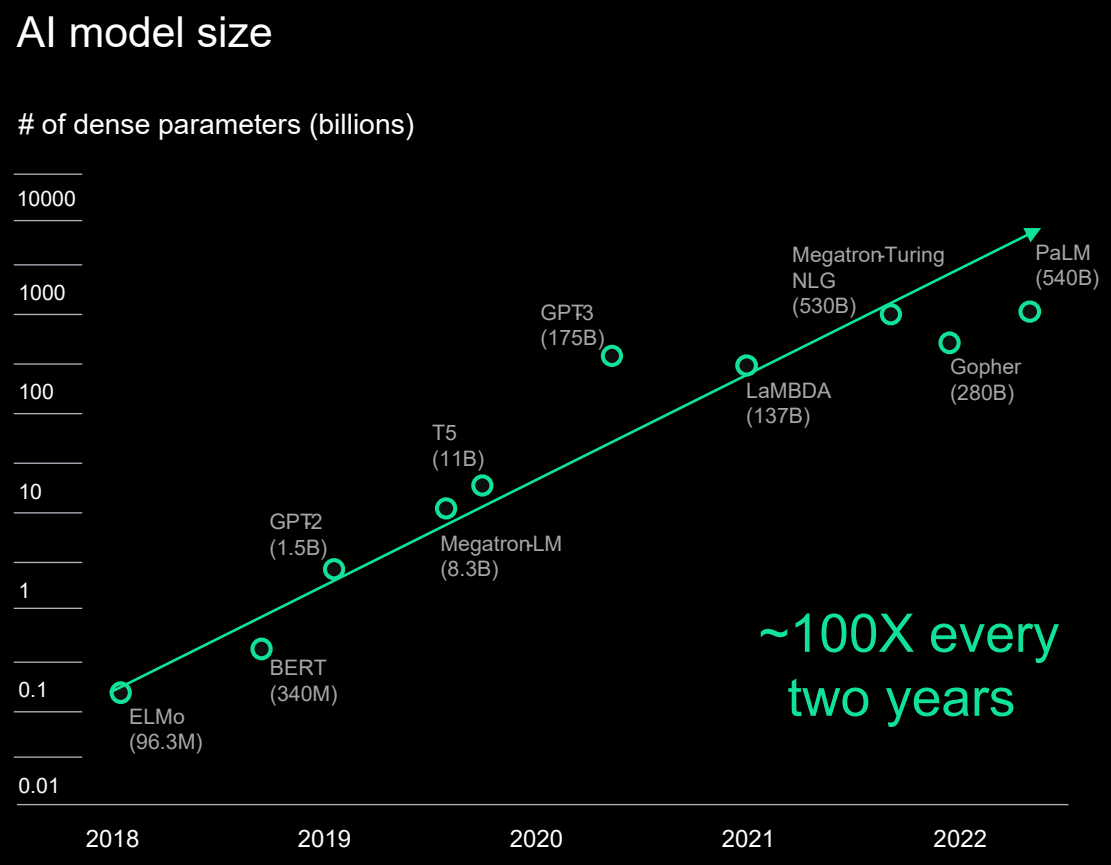
Salience Labs®

An AI revolution requires hardware innovation.

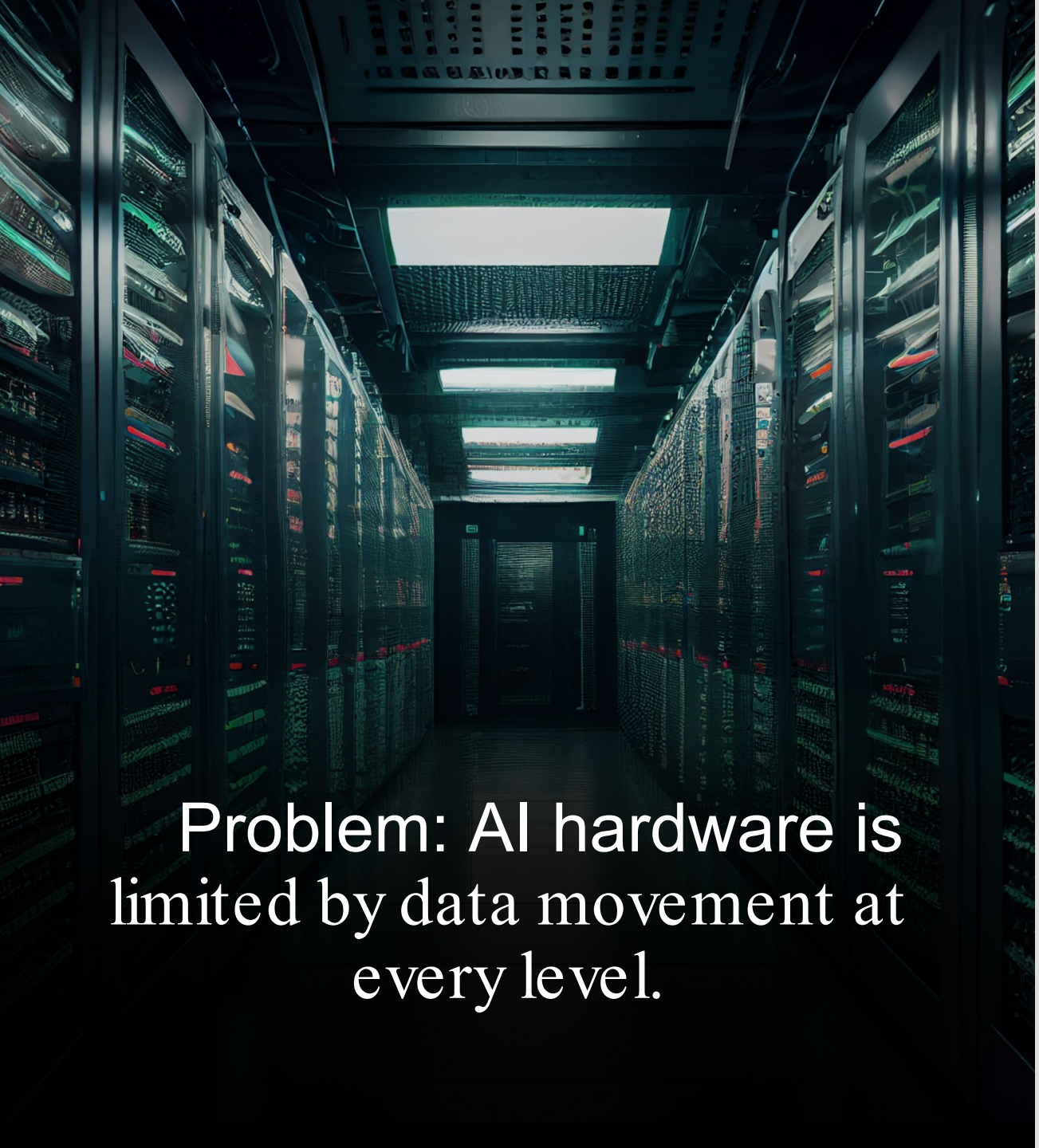
AI model compute requirements grow by 100x every two years.

