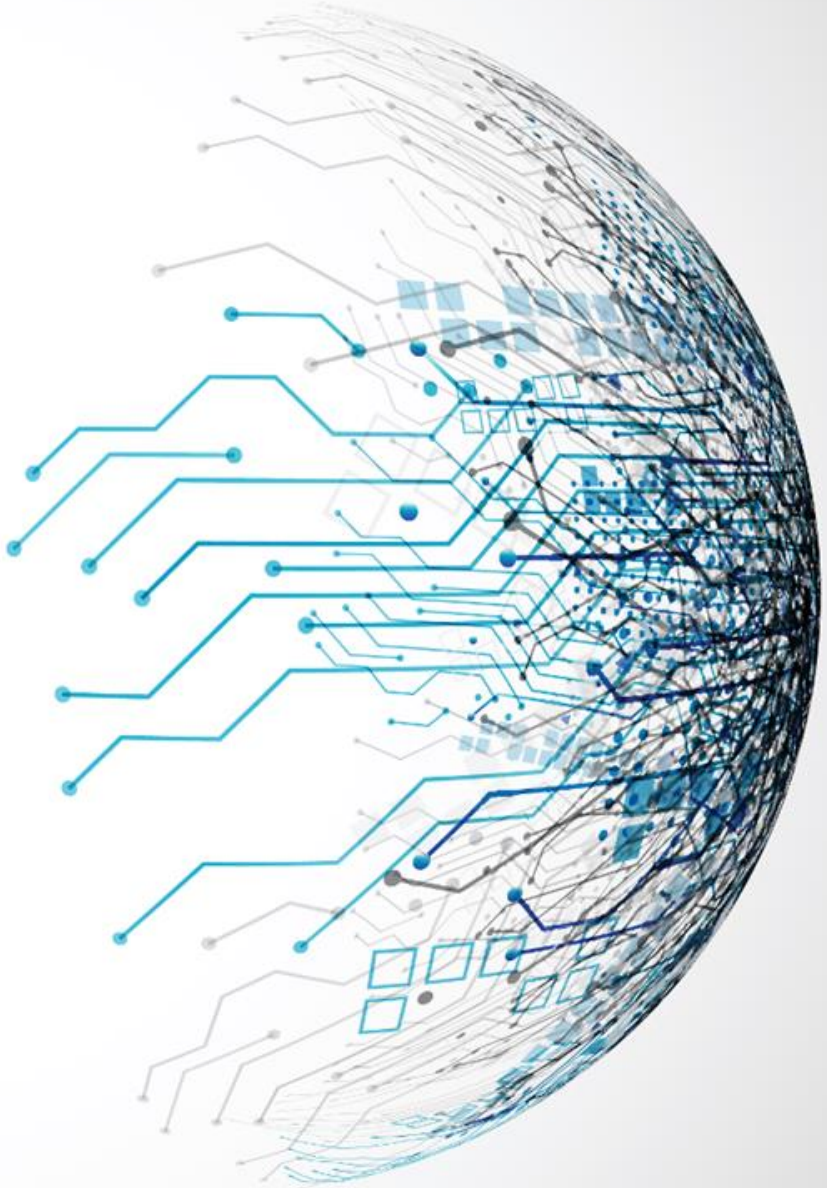
MI architektúrák skálázása: hardver és szoftver megoldások modern GPU infrastruktúrákban

