Good things come in small packages:

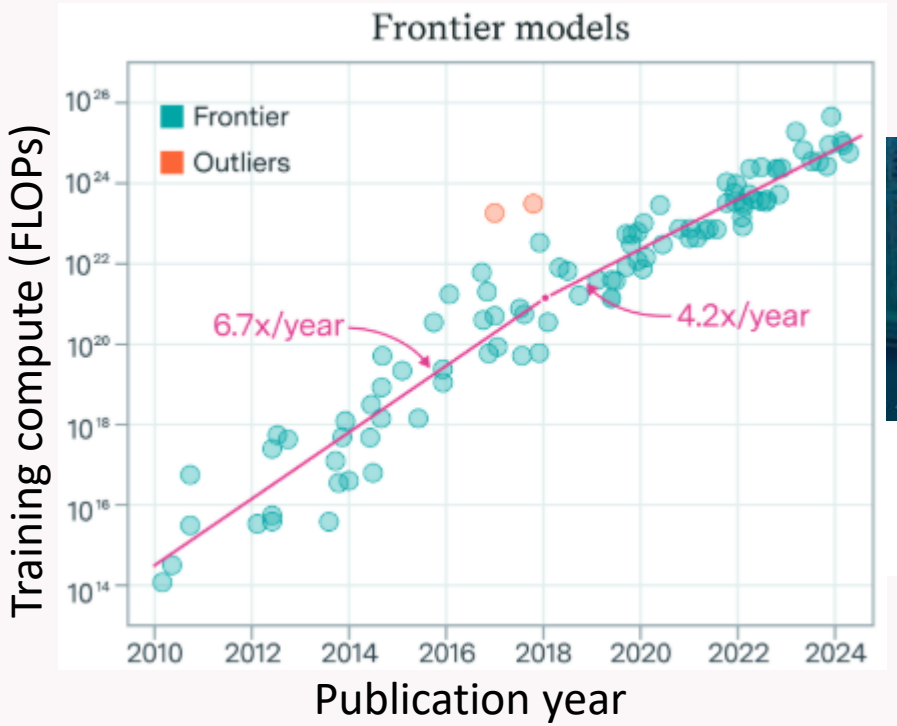
Should we build AI clusters with Lite-GPUs?

Burcu Canakci, Junyi Liu, Xingbo Wu, Nathanaël Cheriere,
Paolo Costa, Sergey Legtchenko, Dushyanth Narayanan, Ant Rowstron

Microsoft Research

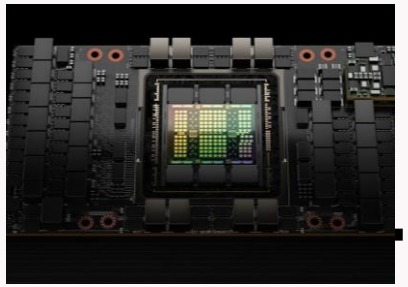
HotOS'25

AI demand and requirements both growing



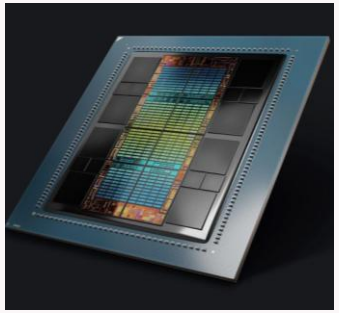
source: epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

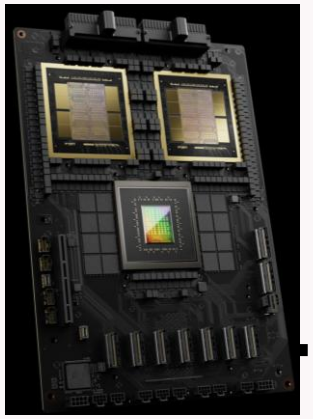
Hardware designers responded by *scaling up*