Required computing growing with model size, plus demand in new applications.

This demand will not be met by incremental improvements to hardware.



SOURCE
[Google]



Problem: AI hardware is limited by data movement at every level.

AI models are too large to fit onto single cards or servers.

Chip to memory, chip to chip, server to server: at every level data movement limits performance.

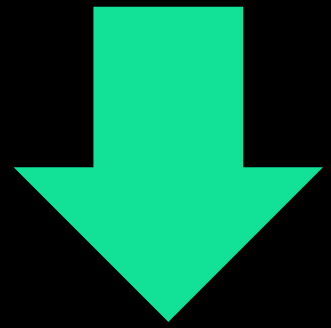
In AI hardware systems today, switches are 40% the heavy chips[1]

1. NVIDIA H100 DGX SUPERPOD HAS 256 GPUS AND 164 SWITCH ASICS.

AI workloads are driving *significant bandwidth growth*

50-60%

Cloud bandwidth growth each year

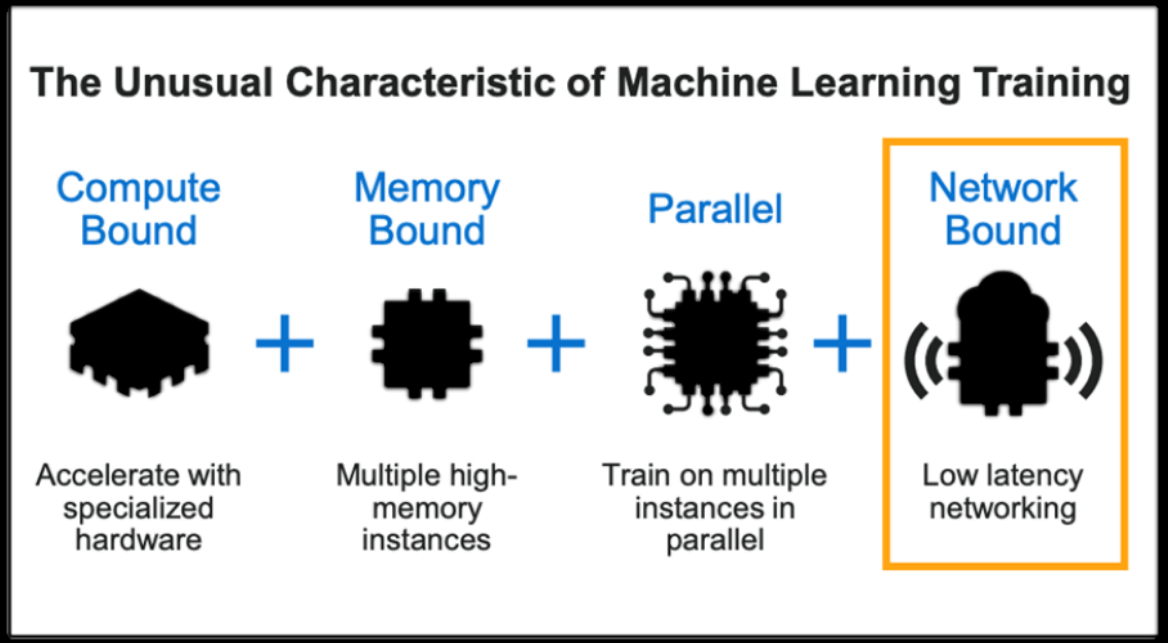


25-30%

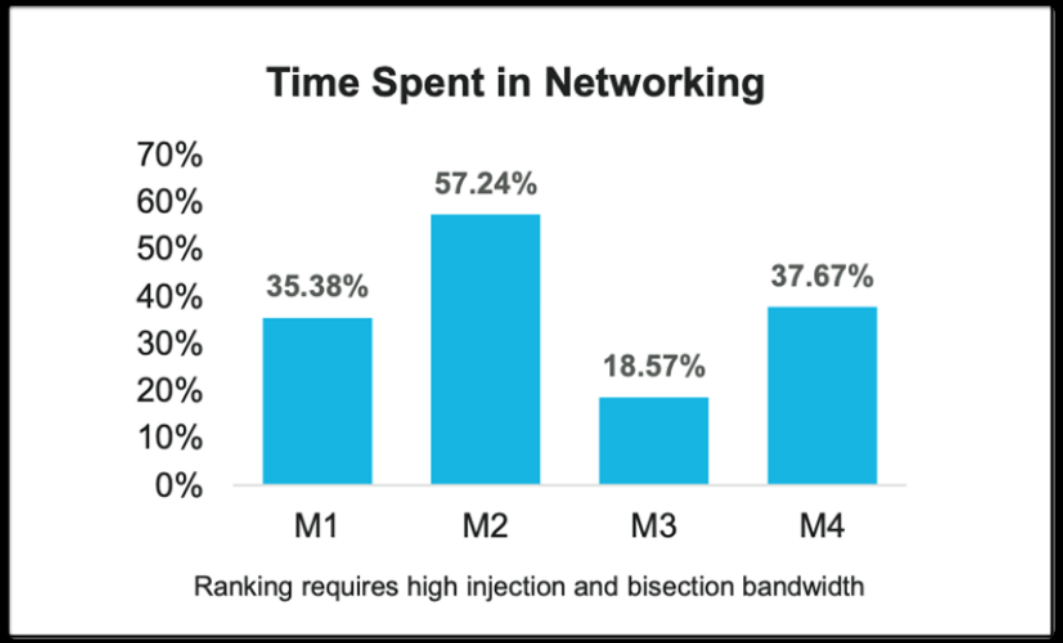
Network power per bit declining each year

And are network limited..

