

# Deep Learning

# **HARDWARE ARCHITECTURE**

Dr. Mohammed Salah Al-Radhi  
(slides by: Dr. Bálint Gyires-Tóth)



# Copyright

Copyright © **Bálint Gyires-Tóth & Mohammed Salah Al-Radhi**, All Rights Reserved.

This presentation and its contents are protected by copyright law. The intellectual property contained herein, including but not limited to text, images, graphics, and design elements, are the exclusive property of the copyright holder identified above. Any unauthorized use, reproduction, distribution, or modification of this presentation or its contents is strictly prohibited without prior written consent from the copyright holder.

**No Recordings or Reproductions:** Attendees, viewers, and recipients of this presentation are expressly prohibited from making any audio, video, or photographic recordings, as well as screen captures, screenshots, or any form of reproduction, of this presentation, its content, or any related materials, whether during its live presentation or subsequent access. Violation of this prohibition may result in legal action.

For permissions, inquiries, or licensing requests, please contact: **{toth.b,malradhi}@tmit.bme.hu**

Unauthorized use, distribution, or reproduction of this presentation may result in civil and criminal penalties. Thank you for respecting the intellectual property rights of the copyright holder.

# Outline

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

A network graph with nodes and edges, overlaid with a white rectangular box containing the text 'Why?'. The graph consists of numerous nodes connected by thin lines, forming a complex web. The nodes are colored in shades of teal and blue. The white box is centered horizontally and contains the word 'Why?' in a bold, black, sans-serif font.

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.

The background of the slide features a complex network diagram. It consists of numerous small, semi-transparent circular nodes connected by thin, light-colored lines. The nodes are distributed across the width of the slide, with some appearing more densely connected than others. A prominent white rectangular box is centered horizontally, containing the text 'Consumer grade GPUs'. The overall aesthetic is clean and technical, typical of a presentation slide.

# 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/>

The image features a complex network diagram with numerous nodes and connecting lines, rendered in shades of teal and light blue. The nodes are scattered across the frame, with some forming dense clusters and others being more isolated. A prominent white rectangular box is centered horizontally, containing the text "Semi-professional GPUs". The background is a light, neutral color with a subtle gradient.

# 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/>



The image features a network diagram with numerous nodes and connecting lines, rendered in shades of teal and light blue. The nodes are arranged in a somewhat horizontal line, with some branching out upwards and downwards. A semi-transparent white rectangular box is centered horizontally across the middle of the image, containing the text "Server grade GPUs" in a bold, black, sans-serif font. The background is a light, neutral color with a subtle gradient.

# 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
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
Allowable up to 30 °C for DGX H100 Systems	A1	59–89.6 °F 15–32 °C	20–80% RH	6
Allowable per ASHRAE for various other classes of data center and telecom environments	A2	50–95 °F 10–35 °C	20–80% RH	6
	A3	41–104 °F 5–40 °C	10.4 °F DP and 8–85% RH -12 °C DP and 8–85% RH	7
	A4	41–113 °F 5–45 °C	10.4 °F DP and 8–90% RH -12 °C DP and 8–90% RH	7
	B	41–95 °F 5–35 °C	8–80% RH	8
	C	41–104 °F 5–40 °C	8–80% RH	8

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

Class	Particle Size <sup>1</sup>					
	> 0.1 μm	> 0.2 μm	> 0.3 μm	> 0.5 μm	> 1 μm	> 5 μm

Table 8. Common distribution schemes compatible with DGX H100 racks

Phase	Distribution Voltage	Line Voltage	Amps	Breaker Derating	Circuit Capacity kW <sup>1</sup>	Maximum Supported DGX H100 Systems per Rack <sup>2,3</sup>	Peak Server Demand per Circuit kW <sup>2</sup>	Stranded Capacity at Peak Demand kW <sup>2</sup>
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

