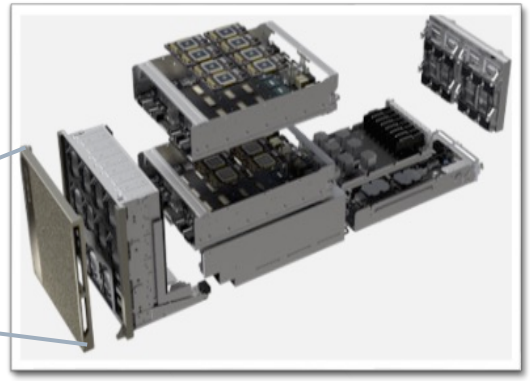
High-Bandwidth Density, Energy-Efficient, Short-Reach Signaling that Enables Massively Scalable Parallelism

John Wilson
Nvidia Research, Durham, NC

Fundamental Motivation

- Energy-cost of communication is a tax on computational performance
- Computational performance … per Watt … is the only thing that matters
- All systems are limited by power

DGX-2



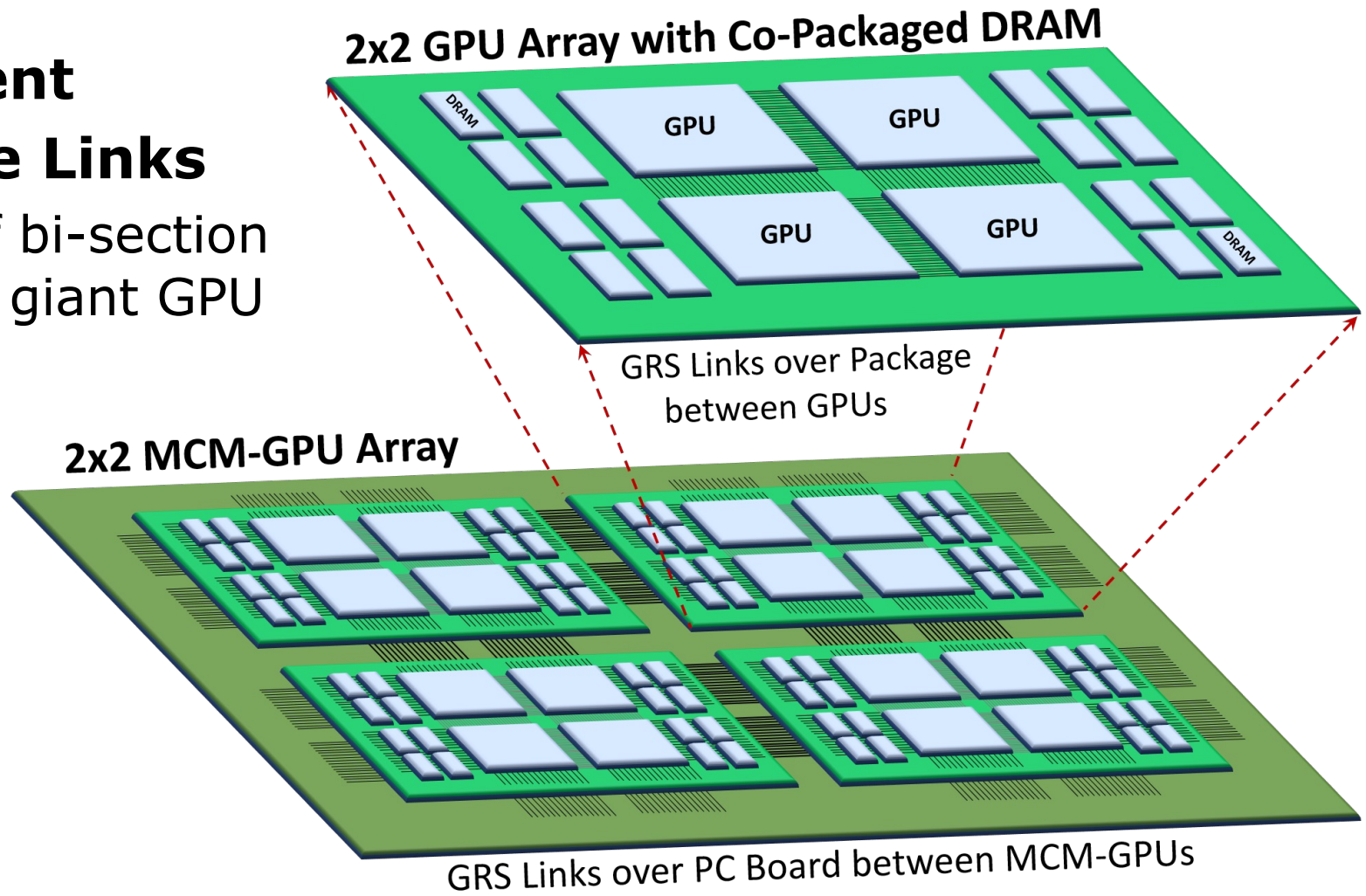
Goal = Increase Computational Performance Per Watt

Systems of the Future

Energy-Efficient

On- & Off-Package Links

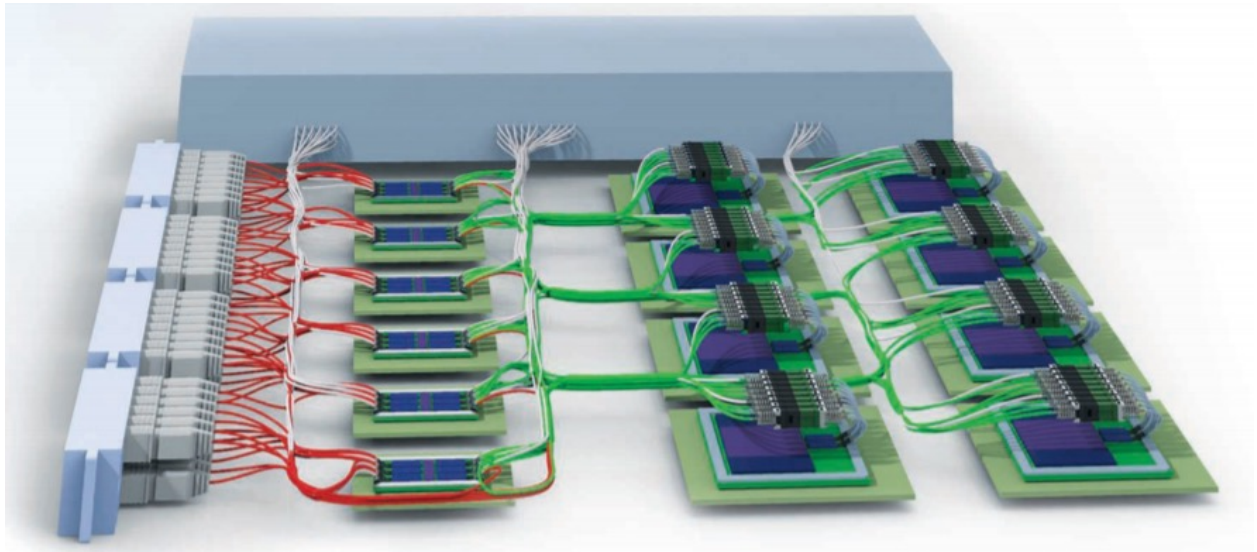
Requires several TB/s of bi-section bandwidth to emulate a giant GPU



We want to enable massively scalable parallelism

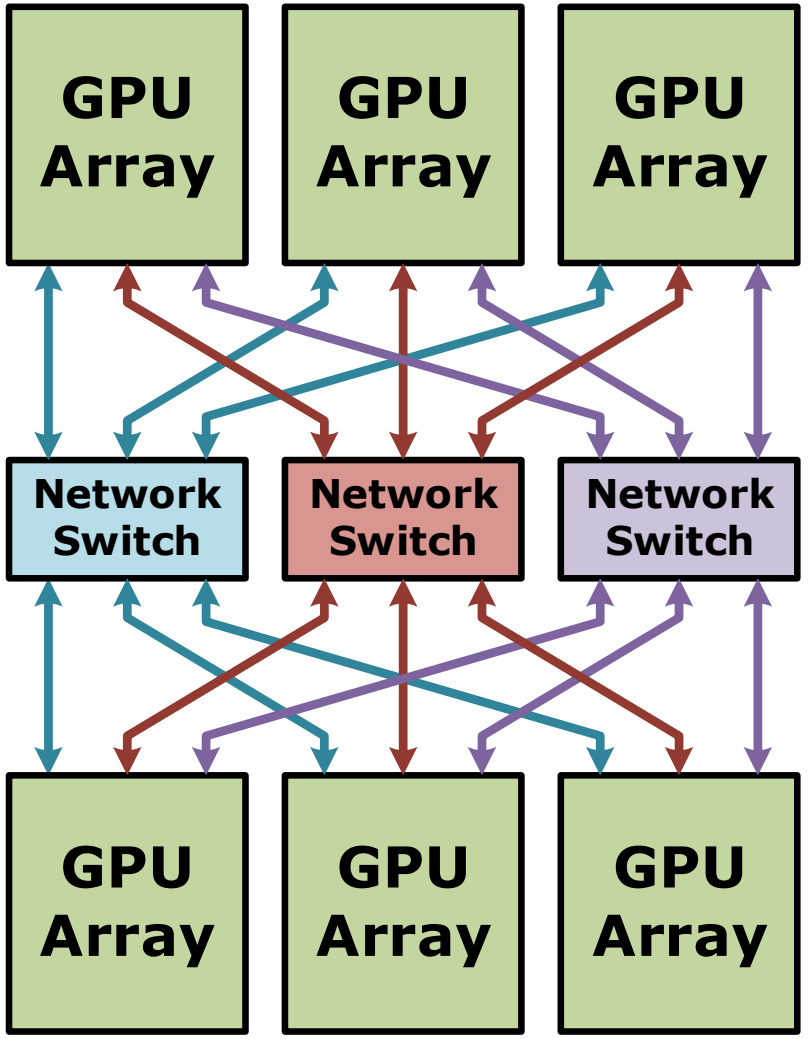
Systems of the Future

GPUs connected via Optical links

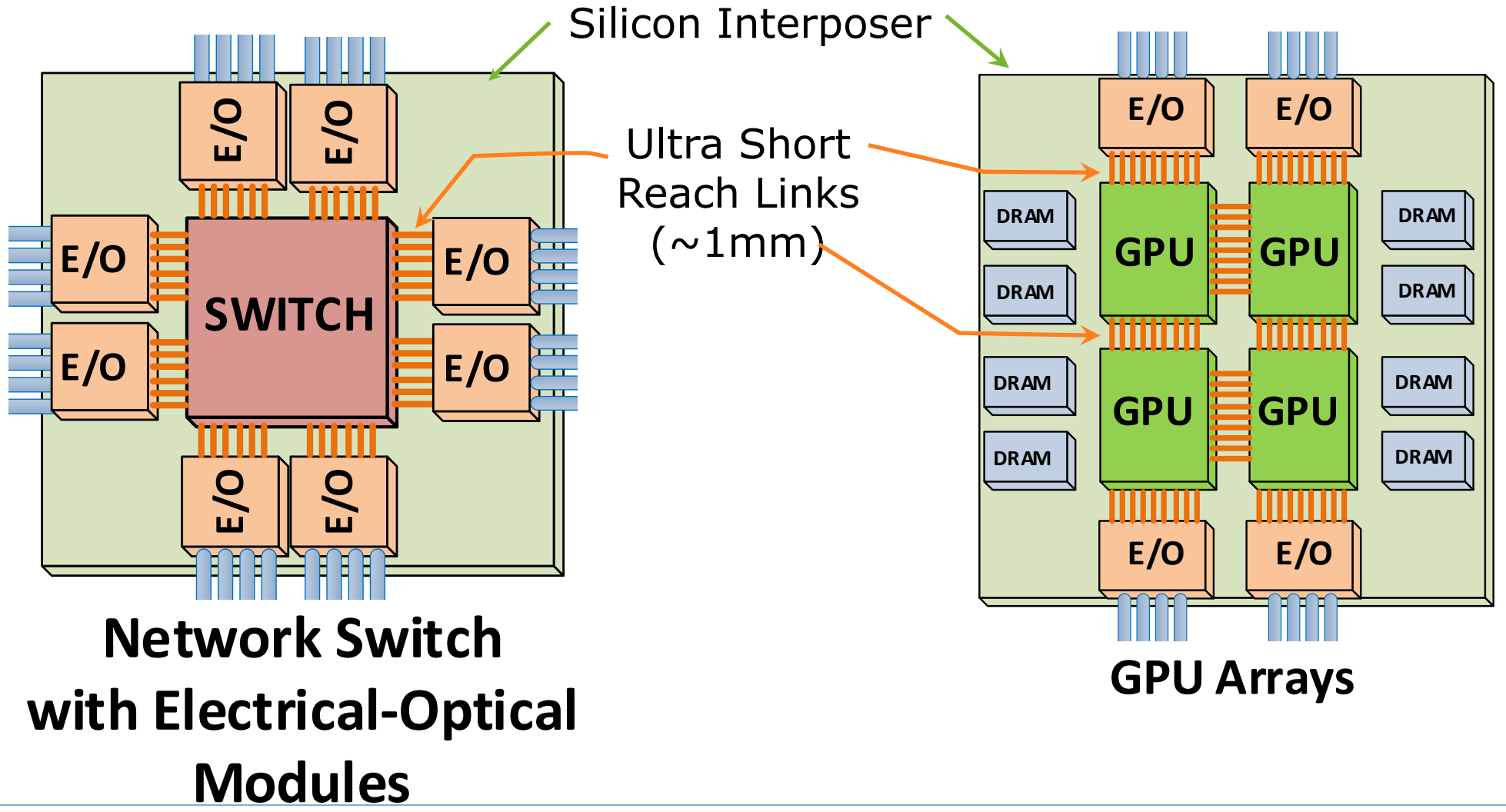


- Networking switches interconnect multiple processors for global communication
- To meet BW and energy challenge, chips will be interconnected with optical links