Gyires-Tóth Bálint



Bálint Gyires-Tóth, PhD

Egyetemi docens

Budapesti Műszaki és Gazdaságtudományi Egyetem

NVIDIA Deep Learning Institute

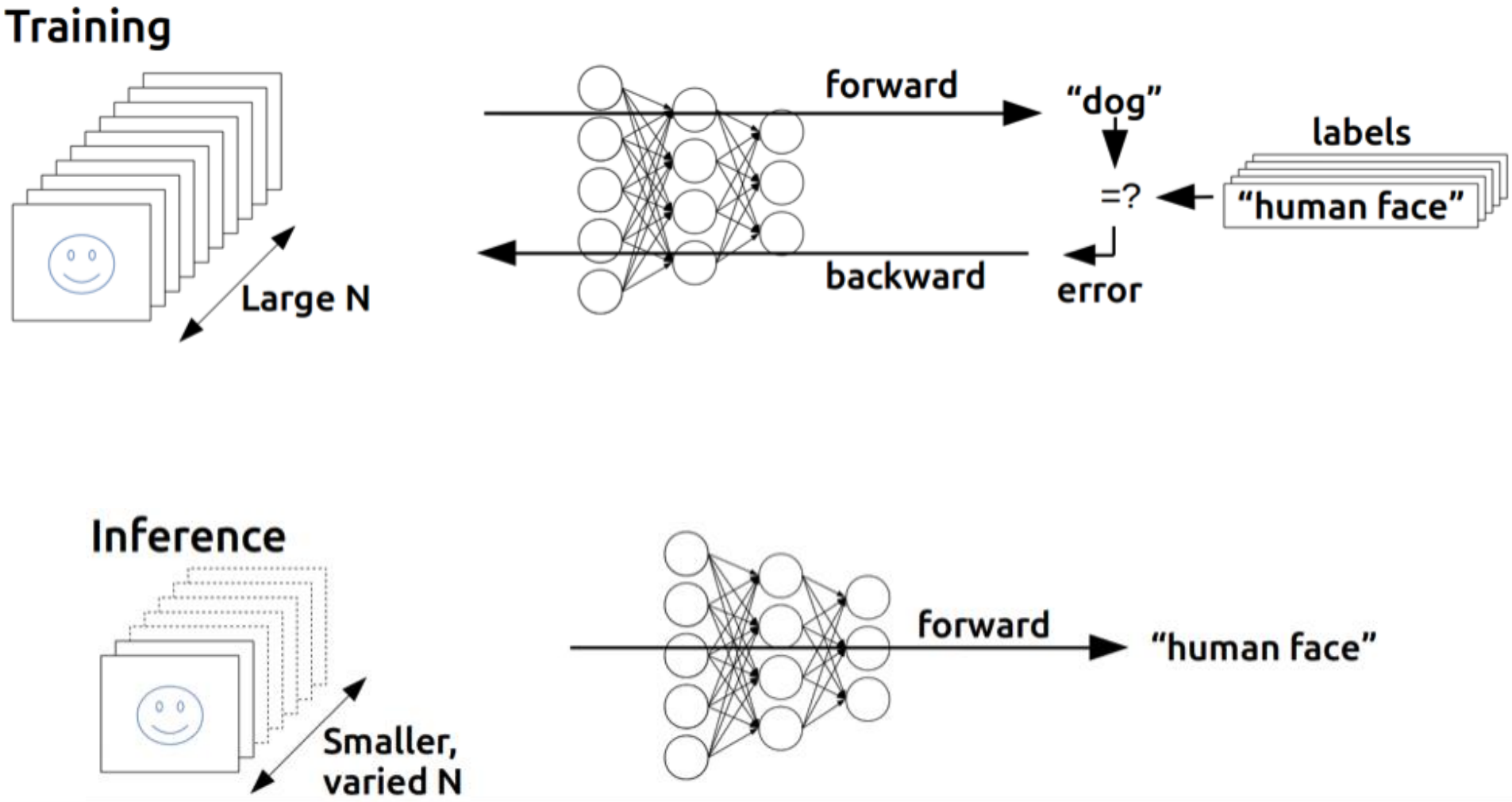
Oktató

Egyetemi nagykövet

Mélytanulás kutatás, fejlesztés, oktatás

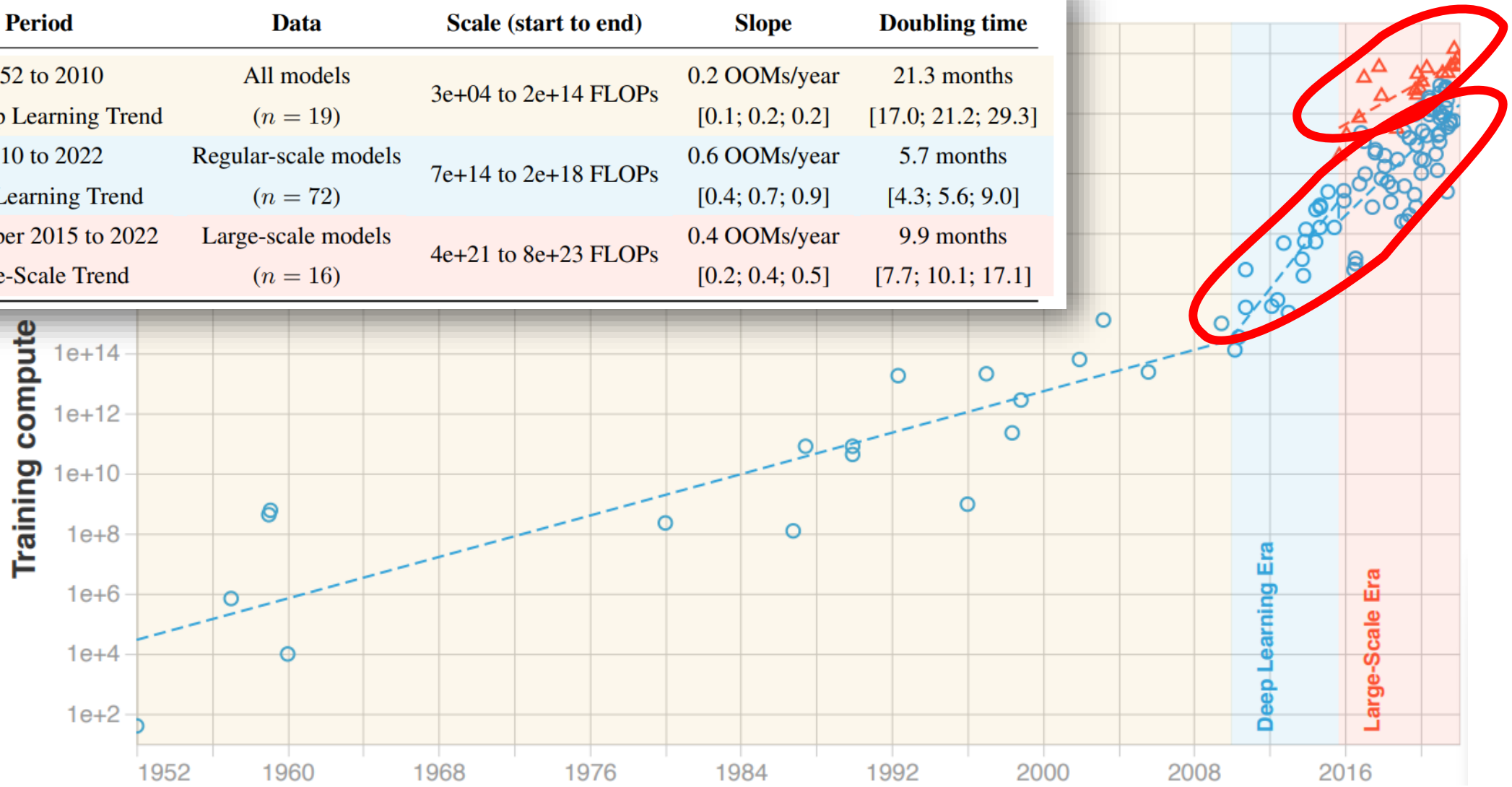
MI skálázás algoritmikus háttere

Gyires-Tóth Bálint (toth.b@tmit.bme.hu)



Forrás: developer.nvidia.com

Period	Data	Scale (start to end)	Slope	Doubling time
1952 to 2010 Pre Deep Learning Trend	All models ($n = 19$)	3e+04 to 2e+14 FLOPs	0.2 OOMs/year [0.1; 0.2; 0.2]	21.3 months [17.0; 21.2; 29.3]
2010 to 2022 Deep Learning Trend	Regular-scale models ($n = 72$)	7e+14 to 2e+18 FLOPs	0.6 OOMs/year [0.4; 0.7; 0.9]	5.7 months [4.3; 5.6; 9.0]
September 2015 to 2022 Large-Scale Trend	Large-scale models ($n = 16$)	4e+21 to 8e+23 FLOPs	0.4 OOMs/year [0.2; 0.4; 0.5]	9.9 months [7.7; 10.1; 17.1]



Forrás: Sevilla, Jaime, et al. "Compute trends across three eras of machine learning." *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022.

