Ethernet For AI: Progress and Challenges

OPEN Compute Project®

VER: 010

ORIG: 012025

REV: 03172025

Bijan Nowroozi

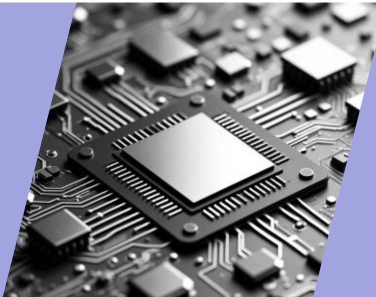
CTO

LIASON INFORMATION

Impacting Data Center IT and Facilities

From hyperscale to distributed edge data center technologies

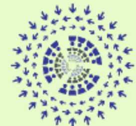
From silicon to systems to physical facilities

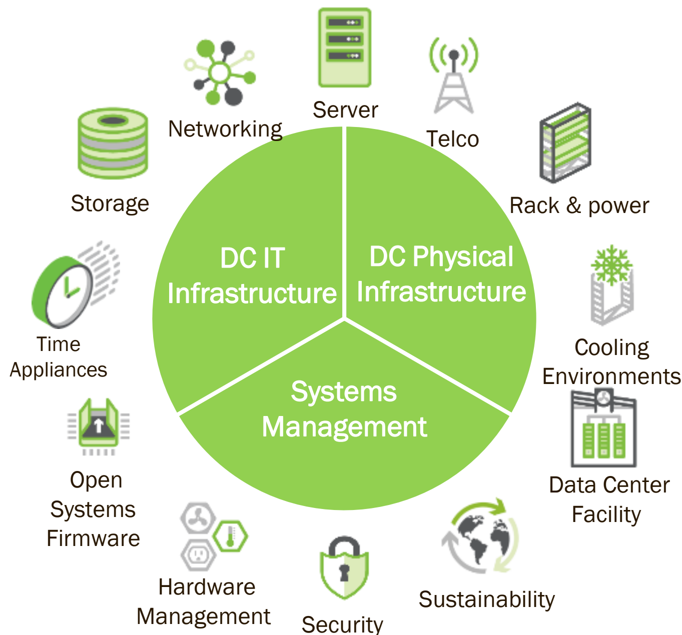


OPEN Compute Project®

OCP: A Collaborative Community Effort

INCUBATOR	COMMUNITY PROJECTS	NEW INITIATIVES	EMERGING MARKETS





TECHNOLOGY

Open Systems for AI

Evenstar OpenRAN

Test & Validation

MARKETS

Photonics

Chiplets

OCP MARKETPLACE

RECOGNITION PROGRAMS

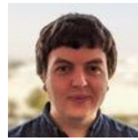
Security I Facilities

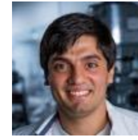
OCP S.A.F.E. OCP READY™

