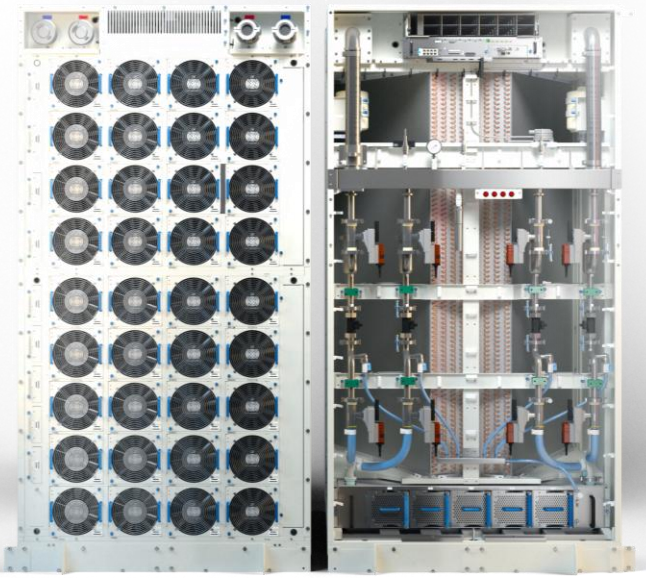
Microsoft Azure

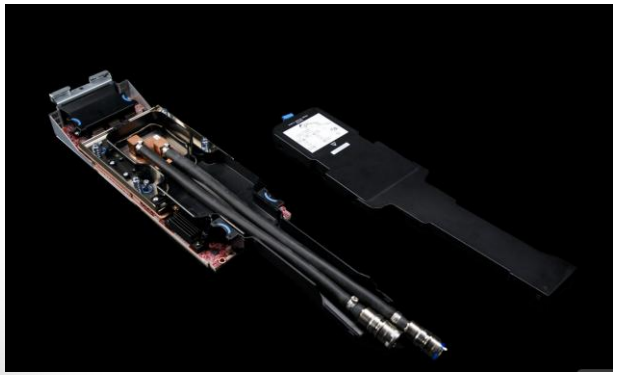
On the Considerations along the path to 448G for Scale up and Scale Out AI Workloads

