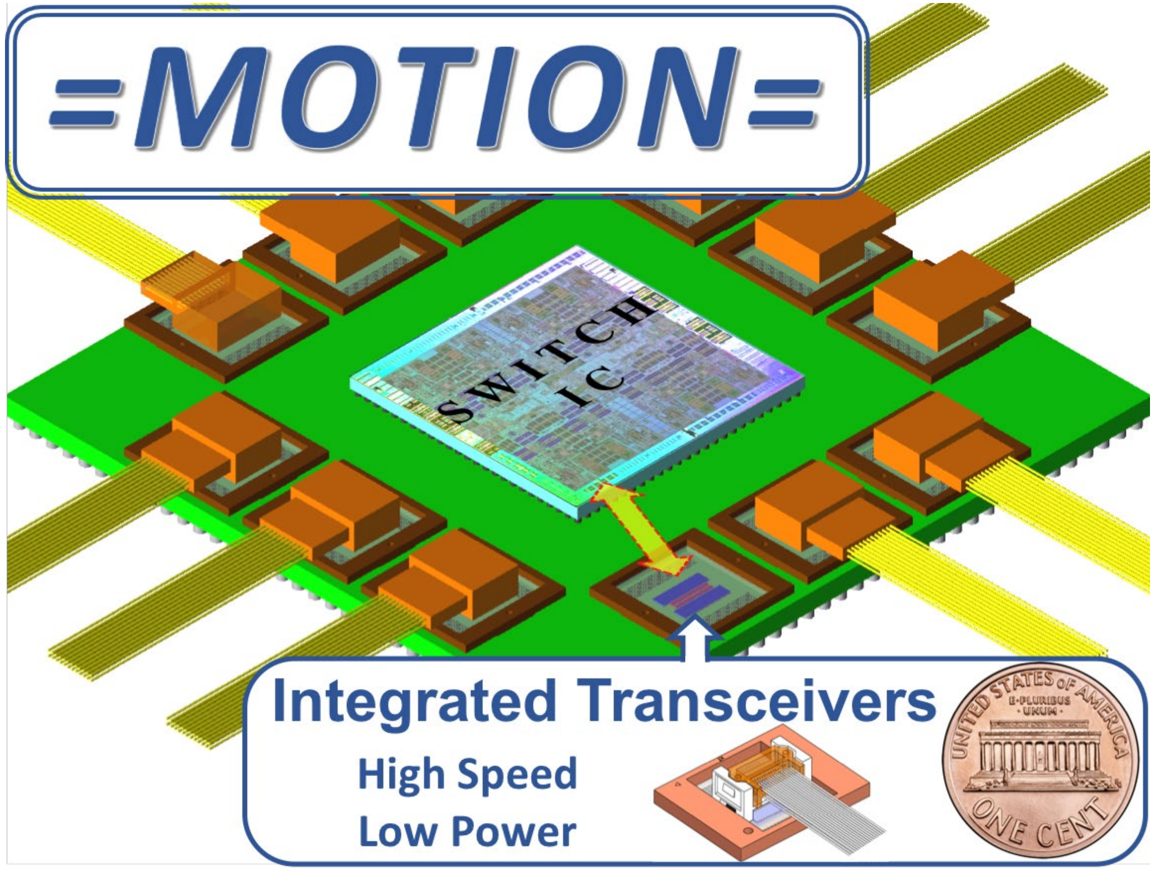
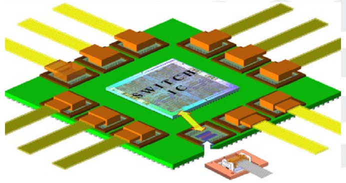


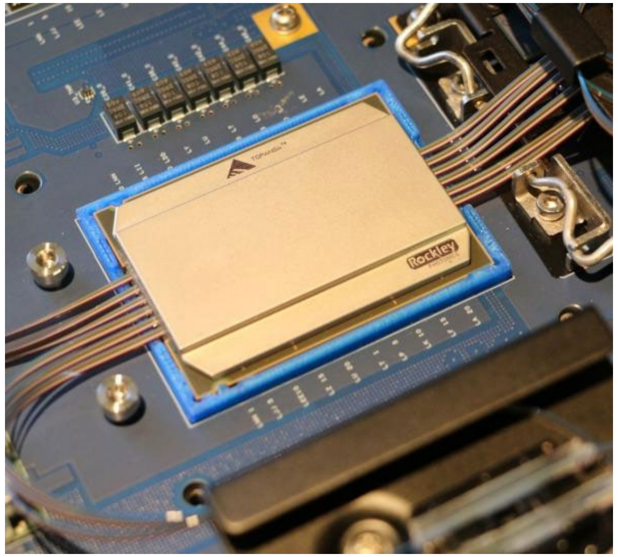
Multi-wavelength Optical Transceivers Integrated On Node

PI: Dan Kuchta, IBM and II-VI (Finisar)

Outline



- IBM P775: 1st Commercial system to use Co-Packaged Optics
- Motivations for Co-Packaging
- Overview of MOTION Project and Technology
- SAFE (Simplified Analog Front End) ICs
- Flip-chip VCSELs and Photodiodes
- Network Modelling & Simulation results
- MOTION Phase 2

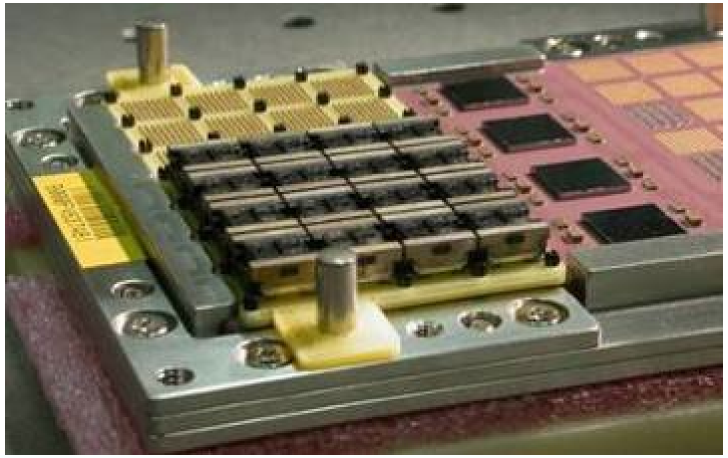


Rockley Photonics c. 2018

1.2+1.2 Tb/s
96 fibers at 25 Gbps

Altera & Avago c. 2012

120+120 Gbps
24 fibers at 10 Gbps



IBM c. 2010

3.4+3.4 Tb/s
672 Fibers at 10 Gbps

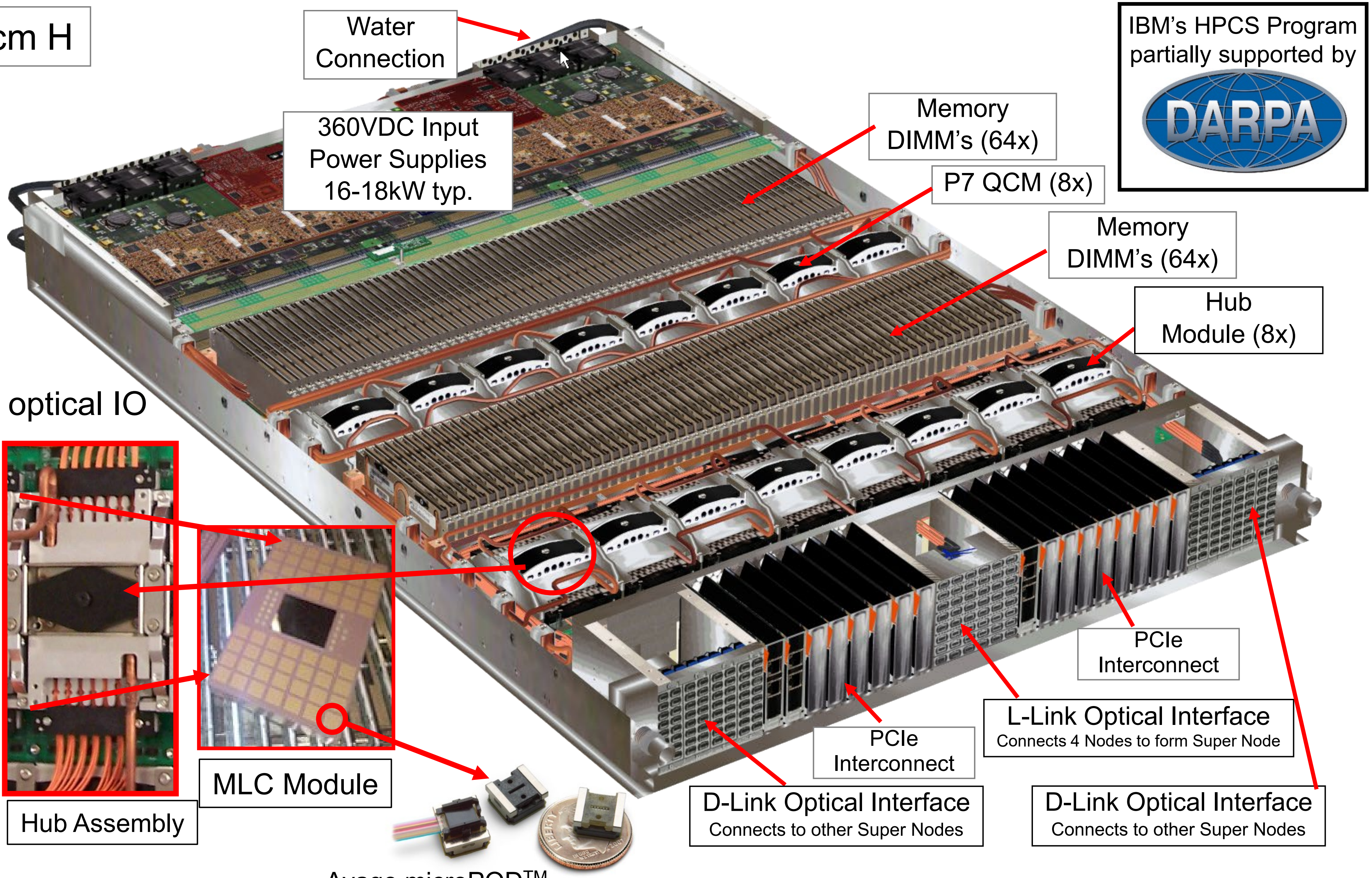
IBM Power 775: 1st system with co-packaged Optics c.2010

IBM

1m W x 1.8m D x 10cm H

IBM's HPCS Program partially supported by DARPA

Each Drawer:
8TF/s, 4TBytes DRAM
24+24 Tb/s co-packaged optical IO



Water Connection

360VDC Input Power Supplies 16-18kW typ.

