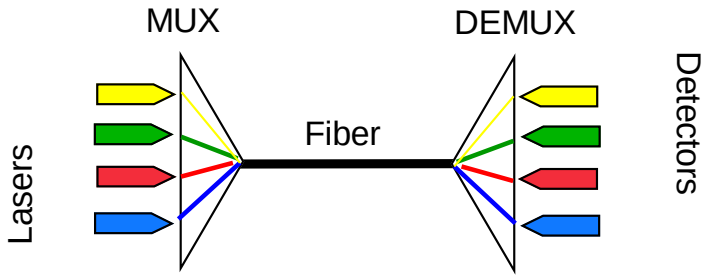
- Brief Intro to Fiber Optics Links
- Fiber Optics in HPC
  - Evolution of optical interconnects in HPC systems
  - System needs and power, cost and density challenges
- Path to Optimizing power and efficiency
  - Packaging Integration
  - Optical-PCB Technology
  - Chip-scale Integration: Generations of Parallel VCSEL Transceivers
  - Optical Link Improvements
  - New Technologies: Si Photonics, Multicore Fiber, CWDM
- Pushing the Limits of Speed and Power
  - Equalization for improved speed and margin
  - Fast SiGe circuits to probe VCSEL speed limits
- Concluding Comments

Two Fiber Optics Camps: Telecom and Datacom/Computercom

- **Telecom links** (10's – 1000's of km)
  - Expensive to install fiber over long distances
  - Wavelength Division Multiplexing (WDM)
    - Maximize use of installed fiber
  - Performance is primary objective
  - Component cost secondary
    - Fiber amplifiers, dispersion compensators
    - EML's, external modulators, APD receivers…
  - Reliability and long operating life is critical

WDM = Wavelength Division Multiplexing

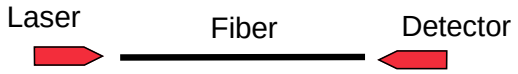
Single optical channel
data carried on separate λ's



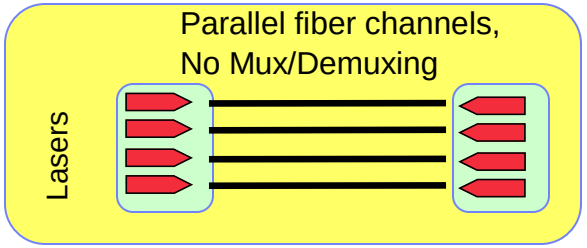
- **Datacom/Computercom links** (100's of meters, or less)
  - Cost is the biggest factor
  - Transceivers are commodities
    - Multimode fiber & optics (relaxed mechanical tolerances)
    - VCSELs, pin receivers
  - Reliability (was) less of an issue: pluggable modules
  - Reach typically not an issue

TDM = Time Division Multiplexing

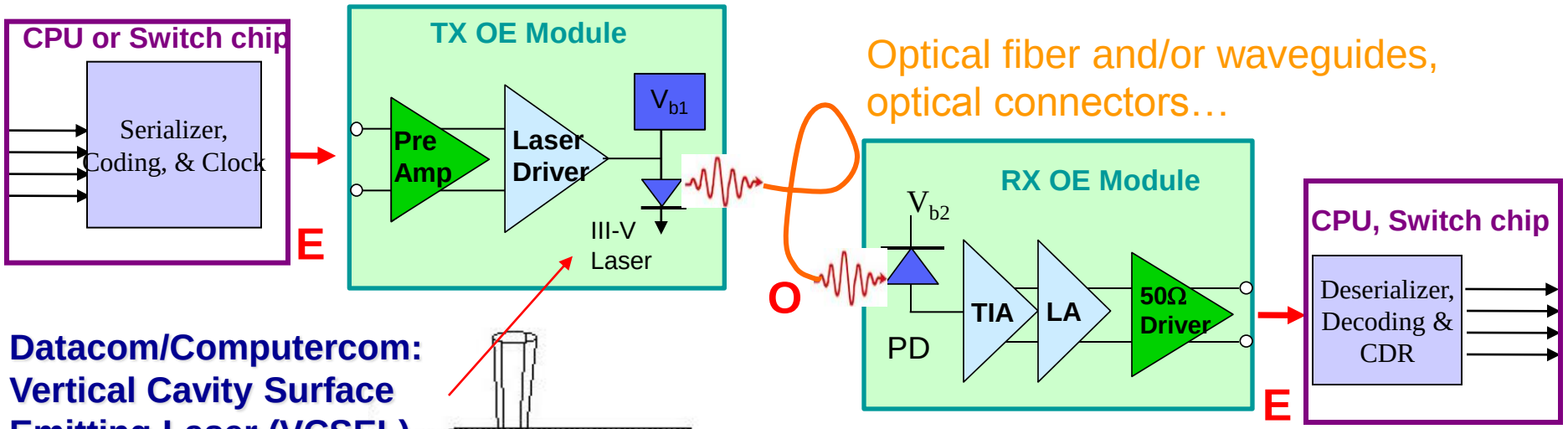
Single optical channel,
Electronic Mux/Demuxing



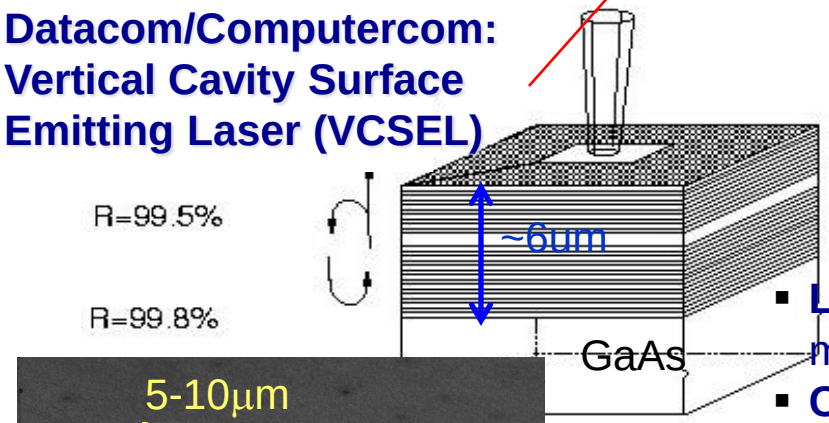
SDM = Space Division Multiplexing

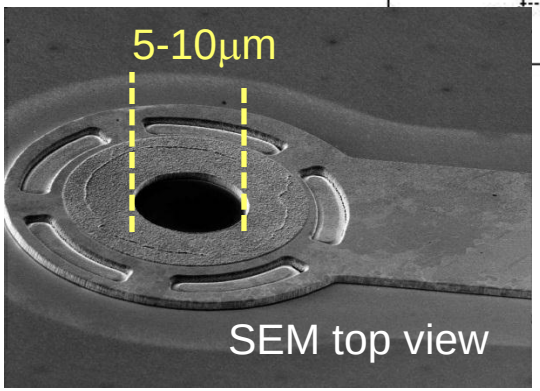


What does an optics link consist of?



Datacom/Computercom: Vertical Cavity Surface Emitting Laser (VCSEL)





- **Low Power Devices:** low-threshold, direct hi-speed modulation
- **Cost:** much cheaper than edge-emitting lasers
  - Wafer-scale fab and test
- **High Density**
  - 2-D arrays possible
- **Temperature control**
  - ~ 40 °C for VCSEL vs. 1 °C for DFB lasers & AWGs
  - No thermoelectric coolers, low power consumption

Today, VCSELs dominate the Datacom/Computer Interconnects: Millions/month Shipping

Datacom VCSELs cost <$1,Optical mouse VCSELs cost “Pennies”

Outline

- Intro to Fiber Optics Links
- **Fiber Optics in HPC**
  - **Evolution of optical interconnects in HPC systems**
  - **System needs and power, cost and density challenges**
- Path to Optimizing power and efficiency
  - Packaging Integration
  - Optical-PCB Technology
  - Chip-scale Integration: Generations of Parallel VCSEL Transceivers
  - Optical Link Improvements
  - New Technologies: Si Photonics, Multicore Fiber, CWDM
- Pushing the Limits of Speed and Power
  - Equalization for improved speed and margin
  - Fast SiGe circuits to probe VCSEL speed limits
- Concluding Comments

Maintaining the HPC Performance Trend

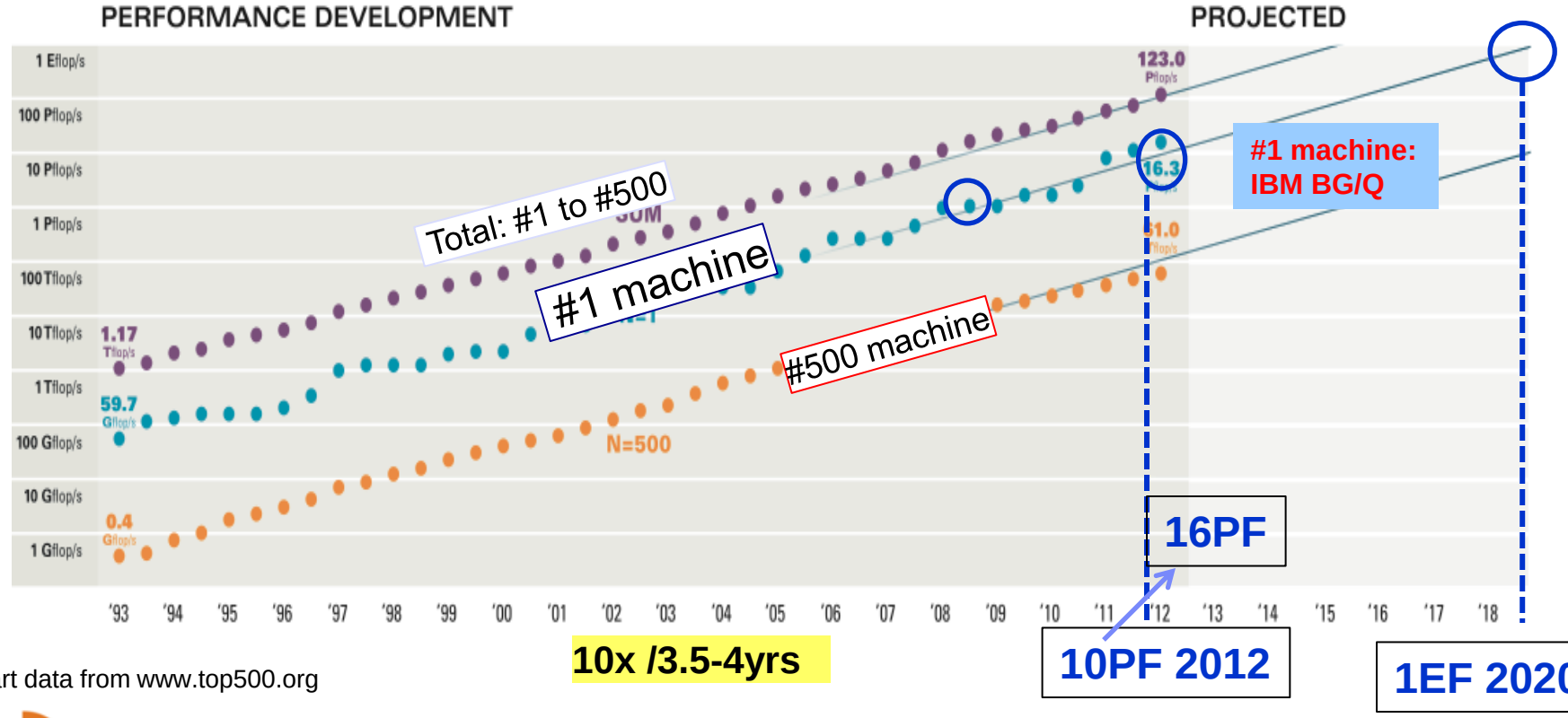
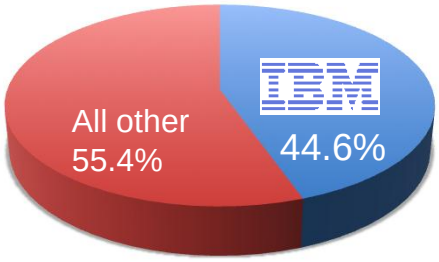


chart data from www.top500.org

TOP500 NOVEMBER 2011

Vendors System Share



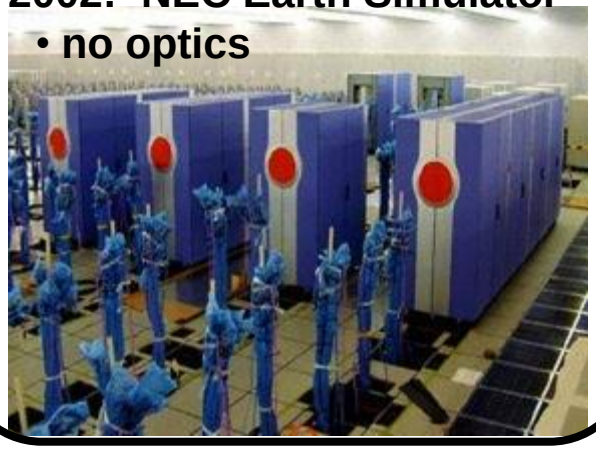
Performance enabled by increased parallelism:

- Processor speed no longer primary driver
- Aggregation of massive numbers of multicore processors
- Challenging interconnect BW demands across system hierarchy
  - Intra-chip, inter-chip, on-board, intra-rack, between racks
  - Communication bottlenecks moving closer to processors
  - Optics displacing copper at ever shorter distance scales

Evolution of Parallel Optics in Supercomputers

2002: NEC Earth Simulator

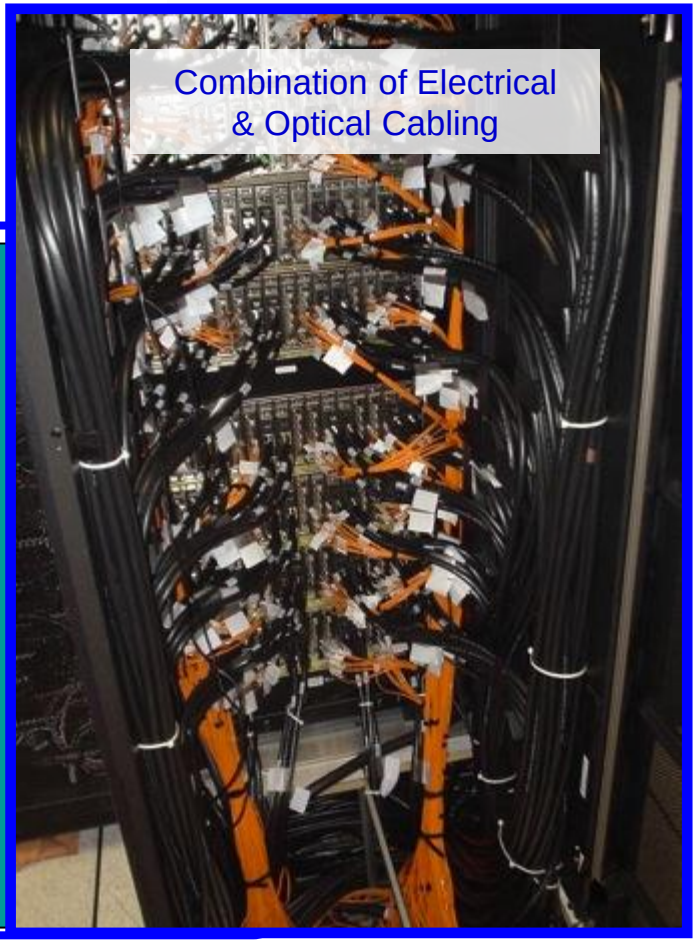
- **no optics**



Snap 12 optical module
12 Tx or Rx at 2.5Gb/s

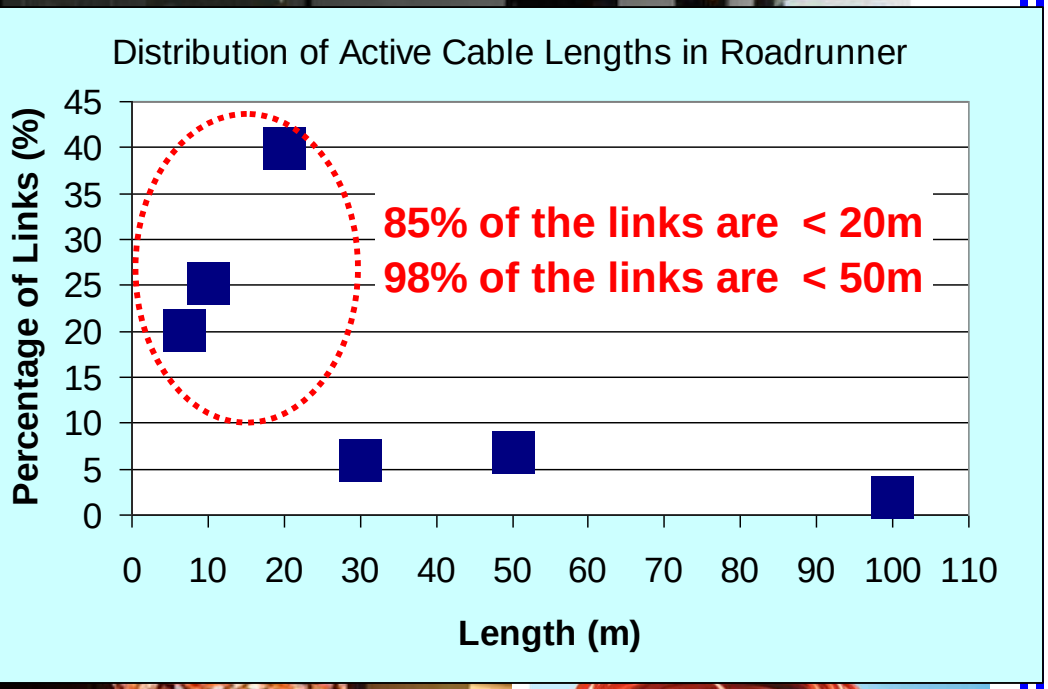
2005:
IBM
ASCI Purple
Server (LLNL)

100 TeraFLOP/s
~3000 parallel links
12+12 @ 2.5Gb/s/ch
Optics for >20m links



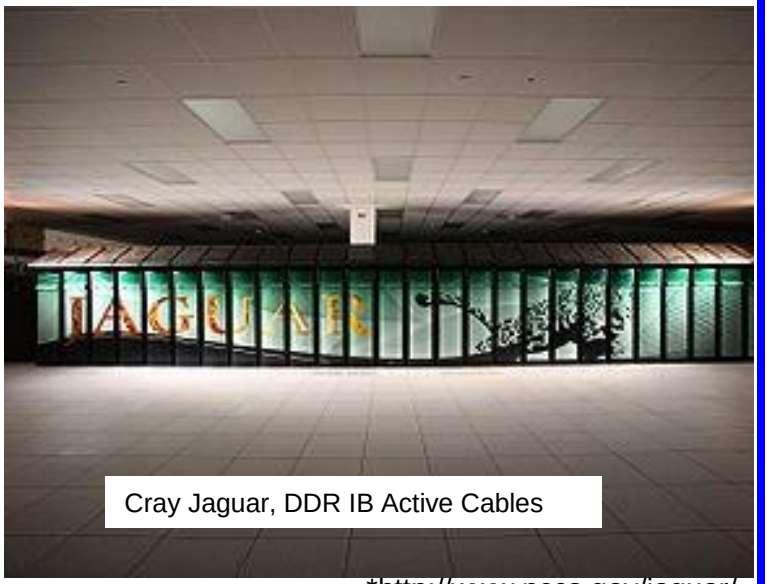
2008: PetaFlop Computers

LANL RoadRunner built by IBM



Fiber to the Rack: 40,000 optical links

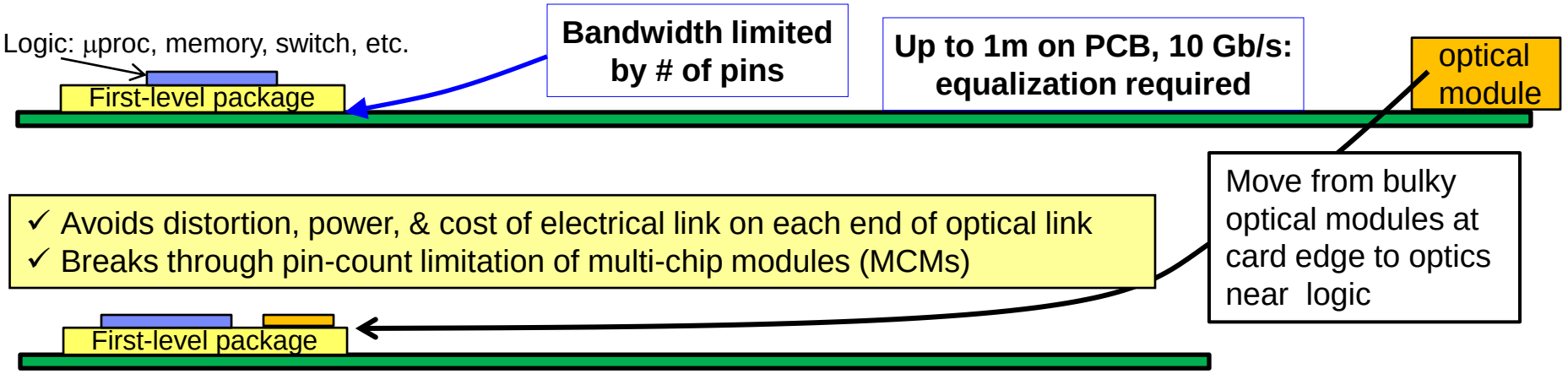
Cray Jaguar



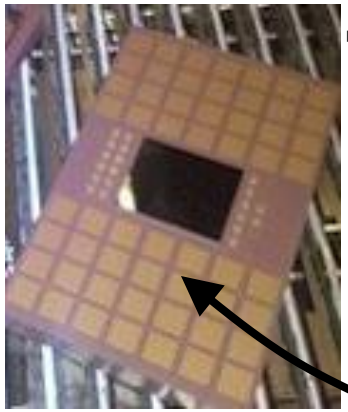
*http://www.nccs.gov/jaguar/

~3000 DDR IB Active Cables
- 3 miles of optical cables
- Up to 60m
- Spread over 2 floors

Optics close to logic, rather than at card edge:

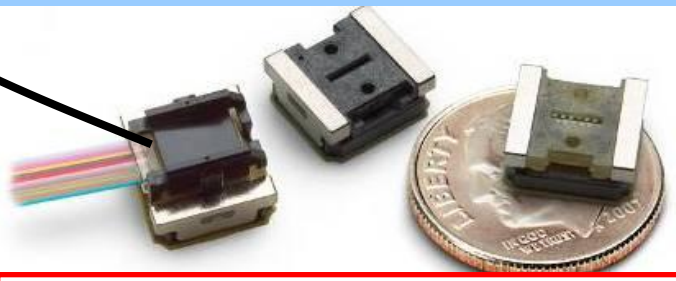


2011: This Packaging Implemented in IBM Power 775 System

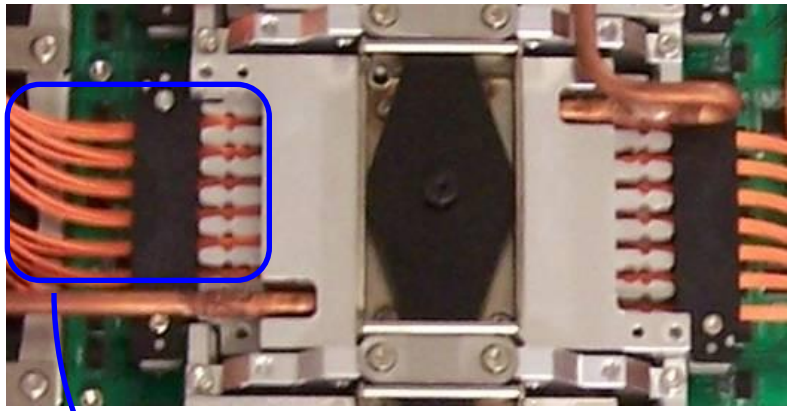


- Hub/switch module, with parallel **optical transmitters & receivers mounted on module surface**

- **Avago microPOD™ modules**
  - **12x10Gb/s parallel**
- **28TX+28RX per hub module**