AWS Keynote at re:Invent2022



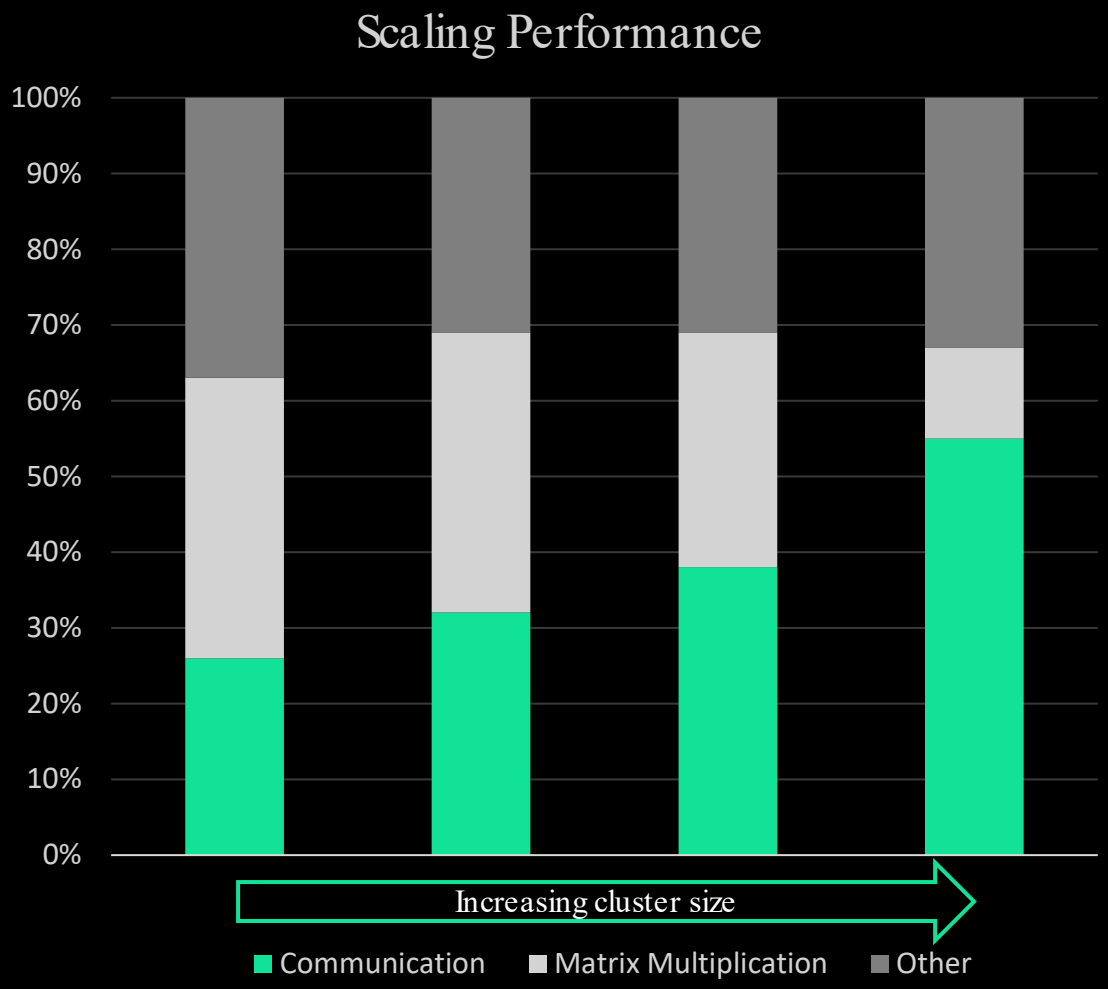
Meta Keynote at OCP Summit 2022



Mx=ML training models

Network bottlenecks fundamentally limit performance and increase model run time