Memory DIMM's (64x)

P7 QCM (8x)

Memory DIMM's (64x)

Hub Module (8x)

PCIe Interconnect

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

PCIe Interconnect

D-Link Optical Interface
Connects to other Super Nodes

D-Link Optical Interface
Connects to other Super Nodes

Hub Assembly

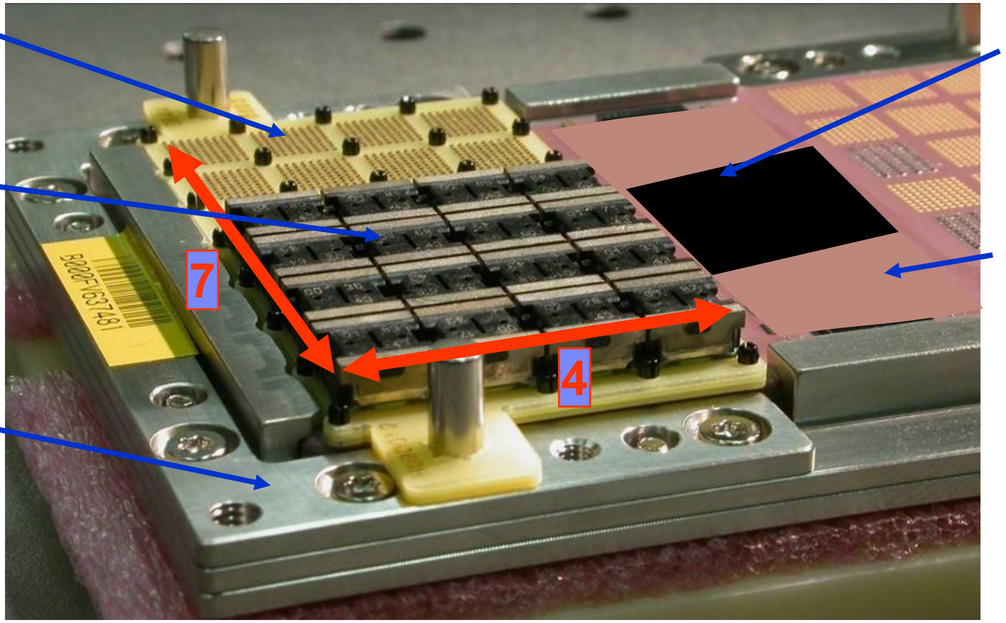
MLC Module

Avago microPOD™

Power775 Hub Assembly with Co-Packaged Optics

- Ceramic with Switch ASIC, **μLGA**, & Optical Modules

•Not Field Replaceable*
•Fail in place Strategy
•~$1Gb/s in 500K volumes in 2010



μLGA Interposer

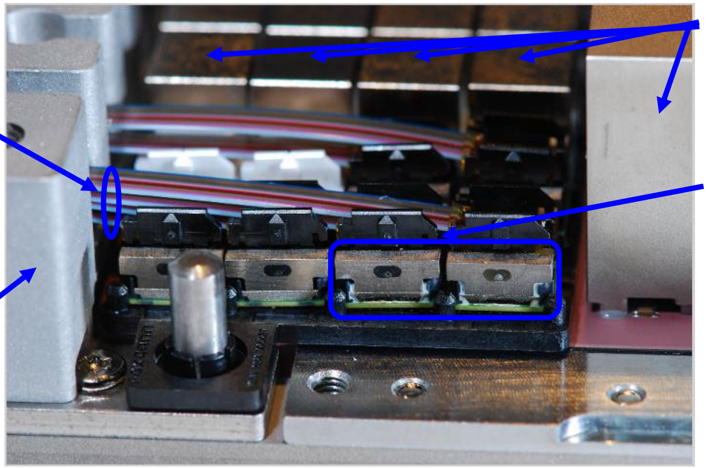
Optical Modules

Frame

Switch ASIC

Ceramic Substrate

- **..and after Fibers & Heat Spreaders are added**



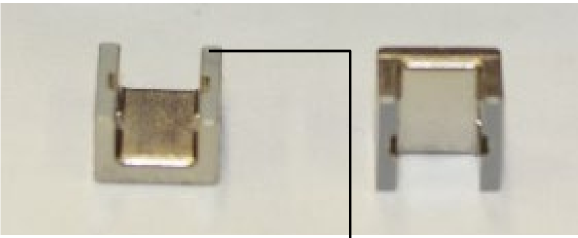
Heat Spreaders

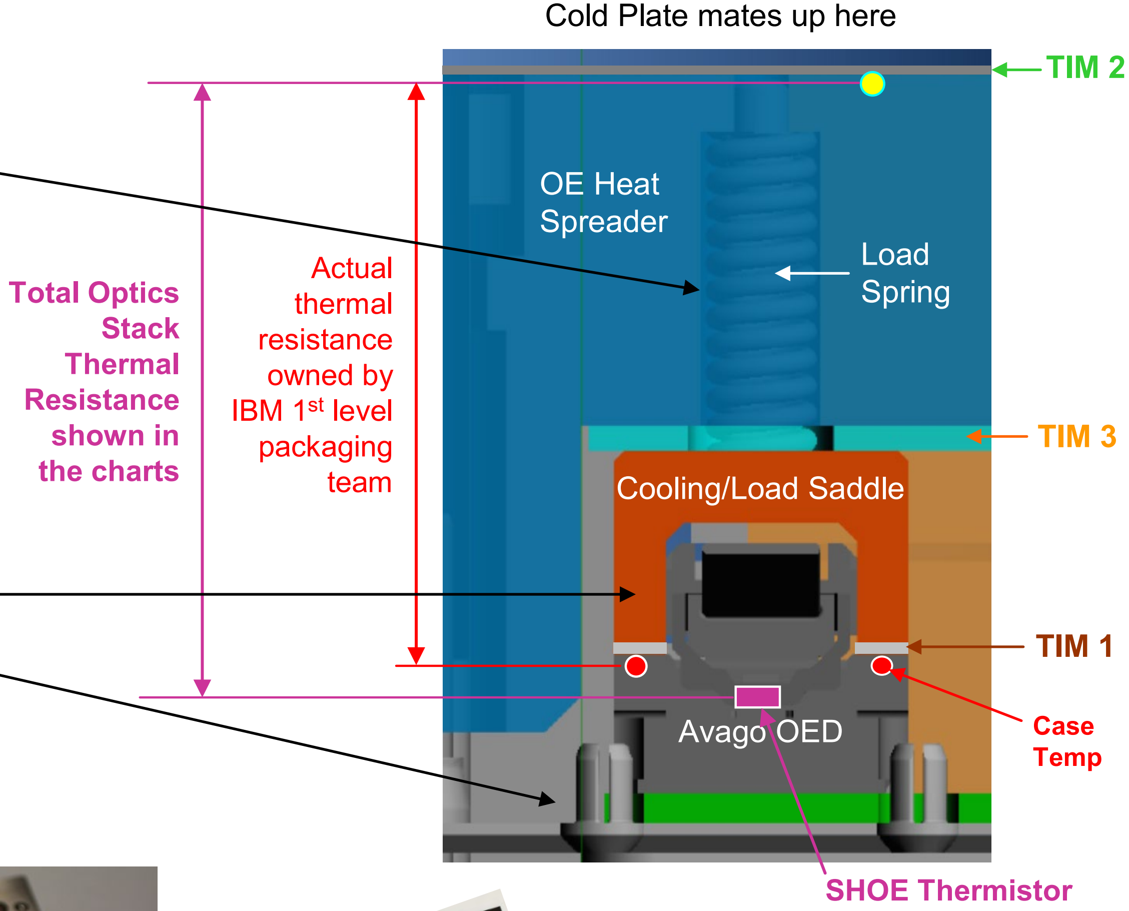
4 Optical Fiber Ribbons
2 incoming, 2 outgoing
12 fibers each

Strain Relief for Optical Fiber Ribbons

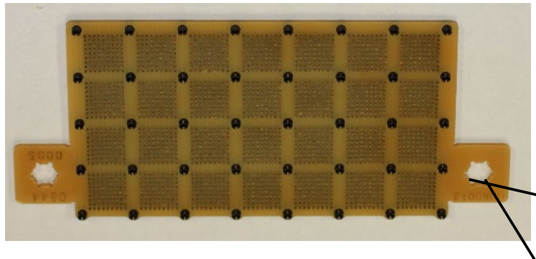
(Tx+Rx) Pair
(12+12) optical channels, 10 Gb/s per fiber
Implements either 1 D-link or 2 LR-links

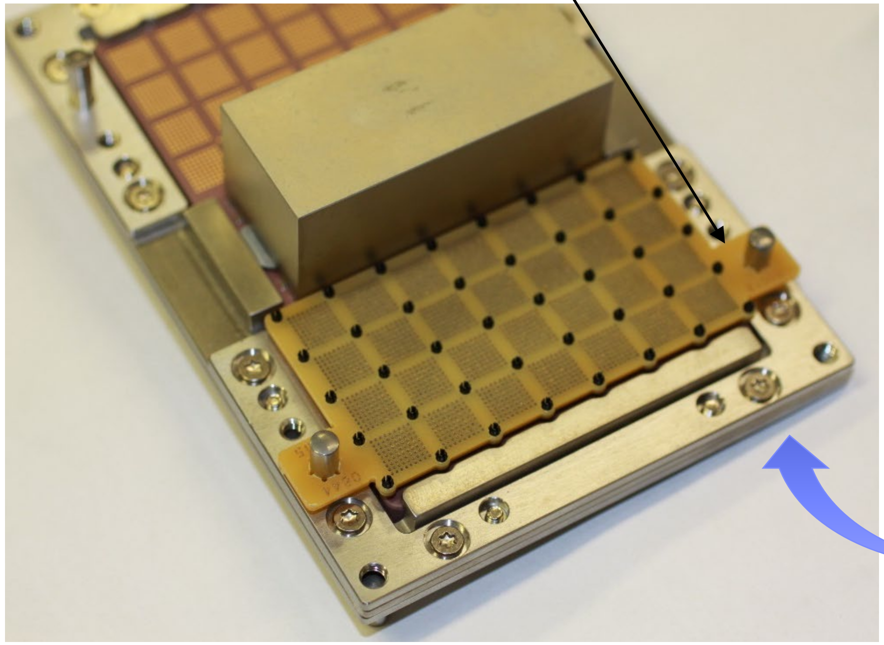
Power775: LGAs and Thermal Interfaces for co-packaged optics

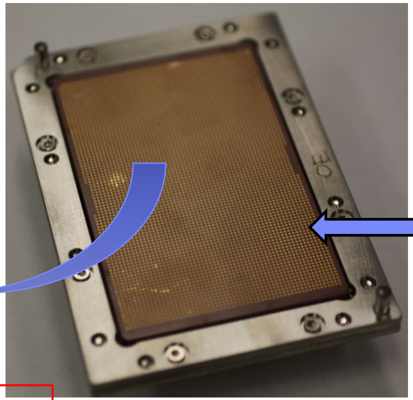


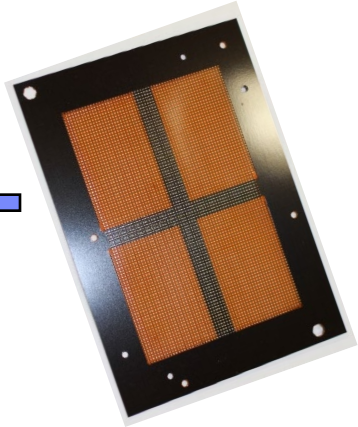

Cold Plate mates up here
TIM 2
OE Heat Spreader
Load Spring
Total Optics Stack Thermal Resistance shown in the charts
Actual thermal resistance owned by IBM 1st level packaging team
TIM 3
Cooling/Load Saddle
TIM 1
Case Temp
Avago OED
SHOE Thermistor

μLGA Interposer for co-packaged optics





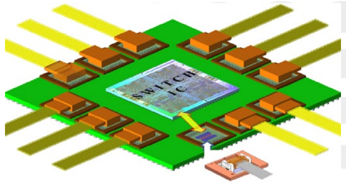




LGA Interposer for hub module

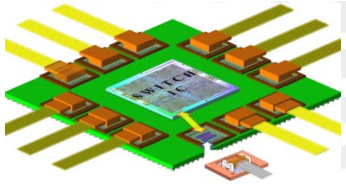
LGAs make Thermal Path and Mounting more challenging….

Why are we *(still)* interested in Co-Packaging?



- Primarily for increasing BW from ASICs
  - Large ASICs are package pin constrained
  - Co-Packaging permits I/O from both sides of the package
- Reduction in Power Consumption
  - Juxtaposed die do not require high power SERDES
  - Lower power I/O cells also use less Si Area
- Reduction in Cost
  - Stripped down optical packages and reduced function ICs *should* cost less
  - Reduced ASIC area will have higher yield
- Expansion of ASIC performance
  - Instead of reducing ASIC Area and Power, other choice is to maintain size and add more functionality to the chip up to the original Power constraint
  - i.e. Additional switch ports or Additional Memory Hubs