Product

OCP ACCEPTED OCP INSPIRED™

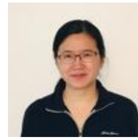
Open Compute Project®



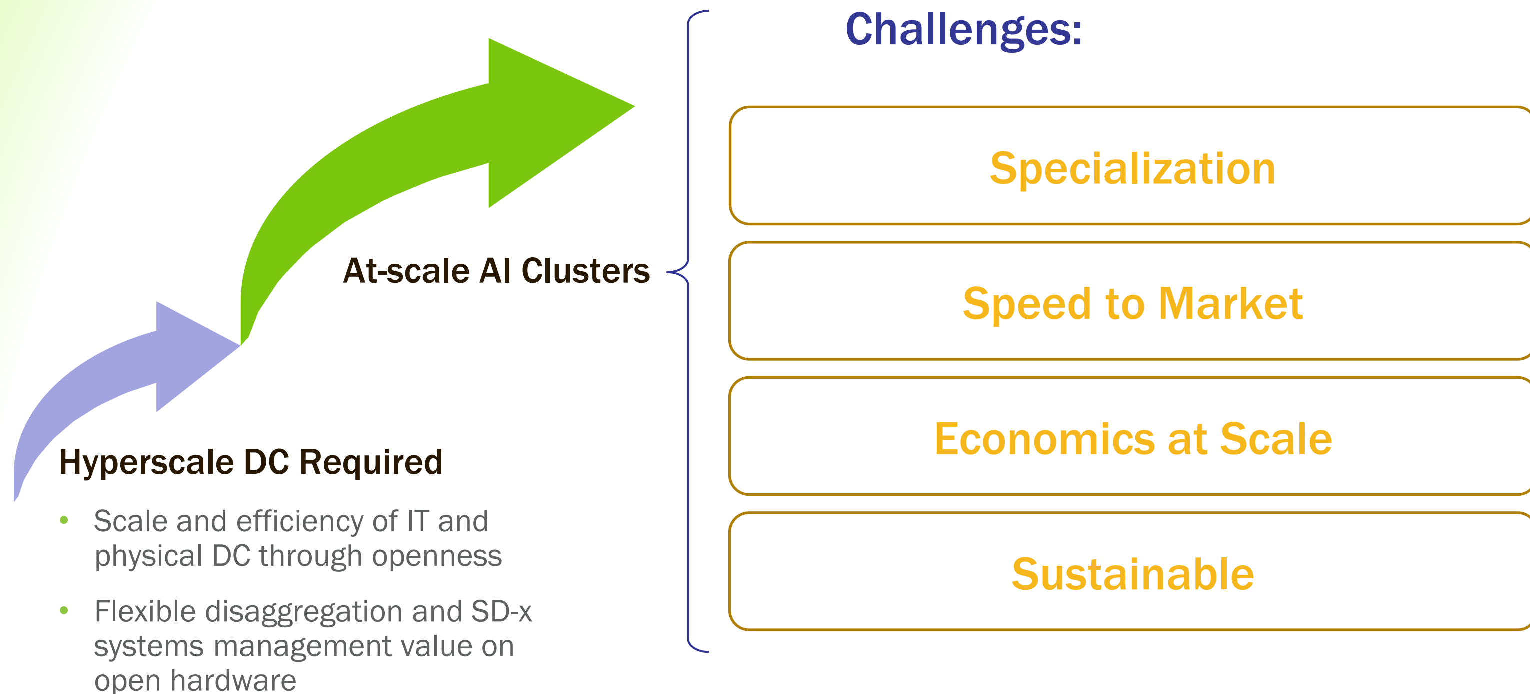


250+ Project Leaders

7,000+ Participants



Tackling the Next DC Technology Shift

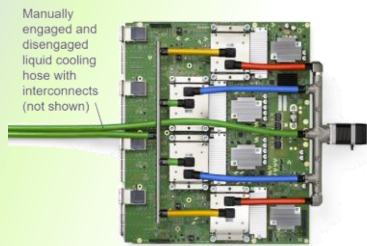


- Scale and efficiency of IT and physical DC through openness
- Flexible disaggregation and SD-x systems management value on open hardware

OPEN Compute Project®

OCP Community: AI Progress

Manually engaged and disengaged liquid cooling hose with interconnects (not shown)



(a) Google TPU (ML) version 4 (2021) Liquid cooled Google server

Blind mate liquid cooling interconnects

(b) Nvidia GB200 NV... Liquid Cooled Blind ...

Announcing GB200 NVL72 Design Contributions to O...

Accelerating open innovation of AI data...