NVIDIA H100
2022

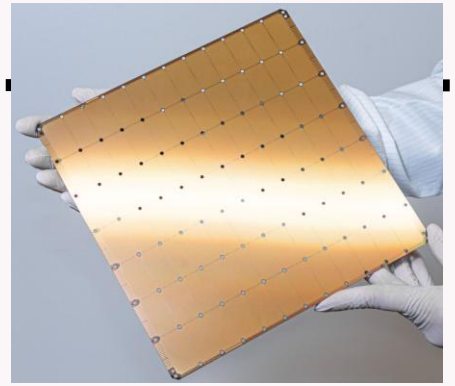
AMD MI300X
2024

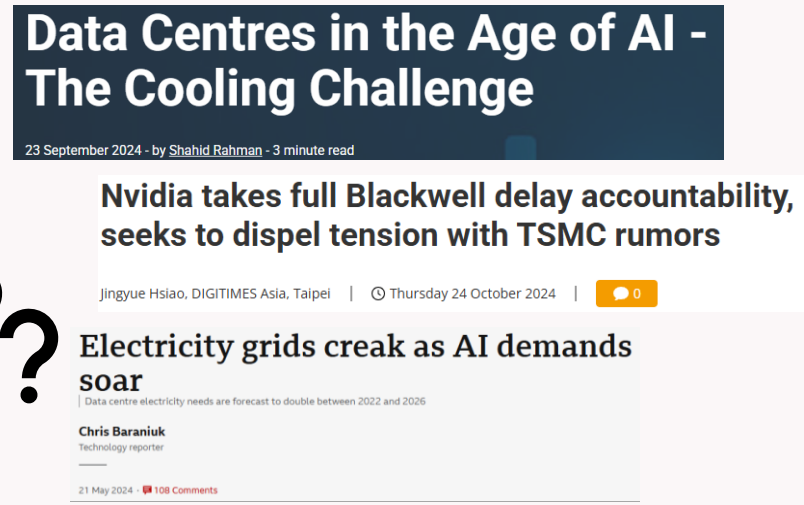




NVIDIA B200
2025

Cerebras W3





Trend has been to build larger and larger accelerators
to tightly pack resources

Large AI accelerators
are facing many limits:

-high manufacturing costs

-poor hardware yield

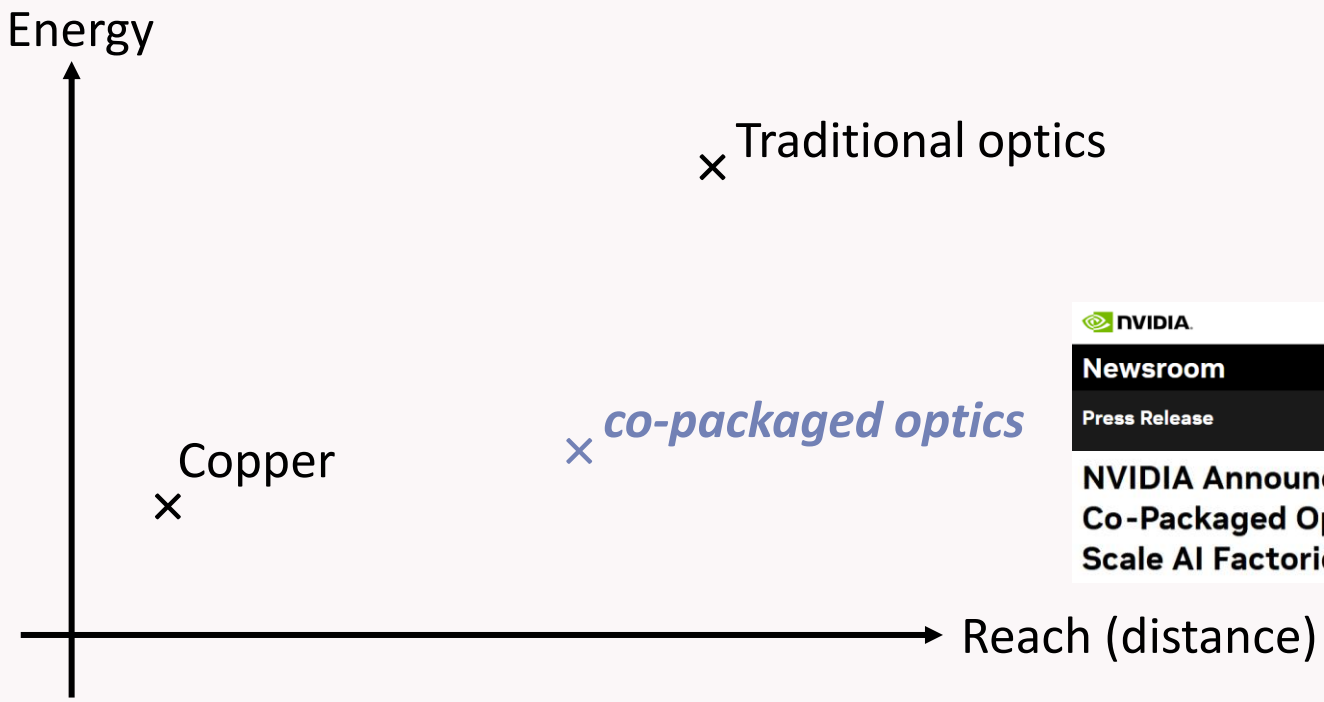
-large failure domain

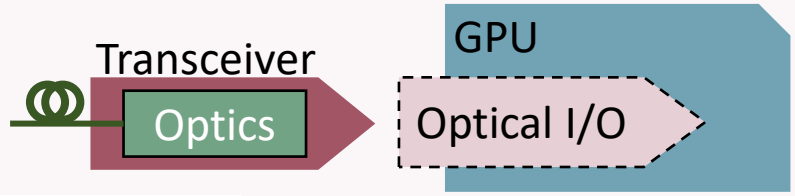
-power & heat challenges

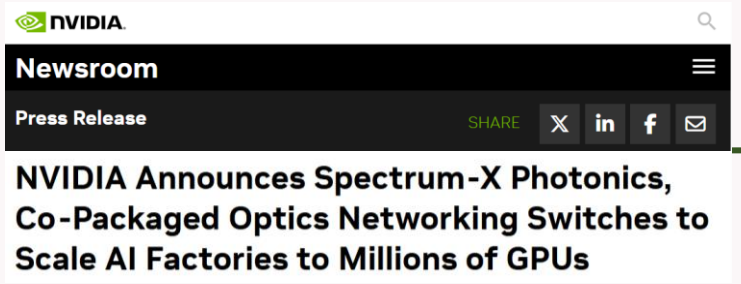
Scaling AI clusters ***out*** is the path forward

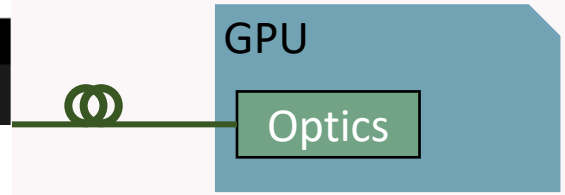
The challenge: AI workloads require very **very high bandwidth**

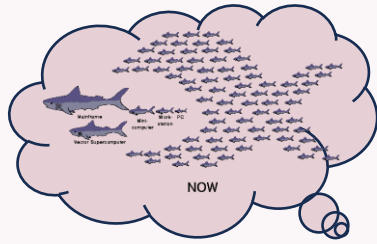
A step change in optical communication is coming





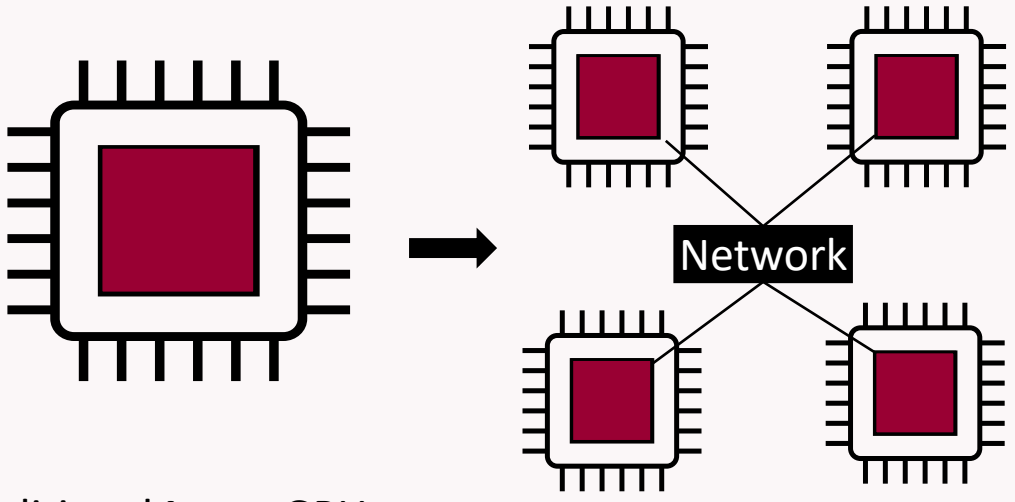






Can we replace large GPUs with clusters of smaller GPUs?