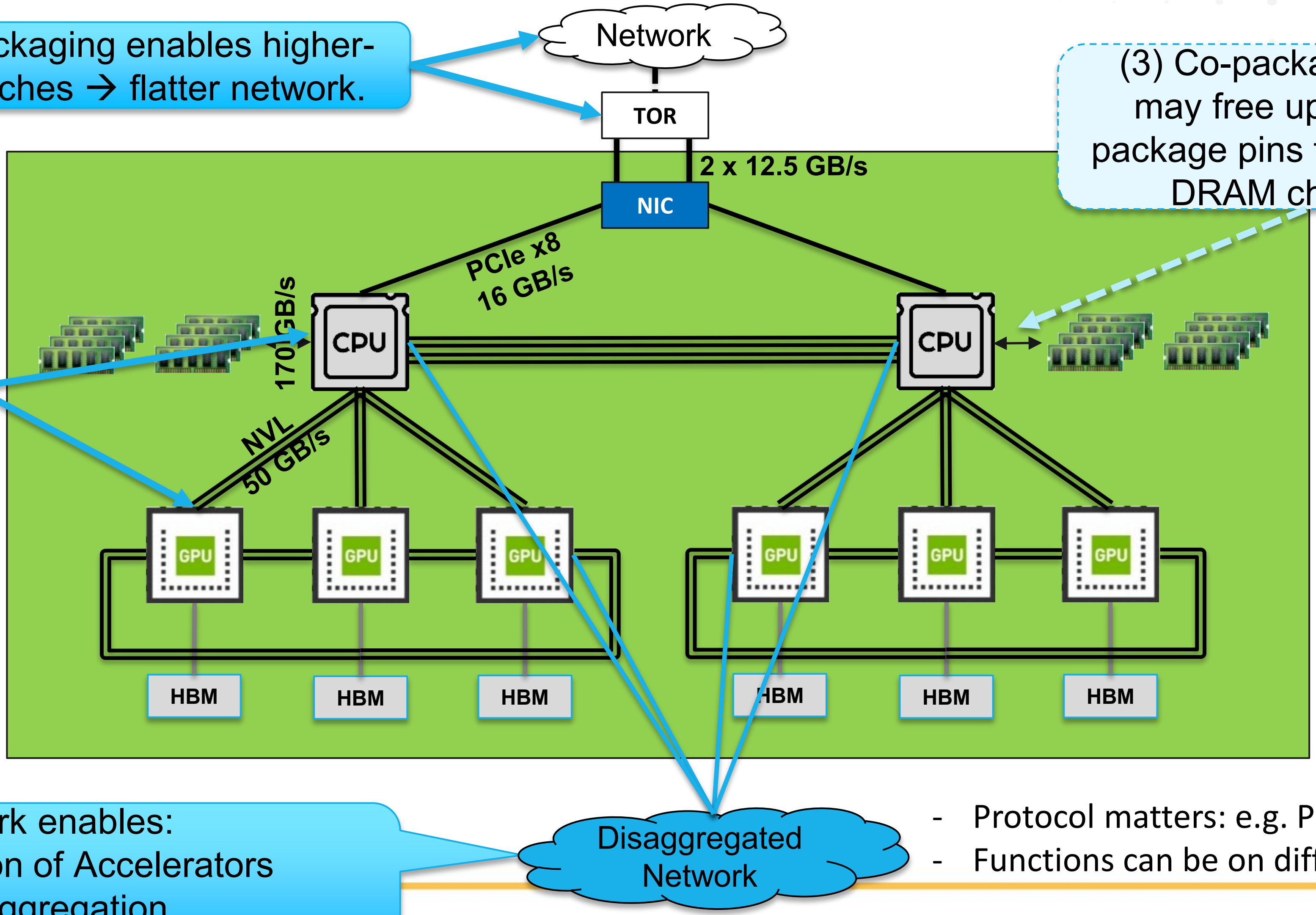
Where can Co-Packaging alleviate bandwidth bottlenecks?



(1) Co-packaging enables higher-radix switches → flatter network.

(2) Optics on CPU & GPU modules enables higher on-node BW, or moving them on different boards.

(3) Co-packaged optics may free up electrical package pins for additional DRAM channels.



(4) Disaggregated network enables:

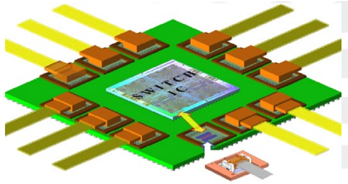
I. On-demand allocation of Accelerators

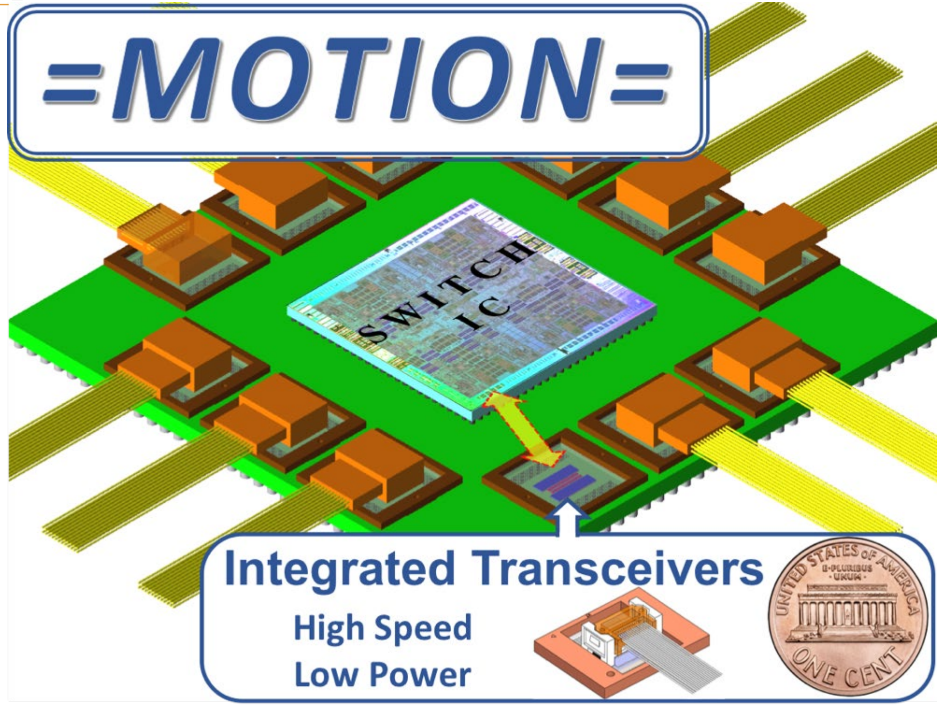
II. NVMe / storage disaggregation

III. Memory “stealing”

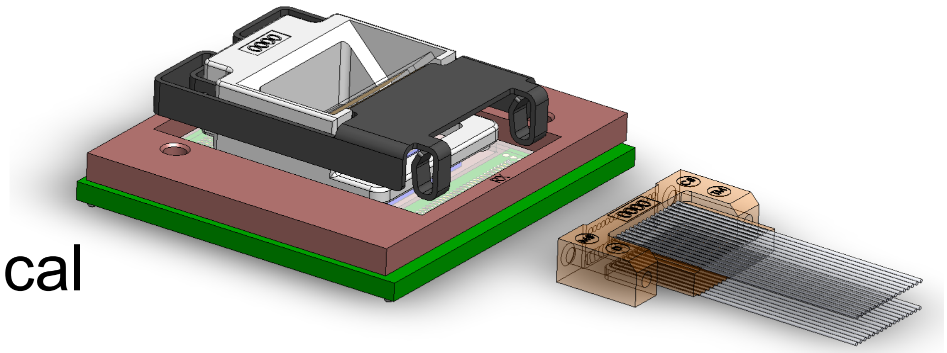
- Protocol matters: e.g. PCIe, CXL, …
- Functions can be on different boards

MOTION Phase 1
Co-packaging for CPU/GPU High-level Specifications

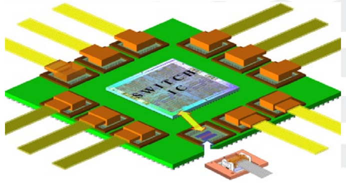




- ARPA-E Sponsored Project on co-packaging
- IBM and Finisar collaboration
- 56GBd NRZ; BER tested to <1E-12 pre-FEC
- 0^{o}C to 70^{o}C Case
- 6dB (electrical) link budget (XSR-like)
- 2 dB optical link margin (30m w/connectors)
- Solderable onto ASIC 1st level substrate
- < 4 pJ/bit (3.2W, 16 channels)
- W:13mm x D:13mm x H:4mm
- 25¢/Gb/s

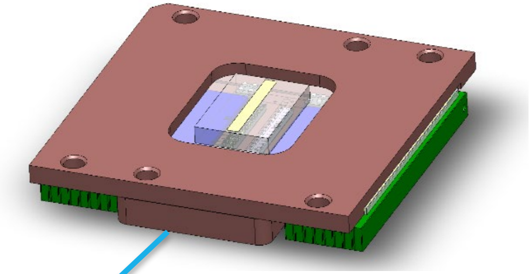


=MOTION=: Multi-wavelength Optical Transceivers Integrated on Node

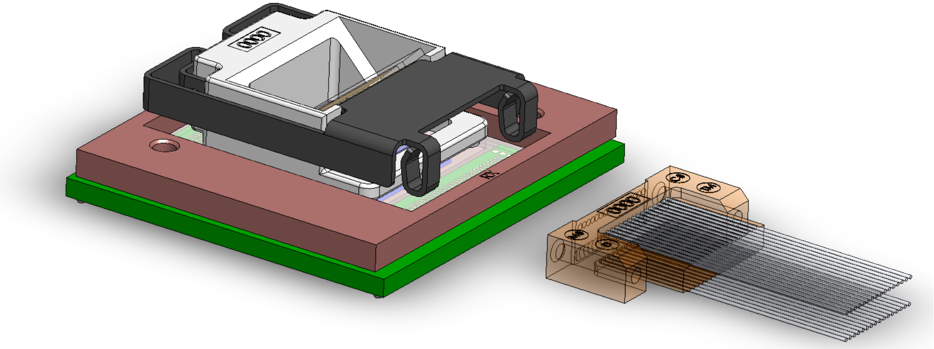


MOTION Transceiver Package Overview

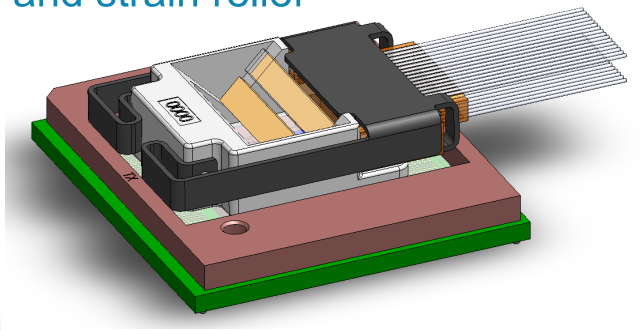
Chip-Scale Optical Package (CSOP)



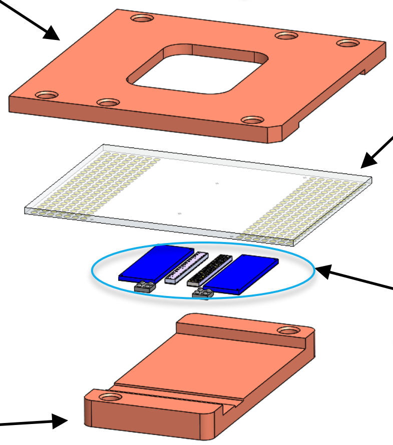
Final Assembly with lens and clip attached

