A Path to 448Gbps

April 15, 2025

Tad Hofmeister
Platforms Infrastructure Engineering (PIE) Optics
Machine Learning, Systems and Cloud AI

Google Cloud
Proprietary

Acknowledgements

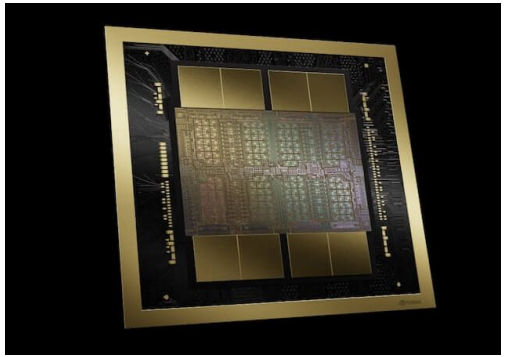
Many thanks to Hong Liu, Cedric Lam, Moray McLaren, and Leesa Noujeim for conversations and material to develop this presentation.

Outline

- Google Cloud ML/AI Clusters
- Why and Where 448Gbps
- Summary

Bigger Models; More Processing

- More powerful TPUs and GPUs recently announced:
  - Google Cloud Next 2025: Ironwood (TPUv7)
  - NVIDIA GTC 2025: Grace Blackwell Ultra, Vera Rubin



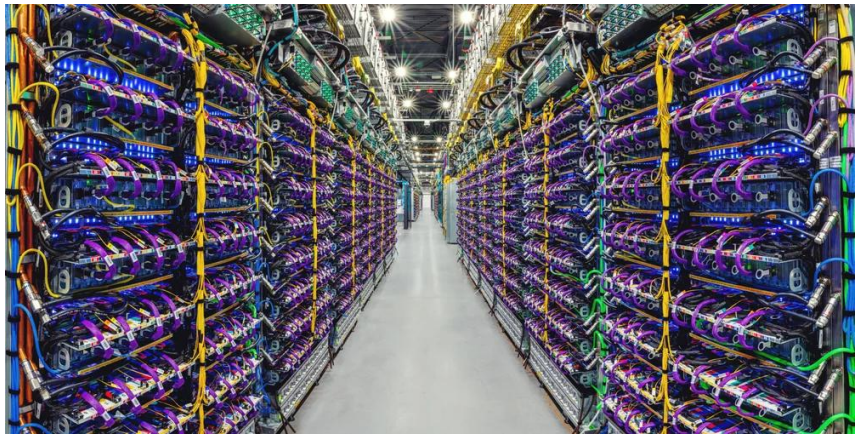
GB200 Superchip

Higher Capacity Interconnect

- As xPUs scale in performance, higher bandwidth interconnect between chips/packages is required
- LLM training of massive models is most efficient with 10s of 1000s of xPUs in a single cluster

Google Offers Both TPU and GPU Clusters

- 10s of 1,000s of xPUs in a cluster
  - Scale-out network across the cluster
  - Scale-up network: 1 to 2 orders of magnitude higher bandwidth than scale-out across a sub-set of xPUs
  - Both scale-up and scale-out terminate on the xPUs: challenging I/O.

