Inside Microsoft AI hardware innovation

Mark Russinovich

CTO, Deputy CISO and Technical Fellow, Microsoft Azure

@markrussinovich

Microsoft AI supercomputer

May 2020

10,000 V100 GPUs
#5 supercomputer

Nov 2023

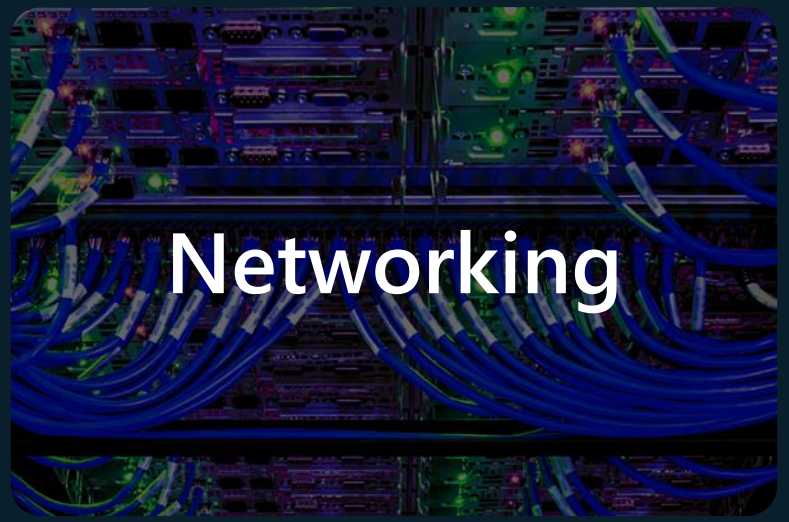
14,400 H100 GPUs
#3 in TOP500

May 2024

30x supercomputers

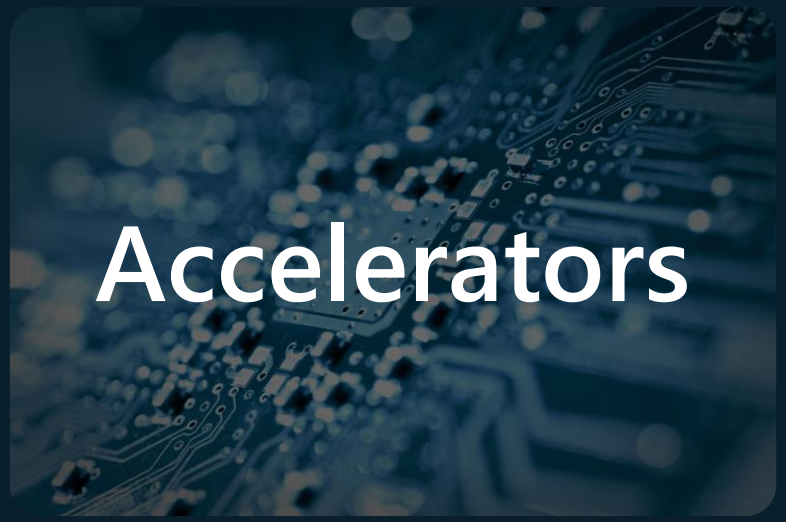

Accelerators


Cooling


Networking

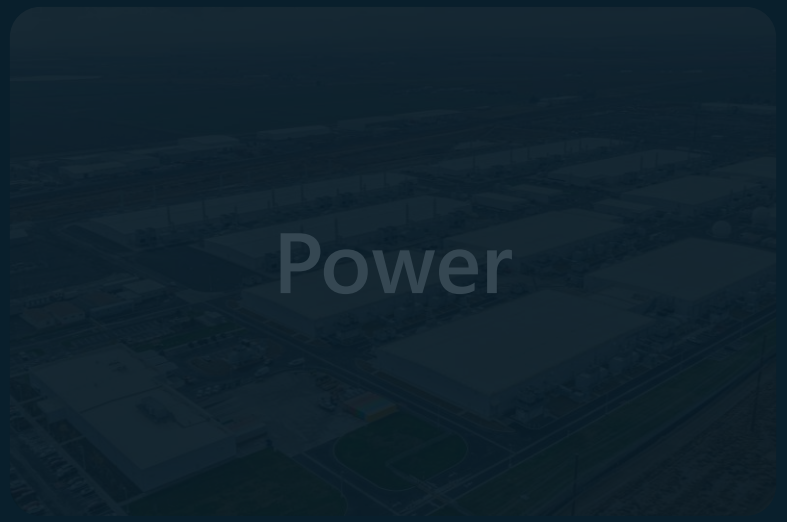

Power


Confidential AI


Accelerators


Cooling


Networking


Power

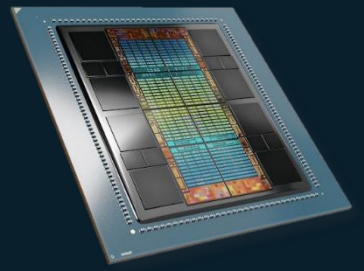

Confidential AI

Diverse accelerators on Azure

NVIDIA

A100, H100
(available today)
H200, GB200
(coming soon)

AMD



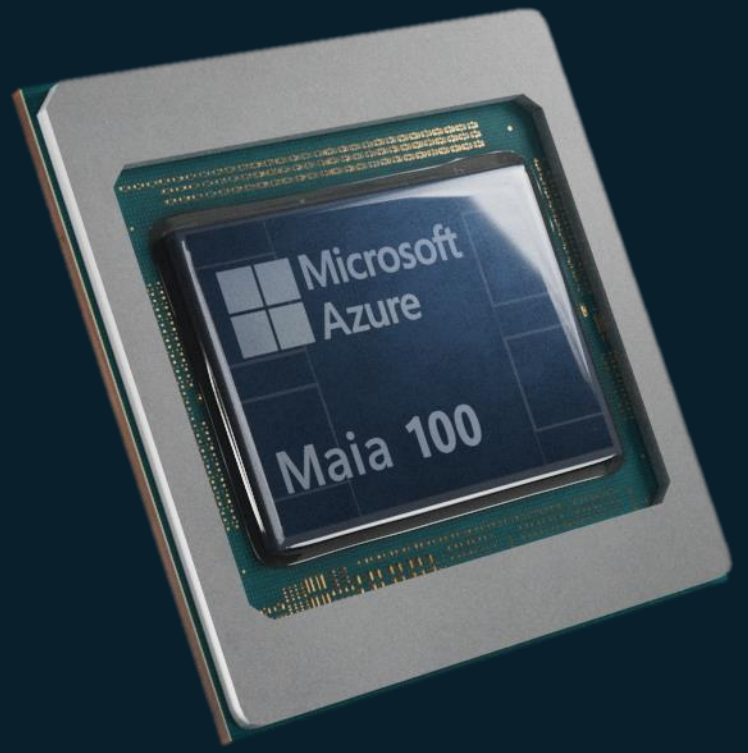
MI300x
(available today)

Microsoft

Azure Maia 100
(Internal)