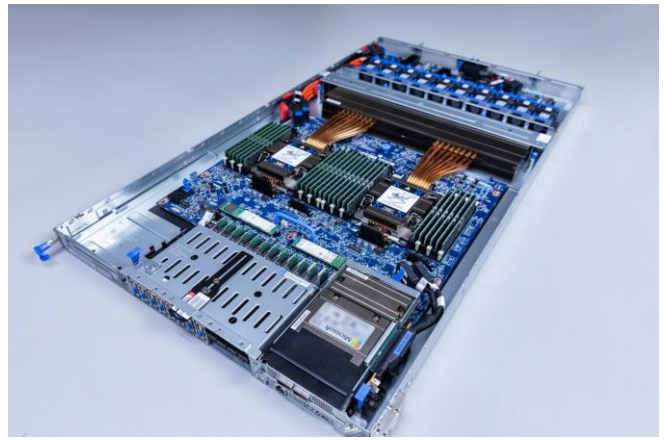
Ashwin Gumaste

Strategic Planning and Architecture

Building the world's computer



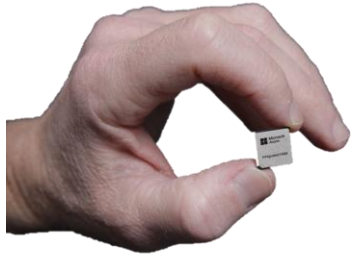


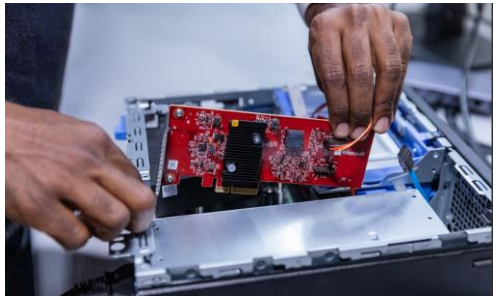



Microsoft
Azure
Cobalt 100


Microsoft
Azure
Maia 100

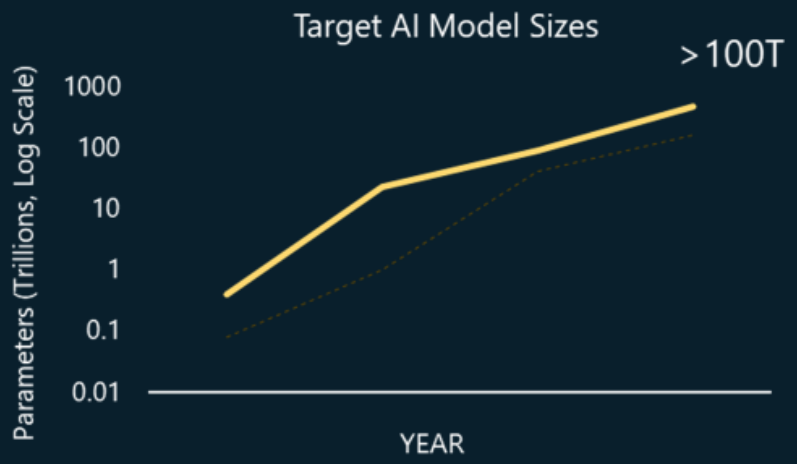

Microsoft
Azure
Boost DPU



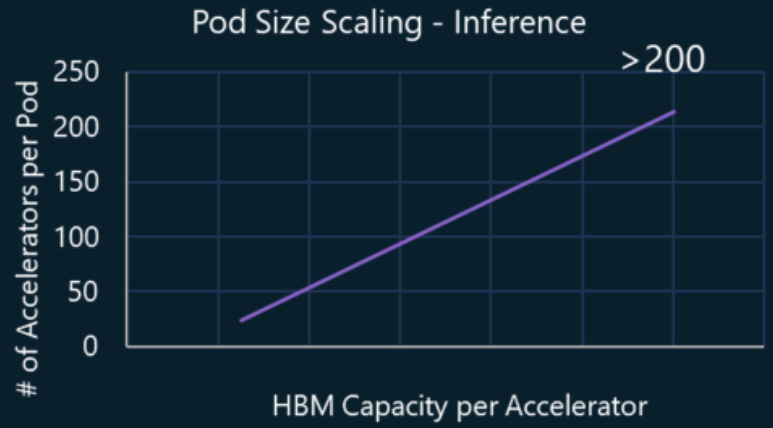


AI Infrastructure Trends

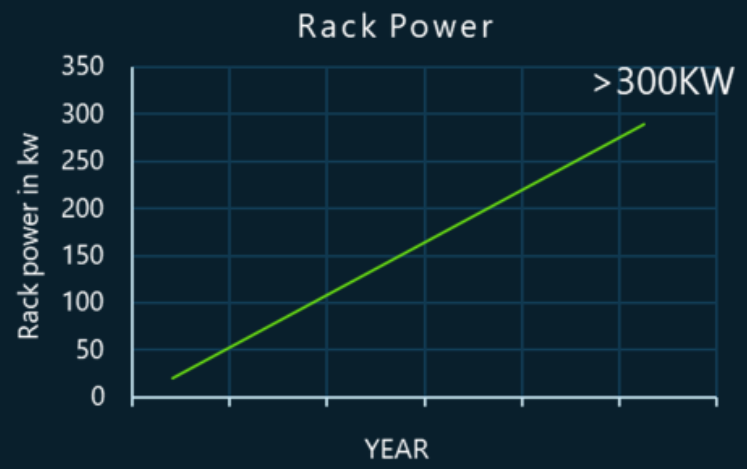
Model sizes driving up demand for GPUs



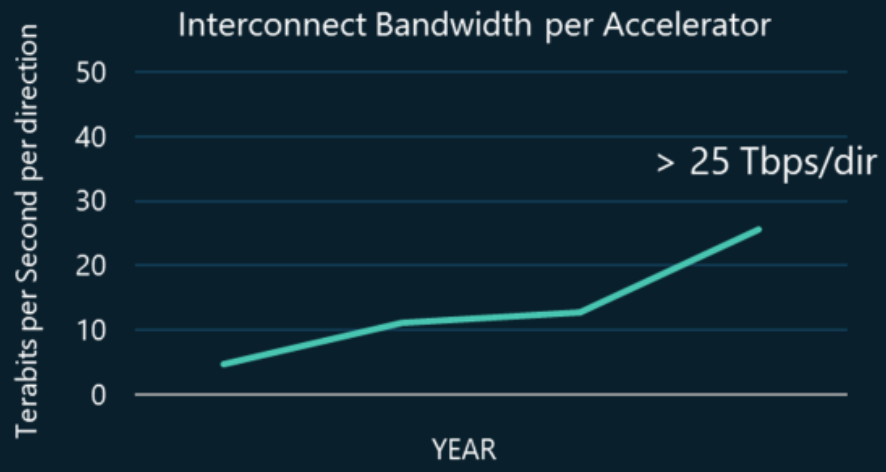
Pod size must scale to support growing HBM capacity needs