Need low-power ultra-short-reach electrical links

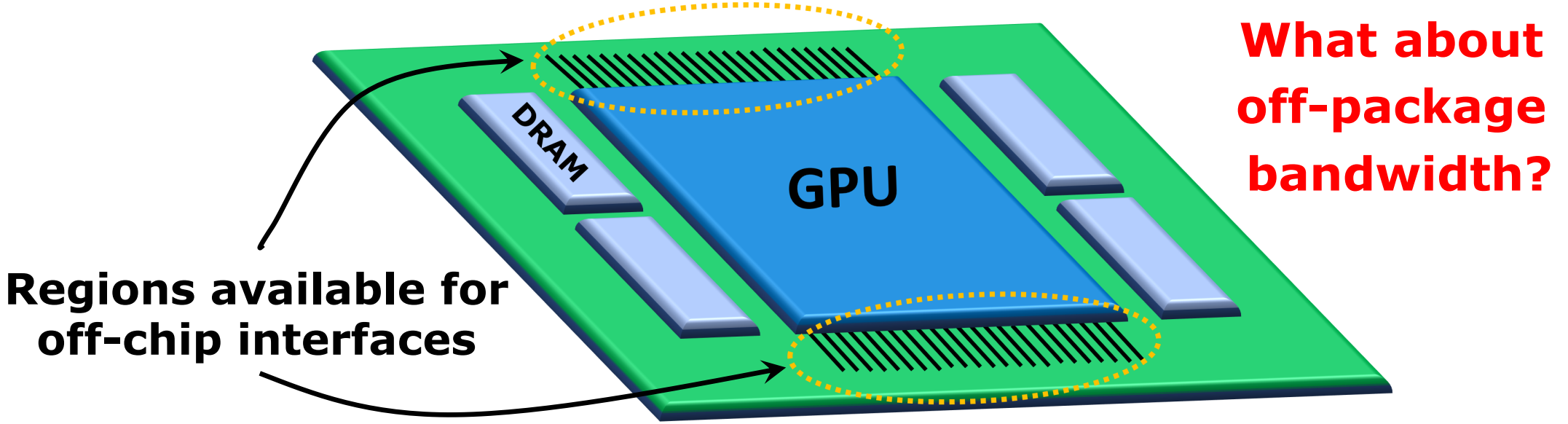


Ultra Short Reach Links on Interposer



What are the Limits to Off-Chip Bandwidth … ?

- Let's assume a simple system with a single GPU & DRAM
  - Reticle limited GPU (25.6 x 32.6mm) w/ 24.25 x 31.25mm usable
  - Both longer sides are for DRAM interfaces
  - 2 x 24.25mm available for off-chip communication

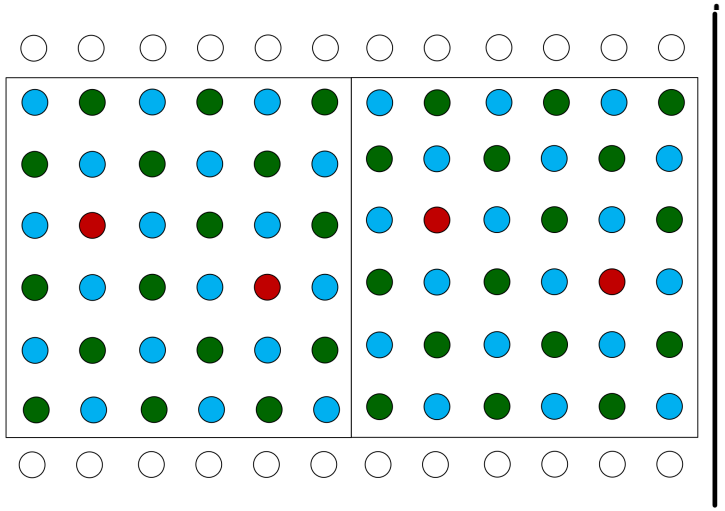


Best Known SE Method: 6x6 Checkerboard

- Total Off-Chip Bandwidth: Escape using 2 Routing Layers above core
  - Use top & bottom edges of a reticle limited die …

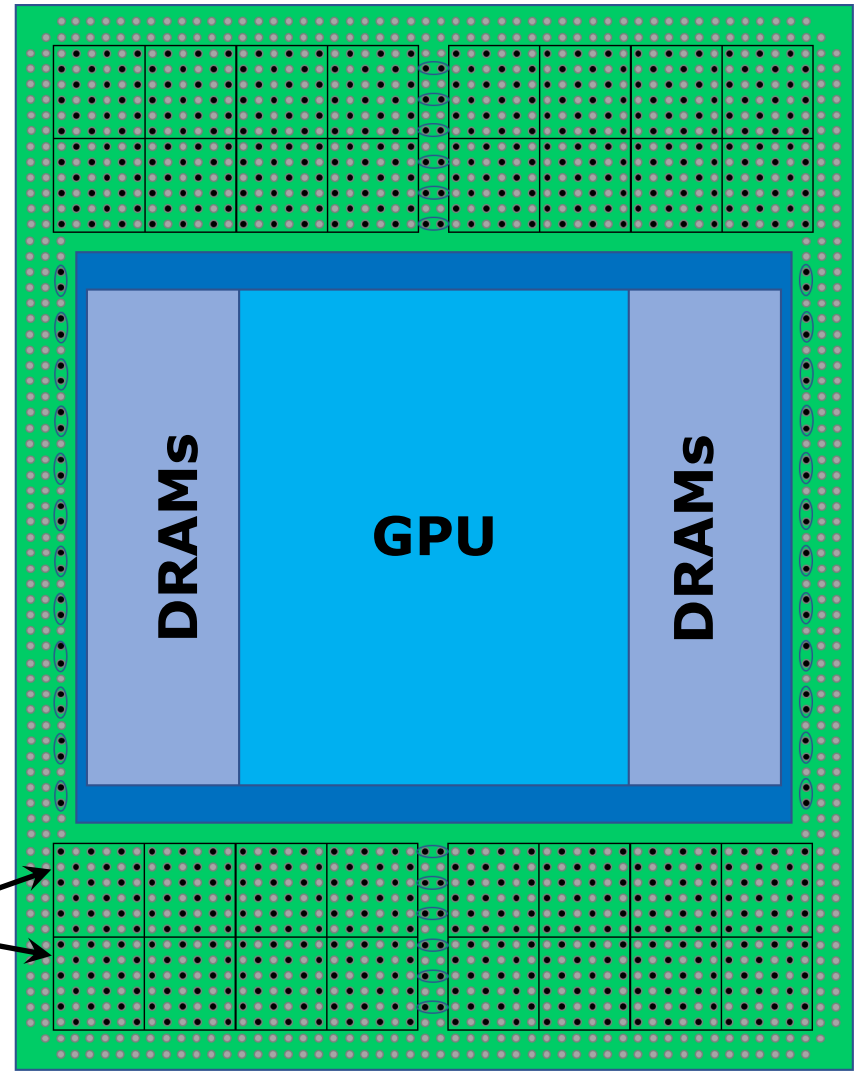
Metric	Single-Ended Signaling
Signaling Style	NRZ
Data Rate	25 Gbps
Output Mux & Clocks	2:1 @ 12.5 GHz
Nyquist Freq.	12.5 GHz
Bump Pitch	130 x 130 um
Edge Length per Byte	0.78 mm
Bandwidth Density per Layer	513 Gbps/mm/Layer
Total Off-Chip BW w/ 2-Layers	**49.6 Tbps**

Escapes two wires between bumps



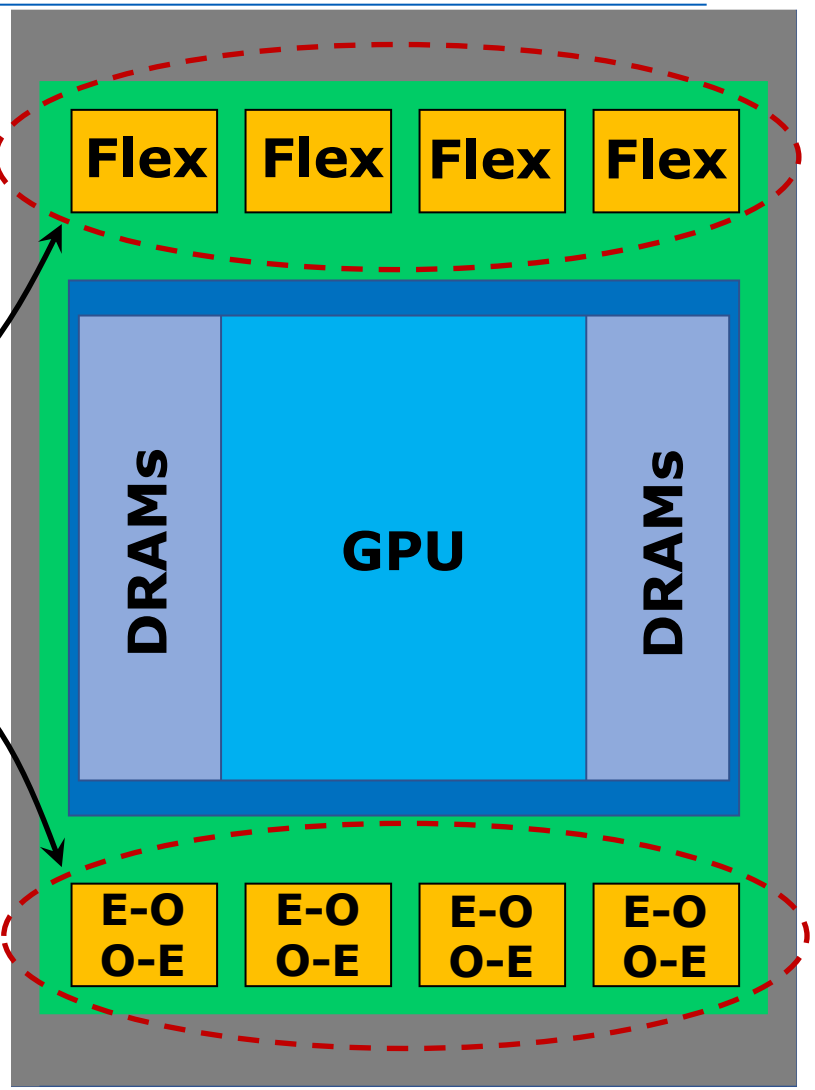
What are the Limits to Off-Package Bandwidth?

- Try to escape from the **bottom** of the package
  - SI Issues: Pkg/PCB core vias & BGA or LGA pins
  - Limited bandwidth: ~1/4 compared to off-chip
    - 25Gbps SE Sys: 12.8Tbps (1.6TBps)
  - Package size is at least 55x70mm
    - Still issues to work through …
    - EPYC Package: 58.5x75.4mm
      - SP3 socket, 4094 LGA



What are the Limits to Off-Package Bandwidth?

- Does a package support 49.6Tbps off-chip BW?
  - Only from **top** of 55x75.4mm package
  - 25Gbps SE: 8 regions x 6Tbps = 48Tbps
  - "Regions" → Interface-pitch to Pkg ≤ 300x300um
    - *Flat flex-cables w/ fine-pitch*
    - *Or each region could be an E-O/O-E*
- Supporting off-chip bandwidth is difficult
  - Only possible from top of package
  - Electrical solutions: Limited reach & high-power
  - Optical: Promising … under development

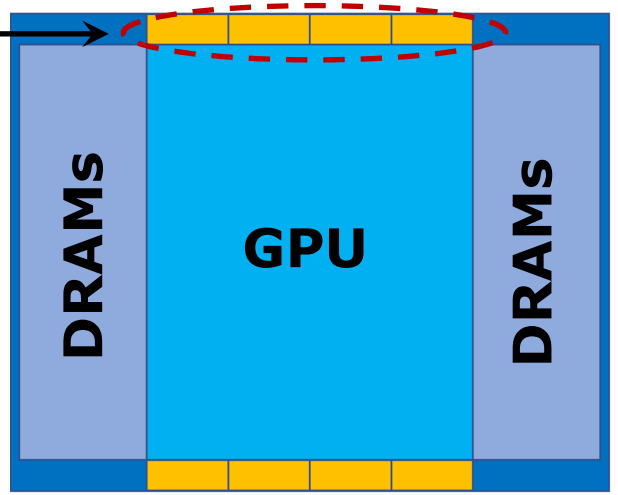


What are the Limits to Interposer Bandwidth?