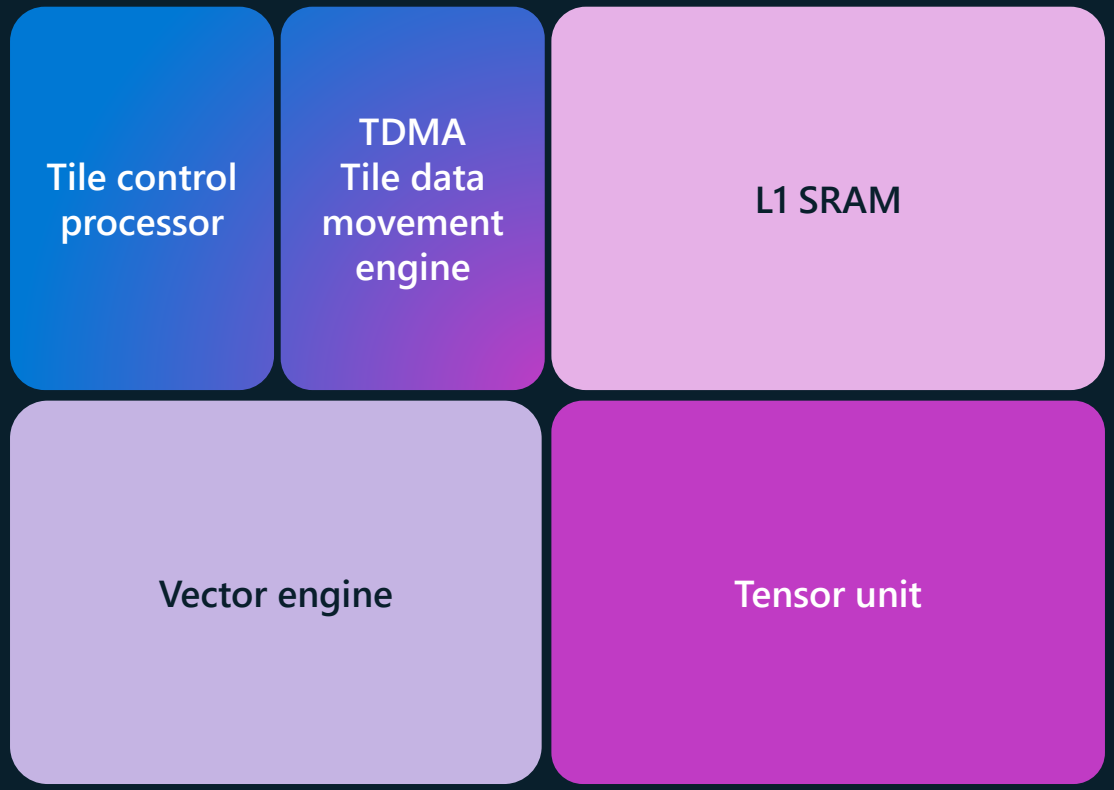
Maia 100 Specs

Maia 100 Specs	
Chip Size	~820mm2@N5
Package/Interposer Technology	TSMC COWOS-S
HBM BW/Cap	1.8TB/s @64GB HBM2E
Peak Dense Tensor POPS	6bit: 3 9bit: 1.5 BF16: 0.8
L1/L2	~500MB
Backend Network BW	600GB/s (12x400gbe)
Host BW (PCIe)	32GB/s PCIe Gen5x8
Design to TDP	700W
Provision TDP	500W