Rack power needs expected to scale > 3x



Increasing scale-up bandwidth demand

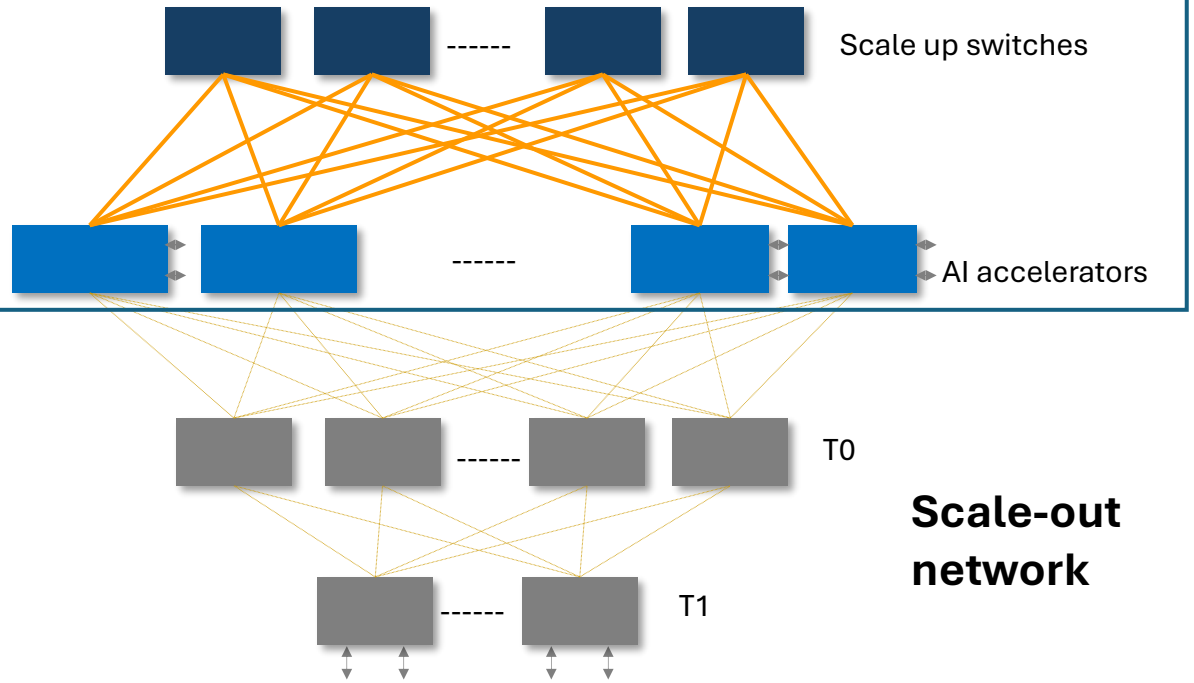


Microsoft Azure

Scale up AI Accelerator Networks

Scale-up network

AI accelerators interconnected in a POD using a single tier of packet switches.



- Highest bandwidth, lowest latency network.
- Symmetric topology: software-friendly, uniform scale-up communication latency
- High Reliability: Communication properties invariant to accelerator or switch hardware failure