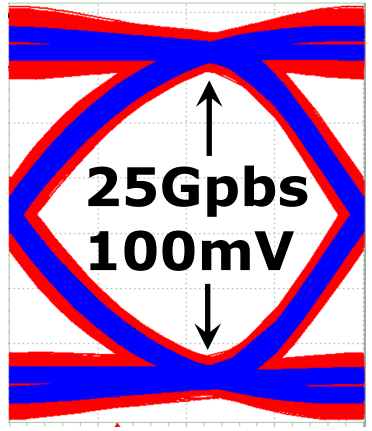
- Signaling limits: **1-2mm long** silicon interposer wires
  - Information escapes optically using **O-E/E-O Interface**

Metric	Single-Ended Signaling
Signaling Style	NRZ
Data Rate	25 Gbps
Output Mux & Clocks	2:1 @ 12.5 GHz
Nyquist Freq.	12.5 GHz
Bump Pitch	36 x 36 um
Total Edge Length	48.5 mm
Bandwidth Density per Layer	2.08 Tbps/mm/Layer
Total Off-Chip BW w/ 4-Layers	**404.2 Tbps**

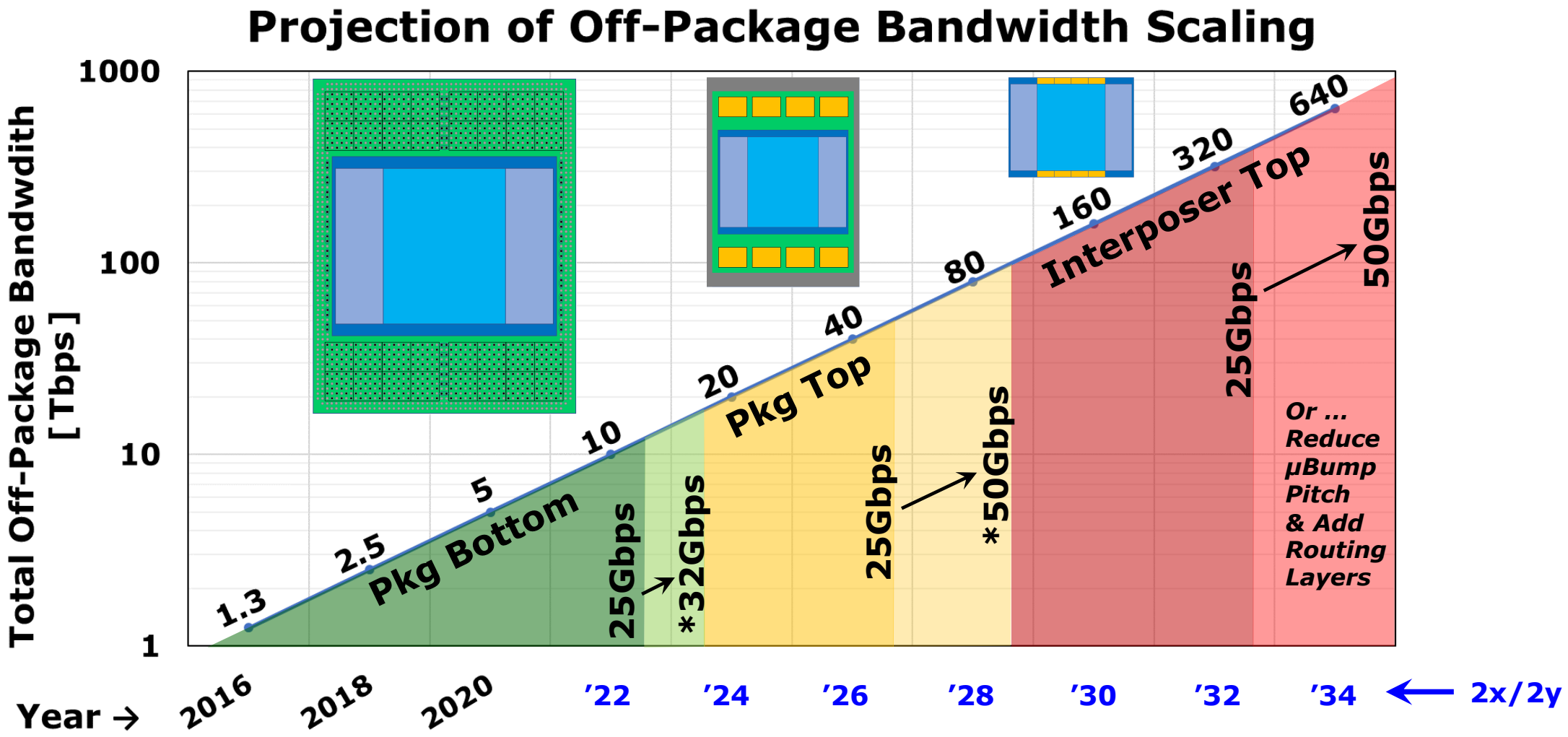
Energy-Efficient Links over Silicon Interposer



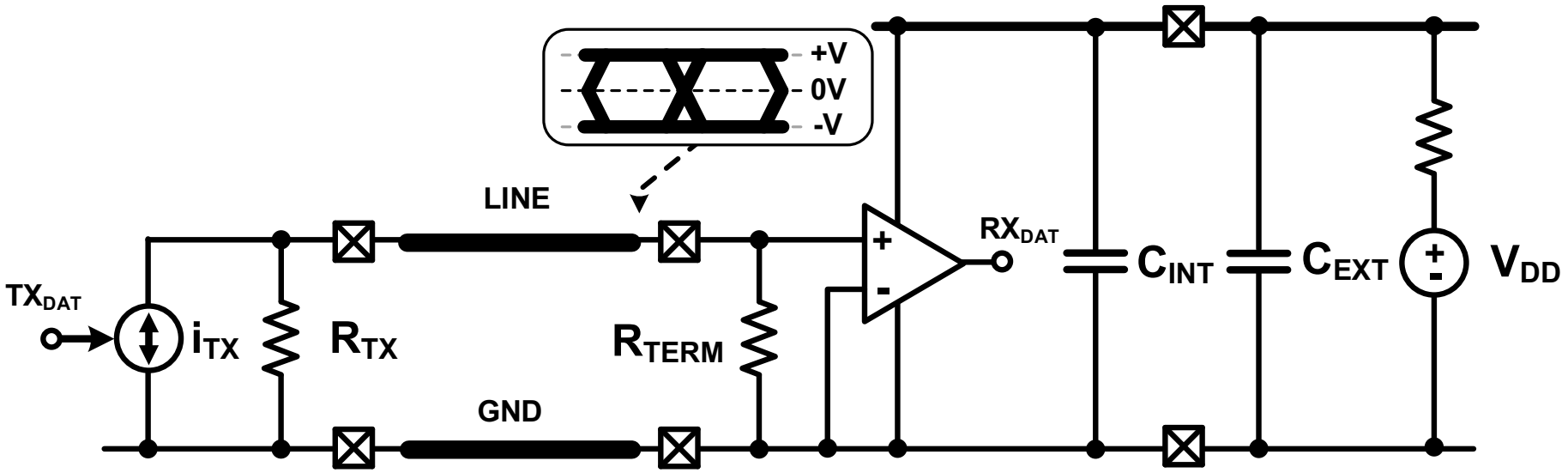
4-L Shielded
G-S-G-S…
S-G-S-G…
G-S-G-S…
S-G-S-G…



What are the Limits to Off-Package Bandwidth?



Ground Referenced Signaling (GRS)



- Ground used as Return Path
  - Lowest impedance node and all systems agree on voltage
- Charge-Pump Transmitter consumes constant current (No SSO Noise)
- Employs a Delay-Matched Clock Forwarding Approach
  - All delays in Data & Clock paths, at both Tx & Rx, track across voltage and temperature variations
  - Very high tracking bandwidth and low power

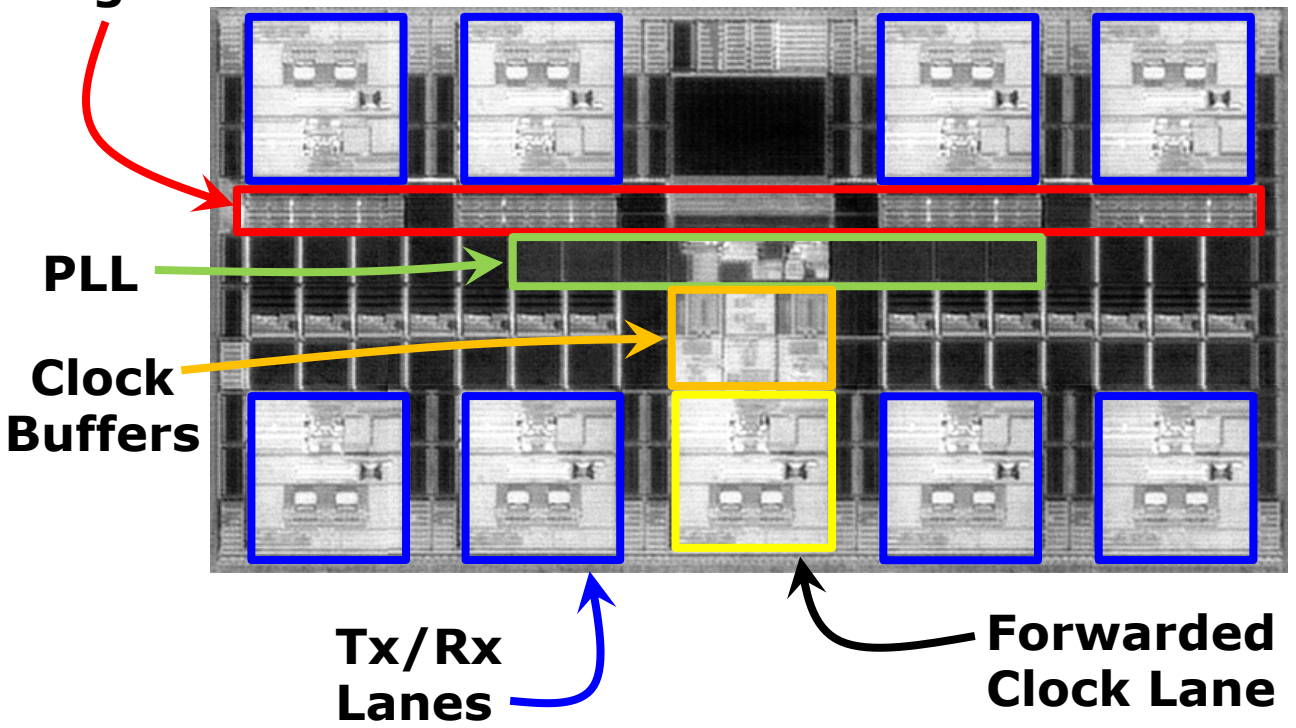
25Gb/s GRS Link Floorplan – 16nm FinFET

x8 GRS Link w/ Bumps
(686x565µm)

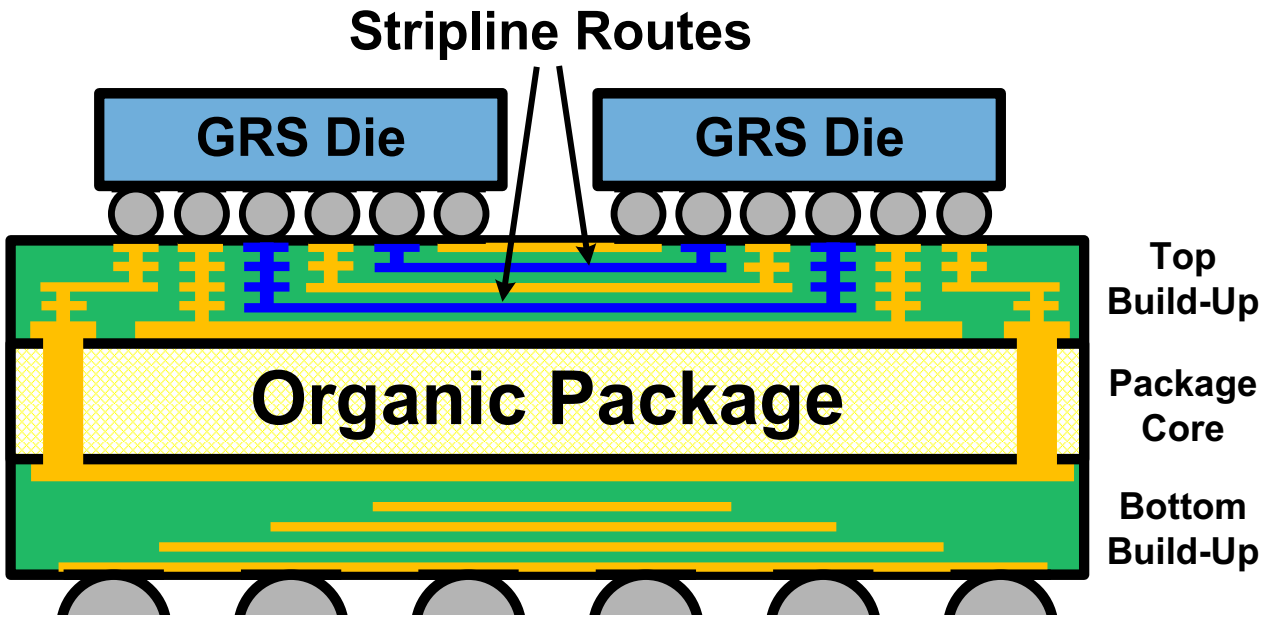
Ground
Signal
Power
Clock

150µm bump pitch

I/O Transceiver Brick
(200x400µm, 16nm FinFET)