29
293
2,930
29,300
93,000







- 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



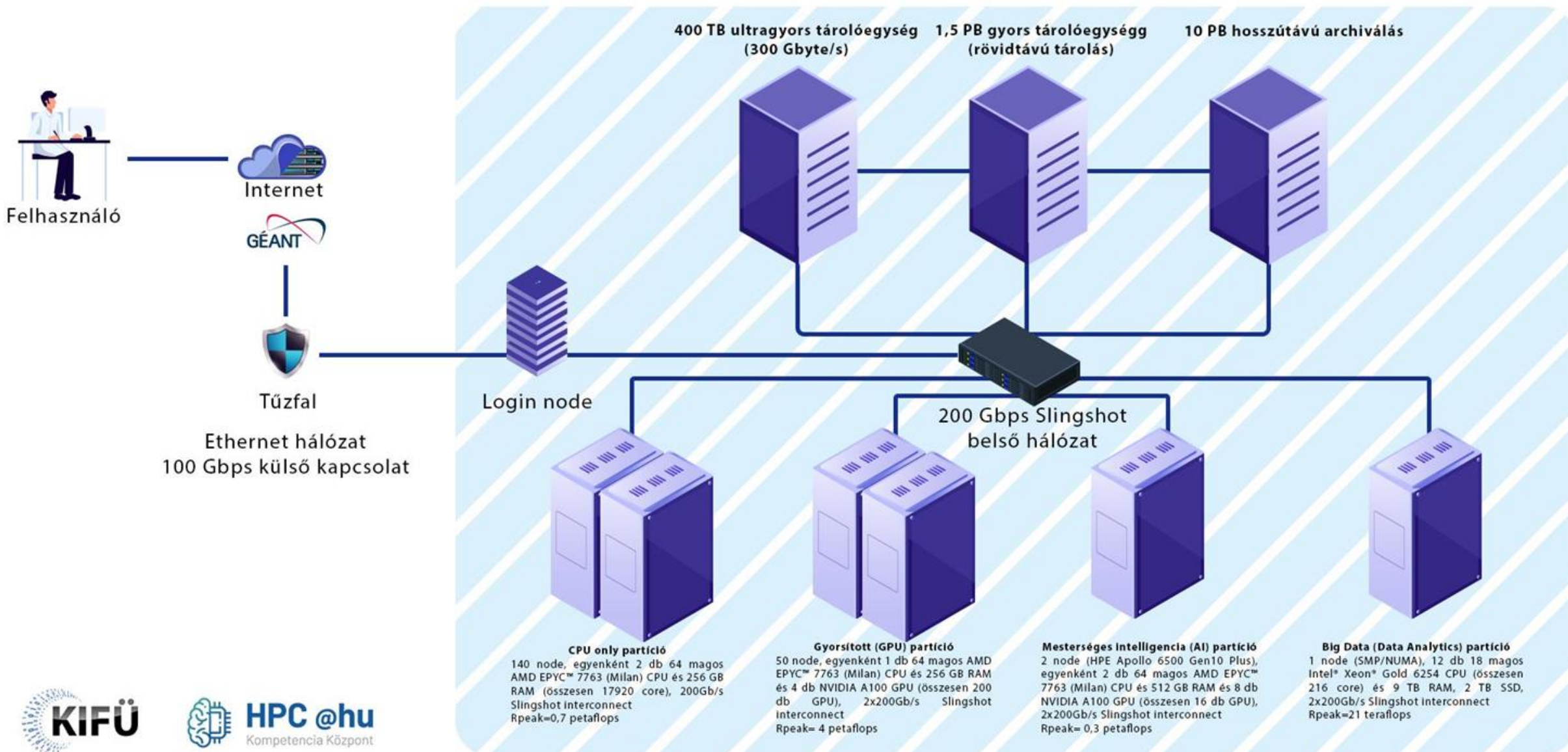
KIFÜ



HPC@hu



# A KOMONDOR SZUPERSZÁMÍTÓGÉP FELÉPÍTÉSE



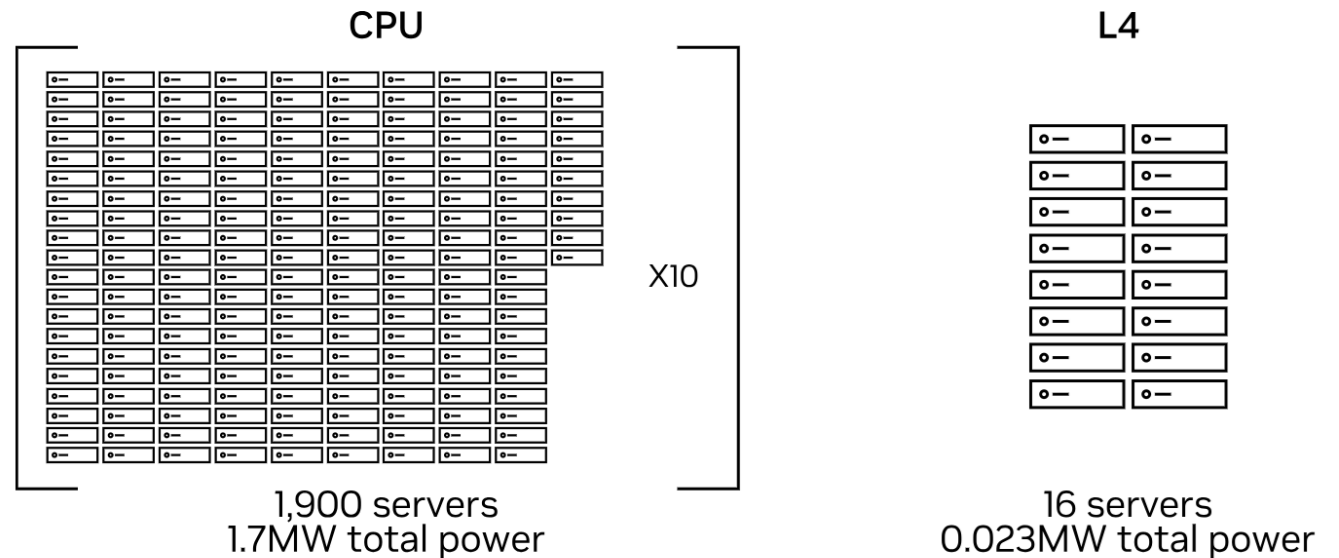


The image features a complex network diagram with numerous nodes and connecting lines, rendered in shades of teal and light blue. A prominent white rectangular box is centered horizontally, containing the text 'Inference servers' in a bold, black, sans-serif font. The background is a light, neutral color with a subtle, abstract pattern of faint lines and dots, suggesting a larger network or data landscape.

# 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)



**99%** less energy  
less rack space

The image features a network diagram with numerous nodes and connecting lines, rendered in shades of teal and light blue. A white rectangular box is superimposed over the center of the diagram, containing the text "GPU in cloud".

# 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/>



The image features a complex network diagram with numerous nodes and connecting lines, rendered in shades of teal and light blue. A prominent white rectangular box is centered horizontally across the middle of the image, containing the text "Cost planning" in a bold, black, sans-serif font. The background is a light, neutral tone with a subtle, abstract pattern of lines and dots, suggesting a digital or data-driven environment.

# Cost planning

Available Cloud	Submitter	System	Processor	#	Accelerator
3.0-2000	NVIDIA+CoreWeave	coreweave_hgxm100_n192_ngc23.04_pytorch	Intel(R) Xeon(R) Platinum 8462Y+	384	NVIDIA H100-SXM
3.0-2001	NVIDIA+CoreWeave	coreweave_hgxm100_n384_ngc23.04_pytorch	Intel(R) Xeon(R) Platinum 8462Y+	768	NVIDIA H100-SXM