Kis modellek tanítása

Kísérletezés

500+ GPU óra tanítás (20+ nap)

TRL 3

1000…3000+ GPU óra tanítás (40…120+ nap)

TRL 4

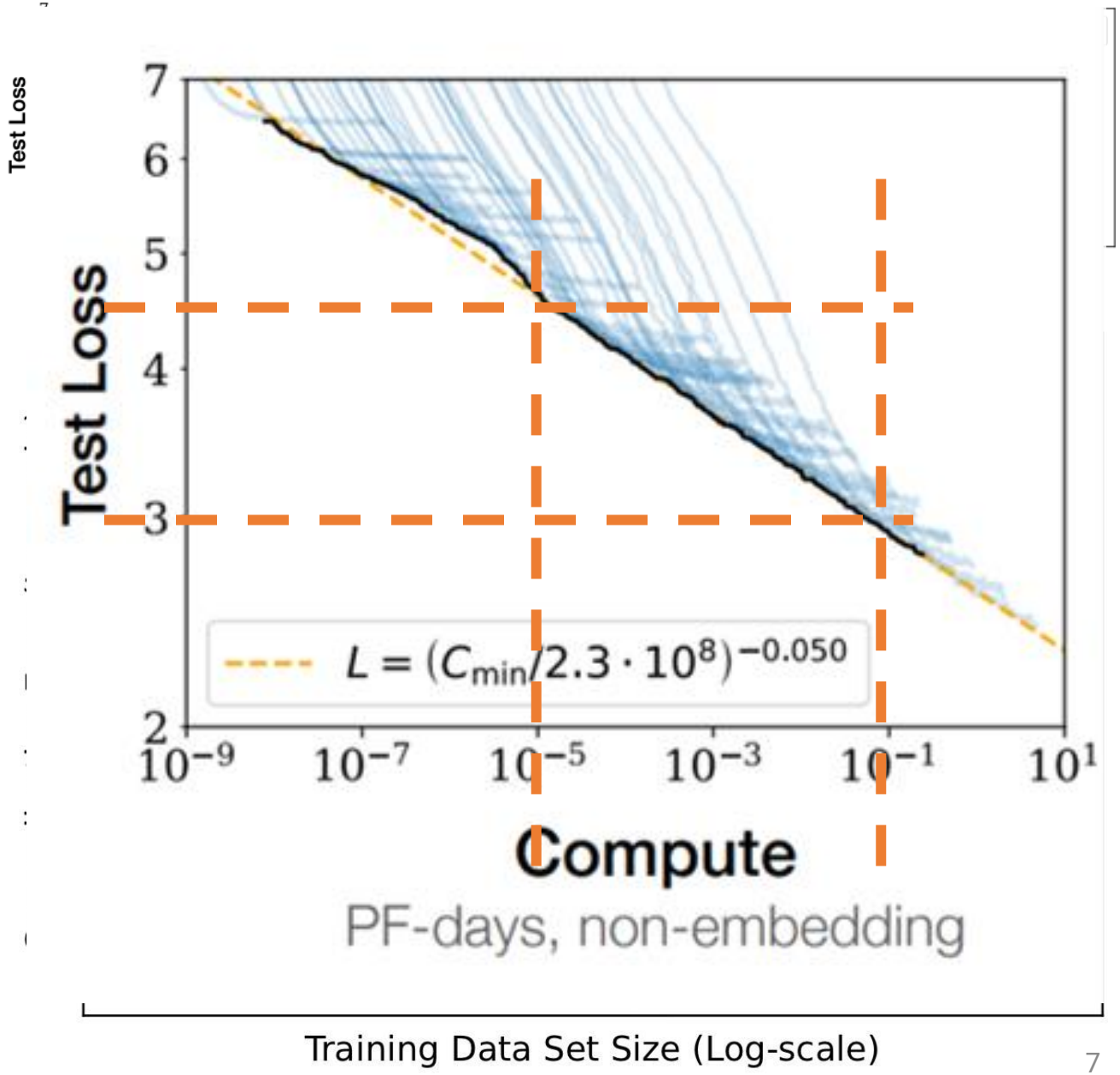
5000…10000+ GPU óra tanítás (~1+ év tanítás)