The push to larger cluster size makes this worse

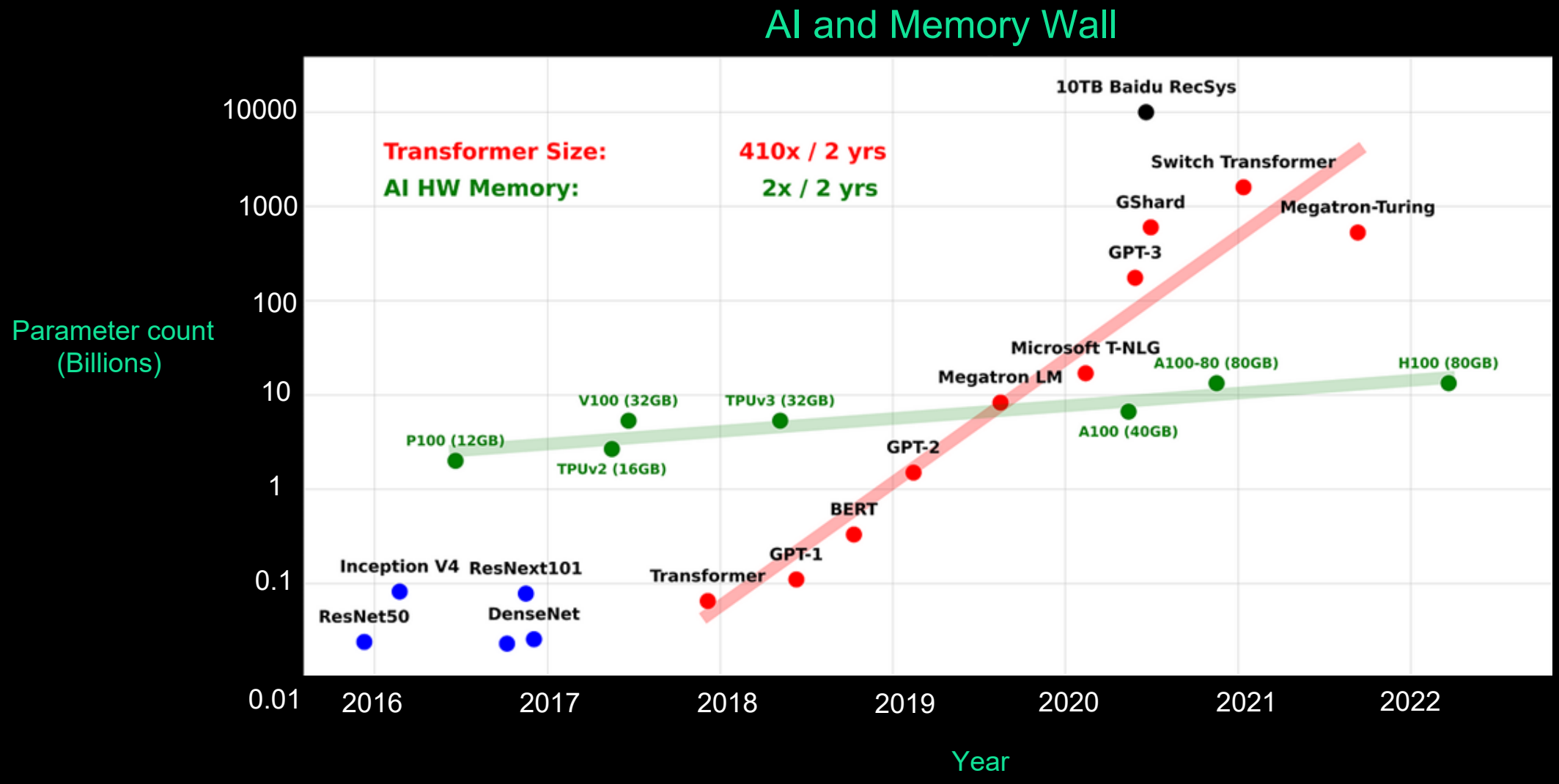


- Communication overhead increases with AI cluster scale
- Ensuring continued high network utilization is critical