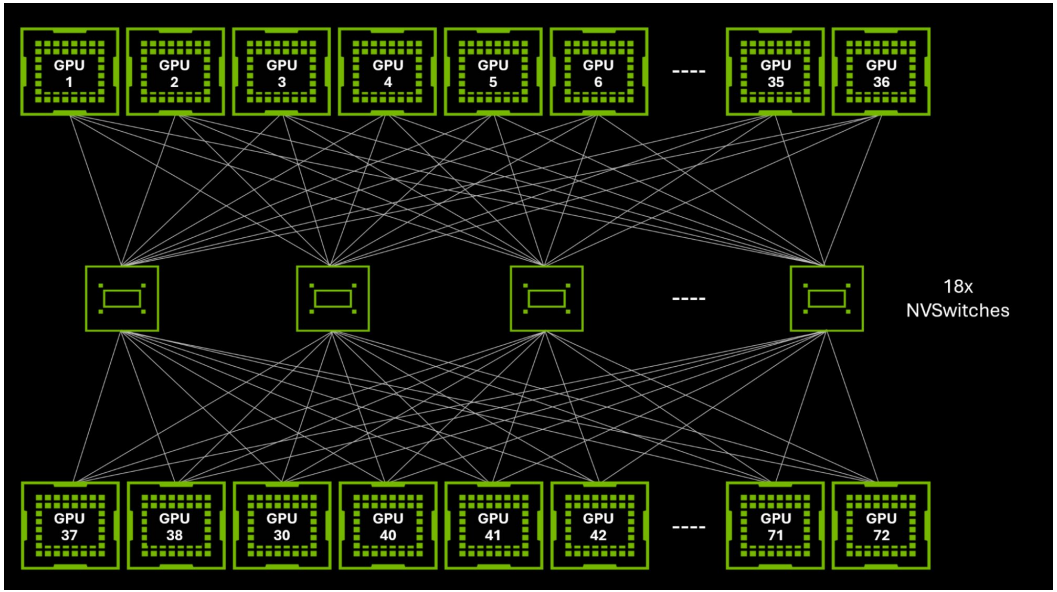


A row of TPU v5p AI accelerator supercomputers

https://blog.google/technology/ai/google-gemini-ai/

Scale Up Network: GPU

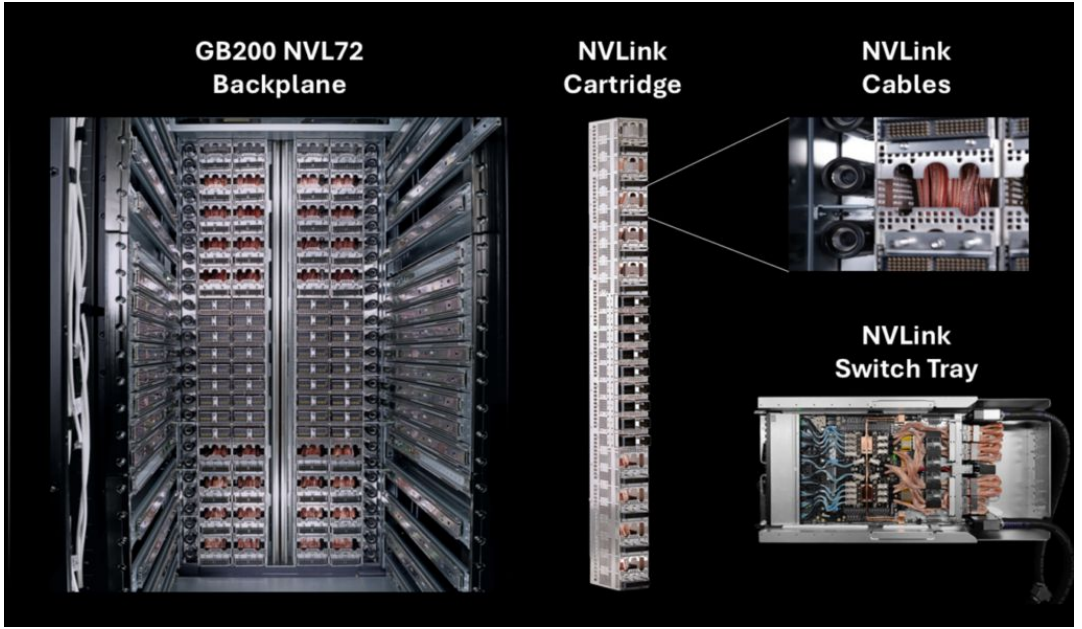
- nvLinks between 72 GPU chips and 18 NVSwitches in a rack
- Copper between GPU trays and switch trays