On-Package … or … MCM Channel



10mm MCM Link

Signal Routing

10mm Package Routing (or more)

LC-Dominated Lines

40Ω Characteristic Impedance

On-Chip RDL Routes

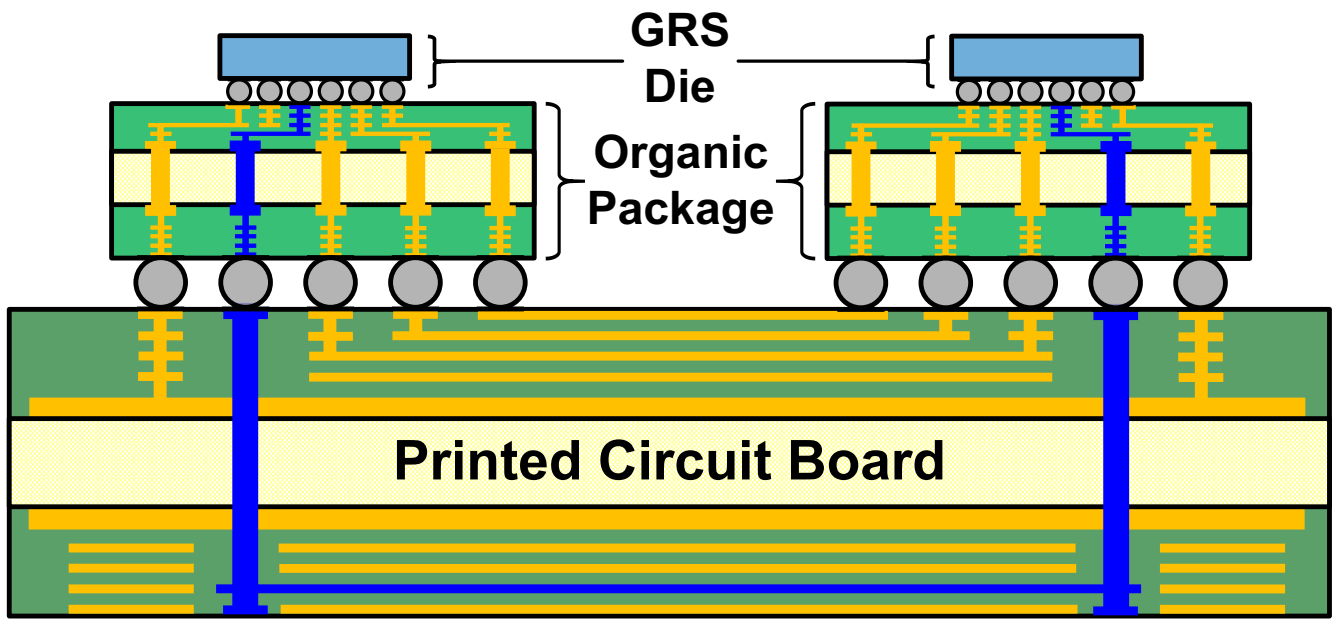
Signal Integrity

Limited to stripline routing above core

GND planes above & below traces

Signal-Ground Checkerboard Pattern

Off-Package … or … PCB Channel



80mm Off-Package PCB Link

Signal Routing

54mm PCB Channels

13mm Package Routing

LC-Dominated Lines

40Ω Characteristic Impedance

On-Chip RDL Routes

Signal Integrity

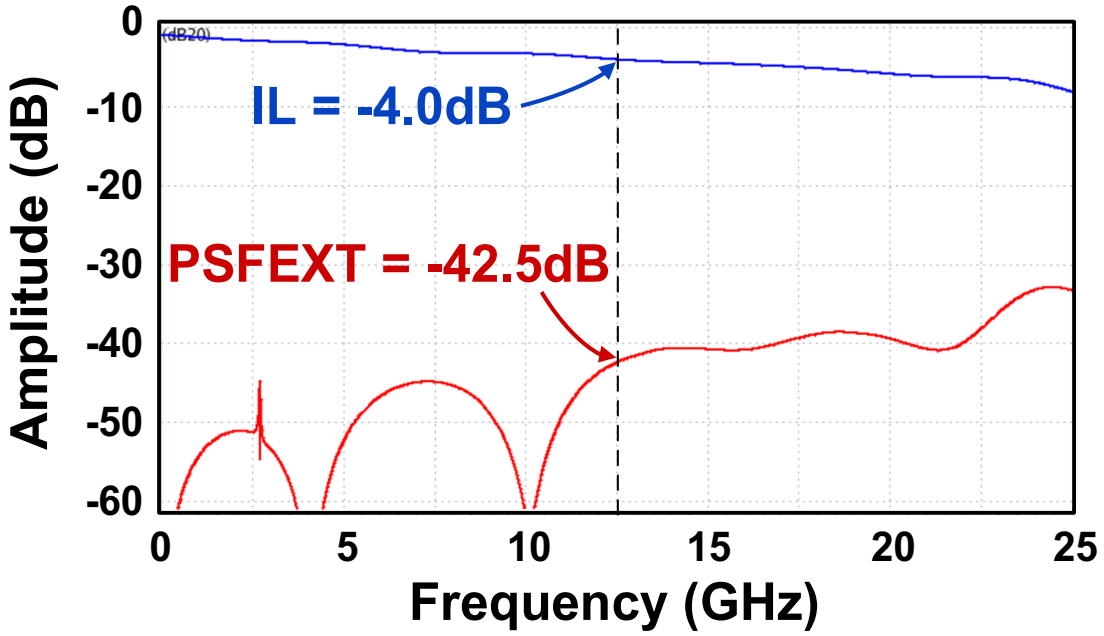
Bottom Stripline PCB Routes

Signal-Ground Checkerboard Pattern

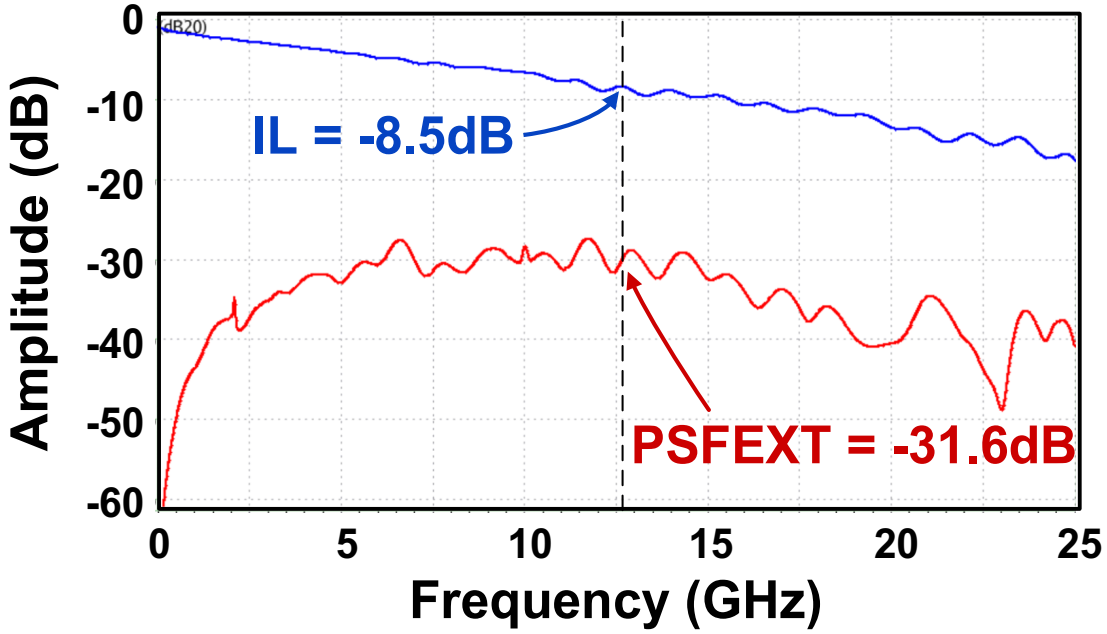
Interleaved GND Traces on PCB

MCM & PCB Channel Response

10mm MCM Link



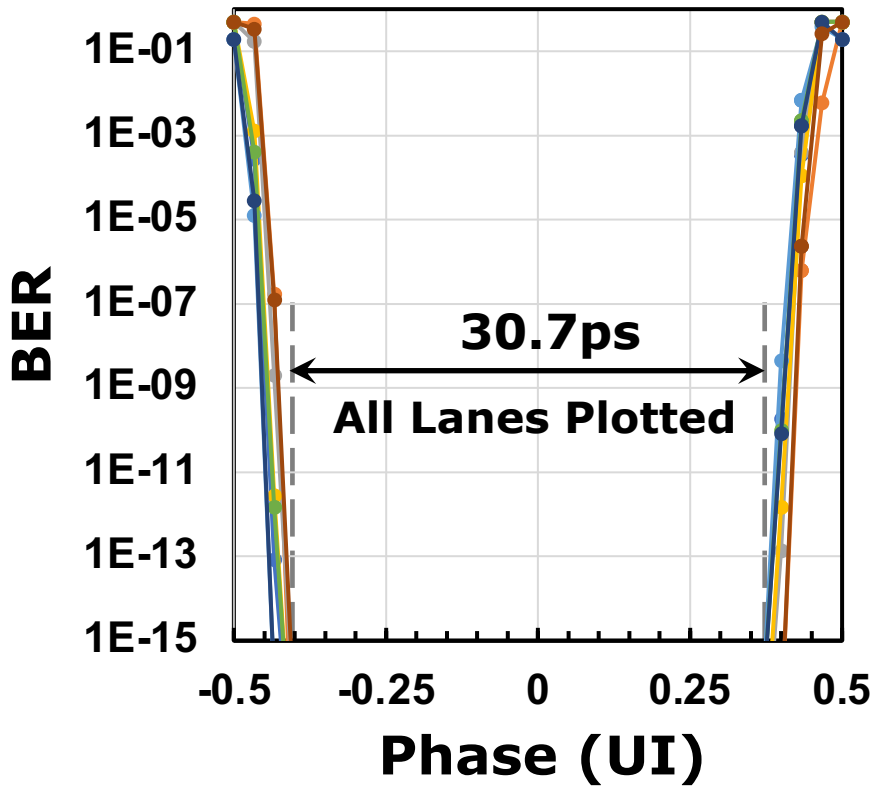
80mm Off-Package PCB Link



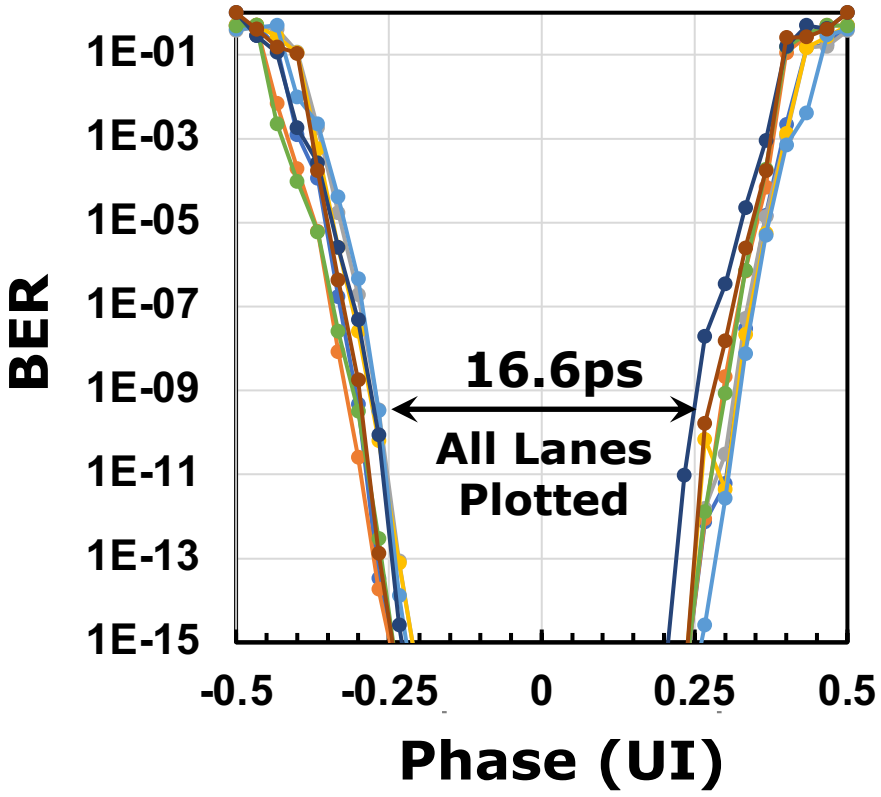
Complete channels models which include: All Circuit, ESD, & Pad Capacitance, 270um on-die RDL trace, T-Coils, and 3D Models for Package/PCB Interconnect

Measured 25Gbps **MCM** & **PCB** Link

10mm MCM Link



80mm Off-Package PCB Link



All 9 lanes operating simultaneously using PRBS-31 with different lanes seeds.