- Opportunity: move to dedicated backend network

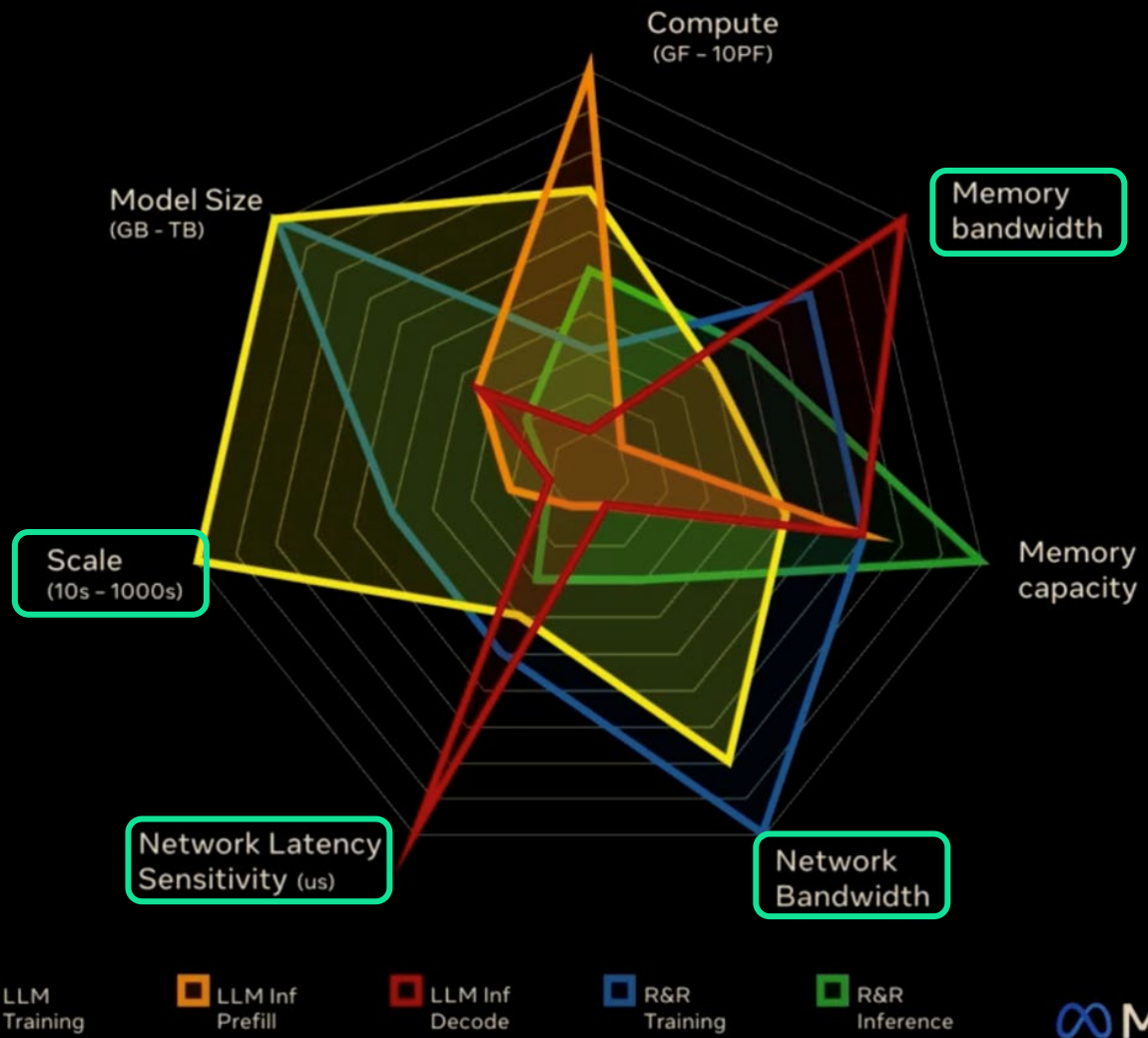
Source: OCP, Meta presentation 2021

Bandwidth is also an issue within the server



Source: Amir Gholami, Zhewei Yao, Sehoon Kim, Michael W. Mahoney, and Kurt Keutzer

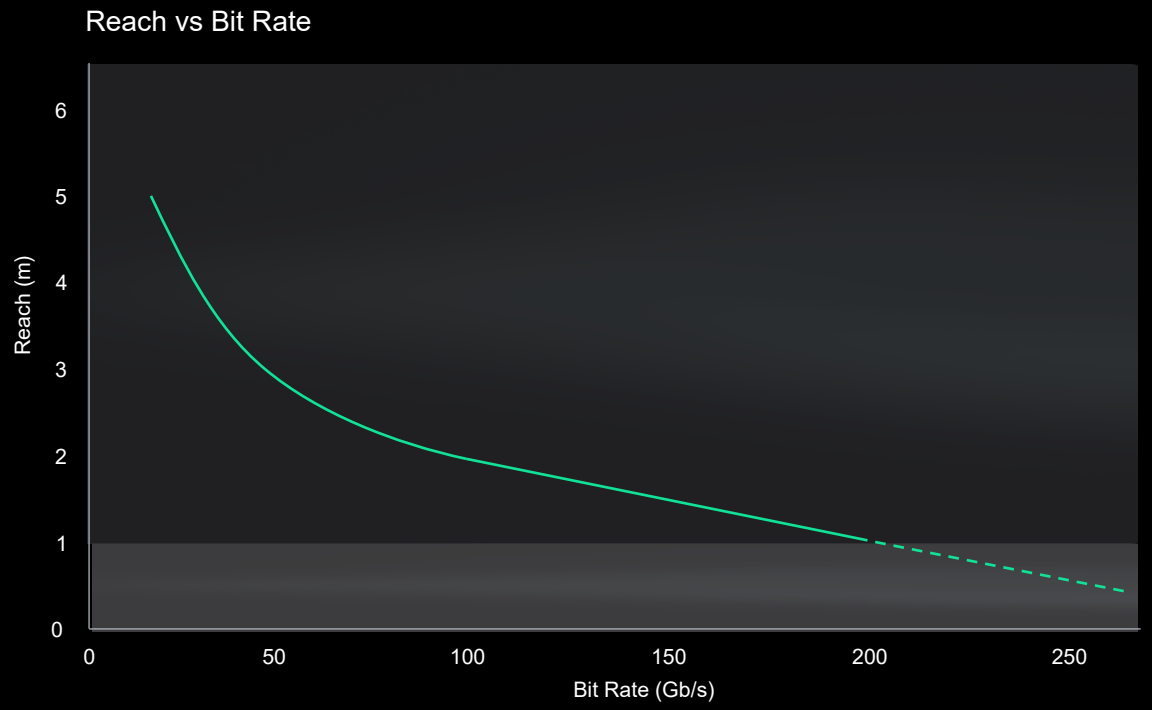
Different models = different needs, but many are network affected



- Meta presentation on different model requirements:
  - LLMs vs Ranking and Recommendation
  - Training vs inference
- Difficult to meet all needs with one solution
- Highlighted: data movement limitations

Meta

Optics will be used to get to high data rates.



Server to server connect is already optical.

Co-packaged optics for chip to chip and chip to memory is under development.