The Lite-GPU



Traditional **Large** GPU

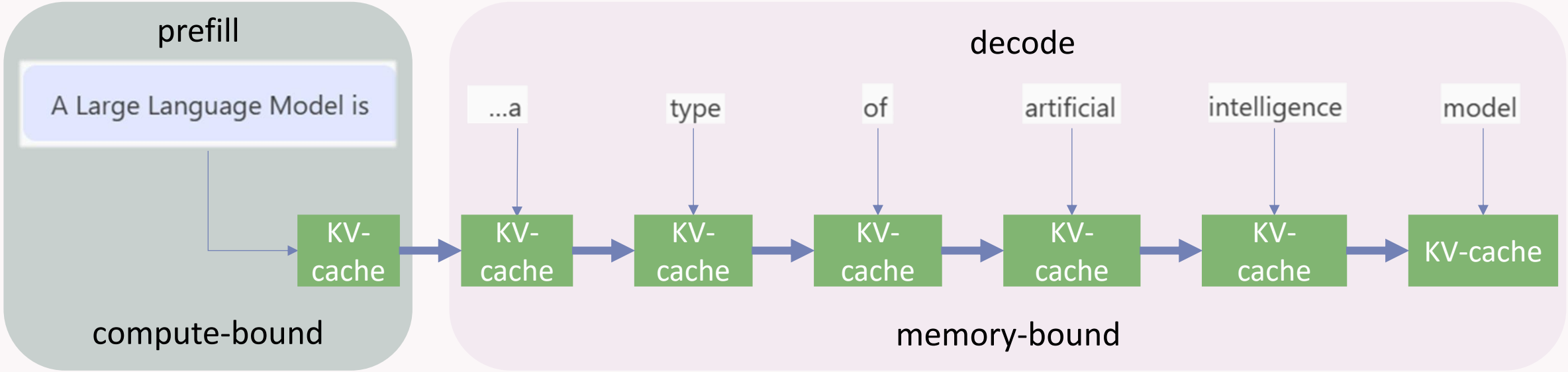
Small GPUs with Optical IO

+ Much lower manufacturing cost
+ Much better yield
+ Easier cooling & power management
+ Higher BW to compute ratio
+ Overclocking potential
+ Smaller failure domain

- More complex network, systems challenges

Case study: LLM inference

- The prominent workload in AI data centres
- Two phases: prompt prefill & decode



- Can optimize for phases separately (Splitwise)

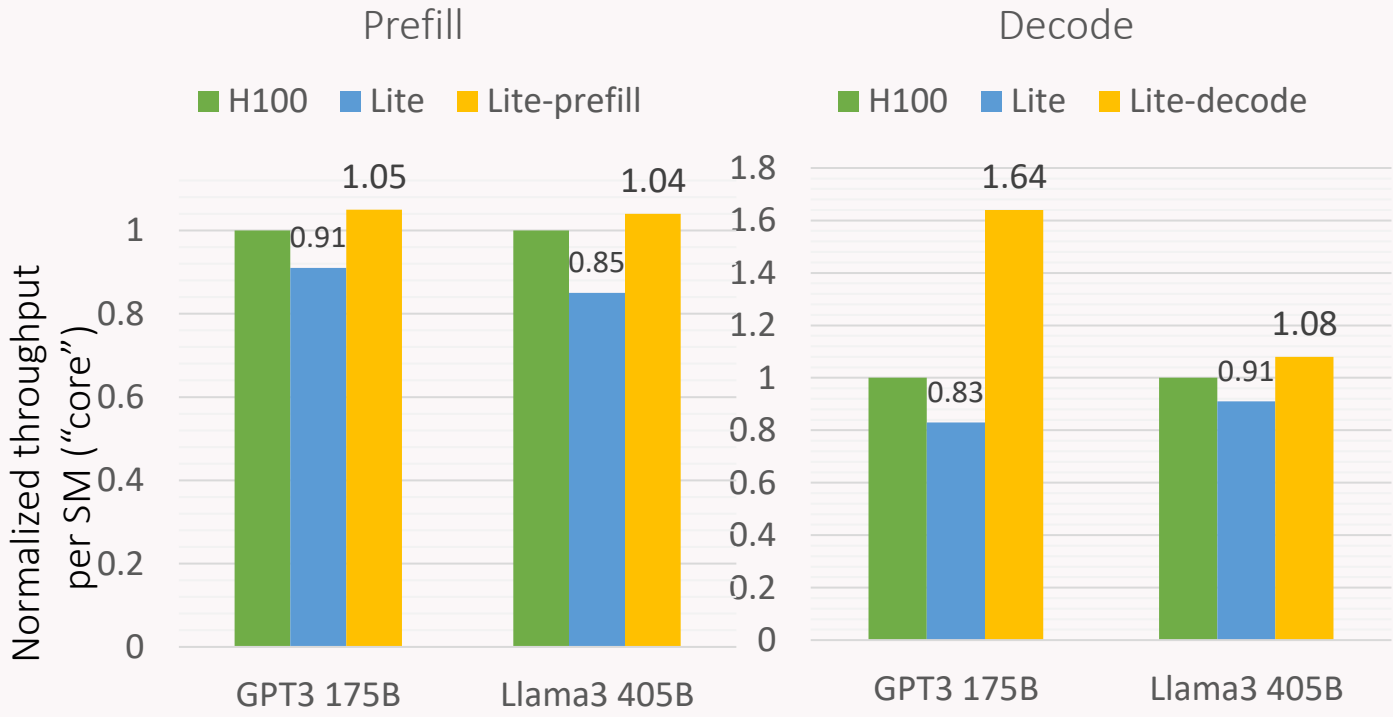
Running LLM inference on a Lite-GPU cluster

