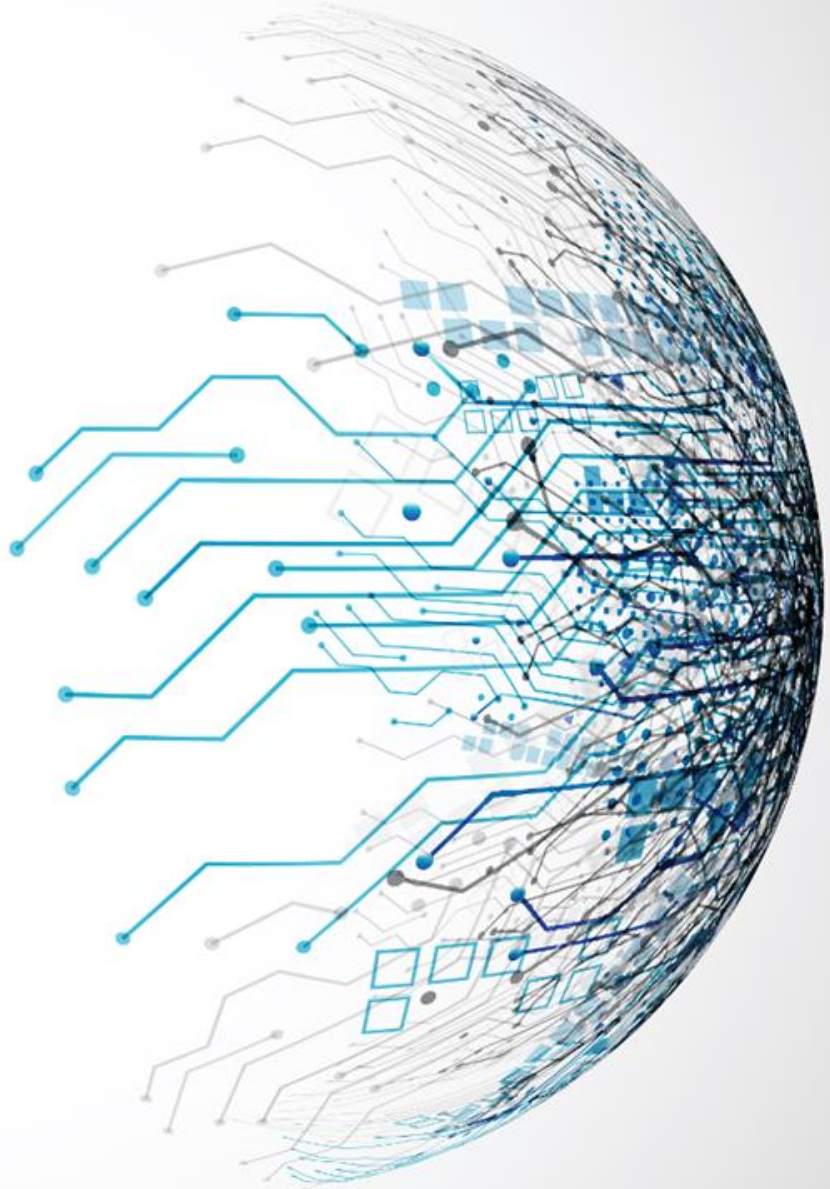
Deep Learning

HARDWARE ARCHITECTURE



Dr. Mohammed Salah Al-Radhi

(slides by: Dr. Bálint Gyires-Tóth)