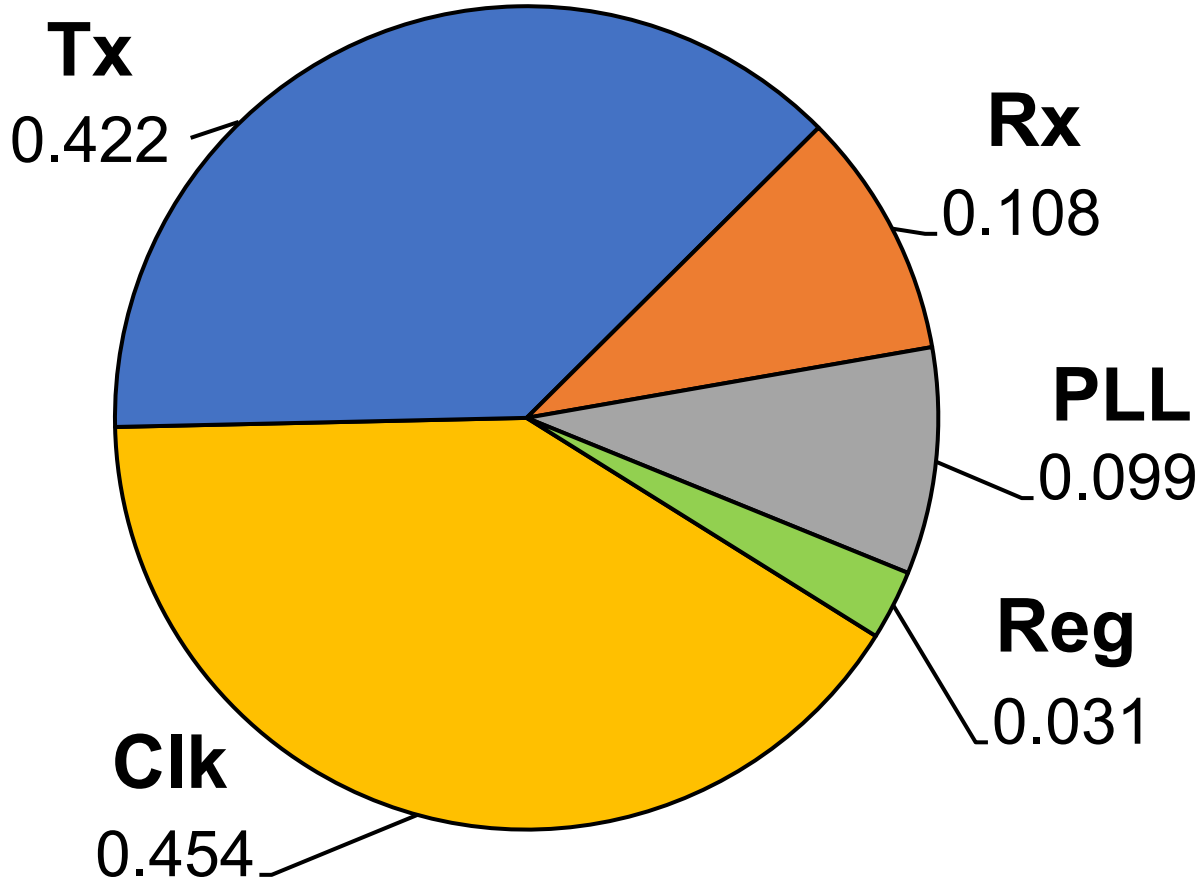
GRS Energy Breakdown: Off-Package (PCB)

1.17pJ/bit

Entire System @ 25GB/s

- 1 Clock & 8 Data Lanes
- 2x PLL
- 2x Regulator
- All clock distribution, etc.

16nm FinFET CMOS Technology
80mm PKG-to-PKG PCB Link

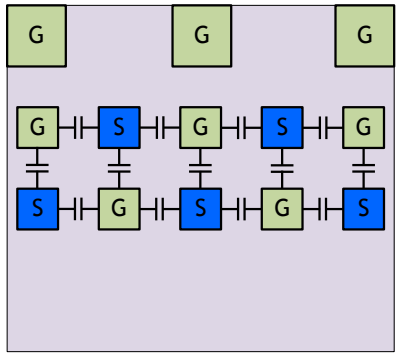


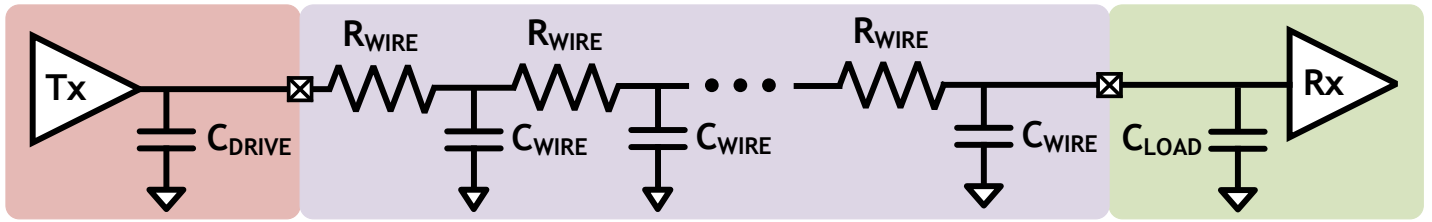
Easily drops to 1pJ/bit for "MCM-only" Link

Silicon Interposers (RC-Dominated Channels)

Keep wire length short!

Si Interposer
Thick Al RDL
2 Cu Layers
GSG routes





Fine Resolution Channels (Very High-Density)

- On-Chip Upper Metal Layers & Silicon Interposers

$$BW_{WIRE} = \frac{1}{2\pi R_{WIRE} C_{WIRE}}$$

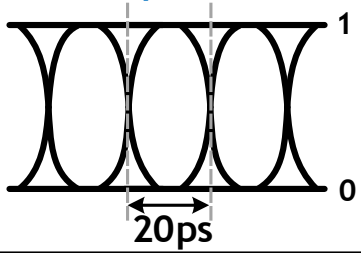
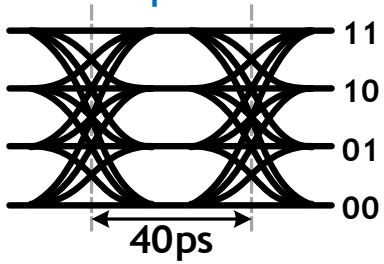
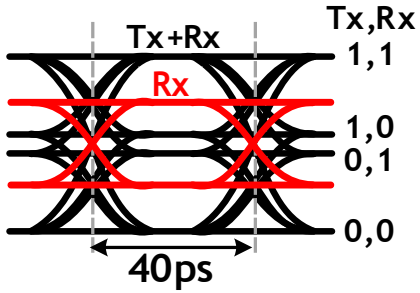
High Wire Resistance & High-Crosstalk

- Signal BW Limitation
- Requires shielding to reduce crosstalk
- Signal Reflections completely attenuated

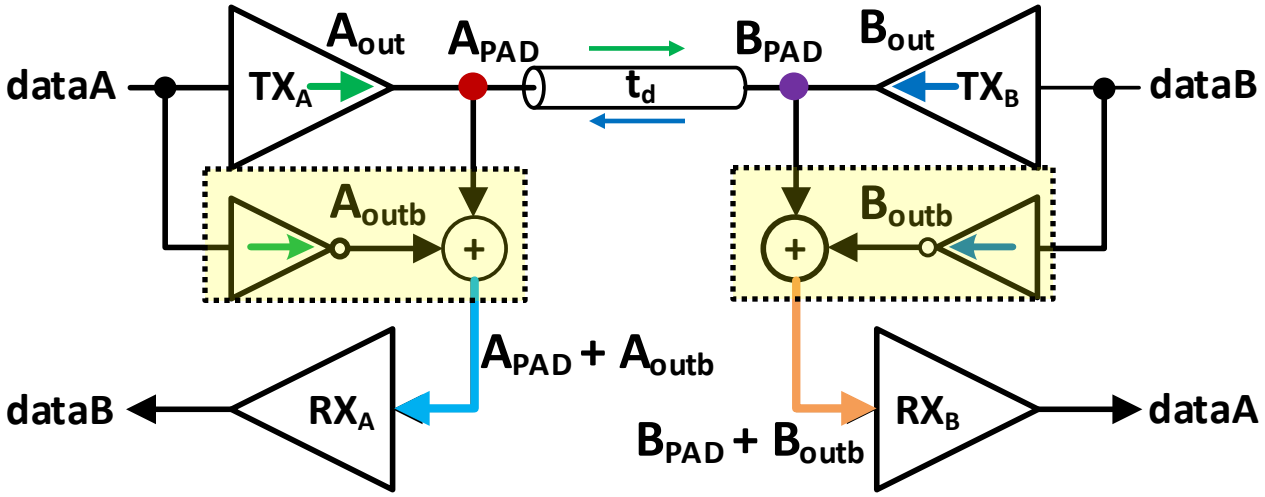
BW decreases quadratically w/ wire length
C_{WIRE} increases linearly w/ wire length
R_{WIRE} increases linearly w/ wire length

Keep wire length very short!

Modulation Scheme Comparison @ 50Gbps

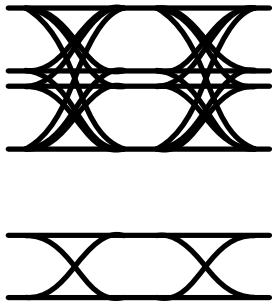
	Pros	Cons
50Gbps NRZ 	• Simple • Excellent tracking between Data & Clk with Clk forwarding	• 25GHz Nyquist rate: higher loss, reflections & crosstalk • High power • Small timing margin
50Gbps PAM4 	• 12.5GHz Nyquist	• Input w/ analog amp or sampler? • Poor Data-Clk tracking • Poor Signal-to-Noise ratio • Inherent ISI
25Gbps x 2 SBD 	• 12.5GHz Nyquist • Wide timing margin • Excellent tracking between Data & Clk with Clk forwarding	Can be resolved by careful channel design • Sensitive to reflections • Sensitive to near-end crosstalk (NEXT) • Difficult for single-ended CMOS

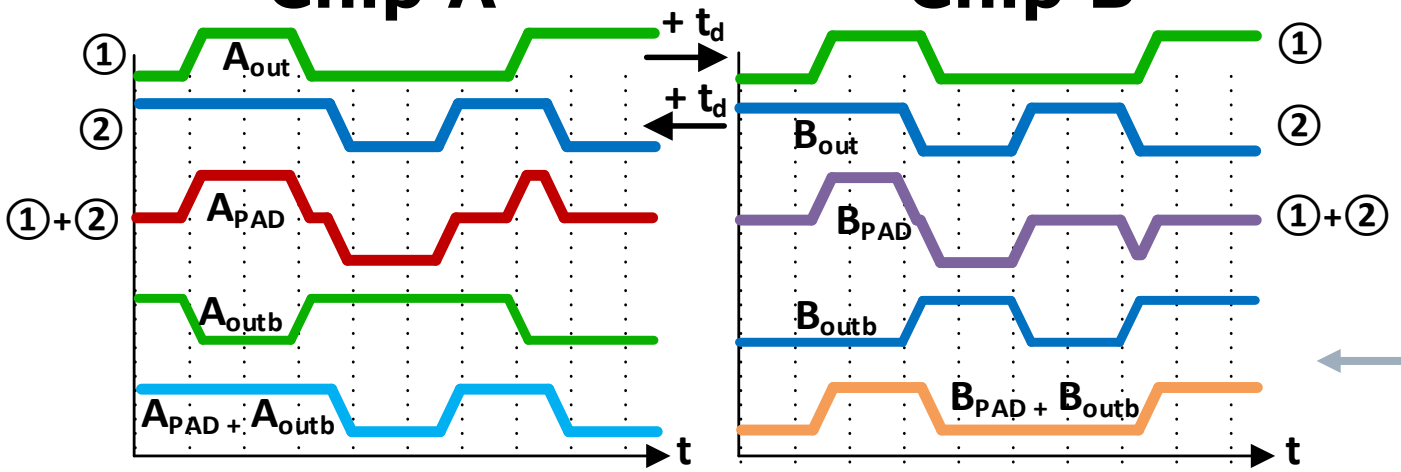
SBD (Simultaneous Bi-Directional) Signaling