Smaller GPUs
- can be overclocked more efficiently
- have higher BW to compute

Baseline: NVIDIA H100

Lite	¼ of H100's FLOPS, memory capacity, BW
Lite-prefill compute heavy	Lite + 10% overclocking + optimize for network BW
Lite-decode memory BW heavy	Lite + optimize for memory and network BW

Analytical model

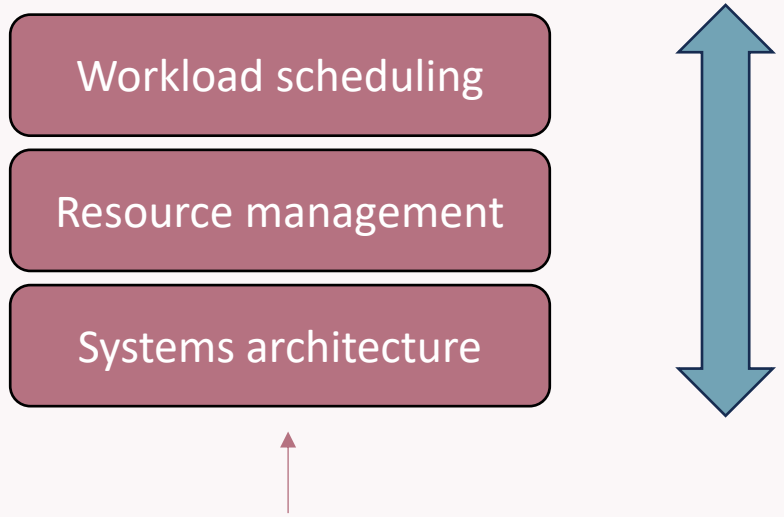


A Lite-GPU cluster can match or even outperform an H100 cluster
In addition to other benefits of Lite-GPUs!

What systems questions do Lite-GPUs prompt?

Complexity previously handled in the hardware is now carried to software

Many challenges are existing distributed systems questions



How to build a Lite-GPU network?

How to manage the Lite-GPU network?