Key Contributions

NVLINK Cable Cartridge — 5,000 NVLink cables

Enhanced Bus Bar — 1,400-amp capacity

Floating Blind Mate — Reduces breakage

Blind Mate Slide Rails — NVLink and LC connections

Rack Reinforcements — +100lb. of steel support

Rear Rack Extenders — Protects manifold fittings

Compute Tray Form Factor — 1U LC: 4 GPUs

Compute Tray DC-SCM — Fits 1RU form factor

Switch Tray Form Factor — 1U LC: 550 NVLINK Lanes

Switch Tray Bus Bar Connector — Narrower 35mm width

Catalina

ORV3 High Power Rack

- Liquid Cooled FLC Compatible
- Retool Orv3 For Higher Power
- Wedge400 Fabric Switch
- Rack Management
- Modular Design

Compute (FP8)	2.7x
HBM Capacity	+2X
HBM BW	2.5X
Scale Up BW	2x
Scale Out BW	2x
Power Density	3-4x

EXPLORE OUR MARKETPLACE SEGMENTS

INFO, SI, OPEN SYSTEMS FOR AI: BLUEPRINT FOR SCALABLE INFRASTRUCTURE

OPEN SYSTEMS FOR AI: BLUEPRINT FOR SCALABLE INFRASTRUCTURE

Contributor: Open Compute Project, Google, Intel, NVIDIA, Denvr Dataworks, Meta, Supermicro, AMD, Dell Technologies

Family: Information

Project: Strategic Initiatives > Open Systems for AI

Release Year: 2025

OPEN Compute Project®

OCP Community is several generations into building AI technology and systems and now is collaboratively working towards OCP Open Systems for AI

Why are we here today?

OCP Community progress and experience building AI Systems has yielded data and feedback from subject matter experts. The feedback to this point has pointed us to common problem areas.

Key feedback points so far:

- Performance is a bottle neck to realizing end user demand
- Technical roadmap/path forward unclear in several important areas
- Risk rising as scale and demand increases
- Economics are not optimal

Early Data Points

- End users: must continue to increase AI intelligence and availability
- Vendors: Increase AI cluster throughput to drive lower unit costs

Multi-task Language Understanding on MMLU

AVERAGE (%)

100
80
60
40
20

Jul '19 Jan '20 Jul '20 Jan '21 Jul '21 Jan '22 Jul '22 Jan '23 Jul '23 Jan '24 Jul '24 Jan '25

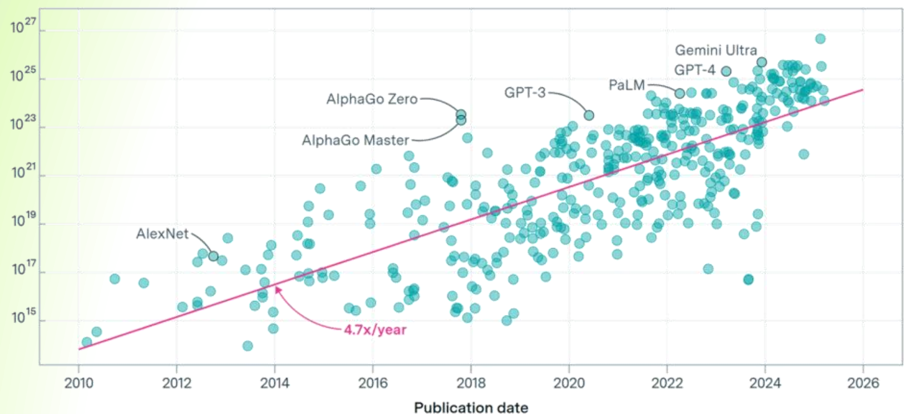
GPT-2-XL 1.5B (fine-tuned)
RoBERTa-base 125M (fine-tuned)
GPT 3
chatgpt/gpt3.5(20B)
LLaMA 65B (fine-tuned)
Claude Instant 1.1 (5-shot)
Qwen1.5 72B (5-shot)
Qwen2-72B-Instruct
ds-r1(671b)