Nvidia and TSMC partner to develop silicon photonics solution

[Reuters News, Sept 13, 2022]

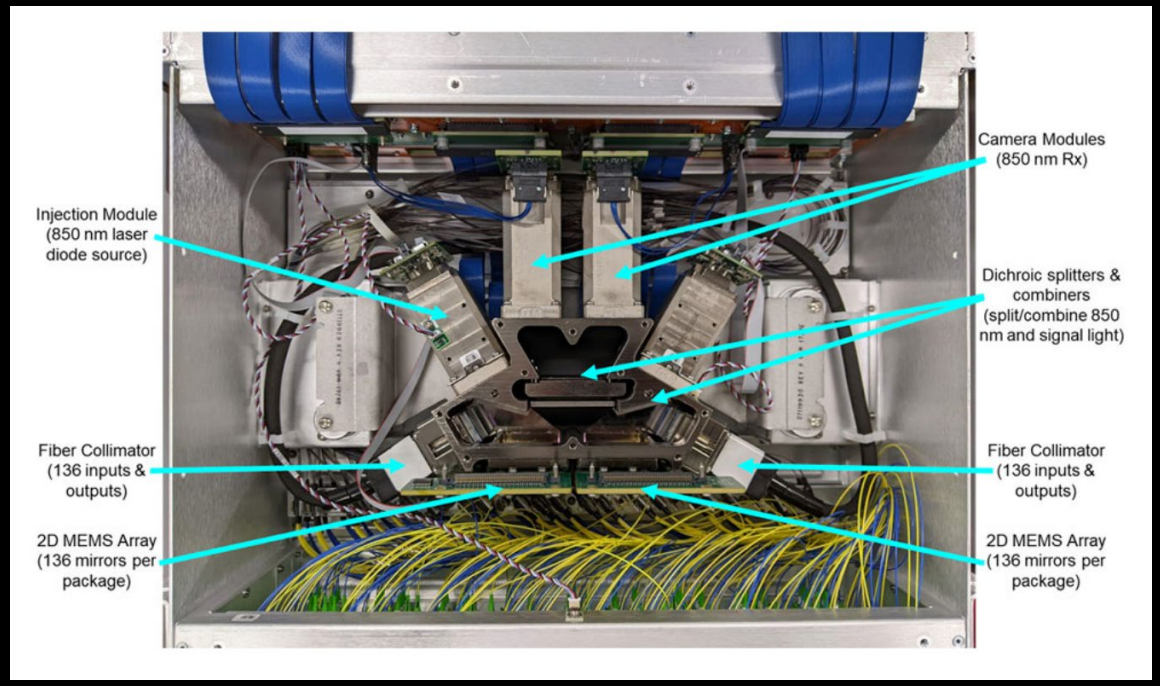
TSMC working with Broadcom and Nvidia to develop co-packaged optics (CPO), after the AI craze has lifted demand for data transmission

[Tech Node, Sept 12, 2023]

Meta Discusses AI Hardware and Co-packaged Optics.

[SemianalysisArticle, Sept 15, 2022]

Optical switches deployed by Google



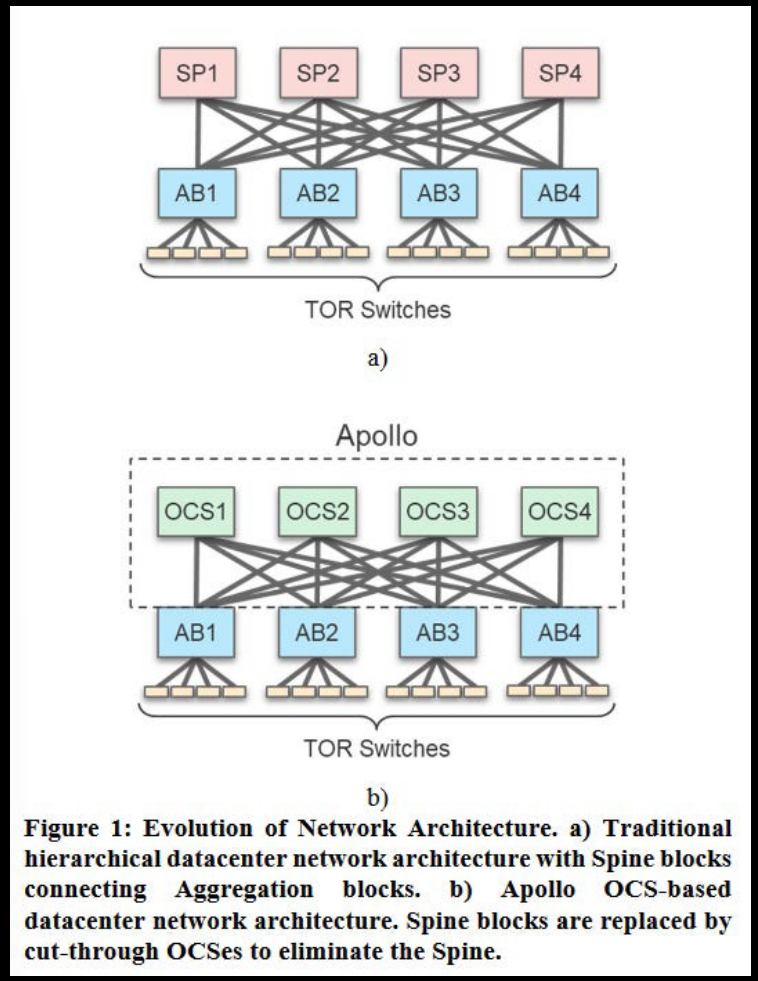
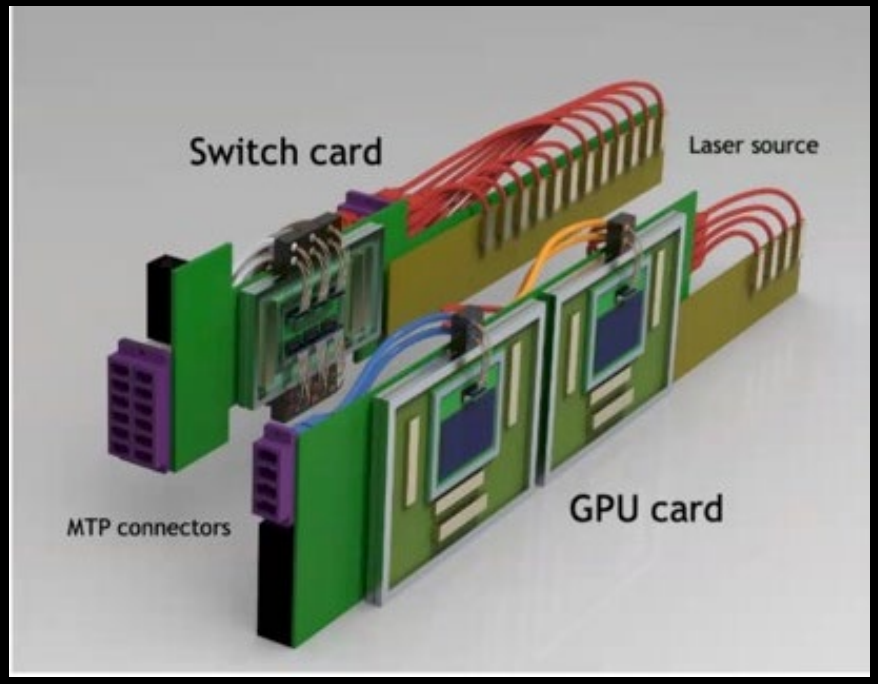


