BROADCOM®

An AI Compute ASIC with Optical Attach to Enable Next Generation Scale-Up Architectures

Manish Mehta
VP Marketing and Operations – Optical Systems Division

August 26th, 2024

Broadcom Co-Packaged Optics (CPO) Background

Our objective

- Build an optical interconnect that provides substantial improvements over current optical modules on cost, power consumption reliability and latency
- Build a high-density optical interconnect that enables up to 1 Tb/s/mm duplex connectivity to support current gen and next gen scale-up and scale-out optical BW density

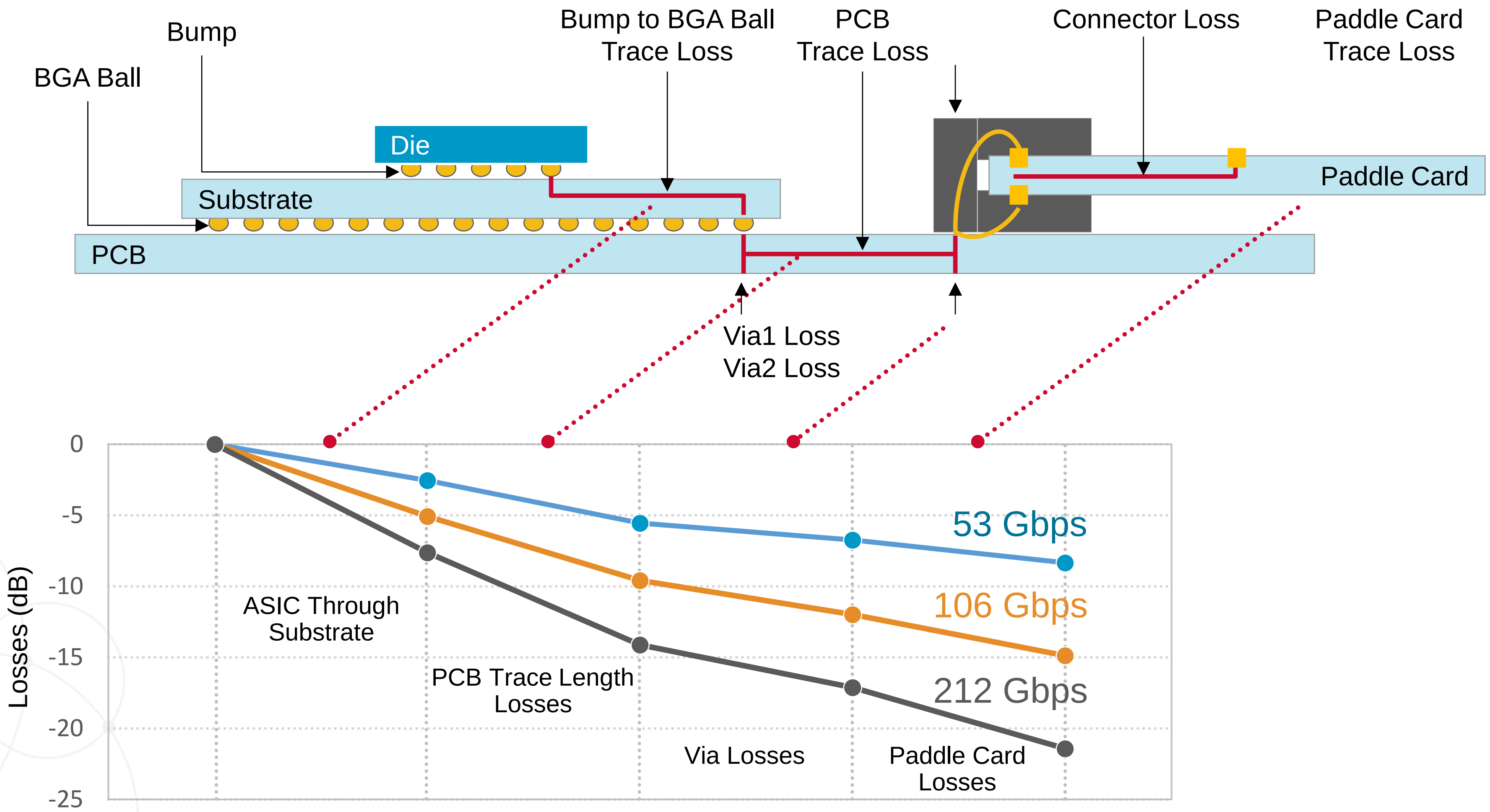
Our Progress

- Shipped first generation proof-of-concept 25.6T CPO switch product
- Currently qualifying and ramping second generation 51.2T CPO switch product
- Demonstrated AI ASIC + CPO functionality: CPO + 2.5D Packaging

What I will share today

- Our journey to build the first Co-Packaged Optics (CPO) deployed in datacenter Front-End and Back-End networks using high density Silicon Photonics (SiPh)
- Implication of CPO Platform on both scale-out and scale-up connectivity

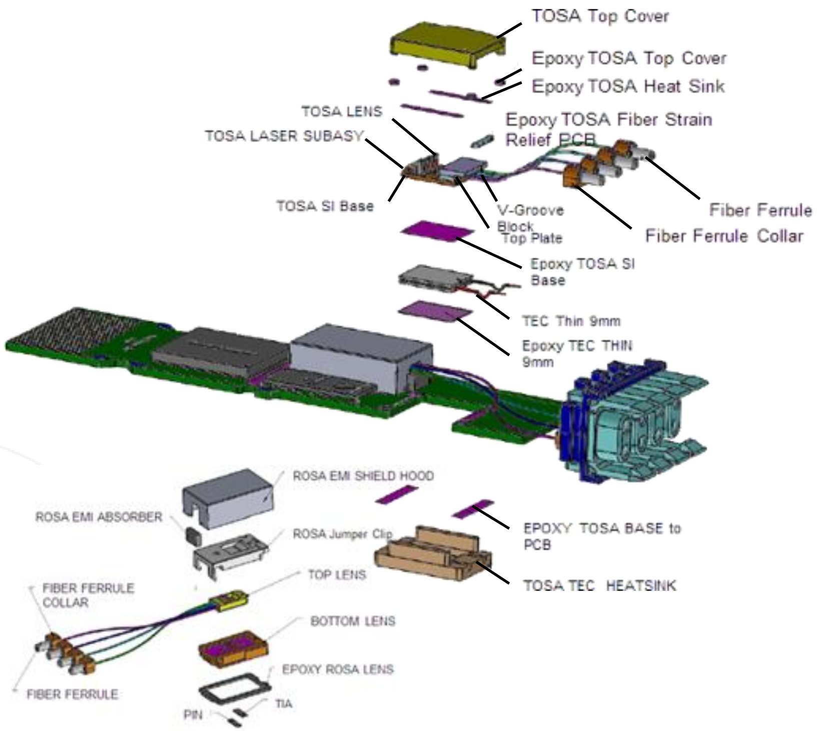
Serdes Migration to 200Gbps Limits Electrical I/O Reach



Mandates Development of Optical Interconnects Co-Packaged with ASIC

Journey to an AI ASIC with CPO: Discrete III-V to SiPh

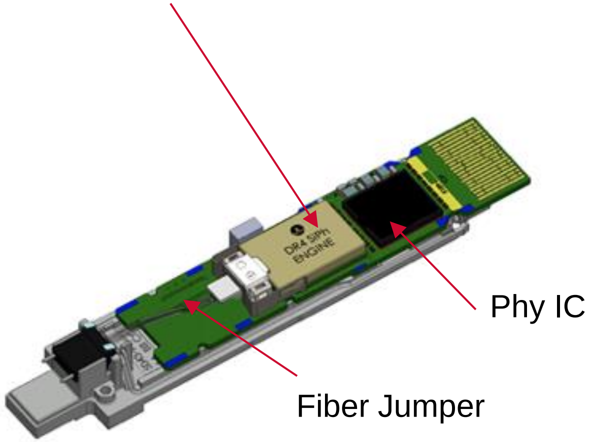
Conventional Module Design



Engineering and manufacturing limits to scale

Modules with Silicon Photonics

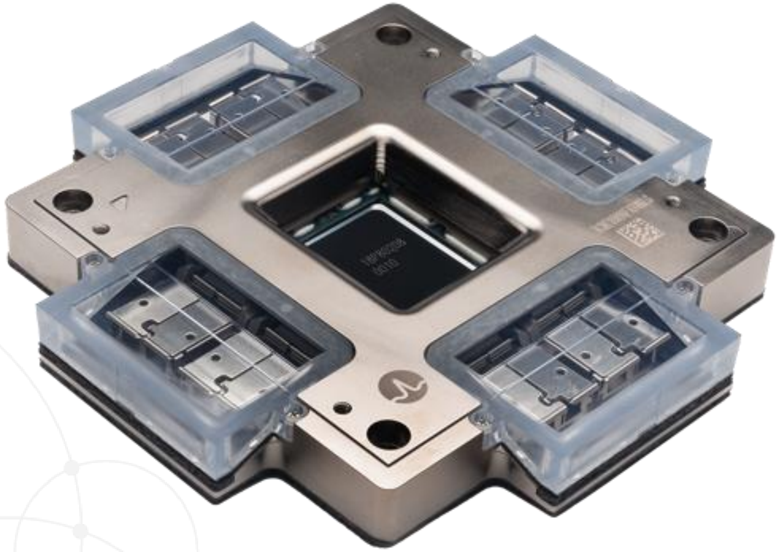
SiPh Chiplets in Module



Module Integration = First step to improved scale

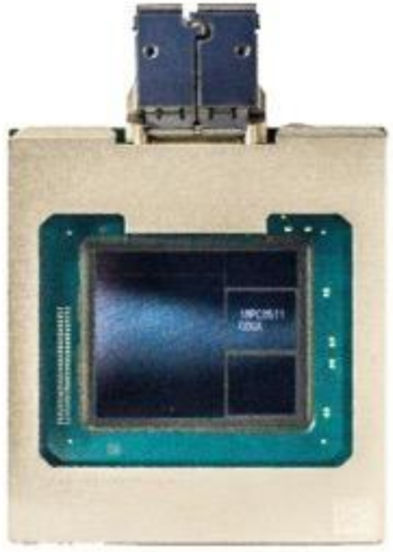
Journey to an AI ASIC with CPO: Switch CPO to GPU CPO