Hybrid

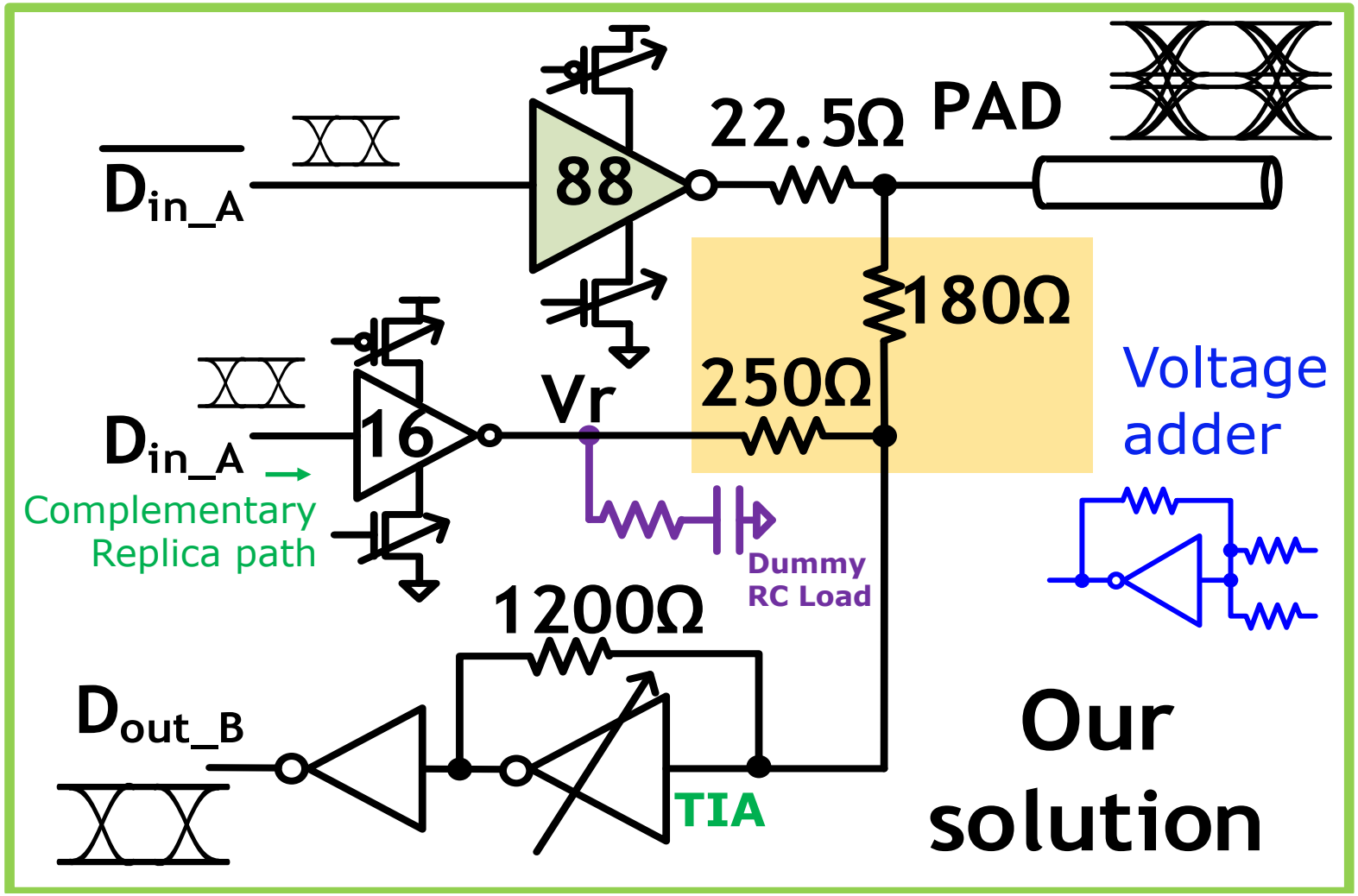
"Signal Separator"



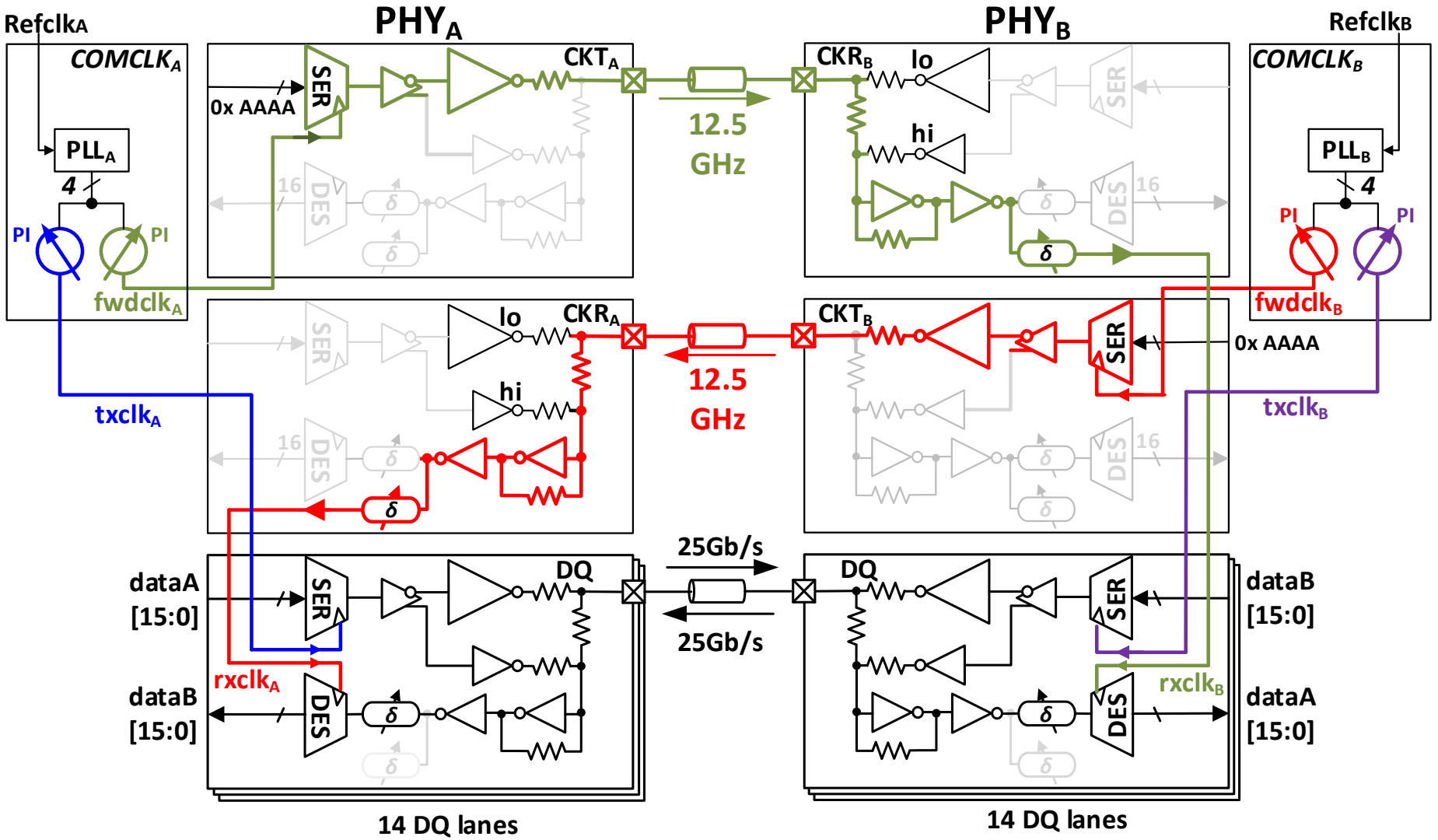


How do we build a ***Hybrid*** *to* perform **Signal Separation** using single-ended CMOS?

Single-Ended Hybrid Design

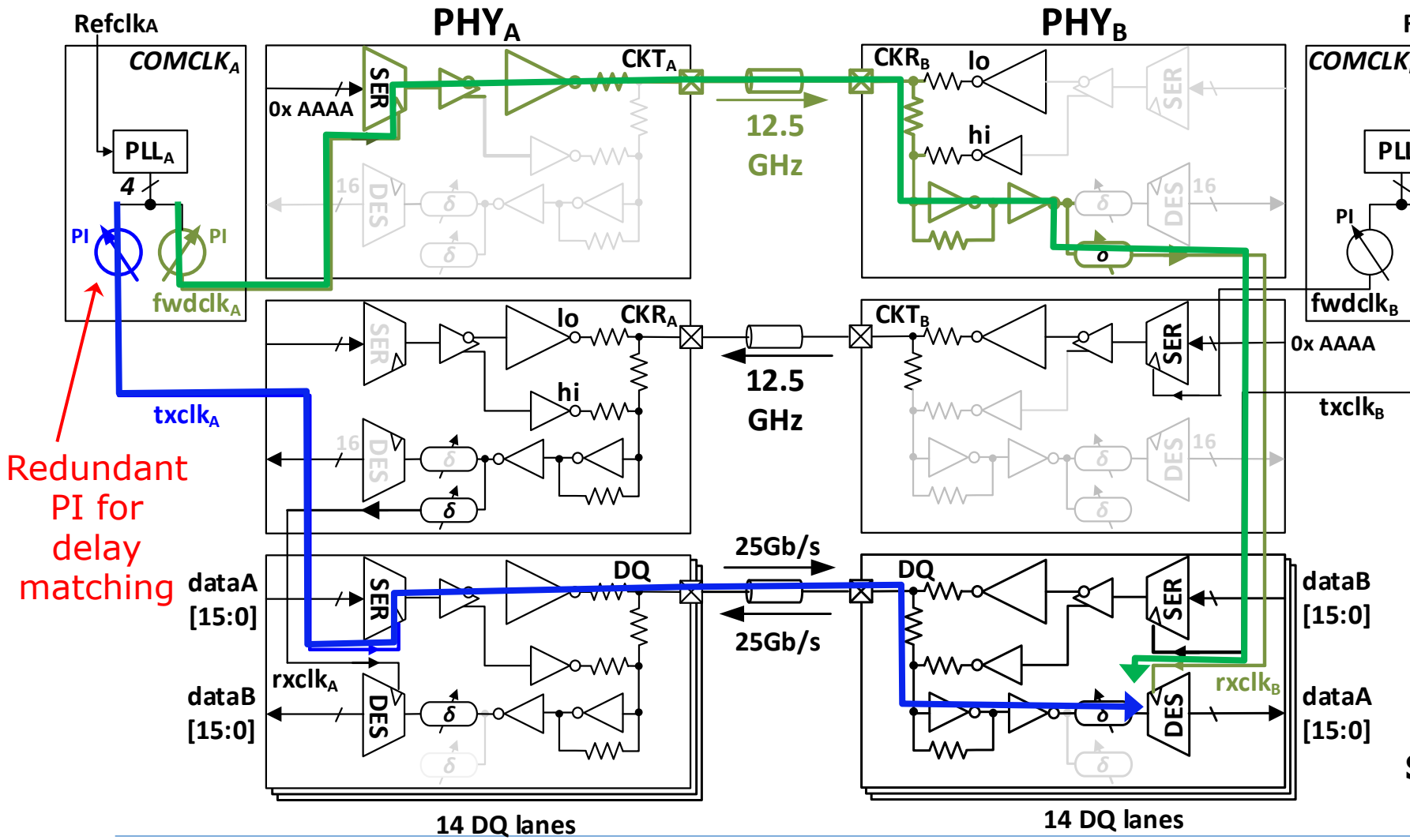


ISR-SBD PHY Architecture



- Clock-forwarding architecture
- 1 PLL, 1 CKT, 1 CKR and 14 DQ
- All 16 lanes are identical
- Data & Clk use near-identical paths

ISR-SBD: Delay Matching



Post-layout Simulation Total Delay @SS 0°C

	0.75V	0.95V
CLK	410.2ps	307.2ps
DATA	387.7ps	289.2ps
Δ	22.5ps	18.1ps

0.75V => 0.95V: -4.4ps

Sufficient to budget 2.2ps for 100mV supply change.

ISR-SBD: Delay Matching: Temp-drift

Total delay Post-layout Simulation

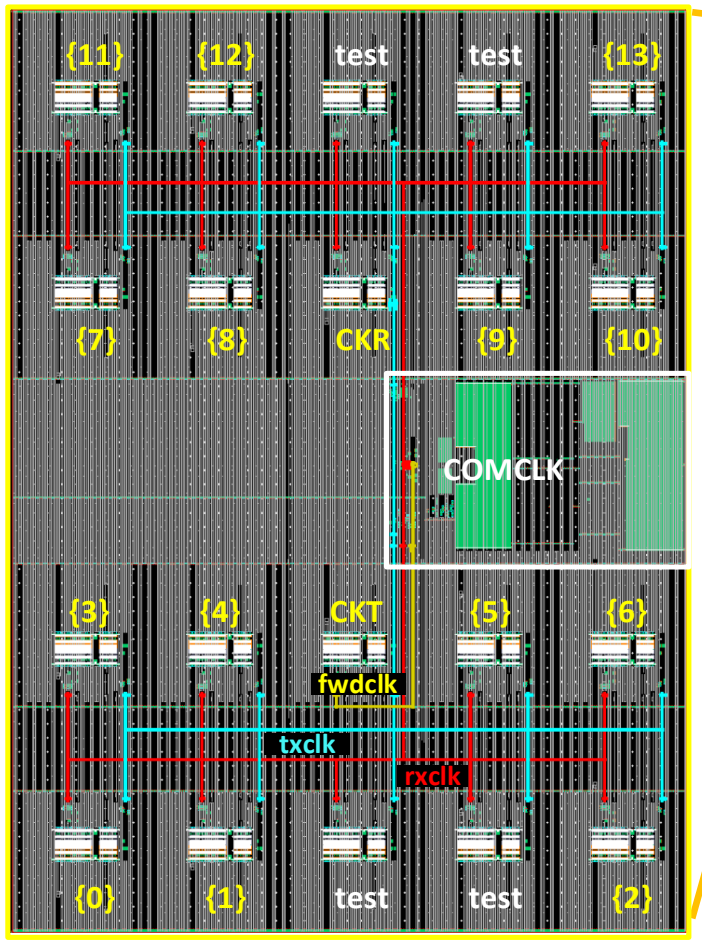
Process/Voltage	SS/0.75V		FF/0.75V	
Temperature	0°C	105°C	0°C	105°C
CLK	410.2ps	397.1ps	285.6ps	293.6ps
DATA	387.7ps	375.3ps	270.4ps	278.3ps
Δ	22.5ps	21.8ps	15.2ps	15.3ps

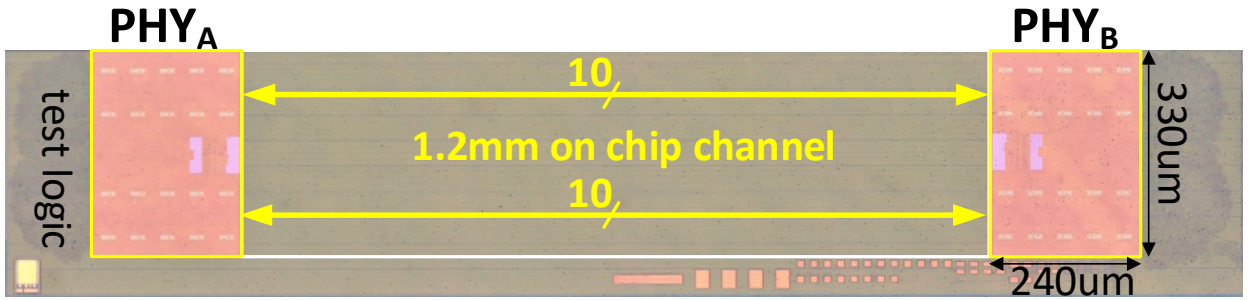
0.7ps **0.1ps**

Phase Interpolator calibration can be done just one time, at any temperature.