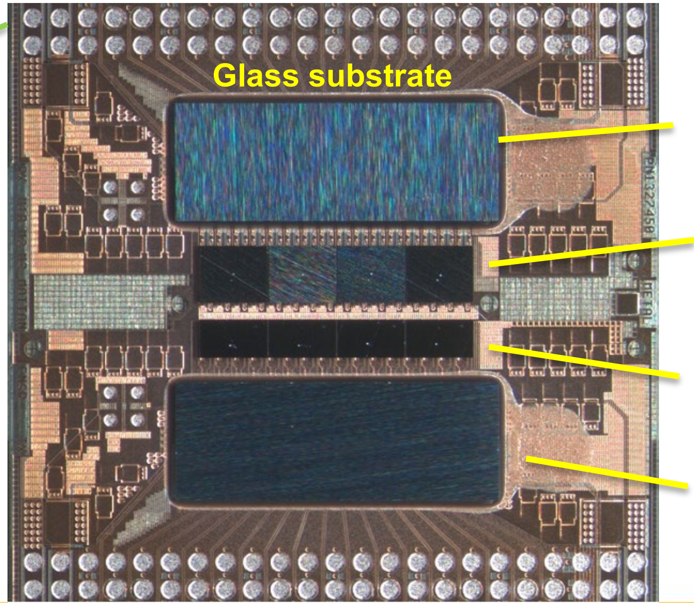


Fully Assembled with fiber cable and strain relief



4mm total height





Chip join process

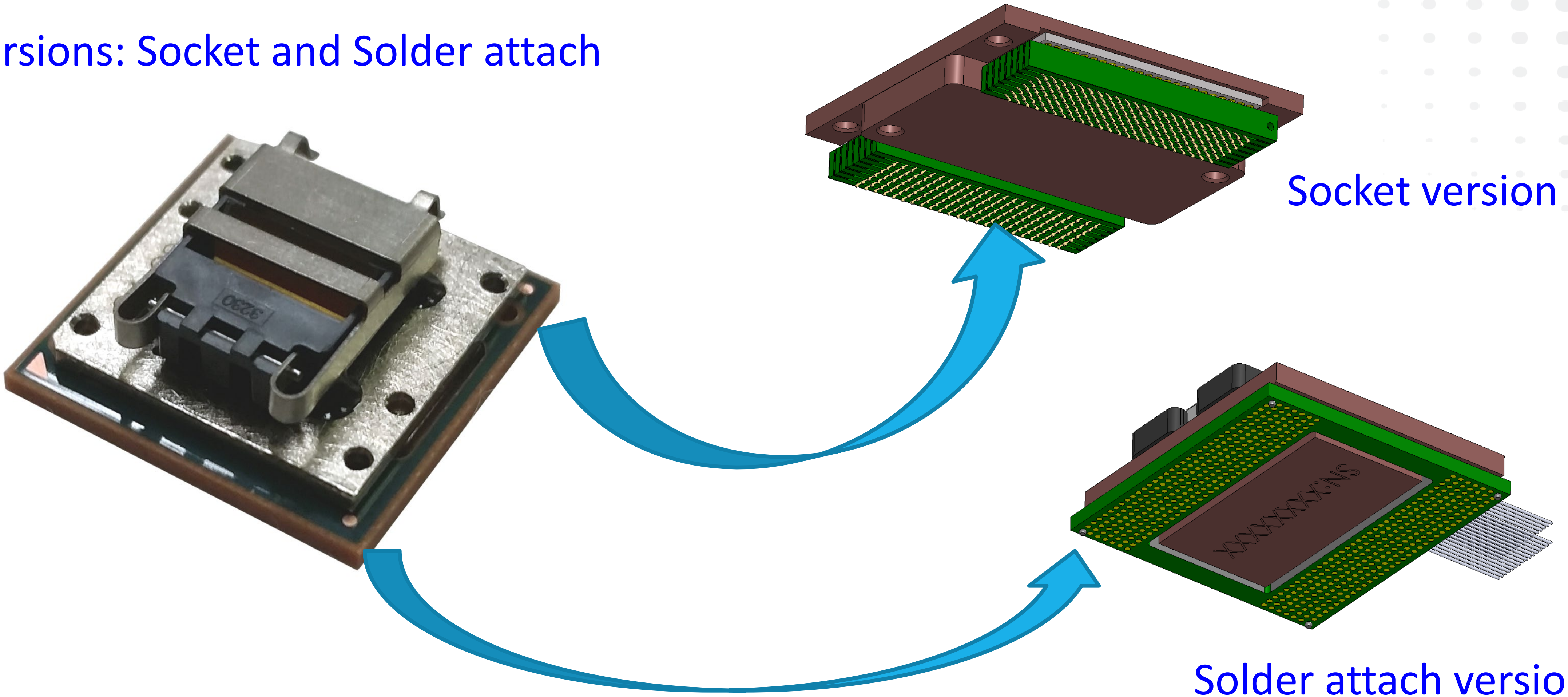
- SAFE chip – Cu pillar w/SnAg cap
- PD/VCSEL – AuSn

Underfill process

- Structural UF for electrical chips
- Epoxy based Optical UF for OEs

A 13mm x 13mm package with socket insert or interposer

Two Versions: Socket and Solder attach

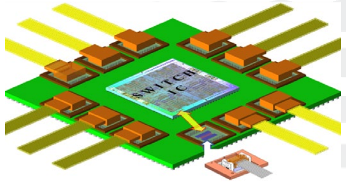


Socket version

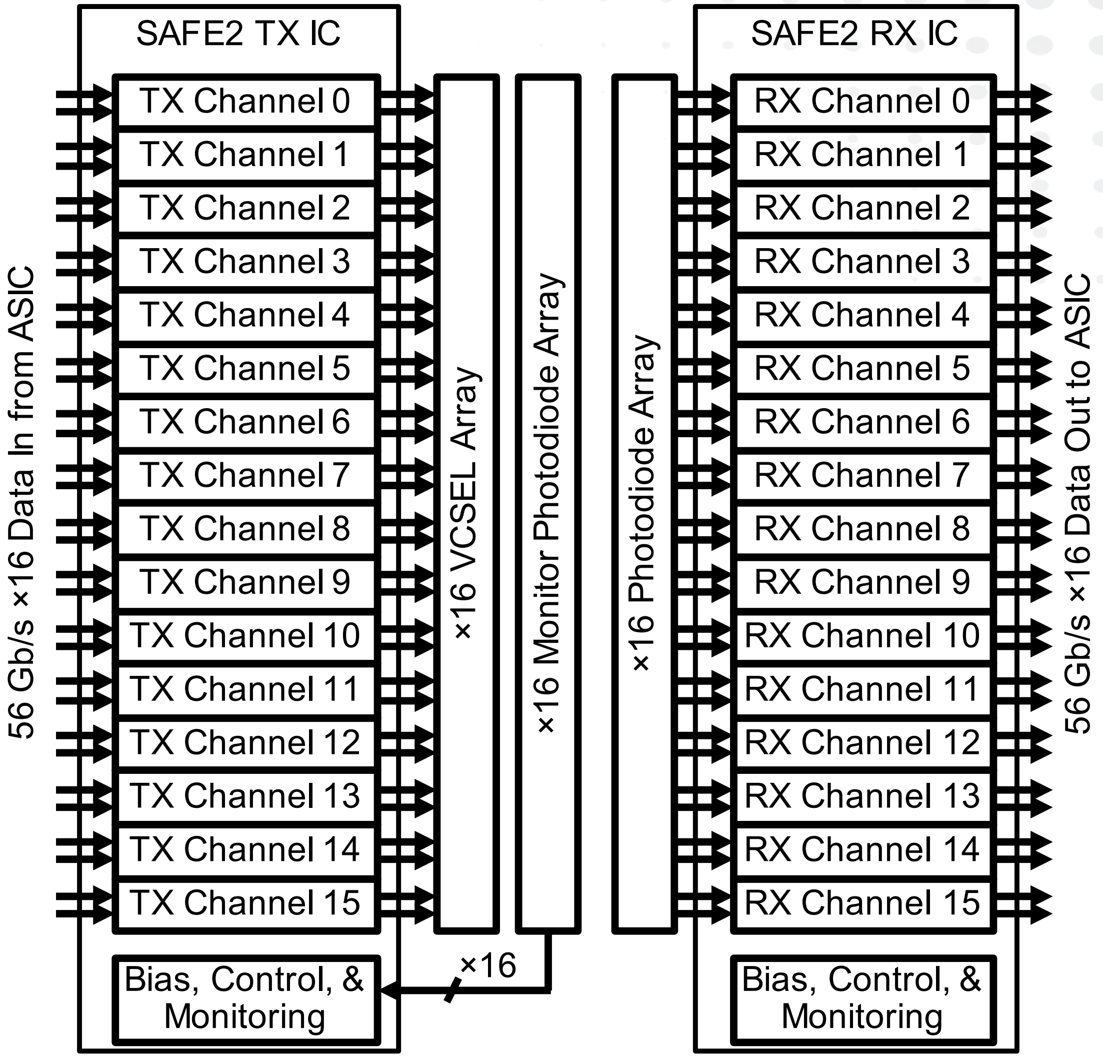
Solder attach version

Sockets may be helpful but incur cost as well as signal & area loss!

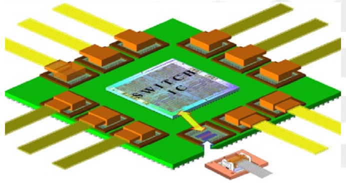
Optical Transceiver Architecture



- Simplified Analog Front End (SAFE)
- SAFE2: 55 nm SiGe BiCMOS
- 16 channels × 56 Gb/s NRZ ~900 Gb/s/IC
- No retiming / CDR
- 4 pJ/bit power consumption
- Fully DC-coupled: passes 64b/66b & PRBS31
- CMOS-compatible electrical signal levels
- Built-in pattern generators and error detectors
- Only 2 Power supply voltages: 1.8v and 3.3v



SAFE2 IC Summary



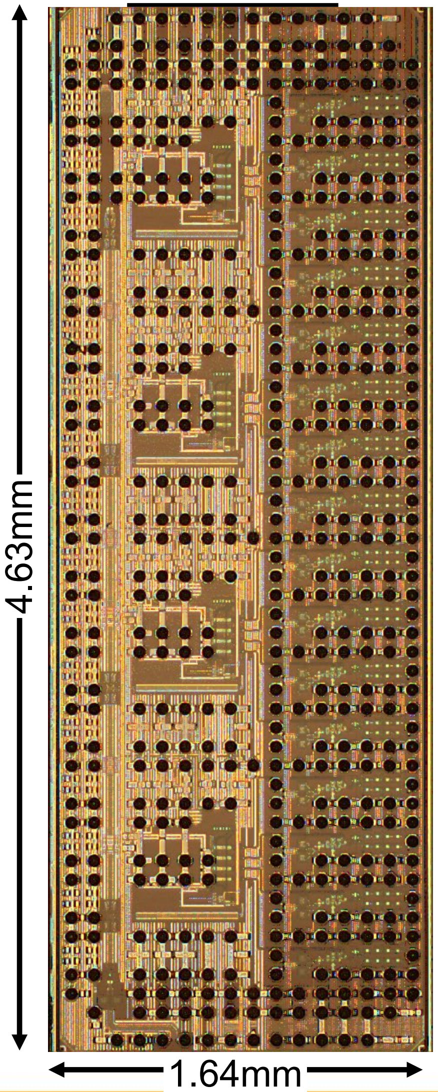
Tx Power Consumption

Power supply	Data mode	PRBS mode
1.8 V	820 mA	1.6 A
3.3 V	256 mA	256 mA
Total power	2.3 W	3.7 W
Energy efficiency	2.5 pJ/bit	4.1 pJ/bit

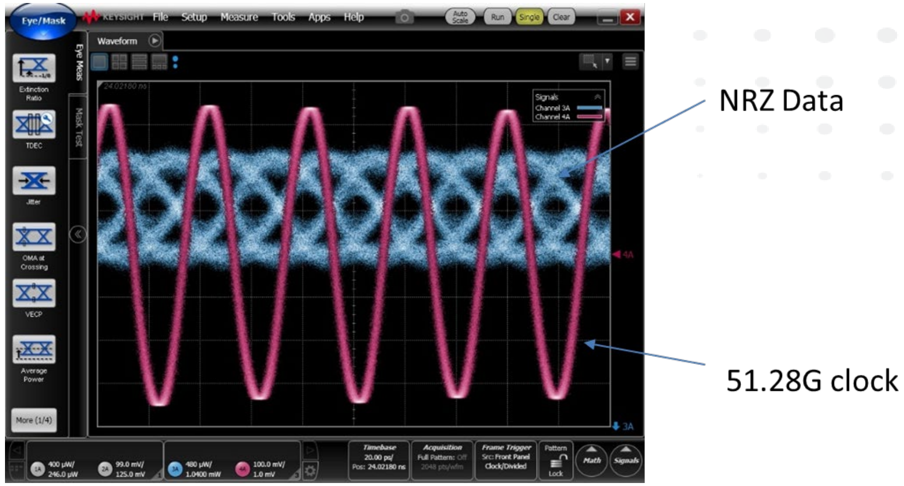
Rx Power Consumption

Power supply	Data mode	PRBS mode
1.8 V	480 mA	1.2 A
3.3 V	120 mA	130 mA
Total power	1.3 W	2.7 W
Energy efficiency	1.5 pJ/bit	3 pJ/bit