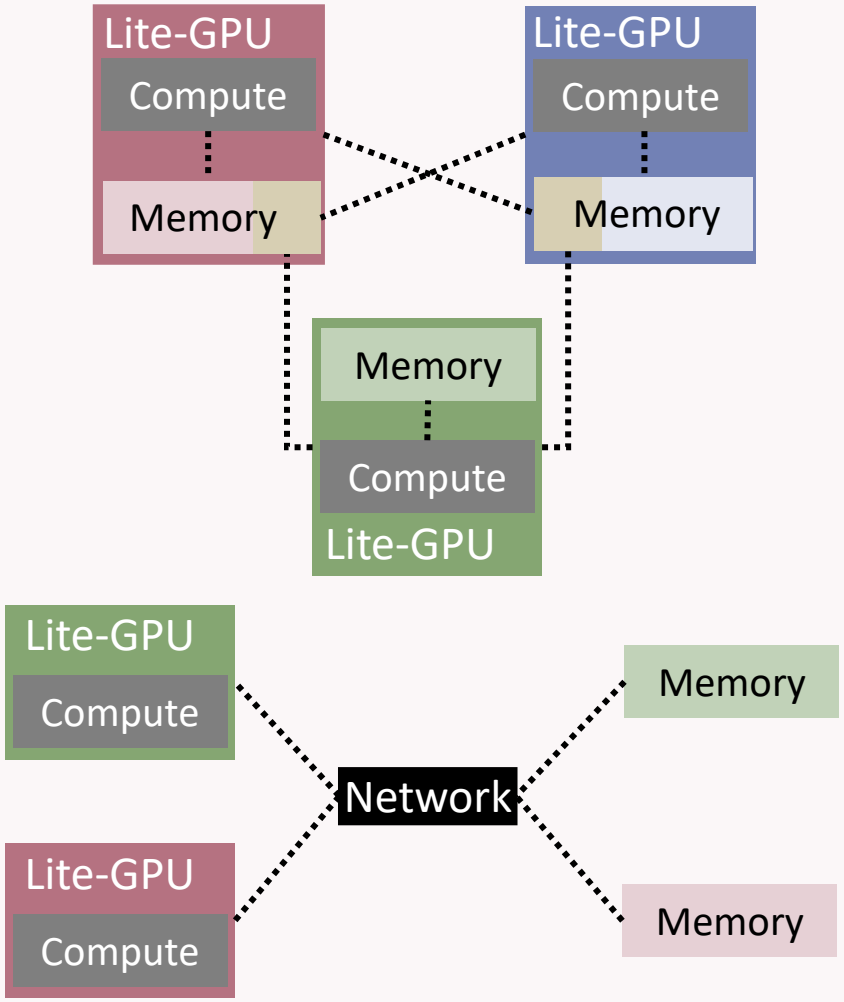
- Needs much higher BW than current networks
- HW/SW co-design is necessary to utilize OCS effectively

Packet switching	Optical circuit switching
+ Traditional + Flexible - Less efficient	+ Faster and much more efficient + More ports at high BW - Less flexible

How to mask the latency overheads?

- e.g., can we exploit predictability in workload?

Memory architecture opportunities



Each Lite-GPU has a fraction of the memory

How should Lite-GPUs access memory?

- memory sharing? ideal semantics?

Where should memory be located?

- within the Lite-GPU? disaggregated?

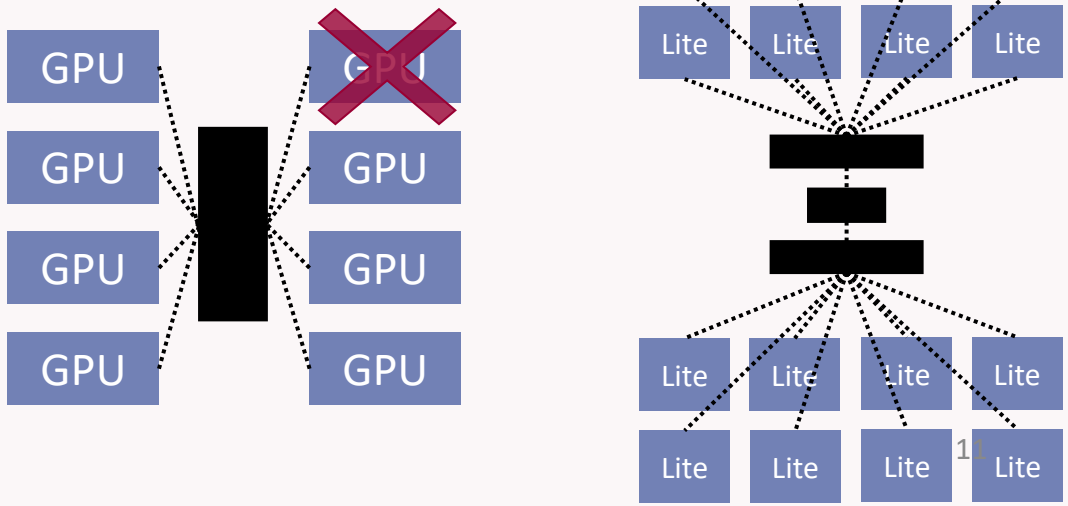
Running AI workloads effectively

Right parallelisation strategies?

- beyond known parallelisms & collectives?

Ensuring improved fault-tolerance?

- smaller failure domain but more devices



To sum up…

- Scaling out: promising alternative to overcome challenges of scaling up
- It is the right time to ask *how* to scale out AI clusters
- Lite-GPUs have a lot of potential as an answer; many systems challenges and opportunities await

Thank you