NVIDIA OCP Contribution

Scale Up Network: GPU

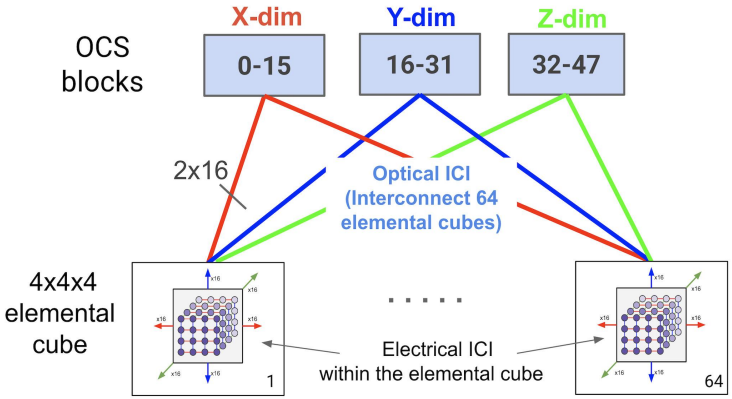
- nvLinks between 72 GPU chips and 18 NVSwitches in a rack
- Copper between GPU trays and switch trays
  - Lower power and lower cost, but limits the radius and number of GPUs
- At GTC 2025, announced NVL576 (144 chips, 576 Rubin GPU dies)
  - Also single rack, Copper interconnect



NVIDIA OCP Contribution, NVL72

Scale Up Network: TPU

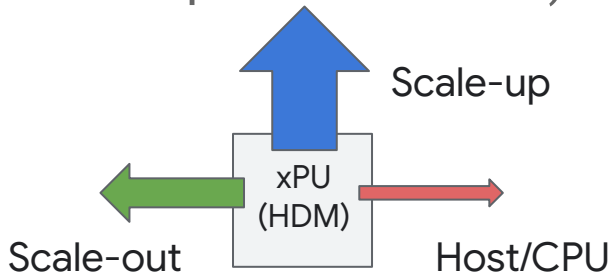
- ICI (Interchip Interconnect) directly between TPUs
  - No external switches in between
  - 3D Torus
  - 64 chips, intra-rack (cube) with Cu
  - Ironwood (7th Gen), up to 144 cubes, 9,216 chips in SuperPod with optics & OCS
- ICI is lower latency and lower power per port for the same bandwidth than with external switches
  - More deterministic traffic patterns are a better fit for ICI and OCS



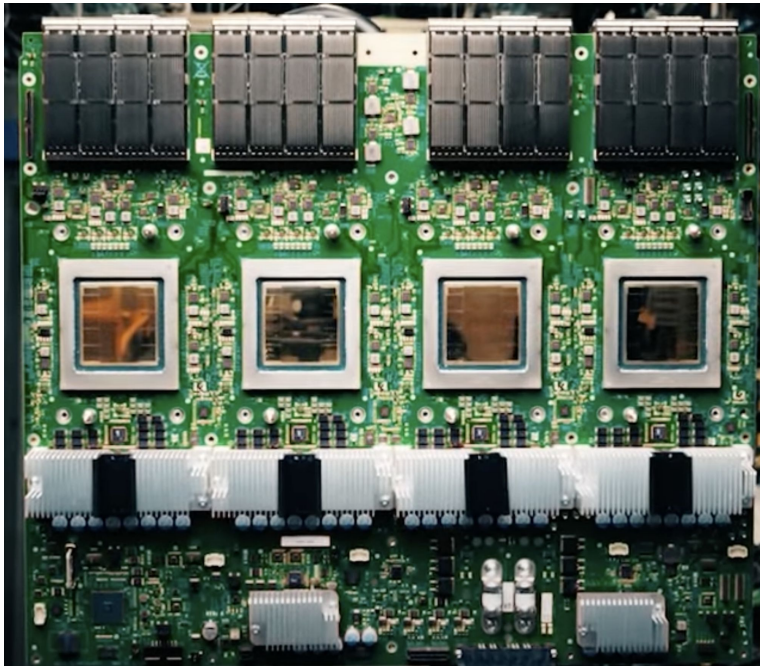
[TPU v4, ISCA'23]

ICI is Mix of Copper and Optics: OSFP Modules

TPU and GPU trays both have copper (for scale-up) and optics (for ICI scale-up and scale-out)



Higher Interconnect demand and limited package escape require higher speed IO
⇒ 448Gbps



Ironwood Tray
with 4 TPUs

Outline

- Google Cloud ML/AI Clusters
- Why and Where 448Gbps
- Summary

xPU Tray Channel Applications

Chip to chip

Chip to module

Module may be optical or copper.