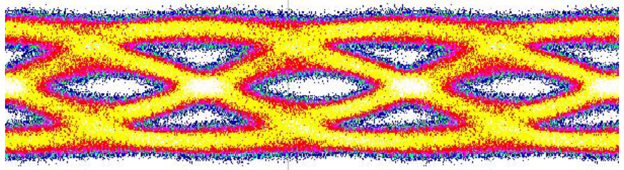
TX IC Photo



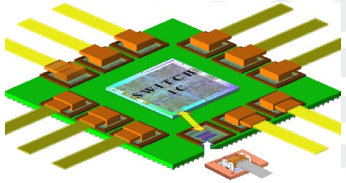
50Gbps NRZ data mode



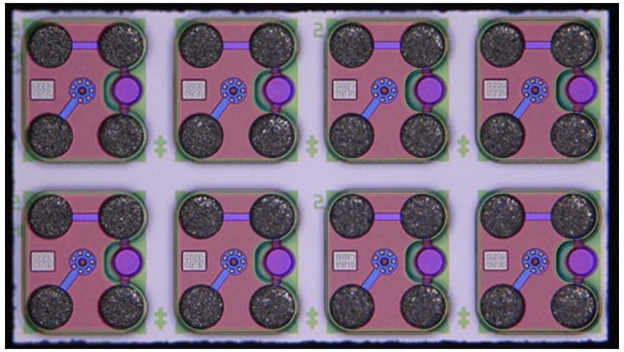
50Gbps NRZ PRBS mode

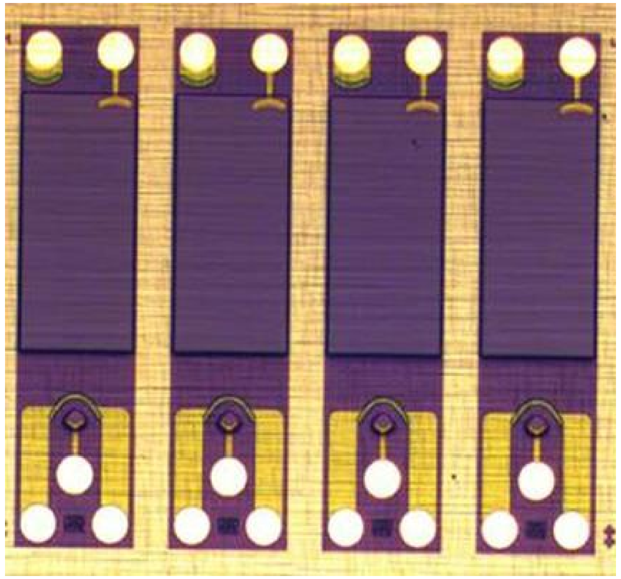


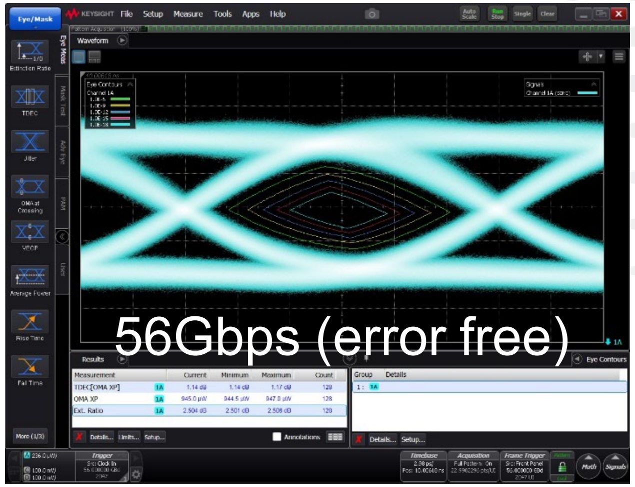
56GBd VCSELs and Photodiodes



- Designed and fabricated flip chip 56GBd 940nm VCSELs on a production epi and wafer fab process
  - Extended to 112G PAM4 for Phase 2
- Implemented a dual aperture structure to realize 1:1 laser sparing
- Designed and Fabricated two different 56GBd Photodiodes structures: GaAs and InP substrate based.

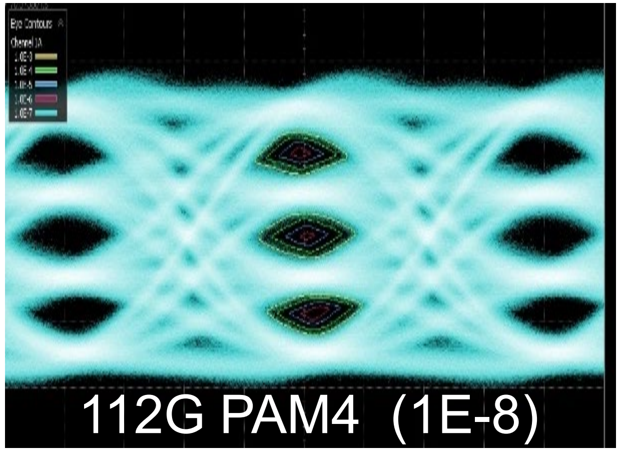






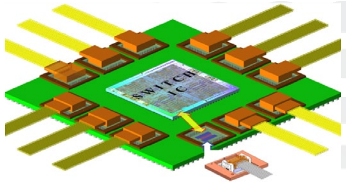
56Gbps (error free)

Measurement		Current	Minimum	Maximum	Count
TDEC[OMA XP]	1A	1.14 dB	1.14 dB	1.17 dB	128
OMA XP	1A	945.0 µW	944.5 µW	947.0 µW	128
Ext. Ratio	1A	2.504 dB	2.501 dB	2.508 dB	128



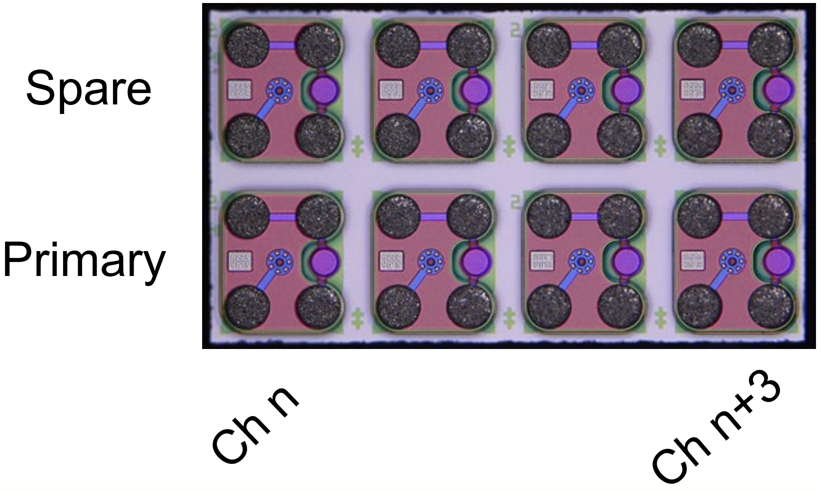
112G PAM4 (1E-8)

MOTION Approach to Reliability: 2-to-1 Sparing

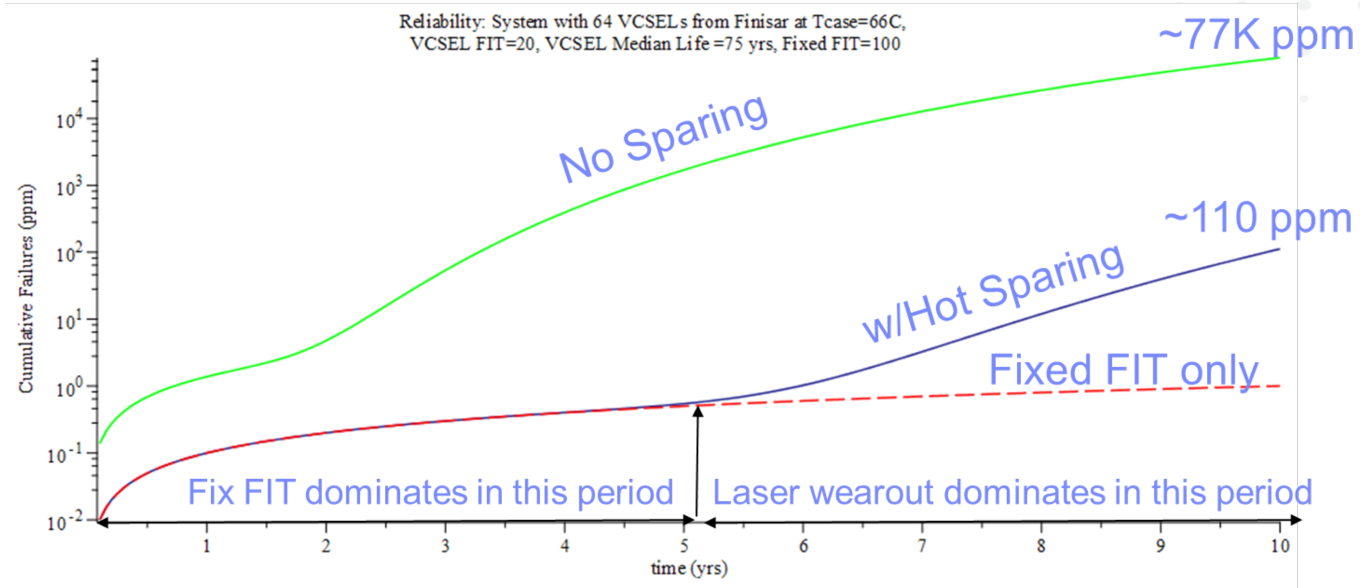


- Lack of field replacement drives stringent reliability requirements
- Laser wearout dominates: Sparing is desired

- MOTION has 2:1 laser redundancy on every channel
- Simulation shows ~1000x improvement in reliability at the end of 10 years of service →



Reliability Simulation for 4 CSOPs on MCC

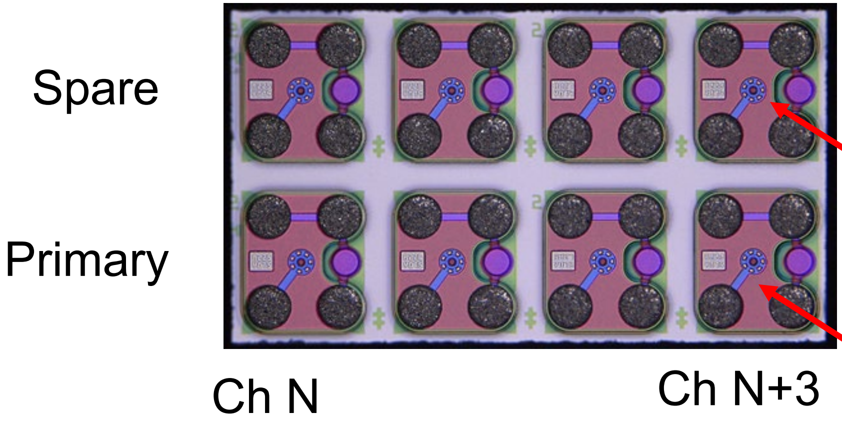


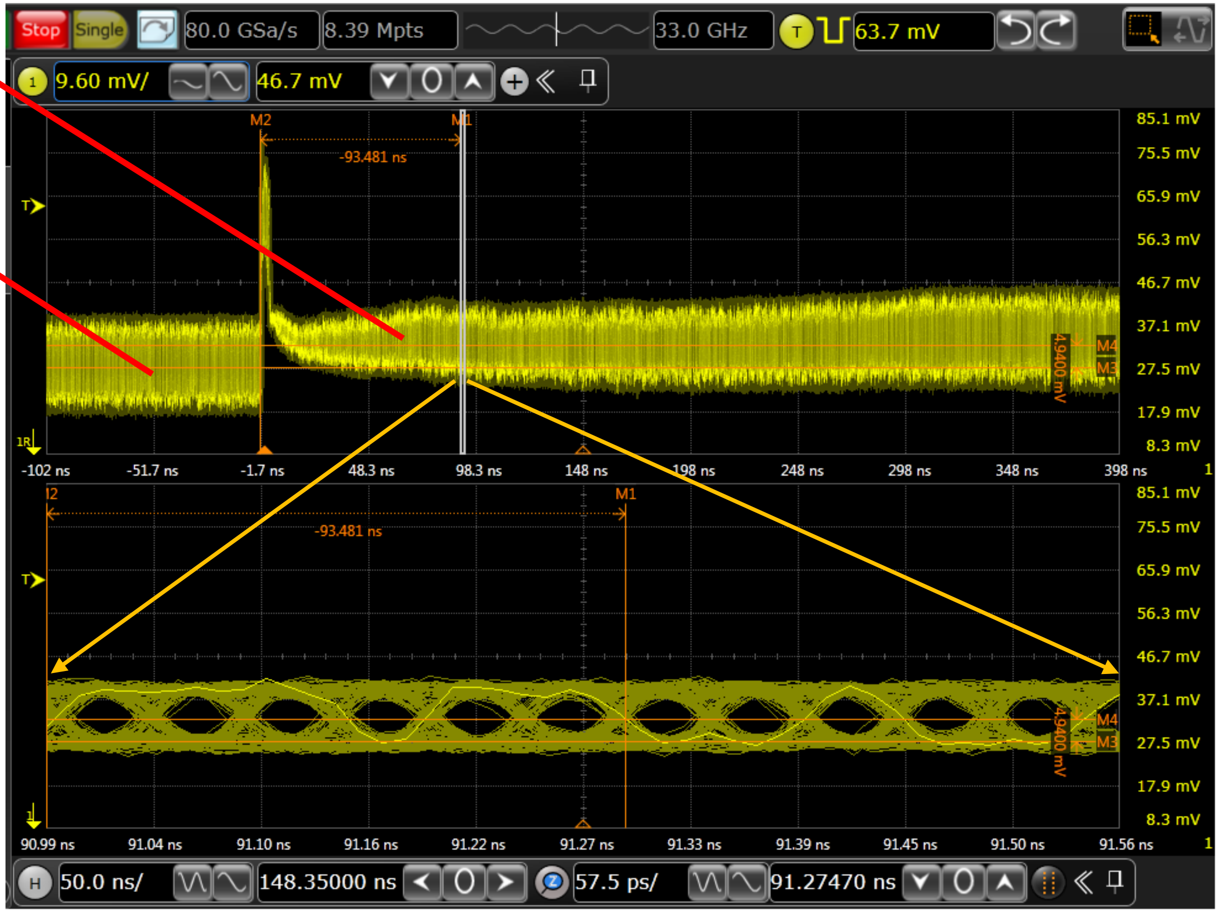
Fixed FIT: package components that can not be spared = Sum of everything with a non-zero FIT

Assumed Parameters:

VCSEL MTTF = 75 yrs, VCSEL FIT = 20
Fixed FIT = 50 FIT per module
Ibias = 9 mA

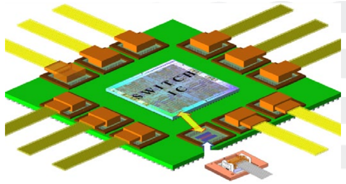
Fast Laser Sparing





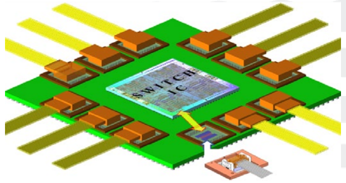
- Switching from Primary VCSEL to Spare is observed to take < 100ns
  - Orders of magnitude improvement over pluggables!
- Waveform shows the sum of both outputs and a brief 2ns period when the primary and spare VCSELs are both on.