Copyright

Outline

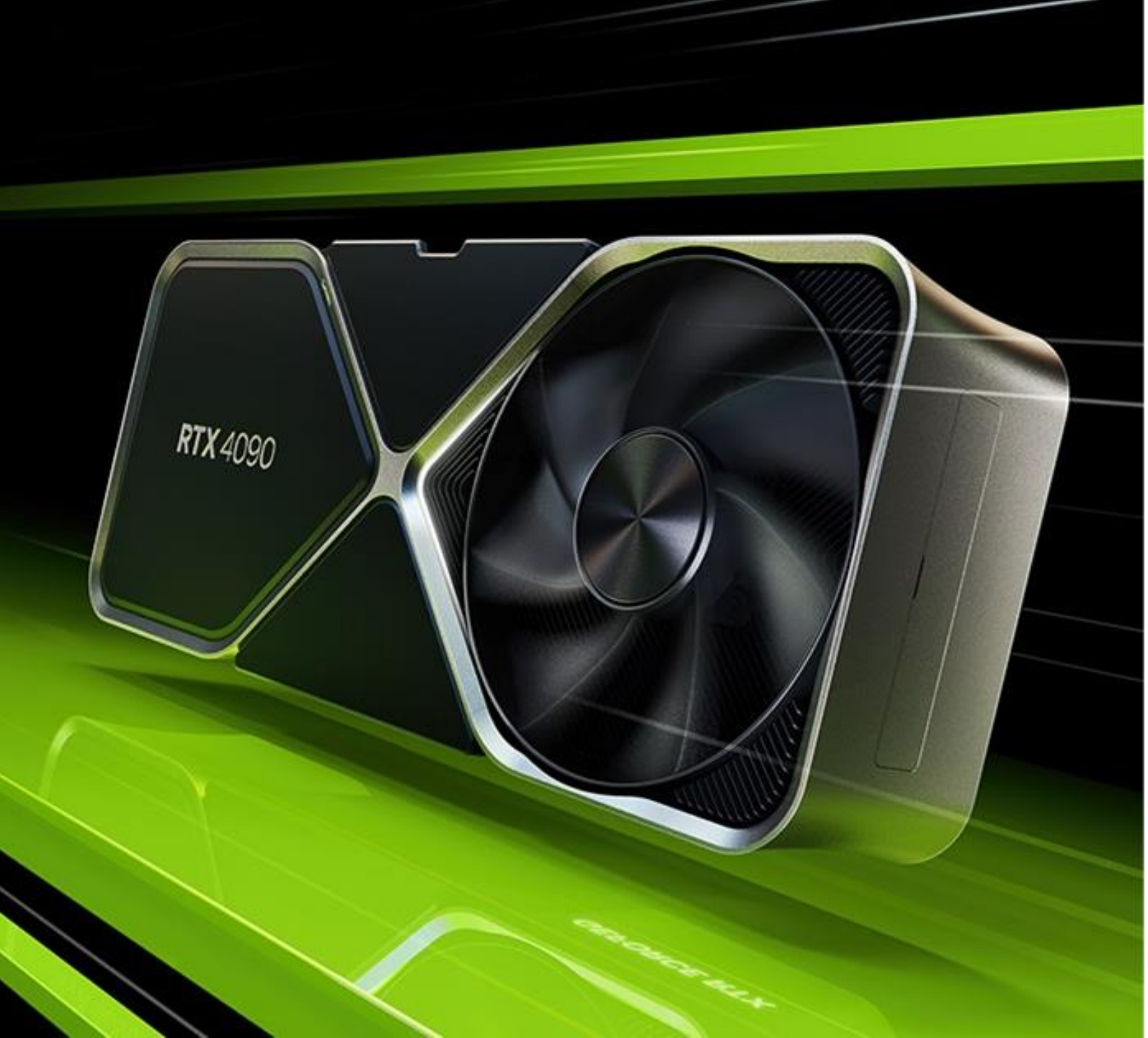
- Why?
- NVIDIA GPUs
  - Consumer grade GPUs
  - Semi-professional GPUs
  - Server grade GPUs
- Inference servers
- Cloud
- Cost planning

Why?

Why?

- Neural networks perform matrix operations in forward and backward steps
- Many of these operations can be done elementwise or in chunks
- These elements / chunks can be assigned to multiple cores
- Modern GPUs have 1000s of cores, which are much faster in parallel computation than CPUs with a few 100 cores.