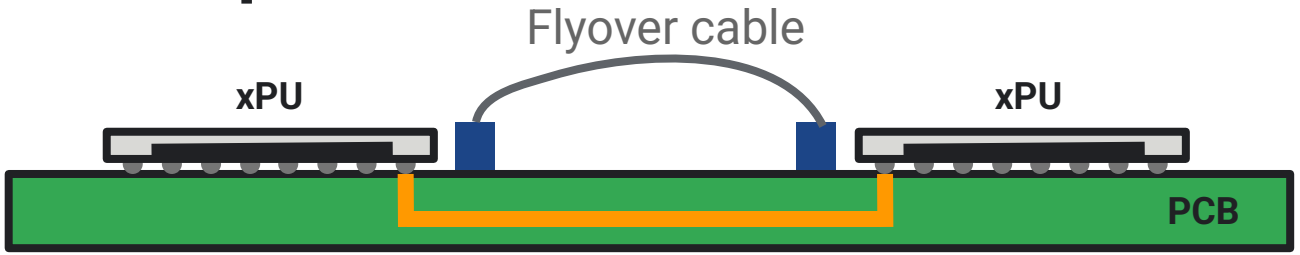
448Gbps SERDES I/O Options

	PAM4 (2 bits/symbol)	PAM6 (2.5 bits/symbol)	PAM8 (3 bits/symbol)	PAM12 (3.5 bits/symbol)
Nyquist frequency @425Gb/s (GHz)	106	85	70	60
SNR requirement for 1E-3 SER (dB)	17	20.8	23.5	27
SNR requirement for 1E-4 SER (dB)	18.8	22.2	25	28.5
SNR requirement for 1E-6 SER (dB)	20.6	24.5	27	30.5

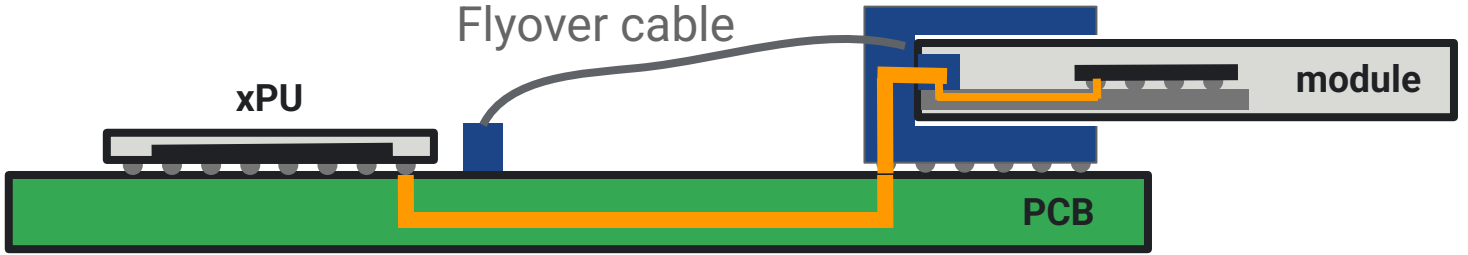
- The bandwidth requirement of chip to module channel is too high for PAM4. Optical channel (fiber) is OK for PAM4 optical.
- The SNR requirement for PAM12 is too stringent, challenging for xtalk and other channel impairments.