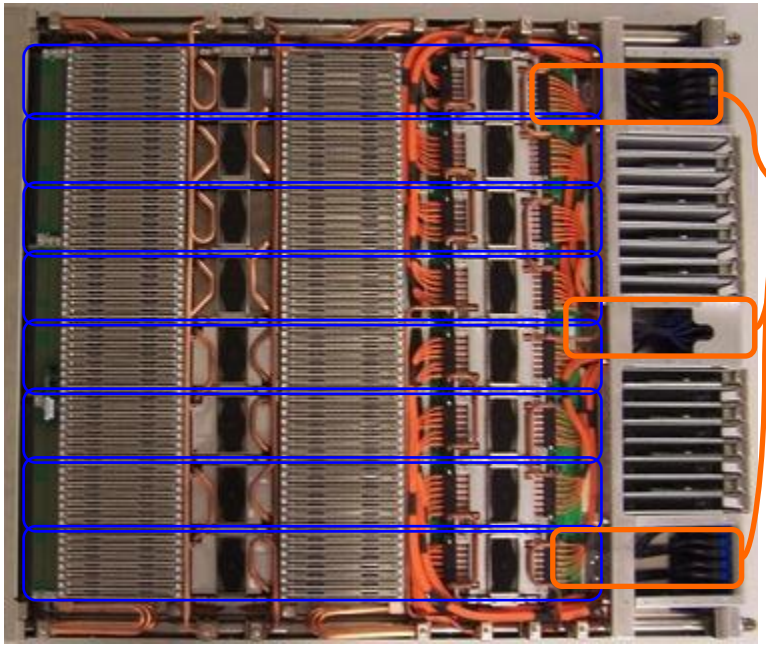


M. Fields, "Transceivers and Optical Engines for Computer and Datacenter Interconnects", OFC 2010



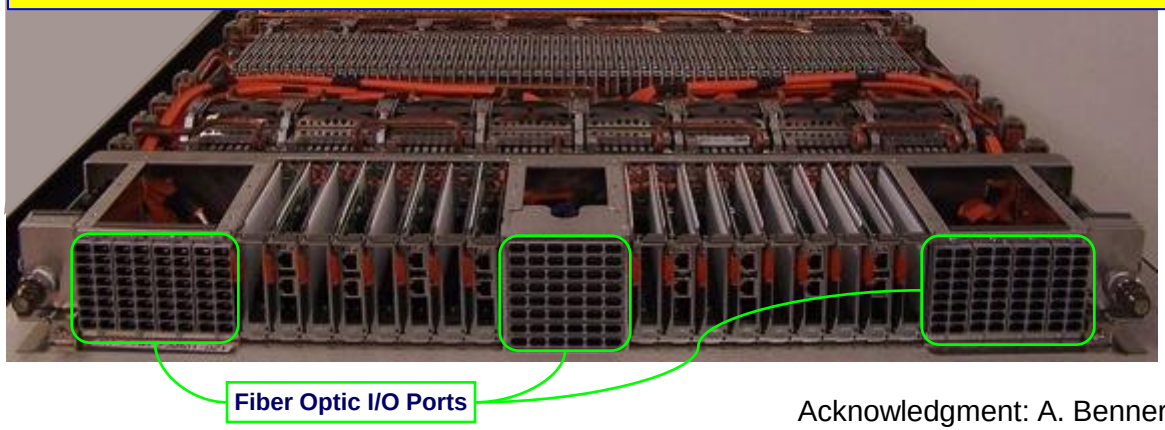
Optical I/Os – Fiber Ribbons

2011: IBM Power 775, Intra-Rack Parallel Optics



256-core Node Drawer

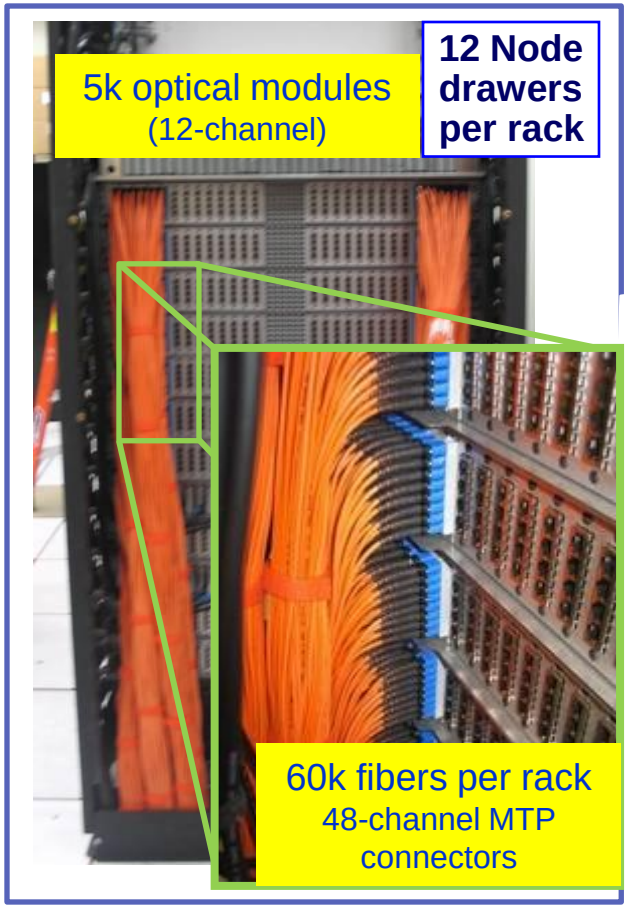
- **Optical transceivers tightly integrated, mounted within drawer**
- **8 Hub/switch modules (8 x 56 optical modules)**



Acknowledgment: A. Benner

P775 Drawer

- **8 32-way SMP nodes**
- **Per SMP node:**
  - 1 TF
  - 128 GB DRAM
  - >512 MB/s memory BW
  - **>190 GB/s network BW**

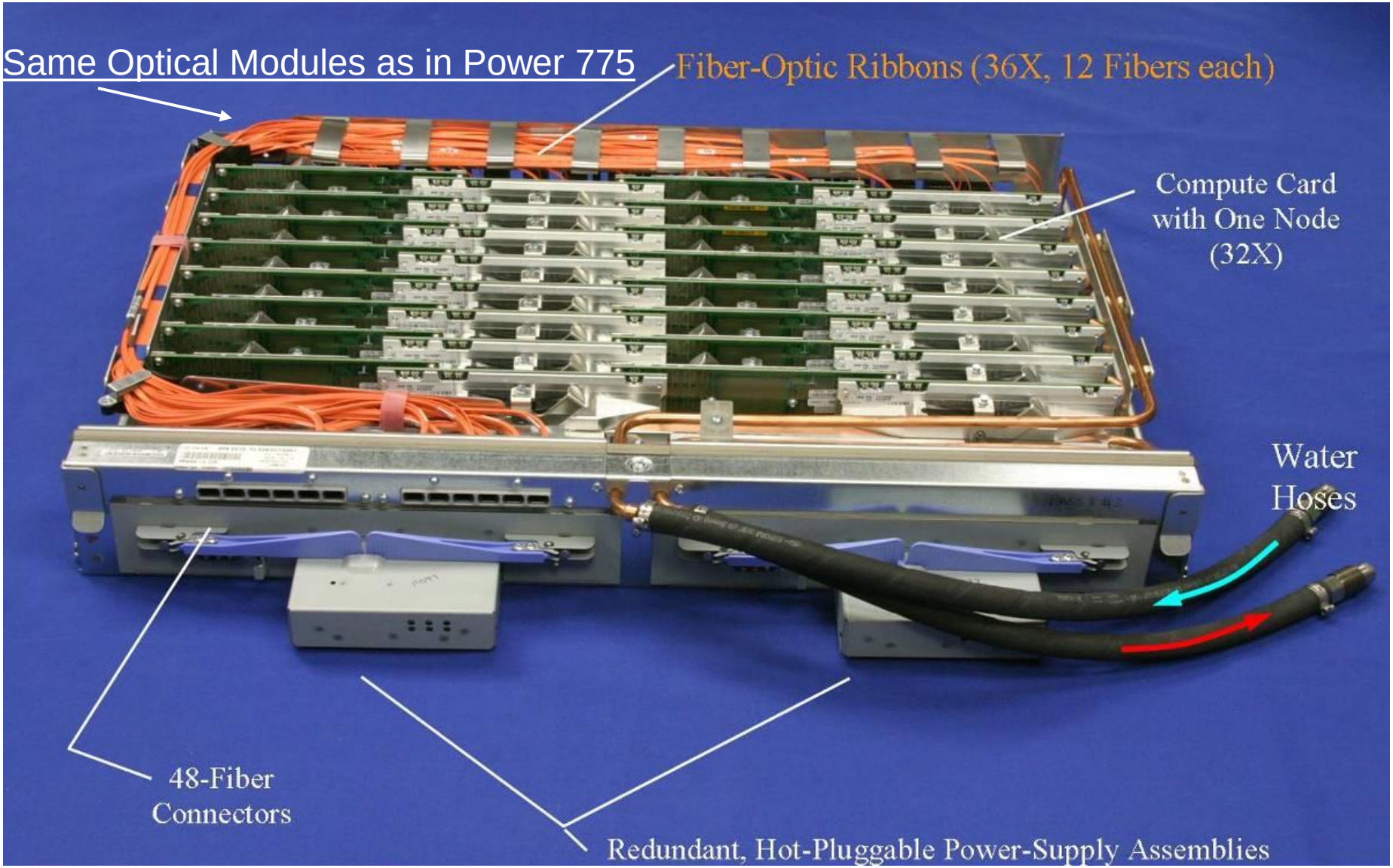


2012: Blue Gene/Q

Sequoia - (96) IBM Blue Gene/Q Racks
20.013 Pflops Peak … 1.572M Compute Cores … ~2026 MFlops/Watt

~8MW

BG/Q Compute Drawer



Exascale Blueprint: U.S. Department of Energy (DOE) RFI

Re-constructed from RFI:
Table 1. Exascale System Goals

Issued 7/11/2011 (1-KD73-I-31583-00)
Available: www.fbo.gov

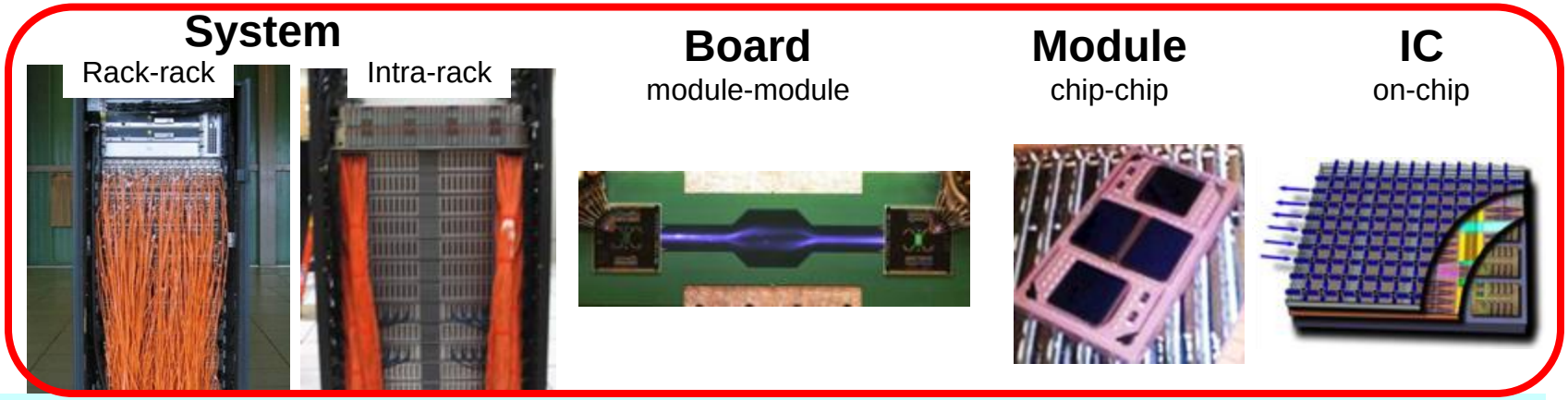
Exascale System	Goal
Delivery Date	2019-2020
Performance	1000 PF LINPACK and 300 PF on to- be-specified applications
Power Consumption*	20 MW
MTBAI** 6 days	6 days
Memory including NVRAM	128 PB
Node Memory Bandwidth	4 TB/s
Node Interconnect Bandwidth	400 GB/s

- 20 MW total system power
- Assume 400 GB/s off-node BW is all Optical
- Assume a relatively lightly interconnected system at **0.1 Byte/F**

- **Every pJ/bit in optical link power results in a total contribution of 0.8 MW to system power**
- **Every 10¢/Gb/s in optical link cost translates into $80M in system cost**

- How much power can be devoted to interconnect?
  - **At todays numbers of ~25 pJ/bit, total network power = system power target = 20MW**
  - **Maybe 5 pJ/bit? Would be 20% of system power…**

Computercom Driving Development and Large-Scale Deployment of Parallel Optical Transceivers



Distance:	~100 m	Few m	~ 1 m	< 10 cm	< 20 mm
Bus width:	10s	100s	**1000's**	**10,000's**	**>10,000's …**

Rack-to-rack,	**intra-rack,**	**module-to-module,**	…	**3-D chip** With photonic layer
Extensively Deployed Today	**2011**	**>2012**		>2020
Conventional Optical Modules Edge of card packaging	**Dense, parallel fiber-coupled modules, Close to CPU**	**Integrated transceivers & Optical-PCBs**		Si Photonics

Short-Reach Optics Optimized for
Power (mW/Gb/s = pJ/bit), Cost ($/Gb/s) & Density

- Future High Performance Computers will demand pJ/bit power efficiencies at ¢/Gb/s

Outline

- Intro to Fiber Optics Links
- Fiber Optics in HPC
  - Evolution of optical interconnects in HPC systems
  - System needs and power, cost and density challenges
- **Path to Optimizing power and efficiency**
  - **Packaging Integration**
  - **Optical-PCB Technology**
  - **Chip-scale Integration: Generations of Parallel VCSEL Transceivers**
  - **Optical Link Improvements**
  - **New Technologies: Si Photonics, Multicore Fiber, CWDM**
- Pushing the Limits of Speed and Power
  - Equalization for improved speed and margin
  - Fast SiGe circuits to probe VCSEL speed limits
- Concluding Comments

Path to Optimizing Link Power Efficiency

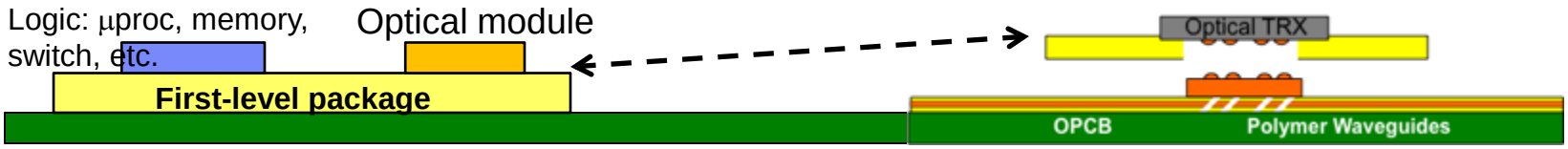
Packaging Integration:

✓ Optics co-packaging

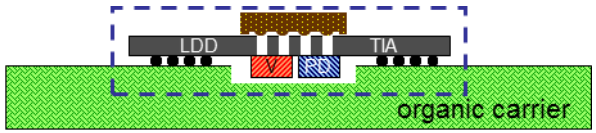
- Minimize power in electrical link from logic to optics
  - drive across chip carrier instead of board
- High BW density electrical/optical interfaces

Optical-PCBs

- **PCBs with integrated polymer waveguides**
- **High BW density optical interfaces**

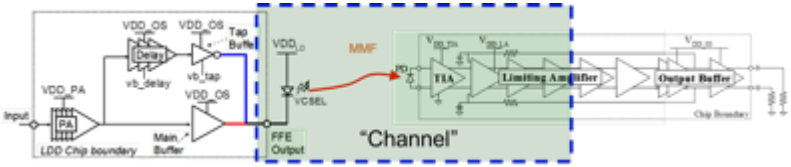


Chip-Scale Integration



- Optochips: chip-like optical transceivers
- Flip-chip packaging enabling dense 2-D arrays
- Direct OE to IC attachment for maximum performance

Optical Link Improvements



- Advanced CMOS for high-speed and low power
- Faster, more efficient VCSELs and PDs
- Equalization to improve link performance and margin

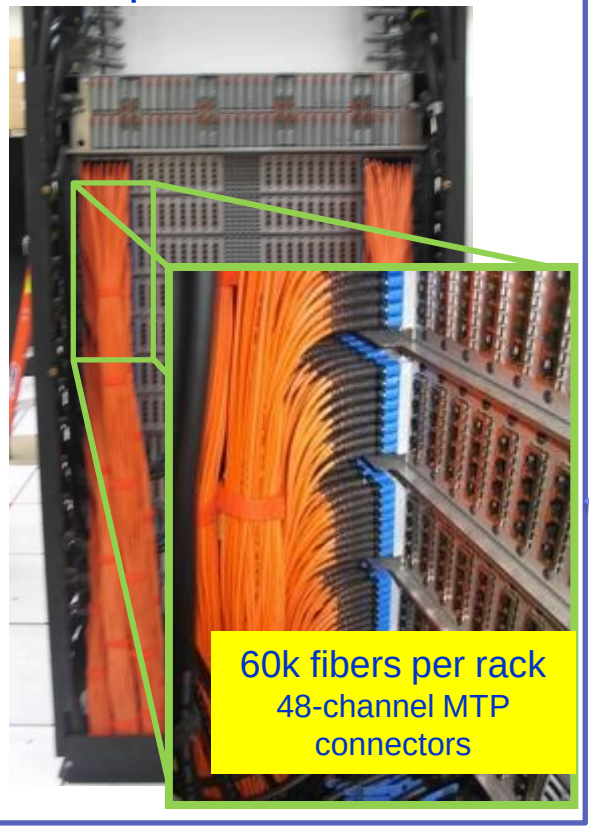
New Technologies, eg. Si Photonics

- Potential for low power, high bandwidth transceivers
- Longer reach through SMF
- Primary advantage is WDM for high BW density

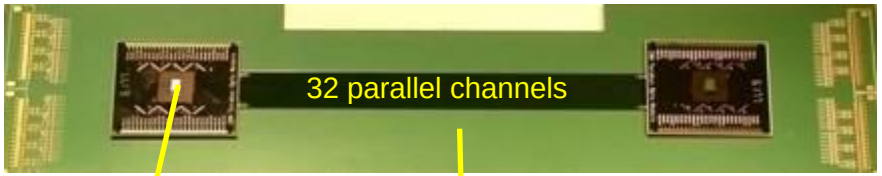
Optical PCB's with Integrated Transceivers: Key to Lower Cost, Tighter Integration

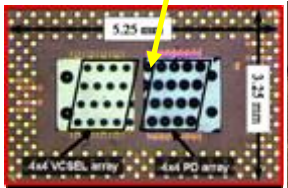
From Fibers and modules…

2011: Wiring with ribbon fiber pushed to the limit

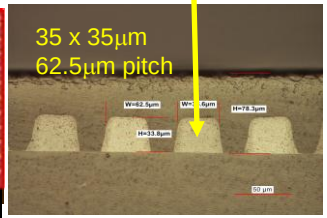


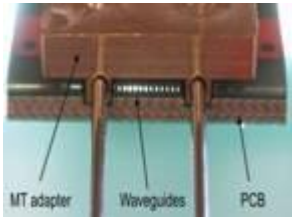
… to integrated waveguides on PCBs with optical components



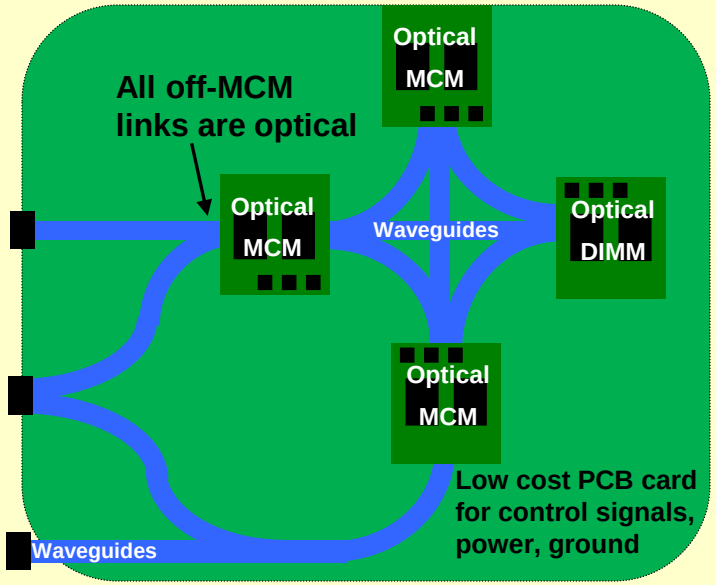


Terabus 160Gb/s TRx (bottom view)





Low Cost Optical Printed Circuit Boards (Polymer Waveguides)

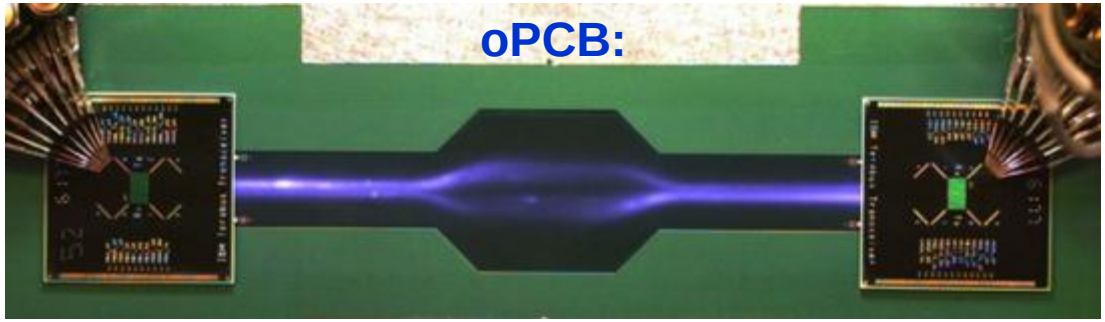


Vision: Optical MCMs

Optics co-packaged with logic

Advantages

- Low cost pick and place assembly
- Passive routing functions: shuffles, splits
- Bring optics close to chips for maximum performance and efficiency
- Enables use of low-cost PCBs – eliminates design challenges for high-speed electrical links



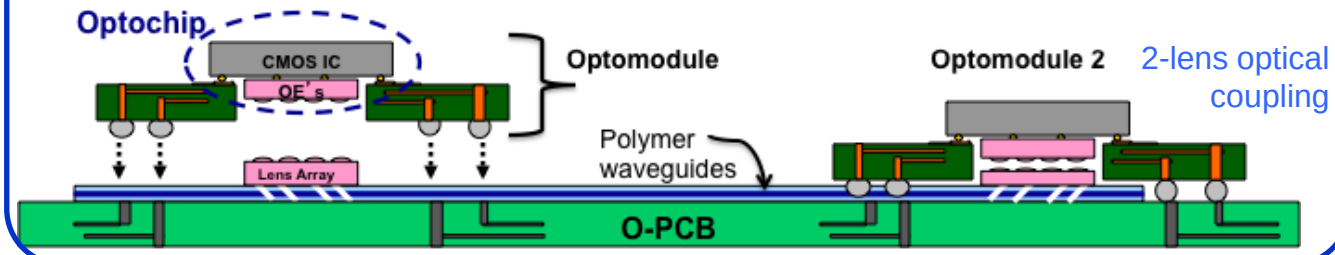
Complete Technology Demonstrated:

PCB with polymer waveguides

Chip-Scale Transceivers "Optochips"

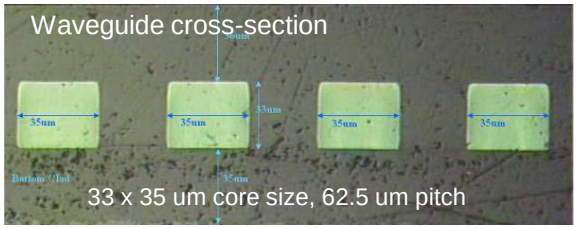
Optical MCMs
Optochips on MCM

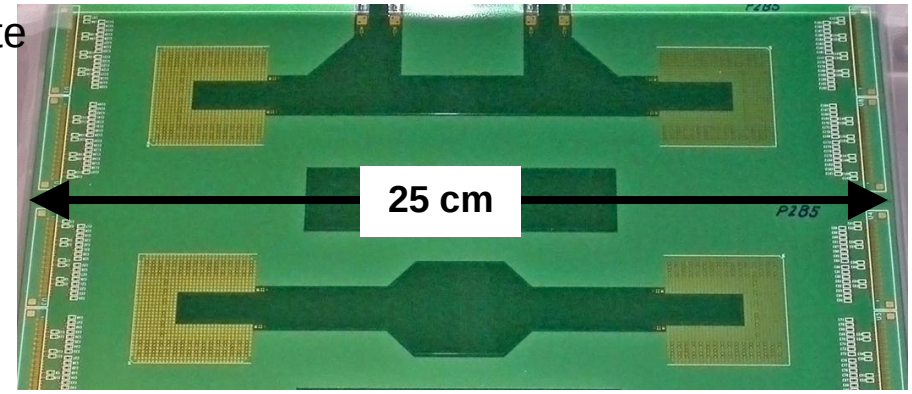
oPCB: Polymer WG on board or flex



Optical-PCB Technology: Waveguides, Turning Mirrors, Lens Arrays

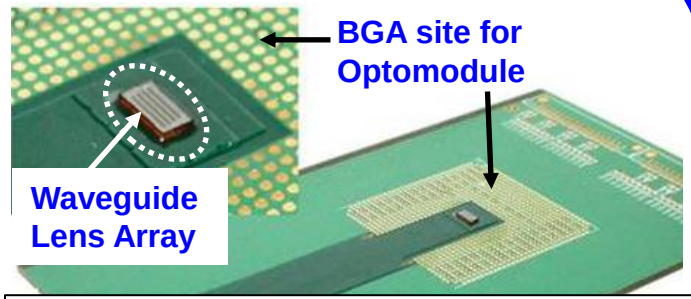
- Polymer waveguides on low-cost FR4 substrate
- Lithographic patterning
  - 48 channels, 35μm core, 62.5 μm pitch