Consumer grade GPUs

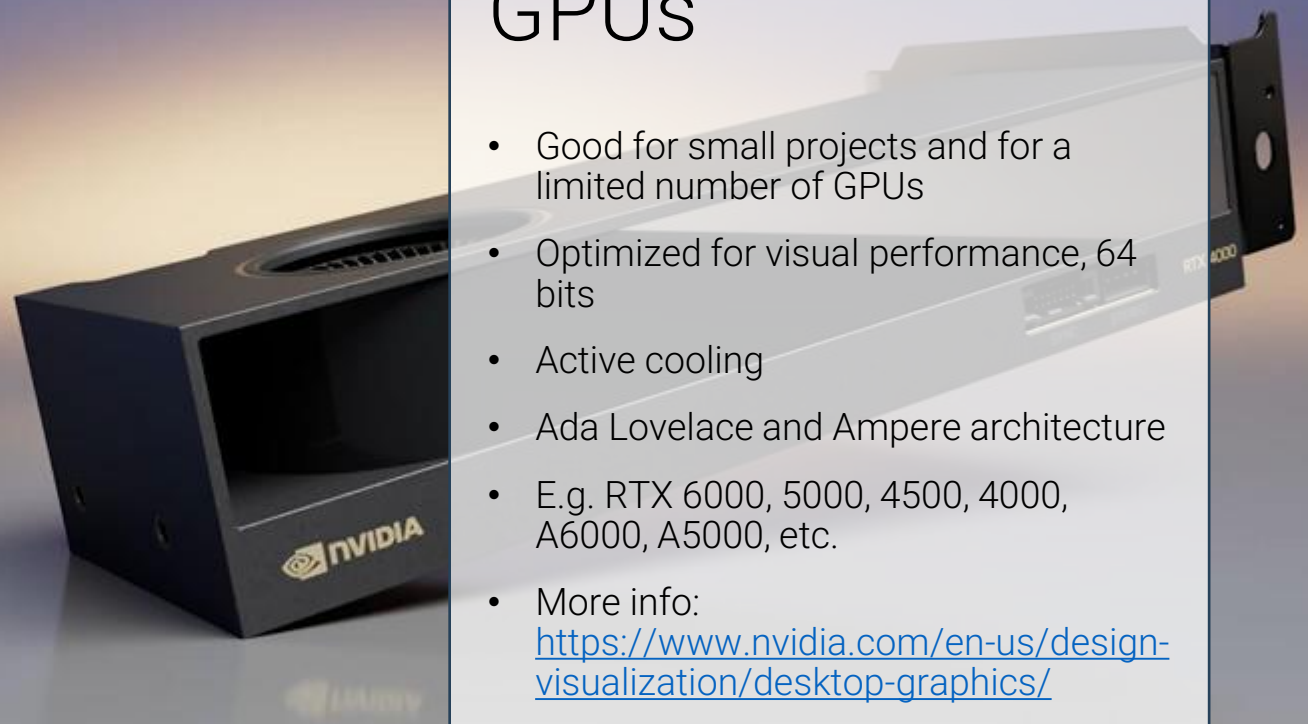


Consumer grade GPUs

- Good for small projects and for a limited number of GPUs
- Limited features and GPU RAM
- Optimized for visual performance
- Active cooling
- Most popular A0B0 series, where A=1,2,3,4 and B=6,7,8,9
- E.g. top pick in 2023: NVIDIA RTX 4090 24GB
- More info: https://www.nvidia.com/en-eu/geforce/graphics-cards/