MOTION Phase 2 Hardware Changes

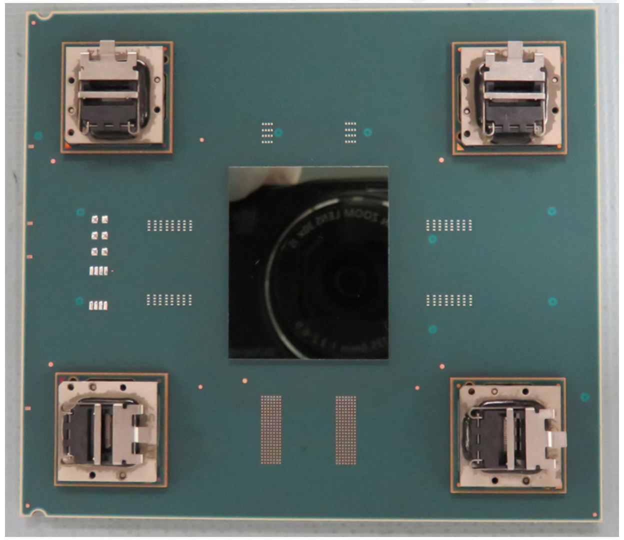


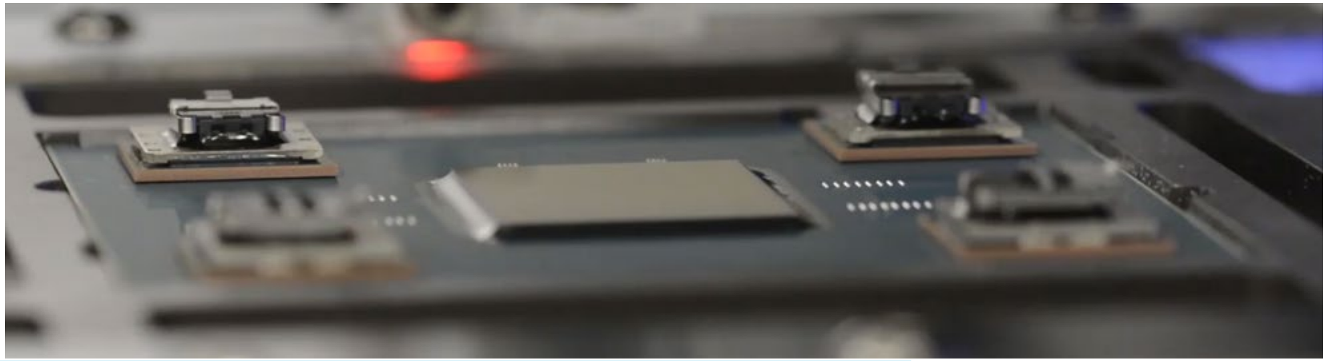
Parameter	Phase 1	Phase 2
Electrical Interface	16 channels @ 56G NRZ	32 channels @ 112G PAM4
IC Technology	SiGe	CMOS
Optical Interface	16 channels @ 56 G NRZ	32 channels @ 112G PAM4
# of Wavelengths	1	2
# of Fibers	16 Tx + 16 Rx	16 Tx + 16 Rx
Fiber Type	50/125 MMF	50/125 MMF
Package I/O Pitch	400um	300um
Glass Carrier Size	13x13mm	13x13mm
Energy consumption	4 pJ/bit	2 pJ/bit
Projected cost	25¢/Gig	TBD but <25¢/Gig
Laminate interface	Soldered or LGA	Soldered or LGA

Phase 2: IBM SYSTEMS TECHNOLOGY EVALUATION



- Goal: To assess the technology readiness of co-packaged optics
  - Optical transceivers soldered directly on the top surface of a laminate package will be built
- The IBM Systems group will perform this technology evaluation focusing on the thermal & mechanical robustness of packaging:
  - Four (4) optical transceivers and one (1) test site die on the top of a single FC-PLGA laminate, assembled with a thermal lid
- Evaluations:
  - Preconditioning
  - Deep Thermal Cycle
  - Accelerated Thermal Cycling
  - Thermal Aging
  - Temperature & Humidity
  - Temperature, Humidity & Bias
  - Power Cycling
  - Low Temperature Storage,
  - Shock & Vibe

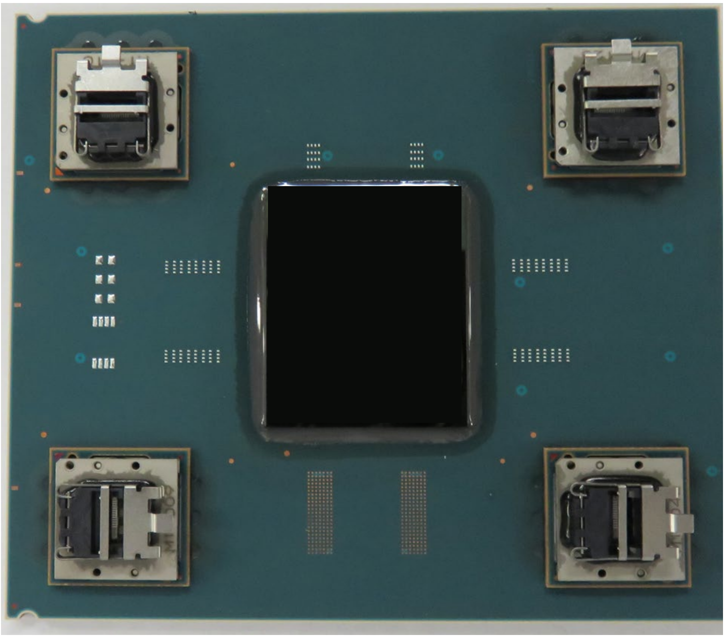




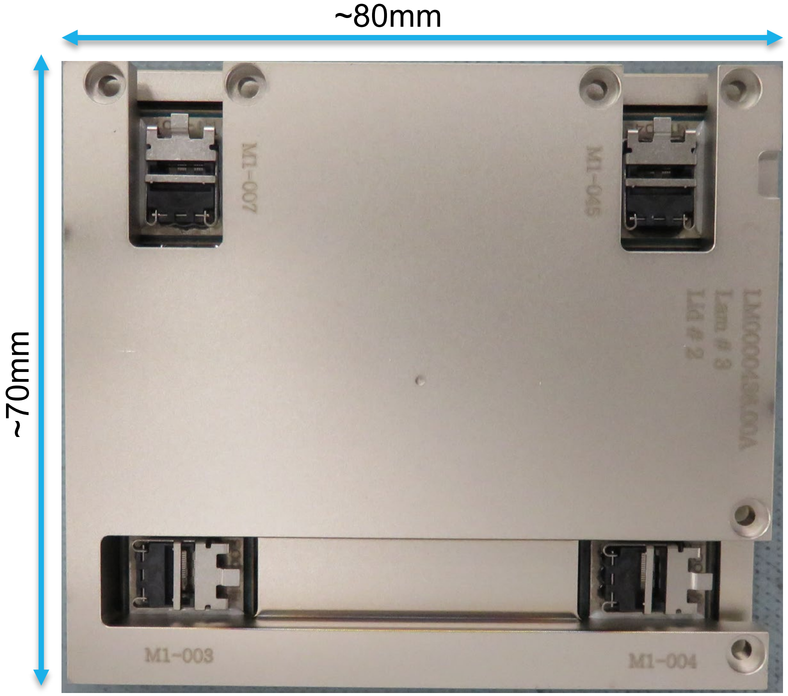
Demonstrating a viable path to system integration using established IBM Server Group processes