Test-site Micrograph

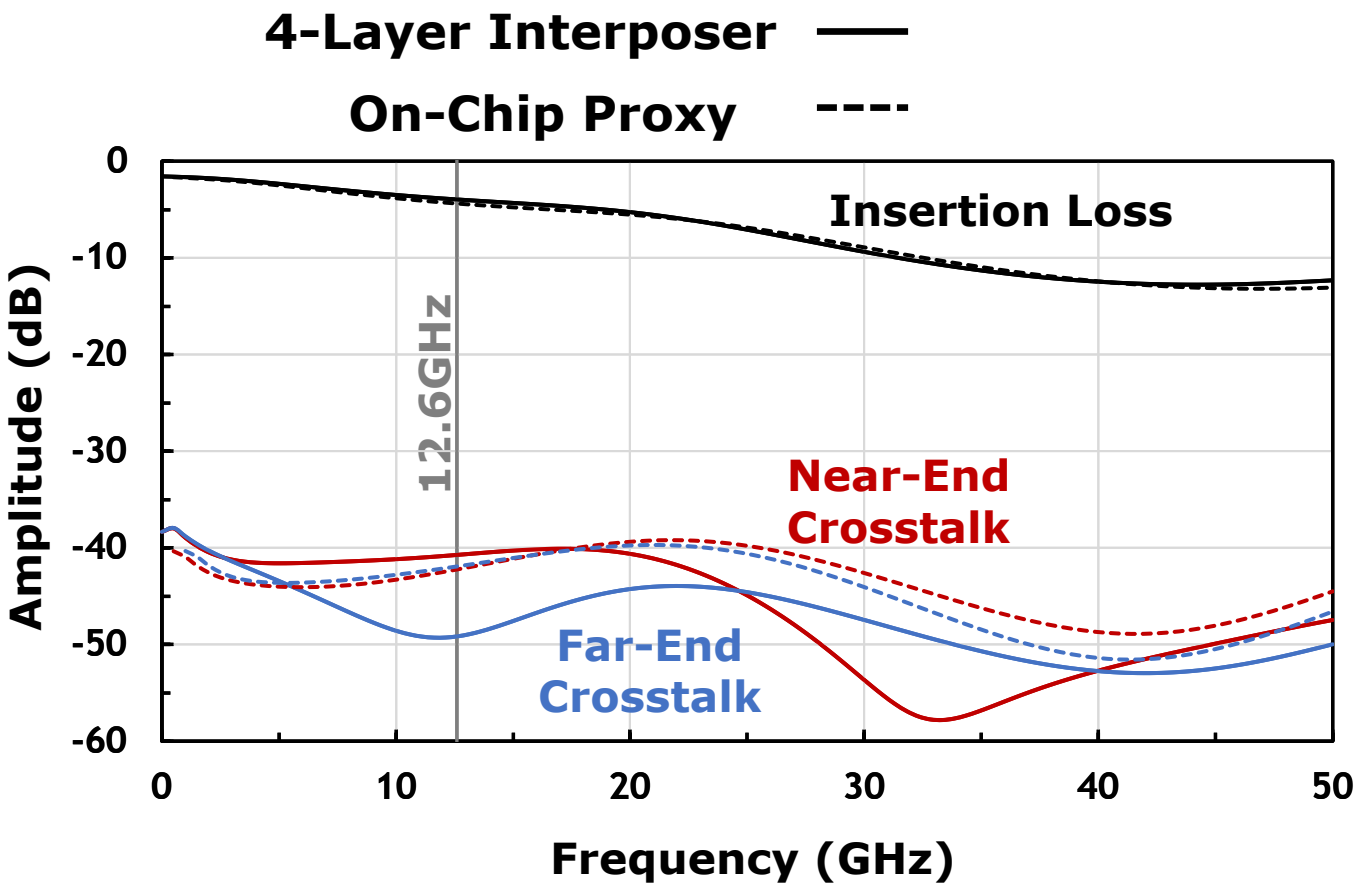
240 x 330um





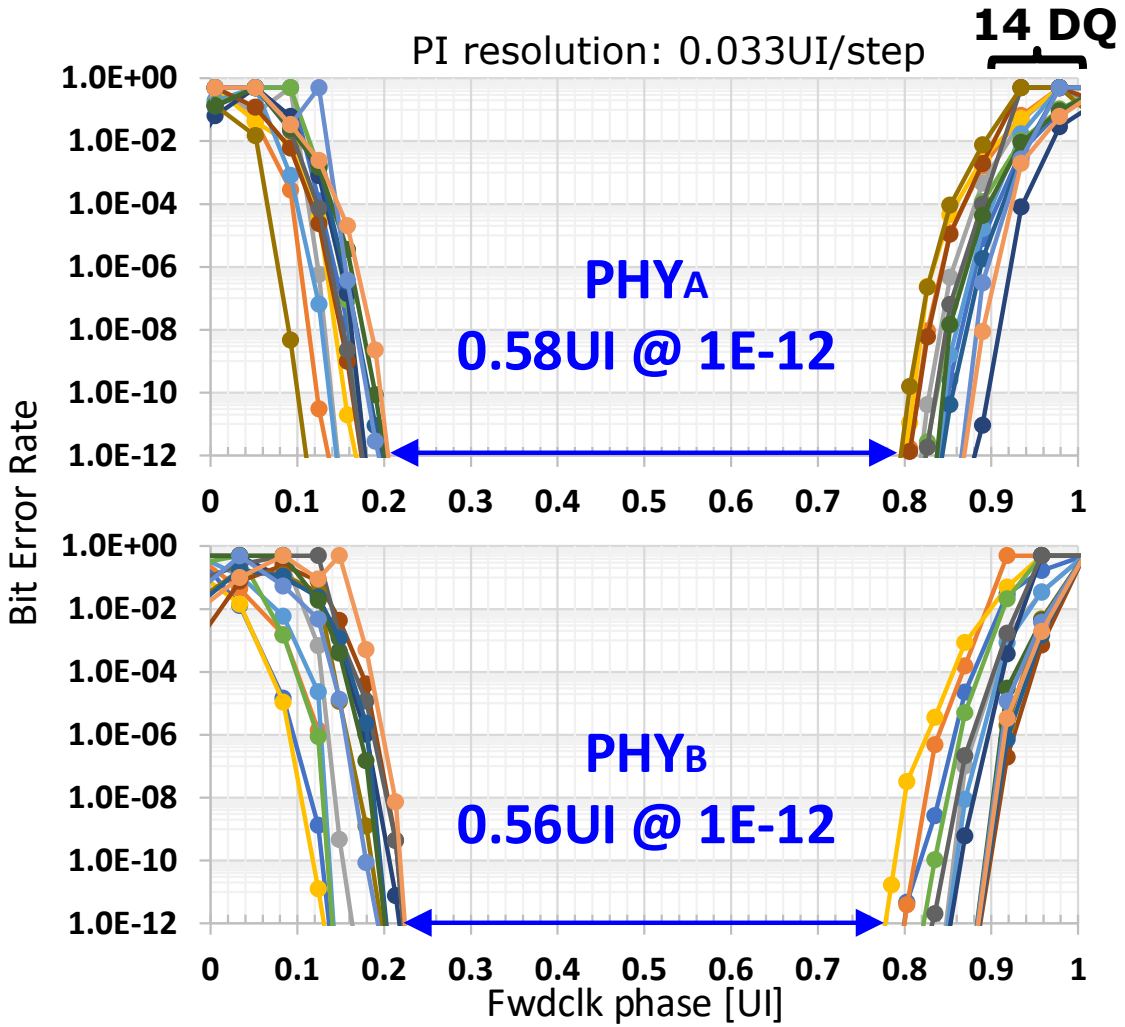
- A test-site on a 5nm production chip under 0.75V core logic supply.
- Two ISR-SBD PHY instances placed 1.2mm apart
- On-chip channel closely emulates a target interposer channel

Channel Response



@12.6GHz		Target Interposer	On-Chip Proxy
Channel Impedance	R	18.1Ω/mm	17.7Ω/mm
	L	364.8pH/mm	339pH/mm
	C	198.2fF/mm	210fF/mm
	Z_o	44.8Ω	42.1Ω
Channel Signal Integrity	IL	-4.0dB	-4.5dB
	FEXT	-49.1dB	-42.2dB
	NEXT	-40.7dB	-41.7dB

Results at 50.4Gbps/wire



- 25.2Gbps/direction, 50.4Gbps/wire
- Using the slope of BER curves, Rj is calculated as 0.75ps rms.
- **Eye Margin@1E-25 is extrapolated as 0.45UI(PHYA) & 0.43UI(PHYB)**

Power Consumption at 50.4Gbps/wire

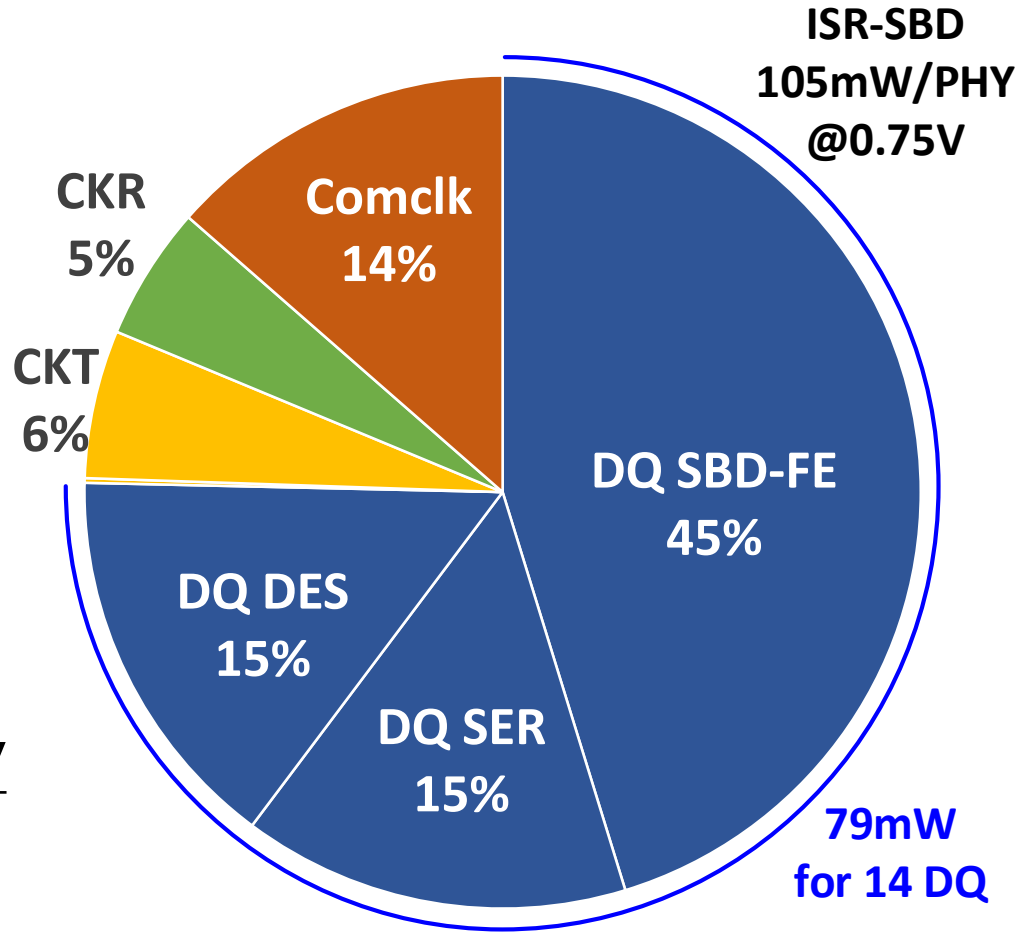
- Power Consumption and breakdown for ***one side*** of a 14-DQ ISR-SBD PHY