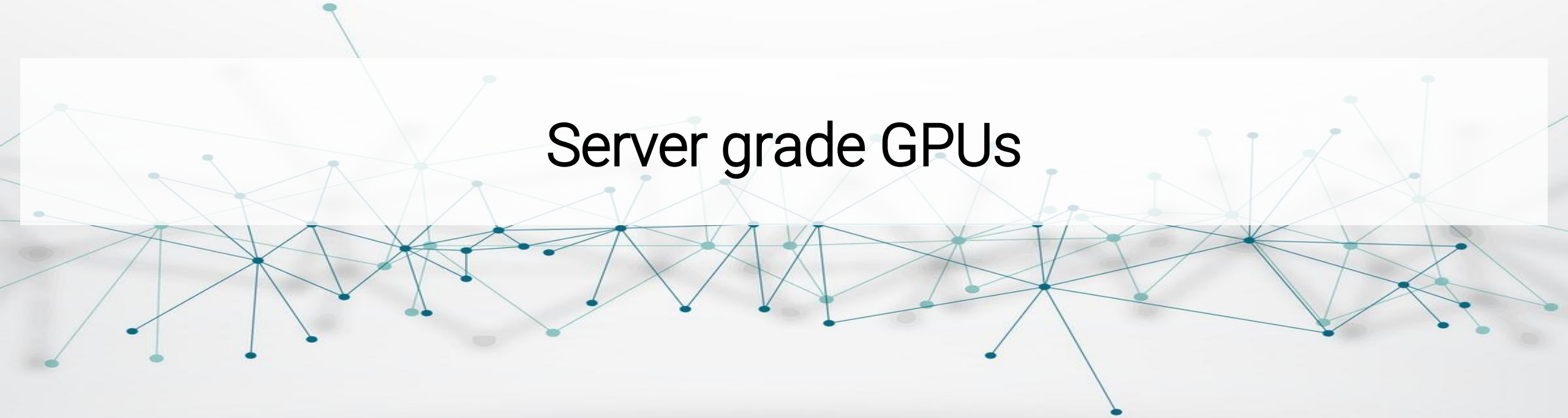
Semi-professional GPUs

Semi professional GPUs



- Good for small projects and for a limited number of GPUs
- Optimized for visual performance, 64 bits
- Active cooling
- Ada Lovelace and Ampere architecture
- E.g. RTX 6000, 5000, 4500, 4000, A6000, A5000, etc.
- More info: https://www.nvidia.com/en-us/design-visualization/desktop-graphics/

Server grade GPUs



Server grade GPUs: Hopper architecture

- „Seamless” integration
  - PCIe 6.0 16 lanes: 128 Gb/s
  - In-GPU: 3.35/2/7.8 TB/s (SMX/PCIe/NVL)
  - Multi GPU: NVSwitch 900 Gb/s
  - Multi node: Infiniband/Ethernet 400Gb/s
- Passive cooling
- DGX H100 off-the-shelf server
- BasePod and SuperPods can be built.
- More info:
  - GPU: https://www.nvidia.com/en-us/data-center/h100/
  - Server: https://www.nvidia.com/en-us/data-center/dgx-h100/

NVIDIA DGX SuperPod

- Off-the-shelf solution for AI infrastructure
- DGX H100 SuperPod: https://docs.nvidia.com/nvidia-dgx-superpod-data-center-design-dgx-h100.pdf

Range	Class	Dry-Bulb Temperature	Humidity Range, Non-Condensing	Maximum D
Recommended	All A	64.4–80.6 °F 18–27 °C	41.9 °F to 60% RH and 59 °F DP 5.5 °C to 60% RH and 15 °C DP	5 1
Allowable up to 30 °C for DGX H100 Systems	A1	59–89.6 °F 15–32 °C	20–80% RH	6 1
Allowable per ASHRAE for various other classes of data center and telecom environments	A2	50–95 °F 10–35 °C	20–80% RH	6 2
	A3	41–104 °F 5–40 °C	10.4 °F DP and 8–85% RH -12 °C DP and 8–85% RH	7 2
	A4	41–113 °F 5–45 °C	10.4 °F DP and 8–90% RH -12 °C DP and 8–90% RH	7 2
	B	41–95 °F 5–35 °C	8–80% RH	8 2
	C	41–104 °F 5–40 °C	8–80% RH	8 2

Table 6. ISO 14644-1 standard for air cleanliness classifications

Class	Particle Size[1]					
	> 0.1 µm	> 0.2 µm	> 0.3 µm	> 0.5 µm	> 1 µm	> 5 µm
						29
						293
						2,930
						29,300
						93,000