3.0-2002	NVIDIA+CoreWeave	coreweave_hgxm100_n448_ngc23.04_mxnet	Intel(R) Xeon(R) Platinum 8462Y+	896	NVIDIA H100-SXM
3.0-2003	NVIDIA+CoreWeave	coreweave_hgxm100_n448_ngc23.04_pytorch	Intel(R) Xeon(R) Platinum 8462Y+	896	NVIDIA H100-SXM
3.0-2004	NVIDIA+CoreWeave	coreweave_hgxm100_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
3.0-2028	Dell	2xR750xax4A100-PCIe-80GB	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	4	NVIDIA A100-PCIe
3.0-2029	Dell	2xR750xax4A100-PCIe-80GB-100g	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	4	NVIDIA A100
3.0-2030	Dell	5xR120xax4A100-PCIe-80GB	Intel(R) Xeon(R) Gold 6338 CPU @ 2.00GHz	4	NVIDIA A100-PCIe
3.0-2031	Dell	16xXE8545x4A100-SXM-40GB	AMD EPYC 7713 64-Core Processor	32	NVIDIA A100-SXM
3.0-2032	Supermicro	SYS-8125GE-TNHR	Intel(R) Xeon(R) Platinum 8480H	2	NVIDIA H100-SXM
3.0-2033	Supermicro	SYS-820GH-TNHR	Intel(R) Xeon(R) Platinum 8380	2	Habana Gaudi2
3.0-2034	Supermicro	SYS-421GU-TNXR	Intel(R) Xeon(R) Platinum 8480+	2	NVIDIA H100-SXM
3.0-2035	Supermicro	SYS-8125GE-TNHR	Intel(R) Xeon(R) Platinum 8480H	2	NVIDIA H100-SXM
3.0-2036	Supermicro	SYS-8125GE-TNHR	AMD EPYC 8634	2	NVIDIA H100-SXM
3.0-2037	Supermicro	SYS-4125GE-TNHR	AMD EPYC 8634 64-Core Processor	2	NVIDIA H100-SXM

# 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



A network diagram consisting of numerous nodes (small circles) connected by thin lines (edges). The nodes are arranged in a roughly horizontal line, with some branching out above and below. The lines are a light teal color. A white rectangular box is superimposed over the center of the diagram, containing the word "References" in a black, sans-serif font.

# 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](mailto:malradhi@tmit.bme.hu)

(slides by: Dr. Bálint Gyires-Tóth)

10 September 2024