CPO for Scale-Out Networking



Greater than 50Tbps of Optics Attached to Switch ASIC

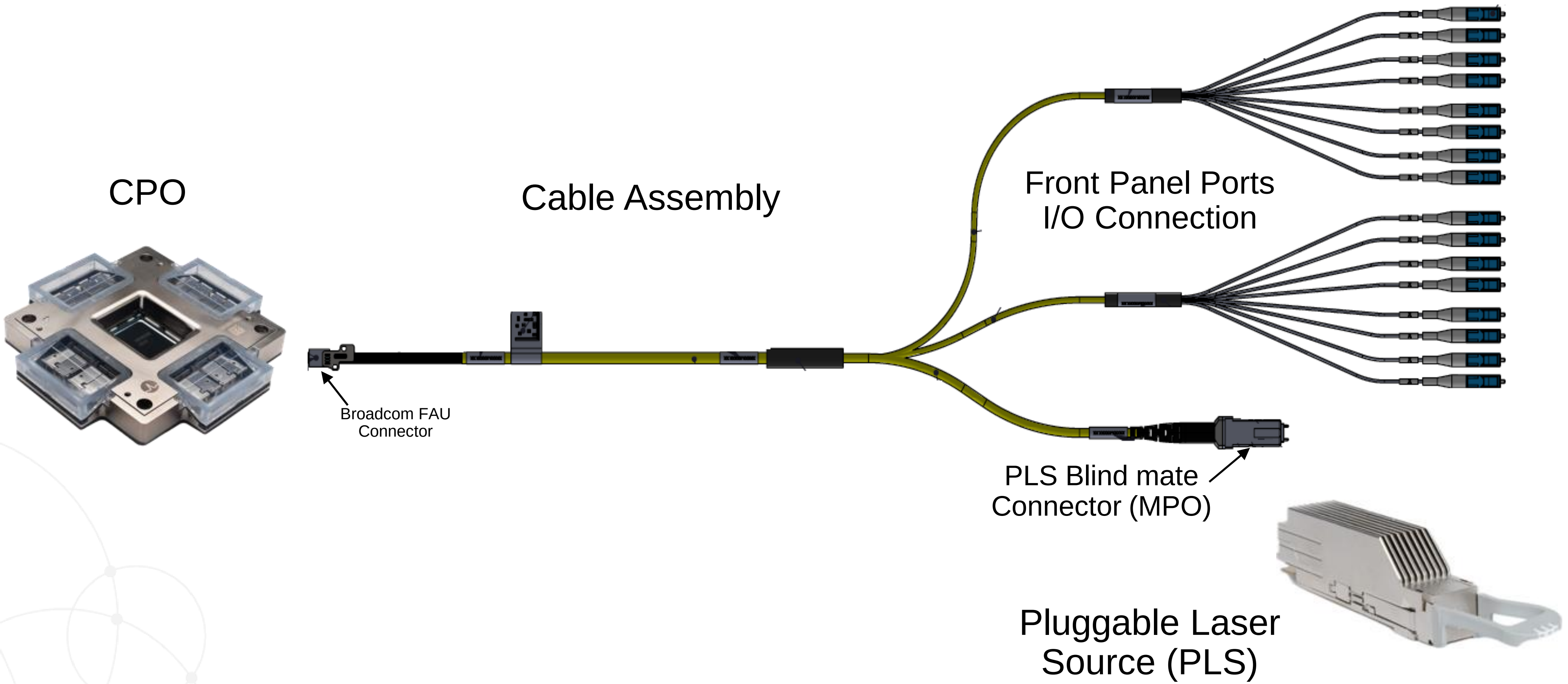
CPO for Scale-Up Compute



Greater than 6.4Tbps of Optics Attached to GPU

What is our CPO Platform

CPO Schematic



CPO

Cable Assembly

Front Panel Ports
I/O Connection

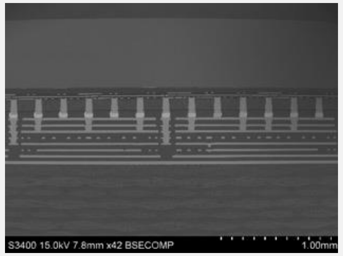
Broadcom FAU
Connector

PLS Blind mate
Connector (MPO)

Pluggable Laser
Source (PLS)

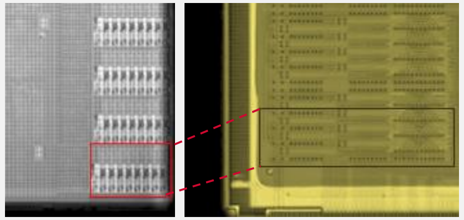
- 51.2Tbps TH5 Switch CPO with 8x6.4T optical engines
- Pluggable Laser Modules x 16 (field serviceable)
- Fiber Cable Assembly

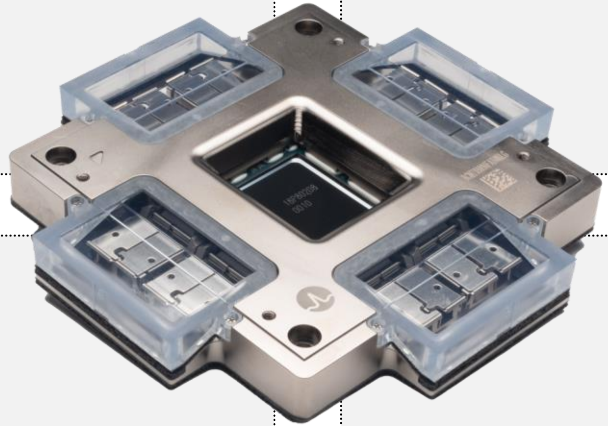
Key Components of CPO

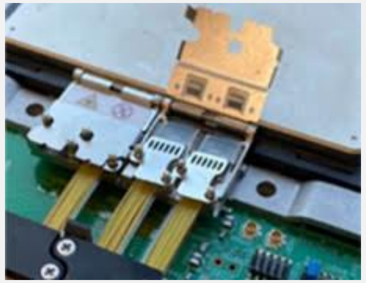


Advanced Packaging

Electrical Integrated Circuit (EIC) with driver and TIA

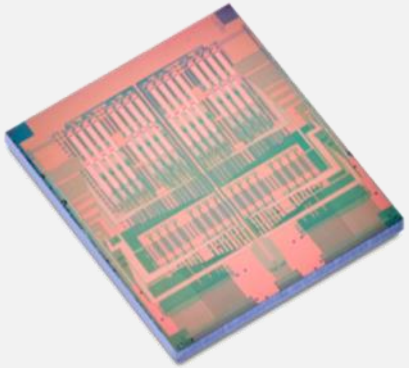






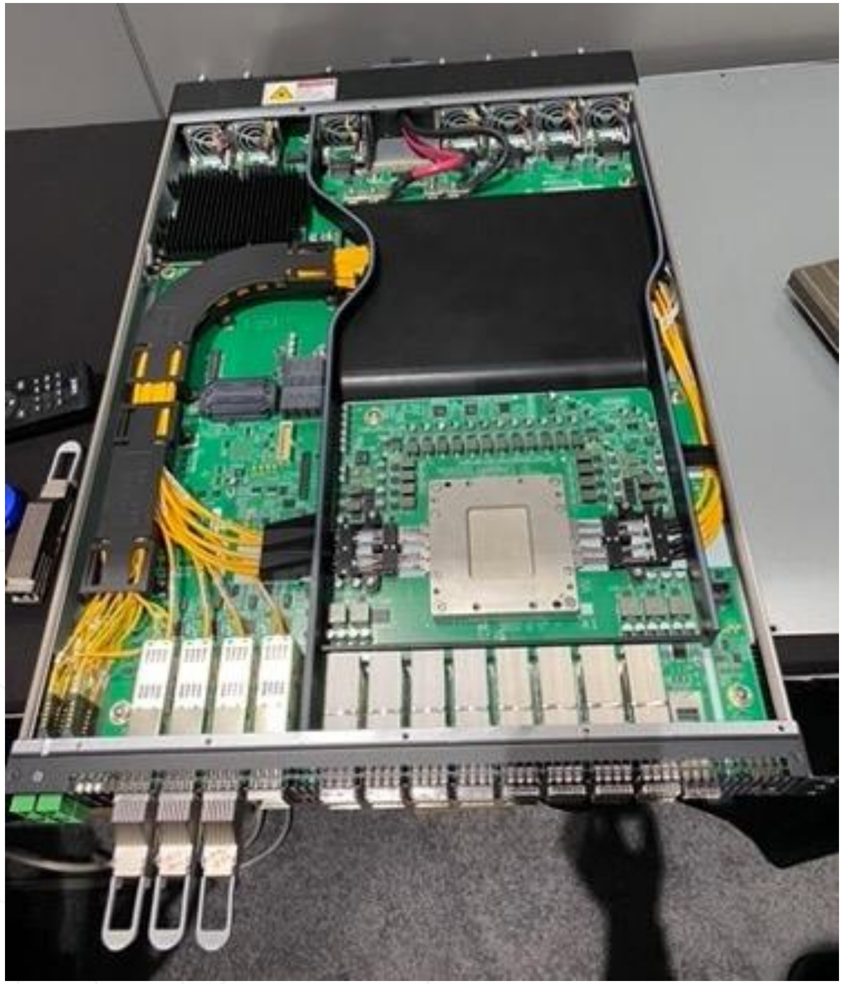
High density fiber connector

Photonic Integrated Circuit (PIC) with modulator and photo diodes



Scale-Out Networking using CPO

TH4-Humboldt: First Generation System

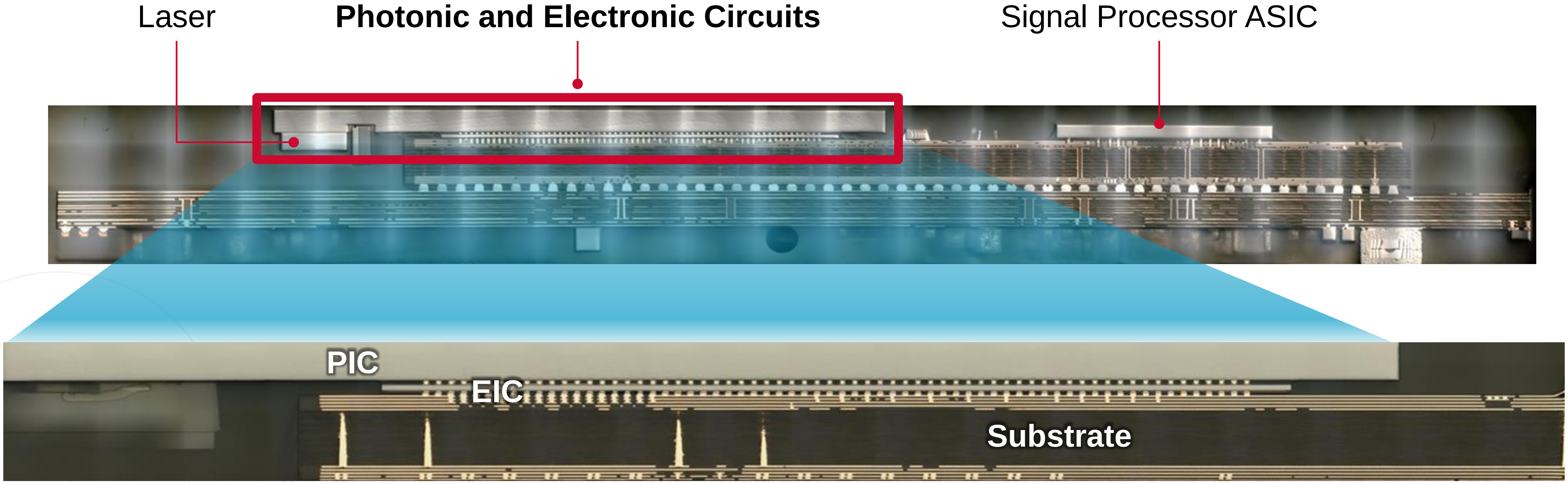


Product Features:

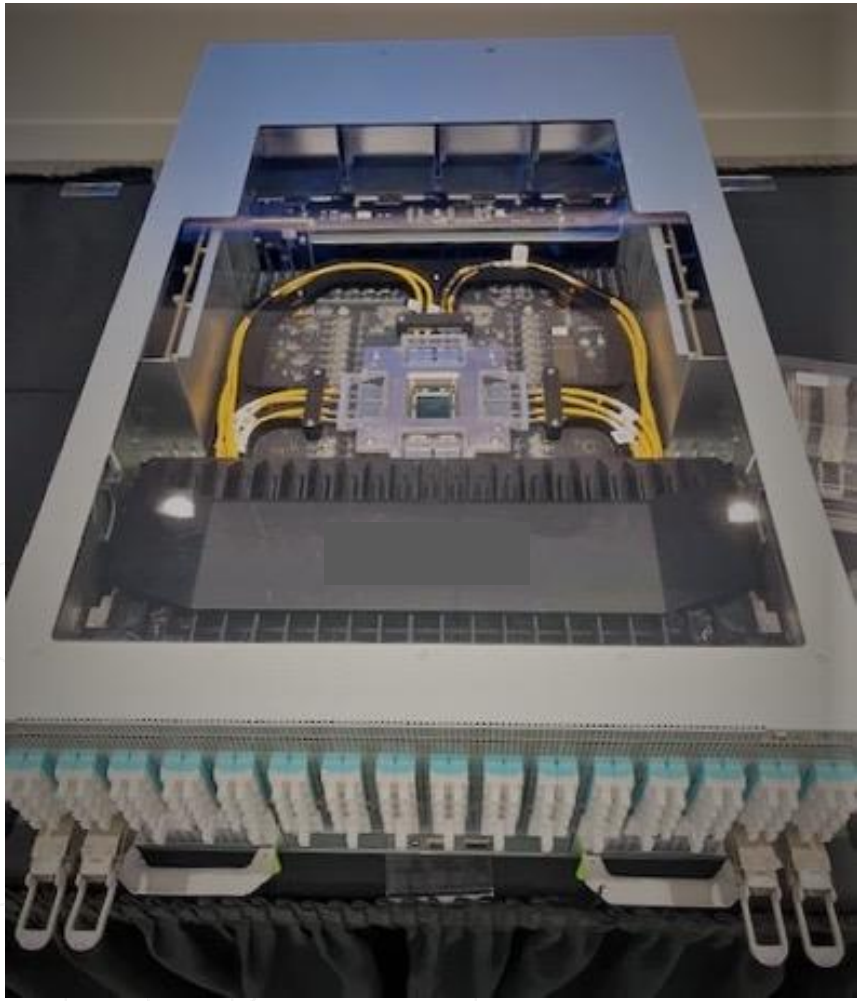
- 25.6T Ethernet Switch
- Half CPO, Half Electrical connectivity
- Four 3.2T optical engines (32x100Gbps DR connectivity)
- Optical engine is a PIC bonded to a SiGe EIC
- Each optical engine has ~ 250 optical components

SiGe dissipates additional 3 pJ/bit power consumption compared to CMOS solutions

TH4-Humboldt: SiPh PIC + SiGe EIC + TSV



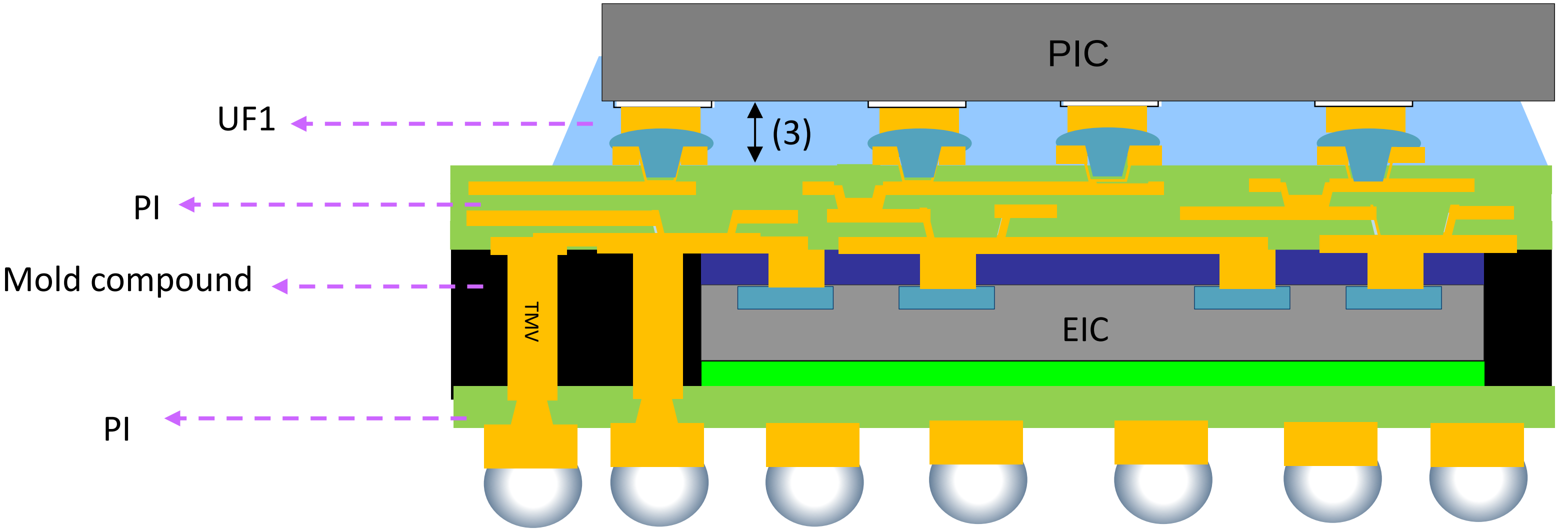
TH5-Bailly: Second Generation System



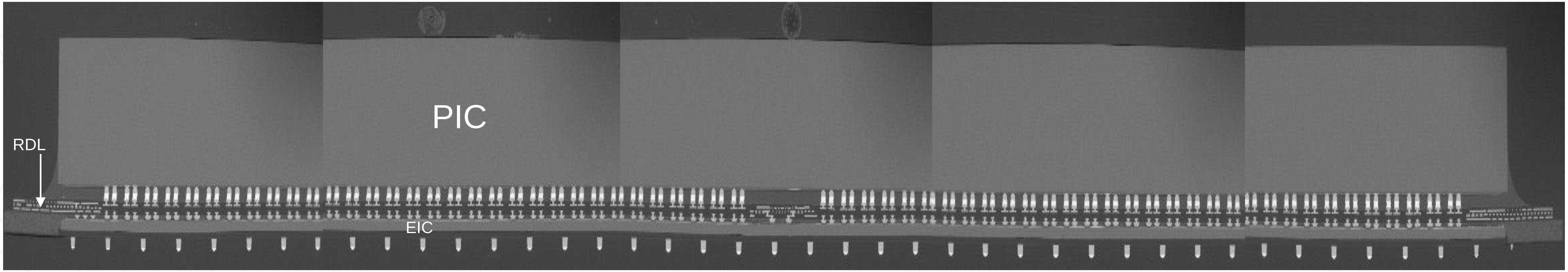
Product Features:

- 51.2T Ethernet Switch
- All Optical CPO connectivity
- Eight 6.4T optical engines (64x100Gbps FR4 connectivity)
- Optical engine is a PIC bonded to a CMOS EIC
- Each optical engine has ~ 1000 optical components