Nagy modellek tanítása

A deep learning erőtörvénye

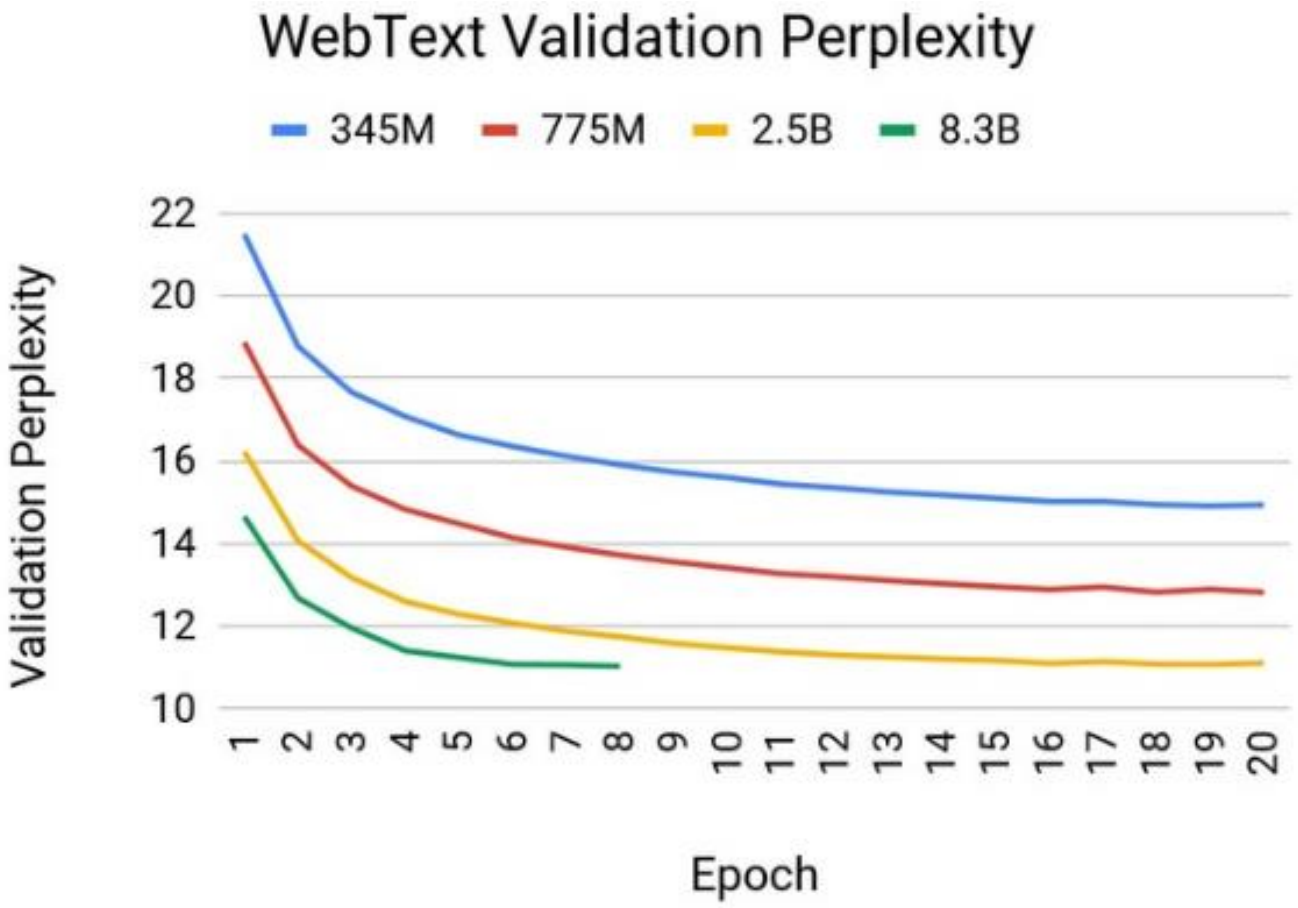
*„These power-law learning curves **exists across all tested domains, model architectures, optimizers, and loss functions**. Further, within each domain, **model architecture and optimizer changes only shift the learning curves** but **do not affect the power-law exponent**—the "steepness" of the learning curve.”*

Forrás: Hestness, Joel, et al. "Deep learning scaling is predictable, empirically." arXiv preprint arXiv:1712.00409 (2017).



Nagy modellek előnyei

- Nagyobb pontosság
- Adathatékonyság + few-shot learning
- Erőforrás-hatékonyság
- Hiperparaméterek jelentősége drasztikusan csökken



Neurális hálózatok tanításának skálázása multi-GPU és multi-node rendszerekre

Adat és modell klónozás

Különböző (hiper)paraméterekkel futtatott modellezési pipeline.

Párhuzamos adatfeldolgozás (data parallelism)

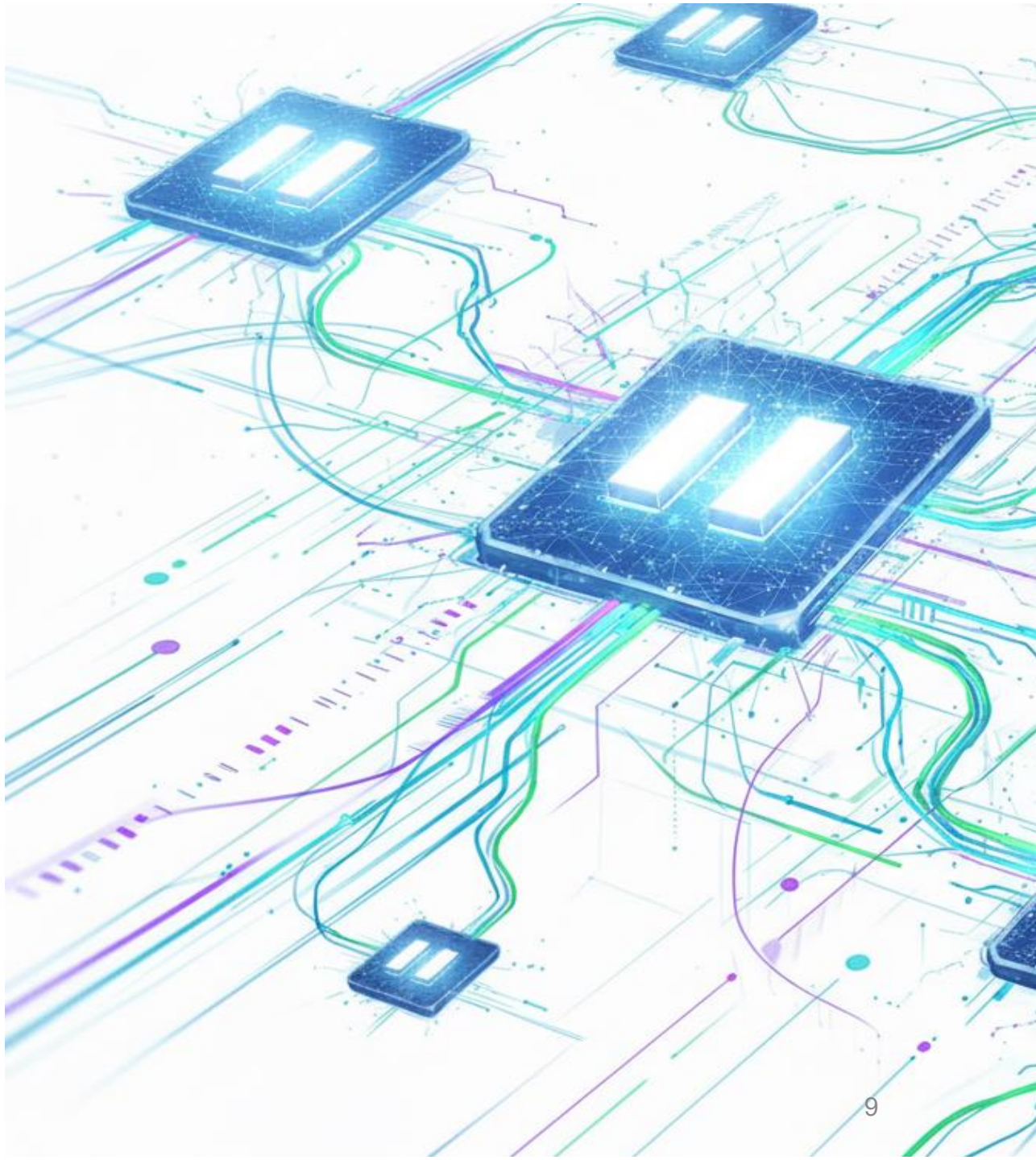
Minden GPU-n külön modell és adatrész.