Pod scale greater than 1 rack to support >100T parameter models

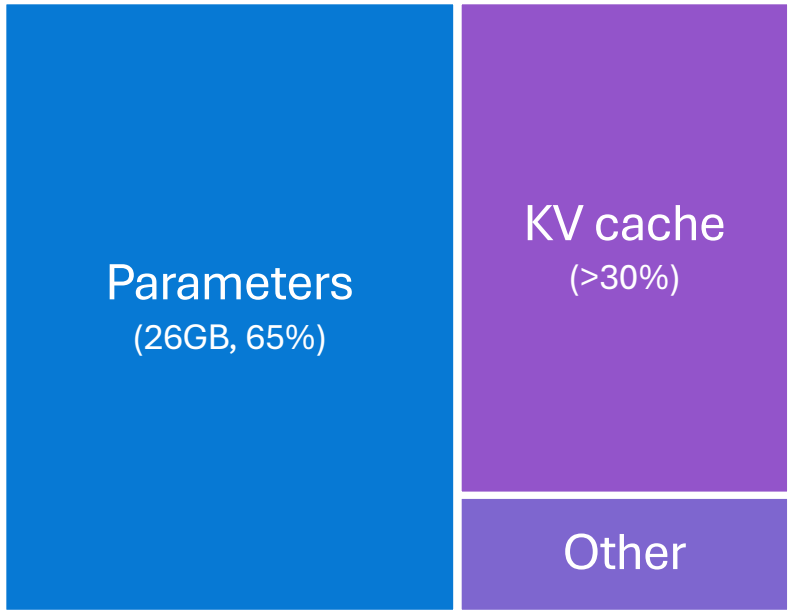
Inference Memory Requirements

Parameter GPU memory

1 FP16/BF16 parameter = **2 Bytes**

1 Billion parameters = **2 GB**

175 Billon parameters = **350 GB**



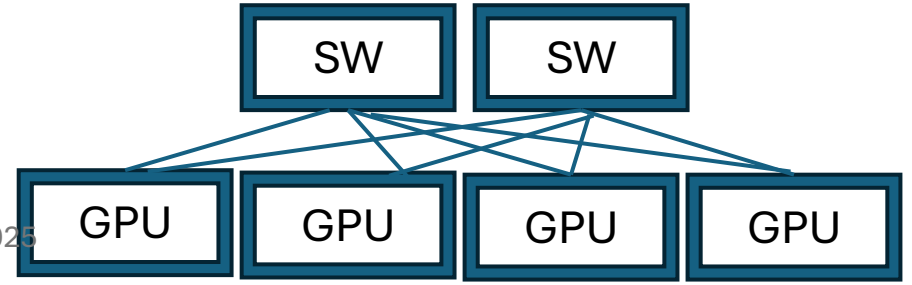
Nvidia A100 40GB

Memory layout with a 13B LLM model

Source: https://arxiv.org/pdf/2309.06180

AI workload characterization -- wish-list for inference

- Inference workloads:
  - All2All, all gather collectives dominate communication
  - Latent vector transfers require low latency dedicated BW
  - Model size in the past 2Qs may be flattening due to MLA
  - Implications of smaller model size – the TPxPP product is relatively modest, implying scale to fit within a rack.
- Rack requirements and SU communication
  - Wide Rack dimensions (mm) – up to 2000+ high, 1200+ deep, 700+wide.
  - Backplane for entire rack. Typical backplane 26AWG twinax. 0.3dB/inch @224G.
- Topologies include:
  - Leaf and spine: multiple accelerator trays and few switch trays – each accelerator connected to every switch across the CBP.
  - Torus topology – an n-ary hypercube mapping with communication from the GPU connected to other GPUs in n-dimension space (n, usually 3)
  - Exotic mapping: dragon fly and switches orthogonal to the rack
- Technology choices:
  - 224G both copper and optics
  - 448G both copper and optics



Scale up requirements for inference

Topology	Worst case path	Path loss	Remarks	Future
Leaf and spine	Accelerator through the tray to the CBP, CBP, switch tray, switch	Could be as bad as 90dB, requiring more than 2 retimers. (@224G)	Various considerations – efficient flyover cable, connectors.	CPC, NPO, OBO, CPO
Torus	Physically separated accelerators – accelerators in trays that are vertically separated.	~70dB, requiring retimed solution	Tough to support scale out from a torus without excessive use of retimers.	Optical scale out. (rate agnostic).
Alternative architecture: Orthogonal switches and dragonfly	In case of orthogonal switches: top of the rack to bottom-most switch. Dragonfly: diameter of the dragonfly network	Orthogonal: 60dB ~ 1 retimer solution Dragonfly – dependent on network diameter	Requirement for specialized CBPs, in-service upgrade challenges, shuffles for twinax in the backplane	CPO/OBO based connectivity.