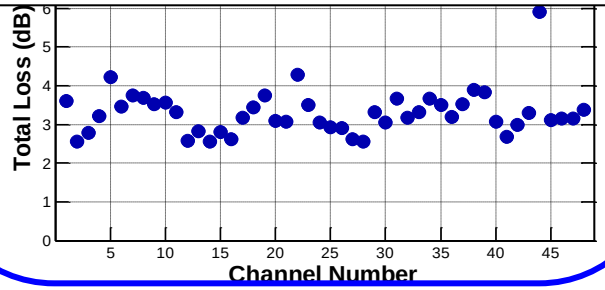


Waveguide on flex

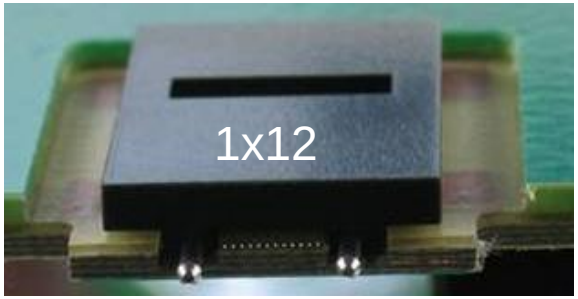
O-PCB

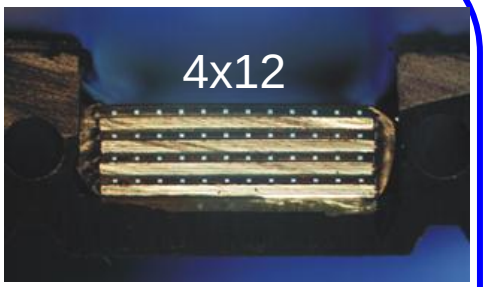


Waveguides, turning mirrors, lens array
Low loss (< .05 dB/cm)
Uniform – 48 WGs



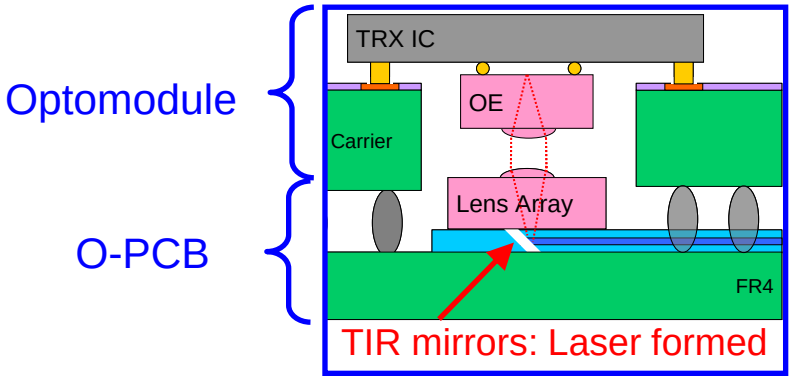
WG-to-MMF connector



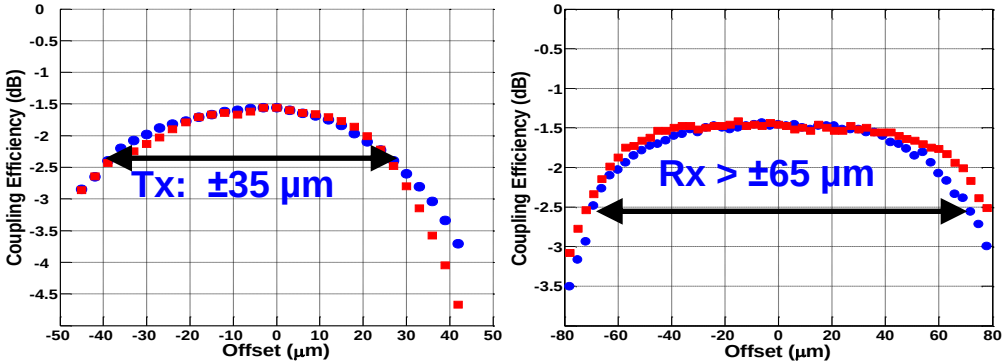


Optical-PCB Technology: Full Link Assembly

2-Lens optical system

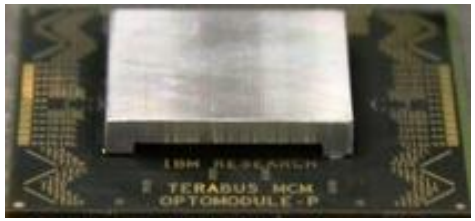


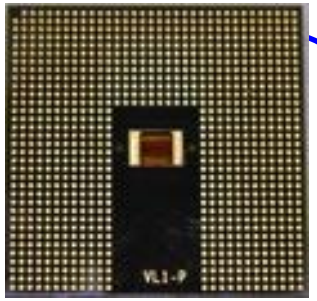
Efficient coupling, relaxed tolerances

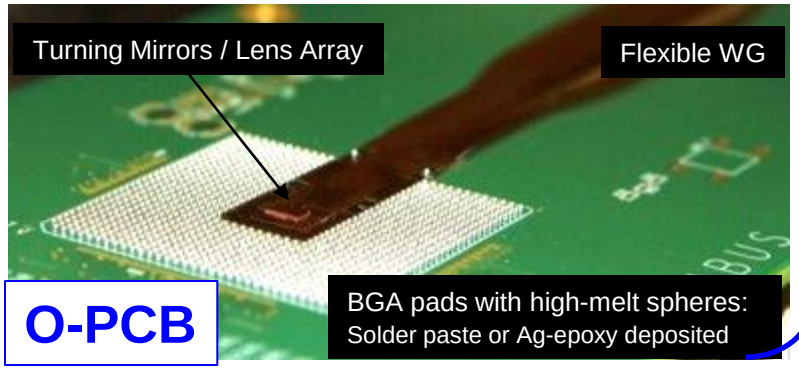


Compatible with pick-and-place tooling (~25um)

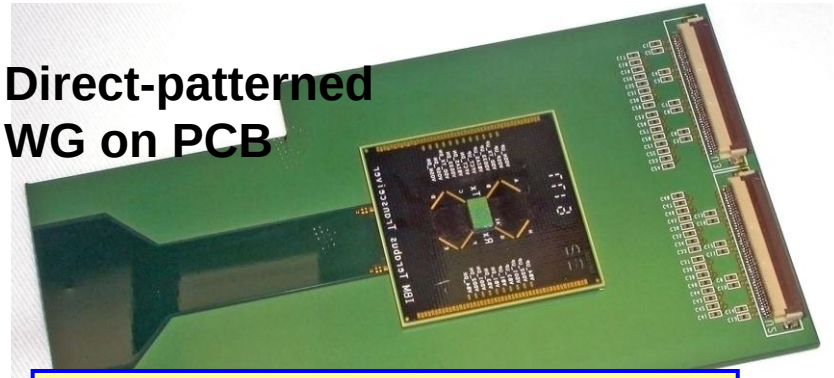
Optomodule: with heat sink and lens array



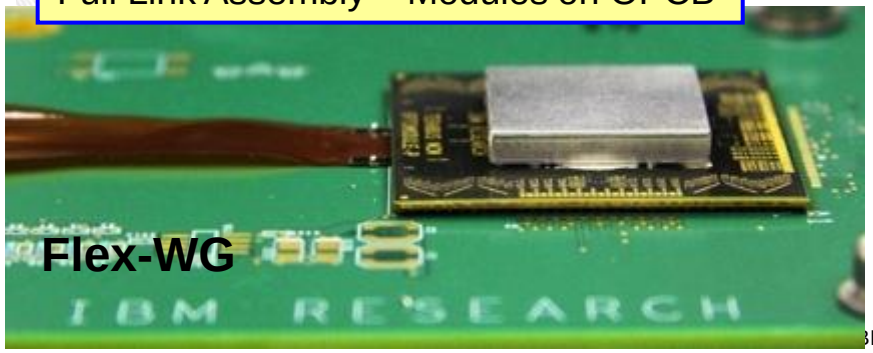




O-PCB



Full Link Assembly – Modules on OPCB



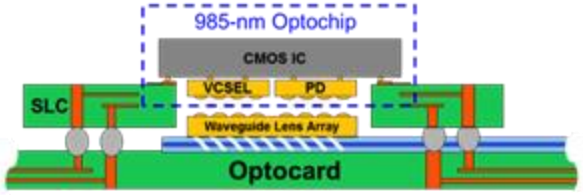
Path to Tb/s Modules: Three Generations of Parallel Transceivers

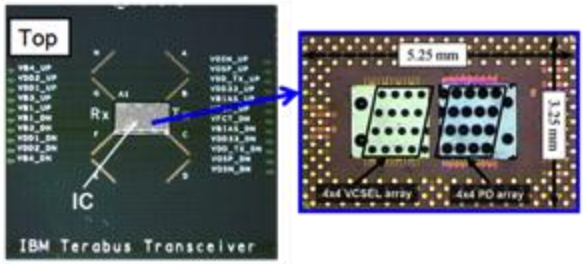
2008: 240 + 240 Gb/s

2010: 360 + 360 Gb/s

2012: 480 + 480 Gb/s

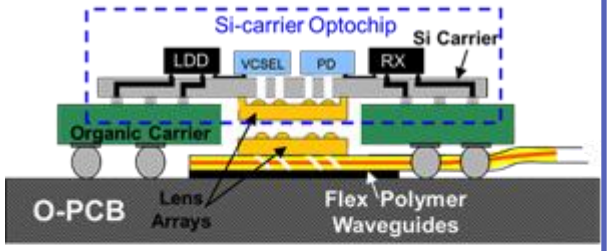
985-nm Optochip

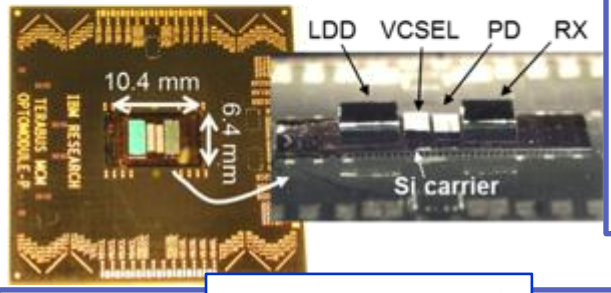




28.1 Gb/s/mm^2

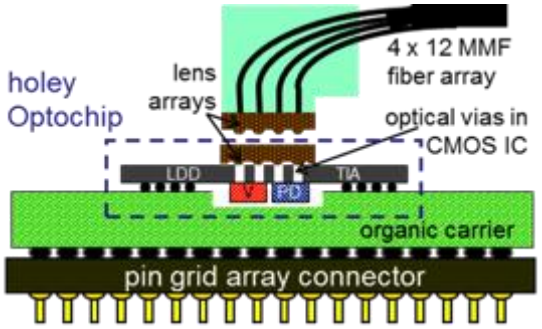
Si-Carrier Optochip

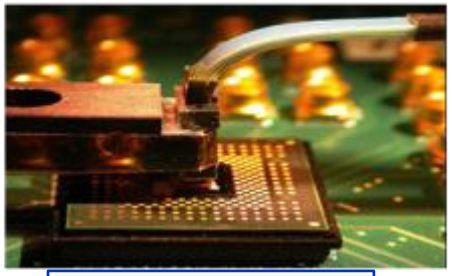




10.8 Gb/s/mm^2

holey Optochip

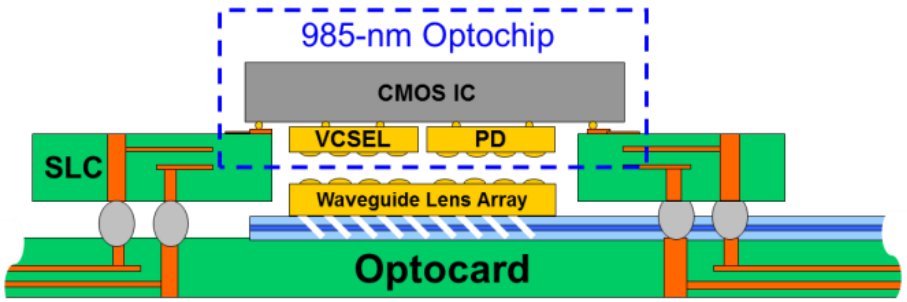




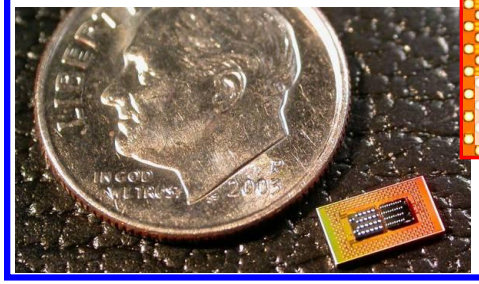
31.8 Gb/s/mm^2

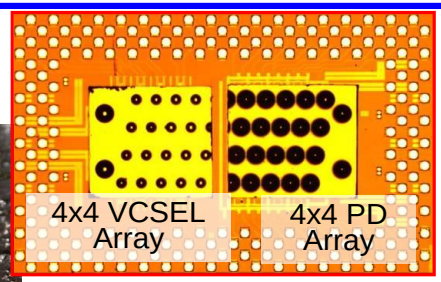
- Exclusive use of flip-chip packaging for maximum performance and density
- Chip-scale packages → Optochips
- Packaging for WG and direct fiber coupling

Optical PCB Realized: 985-nm, 160 Gb/s Bidirectional Link (2008)

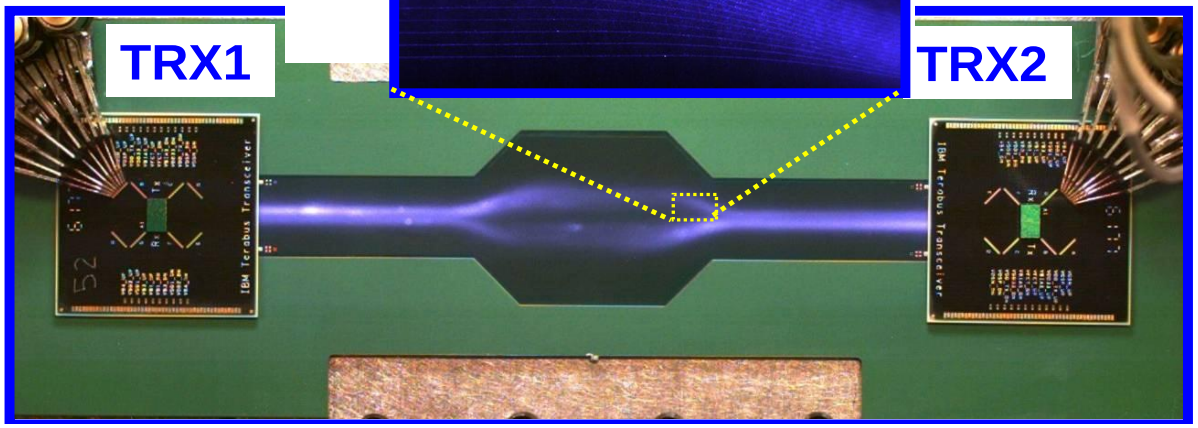


Optochip: CMOS + flip-chip OEs



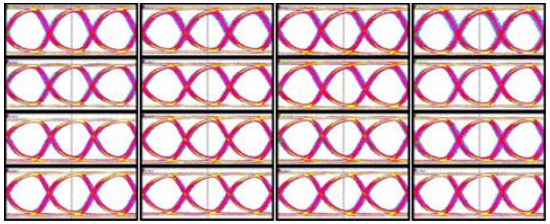


16+16, 985-nm
3mm x 5mm

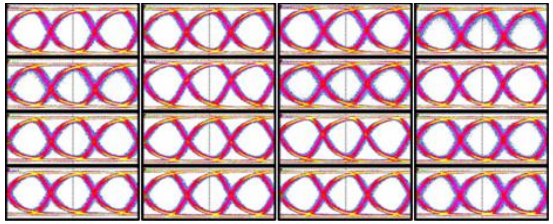


- 10 Gb/s max per channel (thru WG)
- 13.5 pJ/bit
- 130 nm CMOS

16 x10Gb/s TRX1 → TRX2

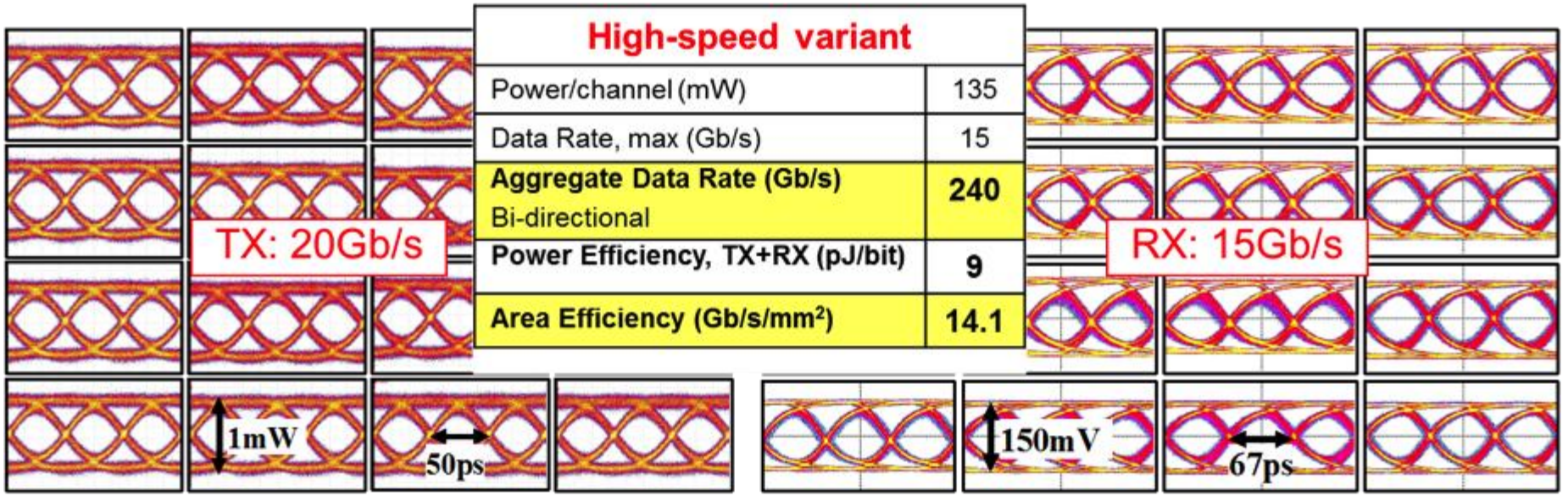


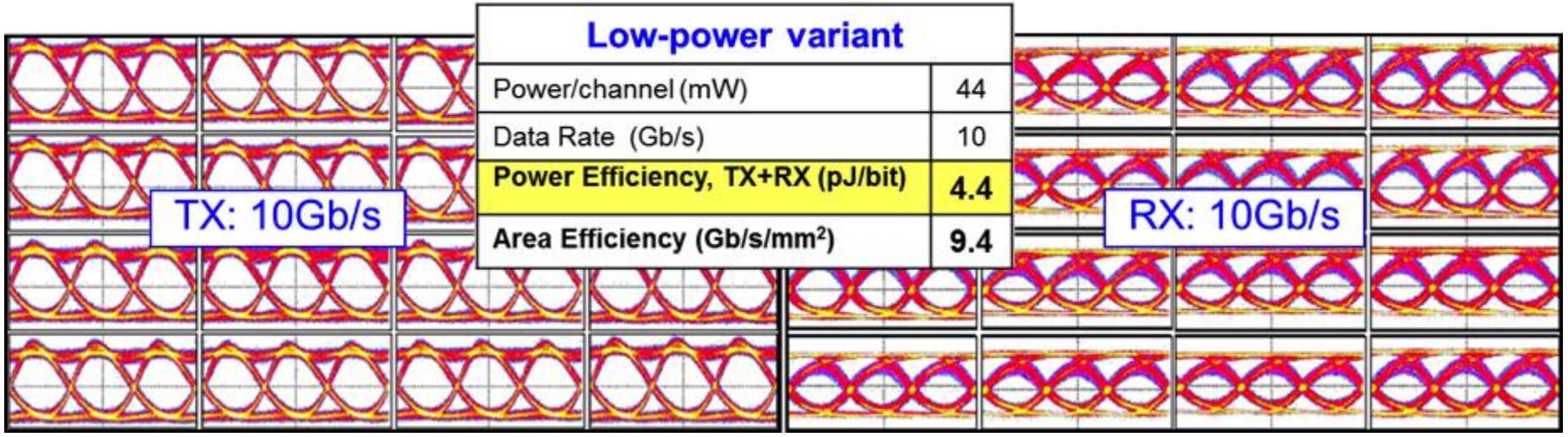
16 x10Gb/s TRX2 → TRX1



F. E. Doany, "160 Gb/s Bidirectional Polymer Waveguide Board-Level Optical Interconnects using CMOS-Based Transceivers," IEEE Adv. Packag., May 2009.

985-nm Transceivers: High-speed, Low power

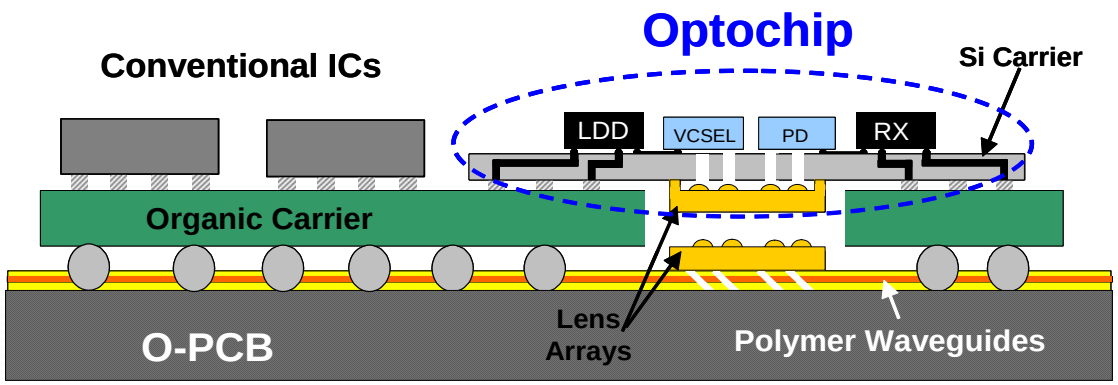




C. L. Schow *et al.*, “A single-chip CMOS-based parallel optical transceiver capable of 240 Gb/s bi-directional data rates,” *IEEE JLT*, 2009.
C. L. Schow *et al.,* "Low-power 16 x 10 Gb/s Bi-Directional Single Chip CMOS Optical Transceivers operating at < 5 mW/Gb/s/link," *IEEE JSSC*, 2009.

Migration to 850-nm Wavelength

- Datacom industry standard wavelength
  - Multiple suppliers, low-cost, optimized MMF fiber bandwidth
- Lower loss in polymer waveguides
  - 0.03dB/cm at 850nm compared to 0.12dB/cm at 985nm
  - **Loss for a 1m link: 850 nm = 3dB, 985 nm = 12dB**
- Retain the highly integrated packaging approach: dense Optomodules that “look” like surface-mount electrical chip carriers
- **Si carrier platform**: high density integration of the electrical and optical components

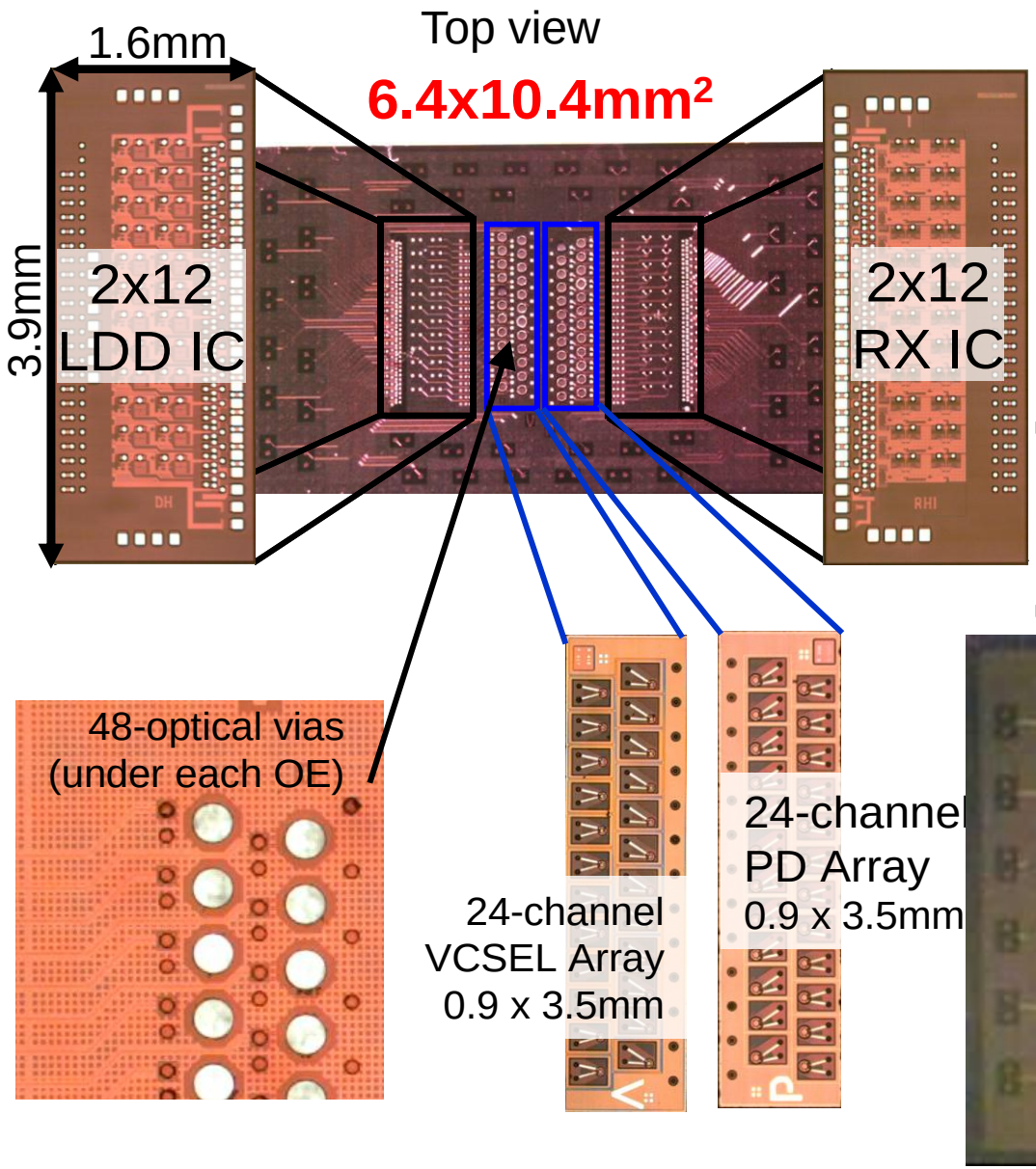


Optically enabled MCM (OE-MCM)