Párhuzamos pipeline (pipeline paralellism)

A neurális hálózat különböző részei az egyes GPU-kon.

Párhuzamos tensor műveletek (tensor parallelism)

Rétegen belüli szétbontása a neurális modellnek.



MŰEGYETEM 1782

TMiT

Hardver és szoftver architektúra

Gyires-Tóth Bálint (toth.b@tmit.bme.hu)

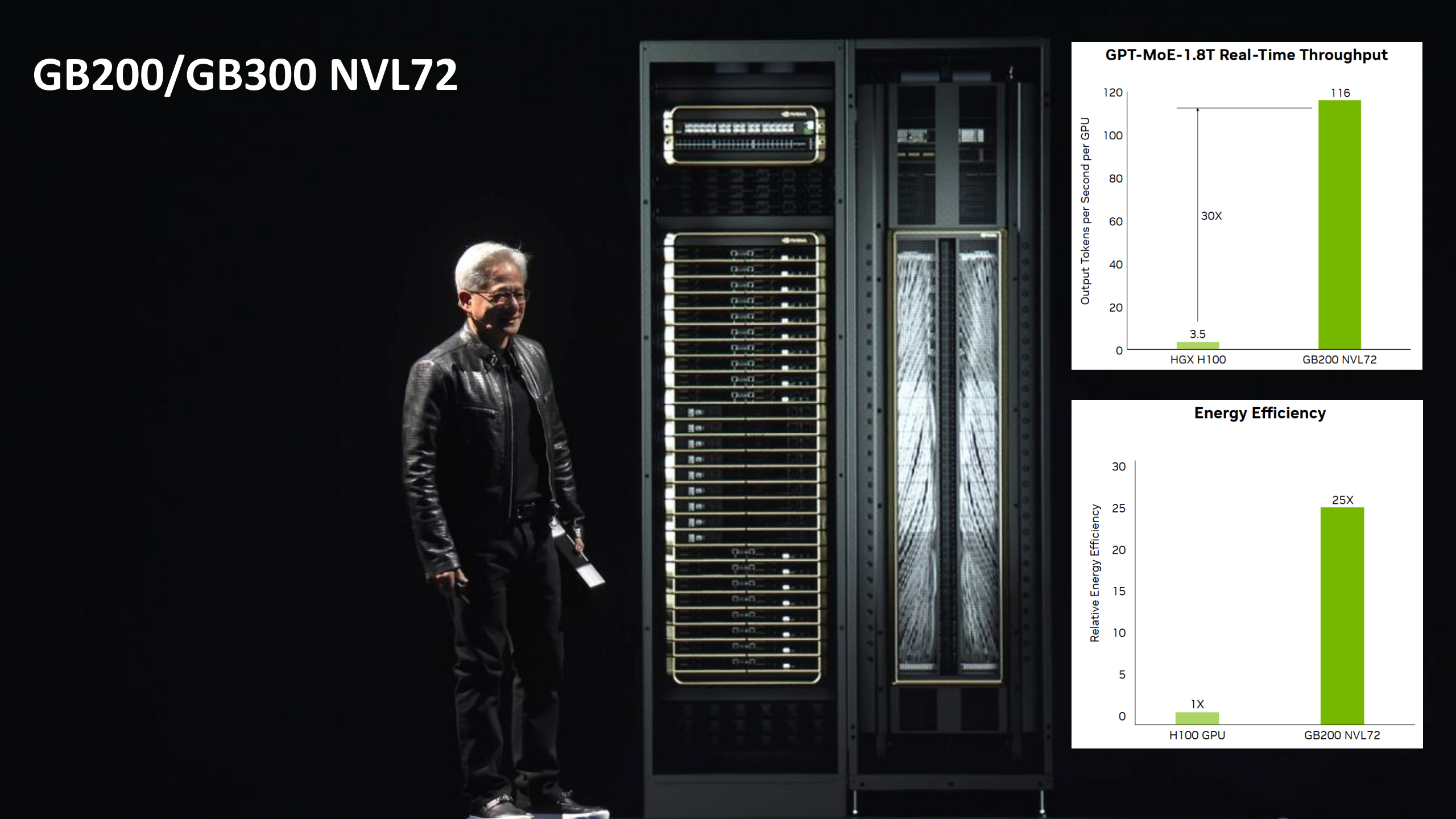
DGX
Spark AI

Gyires-Tóth Bálint (toth.b@tmit.bme.hu)

WORKSTATION
RTX alapon

DGX GB300

GB200/GB300 NVL72



ML Commons

Division/Power	Round	Organization	System Name	Processor	Accelerator	Accelerators Per Node	Software	Availability	MLC Model
Closed	v5.0	(All)	(All)	(All)	(All)	(All)	(All)	(All)	(All)