System Channel Improvements

Chip to chip

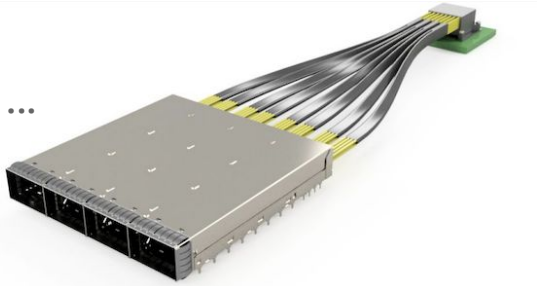


Chip to module

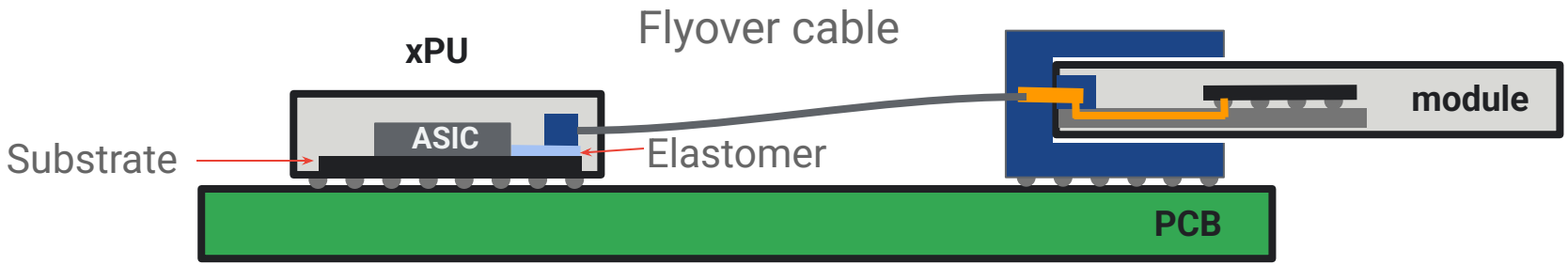


PCB loss, connector stub discontinuities, package ball, vias, …

Will need flyover cables, but still very challenging.



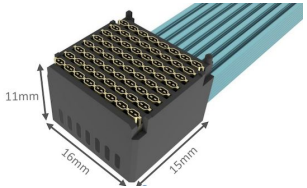
Co-Packaged Copper (CPC)

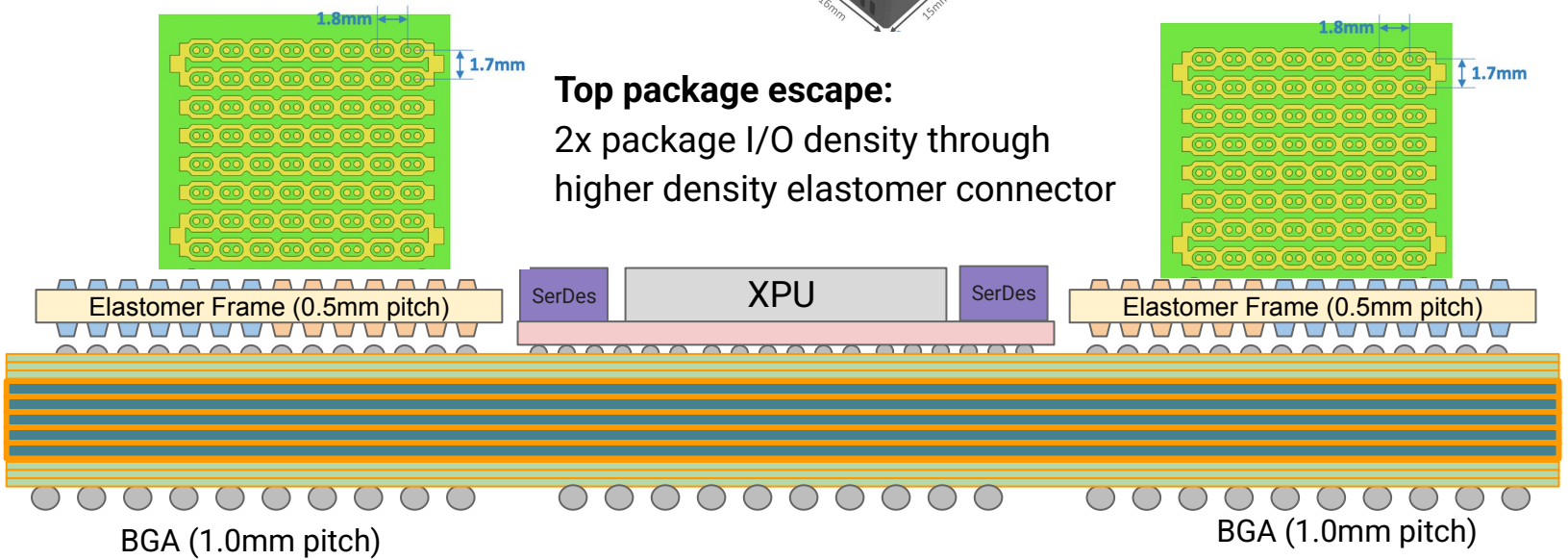


- Top package signal escape
- Better quality channel from bypassing package balls, vias, and PCB
- Double package I/O density