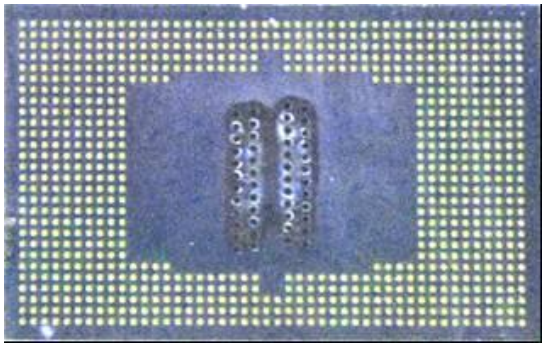
Terabus 850 nm

- 24TX + 24 RX Transceiver
  - 2x12 VCSEL and PD arrays
  - 2 CMOS ICs

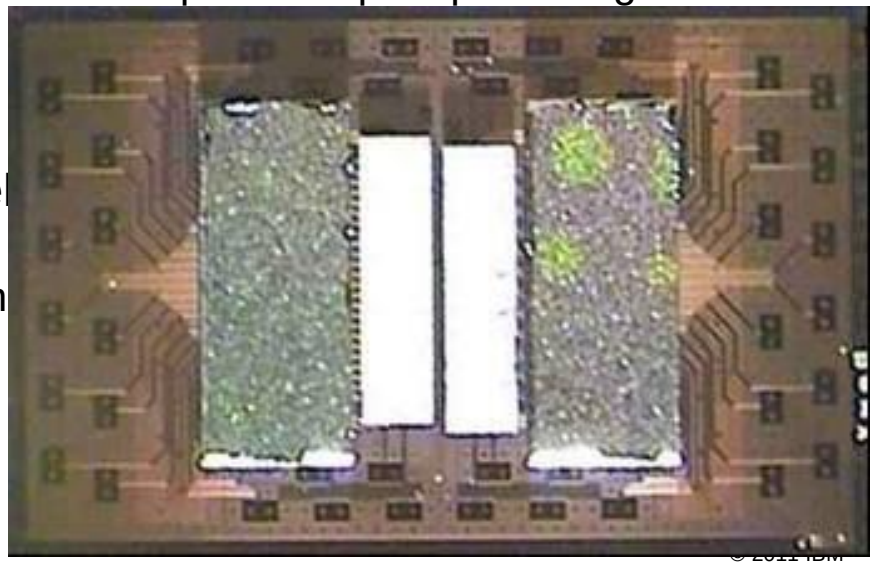
Compact Si-carrier 850-nm Optochips



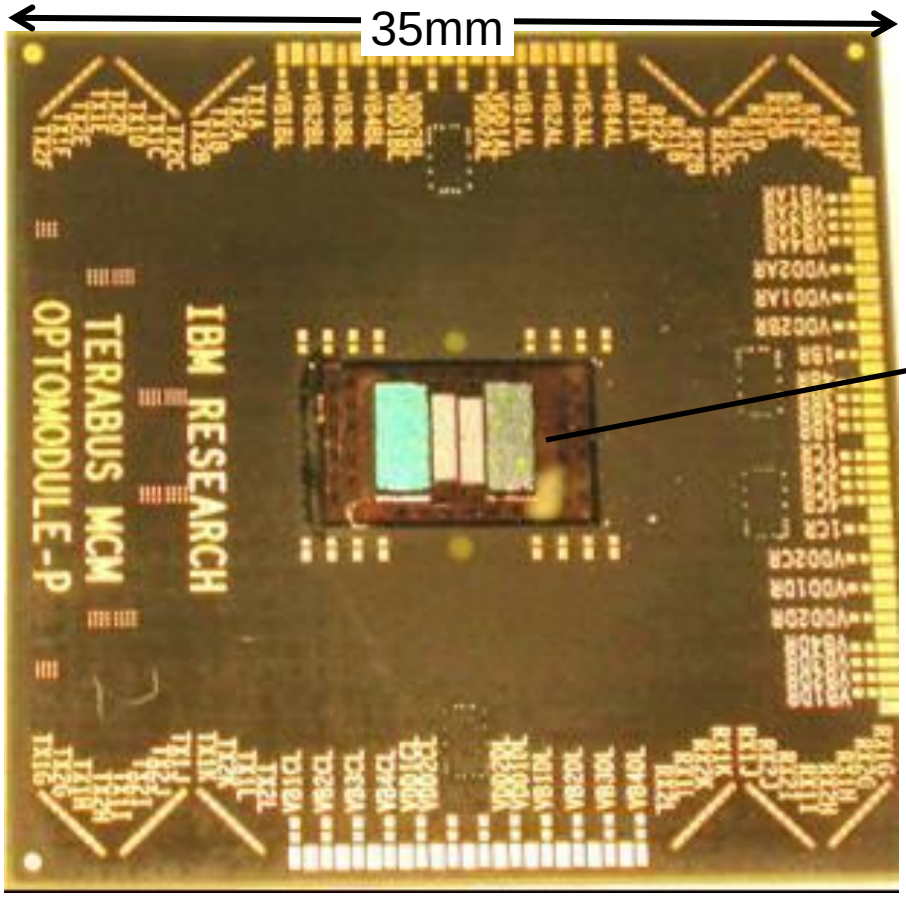
Bottom view

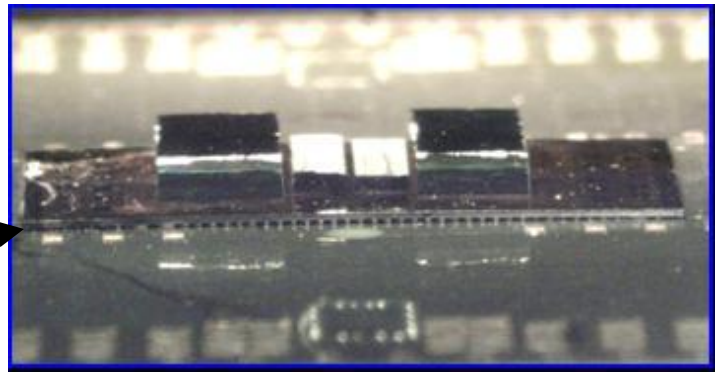


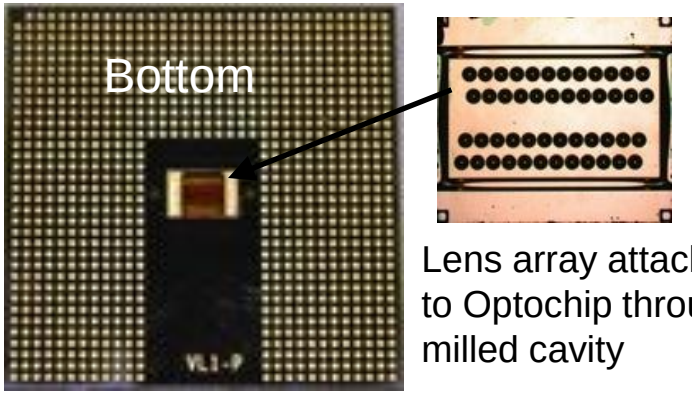
- 150-μm thick Si carrier:
  - 3 surface wiring layers
  - Electrical through-silicon vias (TSVs)
  - 48 Optical vias (ϕ=150μm)
- Sequential flip-chip bonding:



Assembled 850-nm Optomodule



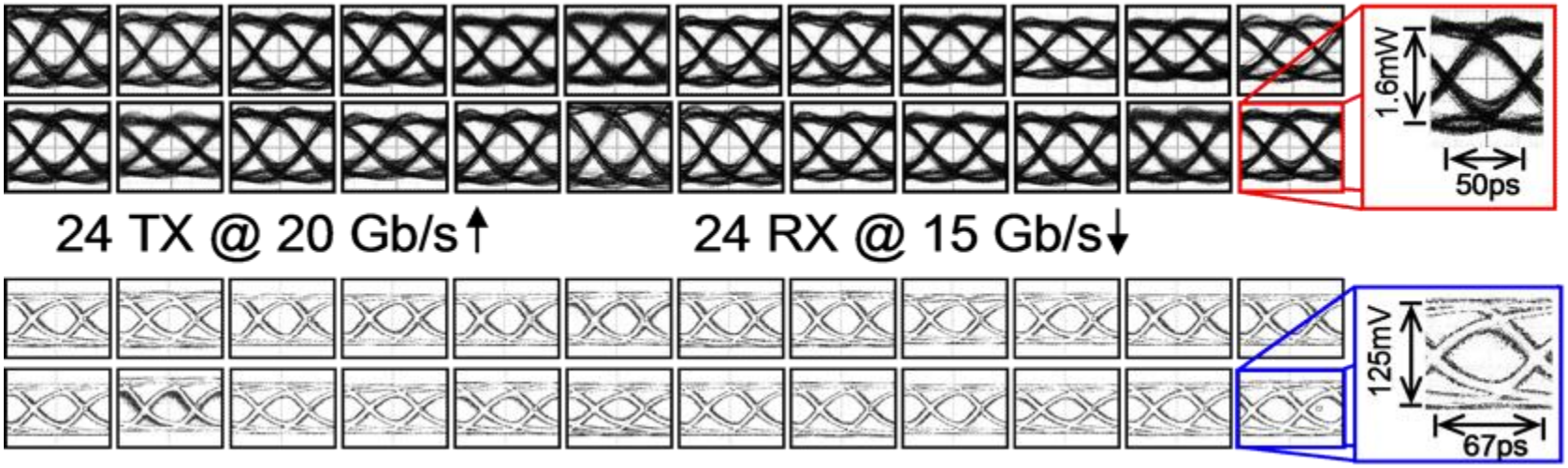


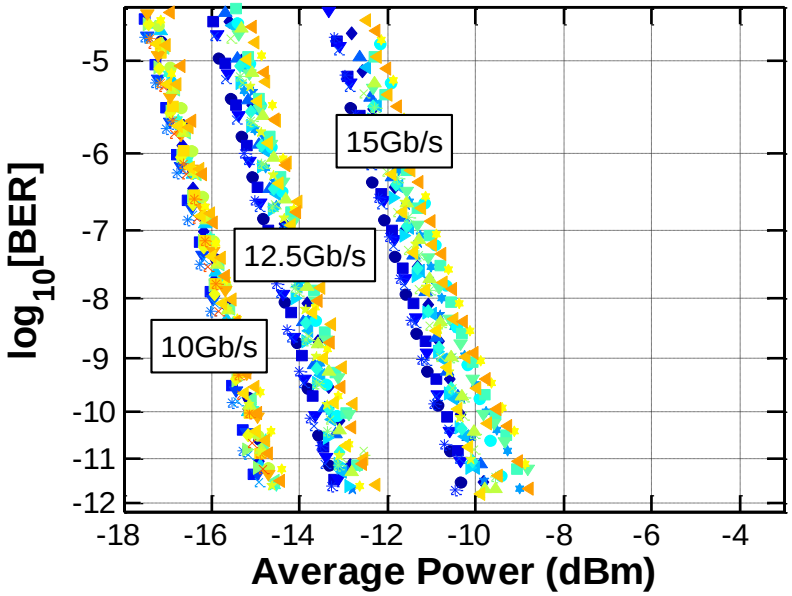


Lens array attached to Optochip through milled cavity

- Optochip soldered onto high-speed organic carrier (EIT CoreEZ™)
- 24 TX + 24 RX high-speed I/O routed to probe sites on the surface

360 Gb/s Bidirectional Optomodules: 24 x 15 Gb/s/ch



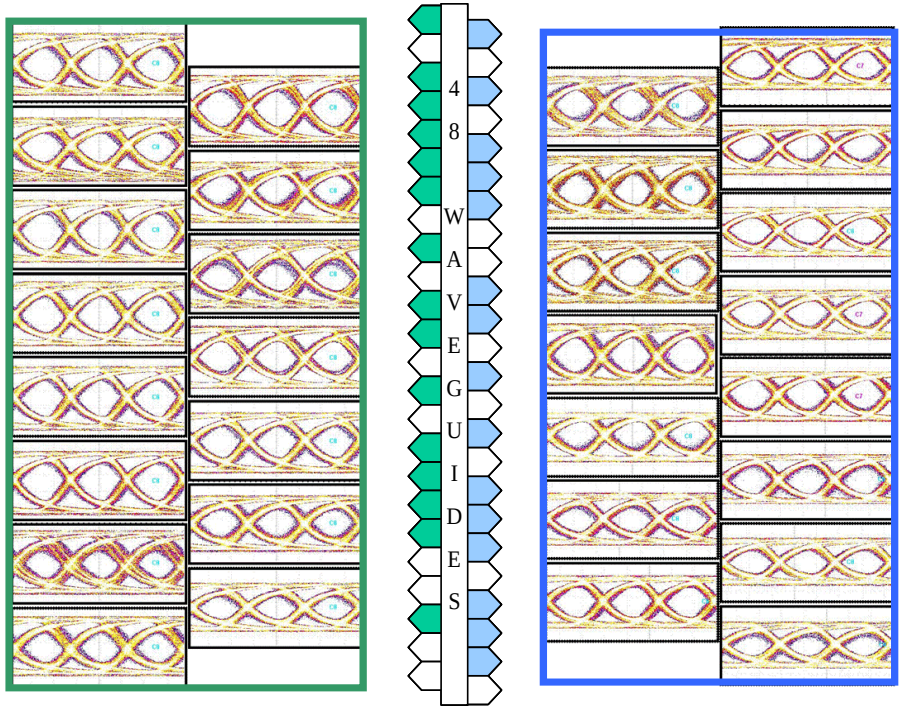


- TX operates up to 20 Gb/s, RX to 15 Gb/s
- Tested with fiber probe
- 360 Gb/s bi-directional total
  - 24 + 24 @ 15 Gb/s
- Uniform performance – RX sensitivity

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.

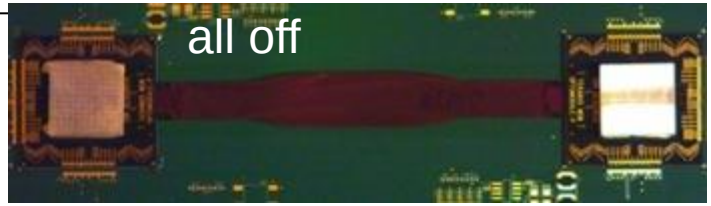
Optical PCB in Operation

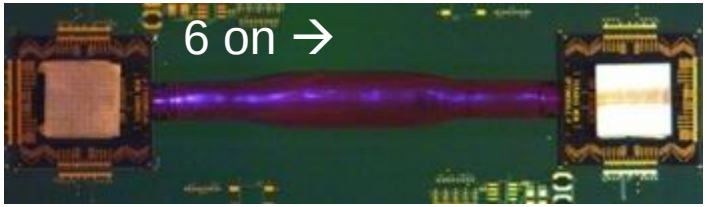
15 Gb/s

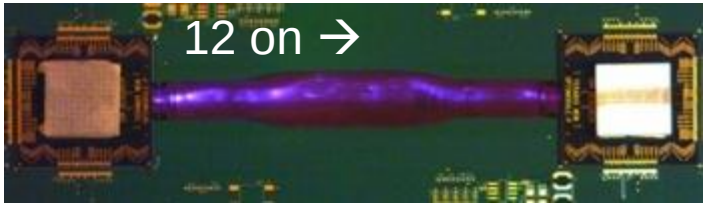


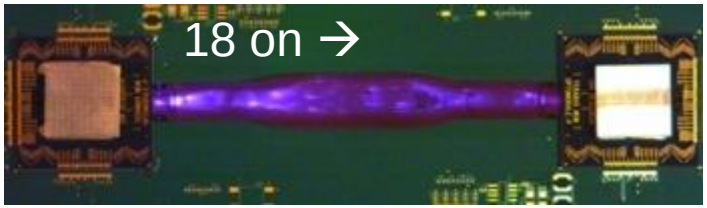
15 + 15 channels

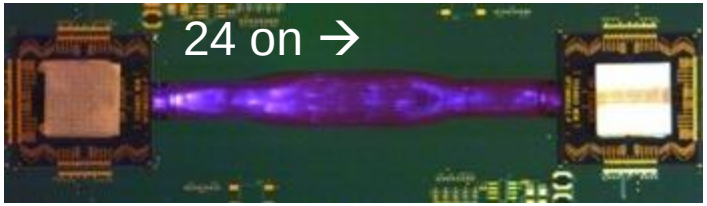
- 15 channels each direction at 15 Gb/s, BER < 10^{-12}
- 225 Gb/s bi-directional aggregate
- 145 mW/link = 9.7 pJ/bit









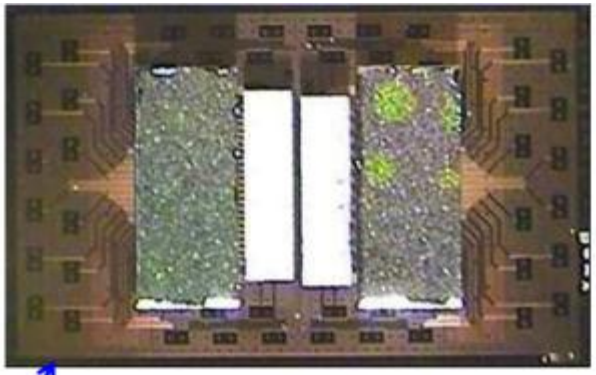


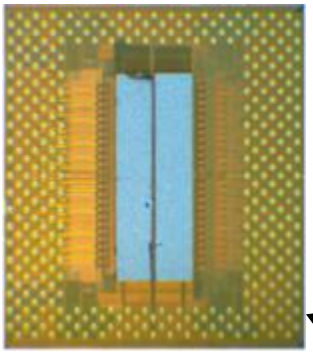
F. E. Doany *et al.*, "Terabit/s-Class Optical PCB Links Incorporating 360-Gb/s Bidirectional 850 nm Parallel Optical Transceivers," *IEEE JLT*, Feb. 2012.

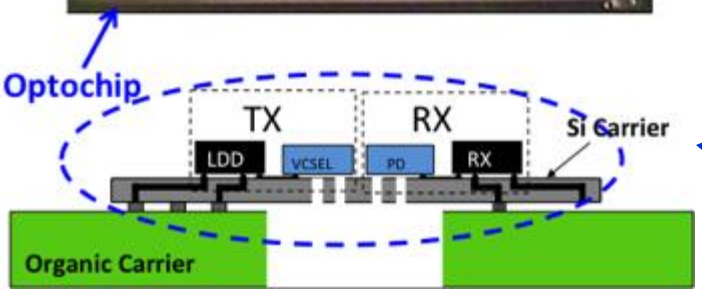
Holey Optochip → **Highly Integrated** 850-nm Transceiver

Single-chip CMOS IC
Integrated optical vias
Flip-chip attached OE arrays

Si carrier-based Optochip

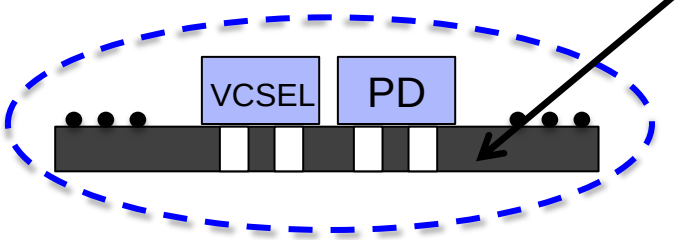






Holey Optochip

CMOS IC



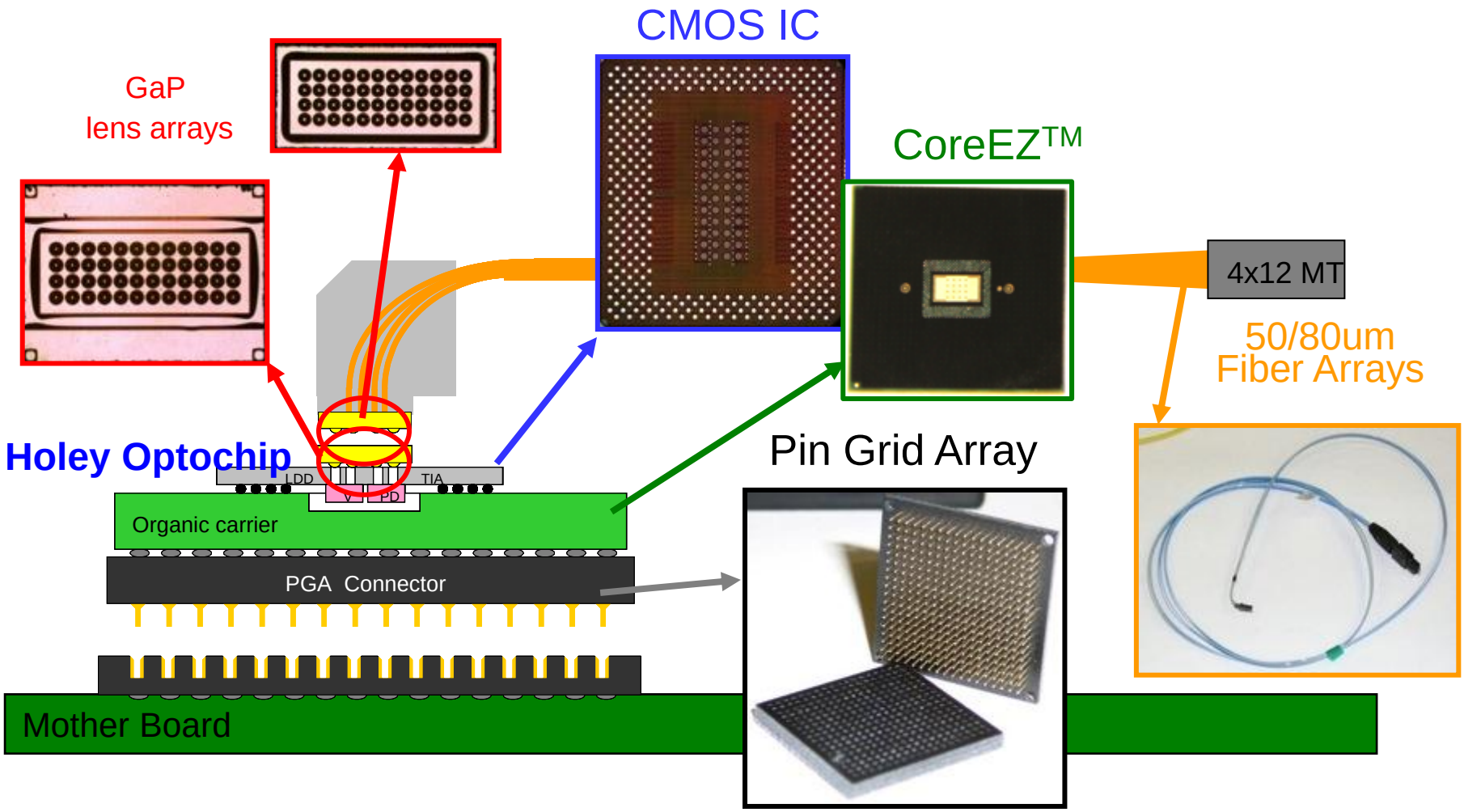
Suitable for fiber or waveguide coupling