MLPerf Results

MLCommons data as of: 6/4/2025 1:38:36 PM

Benchmark: **Training**
Division/Power: **Closed**
Availability: **All**
Round: **v5.0**

										Benchmark / Model MLC / Units						
										Training						
Public ID	Availability	Organization	System Name (Click + for de..)	Total Acc..	Accelerator Model Name	Acceler..	Host Processor Model..	Host Pr..	Software	BERT Latency (In ..	DLRM_dcnv2 Latency (In ..	llama2_70b.. Latency (In ..	llama31_40.. Latency (In ..	RetinaNet Latency (In ..	RGAT Latency (In ..	stable_diffus.. Latency (In ..
	on-premise				(B200-SXM-180GB)		Platinum 8570		25.03							
5.0-0061	Available on-premise	NVIDIA	Nyx (1x NVIDIA DGX B200)	8	NVIDIA Blackwell GPU (B200-SXM-180GB)	8	Intel(R) Xeon(R) Platinum 8570	2	NVIDIA Merlin HugeCTR Release 25.03		2.333					
5.0-0062	Available on-premise	NVIDIA	Nyx (1x NVIDIA DGX B200)	8	NVIDIA Blackwell GPU (B200-SXM-180GB)	8	Intel(R) Xeon(R) Platinum 8570	2	NVIDIA NeMo Framework Release 25.04			11.275				14.076
5.0-0063	Available on-premise	NVIDIA	Nyx (1x NVIDIA DGX B200)	8	NVIDIA Blackwell GPU (B200-SXM-180GB)	8	Intel(R) Xeon(R) Platinum 8570	2	NVIDIA PyTorch Release 25.04	3.502				22.065		
5.0-0064	Available on-premise	NVIDIA	Tyche (2x NVIDIA GB200 NVL72)	144	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA NeMo Framework Release 25.04			1.128				
5.0-0065	Available on-premise	NVIDIA	Tyche (4x NVIDIA GB200 NVL72)	256	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA DGL Release 25.03						0.837	
5.0-0066	Available on-premise	NVIDIA	Tyche (4x NVIDIA GB200 NVL72)	256	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA NeMo Framework Release 25.04				240.335			
5.0-0067	Available on-premise	NVIDIA	Tyche (8x NVIDIA GB200 NVL72)	512	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA NeMo Framework Release 25.04			0.560	121.757			1.039
5.0-0068	Available on-premise	NVIDIA	Tyche (8x NVIDIA GB200 NVL72)	512	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA PyTorch Release 25.04	0.305				1.388		
5.0-0069	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	64	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA Merlin HugeCTR Release 25.03		0.773					
5.0-0070	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	64	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA PyTorch Release 25.04	0.656						
5.0-0071	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	72	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA DGL Release 25.03						1.114	
5.0-0072	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	72	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA NeMo Framework Release 25.04			1.676				2.471
5.0-0073	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	72	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA PyTorch Release 25.04					3.770		
5.0-0074	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	8	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA DGL Release 25.03						4.975	
5.0-0075	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	8	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA Merlin HugeCTR Release 25.03		2.241					
5.0-0076	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	8	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA NeMo Framework Release 25.04			11.145				12.856
5.0-0077	Available on-premise	NVIDIA	Tyche (1x NVIDIA GB200 NVL72)	8	NVIDIA Blackwell GPU (GB200)	4	Neoverse-V2	2	NVIDIA PyTorch Release 25.04	3.424				22.334		
5.0-0078	Available on-premise	QCT	D74H-7U	8	NVIDIA H200-SXM5-141GB	8	Intel(R) Xeon(R) Platinum 8480+	2	NVIDIA PyTorch/MxNet/DGL	5.206	4.005	24.043		34.241	8.128	29.922
5.0-0079	Available on-premise	QCT	D75T-7U	8	AMD Instinct MI325X 256GB HBM3e	8	AMD EPYC 9655	2	PyTorch			22.430				
5.0-0080	Available on-premise	SCITIX	Scitix_n8	64	NVIDIA H100-SXM5-80GB	8	Intel(R) Xeon(R) Platinum 8468	2	NVIDIA NeMo Framework Release 24.09			4.556				
5.0-0081	Available on-premise	Supermicro	AS-4126GS-NMR-LCC	8	AMD Instinct MI325X 256GB HBM3E DLC	8	AMD EPYC 9575F	2	ROCm 6.4.0-47			21.575				

ML Commons

Division/Power	Round	Organization	System Name	Accelerator	Processor	Software	MLC Model	# of Accelerators	Availability	Scenario
Closed	v5.1	(All)	(All)	(All)	(All)	(All)	(All)	(All)	(All)	(All)

MLPerf Results

MLCommons data as of: 9/2/2025 7:20:18 PM

Benchmark: **Inference**
System Type: **Datacenter**
Division/Power: **Closed**
Availability: **All**
Round: **v5.1**