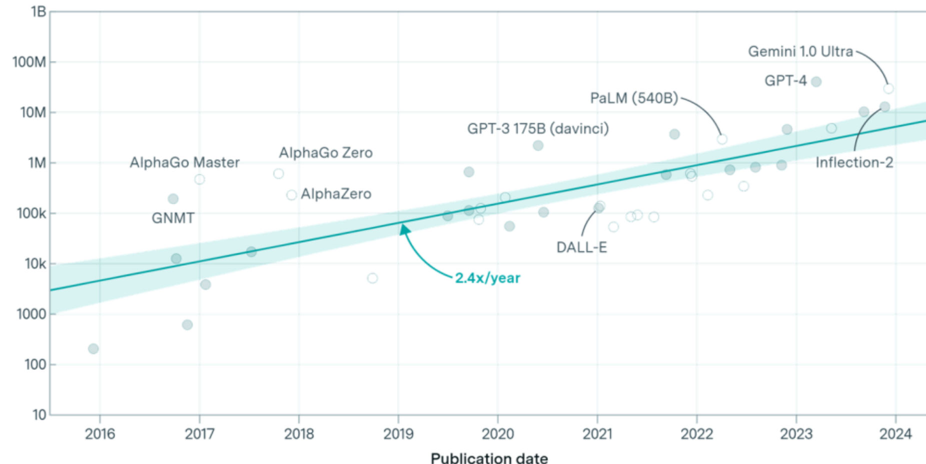
Other models   Models with highest Average (%)

Steady Improvement in intelligence!

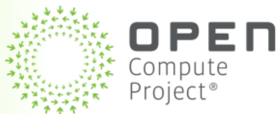
Multi-task Language Understanding on MMLU https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu

Gen AI LLM (Model) Resource Trends





Compute requirements rising at 4.7x/yr and Costs by 2.4x/year



"Data Insights" from https://epoch.ai/data/notable-ai-models

Extrapolating AI Training Cluster Demands

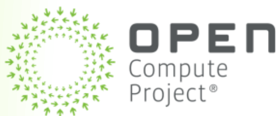
Largest AI Clusters

Year	OOMs	# of H100s-equivalent	Cost	Power	Power reference class
2022	~GPT-4 cluster	~10k	~$500M	~10 MW	~10,000 average homes
~2024	+1 OOM	~100k	$billions	~100MW	~100,000 homes
~2026	+2 OOMs	~1M	$10s of billions	~1 GW	The Hoover Dam, or a large nuclear reactor
~2028	+3 OOMs	~10M	$100s of billions	~10 GW	A small/medium US state
~2030	+4 OOMs	~100M	$1T+	~100GW	>20% of US electricity production

Scaling the largest training clusters, rough back-of-the-envelope calculations.

OOM = order of magnitude, 10x = 1 order of magnitude. Roughly **~0.5 OOMs/year** trend growth of AI training compute

Scale everything!



SITUATIONAL AWARENESS: The Decade Ahead
Leopold Aschenbrenner, June 2024 https://situational-awareness.ai/

Seeking OOMs Increase in Compute

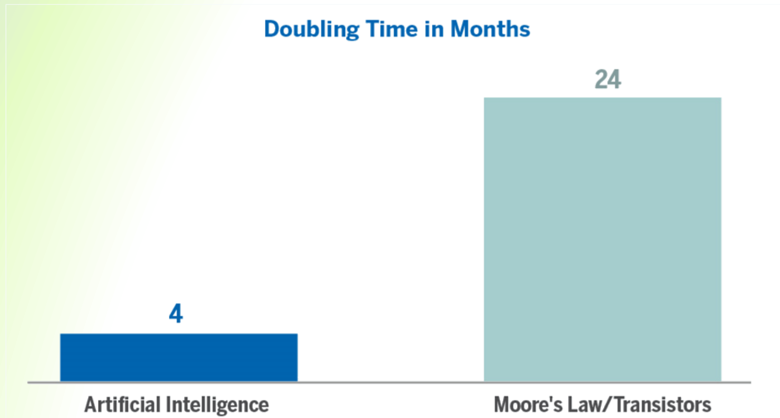
Non exhaustive, focusing on Existing ISAs:

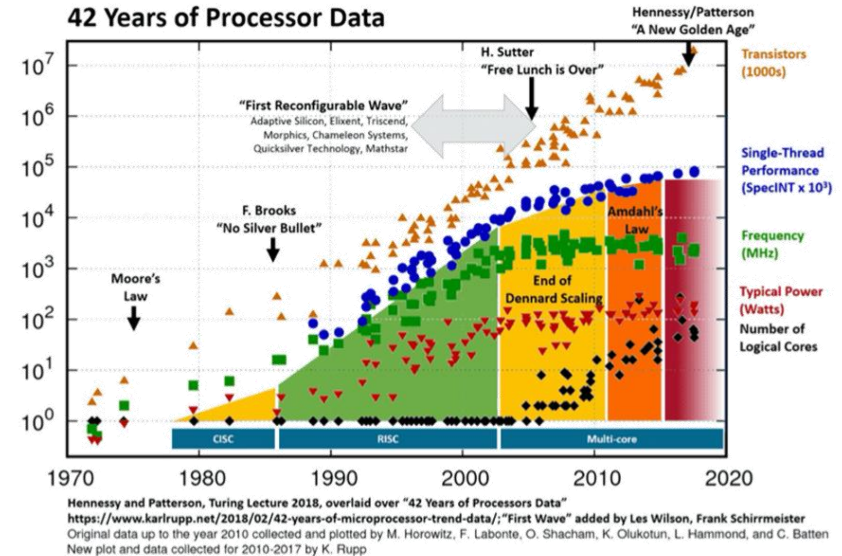
- **More pipeline throughput** add logic/data path width via increasing transistor count (Moore)
- **Speed Up** increase the clock speeds of the ICs. (assuming caches are optimized).
- **Parallelism** distribute compute in various ways.

Note other approaches such as addressing the memory wall (ex: CXL) are valid in improving performance but don't accelerate compute by an OOM.

** Note: these are often overloaded terms*

Improve with Processor/IC Scaling?





Moore's Law: compute doubles for constant cost roughly once every two years. Quantum effects are beginning to interfere in electronic devices as they shrink.

Dennard Scaling: "Power Wall" has limited practical processor frequency to around 4 -6 GHz since 2006.

IC process scaling for logic is not going to help much. Effective gain (step function) has been realized by adoption of HBM this is data path acceleration.

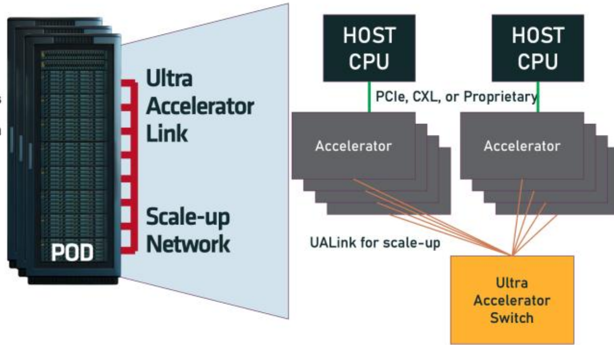
Dennard chart, text from 2023 OCP Keynote, "Quantum Computing – What, why, and what for?" Dr. Angie Qarry

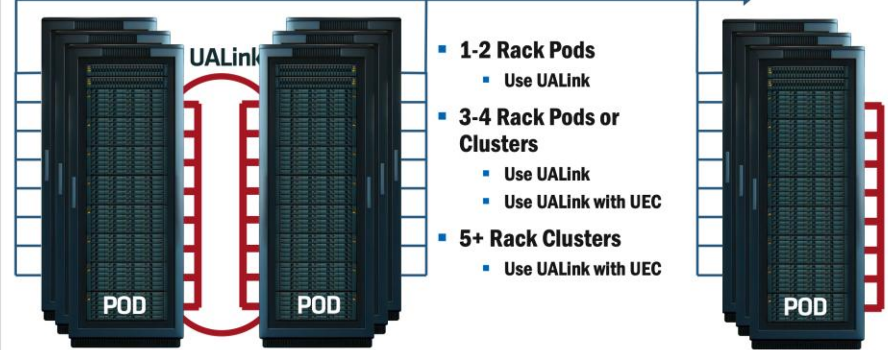
AI Scaling With Networking?