How did we get here

- 3.125Gb/s → 28G → 56G →112G → 224G

- The challenges at 224G
  - The illusion of 40dB
- Making 224G work
  - Retimer
  - Retimer design constraints:
    - Loss (2-3dB, equalization, software support)
- Modulation formats
  - NRZ
  - CS RZ
  - PAM4
  - PAM X (PAM6 vs PAM8)

CBP		26 AWG loss (dB/m)	32 AWG loss (dB/m)
224	PAM 4	7.8	13.4
448*	PAM 8	9.3	15.6
448*	PAM 6	10.2	17.2
448*	PAM 4	11.4	19.3

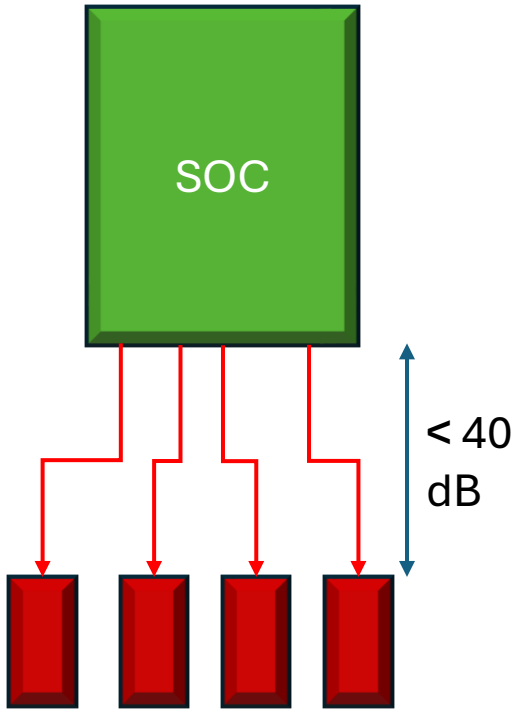
* Marvell

AI workload desires from 448G (Electrical)

- 40dB end to end IL.
- Low power mode for short range links
- Interop with optics (low power mode, CPC, optics interface)
- Ecosystem for connectors and flyover cables
- PCB materials at relatively low loss
- FEC power requirements.
- Low power retimer
- PAM4/6/8 requirements to make 448G work: 50% improvement in key COM parameters
- Reach ~ 1m without retimers.

448G and optics intercept

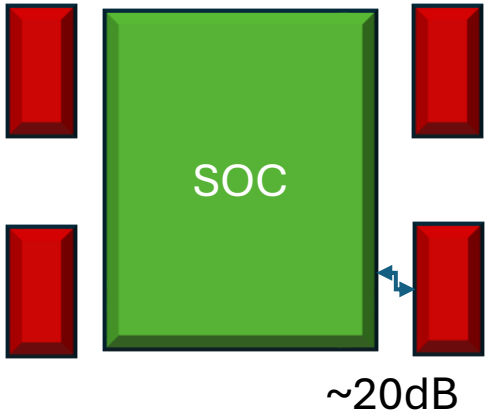
Pluggable Optical modules



1.6T/3.2T optical modules (OSFP)

- Restricted by frontplate capacity. 20-100m
- Cost of optical modules
- Fiber cabling in the front
- Standardization of 448G interface
- Intercept to LPO and LRO

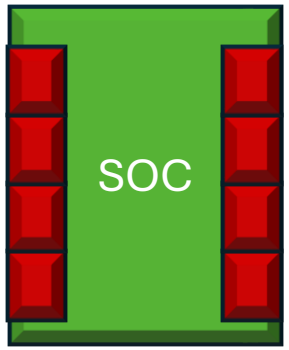
OBO modules



- OBO or NPO: Up to 6.4T or even 12.8T per module.
- ELS for optics
- Large port count module
- Expected power: <10pj/bit and target of ~4.5Tb/s /mm (TX+RX) (SerDes density)
- 20-100m