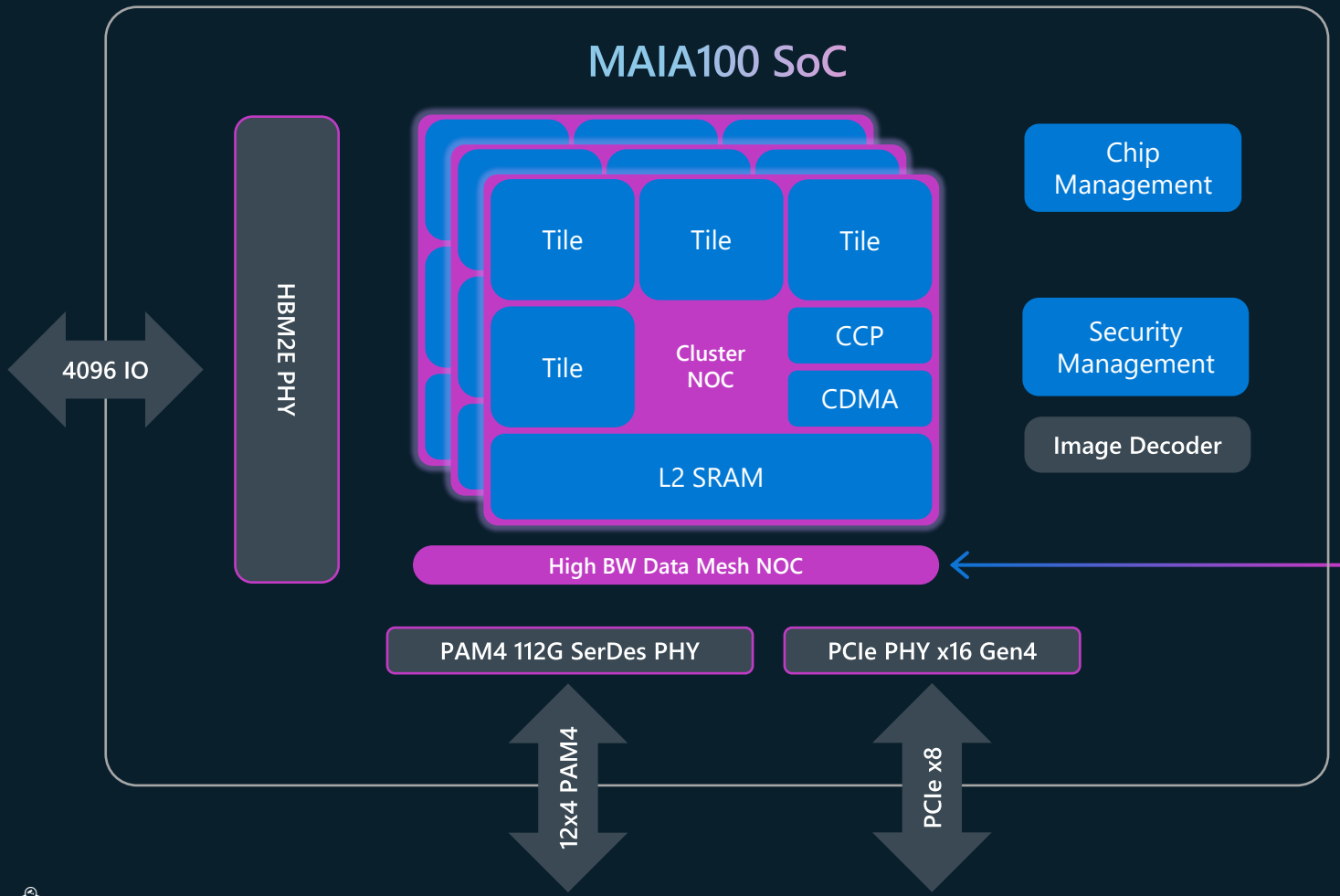


Inside Maia 100

Single tile



Inside Maia 100
SoC



4 Tiles per Cluster

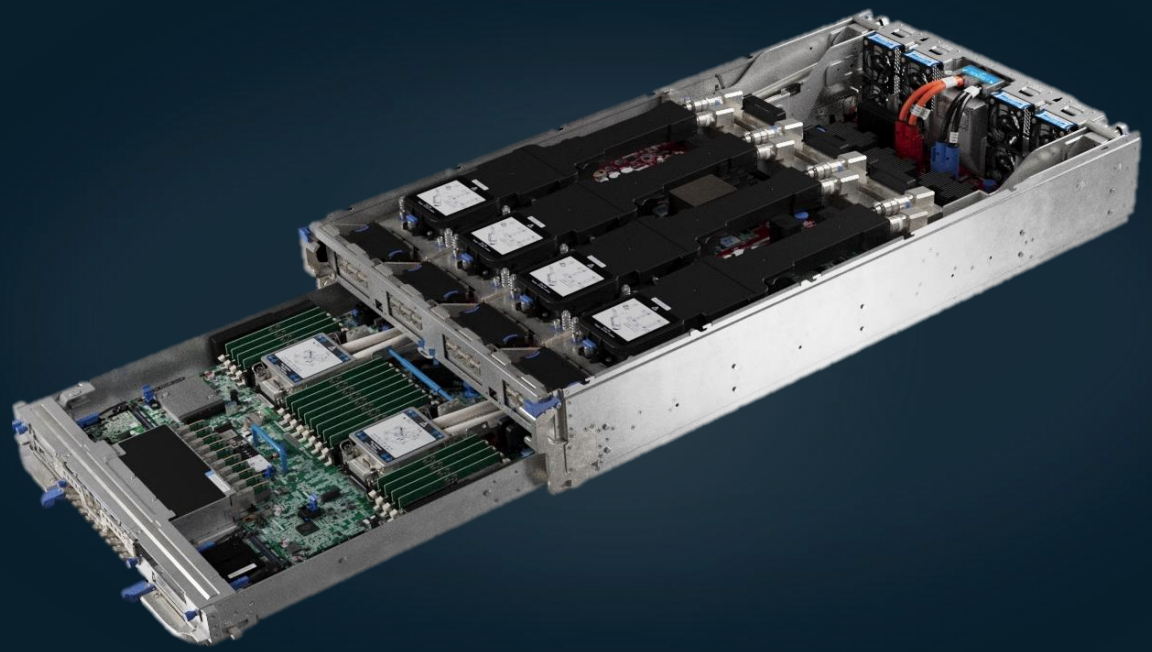
16 Clusters per Soc

Mesh-like NOC topology, with features optimized for ML

Great perf/W and performance

Maia 100

Server chassis



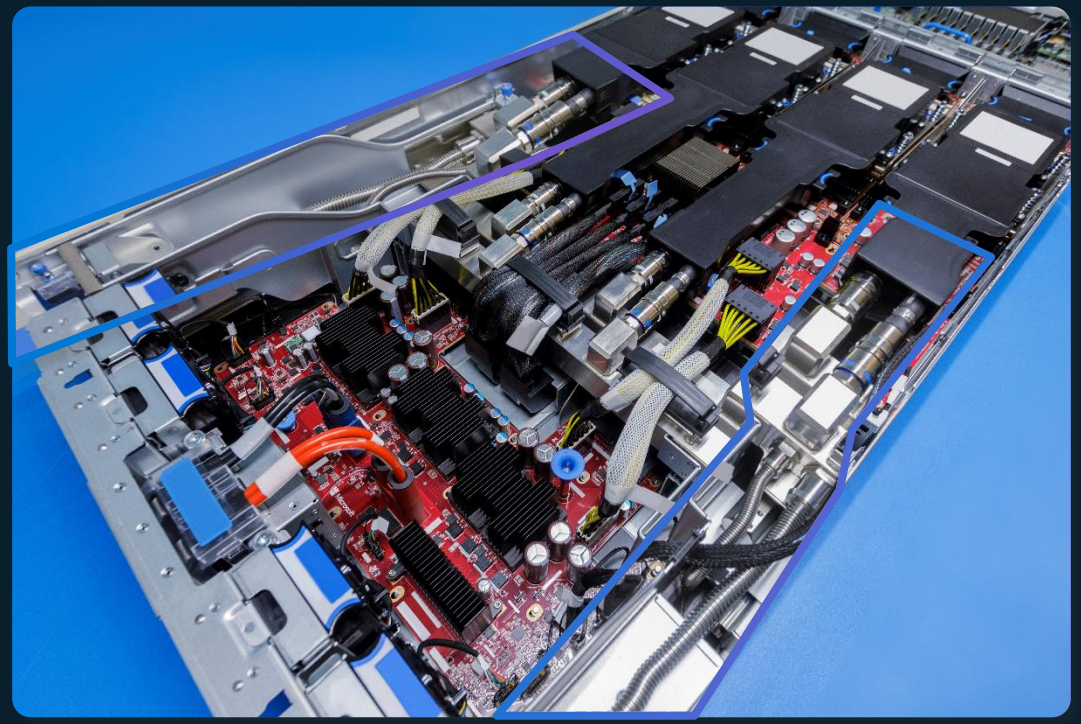
Accelerators