Holey Optochip enables dense integration with simplified packaging

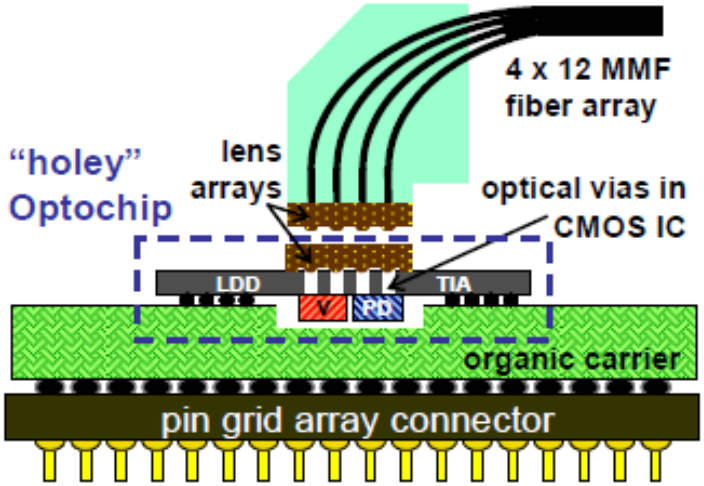
Holey Optochip Transceiver Module

24+24 channel 850-nm optical transceiver based on “holey” CMOS IC

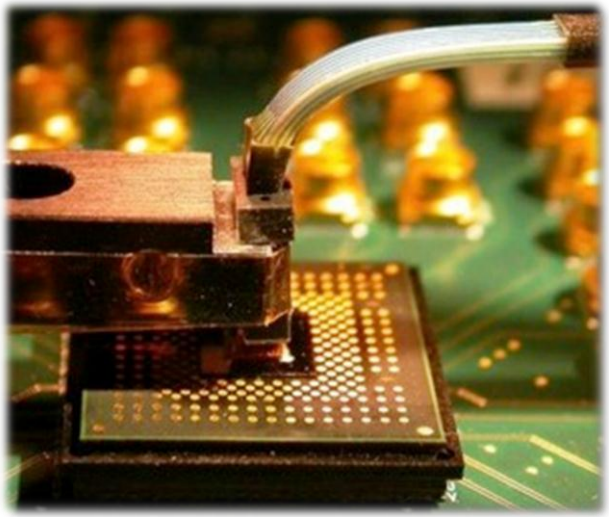
Fiber-coupled version

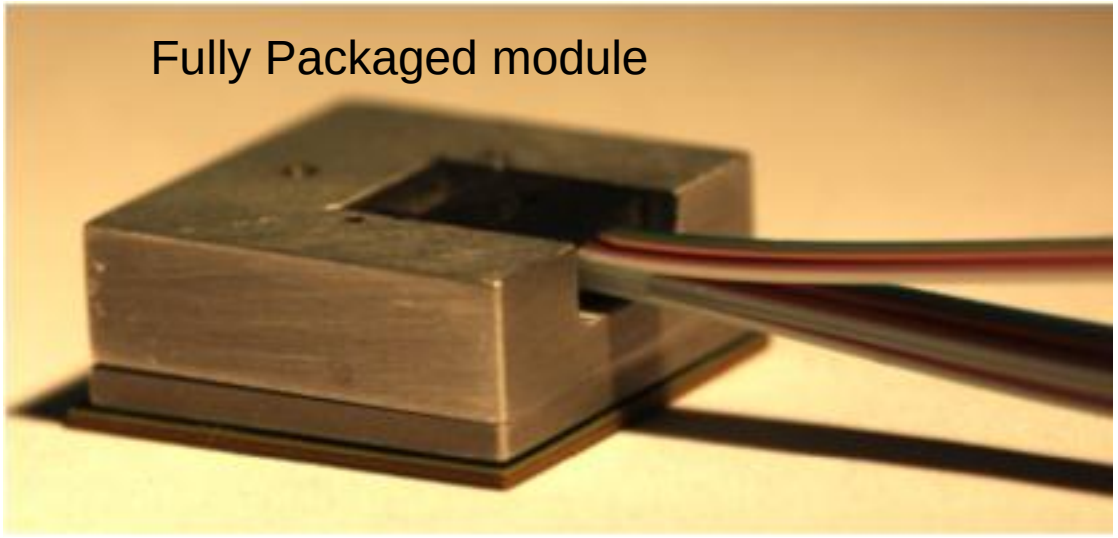


“Holey” Transceiver Module: Tb/s Chip-Scale Transceivers



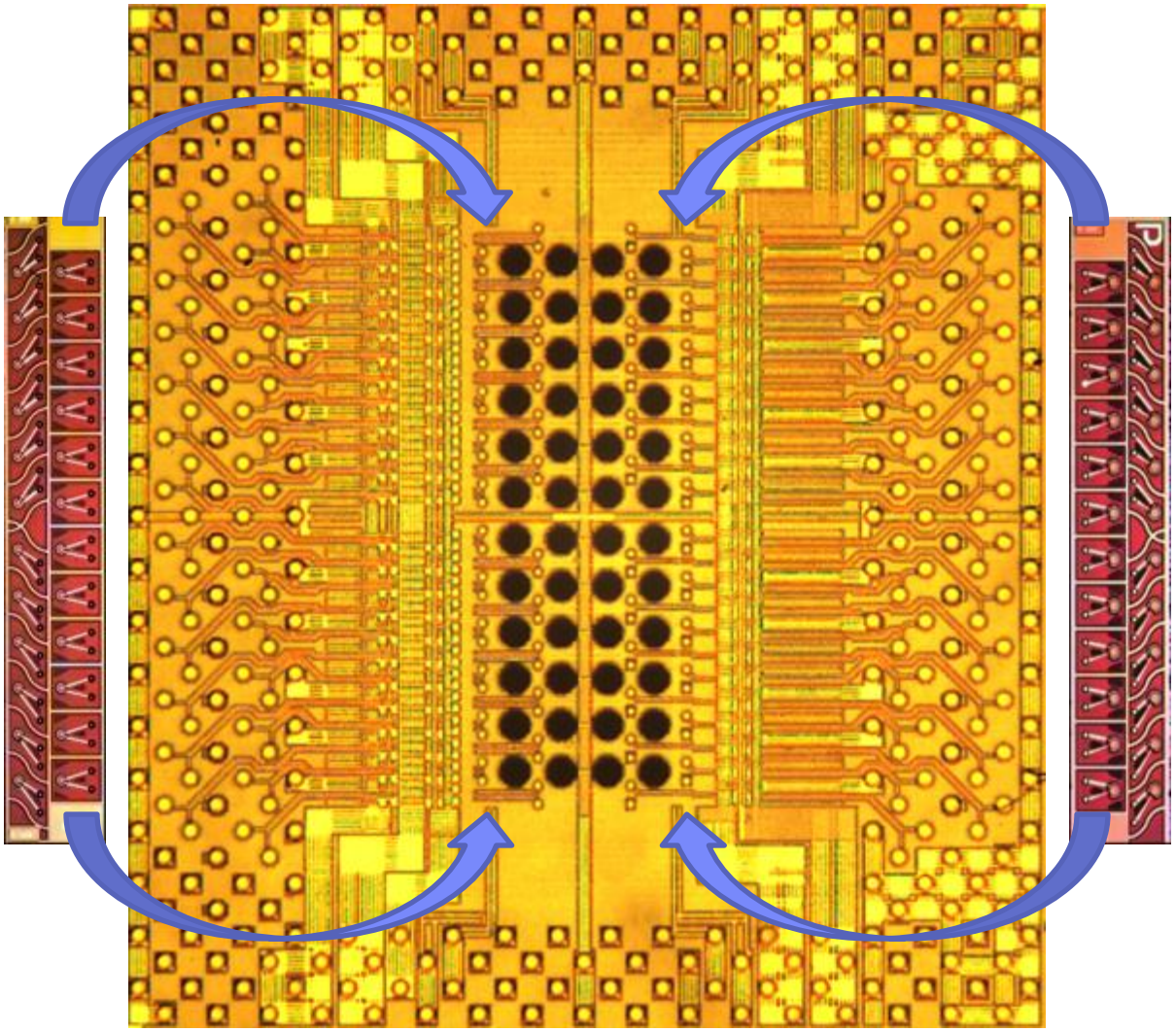
- **Tb/s target** → 24 TX + 24 RX @ 20 Gb/s = 0.96 Tb/s
- Circuit design focus on power efficiency, targeting 5 pJ/bit
- Single “holey” CMOS IC -- bulk CMOS process + wafer-level post-processing for optical vias
- Dual-lens system → relaxed tolerances & efficient coupling

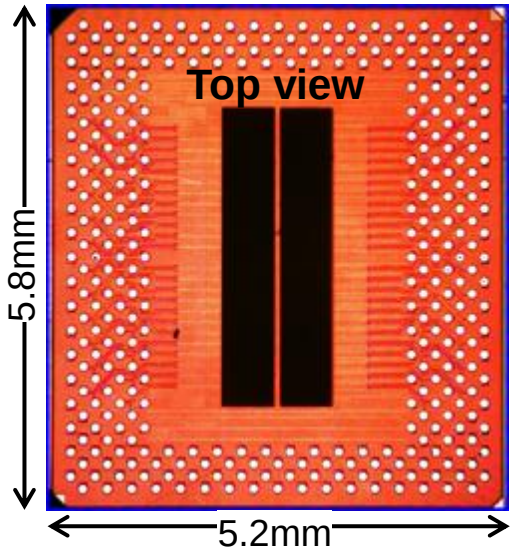


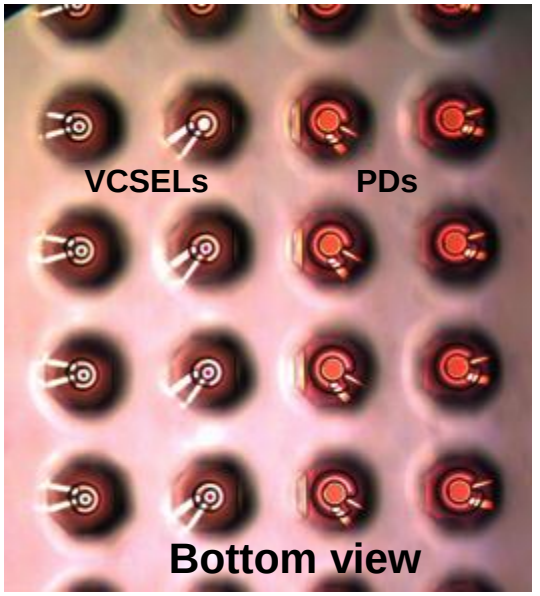


- F. E. Doany *et al.*, "Dense 24 Tx + 24 Rx Fiber-Coupled Optical Module Based on a Holey CMOS Transceiver IC," *ECTC 2010*, pp. 247–255.
- C.L. Schow *et al.*, "A 24-Channel 300Gb/s 8.2pJ/bit Full-Duplex Fiber-Coupled Optical Transceiver Module Based on a Single “Holey” CMOS IC," IEEE *JLT*, Feb 2011.

Holey Optochips: Direct OE to IC packaging at 850 nm





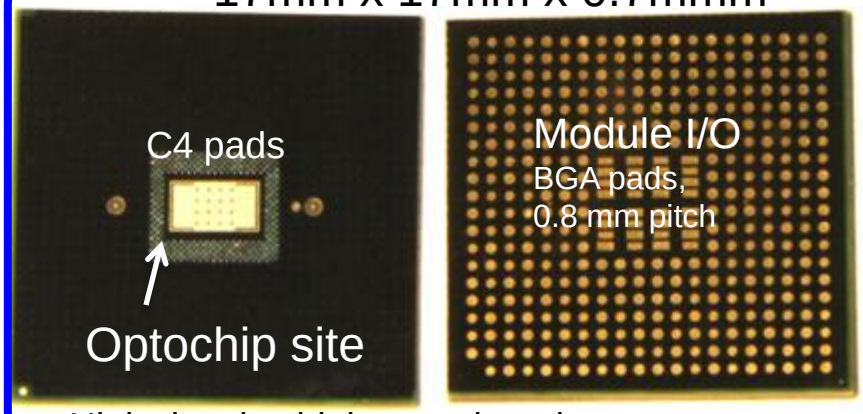


- Single 90-nm CMOS IC
  - Wafer-scale process for optical vias and Ni/Au pad plating
- OE arrays (Emcore) flip-chipped directly onto CMOS

N. Li *et al.*, "High-Performance 850 nm VCSEL and Photodetector Arrays for 25 Gb/s Parallel Optical Interconnects," *OFC 2010*, paper OTuP2.

Optochip Packaging: Pluggable Module

17mm x 17mm x 0.7mmm

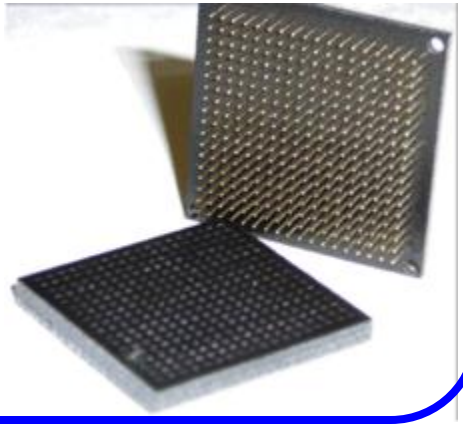


High-density, high-speed carrier (EIT CoreEZ™)

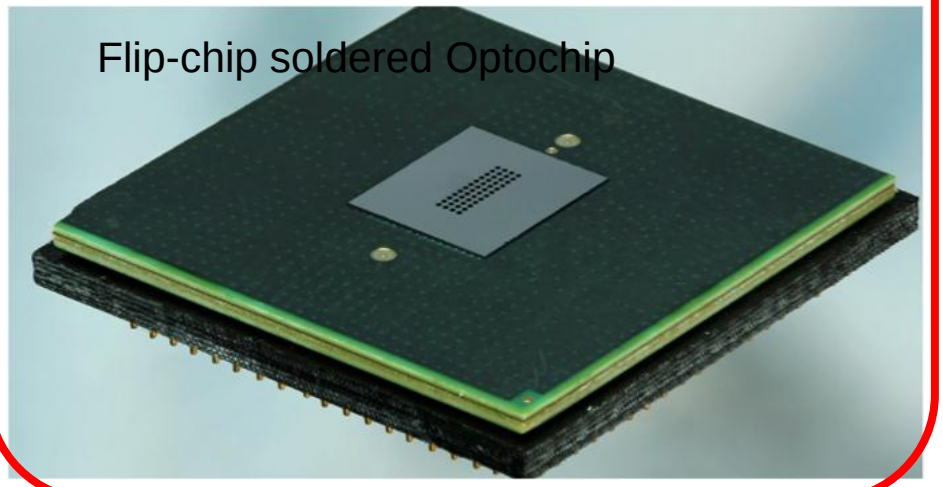
Low-profile, high-speed connector:

- ISI HiLo, 0.8 mm pitch

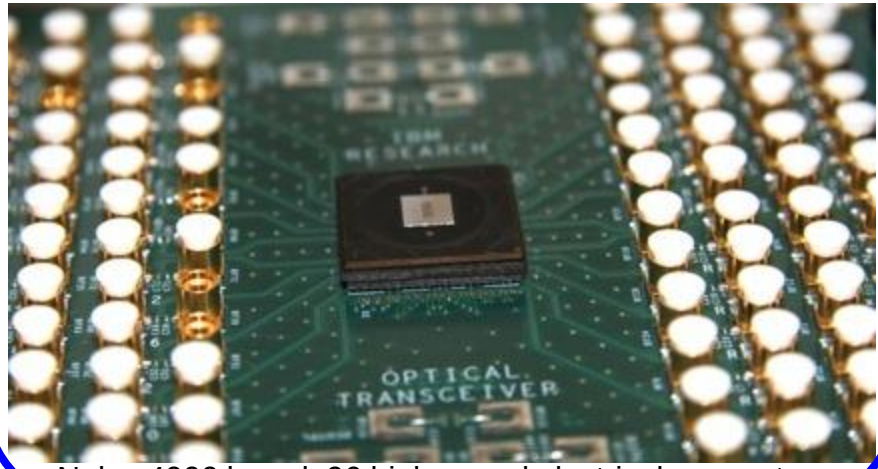
Optomodules can be swapped into and out of a socket on a motherboard



Complete Optomodule: Optochip-Organic carrier-PGA Connector



Transceiver Optomodule plugged into test board

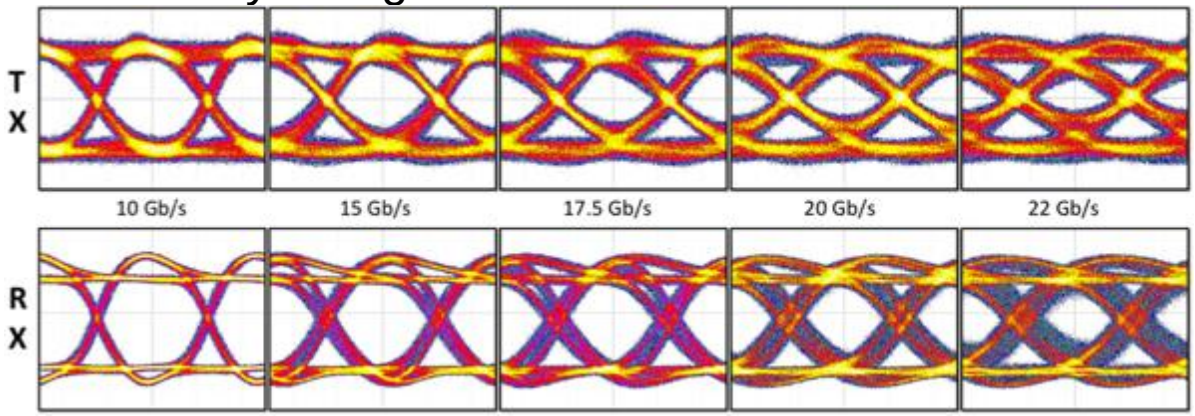


Nelco 4000 board; 96 high-speed electrical connectors

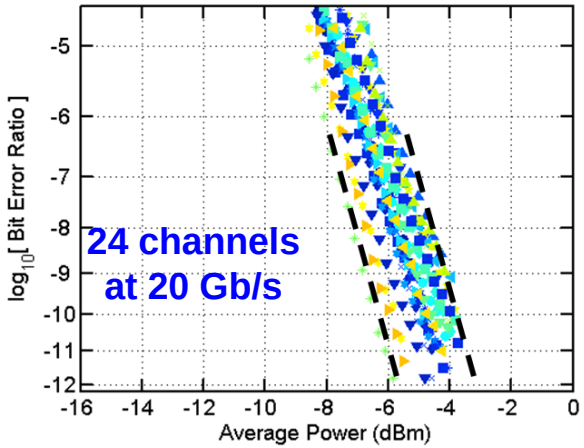
Holey Optomodule: First Terabit/sec Multimode Optical Module

IBM

Eye Diagrams at Various Data Rates



T X 20 Gb/s



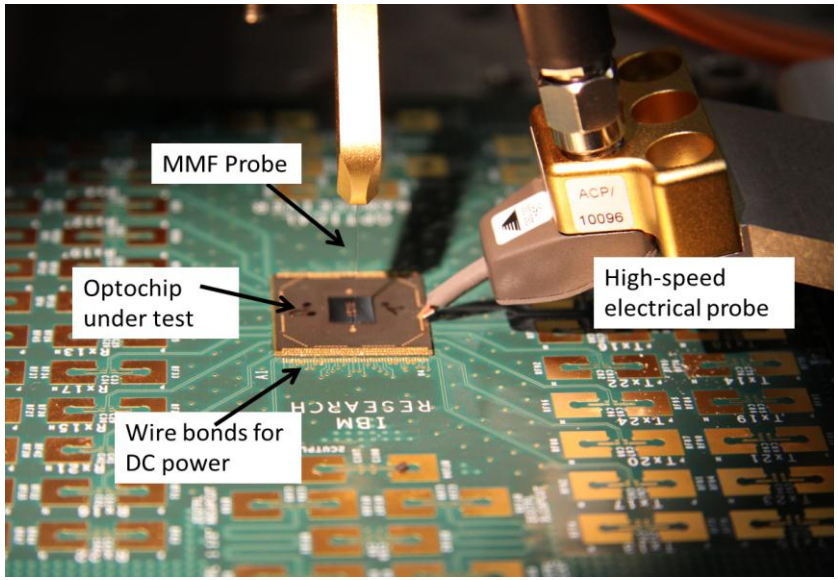
Error-free (BER < 10^{-12})

R X 20 Gb/s

480 + 480 Gb/s (24 + 24 @ 20 Gb/s)

- **7.3 pJ/bit** (79 mW RX and 67 mW TX)

Probe-able Holey Optomodule: 20 Gb/s @ 4.9 pJ/bit Link Efficiency



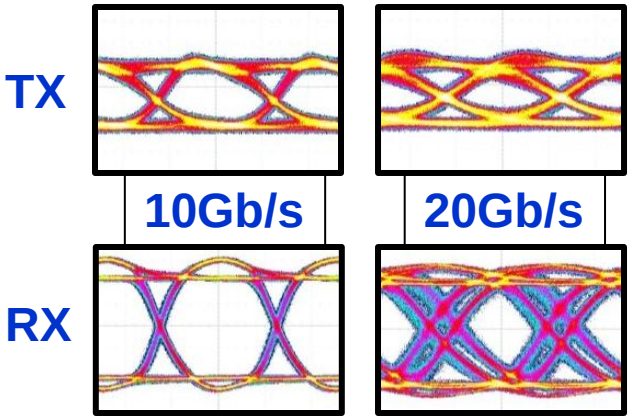
Low-Power optimization

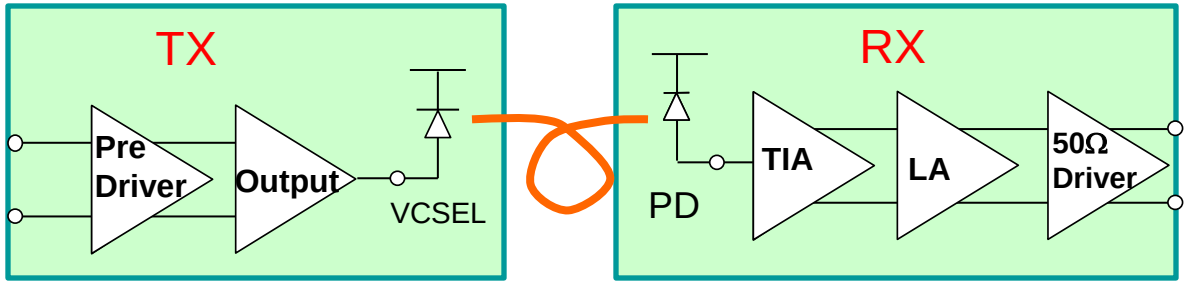
- Probe-able version of chip carrier
  - Intrinsic Optochip performance
- BER < 10^{-12} for 18 RX links
- Wall-plug power counting all contributions

Link Component	Supply Voltage	Power Dissipation (mW)
LDD	1.8	26
VCSEL	2.6	10
TIA	1.7	14
RX_LA	1.8	39
RX_IO	1.0	9
Total		98

TX

RX





1 Tb/s Data Transfer Comparison

- Holey Optochip is complete transceiver providing Tb/s data transfer in ~ 30mm^2
  - Potential for direct flip-chip packaging to MCM
  - Current packaged implementation limited by BGA pitch of PCB
- Best commercial modules: requires 8 modules with ~600mm^2 footprint

1 Tb/s

~ 30mm^2

Holey Optochip

480 + 480 Gb/s

(24 + 24 @ 20 Gb/s)

- **31.8 Gb/s/mm^2**

Path to Optimizing Link Power Efficiency

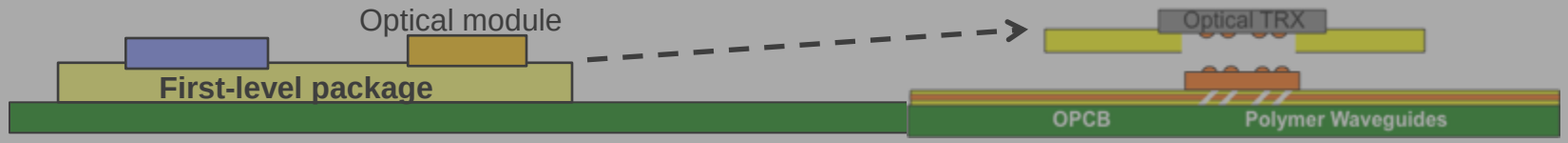
Packaging Integration:

Optics co-packaging

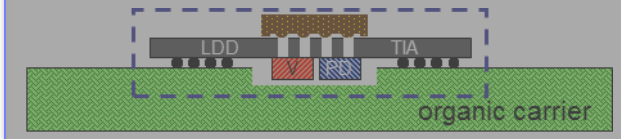
- Minimize power in electrical link from logic to optics
  - drive across chip carrier instead of board
- High BW density electrical/optical interfaces

Optical-PCBs

- **PCBs with integrated polymer waveguides**
- **High BW density optical interfaces**
- Efficient optical coupling systems with relaxed tolerances

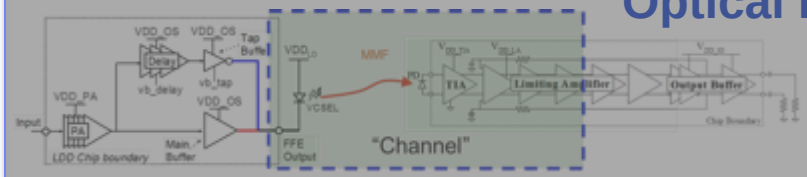


Chip-Scale Integration



- Optochips: chip-like optical transceivers
- Flip-chip packaging enabling dense 2-D arrays
- Direct OE to IC attachment for maximum performance

Optical Link Improvements



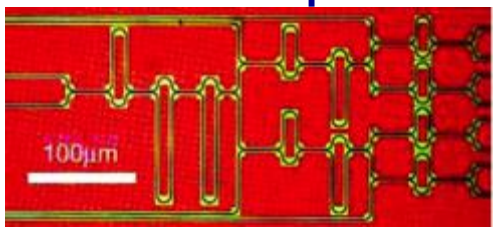
- Advanced CMOS for high-speed and low power
- Faster, more efficient VCSELs and PDs
- Equalization to improve link performance and margin

New Technologies: Si Photonics

Light modulation:

- Mach Zehnder interferometers
- Ring Resonators

External laser input

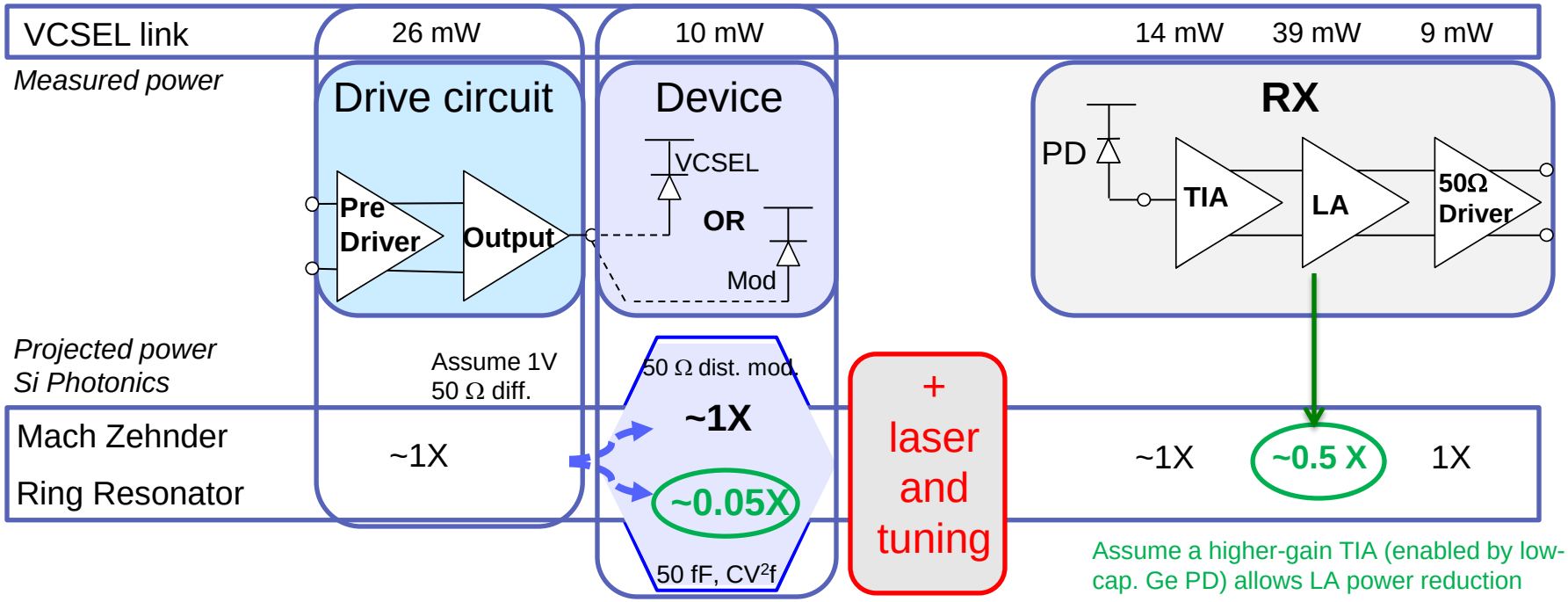


Potential for low power, high bandwidth transceivers

- Integrated Si nano-photonics
  - High density but µm-alignment challenges
  - Temperature stabilization for resonant devices
- Longer reach through SMF
- WDM → high BW density
- Low-power devices, but must consider full link power:
  - modulator + drive circuits + laser