Defining 448G for ~VSR/XSR

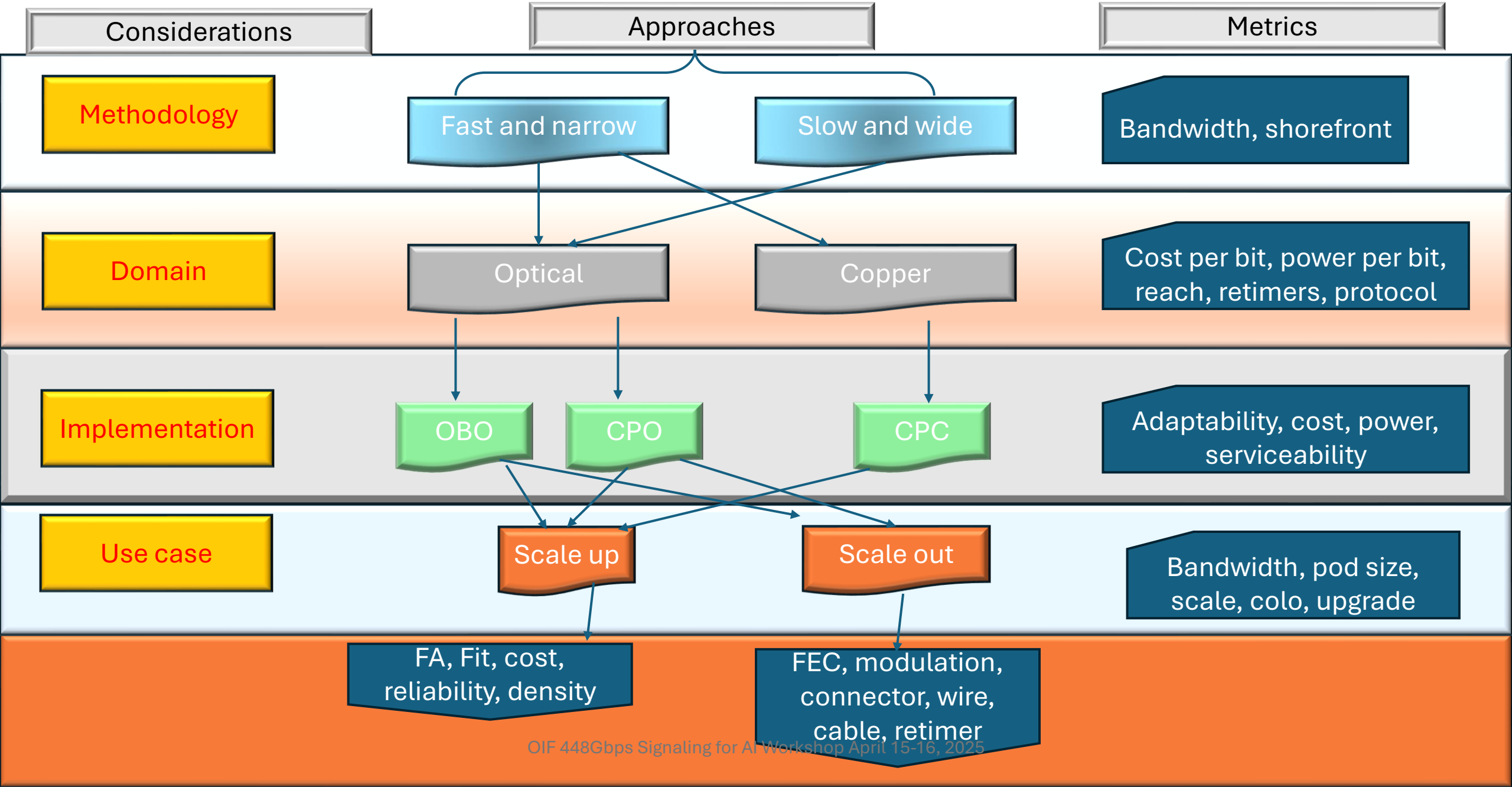
CPO modules integrated in the SOC



CPO intercept

- Integrated in the SOC
- ELS for optics
- Well integrated in the accelerators
- Expected power: <<10pj/bit and target of ~4.5Tb/s/mm (TX+RX) (SerDes density)
- Reach of 20-100m.
- Support for both 224G and 448G optics, plausible support for UETS.
- Support for slow and wide as another option.

Paths at 448G



448G – the way forward

- The scale up network for next gen inference
- Power optimized (pj/bit)
- Cost optimized ($/bit)
- Copper vs Optical technologies
- Optical intercept for 448G
- Modulation technologies
- Reach of new SerDes

Metric	Good for?	Remarks
Reach	1 m	Copper
Reach	5-20m	Optics
Latency	<500ns	Copper
Power	<10pj/bit	Incl Serdes
Shoreline	5Tb/s/mm	Copper/optics