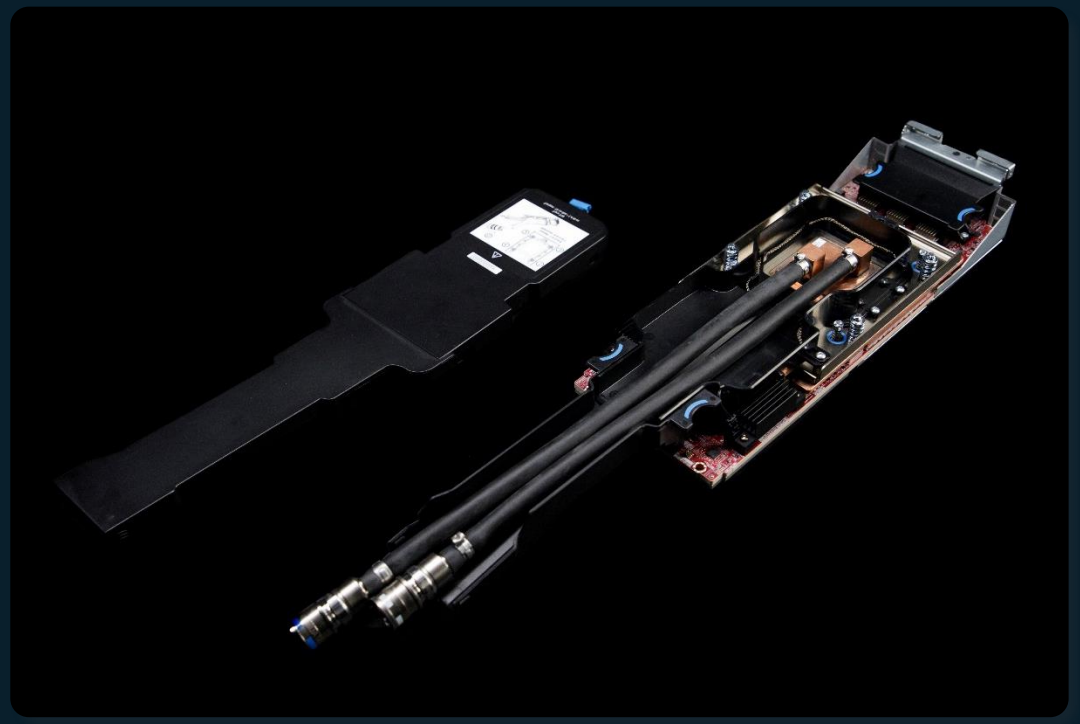
Networking

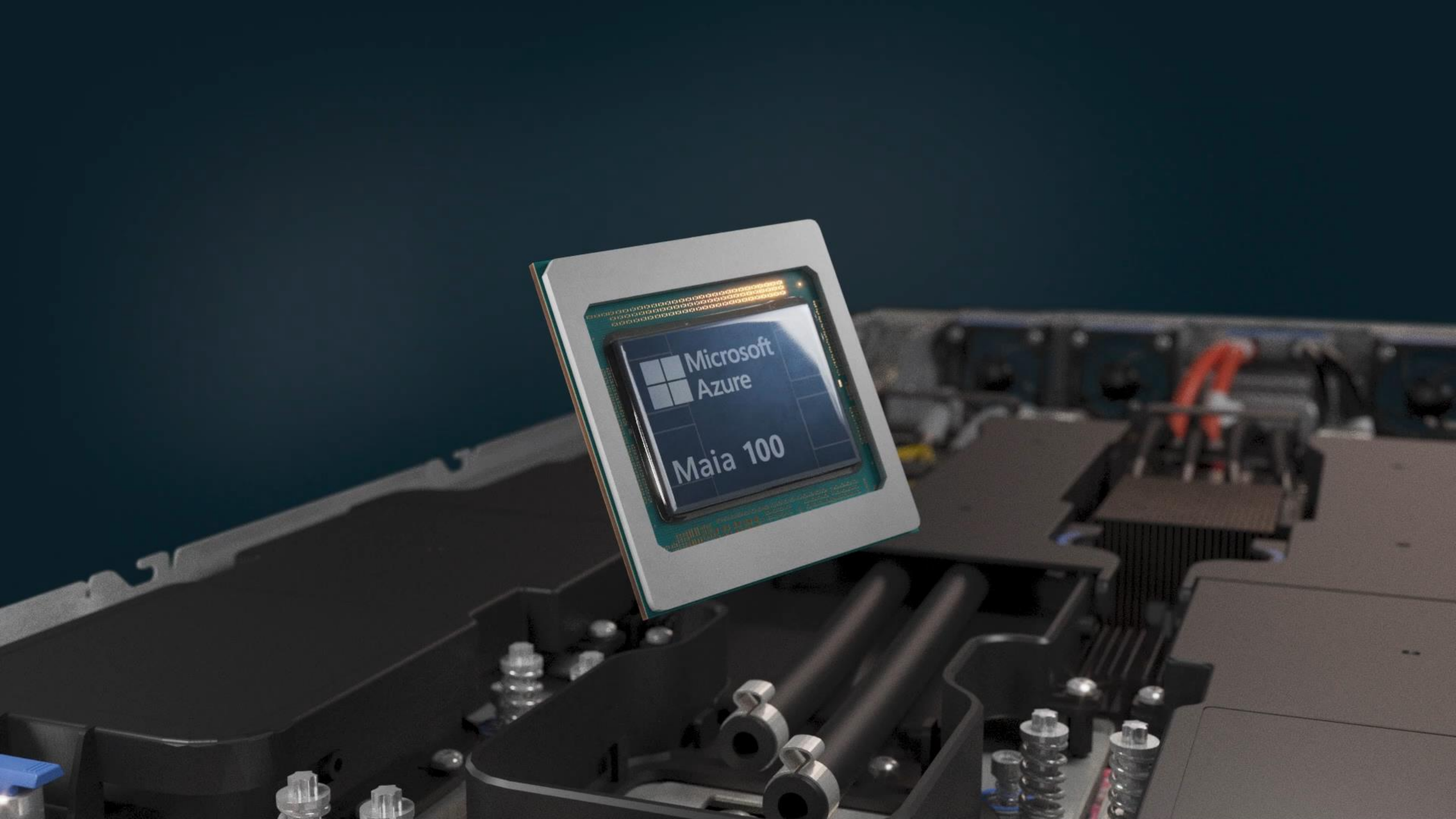
Power

Confidential AI

Maia liquid cooling

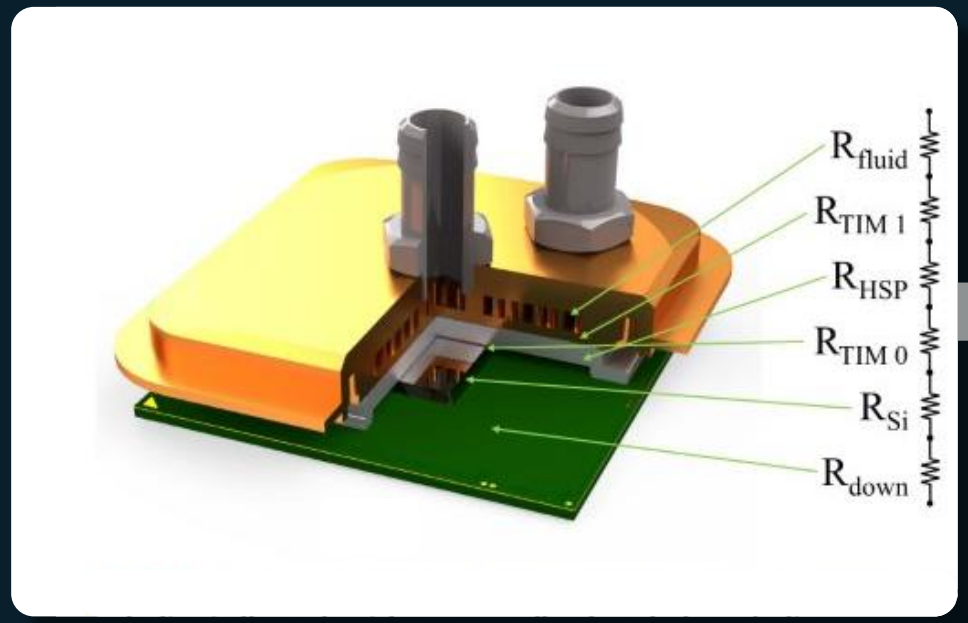




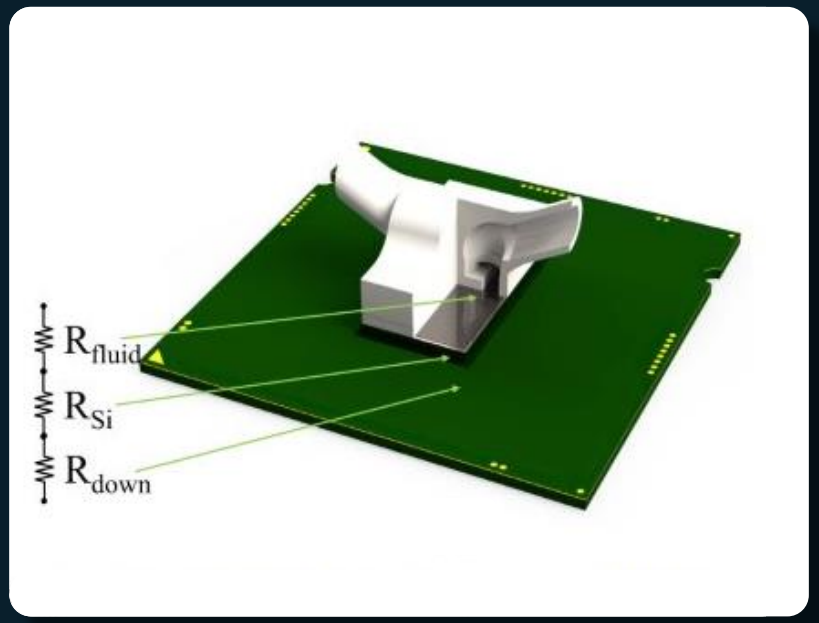

Microsoft
Azure
Maia 100

Microfluidics cooling

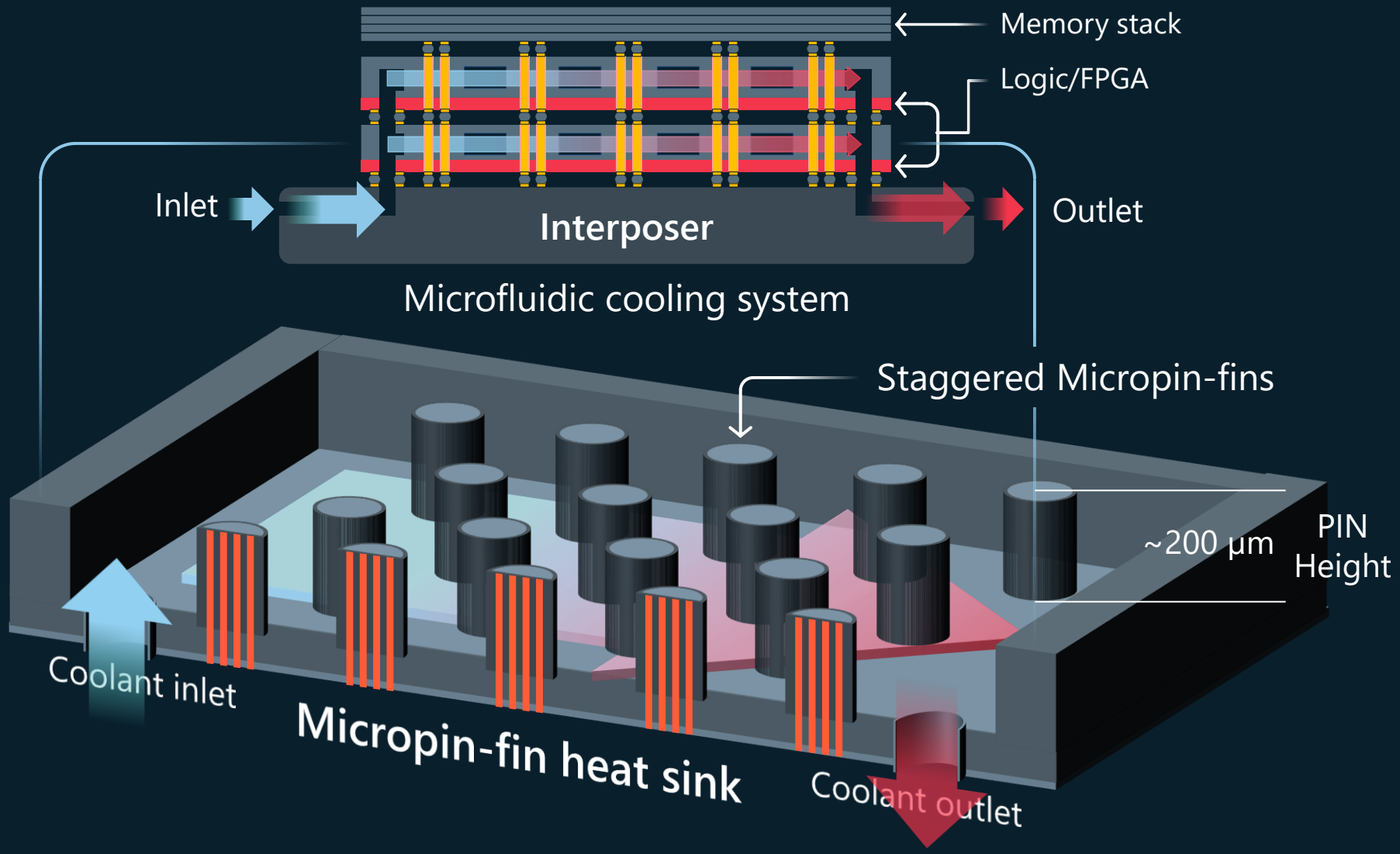
Cold plate



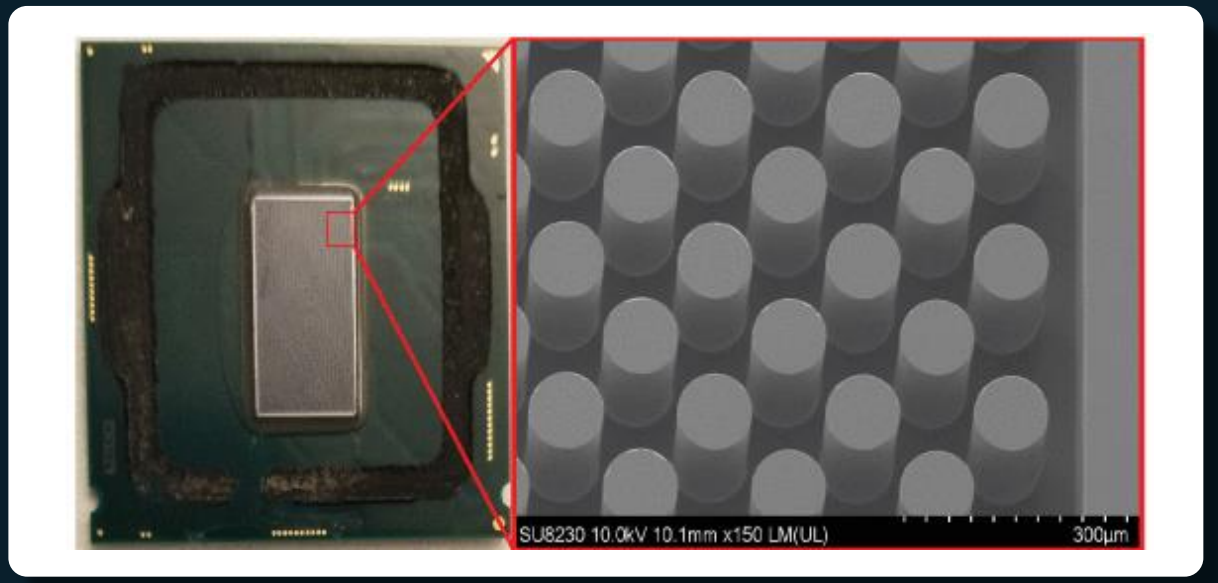
Microfluidics cooling