IBM Processor module with MOTION Devices

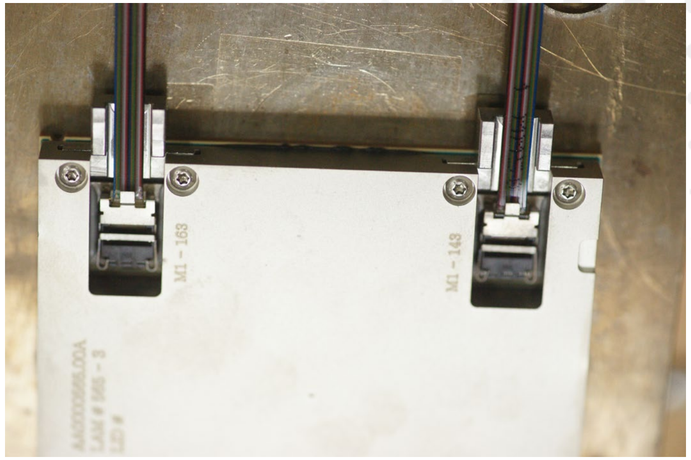
Phase 2 hardware

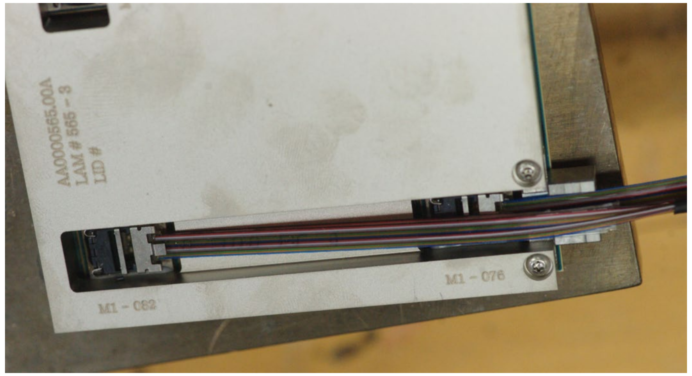


Processor Laminate with 4 MOTION devices, w/o heat spreader



Processor Laminate with 4 MOTION devices, w/ heat spreader

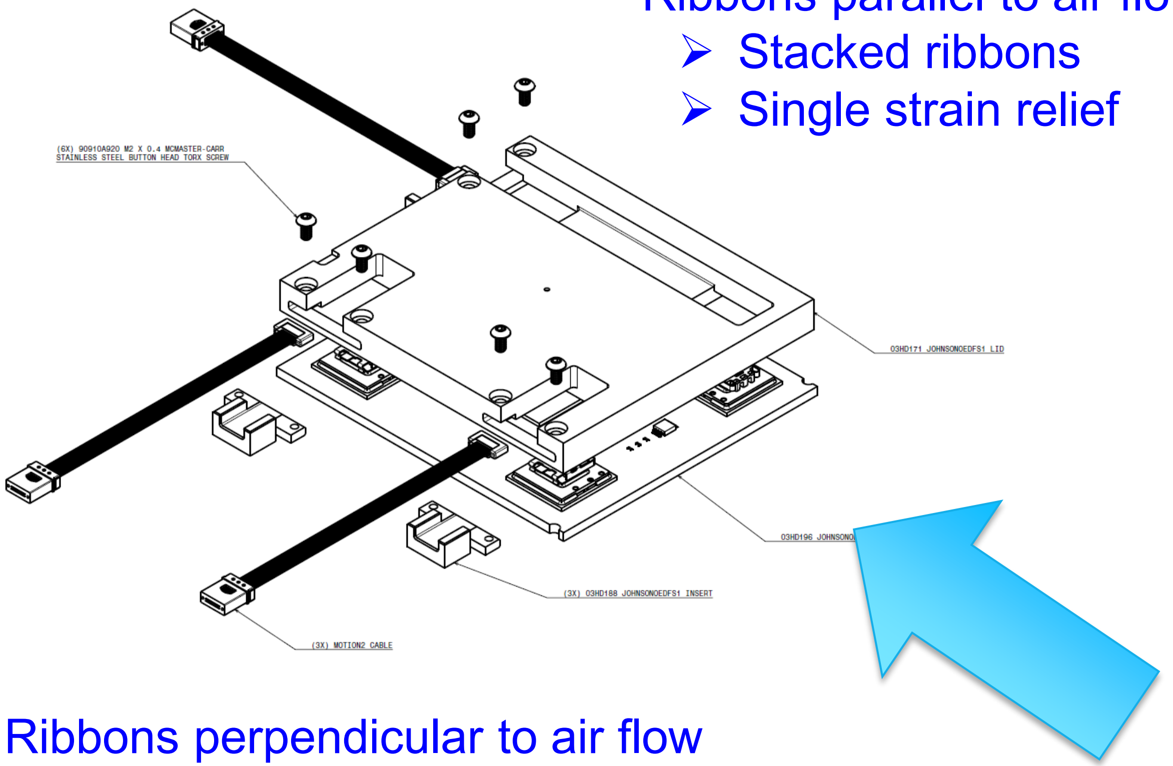




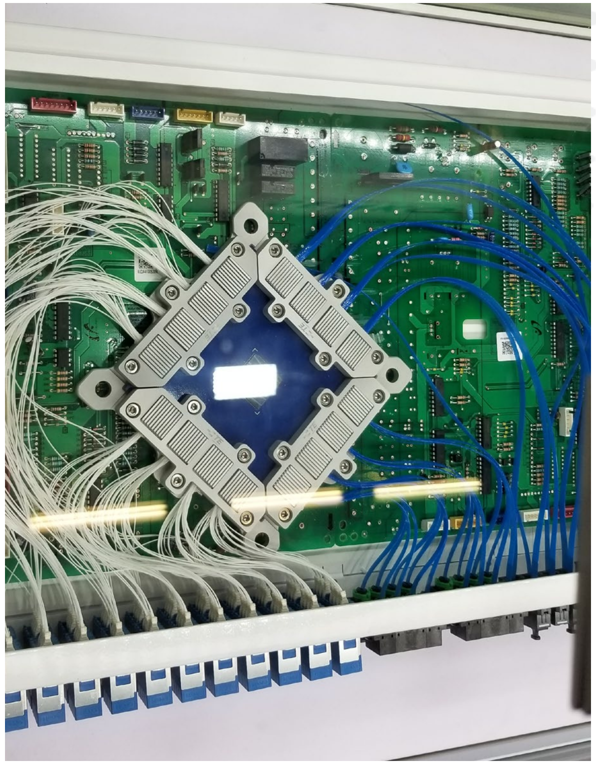
- Large IC and MOTION devices are assembled with a single solder reflow step.
- Two different orientations for fiber escape: Overlapping and Non-overlapping

Ribbon Fiber Orientation experiment

- Ribbons parallel to air flow
  - ➢ Stacked ribbons
  - ➢ Single strain relief

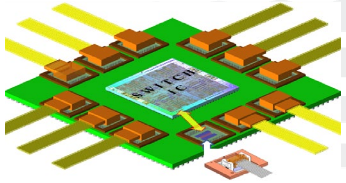

(6X) 90910A920 M2 X 0.4 MCMASTER-CARR STAINLESS STEEL BUTTON HEAD TORX SCREW
03HD171 JOHNSONOEDFS1 LID
03HD196 JOHNSONO
(3X) 03HD188 JOHNSONOEDFS1 INSERT
(3X) MOTION2 CABLE
Air Flow

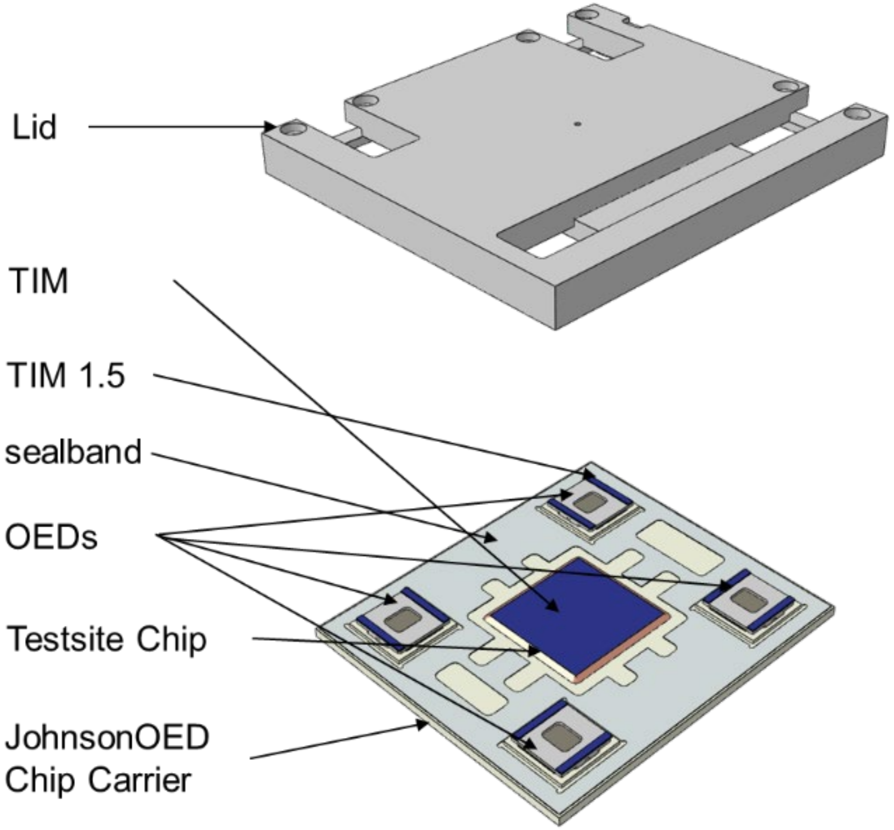
- Ribbons perpendicular to air flow
- ➢ Individual Strain Relief



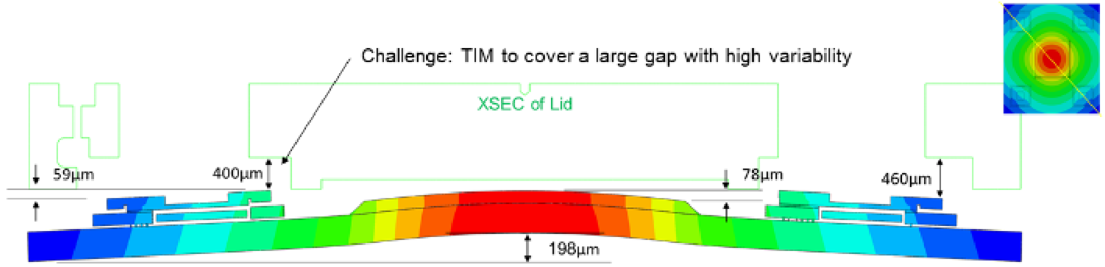
Data Center Switches do not have this option

Thermomechanical package simulations



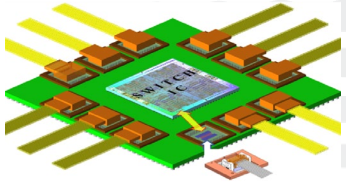


Components in Finite Element Analysis simulation



CTE mismatches between ASIC, laminate, and OED result in considerable warpage prior to the lid attach but the range is still acceptable

Summary of Reliability Testing



Thermal readout at different interval of stressing, the results are reasonably stable

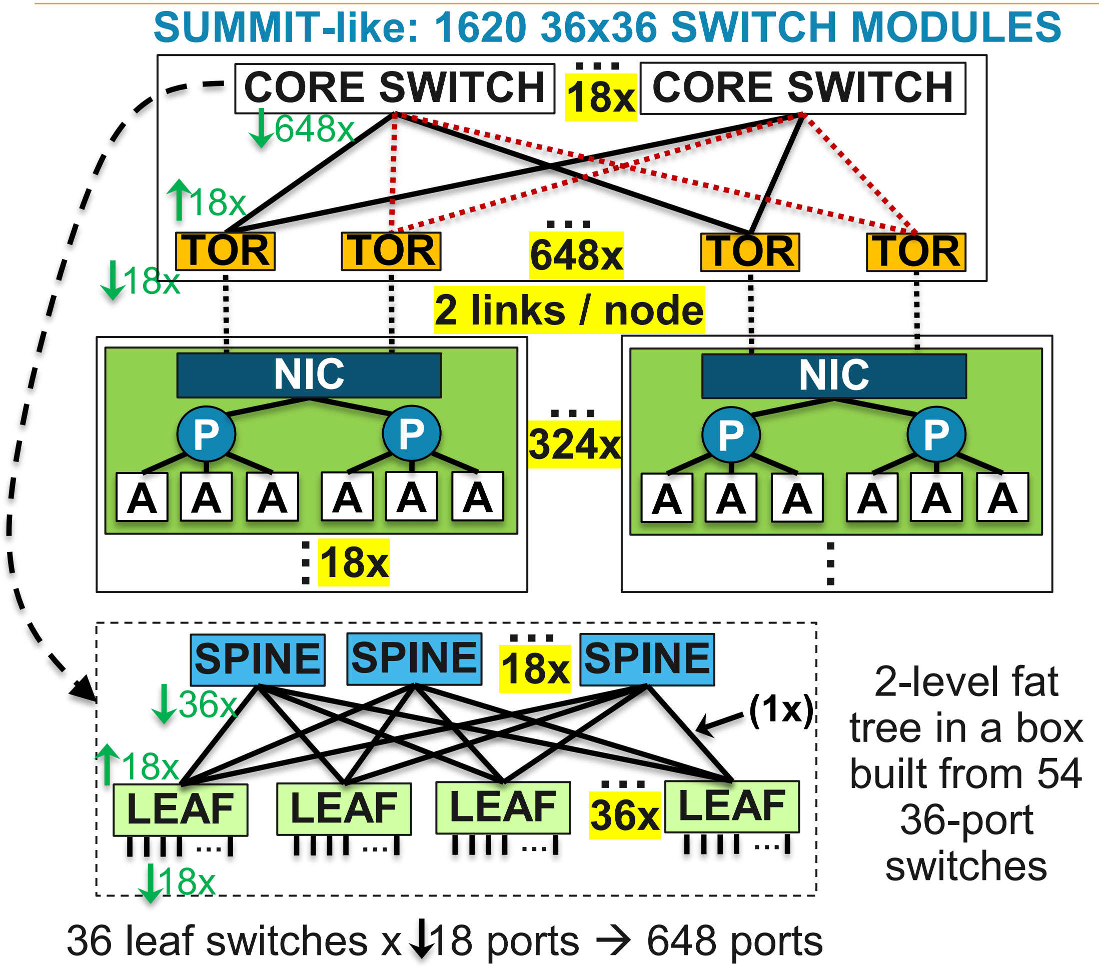
Stress	Samples	T/S	Time 0	Readout 1	Readout 2	Readout 3	Readout 4	Readout 5	Readout 6
		-40/60°C							
DTC -40/125°C	5	5X	T, R	250 c	500 c	750 c	1000 c	1250 c	1500 c
ATC 0/100°C	5	5X	T, R	500 c	1000 c	1500 c	2000 c	2500 c	3000 c
HTS 125°C	5	5X	T	494 hrs.	989 hrs.	1486 hrs.	2009 hrs.		
T&H 85°C/85% RH	5	5X	T	279 hrs.	438 hrs.	721 hrs.	1002 hrs.		

Laminate feature fail rate vs DTC cycles.