Power Efficiency for Analog Links: VCSELs Versus Si Photonics

Example: Basic Analog Link, 20 Gb/s, 90-nm CMOS



- **Compared to VCSEL links (not including laser and tuning power):**
  - MZ modulators comparable
  - RR potentially ~30% lower (without laser)
    - Require precise temperature stabilization
- **Primary advantage for Si photonics is WDM capability and density potential**
  - **MUST** be implemented cost effectively and with low optical loss
- Sub-pJ/bit Si photonic TX and RX demonstrated at 10Gb/s
  - Using digital clocked circuits, typically limited to lower speeds

Laser not included

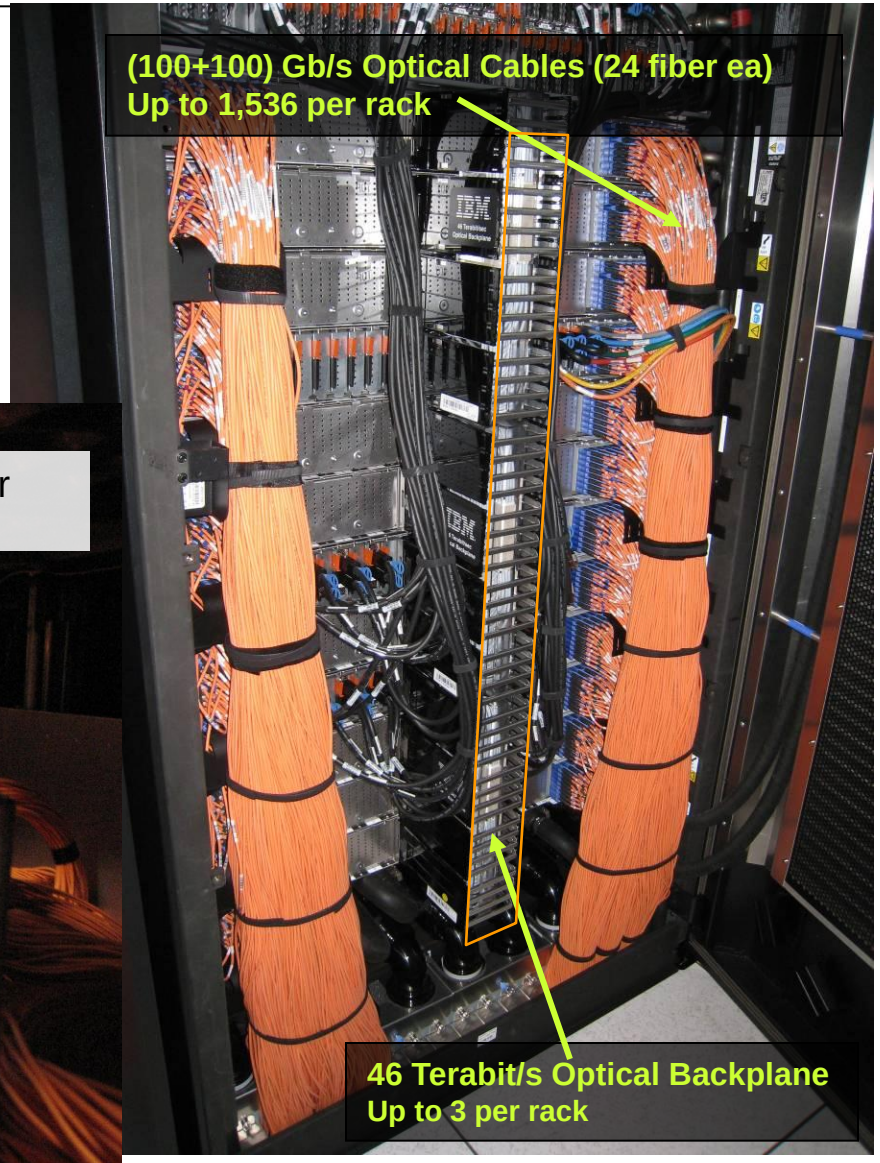
*X. Zheng, "Ultralow Power 80 Gb/s Arrayed CMOS Silicon Photonic Transceivers for WDM Optical Links," JLT, Feb.15, 2012

New Technologies: More BW per Fiber

- Si Photonics with WDM
  - – Can alleviate fiber management issues

Potential VCSEL-Based Transceiver Technologies:

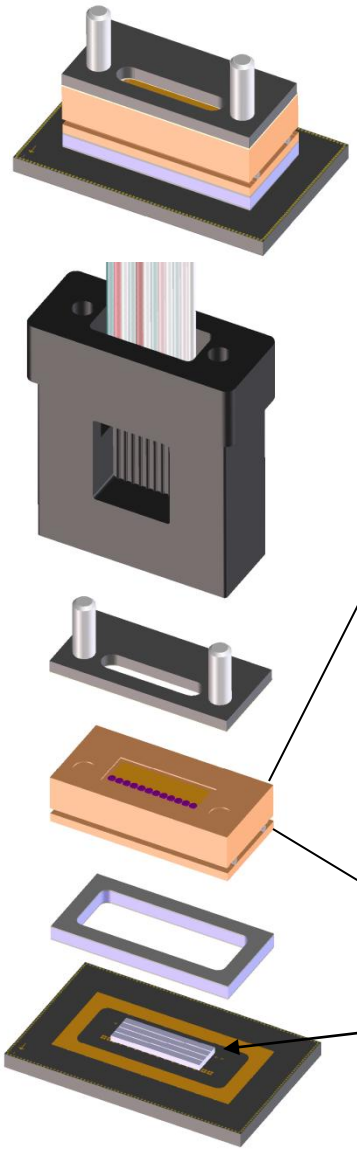
- Coarse WDM (CWDM)
- Multicore Fiber



Power 775 System

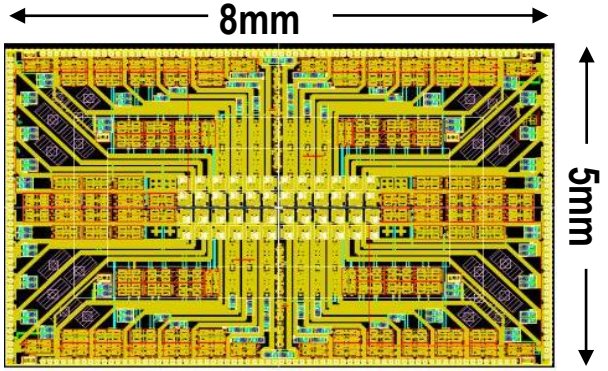
MAUI: 4λ CWDM, 48ch, 12 Fibers, 0.5Tb/s, ~6pJ/bit

IBM



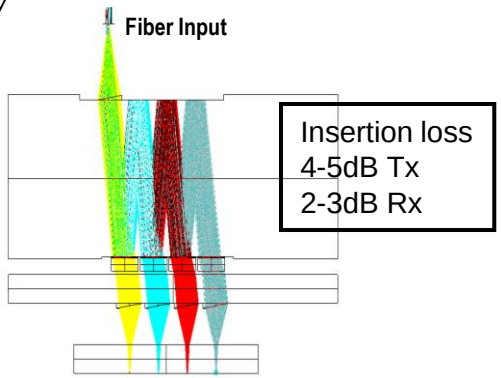
Prototype Demonstration

Assembled Tx

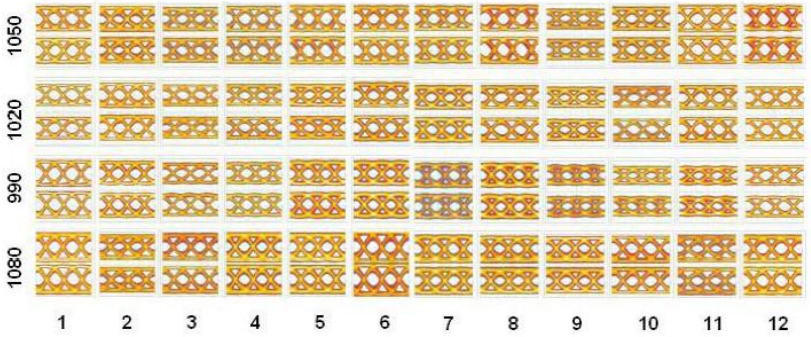


75GHz SiGe Technology
3.3W Total Tx+Rx @ 500Gb/s = 6.6pJ/bit

MicroOptical Mux/Demux



Insertion loss
4-5dB Tx
2-3dB Rx



48ch, 10Gb/s/ch

Bottom-emitting VCSELs

4 separate VCSEL Arrays flip-chip mounted on IC
CWDM @ 30nm spacing 990, 1020, 1050, and 1080nm

Using today's 25Gb/s VCSELs, this technology could realize 1.2Tb/s over 12 fibers

Agilent Labs:

G. Panotopoulos, Workshop on Interconnections Within High-Speed Digital Systems Santa Fe, May 2004

B. Lemoff et. al. IEEE LEOS 2005

An Alternative to CWDM: Multicore Fiber

MCF = Multiple Cores in a single fiber strand

- 7 lasers coupled to MCF → packaging challenge
- 7 wavelengths in a single fiber → Manufacturing, Mux/Demux challenge

Refractive Index Δn 2D profile

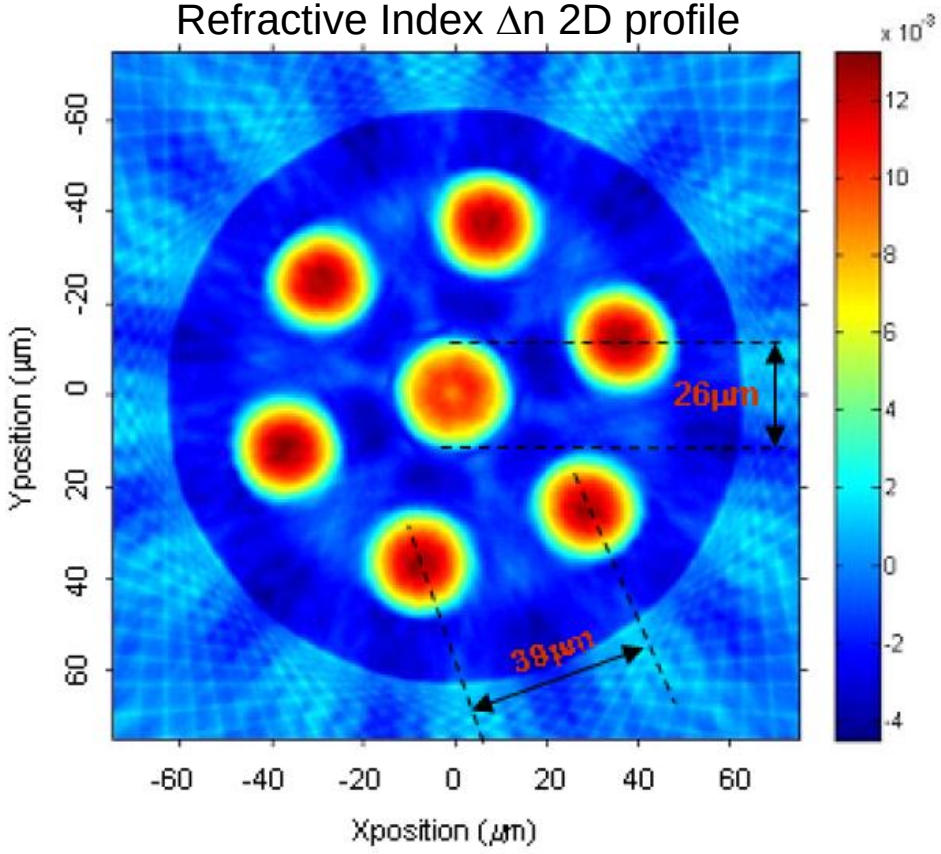
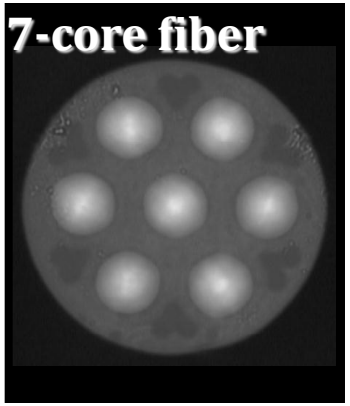


Table 2. Fiber properties

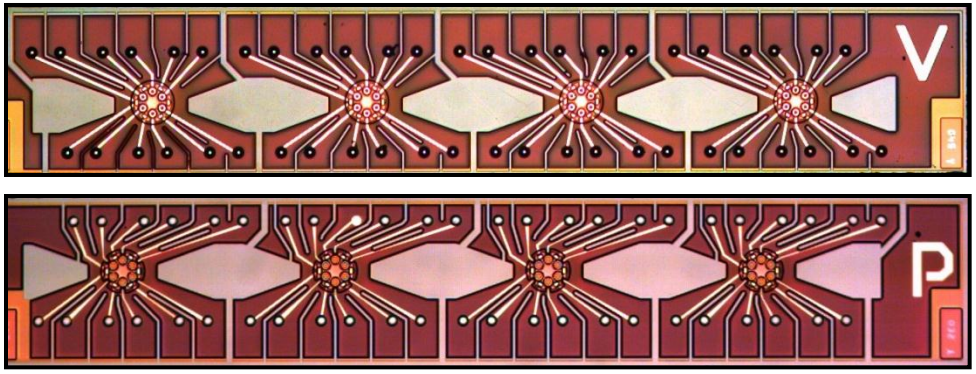
Fiber parameters	Typical values
Core diameter	26 μm
Core-to-core pitch	39 μm
Clad diameter	125 μm
Coating diameter	250μm
Attenuation at 850nm	~2.2 dB/km
Attenuation at 1310nm	~0.5 dB/km
Crosstalk *	<-40 dB
DMD **	~0.12 ps/m



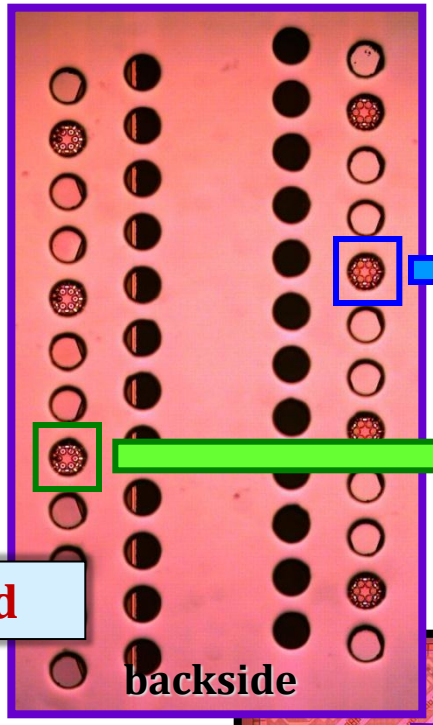
Smaller cores have higher BW

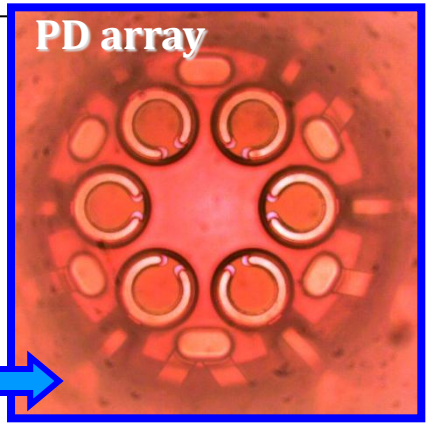
4-Fiber 24-Core Optical Transceiver

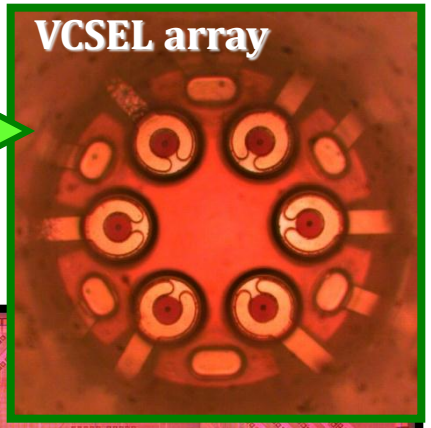
Custom VCSEL/PD Arrays Matched to 4 Multicore Fibers



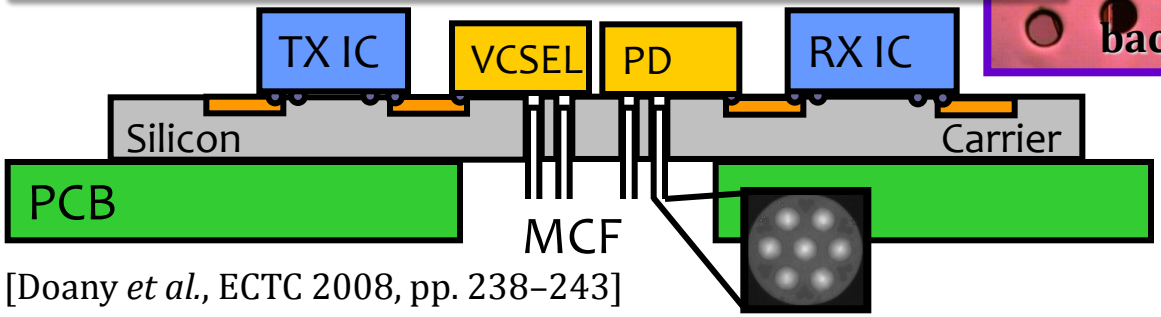
Fabricated by Emcore Corporation



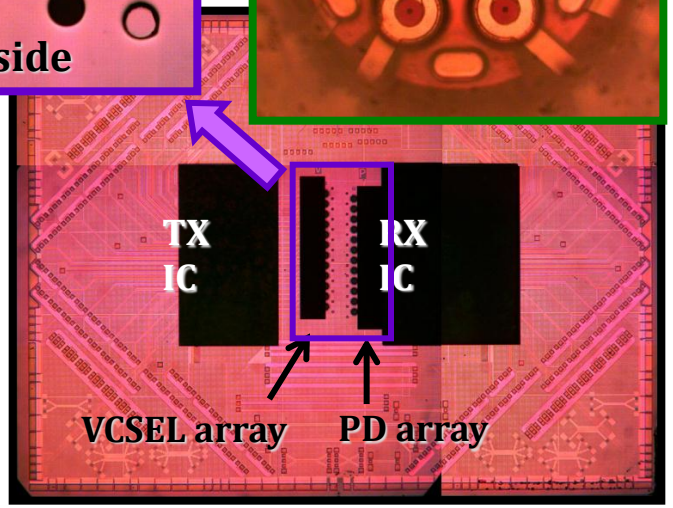




120 Gb/s over 100-m using one MMF strand



[Doany *et al.*, ECTC 2008, pp. 238–243]



- Custom OE chips designed to fit into existing configuration of Terabus project—Match silicon carrier designed for 24-channel polymer optical waveguide transmitter.

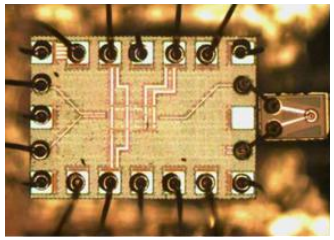
Outline

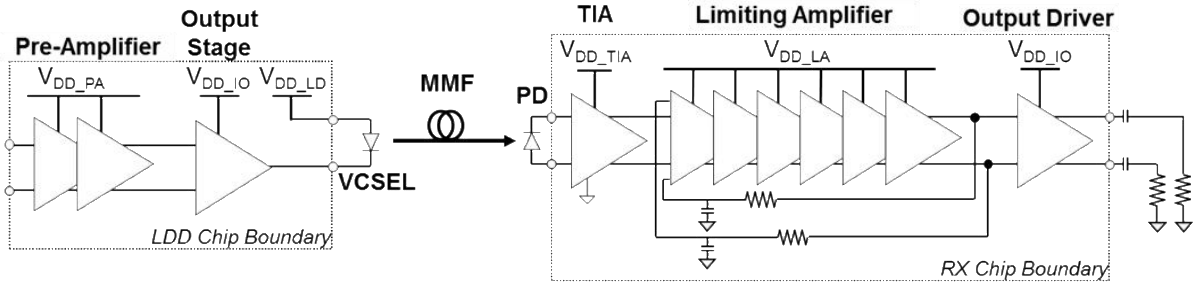
- Intro to Fiber Optics Links
- Fiber Optics in HPC
  - Evolution of optical interconnects in HPC systems
  - System needs and power, cost and density challenges
- Path to Optimizing power and efficiency
  - Packaging Integration
  - Optical-PCB Technology
  - Chip-scale Integration: Generations of Parallel VCSEL Transceivers
  - Optical Link Improvements
  - New Technologies: Si Photonics, Multicore Fiber, CWDM
- **Pushing the Limits of Speed and Power**
  - **Equalization for improved speed and margin**
  - **Fast SiGe circuits to probe VCSEL speed limits**

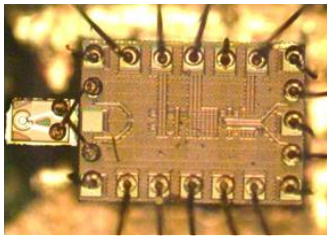
Single-Channel Transceiver Studies to Determine Technology Limits and Future Directions

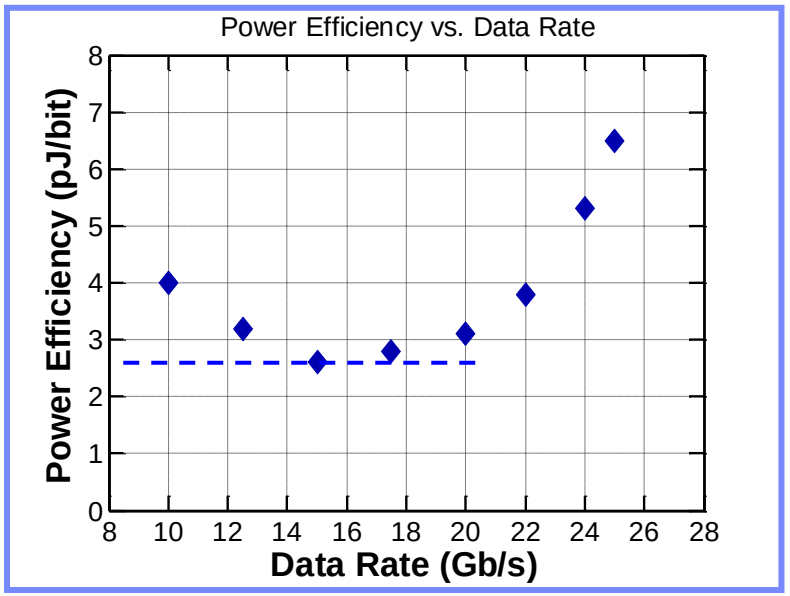
- Concluding Comments

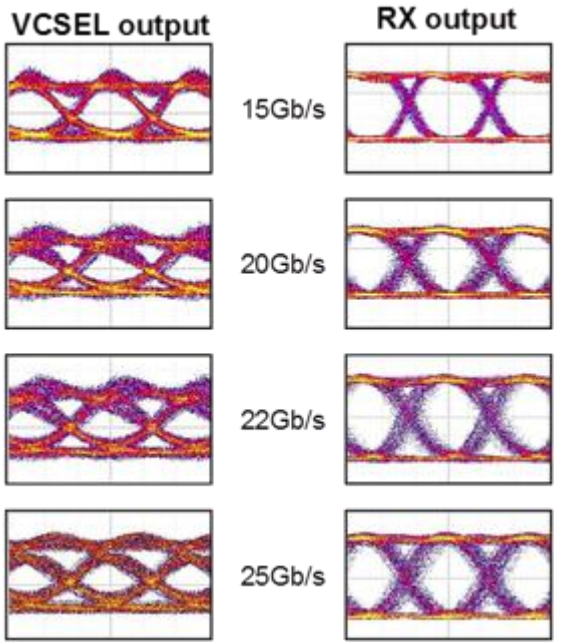
Un-equalized CMOS Links Achieve 25 Gb/s, Record Efficiency