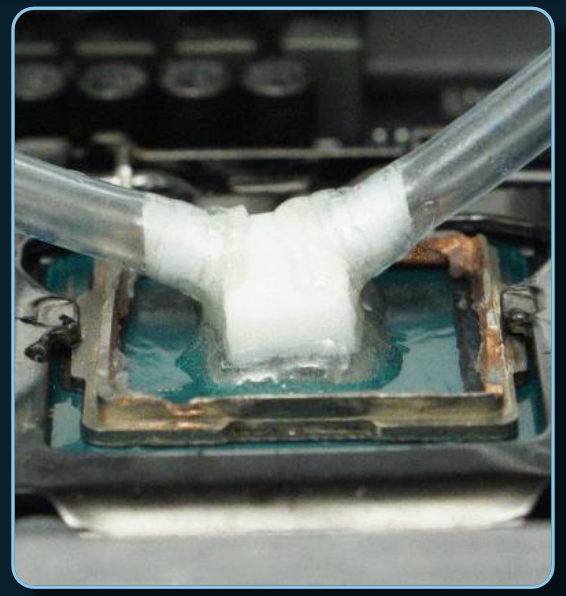
Microfluidics cooling



Microfluidics cooling



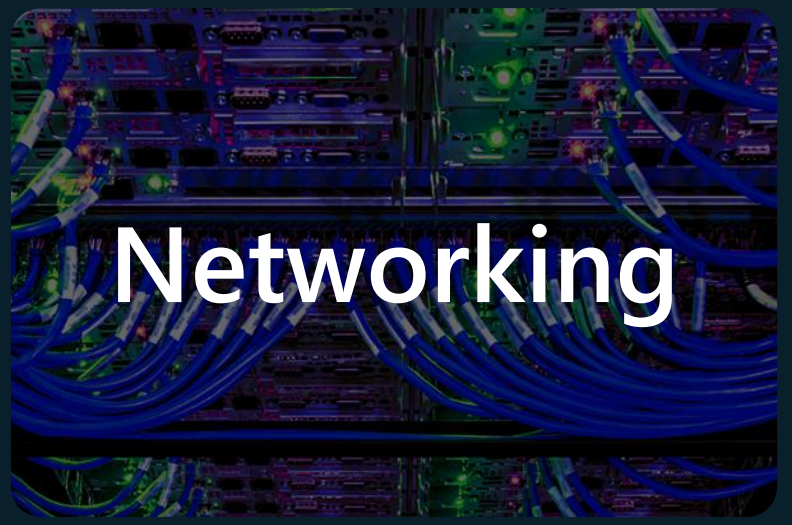
Micropin-fins
on CPU



Fluid
delivery tubes
connected

Accelerators

Cooling

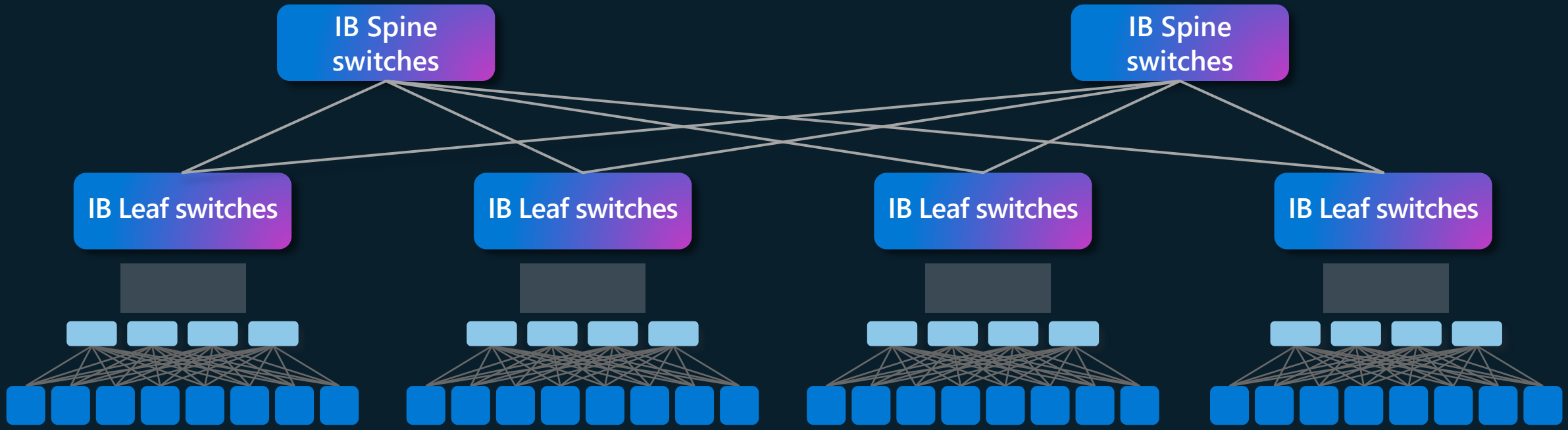


Power

Confidential AI

Infiniband networking

Intra-cluster networking

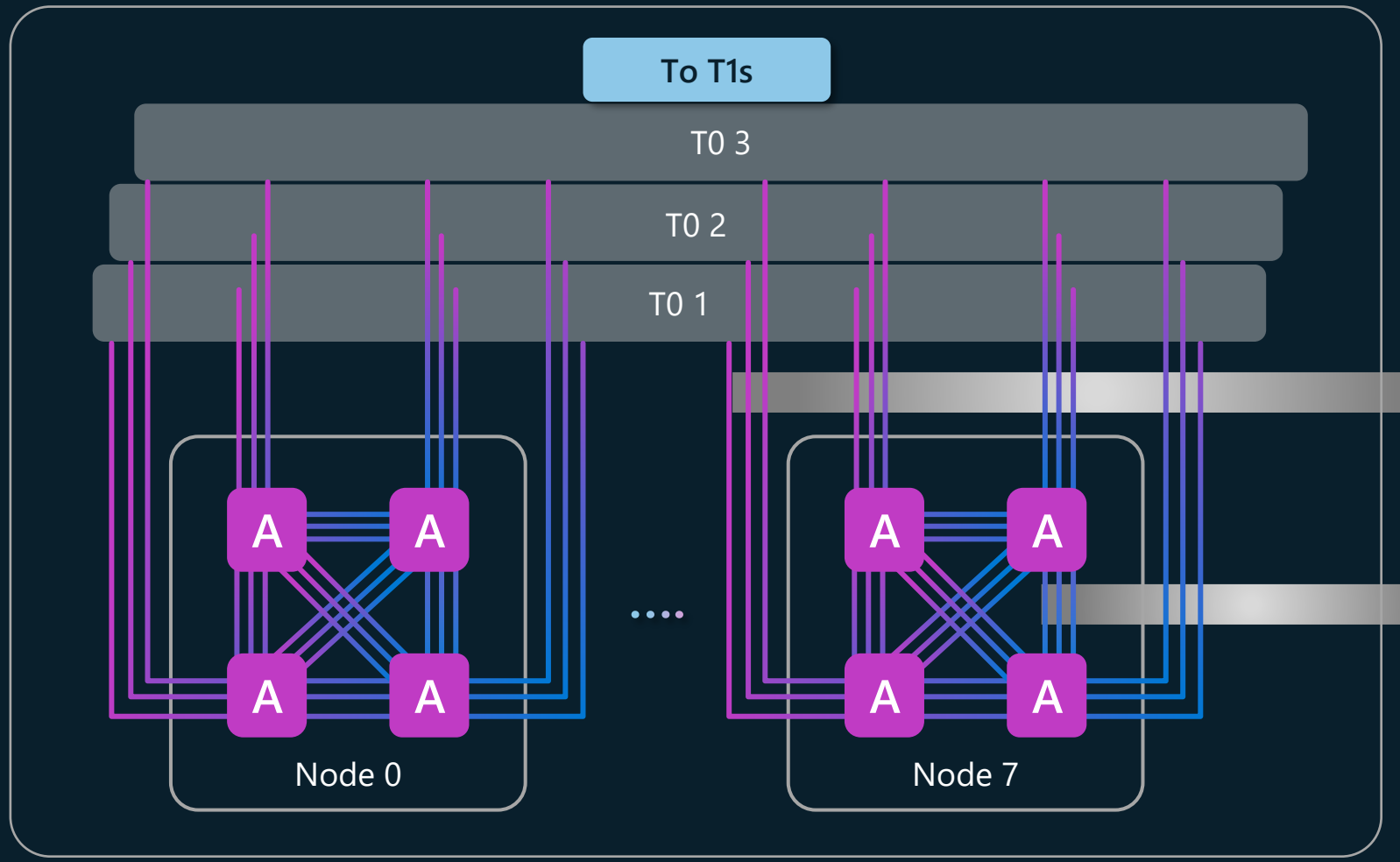


Ethernet

Ultra Ethernet Consortium

Open / multivendor	Switches, NICs, cables, optics, tools, software
Scalable	Addressing and routing for rack-, building-, DC- scale networks
Tools	Many tools for testing, operations, measurements
Cost	Economies of scale and competitive market
Supporting standards	Regular progress in IEEE, for many technologies, across layers