TH5-Bailly: SiPh PIC + 7nm CMOS EIC + FOWLP

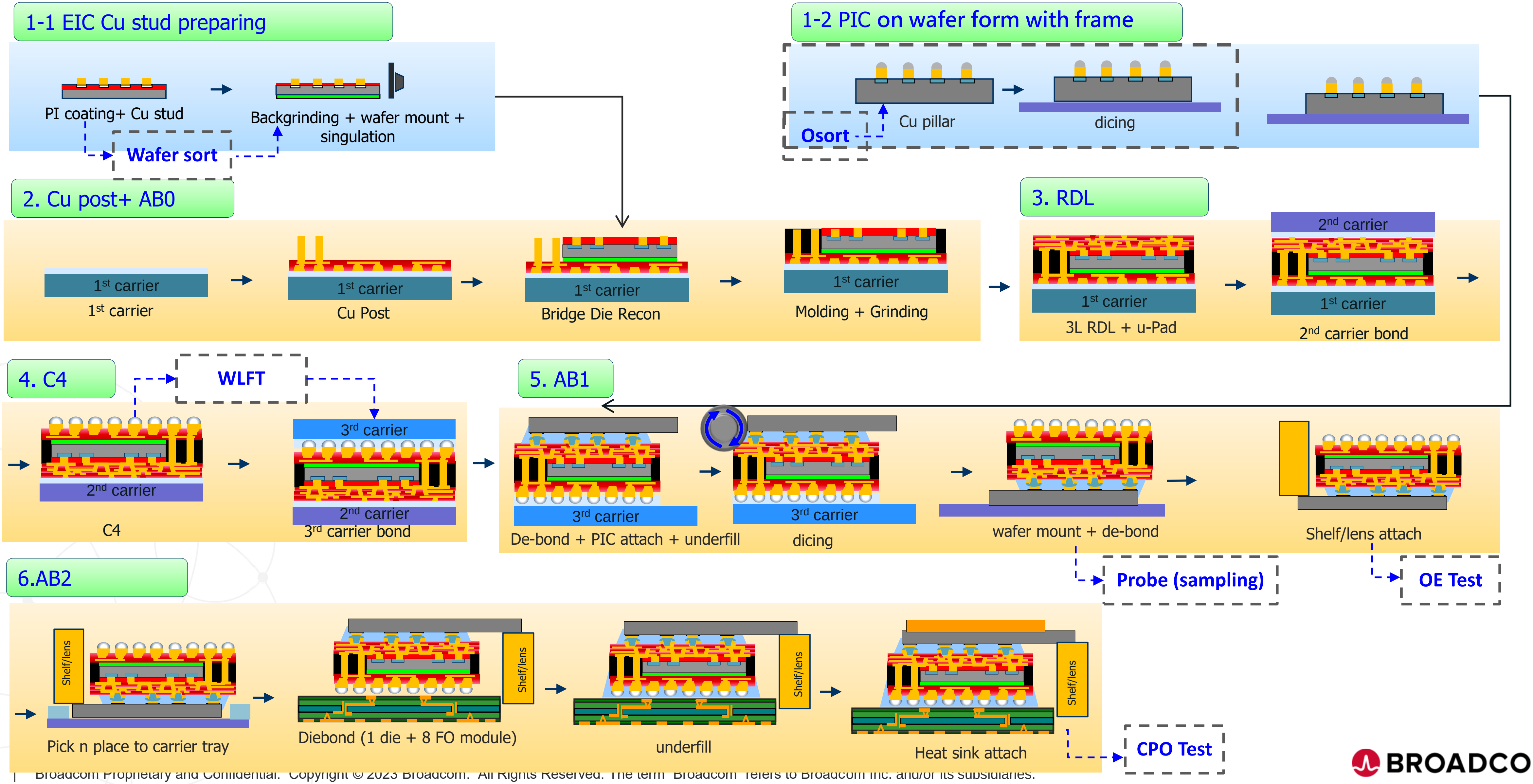


FOWLP: Fan-out Wafer-level Packaging



FOWLP Improved Scalability of PIC to EIC bonding

FOWLP Innovation: Dual-Side Attach for Co-Packaged Optics



1-1 EIC Cu stud preparing

PI coating+ Cu stud → Backgrinding + wafer mount + singulation

Wafer sort

1-2 PIC on wafer form with frame

Cu pillar → dicing

Osort

2. Cu post+ AB0

1st carrier → Cu Post → Bridge Die Recon → Molding + Grinding

3. RDL

3L RDL + u-Pad → 2nd carrier bond

4. C4

WLFT

C4 → 3rd carrier bond

5. AB1

De-bond + PIC attach + underfill → dicing → wafer mount + de-bond → Shelf/lens attach

Probe (sampling)

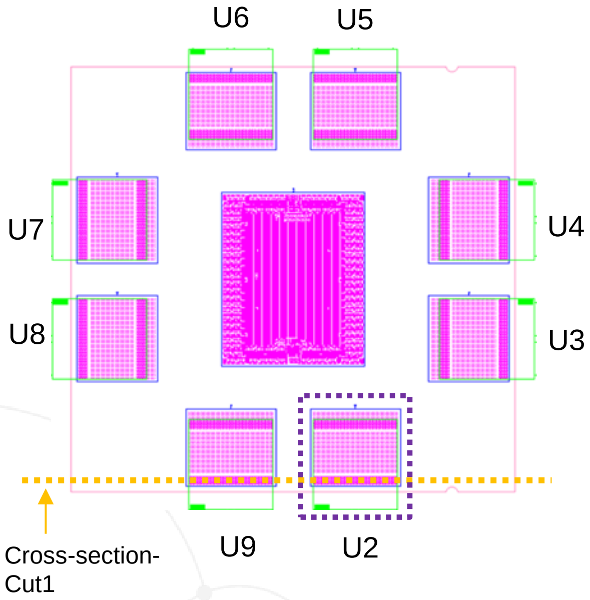
OE Test

6.AB2

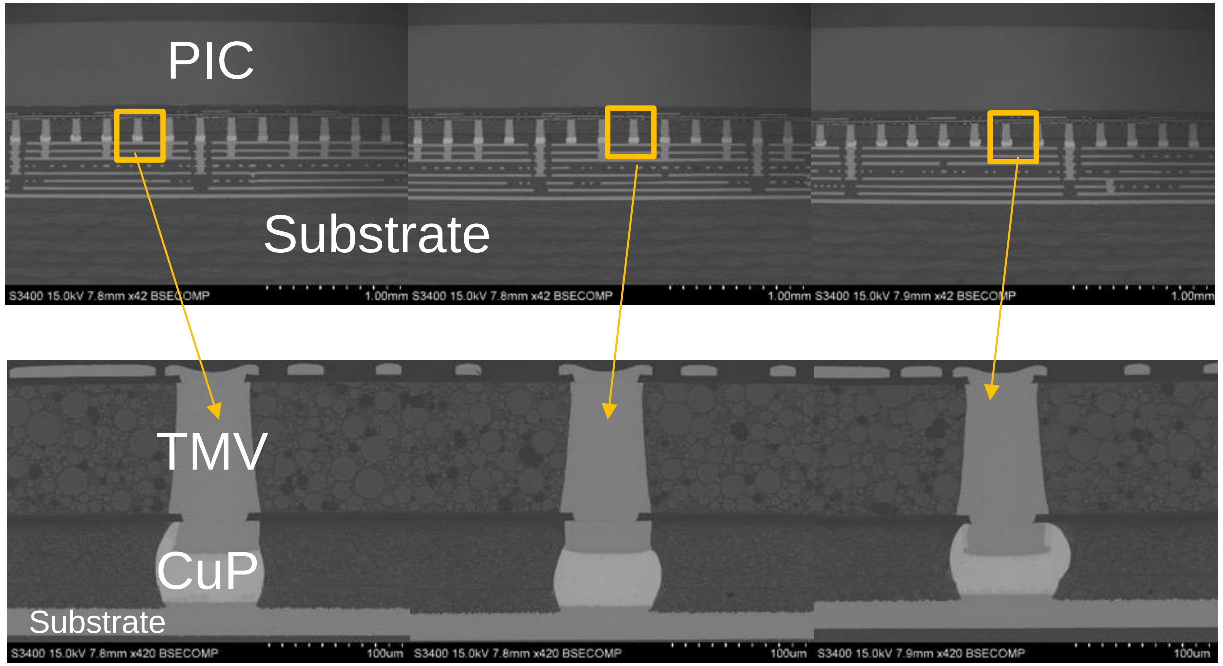
Pick n place to carrier tray → Diebond (1 die + 8 FO module) → underfill → Heat sink attach

CPO Test

Optical Engine Cross Sectional Images



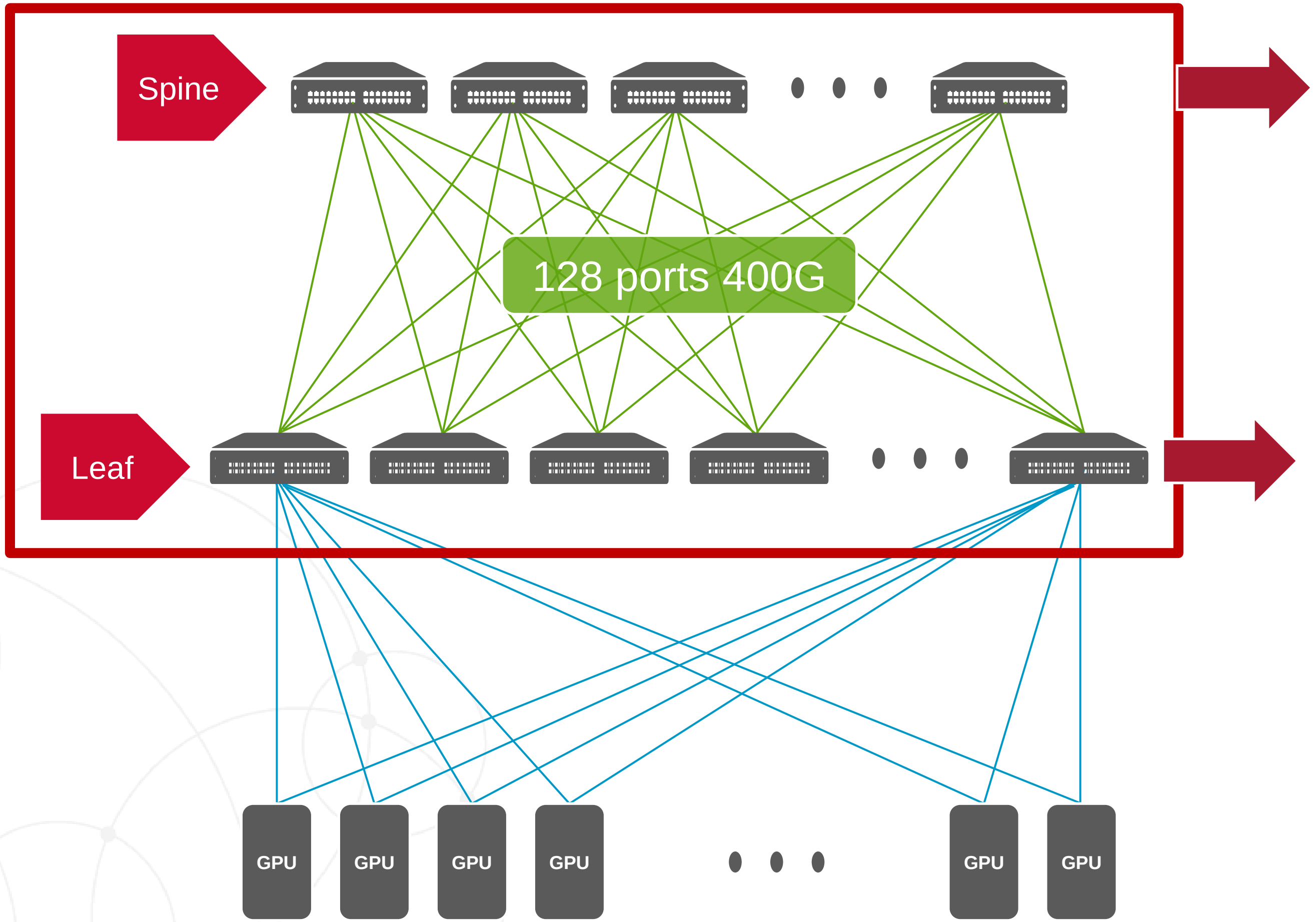
□ **Cross-section on U2**


PIC
Substrate
TMV
CuP
Substrate
S3400 15.0kV 7.8mm x42 BSECOMP 1.00mm
S3400 15.0kV 7.8mm x42 BSECOMP 1.00mm
S3400 15.0kV 7.9mm x42 BSECOMP 1.00mm
S3400 15.0kV 7.8mm x420 BSECOMP 100um
S3400 15.0kV 7.8mm x420 BSECOMP 100um
S3400 15.0kV 7.9mm x420 BSECOMP 100um