Table 8. Common distribution schemes compatible with DGX H100 racks

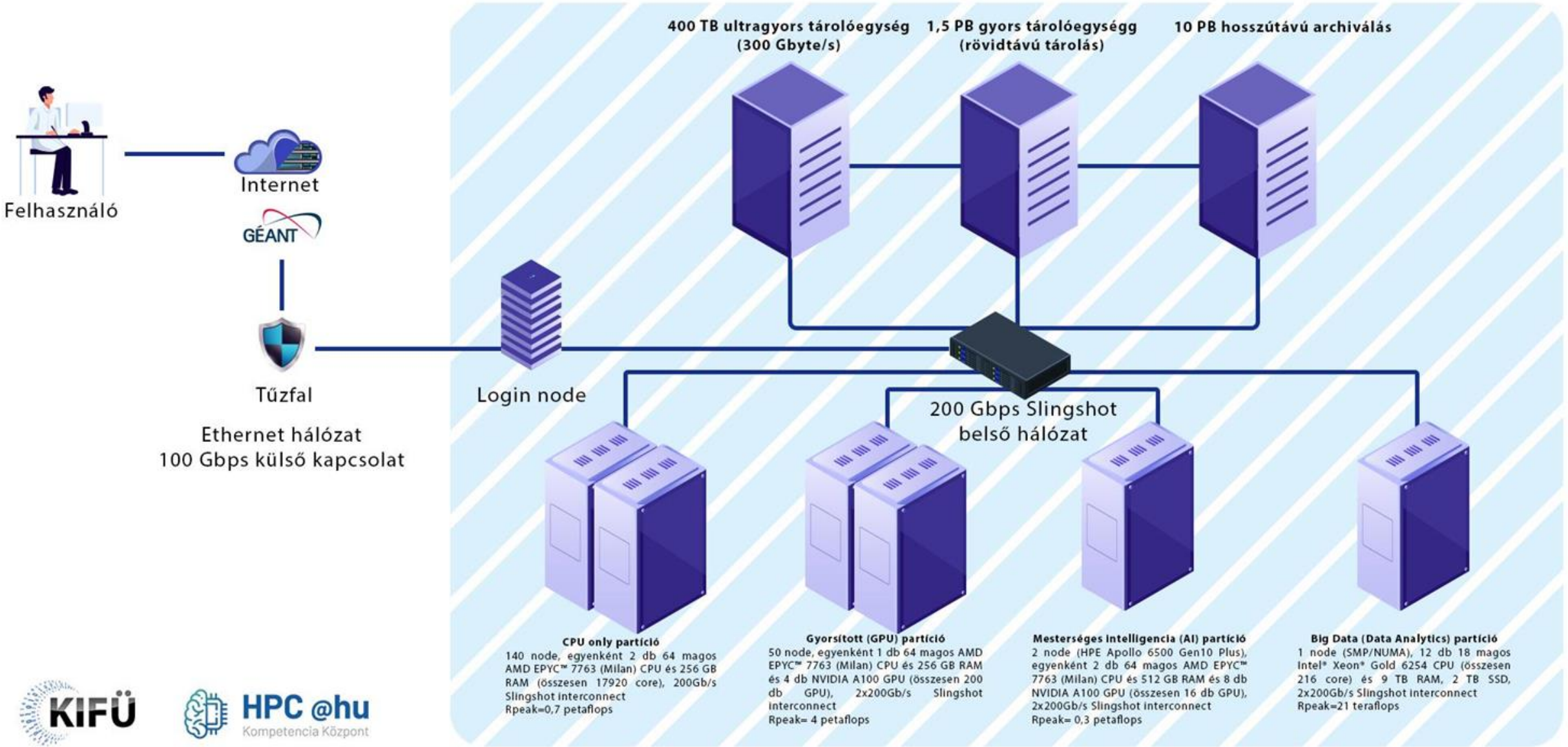
Phase	Distribution Voltage	Line Voltage	Amps	Breaker Derating	Circuit Capacity kW[1]	Maximum Supported DGX H100 Systems per Rack[2,3]	Peak Server Demand per Circuit kW[2]	Stranded Capacity at Peak Demand kW[2]
1Φ	230	230	63	100%	13.7	2	10.2	3.5
3Φ Delta	208	208	60	80%	32.8	4	20.4	12.4
3Φ Wye	400	230	32	100%	21	4	20.4	0.6
3Φ Wye	415	240	32	100%	21.8	4	20.4	1.4
3Φ Wye	415	240	60	80%	32.7	4	20.4	12.3


summit
IBM

- 760 NVIDIA DGX A100
- 6,080 NVIDIA A100 GPU
- linked on an NVIDIA Quantum 200Gb/s InfiniBand network
- 1,895 petaflops of TF32 performance
- 5 exaflops of mixed precision AI performance