Laminate features							
C4 nets, LW/LS, RFPs, Die edge crack sensors, FC&BC side resin cracking sensor, Power and Ground							
	250	500	750	1000	1250	1500	1750
DTC	0/12	0/12	0/12	0/12	3/10	3/10	
DTC on card	0/5	0/5	0/5	0/5	0/5	0/5	3/5

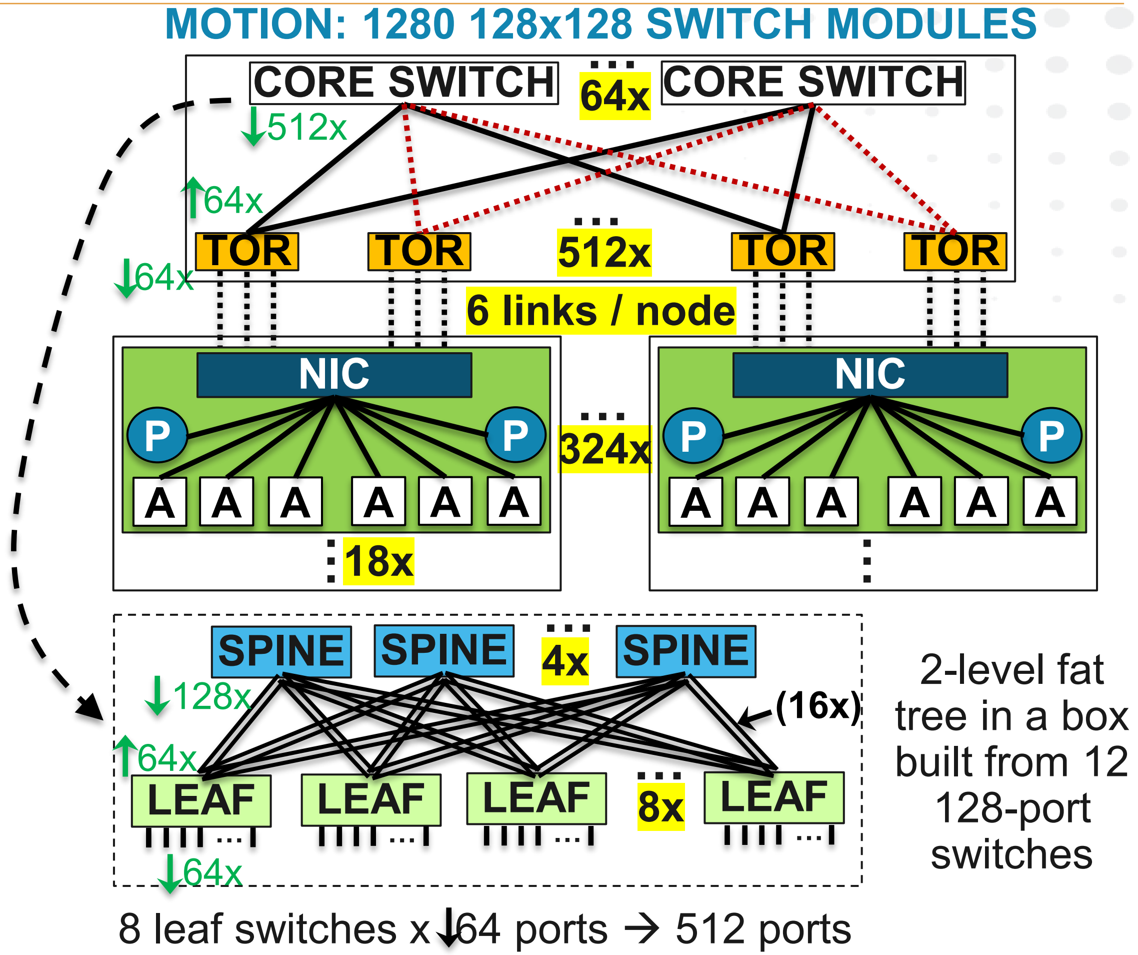
The first fail occurs after 1000 cycles of DTC which is not a reliability concern (DTC is a highly accelerated stress as opposed to field application).

The benefits of higher-radix switches enabled by MOTION



2-level fat tree in a box built from 54 36-port switches

36 leaf switches x ↓18 ports → 648 ports



2-level fat tree in a box built from 12 128-port switches

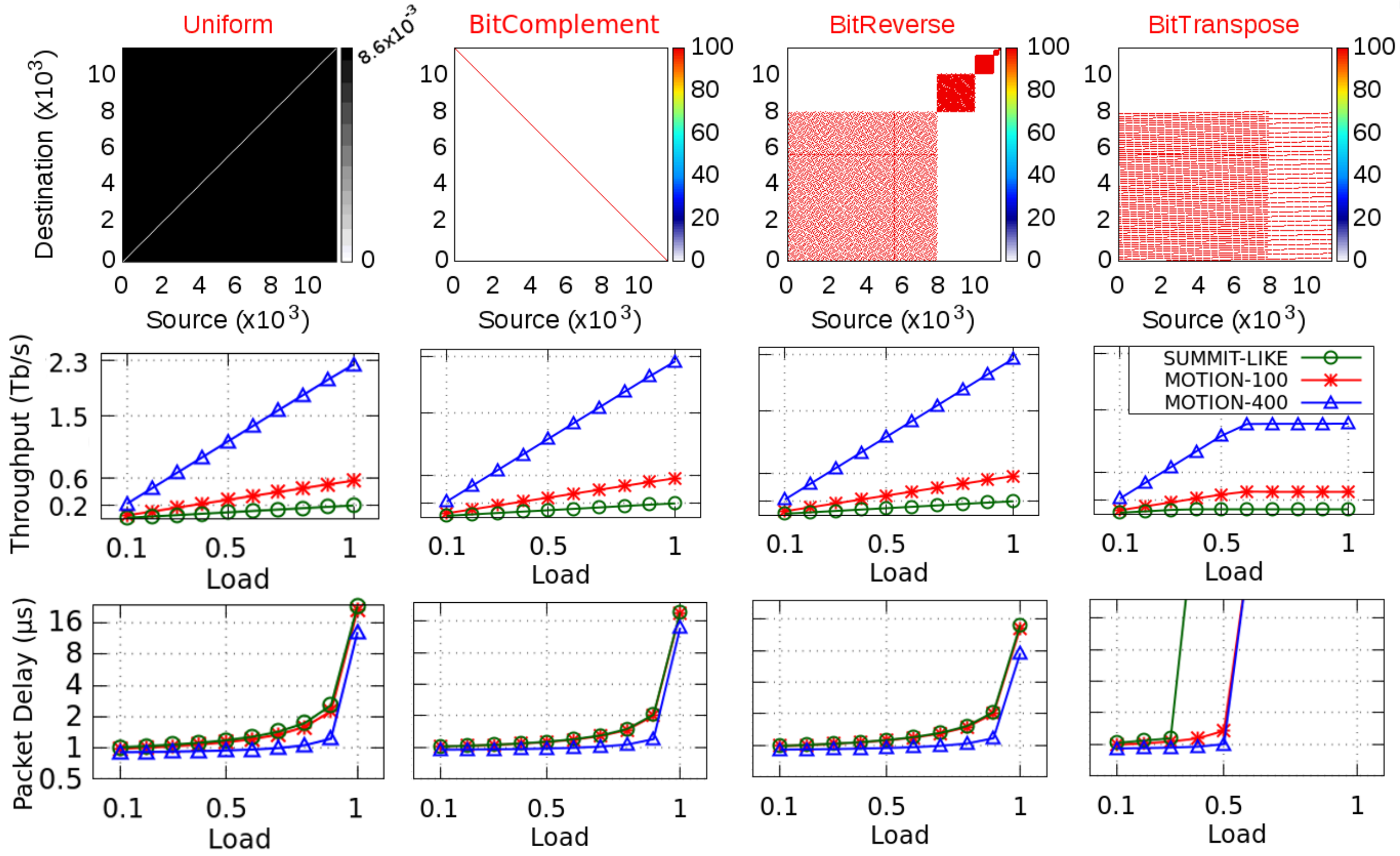
8 leaf switches x ↓64 ports → 512 ports

- 3x more network end points for 21% fewer switch modules
- 2.8x higher bisection BW for 100 Gb/s per port (11.2x for 400 Gb/s)
- MOTION opens the way to direct-network-attached accelerators

Network performance analysis: MOTION vs. Summit

Venus discrete event network simulator

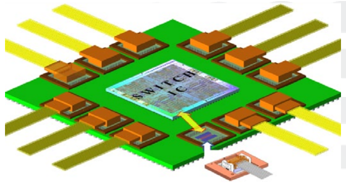
Traffic and Network	
Generator	
RX buffer size	Infinite
Message size	1024 B
Generation distribution	Bernoulli
Data rate	100/400 Gbps
Load	[0.1-1]
Adapter/Switch	
Type	InfiniBand
Data rate per link	100/400 Gbps
Delay	100 ns
Switch Buffer per port	128 KB
Packet size	1024 B
Routing algorithm	Random



- *Uniform, BitComplement, BitReverse*: Linear throughput increase for all systems → 2.8x and 11.2x higher throughput for 100 and 400 Gb/s data rates
- *BitTranspose*: earlier saturation due to the significantly fewer destination nodes → 4.3x and 17.2x higher throughput for 100 and 400 Gb/s data rates
- MOTION-400: best mean packet delay for all cases

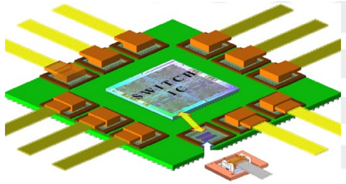
arpa·e
CHANGING WHAT'S POSSIBLE

Challenges for Co-packaged Optics



- Reliability
- Field Replacement/Serviceability
  - Fail-in-place Strategy?
- Yield
  - Who is responsibility for final yield?
- Assembly
  - Who does what and when?
- Standards and/or MSAs
  - Minimum Time for Standards seems to be > 2 years
  - Proprietary solutions likely to emerge first
- Compatible Technologies
  - MMF or SMF
- Field Upgradeable Firmware!

Acknowledgements: MOTION Phase 1 & 2 Team members & Sponsor



- IBM Research
  - C. Baks, A. Benner, R. Budd, T. Dickson, F. Doany, W. Lee, M. Meghelli, P. Pepeljugoski, J. Proesel, M. Taubenblatt, L. Schares, M. Schultz, P. Maniotis, P. Stark, H. Ainspan, Z. Toprak Deniz, S. Dhawan, T. Dickson, N. Dupuis, P. Francese, B. Sadhu, M. Kossel, T. Morf, M. Brändli, S. Rylov, M. Cochet, C. Ozdag, A. Watanabe, H. Rahmani, D. Kuchta,
- IBM Bromont
  - L-M. Achard, P. Fortier, C. Dufort, E. Tucotte, C. Bureau, M. Pion, Y. Cossette, P. Ducharme, S. Desputeau, A. Janta-Polczynski, P. Minier, G. Jutras, P. McInnes, S. Whitehead, B. Sow
- IBM Server Packaging

  B. Parikh, S. Ostrander, S. Li, C. Setzer, H. Toy, J. Ross, K. Lange, M. Kapfhammer, B. Meiswinkel, C. Muzzy, E. Steiner, D. Smith, T. Saunders, G. Pomerantz, J. Coffin, K. Marston, K. Smith, T. Ahmed, L. Rapp, Y.Yao, T. Wassick, M. Warbrick, T. Lombardi, B. Singh, C. Walker, S. Iruvanti, D. Yannitty, P. Ramaglia, T. Weiss, M. Interrante, J. Sorbello, C. Arvin, M. Stalter, A. Perez, P. Torbet, T. Olowofela, J. Bunt, M. Fisher, T. Childress, J. Mingo, C. Savoy, S. Ruiz, D.Kohler, R. Seifts, R. Olson, H. Polgrean, J. Rowland, C. Thomas, E. Kastberg, A. Schetter, D. Babcock, A. Greenberg, D. Lord, R. Rodriguez, C. Taylor
- IBM Server Development
  - D. Becker, R. Laning, D. Dreps, M. Hoffmeyer, J. Eagle, F. Gholami, K. O'Connell, S. Canfield, S. Chun, R. Frota
- IBM Supply Chain Engineering
  - H Bagheri, K. Akasofu, C. Grosskopf, A. Tiano, T. Sass, E. Mallery
- II-VI Finisar Corporation

  F. Flens, D. Case, P. Chen, J. Glover, C. Kocot, K. Koski, G. Light, T. Nguyen, S. Pandy, S Quadery, K. Szczerba, B. Wang, P Westbergh, S. Pandey, H. Hayashigatani (131)
- Texas A&M University
  - N. Kim, A. Kumar, T. Liu, I. Yi, S. Palermo

Acknowledgment: “The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000846. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.”