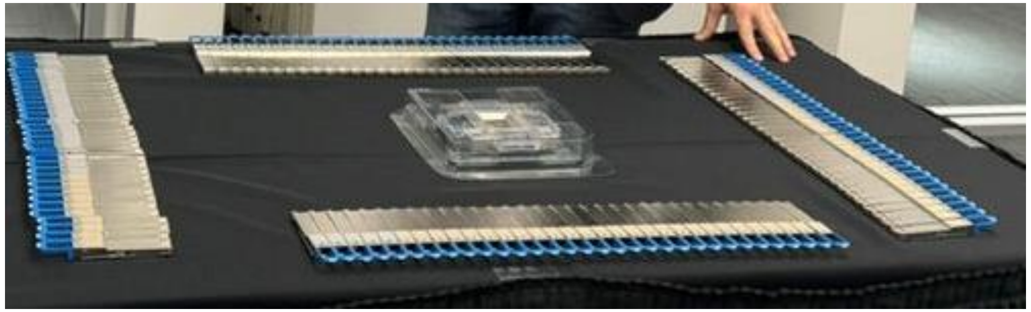
Cross-sections taken AFTER eight (8) optical engines bonded to substrate

Network Placement and Module Substitution

Scale-Out Network



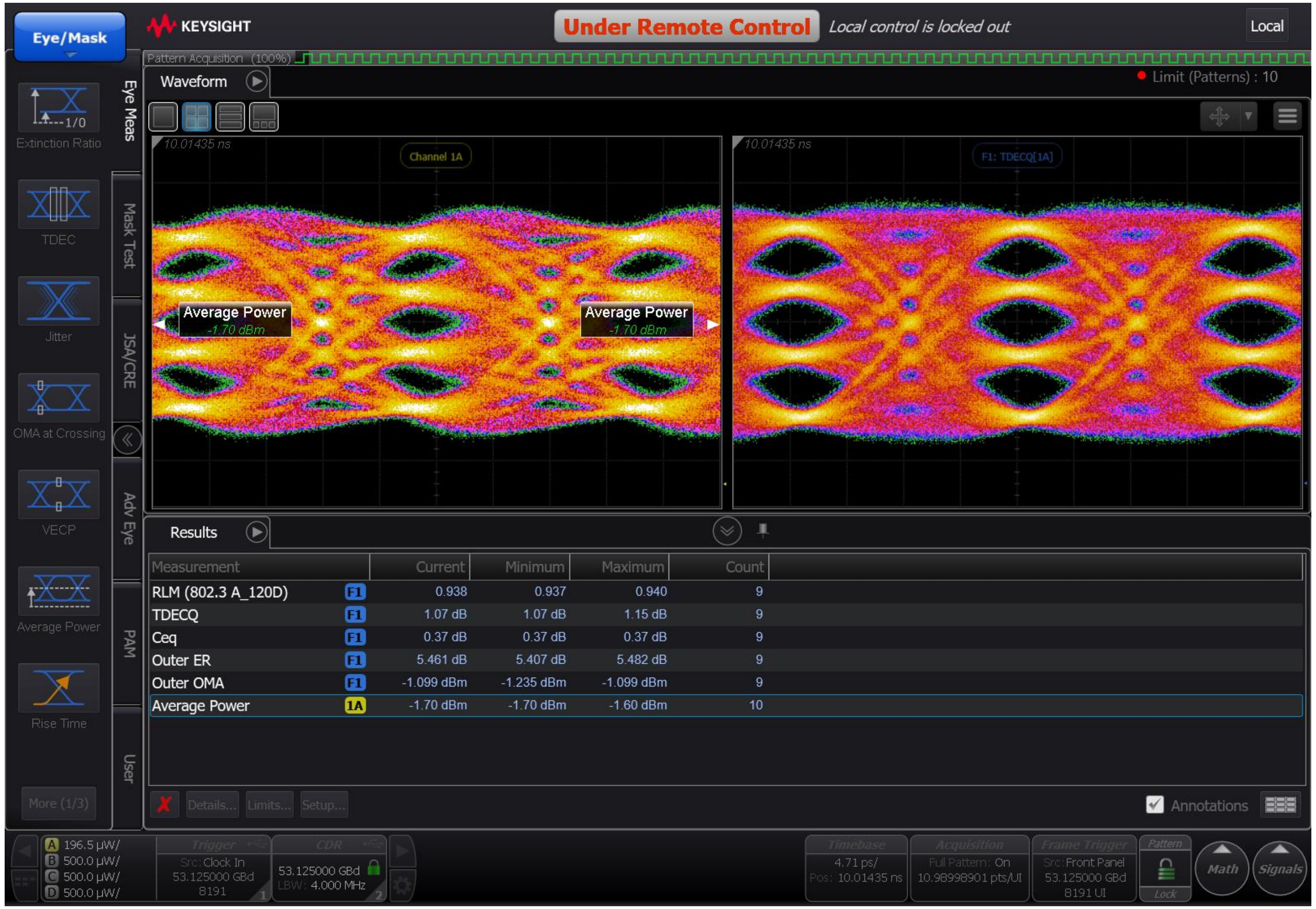
Modules vs CPO



51.2T TH5-Bailly Demonstration

Fully Functional 51.2T TH5-Bailly inside 4RU MP3 Chassis

800G Interconnect (Watts)
4.8



Under Remote Control *Local control is locked out*

Results

Measurement		Current	Minimum	Maximum	Count
RLM (802.3 A_120D)	F1	0.938	0.937	0.940	9
TDECQ	F1	1.07 dB	1.07 dB	1.15 dB	9
Ceq	F1	0.37 dB	0.37 dB	0.37 dB	9
Outer ER	F1	5.461 dB	5.407 dB	5.482 dB	9
Outer OMA	F1	-1.099 dBm	-1.235 dBm	-1.099 dBm	9
Average Power	1A	-1.70 dBm	-1.70 dBm	-1.60 dBm	10

A 196.5 µW/
B 500.0 µW/
C 500.0 µW/
D 500.0 µW/

Trigger: Src: Clock In, 53.125000 GBd, 8191

CDR: 53.125000 GBd, LBW: 4.000 MHz

Timebase: 4.71 ps/, Pos: 10.01435 ns

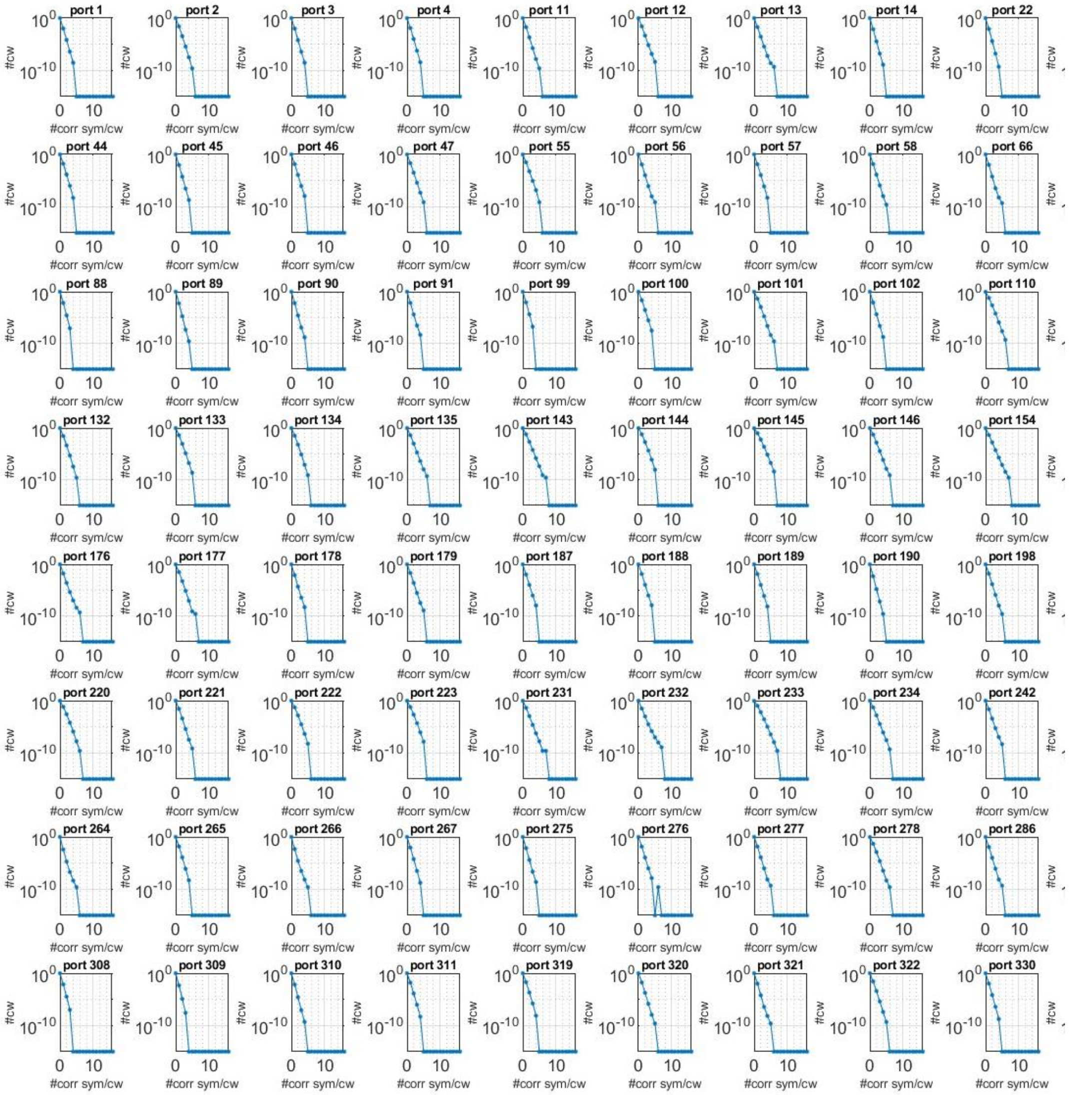
Acquisition: Full Pattern: On, 10.98998901 pts/UI