komondor
KIFÜ
HPC @hu

A KOMONDOR SZUPERSZÁMÍTÓGÉP FELÉPÍTÉSE



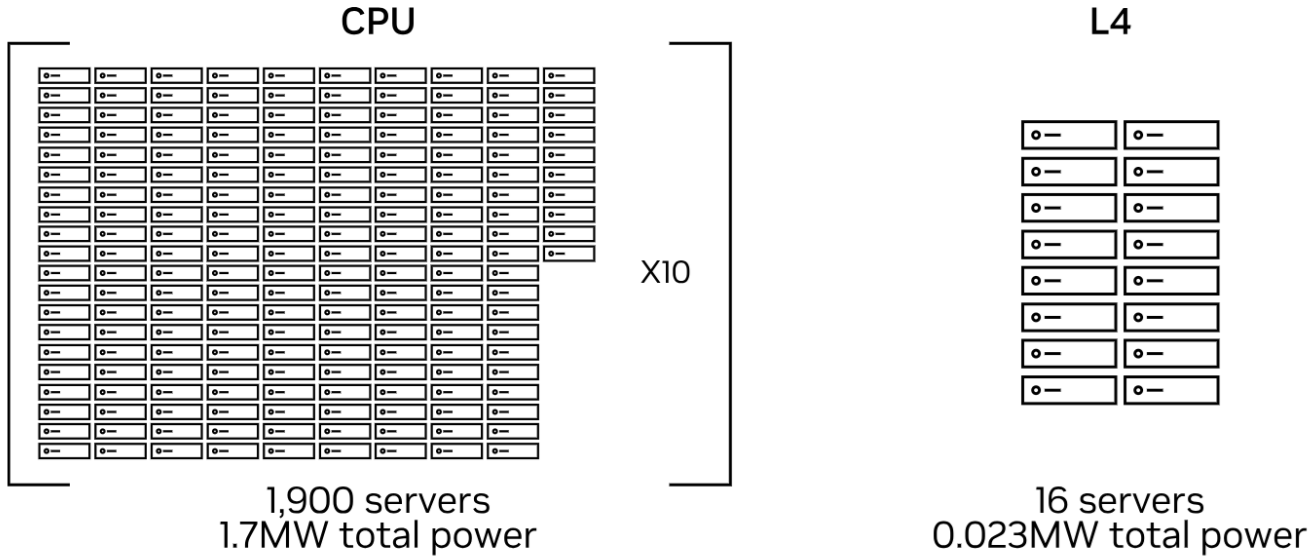
KIFÜ

HPC @hu
Kompetencia Központ

Inference servers

Inference servers

- Increasing demand as AI solutions go to production phase
- General HW vs dedicated inference specific HW
- Example 1: NVIDIA H100 NVL
  - 2x94 GB RAM, 7.8 TB/s GPU memory bandwidth
  - Fits 175B LLMs in GPU memory
- Example 2: L4 GPU
  - Energy efficient (72W max, 1-slot)



GPU in cloud

GPUs in cloud

- Google Cloud, AWS, MS Azure, Oracle, NVIDIA DGX Cloud, Others.
- Flexible, good for scaling
- Large-scale trainings are non-trivial
- Great for inference options
- Comparison:
  - https://fullstackdeeplearning.com/cloud-gpus/
  - https://cloud-gpus.com/