Parallelism with processor-to-processor links and the use of collectives provided a large gain. IEEE 802.3dj speeds (200Gb/s lane) FTW to unleash AI cluster performance.

OPEN Compute Project®

AI Scaling With Ethernet Fabrics

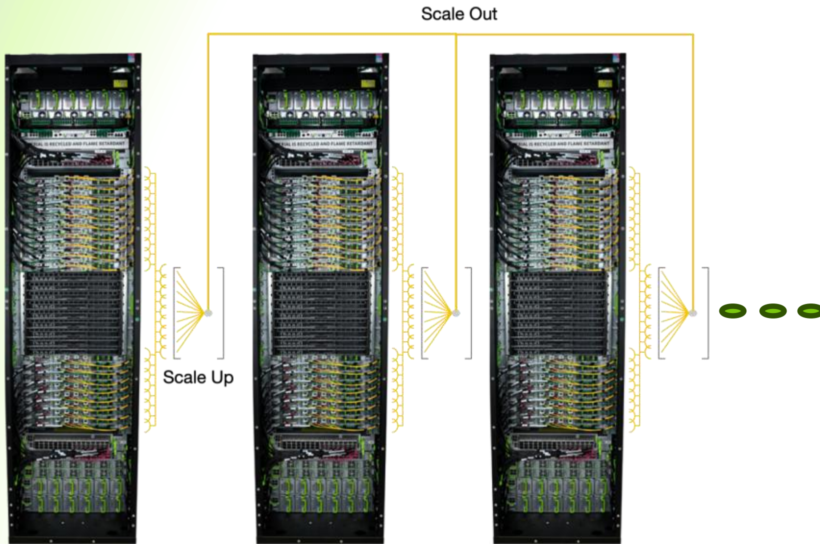


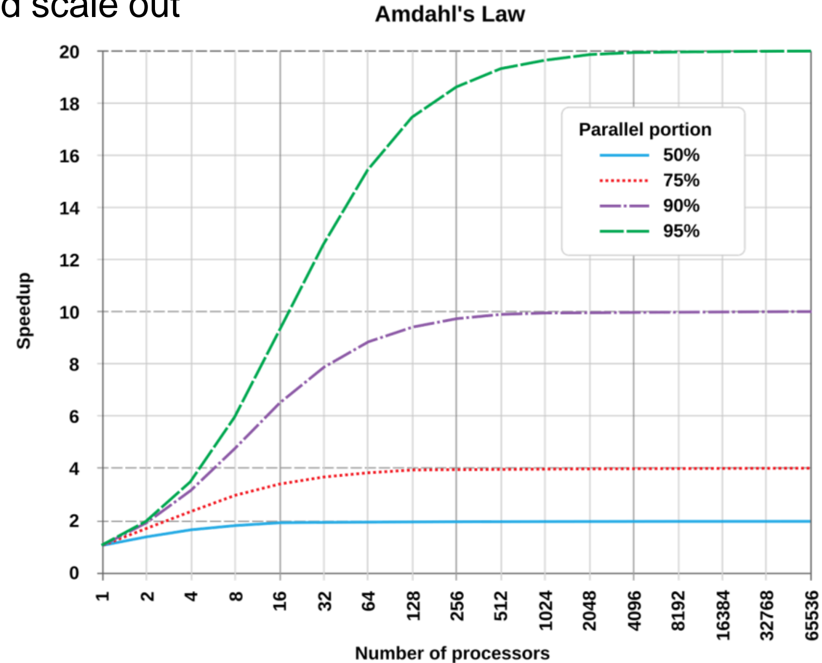


Scaling up via short reach networking to create an Ethernet fabric to interconnect multiple processors on a system tray, and multiple trays in a rack for a single ‘node’ or ‘pod’ via Ethernet switching. Scaling out is a function of meshing the node/pod switches for a cluster.