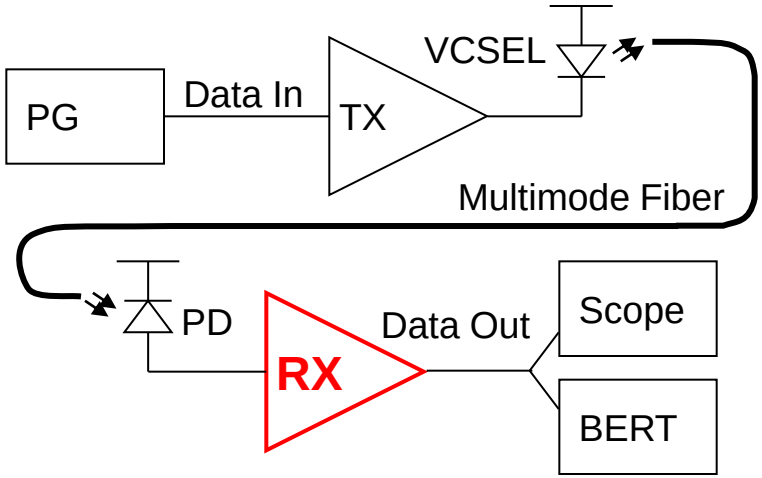


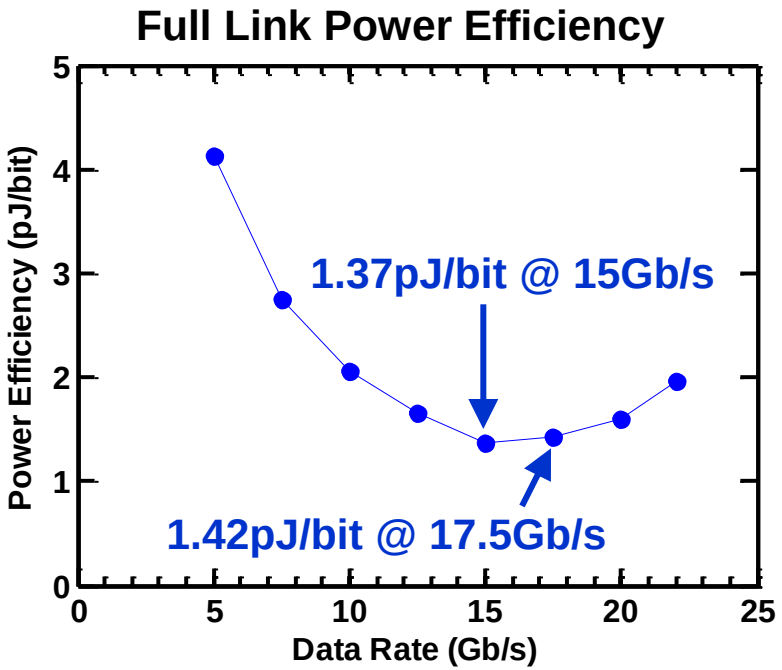
- Links operate up to 25 Gb/s: a first for CMOS
- Record power efficiencies: 2.6pJ/bit @ 15 Gb/s, 3.1 pJ/bit @ 20 Gb/s
- Transmitter equalization will likely yield further improvement

C. L. Schow *et al.*, "A 25 Gb/s, 6.5 pJ/bit, 90-nm CMOS-Driven Multimode Optical Link," *IEEE PTL*, 2012, in press.

90-nm CMOS Inverter-Based RX without TX Equalization

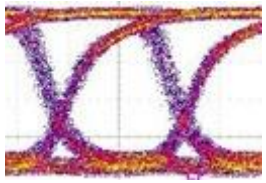
- Record low power for an optical link in any technology
- Power consumption is on the order of exascale requirements

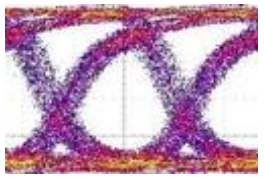




IBM Optical Link Efficiency

March 2011	4.6pJ/bit @ 15Gb/s
March 2012	1.37pJ/bit @ 15Gb/s

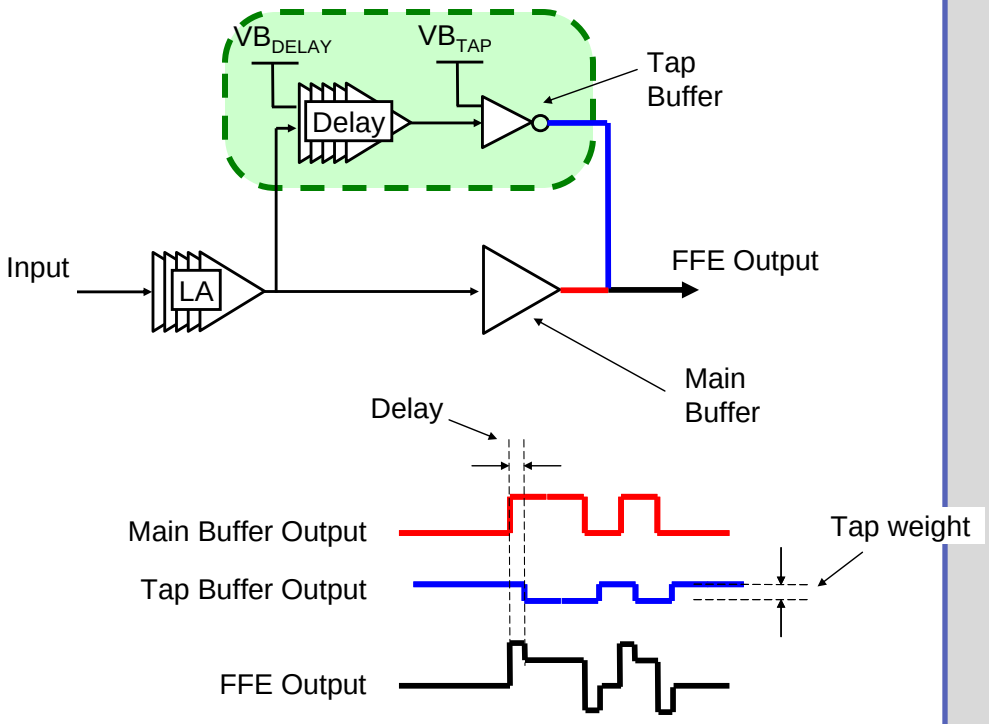
15Gb/s 

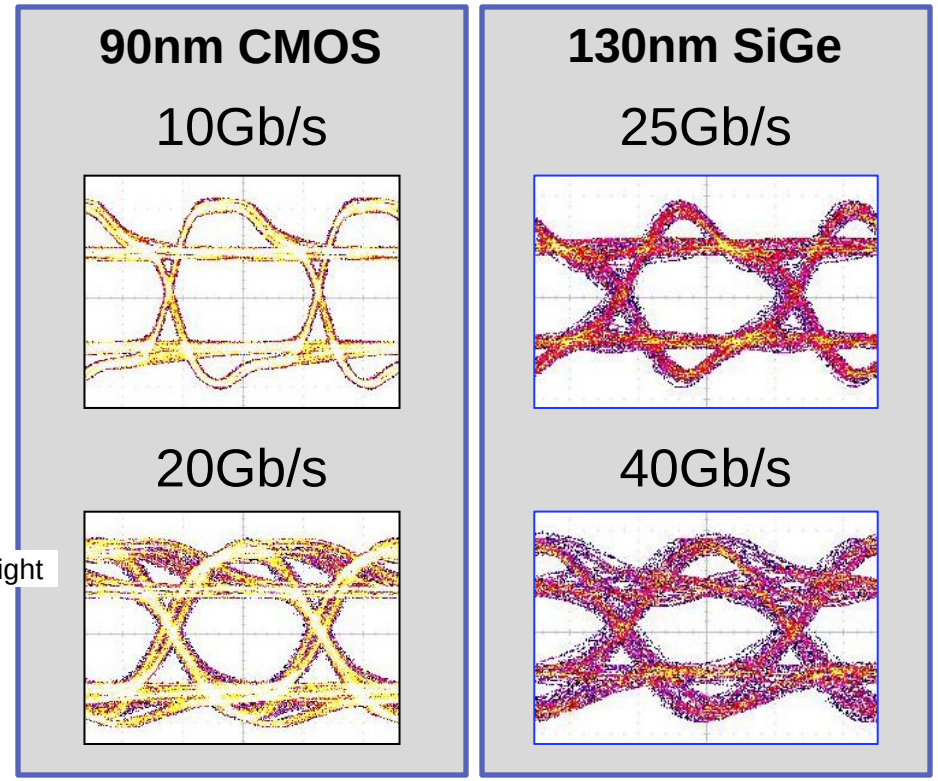
20Gb/s 

J. Proesel, C. Schow, A. Rylyakov, "Ultra Low Power 10- to 25-Gb/s CMOS-Driven VCSEL Links," *OFC 2012*, paper OW4I.3.

Transmitter and Receiver Equalization

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





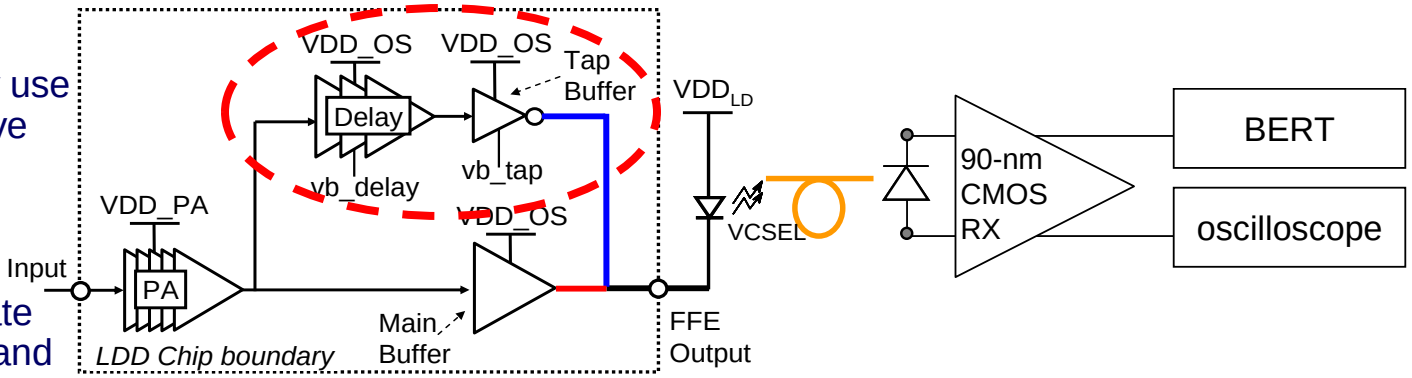
- **Feed-Forward Equalizer (FFE) leveraging extensive electrical serial link design**
- **Equalization heavily applied to VCSEL outputs for end-to-end link Optimization**

Applying Signal Processing to Low Power Optical Links

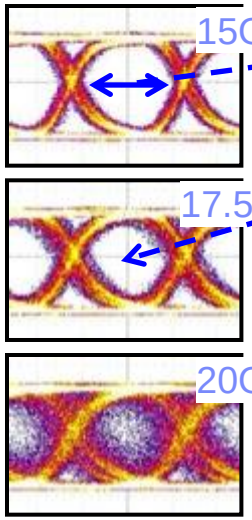
Electrical links increasingly use signal processing to improve performance…

– optics can do this too

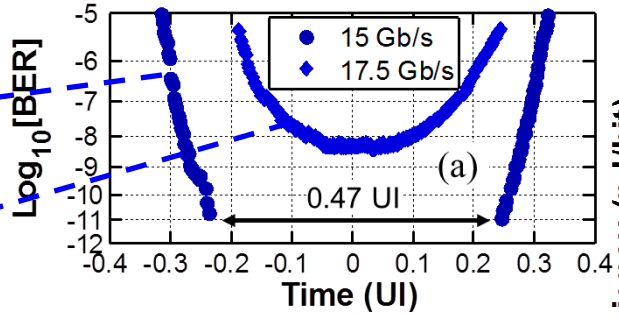
Pre-distortion to compensate for combined VCSEL, TIA and LA bandwidth limitations

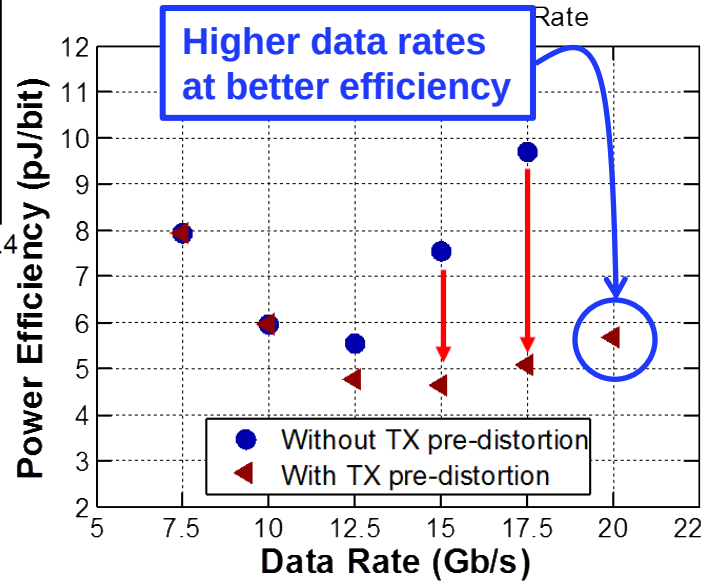


No FFE



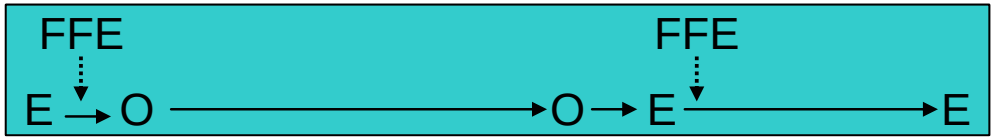
Timing Margin

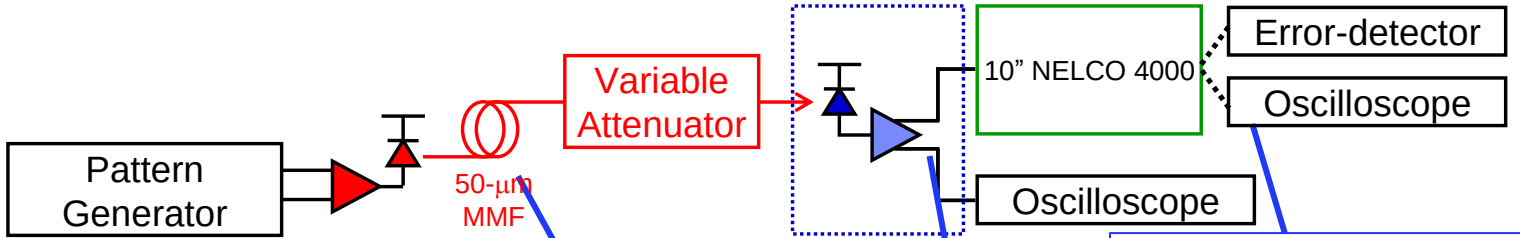




- A. V. Rylyakov *et al.*, "Transmitter Pre-Distortion for Simultaneous Improvements in Bit-Rate, Sensitivity, Jitter, and Power Efficiency in 20 Gb/s CMOS-driven VCSEL Links," *J. of Lightwave Technol.*, 2012.

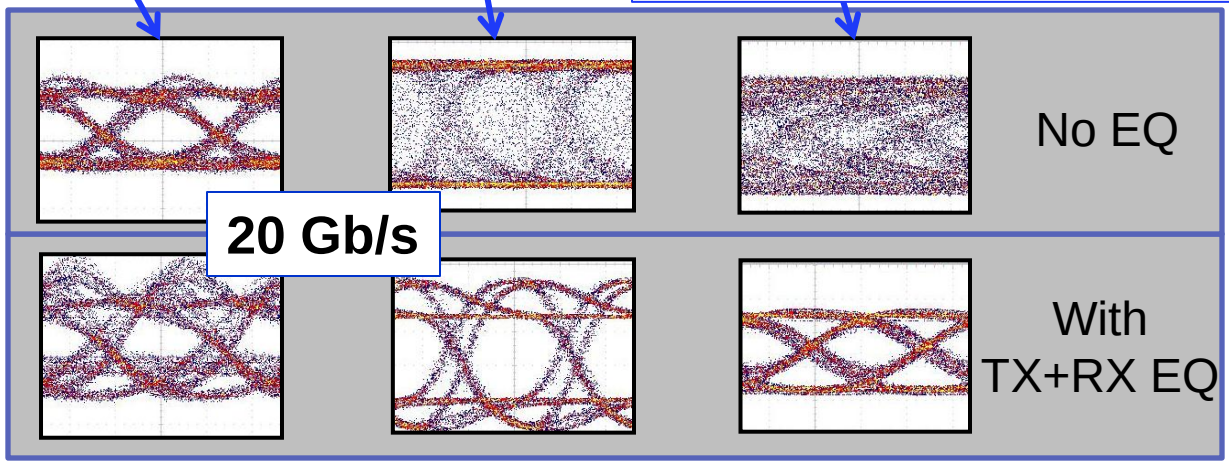
TX & RX Equalization for End-to-End Link Optimization





Equalize not to improve TX output or RX output

Equalize to improve signal quality at data destination



	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**

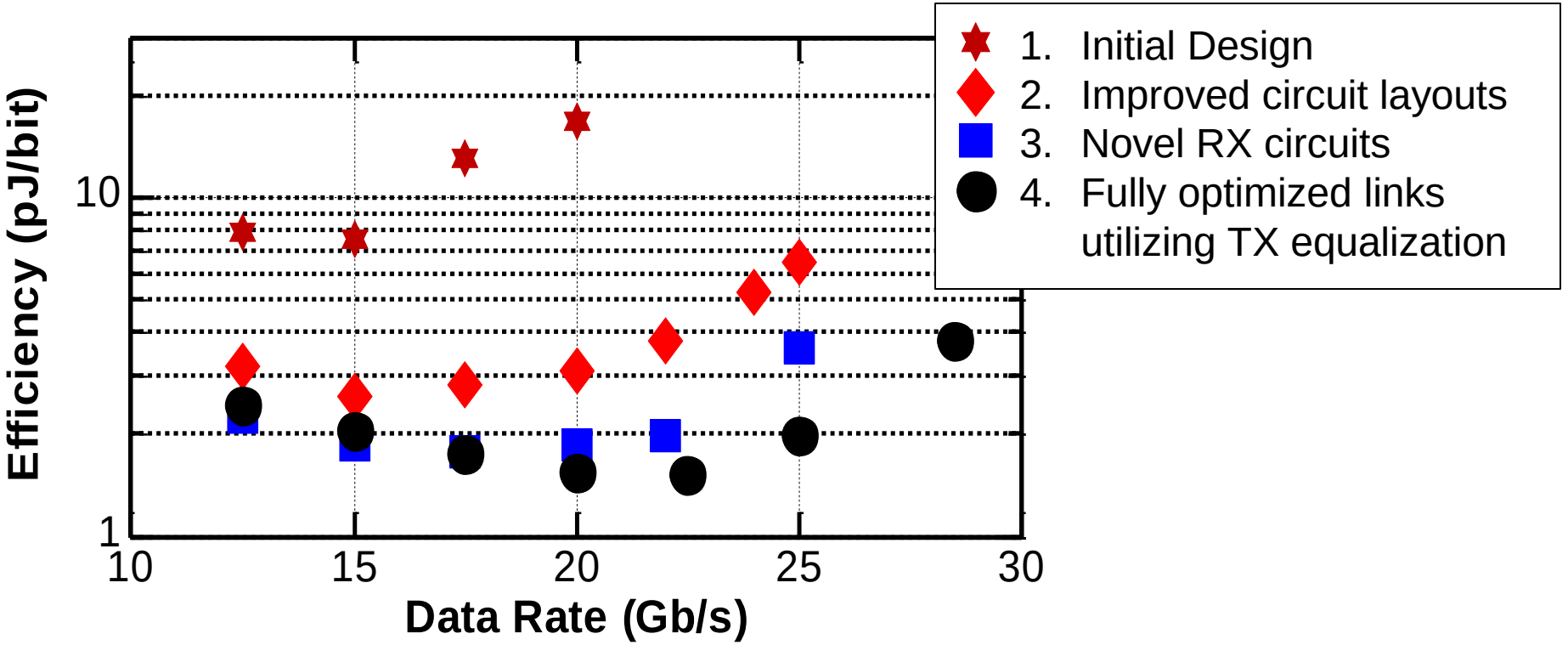
<5%

- **Heavy equalization to optimize the full link, NOT to beautify the TX output**
- **Minor contributor to total power → significant benefit to power efficiency**

- 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.

Summary of Power Efficiency Trends

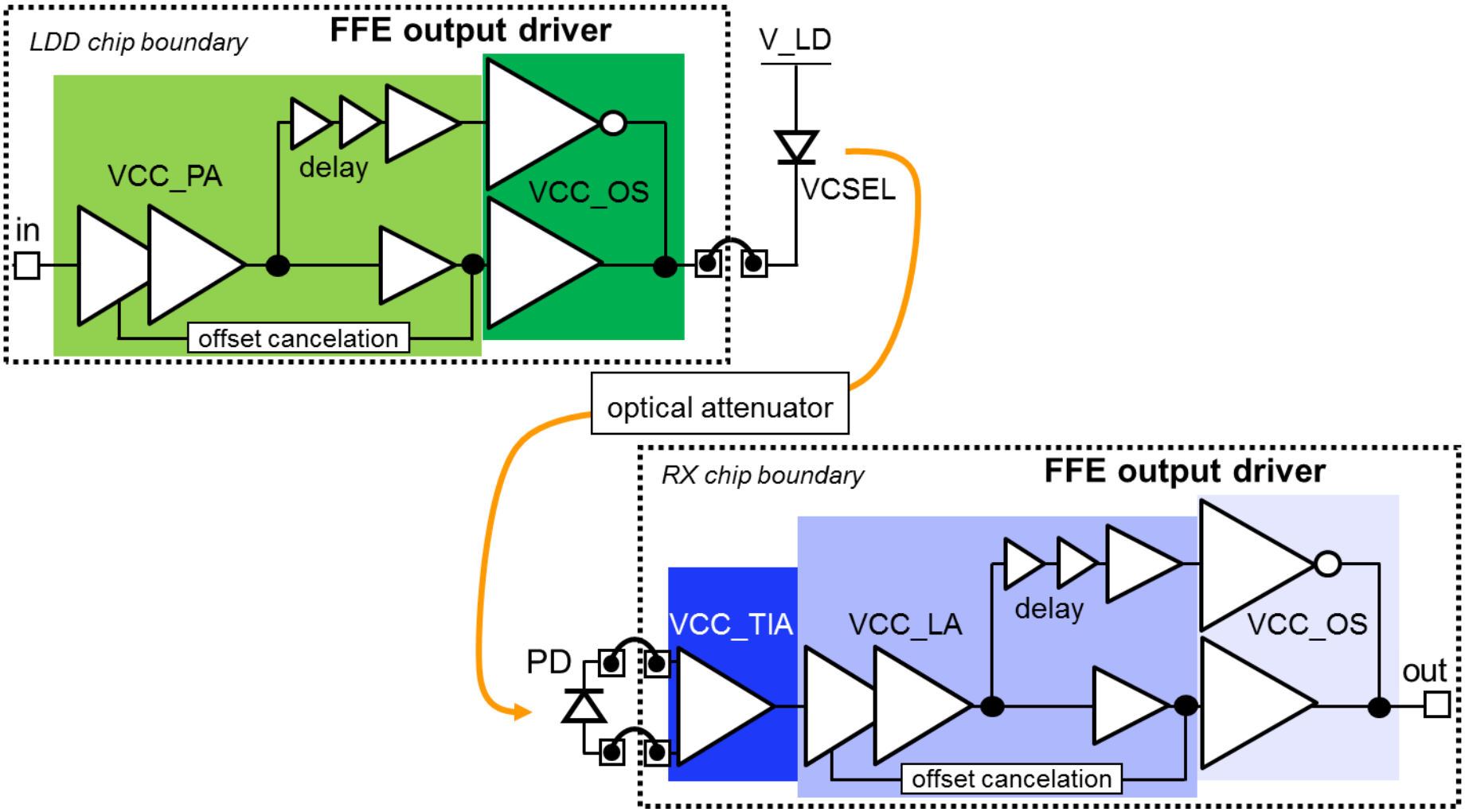
- 4 generations of 90-nm CMOS-driven optical links



10X improvement in power efficiency

[1] C. P. Lai *et al.*, “20-Gb/s Power-Efficient CMOS-Driven Multimode Links,” Optical Fiber Communication (OFC) Conference 2011, Los Angeles, CA, Mar. 2011.
[2] C. L. Schow *et al.,* "A 25 Gb/s, 6.5 pJ/bit, 90-nm CMOS-driven multimode optical link,” *IEEE Photon. Technol. Lett.*, vol. 24, no. 10, May 2012.
[3] J. E. Proesel, C. L. Schow, A. V. Rylyakov , "Ultra low power 10- to 25-Gb/s CMOS-driven VCSEL links," *Proc. Optical Fiber Communication (OFC) Conference 2012*, Los Angeles, CA, Mar. 2012.
[4] J. E. Proesel *et al.*, “Ultra low power 10- to 28.5-Gb/s CMOS-driven VCSEL-based optical links,” *OSA J. of Optical Comm. and Networking*, in press.