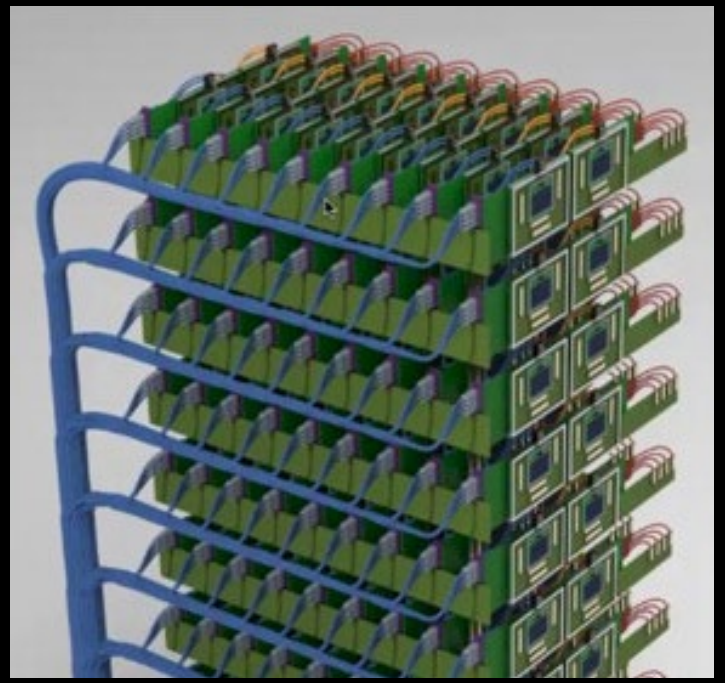
Figure 1: Evolution of Network Architecture. a) Traditional hierarchical datacenter network architecture with Spine blocks connecting Aggregation blocks. b) Apollo OCS-based datacenter network architecture. Spine blocks are replaced by cut-through OCSes to eliminate the Spine.

Source: Google Transforming Google Datacenter Network via Optical Circuit Switches and Software-Defined Networking

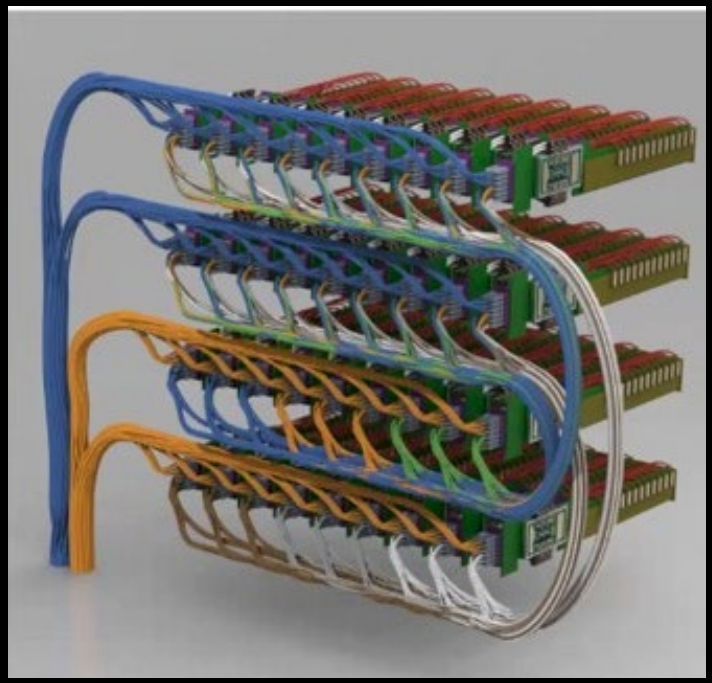
Nvidia concept on optically connected GPUs