Networking for the next OOM Improvement

Problem 1: We are physically limited in scale up and scale out





Amdahl's Law shows 1024-2048 nodes max. Economically, deploying 2048 racks doesn't reduce costs and might require 100% utilization for adequate ROI on resources.

Diagram of Amdahl's Law from https://en.wikipedia.org/wiki/Amdahl%27s_law#/media/File:AmdahlsLaw.svg

Achieving OOM Improvement: 802.3dj

Where do we stand?

Advantages of scaling with 802.3dj

- Double throughput results in far greater xPU performance over 100Gb/s /lane
- Usable Cu cable lengths for AI Clusters
- Multilane to 1.6Tb/s
- Connectors not hard
- Coming soon!

Scaling Challenges

- Increasing to multilane is expensive Radix hit, power wise and thermally
- Max throughput (1.6Tb/s) doesn't match HBM rates (~*8Tb/s HBM3e)
- More lanes = more complexity to maintain

Moving to 200Gb/s is great step forward utilizing Ethernet in AI Clusters. Although this is entering the market in 2H 2025, AI performance demands more throughput to reduce bottlenecks.

Achieving OOM Improvement: 802.3 {NG}

What do we need?

++OOM Ethernet Fabric for Scaling

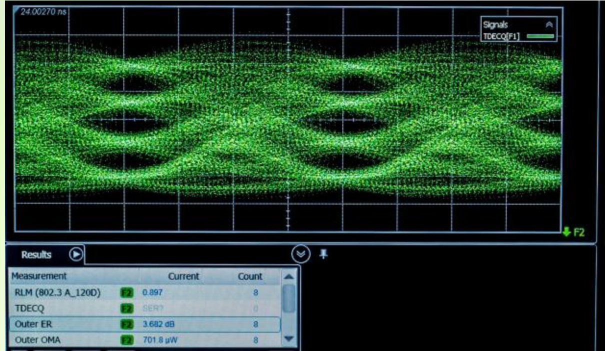
- *Minimum* 2X throughput from 802.3dj
- Cu not a hard requirement. Nice if reach from server tray to tray.
- Capability to reach multilane to 3.2Tb/s (eventually)
- TBD reasonably thermally performant/reliability connectors
- 18-24 months?

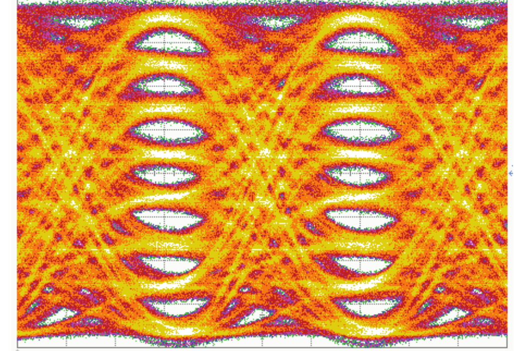
Scaling Challenges

- Faster modulation = greater complexity (FEC, etc…)
- Realistically requires faster SERDES (min 224Gb/s)
- Max throughput (3.2Tb/s) doesn't match HBM rates (~*16Tb/s HBM4)
- Uncertainty (Ex is 2X enough?)

Moving to 400Gb/s is great step forward utilizing Ethernet in AI Clusters. The power reduction will be helpful to 2025 workloads, however as scale increases, other/further optimizations may be highly desirable.

Next Steps





Development of use cases and Outreach

- Increase awareness, communicate the vision and identify users and use cases
- We are presently workshopping design for OCP Open Systems for AI. Additional details will help drive clarity on requirements
- Capture requirements from communities and build consensus
- Communicate and meet often!

OPEN Compute Project®

Thank You