CPC Example

64 Diff Pairs in 240 mm^2

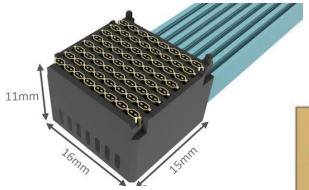


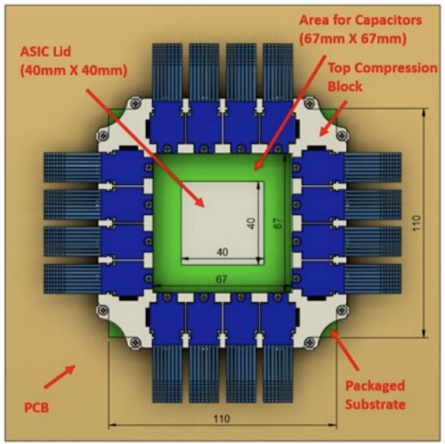


Top package escape:
2x package I/O density through higher density elastomer connector

Bottom package escape through regular BGA

CPC for Switch Chip

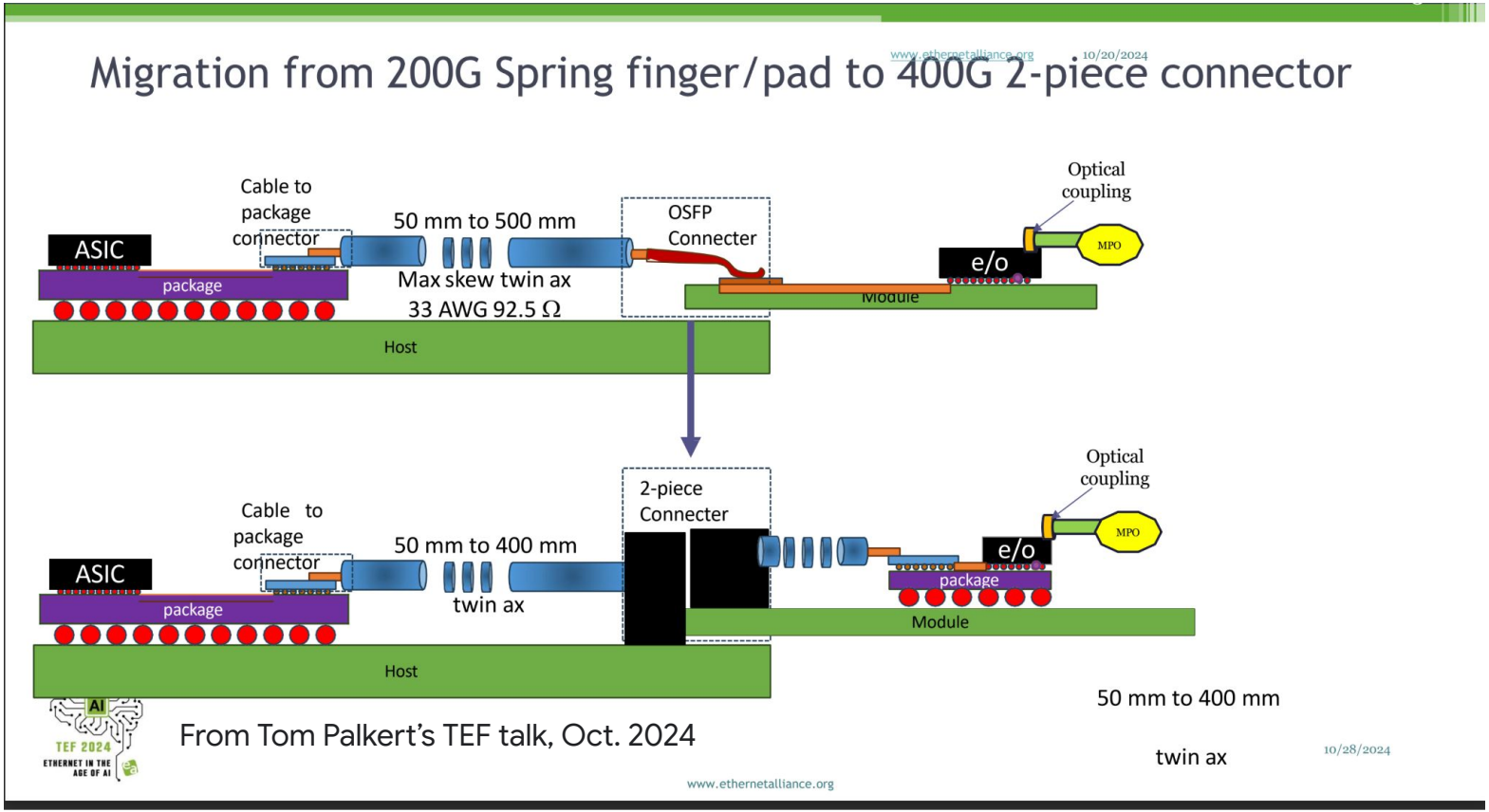




- Applicable to switch Chip too
- Shoreline for 16x 64 diff pair connectors

⇒ sufficient for 128x 3.2T port (400Tbps) switch at 448Gpbs per diff pair!

OSFP1600 Connector Needs Improvement



OSFP3.2T or Other MSA

- Connector solution for 8x 448Gbps
- DAC, LPO, TRO are not possible
  - Link budget and/or need for gearbox
- AEC, AOC, or similar for intra-rack connections
- IM-DD will be very reach limited (<1km?)
- Coh-lite support required
- Module form factor must support 50W electrical power
  - 12v Vcc in place of 3.3v

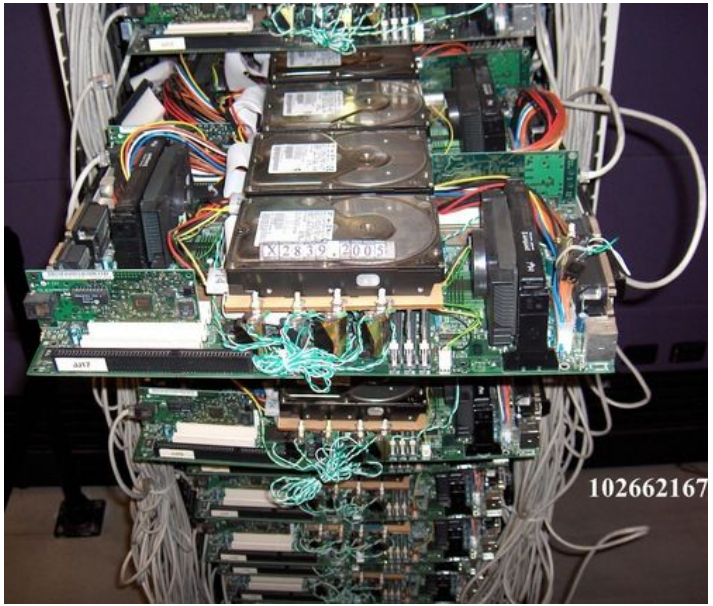
Do not need backward compatibility for scale-up and scale-out networks. Better to optimize form factor to maximize SI margin and simplify power delivery to module.

Serviceability

- We will be deploying 100,000s of xPU trays, millions of 448G links
- Need a reliable solution to yield installation and turn-up
  - Signal link margin
  - Mechanical resilience: ship fully populated racks
- Rack must be serviceable when components fail
  - Not practical to return the entire assembled rack
  - Minimize mean time to repair

25 Years of Data Center Operations at Scale

- Google has over 25 years of experience designing systems for data centers
- Numerous innovations to improve energy efficiency, system reliability, and serviceability
- xPU clusters for AI present new challenges in capacity and density and must continue to meet operational efficiency



Google corkboard server, 1999
Photo from Computer History Museum
https://www.computerhistory.org/collections/catalog/102662167

Outline

- Google Cloud ML/AI Clusters
- Why and Where 448Gbps
- Summary

Summary

- AI demands require 448Gbps links
- Front panel pluggables for xPU trays and switch trays can meet the density requirements and are preferred for lower operational risk and flexibility
- Need to improve on chip to module channel to support the higher speed 448Gbps links

Thank You

Thank you Standards Development Organizations and Multi-Source Agreement participants for your efforts to get us to 448Gbps deployments!

Google Cloud

Proprietary