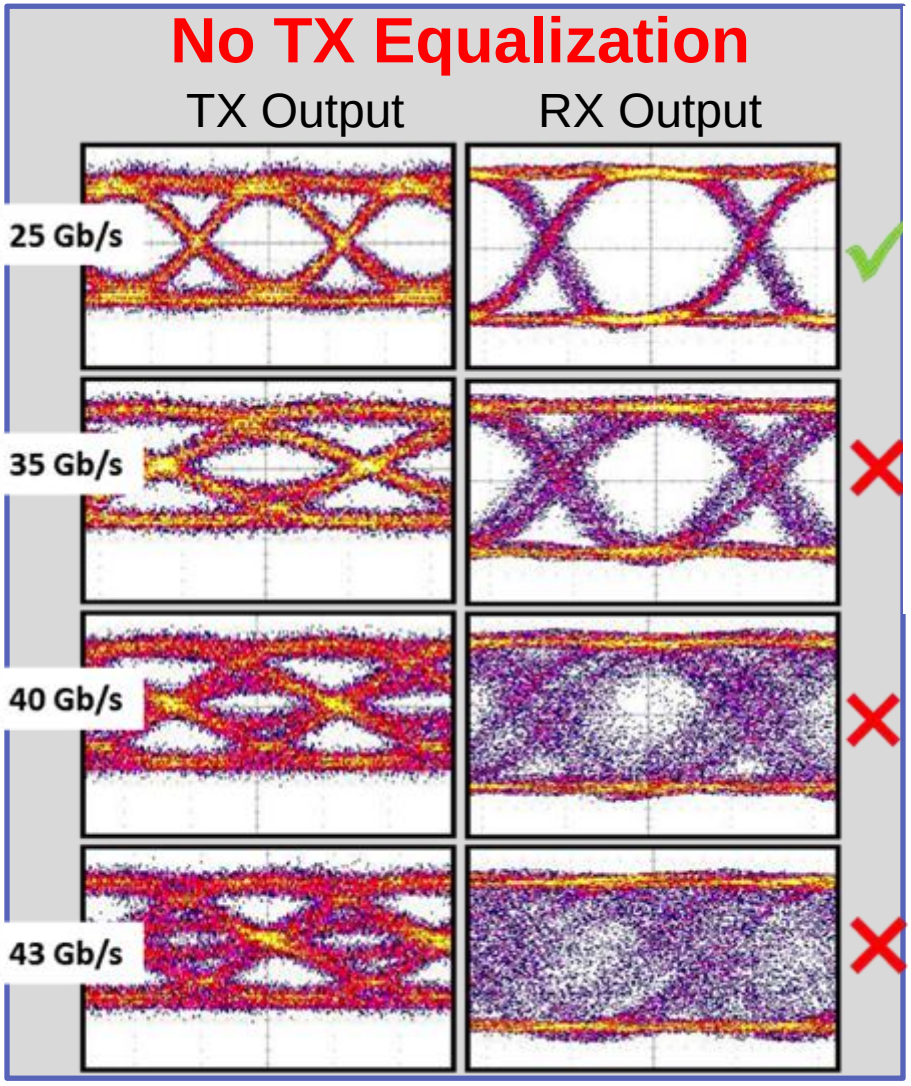
SiGe 8HP (130-nm BiCMOS): Pushing Speed Limits of VCSEL Links



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

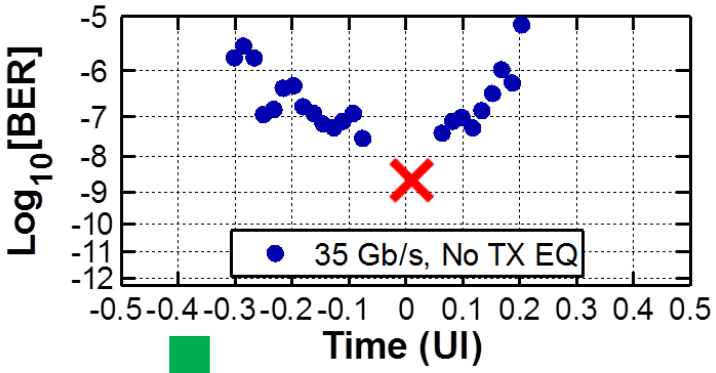
Transmitter-side FFE (TX EQ) is Dramatically Effective

40 Gb/s link using a 20 Gb/s VCSEL

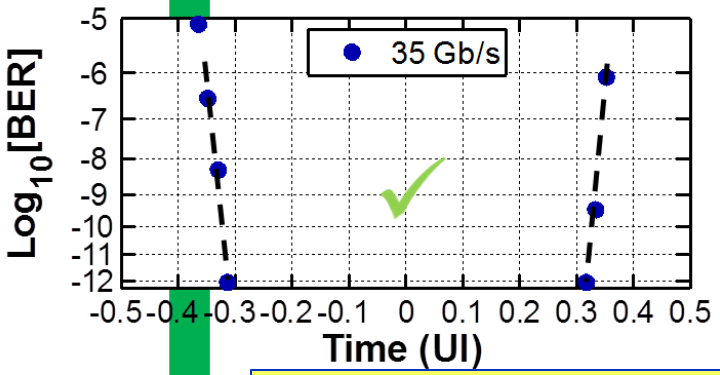


A. V. Rylyakov *et al.* "A 40-Gb/s, 850-nm, VCSEL-Based Full Optical Link," *OFC 2012*, paper OThE1.1, Mar. 2012.

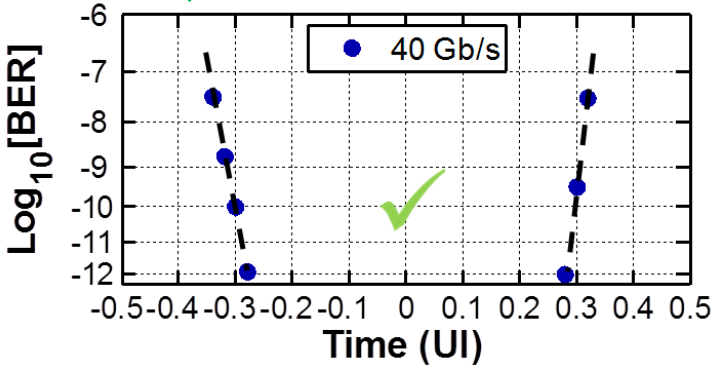
BER Proves Robust Operation at 35- and 40 Gb/s

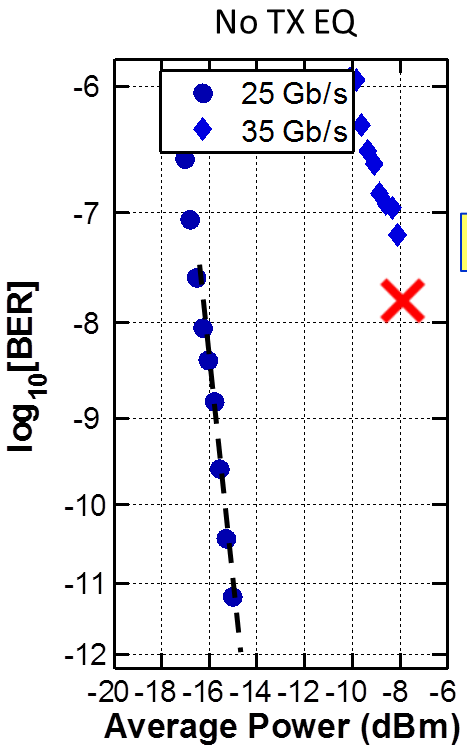


TX EQ Enables 35 Gb/s

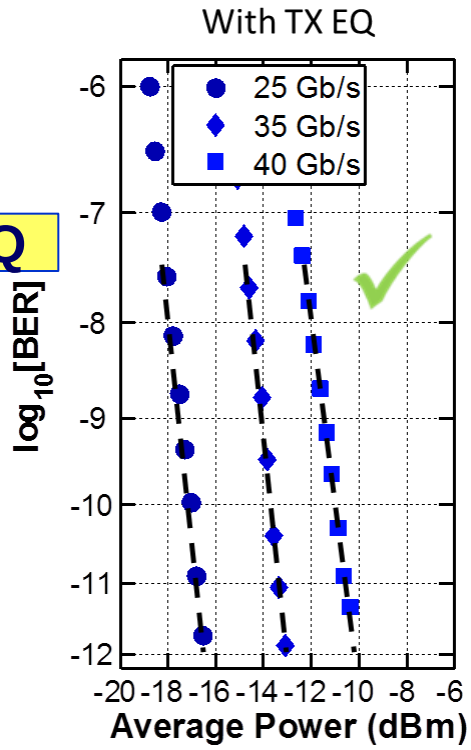


TX EQ Extends link to > 40Gb/s





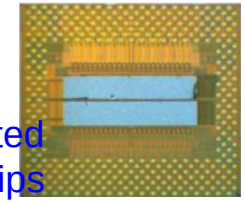
TX EQ



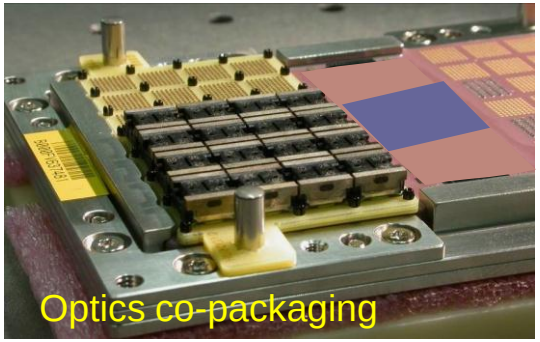
- **The link that is broken at 35 Gb/s has wide margin at data rates > 40 Gb/s with EQ enabled**
- **Power efficiency 22 pJ/bit @ 40 Gb/s, expect future improvements in SiGe, mapping to CMOS**

Summary: Path to High-Speed, Low-Power, Dense Parallel Optical Transceivers

- **Dense hybrid integration can achieve very high performance**
  - Optics near CPU
  - Optics co-packaging
  - Integrated Optochips

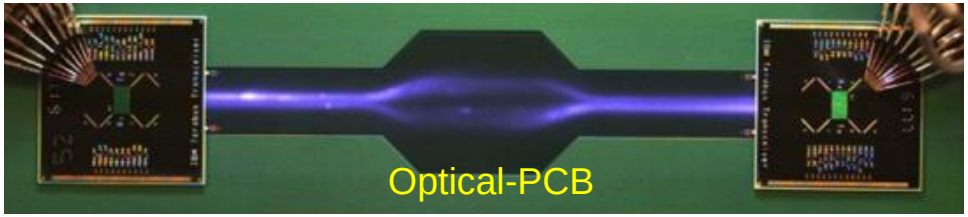


Integrated Optochips

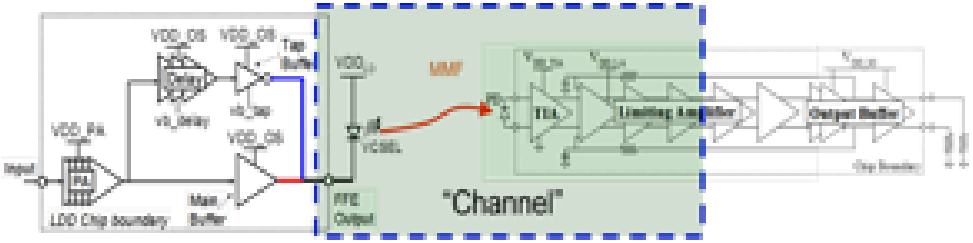


- **Optical-PCB**
  - PCBs with integrated polymer waveguides
  - On-board module-to-module high BW density optical interfaces

Optically-enabled MCMs



- **Advanced Circuits - CMOS scaling**
  - Expect future technologies to offer benefits in speed and power
  - SiGe for ultimate speed
  - Continuous VCSEL improvements also critical in efficiently pushing to higher speeds



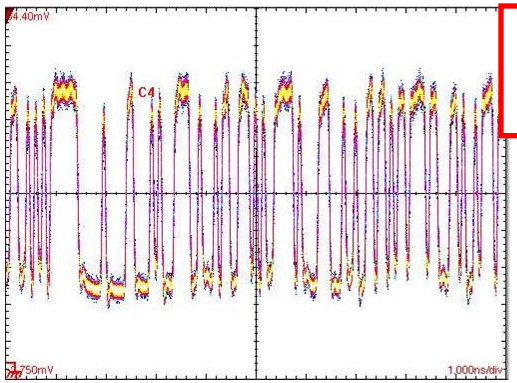
Advanced CMOS

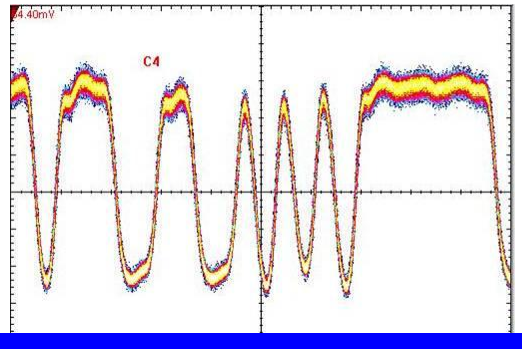
Comments and Outlook

- **VCSEL transceivers evolving to meet exascale needs**
  - Power, speed, density, cost …
- **Coexistence for VCSELs and Si Photonics**
  - No one-size-fits-all technology for optics
  - Incumbent technologies are not easily displaced
  - Critical challenge for Si photonics is packaging.
- **New approaches and innovation required**
  - Shrinking margins to meet aggressive power numbers
  - Equalization and more advanced circuits
  - o-PCB, CWDM, multicore fibers…
- **Future Computercom cost/density/performance targets require holistic approach to design and manufacture of optical interconnects**
  - Closer collaboration between system manufacturer and suppliers of TRX, OPCB …
- **Daunting challenges open many opportunities for innovation**

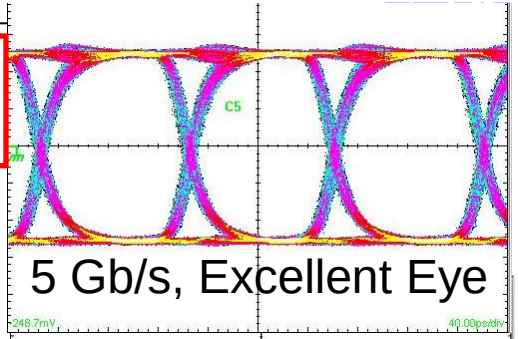
Back up

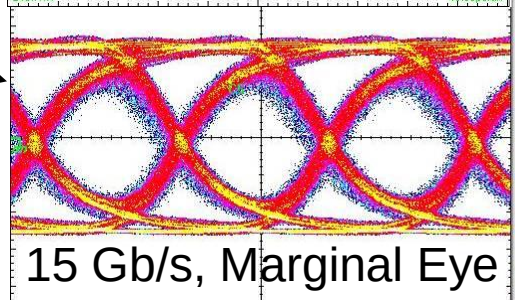
Signal Integrity: Eye Diagrams





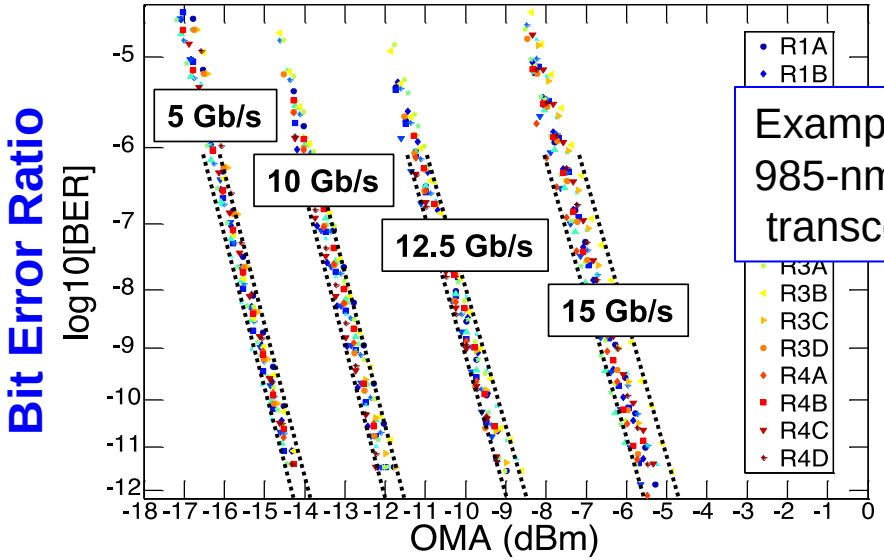
Eye Diagrams = snapshot of performance: amplitude, speed, noise, jitter, distortion…





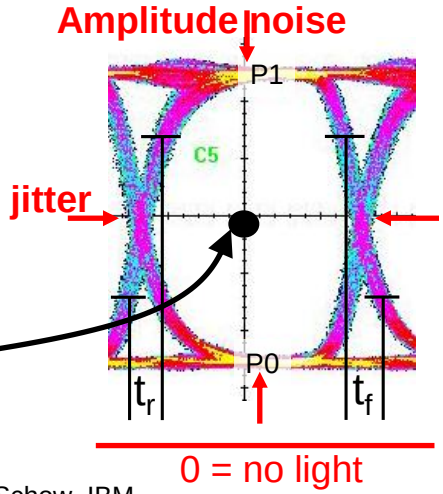
Receiver Sensitivity

- Min optical power required for specified BER (often 10^{-12})
- Degrades at higher data rates due to bandwidth limitations



Example: 985-nm Terabus transceivers

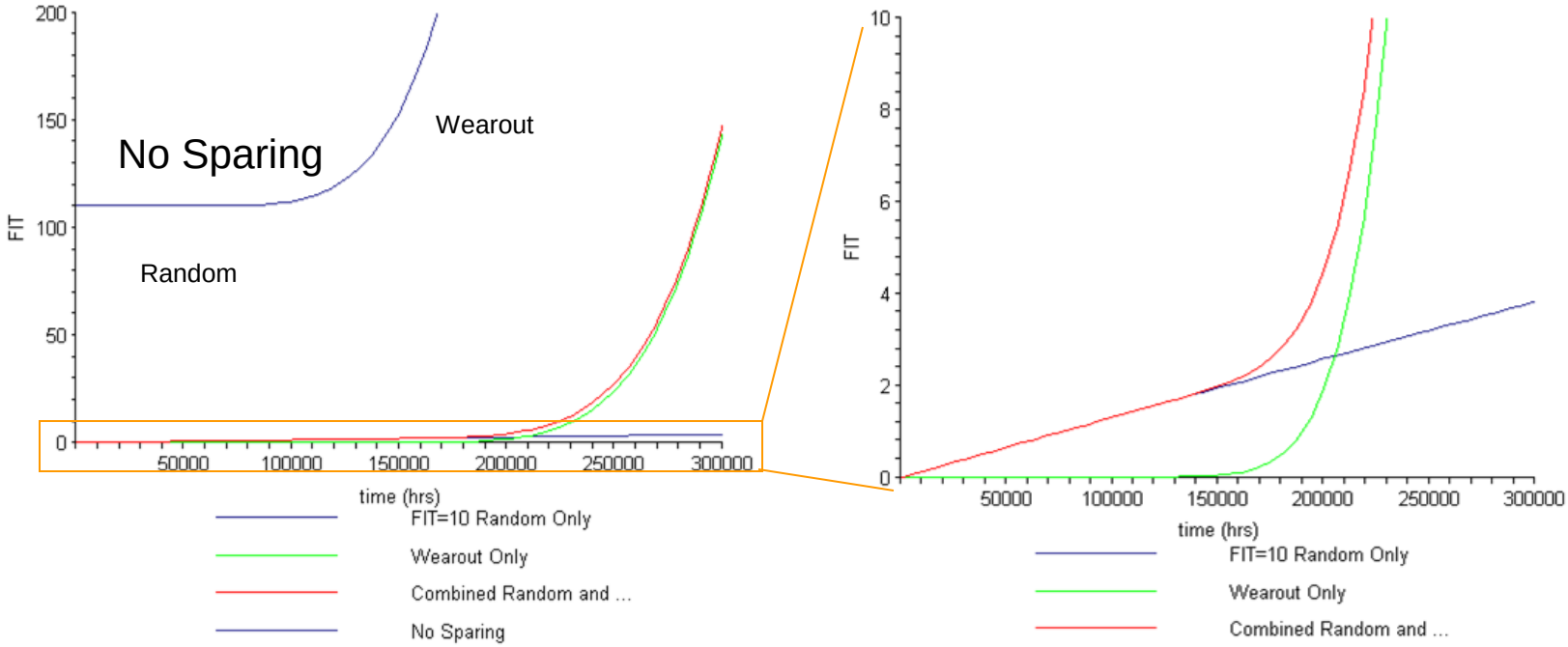
ISI: Inter-symbol interference, eye depends on bit history



Extinction Ratio, P1/P0

Courtesy Clint Schow, IBM

Optical Link Reliability for PetaScale+: Sparing (11+1)



- Effective FIT rate for 11ch 'link' with typical VCSEL Wearout and Random FIT=10/device
- Sparing (12th device) reduces Eff. FIT to low levels.

# of Links	No Spare VCSEL + 50 FIT unspared	Spare VCSEL + 50 FIT unspared [time to 1st fail]	Spare VCSEL only [time to 1st fail]
1K	1.5 fails/year	20Khrs	252Khrs
10K	1 fail/month	2Khrs	174kHrs
100K	2.68 fails/week	200hrs	39Khrs
1M	3.8 fails/day	20hrs	12Khrs

ExaScale will need sparing + ultra-reliable components

Example: Two Fabricated Si Photonic Switches

- **Mach-Zehnder: less T sensitivity, more power, larger area**
  - **Bandwidth >100nm**
- **Ring Resonator: high Q for lower power but T sensitivity, also much smaller**
  - **Bandwidth < 1nm**

Ring Resonator (2RRS)

50-µm scale bars

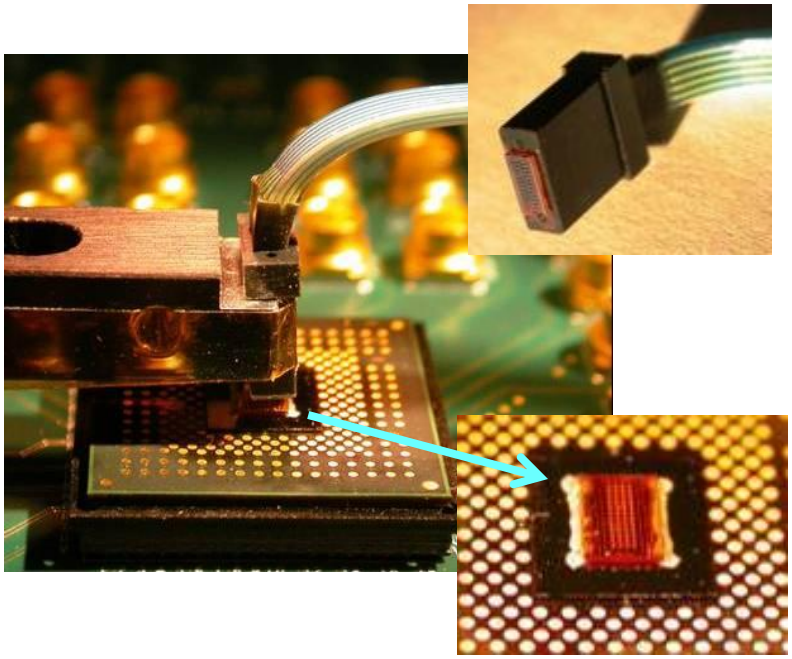
Wavelength-Insensitive Mach Zehnder (WIMZ)[1]

[1)] J. Van Campenhout, *Optics Express*, **17** (26) 24020.

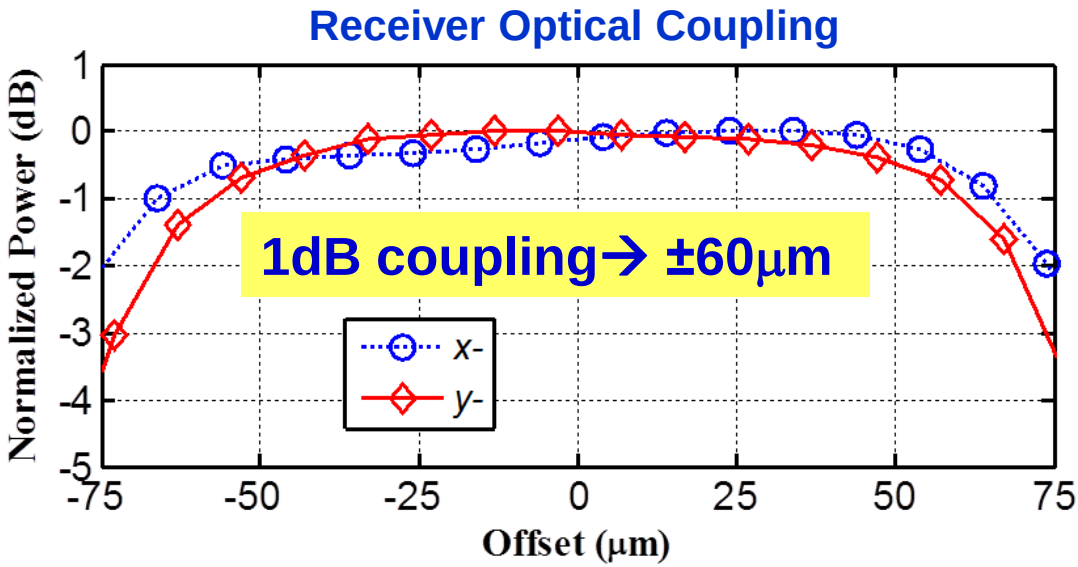
Challenges:

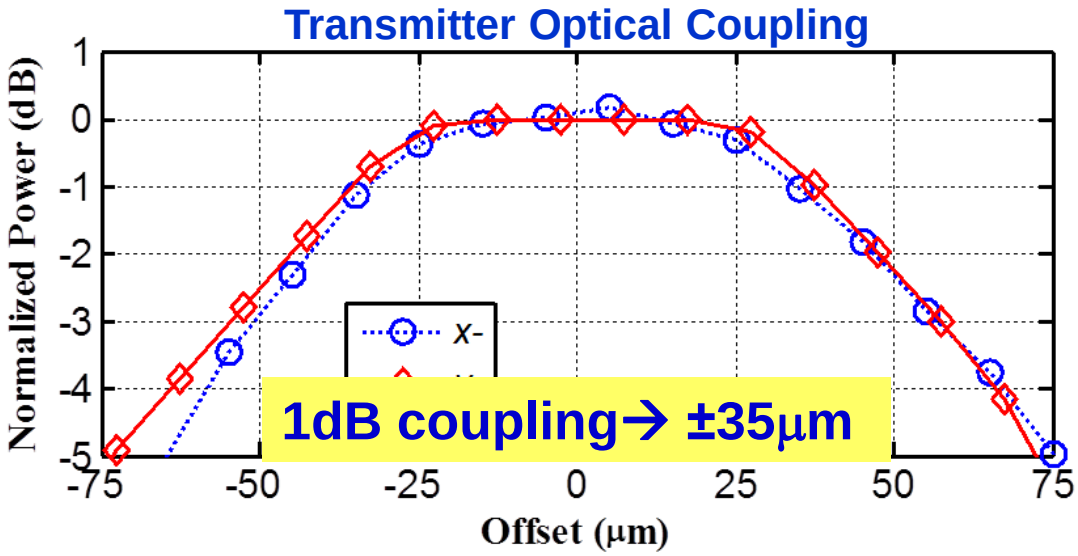
- Packaging
- Reliability
- Thermal
- Polarization issues

Key Advantage for Multimode Optics: Alignment Tolerance



- Two-lens system
- **Achieves relaxed tolerances compatible with pick and place tools**
- Similar alignment tolerances achieved for waveguide coupled links





C.L. Schow *et al.*, "A 24-Channel 300Gb/s 8.2pJ/bit Full-Duplex Fiber-Coupled Optical Transceiver Module Based on a Single "Holey" CMOS IC," *IEEE JLT*, Feb 2011.