- Power Consumption

  105mW/PHY

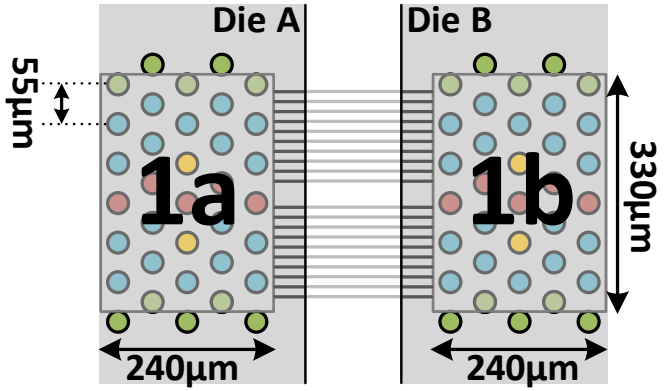
- **Overall Link Energy Efficiency**

  0.297pJ/bit



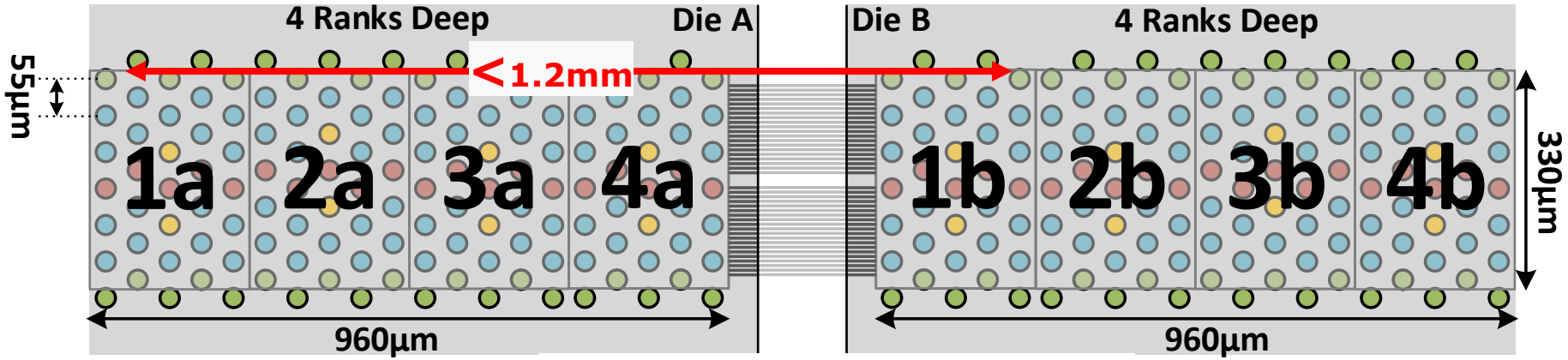
Application to Multi-Rank System

Common µbump pitch used by HBM



18DQ & 2 Clocks on 1 Layer

- 14DQ: 2.14Tb/s/mm
- 18DQ: 2.75Tb/s/mm
- 72DQ: 11.0Tb/s/mm



72DQ & 8 Clocks on 4 Layers

72DQ @ 40um pitch provides >15Tb/s/mm

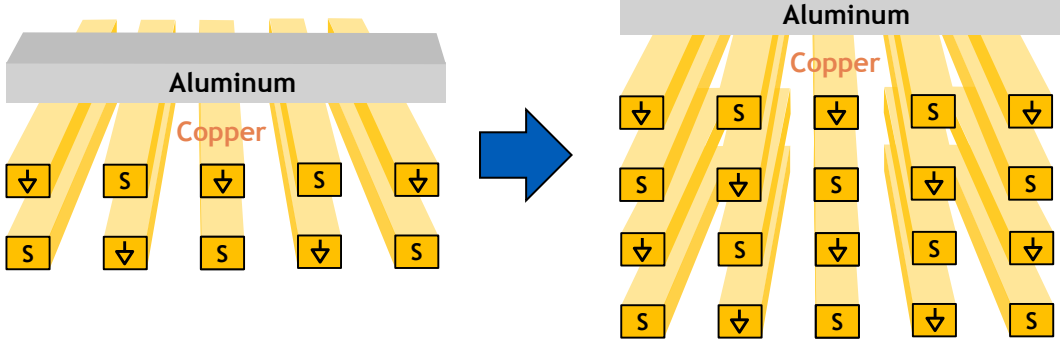
Conclusion

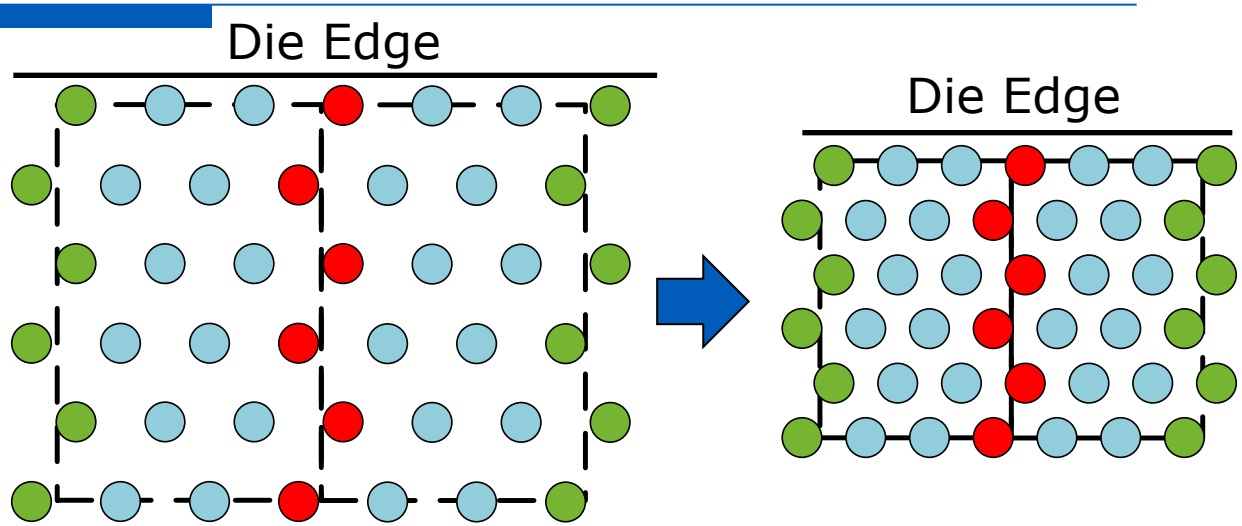
- Energy-Cost of Communication
  - Tax on computational performance … and it's all that matters
- Limitations to Off-Chip & Off-Package Bandwidth
- Single-Ended Signaling Methods for Scaling Chip-to-Chip Bandwidth
  - Organic Package (MCM) and PCB Applications using Ground-Referenced Signaling
  - Interposer Short-Reach Signaling using Simultaneous Bidirectional Signaling
- Plenty of room to scale chip-to-chip bandwidth for 2.5D systems
  - Main challenge is reducing power consumption of data movement

References

- J. Poulton, J. Wilson, W. Turner, B. Zimmer, X. Chen, S. Kudva, S. Song, S. Tell, N. Nedovic, W. Zhao, S. Sudhakaran, C. T. Gray, W. Dally, *"A 1.17-pJ/bit 25-Gb/s/pin Ground-Referenced Single-Ended Serial Link for Off- and On-Package Communication Using a Process- and Temperature-Adaptive Voltage Regulator", IEEE Journal of Solid-State Circuits, Vol. 54, Issue 1, Digital Object Identifier: 10.1109/JSSC.2018.2875092, Jan. 2018.*
- W. Dally, T Gray, J. Poulton, B. Khailany, J. Wilson, L. Dennison, *"Hardware-Enabled Artificial Intelligence",* Symposium on VLSI Circuits, IEEE, June 18-22, 2018. (Plenary Talk)
- J. Wilson**,** W. Turner, J. Poulton, B. Zimmer, X. Chen, S. Kudva, S. Song, S. Tell, N. Nedovic, W. Zhao, S. Sudhakaran, C. T. Gray, W. Dally, *"A 1.17pJ/bit 25Gb/s/pin Ground Referenced Single-Ended Serial Link for Off- and On-Package Communication in 16nm CMOS Using a Process- and Temperature-Adaptive Voltage Regulator", IEEE* International Solid-State Circuits Conference, ISSCC'18, Vol. 61, pp. 276-277, Paper 16.8, Feb. 2018.
- J. Wilson, M. Fojtik, J. Poulton, X. Chen, S. Tell, T. Greer, C. T. Gray, W. Dally, *"A 6.5-to-23.3fJ/bit/mm Balanced Charge-Recycling Bus in 16nm FinNET CMOS at 1.7-to-2.6Gb/s/wire with Clock Forwarding and Low-Crosstalk Contraflow Wiring", IEEE* International Solid-State Circuits Conference, ISSCC'16, Vol. 59, pp. 156-157, Paper 8.6, 2016.
- J. Poulton, W. Dally, X. Chen, J. Eyles, T. Greer, S. Tell, J. Wilson, C. T. Gray, *"A 0.54 pJ/bit 20 Gb/s Ground-Referenced Single-Ended Short-Reach Serial Link in 28nm CMOS for Advanced Packaging Applications",* IEEE Journal of Solid-State Circuits, Vol. 48, Issue 12, pp 3206-3218, 2013.
- Y. Nishi, J. Poulton, X. Chen, S. Song, B. Zimmer, W. Turner, S. Tell, N. Nedovic, J. Wilson, W. Dally, C. T. Gray, "A 0.297-pJ/bit 50.4-Gb/s/wire Inverter-Based Short-Reach Simultaneous Bidirectional Transceiver for Die-to-Die Interface in 5nm CMOS", IEEE Symposium on VLSI Circuits, pp. 154-155, June 2022.
- Y-Y. Hsu, P-C. Kuo, C-L. Chuang, P-H. Chang, H-H. Shen, C-F. Chiang, "A 7nm 0.46pJ/bit 20Gbps with BER 1E-25 Die-to-Die Link Using Minimum Intrinsic Auto Alignment and Noise-Immunity Encode", IEEE Symposium on VLSI Circuits, pp. 1-2, June 2021.
- M-S. Lin, T-C. Huang, C-C. Tsai, K-H. Tam, C-H. Hsieh, T. Chen, W-H. Huang, J. Hu, Y-C. Chen, S. K. Goel, C-M. Fu, S. Rusu, C-C. Li, S-Y. Yang, M. Wong, S-C. Yang, F. Lee, "A 7nm 4GHz Arm®-core-based CoWoS® Chiplet Design for High Performance Computing", IEEE Symposium on VLSI Circuits, pp. C28-C29, June 2019.
- B. Dehlaghi, A. C. Carusone, "A 0.3 pJ/bit 20 Gb/s/Wire Parallel Interface for Die-to-Die Communication", IEEE Journal of Solid-State Circuits, vol. 51, no. 11, pp. 2690-2701, Nov. 2016.

Routing Layers & Bump Pitch





Double Number of Routing Layers

- Instant 2x increase in die-edge BW
- Maintain per-channel signal-rates
- Higher cost interposer
- Doubles the I/O bump array height
- Limitations from large vias *(InFO or CoWoS-R)*

Reduce I/O µBump Pitch

- $\sqrt{2}$ pitch reduction doubles BW density
- Maintain per-channel signal-rates
- Optimum is to improve both µBump and routing pitches
- Area available for PHY becomes a concern along with thermal and power delivery

Comparison

	Our work	Y-Y Hsu VLSI21	M-S Lin VLSI19	B.Dehlaghi JSSC16
Technology	5nm	7nm	7nm	28nm
µbump pitch	55µm*	40µm ✓	40µm ✓	100µm
Interposer Channel	1.2mm**	1.0mm	0.5mm	2.5mm ✓
Supply[V]	0.75 ✓	0.8	0.8,0.3	NA
Data Rate/wire [Gb/s]	50.4 (SBD) ✓	20 (NRZ)	8 (NRZ)	20 (NRZ)
Energy Efficiency [pJ/bit]	0.297 ✓	0.46	0.56	0.3***
Areal Density [Tb/s/mm^2]	4.45 ✓	2.25	0.8	NA
Edge Density [Tb/s/mm]	2.14	5.31 ✓	0.67	NA

** Target application*
*** On-chip replica channel*

****Clocking not included*

Comparison (18-DQ, 4 ranks)

	Our work	Y-Y Hsu VLSI21	M-S Lin VLSI19	B.Dehlaghi JSSC16
Technology	5nm	7nm	7nm	28nm
µbump pitch	55µm*	40µm ✓	40µm ✓	100µm
Interposer Channel	1.2mm**	1.0mm	0.5mm	2.5mm ✓
Supply[V]	0.75 ✓	0.8	0.8,0.3	NA
Data Rate/wire [Gb/s]	50.4 (SBD) ✓	20 (NRZ)	8 (NRZ)	20 (NRZ)
Energy Efficiency [pJ/bit]	0.281 ✓	0.46	0.56	0.3***
Areal Density [Tb/s/mm^2]	5.73 ✓	2.25	0.8	NA
Edge Density [Tb/s/mm]	11.0 ✓	5.31	0.67	NA
Edge Density [Tb/s/mm]	2.14	5.31	0.67	NA

** Target application*
*** On-chip replica channel*
18-DQ, 4-rank extrapolation

****Clocking not included*