Azure Maia Ethernet based networking



All-gather/scatter-reduce at 4800Gbps

Any-to-any at 1200Gbps

400Gbps each line

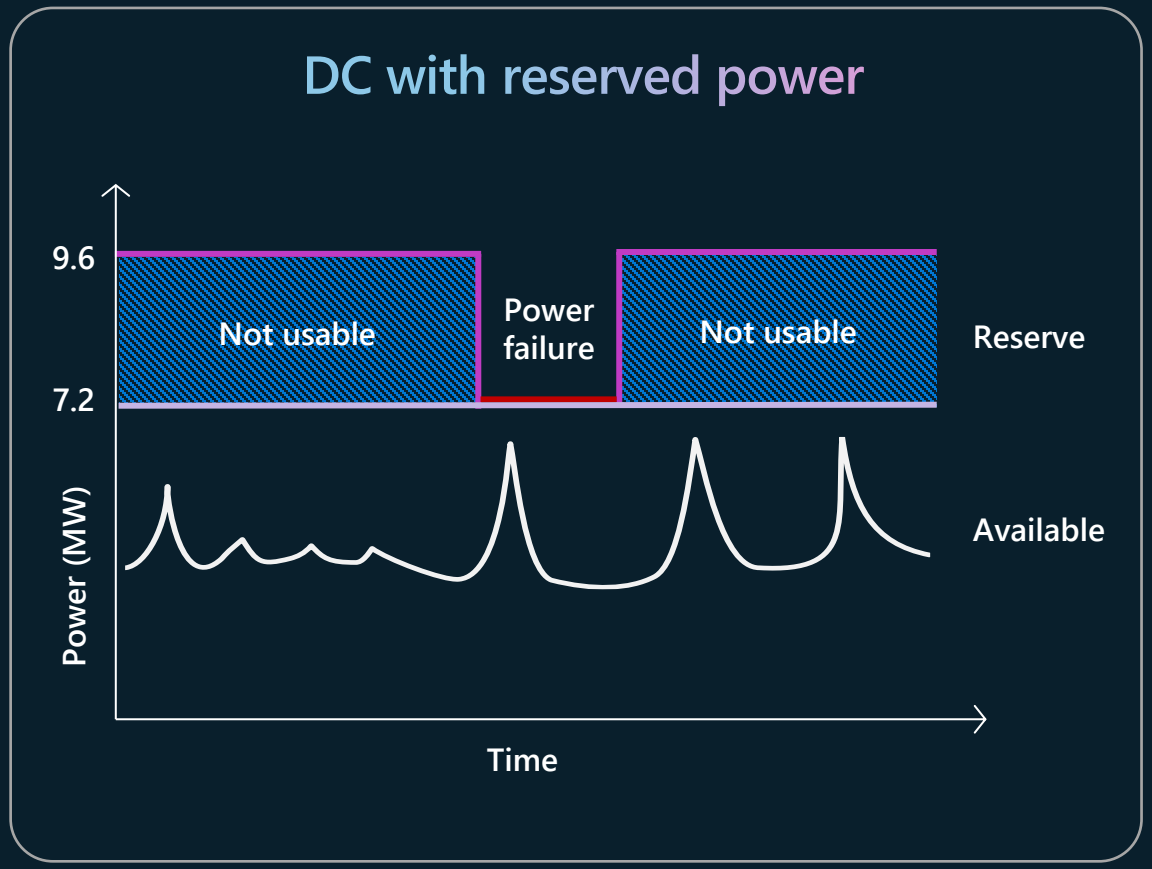
Accelerators

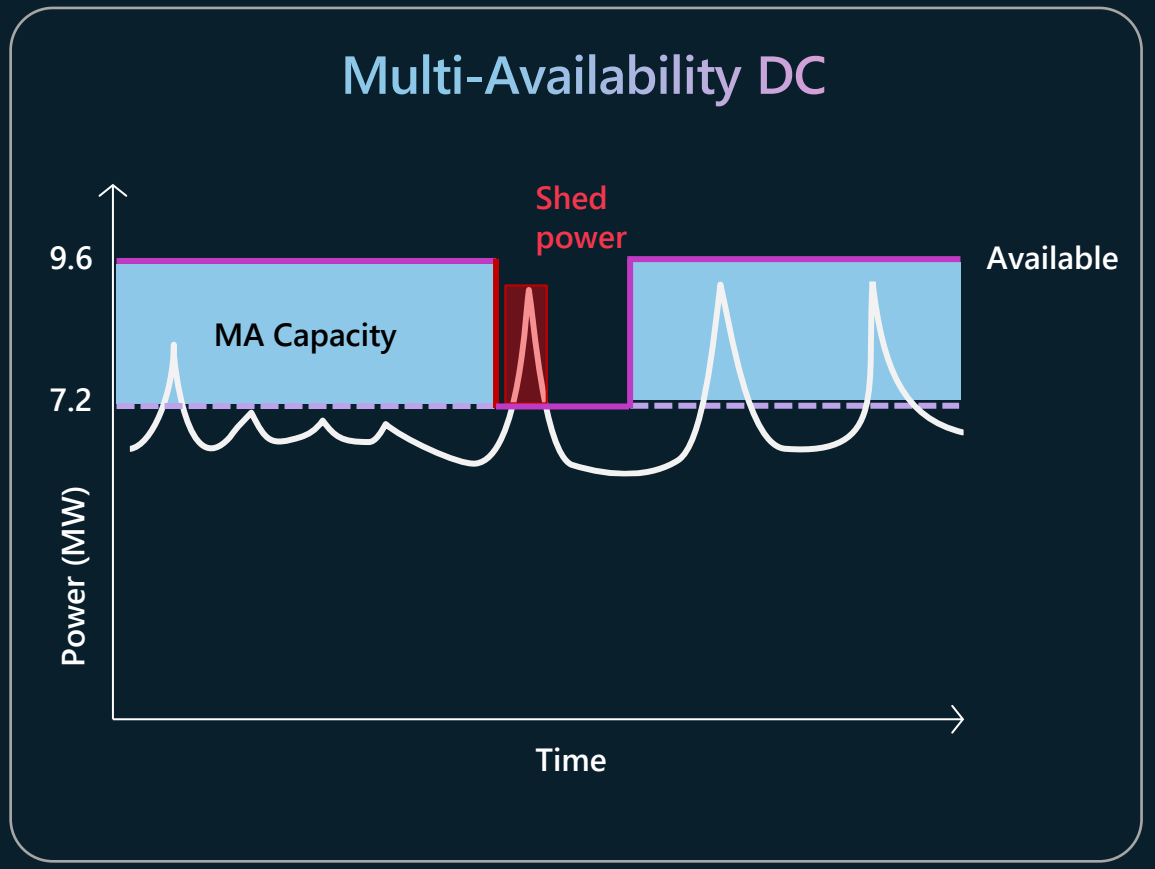
Cooling

Networking

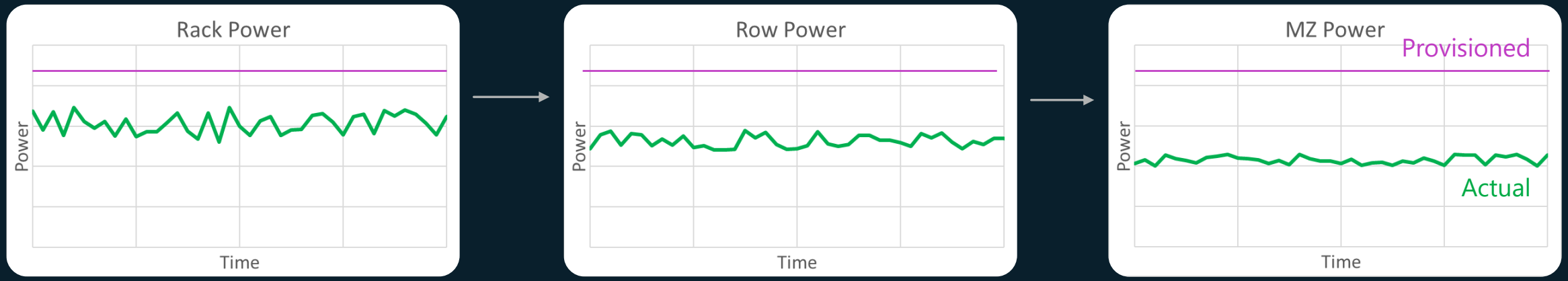
Confidential AI

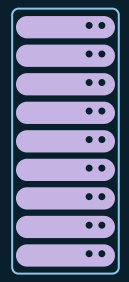
Multi-Availability Datacenters



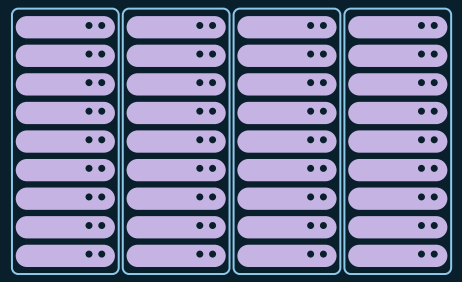


Datacenter Power Slack

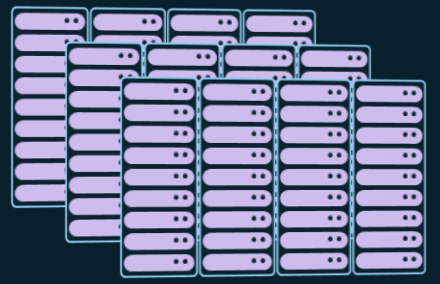




Rack

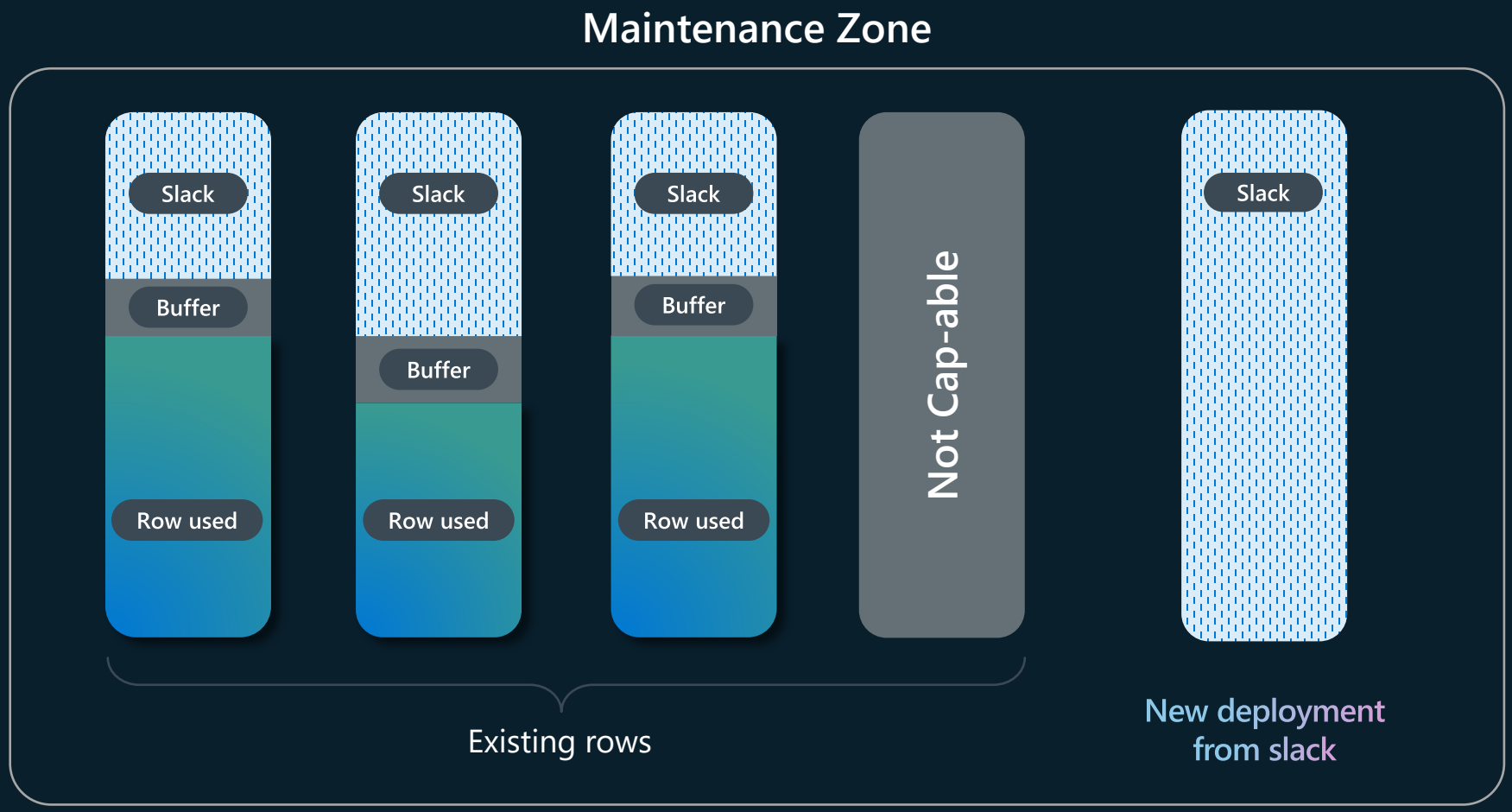


Row



Maintenance Zone

Hierarchical power management



If row power gets to limit, cap by throttling servers

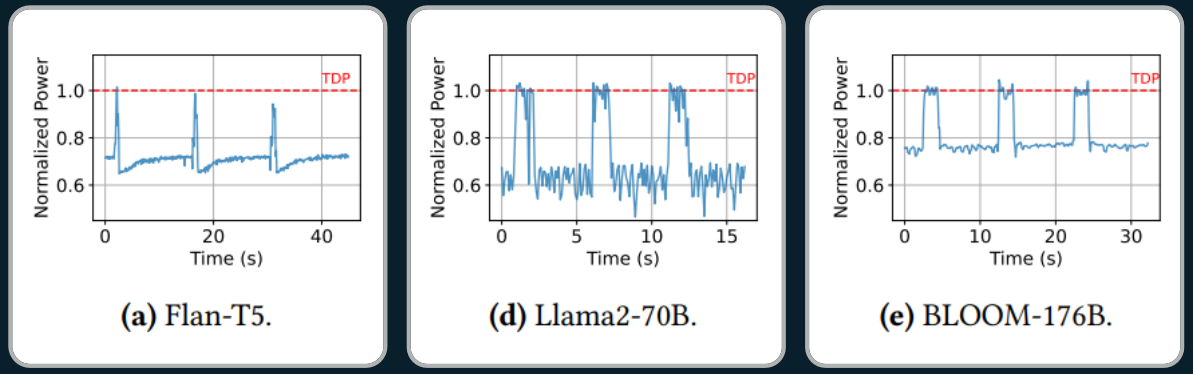
Project POLCA

Power oversubscription for inference

GPU servers are provisioned for peak power draw

3% power headroom for training
21% power headroom for inference