									Benchmark / Model MLC / Scenario / Units							
									Inference							
									llama3.1-8b-datacenter			llama3.1-405b			mixtral-8x7b	
									Offline	Server	Interactive	Offline	Server	Interactive	Offline	Ser
Public ID	Organization	Availability	System Name (click + for details)	# of Nodes	# of Processors	Accelerator	# of Accelerators	Code	Samples/s	Queries/s	Queries/s	Tokens/s	Tokens/s	Tokens/s	Tokens/s	Toke
5.1-0065	MangoBoost	available	MangoBoost Heterogeneous Cluster (4x8x..	6	2	AMD Instinct MI300X 19..	8	https://github.com/mlcomm..								
5.1-0066	MangoBoost	available	Supermicro AS-8125GS-TNMR2 (8x MI30..	1	2	AMD Instinct MI300X 19..	8	https://github.com/mlcomm..								
5.1-0067	MangoBoost	available	MangoBoost (8x MI325X-256GB LLMBoost)	1	2	AMD Instinct MI325X 25..	8	https://github.com/mlcomm..								
5.1-0068	MiTAC	available	G8825Z5	1	2	AMD Instinct MI325X 25..	8	https://github.com/mlcomm..							67,964.40	57,
5.1-0069	NVIDIA	available	NVIDIA DGX B200 (8x B200-SXM-180GB ..	1	2	NVIDIA B200-SXM-180..	8	https://github.com/mlcomm..	136,752.00	128,541.00	102,903.00	1,624.11	1,240.42	751.37		
5.1-0070	NVIDIA	available	(4x GB200-186GB_aarch64 TensorRT)	1	2	NVIDIA GB200	4	https://github.com/mlcomm..								
5.1-0071	NVIDIA	available	(72x GB200-186GB_aarch64 TensorRT)	18	2	NVIDIA GB200	4	https://github.com/mlcomm..				14,774.30	11,614.30	9,921.02		
5.1-0072	NVIDIA	available	(72x GB300-288GB_aarch64 TensorRT)	18	2	NVIDIA GB300	4	https://github.com/mlcomm..				16,104.10	12,248.40			
5.1-0073	Nebius	available	Nebius B200 (8x B200-SXM-180GB Tenso..	1	2	NVIDIA B200-SXM-180..	8	https://github.com/mlcomm..				1,659.50	1,279.53	770.67		
5.1-0074	Nebius	available	Nebius GB200 (4x GB200-186GB_aarch6..	1	2	NVIDIA GB200	4	https://github.com/mlcomm..				855.82	596.11			
5.1-0075	Nebius	available	Nebius H200 (8x H200-SXM-141GB Tens..	1	2	NVIDIA H200-SXM-141..	8	https://github.com/mlcomm..				552.75	295.36	202.78		
5.1-0076	Oracle	available	BM.GPU.B200.8	1	2	NVIDIA B200-SXM-180..	8	https://github.com/mlcomm..	145,789.00	128,649.00		1,615.85	1,243.67			
5.1-0077	Oracle	available	BM.GPU.GB200.4	1	2	NVIDIA GB200	4	https://github.com/mlcomm..								
5.1-0078	Quanta_Cloud_..	available	1-node-2S-GNR_86C	1	2	N/A	0	https://github.com/mlcomm..								
5.1-0079	Quanta_Cloud_..	available	QuantaGrid D74H-7U (8x H200-SXM-141..	1	2	NVIDIA H200-SXM-141..	8	https://github.com/mlcomm..	56,917.30	57,970.40	54,117.60	550.11	295.52			
5.1-0080	Quanta_Cloud_..	available	D75E-4U_H200-NVL-141GBx4	1	2	NVIDIA H200-NVL-141..	4	https://github.com/mlcomm..	25,180.80	25,753.80	21,944.10					
5.1-0081	Quanta_Cloud_..	available	D75T-7U_8xMI325X	1	2	AMD Instinct MI325X 25..	8	https://github.com/mlcomm..							68,218.70	59,
5.1-0082	RedHat	available	IBM Cloud gx3 instance 1xL40S-PCIE-48..	1	2	NVIDIA L40S	1	https://github.com/mlcomm..	1,642.22	1,207.14						
5.1-0083	RedHat	available	Dell PowerEdge XE8640 (1xH100-SXM-80..	1	2	NVIDIA H100-SXM-80GB	1	https://github.com/mlcomm..	5,777.08	5,103.99						
5.1-0084	Supermicro	available	1-node-2S-GNR_128C	1	2	N/A	0	https://github.com/mlcomm..	1,195.93	450.87						
5.1-0085	Supermicro	available	AS-8126GS-TNMR (8x MI325X)	1	2	AMD Instinct MI325X 25..	8	https://github.com/mlcomm..								
5.1-0086	Supermicro	available	AS-4126GS-NBR-LCC (8x B200-SXM-180..	1	2	NVIDIA B200-SXM-180..	8	https://github.com/mlcomm..	142,955.00	128,701.00	122,269.00					
5.1-0087	Supermicro	available	SYS-422GS-NBRT-LCC (8x B200-SXM-1..	1	2	NVIDIA B200-SXM-180..	8	https://github.com/mlcomm..	141,056.00	122,331.00		1,650.51	1,244.69	750.94		
5.1-0088	Supermicro	available	SYS-A21GE-NBRT (8x B200-SXM-180GB..	1	2	NVIDIA B200-SXM-180..	8	https://github.com/mlcomm..				1,621.72	1,242.84	750.41		
5.1-0089	Supermicro	available	SYS-422GA-NBRT-LCC (8x B200-SXM-1..	1	2	NVIDIA B200-SXM-180..	8	https://github.com/mlcomm..				1,641.34	1,240.44	751.39		
5.1-0090	Supermicro_M..	available	Supermicro AS-8126GS-TNMR (2x8x MI3..	2	2	AMD Instinct MI325X 25..	8	https://github.com/mlcomm..								
5.1-0091	Supermicro_M..	available	Supermicro AS-8126GS-TNMR (2x8x MI3..	3	2	AMD Instinct MI325X 25..	8	https://github.com/mlcomm..								
5.1-0092	TheStage AI	available	Nebius H100 (8x H100-SXM-80GB TheSta..	1	2	NVIDIA H100-SXM-80GB	8	https://github.com/mlcomm..								
5.1-0093	University-of-Fl..	available	NVIDIA DGX B200 (1x B200-SXM-180GB ..	1	2	NVIDIA B200-SXM-180..	1	https://github.com/mlcomm..								
5.1-0094	University-of-Fl..	available	NVIDIA DGX B200 (1x B200-SXM-180GB ..	1	2	NVIDIA B200-SXM-180..	8	https://github.com/mlcomm..								
5.1-0095	Vultr	available	Supermicro AS -8126GS-TNMR (8x MI325..	1	2	AMD Instinct MI325X 25..	8	https://github.com/mlcomm..								
5.1-0096	Vultr	available	Supermicro AS -8126GS-TNMR (8x MI325..	1	2	AMD Instinct MI325X 25..	8	https://github.com/mlcomm..							70,035.60	56,
5.1-0097	CoreWeave	preview	(8x GB300-288GB_aarch64 TensorRT)	2	2	NVIDIA GB300	4	https://github.com/mlcomm..								