Cost planning

	Submitter	System	Processor	#	Accelerator
Available Cloud					
3.0-2000	NVIDIA+CoreWeave	coreweave_hgxh100_n192_ngc23.04_pytorch	Intel(R) Xeon(R) Platinum 8462Y+	384	NVIDIA H100-SXM
3.0-2001	NVIDIA+CoreWeave	coreweave_hgxh100_n384_ngc23.04_pytorch	Intel(R) Xeon(R) Platinum 8462Y+	768	NVIDIA H100-SXM
3.0-2002	NVIDIA+CoreWeave	coreweave_hgxh100_n448_ngc23.04_mxnet	Intel(R) Xeon(R) Platinum 8462Y+	896	NVIDIA H100-SXM
3.0-2003	NVIDIA+CoreWeave	coreweave_hgxh100_n448_ngc23.04_pytorch	Intel(R) Xeon(R) Platinum 8462Y+	896	NVIDIA H100-SXM
3.0-2004	NVIDIA+CoreWeave	coreweave_hgxh100_n96_ngc23.04_pytorch	Intel(R) Xeon(R) Platinum 8462Y+	192	NVIDIA H100-SXM
Available On premise					
3.0-2004	ASUSTeK	ESC4000-E11-4xA100-PCIE-80GB	Intel(R) Xeon(R) Platinum 8462Y+	2	NVIDIA A100-PCIe
3.0-2005	ASUSTeK	ESC4000-E11-4xA100-PCIE-80GB-NVBridge	Intel(R) Xeon(R) Platinum 8462Y+	2	NVIDIA A100-PCIe
3.0-2006	ASUSTeK	ESC8000A-E12-8xH100-PCIE-80GB	AMD EPYC 9654 96-Core	2	NVIDIA H100-PCIe
3.0-2007	H3C	R4900G6x2A30-PCIE-24GB	Intel(R) Xeon(R) Platinum 8490H CPU @ 1.90GHz	2	NVIDIA A30-PCIE-
3.0-2008	H3C	R5300G6x8A30-PCIE-24GB	Intel(R) Xeon(R) Platinum 8458P	2	NVIDIA A30-PCIE-
3.0-2009	H3C	R5350G6x8A30-PCIE-24GB	AMD EPYC 9754 128-Core Processor	2	NVIDIA A30-PCIE-
3.0-2010	H3C	R5350G6x8A30-PCIE-24GB	AMD EPYC 9754 128-Core Processor	2	NVIDIA A30-PCIE-
3.0-2011	Intel	16-nodes-SPR-pytorch	Intel(R) Xeon(R) Platinum 8480+ @ 2.00GHz	32	N/A
3.0-2012	Intel	8-nodes-SPR-pytorch	Intel(R) Xeon(R) Platinum 8480+ @ 2.00GHz	16	N/A
3.0-2013	Intel-HabanaLabs	HLS-Gaudi2-N32-PT	Intel(R) Xeon(R) Platinum 8380	64	Habana Gaudi2
3.0-2014	Intel-HabanaLabs	HLS-Gaudi2-N48-PT	Intel(R) Xeon(R) Platinum 8380	96	Habana Gaudi2
3.0-2015	Intel-HabanaLabs	HLS-Gaudi2-N8-PT	Intel(R) Xeon(R) Platinum 8380	16	Habana Gaudi2
3.0-2016	Intel-HabanaLabs	HLS-Gaudi2-PT	Intel(R) Xeon(R) Platinum 8380	2	Habana Gaudi2
3.0-2017	Intel-HabanaLabs	HLS-Gaudi2-TF	Intel(R) Xeon(R) Platinum 8380	2	Habana Gaudi2
3.0-2018	Lenovo	Lenovo ThinkSystem SR670 V2 Server with 4x 40GB SXM4 A100	Intel(R) Xeon(R) Platinum 8360Y CPU @ 2.40GHz	2	NVIDIA A100-SXM
3.0-2019	Lenovo	Lenovo ThinkSystem SR670 V2 Server with 8x 80GB PCIe A100	Intel(R) Xeon(R) Platinum 8360Y CPU @ 2.40GHz	2	NVIDIA A100-PCIe
3.0-2020	Lenovo	Lenovo ThinkSystem SR670 V2 Server with 8x 80GB PCIe H100	Intel(R) Xeon(R) Platinum 8360Y CPU @ 2.40GHz	2	NVIDIA H100-PCIe
3.0-2021	Supermicro	AS-4125GS-TNRT	AMD EPYC 9554 64-Core Processor	2	NVIDIA H100-PCIe
3.0-2022	Supermicro	AS-8125GS-TNHR	AMD EPYC 9634	2	NVIDIA H100-SXM
3.0-2023	Supermicro	SYS-421GU-TNX	Intel(R) Xeon(R) Platinum 8460H	2	NVIDIA H100-SXM
3.0-2024	Supermicro	SYS-421GU-TNXR	Intel(R) Xeon(R) Platinum 8480+	2	NVIDIA H100-SXM
3.0-2025	Supermicro	SYS-820GH-TNR2	Intel(R) Xeon(R) Platinum 8380	2	Habana Gaudi2
3.0-2026	Supermicro	SYS-821GE-TNHR	Intel(R) Xeon(R) Platinum 8490H	2	NVIDIA H100-SXM
3.0-2027	Dell	16xXE8545x4A100-SXM-40GB	AMD EPYC 7713 64-Core Processor	32	NVIDIA A100-SXM
0-2028	Dell	2xR750xax4A100-PCIE-80GB	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	4	NVIDIA A100-PC
029	Dell	2xR750xax4A100-PCIE-80GB-1npa	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	4	NVIDIA A100

Cost planning

Example requests

- 1400 traffic cameras, real-time object detection
- Training language models for Named Entity Recognition, cca. with 100k A4 pages

MLCommons.org (https://mlcommons.org/en/) founded by AI stakeholders. Standardized setup for measuring training and inference speeds – instead of comparing FLOPs and OPs.