Frame Trigger: Src: Front Panel, 53.125000 GBd, 8191 UI

Pattern: Lock

TDECQ: 1.07

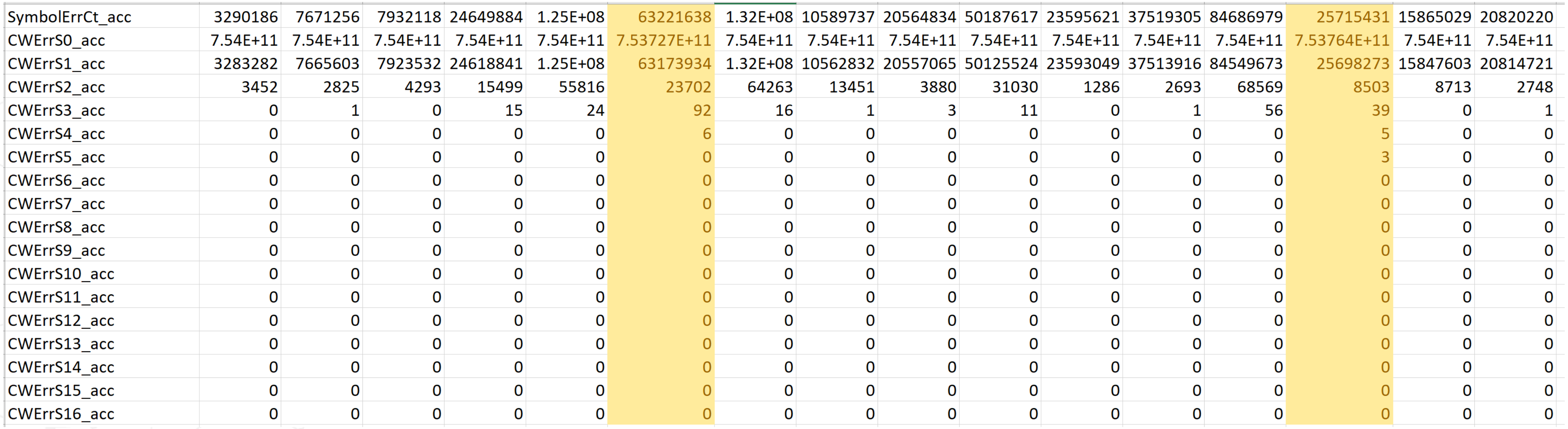
Error Free Operation on a Fully Integrated System



- FEC tail distributions collected on all 128 optical ports of a fully integrated 51.2T switch
- Production chassi assembled by ODM in a manufacturing environment
- Image to the left shows FEC tail decay curves for 72 ports

Further Optimization at Lab Level

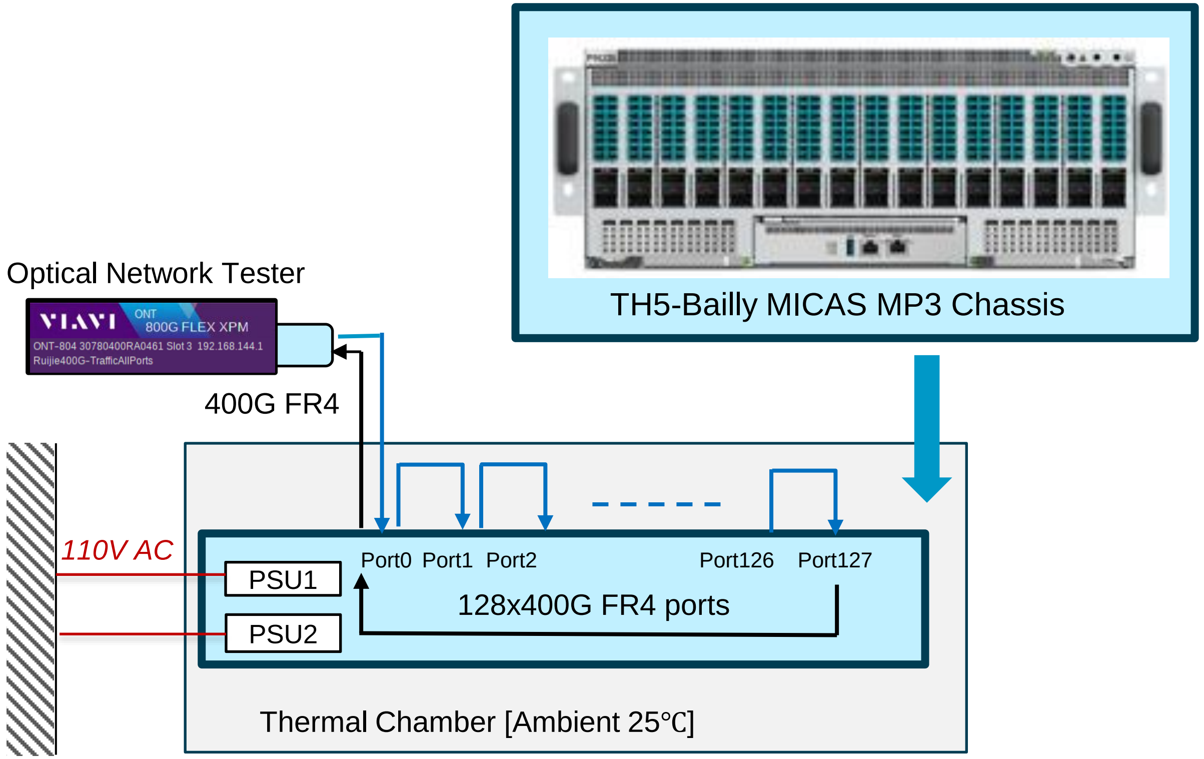
- Data on a single 64 channel optical engine (6.4T)
- Recent optimizations (still on-going) demonstrate even more rapid FEC tail decay achievable



SymbolErrCt_acc	3290186	7671256	7932118	24649884	1.25E+08	63221638	1.32E+08	10589737	20564834	50187617	23595621	37519305	84686979	25715431	15865029	20820220
CWErrS0_acc	7.54E+11	7.54E+11	7.54E+11	7.54E+11	7.54E+11	7.53727E+11	7.54E+11	7.54E+11	7.54E+11	7.54E+11	7.54E+11	7.54E+11	7.54E+11	7.53764E+11	7.54E+11	7.54E+11
CWErrS1_acc	3283282	7665603	7923532	24618841	1.25E+08	63173934	1.32E+08	10562832	20557065	50125524	23593049	37513916	84549673	25698273	15847603	20814721
CWErrS2_acc	3452	2825	4293	15499	55816	23702	64263	13451	3880	31030	1286	2693	68569	8503	8713	2748
CWErrS3_acc	0	1	0	15	24	92	16	1	3	11	0	1	56	39	0	1
CWErrS4_acc	0	0	0	0	0	6	0	0	0	0	0	0	0	5	0	0
CWErrS5_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
CWErrS6_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS7_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS8_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS9_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS10_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS11_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS12_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS13_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS14_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS15_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CWErrS16_acc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

51T Switch Chassis Power Consumption with CPO

CPO Power Measurement Test Set-up


Optical Network Tester
VIAVI
ONT
800G FLEX XPM
ONT-804 30780400RA0461 Slot 3 192.168.144.1
Ruijie400G-TrafficAllPorts
400G FR4
TH5-Bailly MICAS MP3 Chassis
110V AC
PSU1
PSU2
Port0 Port1 Port2
Port126 Port127
128x400G FR4 ports
Thermal Chamber [Ambient 25°C]

Power Measurement Results

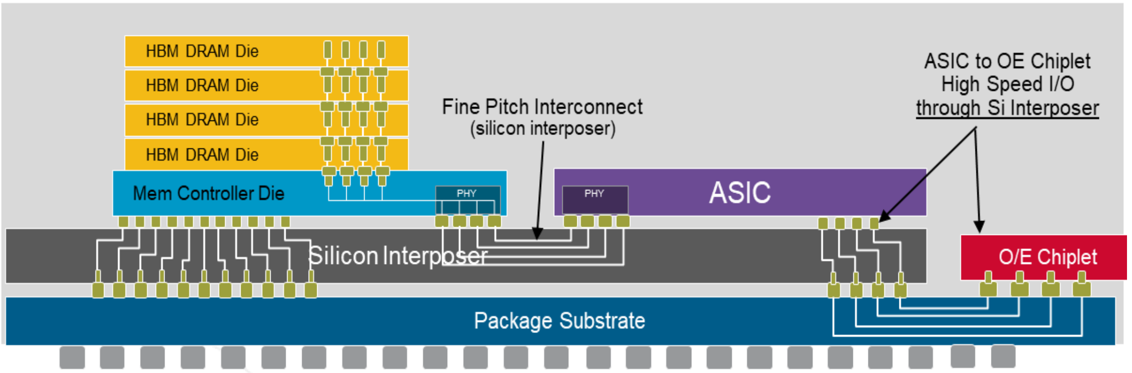
51T Switch Box Total Power			
	Bailly CPO	Pluggable LPO	Pluggable w/DSP
8x OE Chiplets	241	630	1024
16x PLS	118		
ASIC, CPU, Other	975	975	975
Total (Watts)	**1,334**	**1,605**	**1,999**

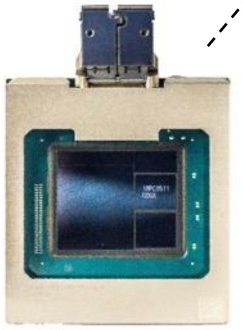
- Optical Interconnect using Bailly CPO is 70% lower power
- Total Switch Box is ~30% lower power using Bailly CPO