GPU card with CPO



GPU rack

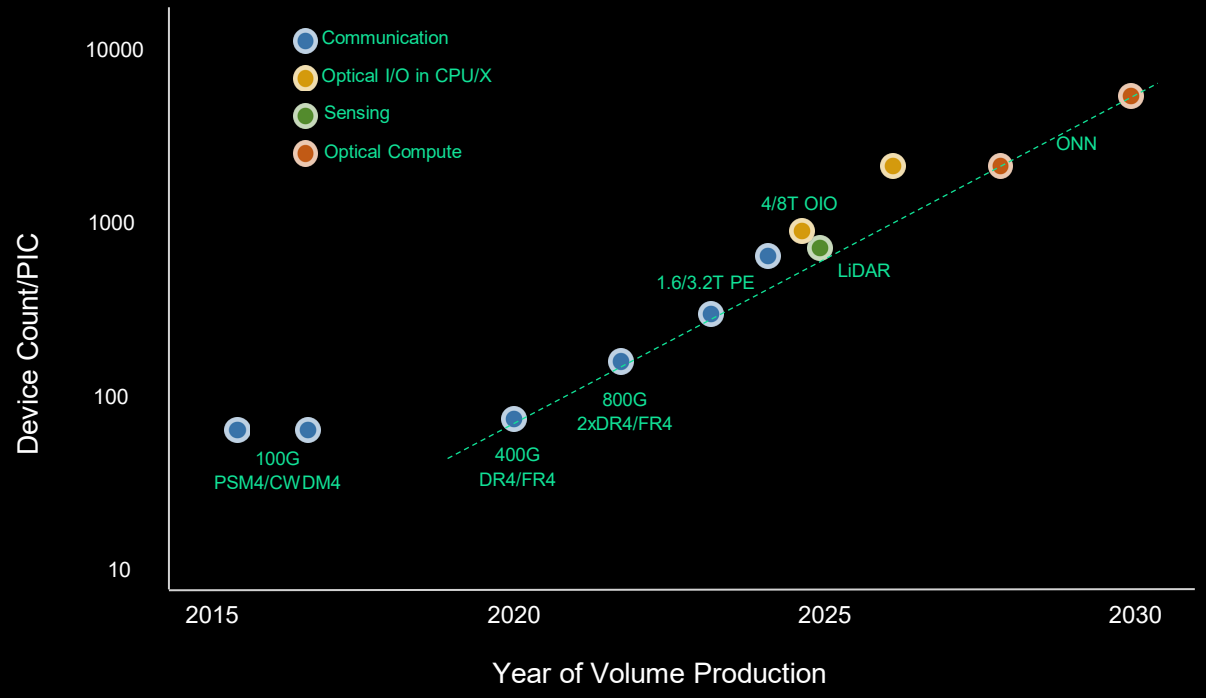


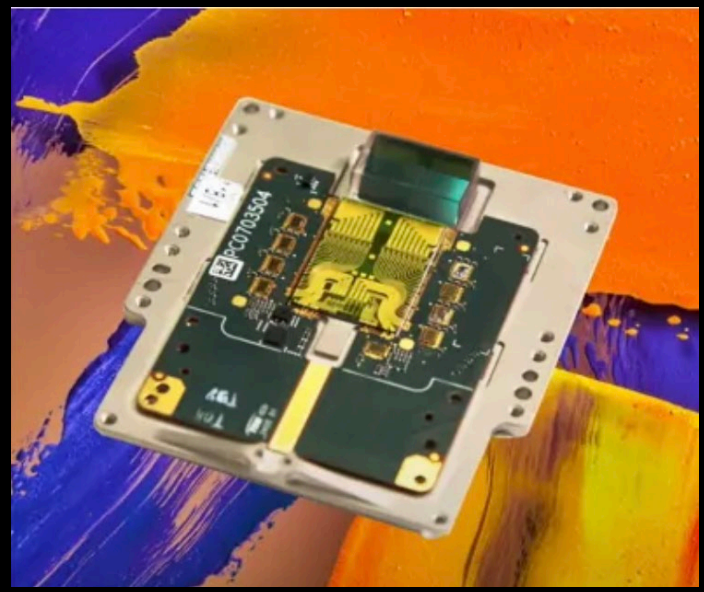
Switch rack

Source: Bill Dally presentation, OFC 2022

Silicon photonics fabrication has come a long way..

Silicon Photonics “Moore’s Law” Scaling





Intel: FMCW Lidar on a chip: integrated 6000 active and passive components on chip for high volume manufacturing

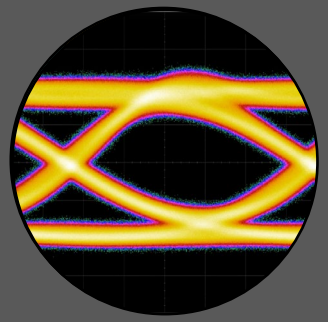
Source: Intel Silicon Photonics: An Overview

However, there are still some challenges faced

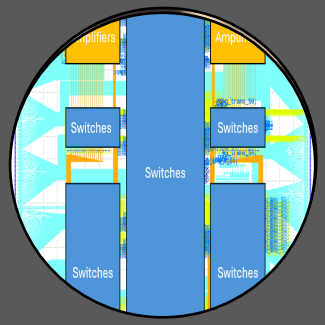
- Yield on devices: talked about as a generic “silicon photonics issue”, however needs to be broken down into its drivers
  - For example, can be a device issue very specific to each active component
  - Intel drove down to 7 failures per billion on integrated InP lasers
- Packaging:
  - Yield on fibre coupling: strong progress has been made, but yield here is critical for volume deployment
- Tools and flow are still under development
- Significant timeline to development
  - Several fabrication-test cycles required for characterization

End-to-end ecosystem co-development is required

Foundry PDK, tools and process flow



Materials and component development



Photonic integrated circuit design

System implementation and customer requirements

End-to-end co-design and cooperation


Let’s shape the
future together.