- Training Working Group
- Inference Working Group

References

References

- Consumer grade GPUs: https://www.nvidia.com/en-eu/geforce/graphics-cards/
- Semi-professional GPUs: https://www.nvidia.com/en-us/design-visualization/desktop-graphics/
- Server grade GPUs: https://www.nvidia.com/en-us/data-center/h100/
- GPU server: https://www.nvidia.com/en-us/data-center/dgx-h100/
- DGX H100 SuperPod: https://docs.nvidia.com/nvidia-dgx-superpod-data-center-design-dgx-h100.pdf
- KIFÜ Komondor: https://hpc.kifu.hu/hu/komondor
- ML Commons benchmarks: https://mlcommons.org/en/

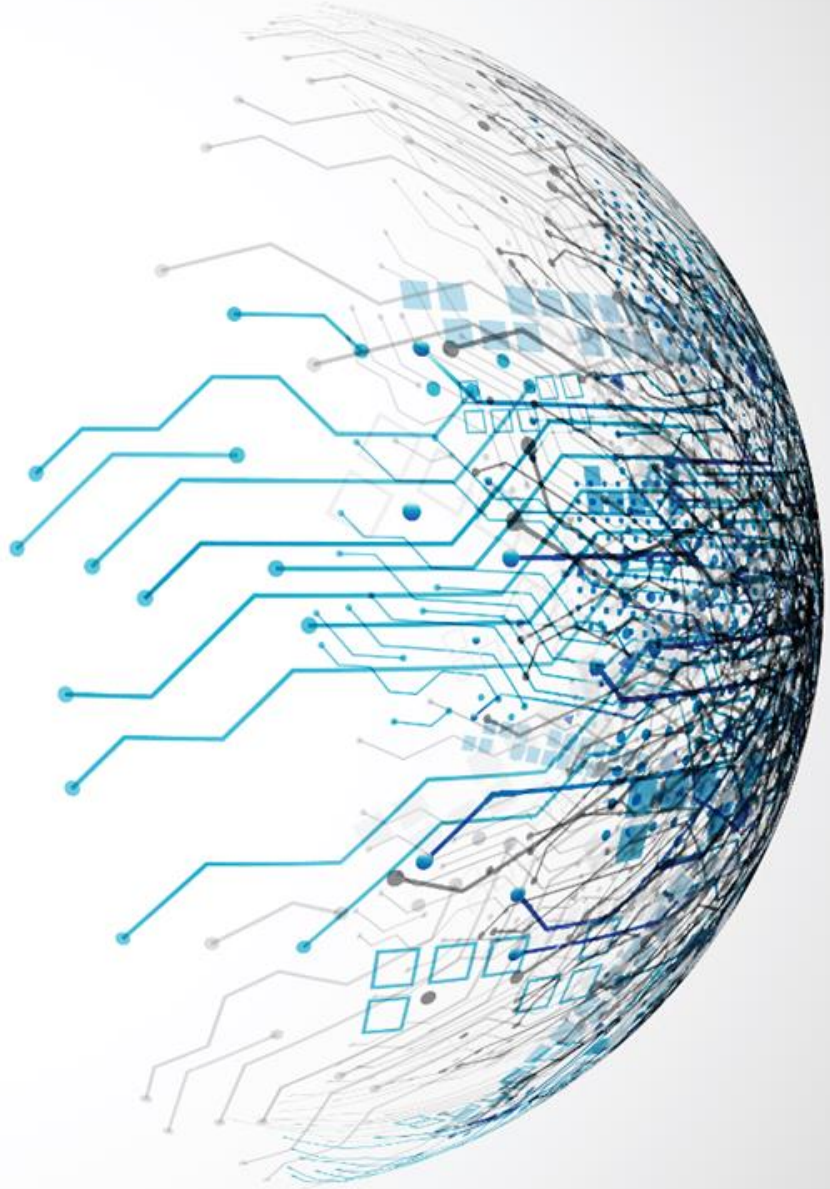
Please, don’t forget to send feedback:

https://bit.ly/bme-dl

Thank you for your attention

Dr. Mohammed Salah Al-Radhi
malradhi@tmit.bme.hu

(slides by: Dr. Bálint Gyires-Tóth)



10 September 2024