30% more servers in existing clusters with minimum performance loss

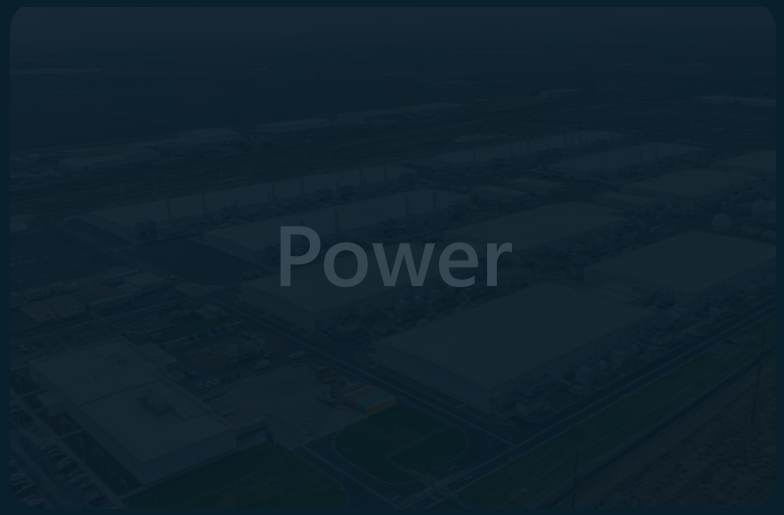


Inference GPU power usage


Accelerators


Cooling


Networking


Power


Confidential AI

Confidential computing

Existing encryption

Data at rest

Encrypt inactive data when stored in blob storage, database, etc.

Data in transit

Encrypt data that is flowing between untrusted public or private networks

Confidential computing

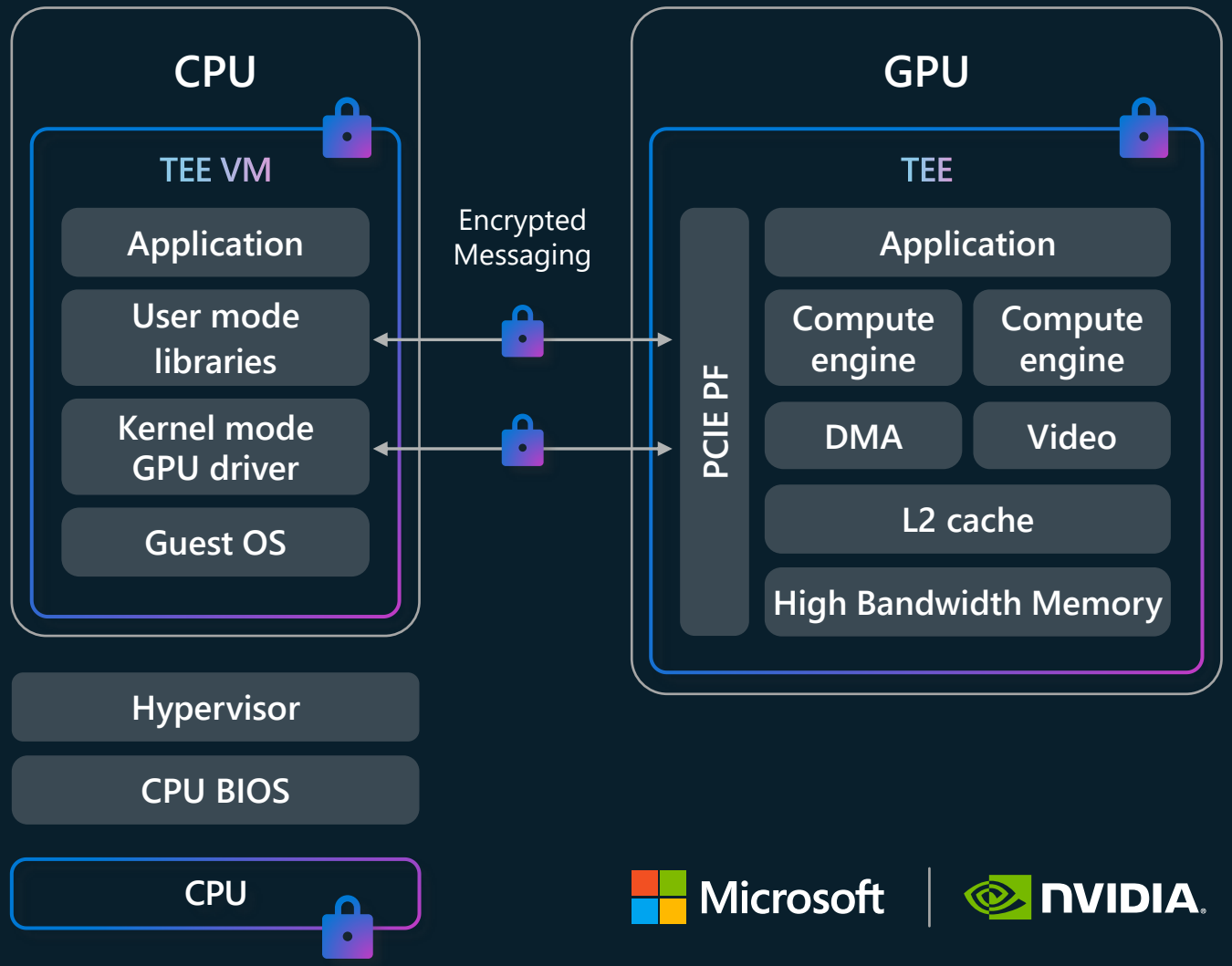
Data in use

Protect/encrypt data that is in use, while in RAM, and during computation

Preview

Azure Confidential GPU VMs powered by NVIDIA

Confidential VMs with NVIDIA Tensor Core H100 GPU



Microsoft | NVIDIA

Confidential AI

Model

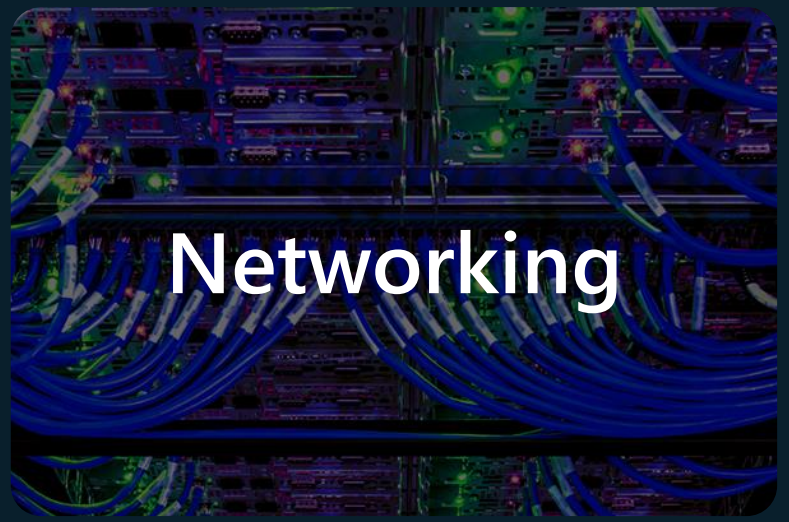
Data

Training / Fine-tuning / Inference

Multi-party sharing


Accelerators


Cooling


Networking


Power


Confidential AI