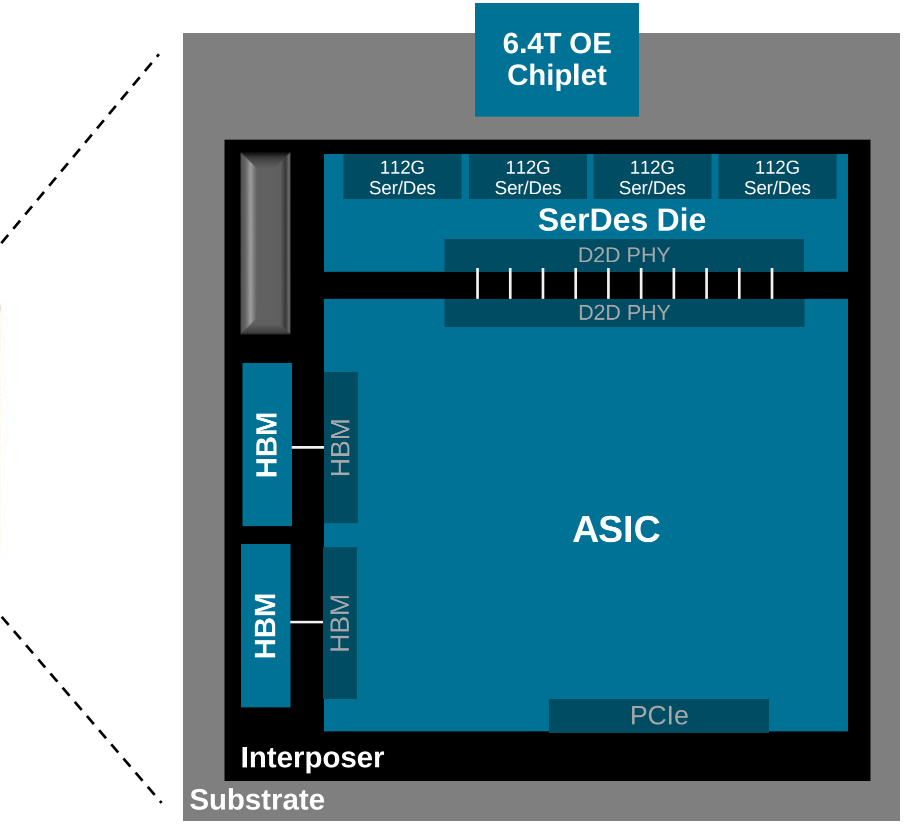
For a 32k GPU cluster, achieve between > 1MW power consumption savings

Scale-Up Compute using CPO

Stage 3: Compute ASICs with CPO





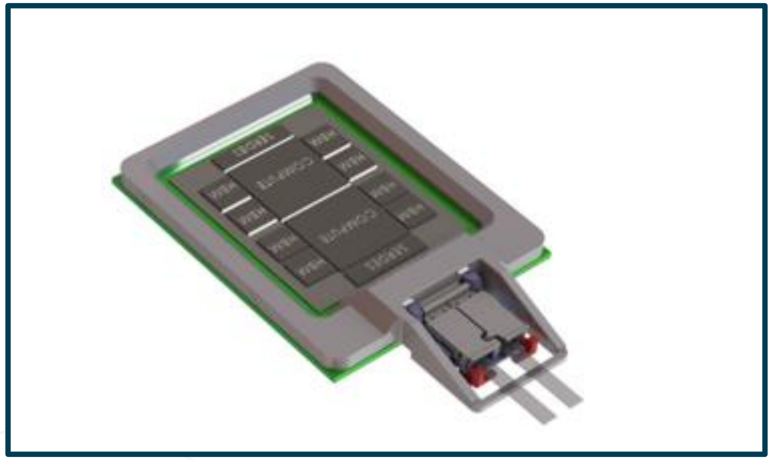


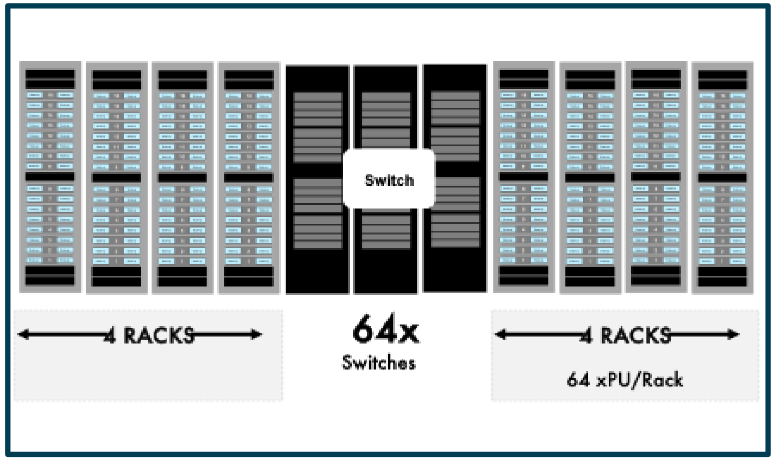
AI scaling with 2.5D multi-die packaging

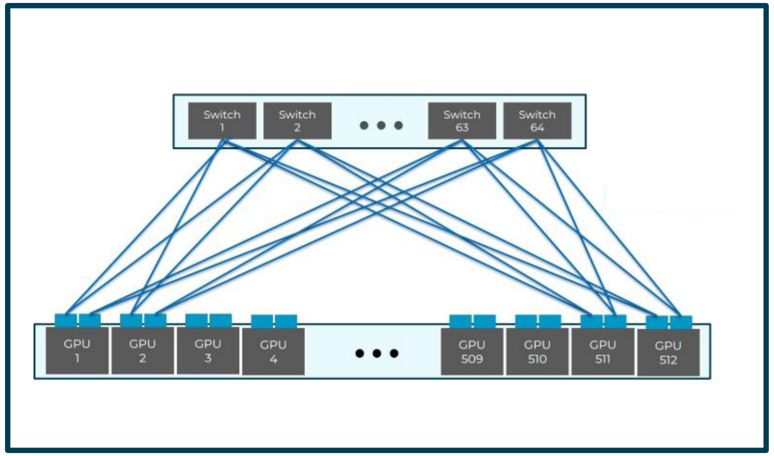
CoWoS Package with Si Interposer, O/E chiplets and HBM

CPO with 6.4Tbps I/O BW per optical Engine

Single-Stage Scale-up Fabric with All-to-All 512 GPU Connectivity







CPO GPU Attach

Physical Diagram

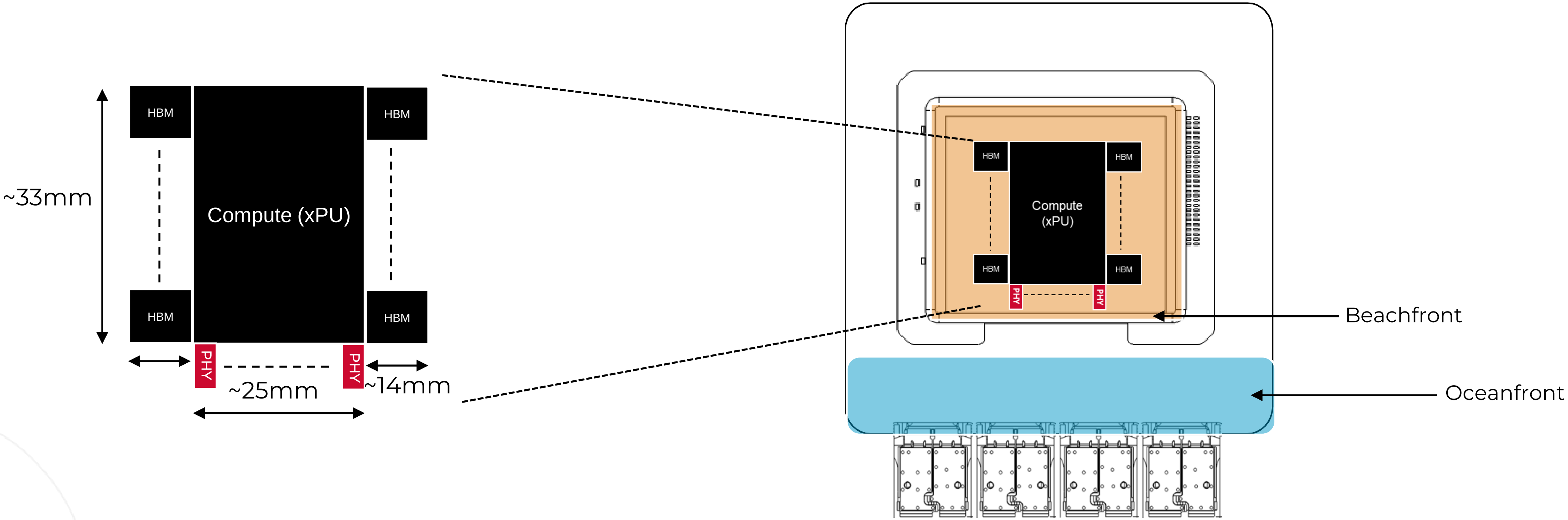
- 512 GPU in single row of connectivity
- Optical links 5m-30m (single layer)

Logical Diagram

- 512 GPUs, 64 high radix switches
- Each GPU connects to all 64 switches via CPO optics

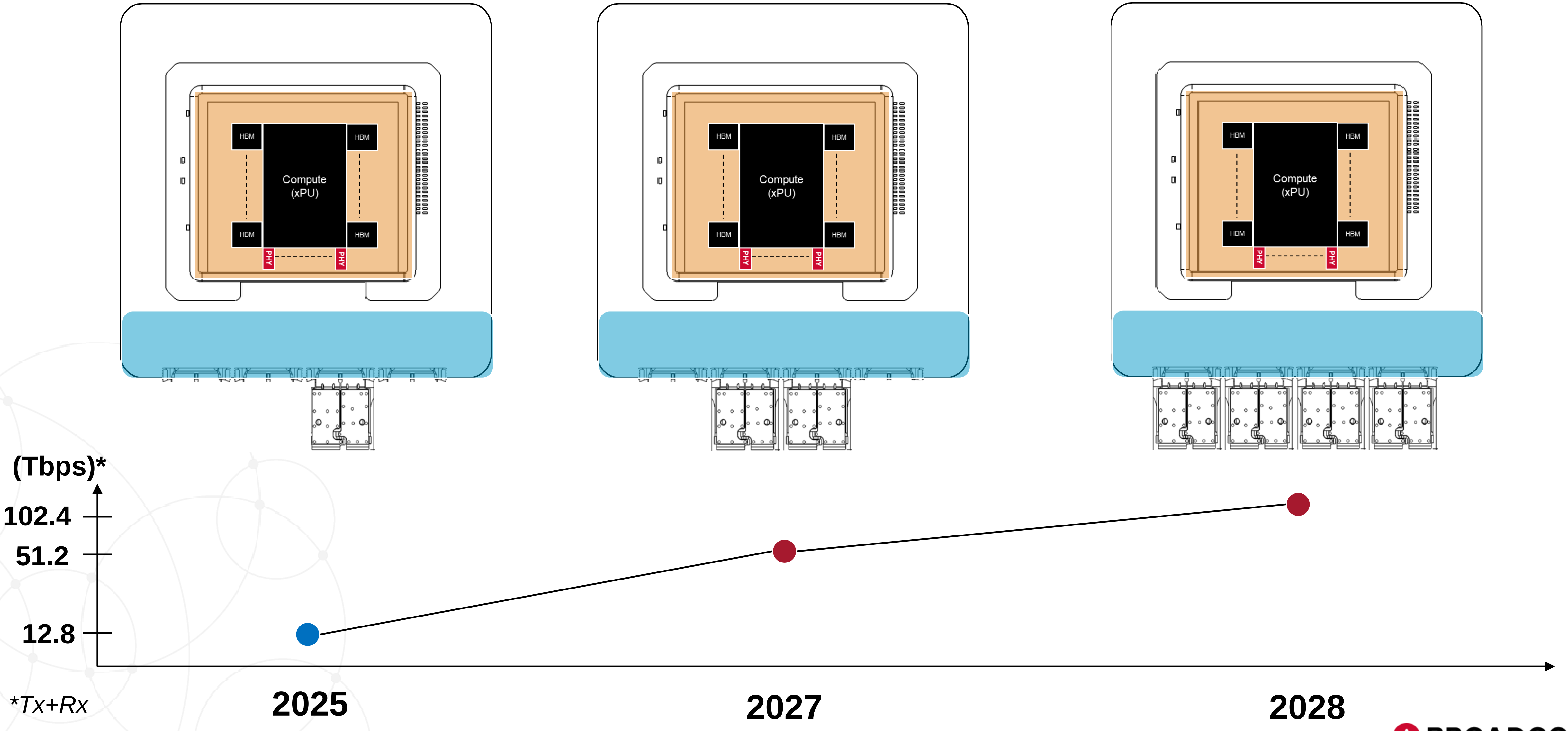
CPO can enable even larger scale-up domains

Beachfront vs Oceanfront: Utilizing Fan-Out



- **Can escape four optical engines along single oceanfront**
- **Much more reliable and cost effective**
  - Optics is farther away from high power dissipation GPU
  - Known Good Optical Engines can be attached last to the package → high manufacturing yield

Scale-Up Optical Oceanfront Density Roadmap



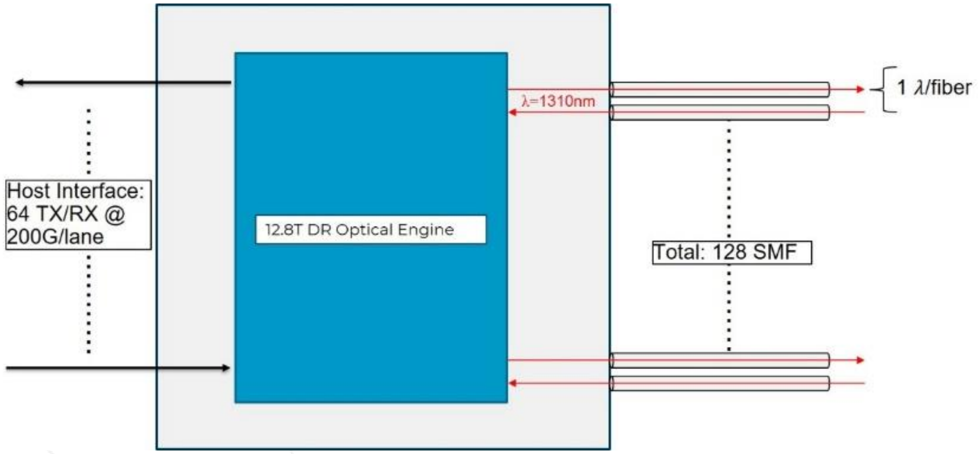
*Tx+Rx

Using Integration and CPO to Minimize Network Cabling Cost in High Radix Clusters

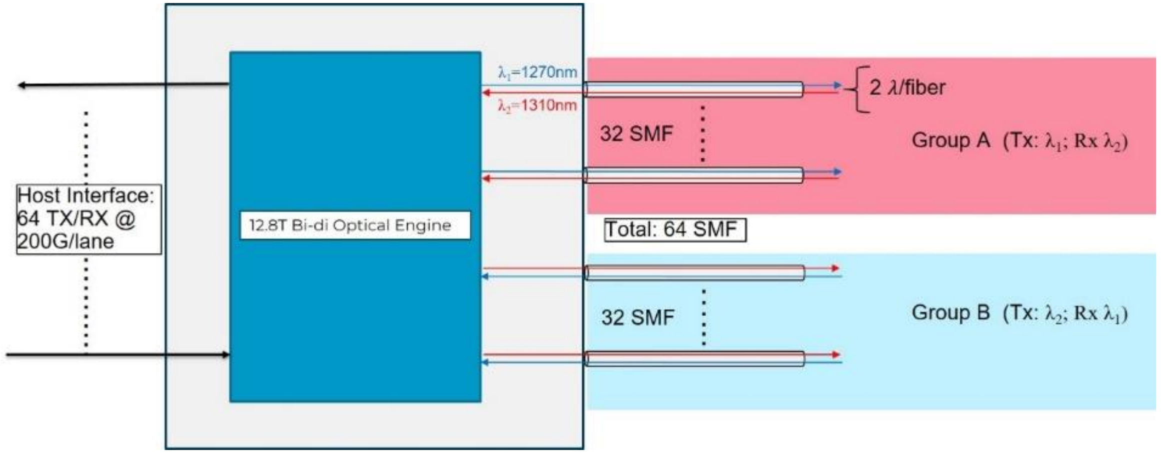
BROADCOM®

The Lowest Cost High Radix Solution: Bidi Optics

12.8T optical engine fiber I/O with DR optics



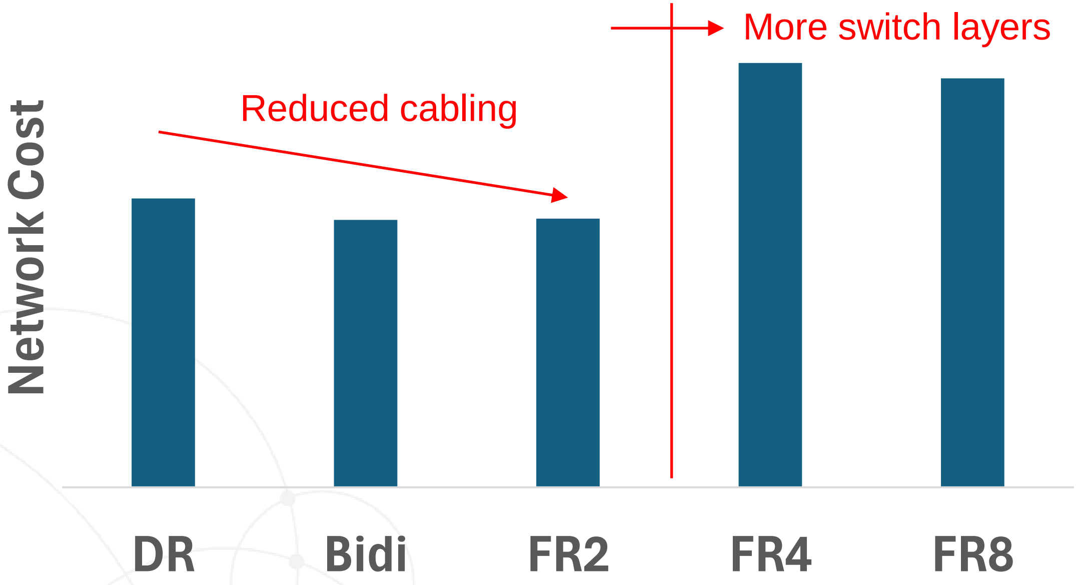
12.8T optical engine fiber I/O with Bidi optics


λ_1=1270nm
λ_2=1310nm
2 λ/fiber
32 SMF
Group A (Tx: λ_1; Rx λ_2)
Host Interface:
64 TX/RX @
200G/lane
12.8T Bi-di Optical Engine
Total: 64 SMF
32 SMF
Group B (Tx: λ_2; Rx λ_1)

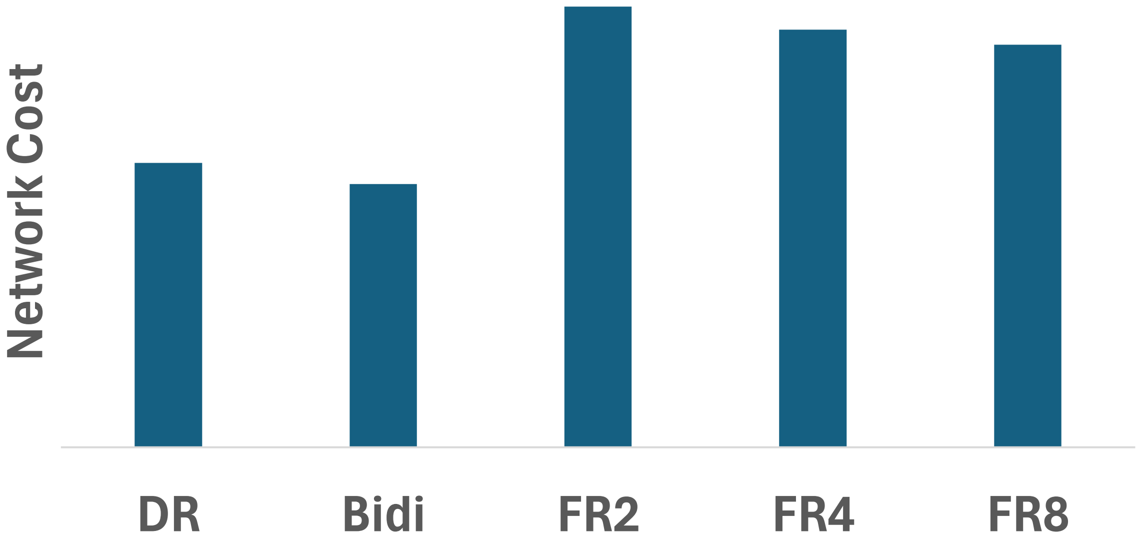
Bidi widely deployed in FTTx for 20 years, now an attractive solution for AI

Implication of Radix and Cabling Selection on Network Cost

32k GPU cluster, 512 radix capable switch



64k GPU cluster, 512 radix capable switch



High Radix + Efficient Cabling = Lowest Latency and Lowest Cost

Thank You

BROADCOM®

connecting everything®