View on Tableau Public

Tervezz egy kirándulást Mátraháza körül, körülbelül 2 órás legyen.

Szia! Mátraháza körül van egy szép, könnyen elérhető kirándulóösvény, amely kb. 2 órát vesz igénybe. A túra a Mátraházai-víztározó partján kezdődik, és a hegyi erdőkön és vízmosásokon vezet át.

Az út első része a víztározó partján tart, ahol szép kilátás nyílik a tározóra és a környező hegyekre. Később az út átkel a vízmosásokon, ahol szűk földutakon és kőlépcsőkön kell átjutnunk.

A túra második része a hegyi erdőben vezet, ahol számos fajta fa és cserje található. Ebben a részben lehetőség van pihenésre és a környező kilátásokban gyönyörködni.

Ha sikerül időben visszaérnünk Mátraházára, akkor érdemes megállni a Vörös-szék étteremben, ahol finom ételeket és italokat kínálnak.

Llama 3.1. 8b

Mátraháza körül számos kirándulási lehetőség található. Egy 2 órás kirándulásra ajánlom a következő útvonalat:

- Indulás Mátraháza központjából

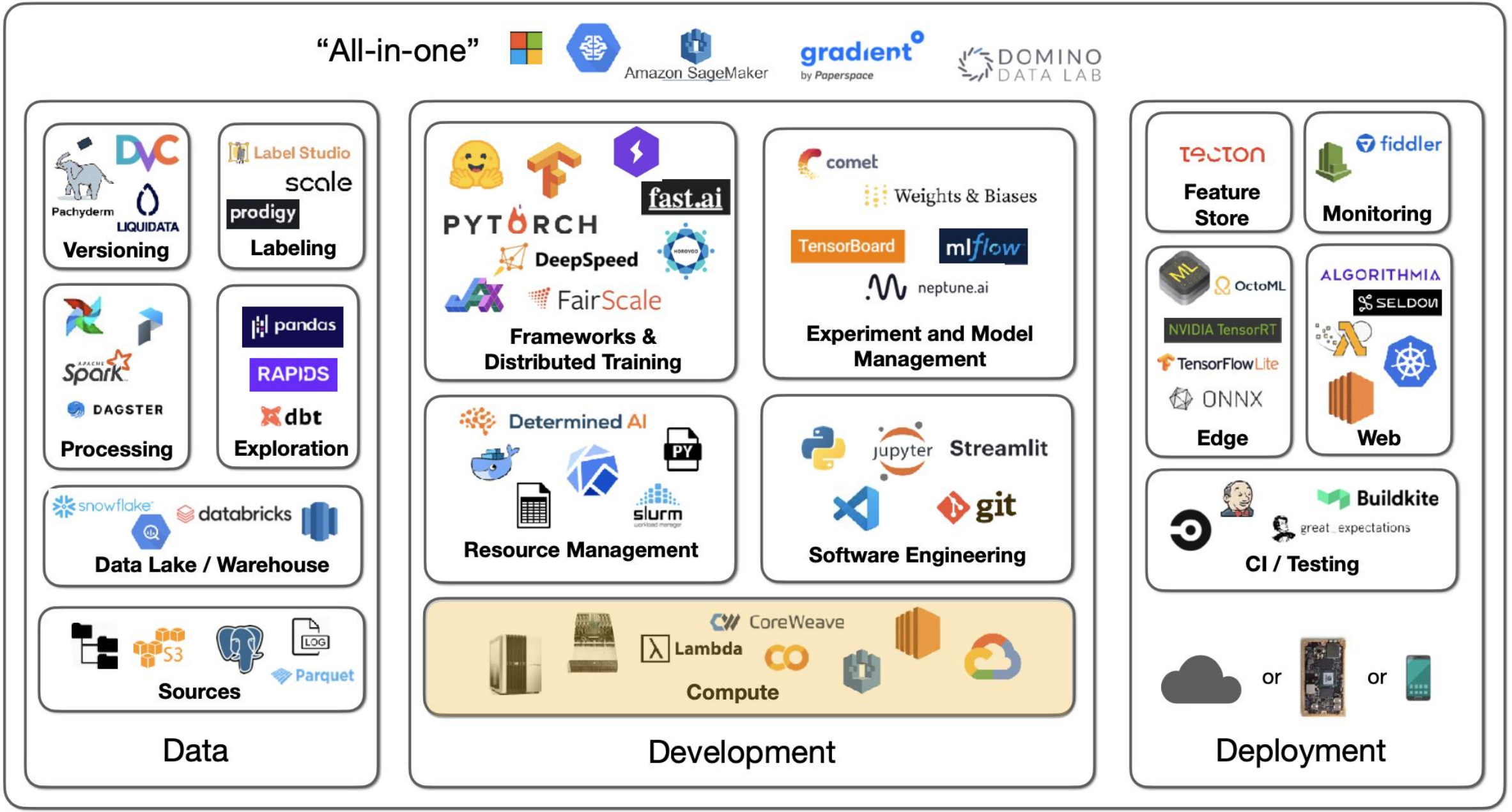
- Először menjünk a Mátraháza és Mátraszentimre közötti erdészeti úton (kb. 20 perc)

- Az úton haladva érjünk el a Kékes-tetőhöz (kb. 45 perc), amely Magyarország legmagasabb pontja

- A Kékes-tető megmászása után térjünk vissza Mátraházára a Szent Imre-úton keresztül (kb. 35 perc)

Ez a túra körülbelül 2 órát vesz igénybe, és végig szép erdőkben sétálhatunk, élvezhetjük a természetet és a panorámát. Ne felejtsd el a túracipőt, és ha szükséges, ételt és italt is vigyél magaddal!

Llama 3.1. 405b

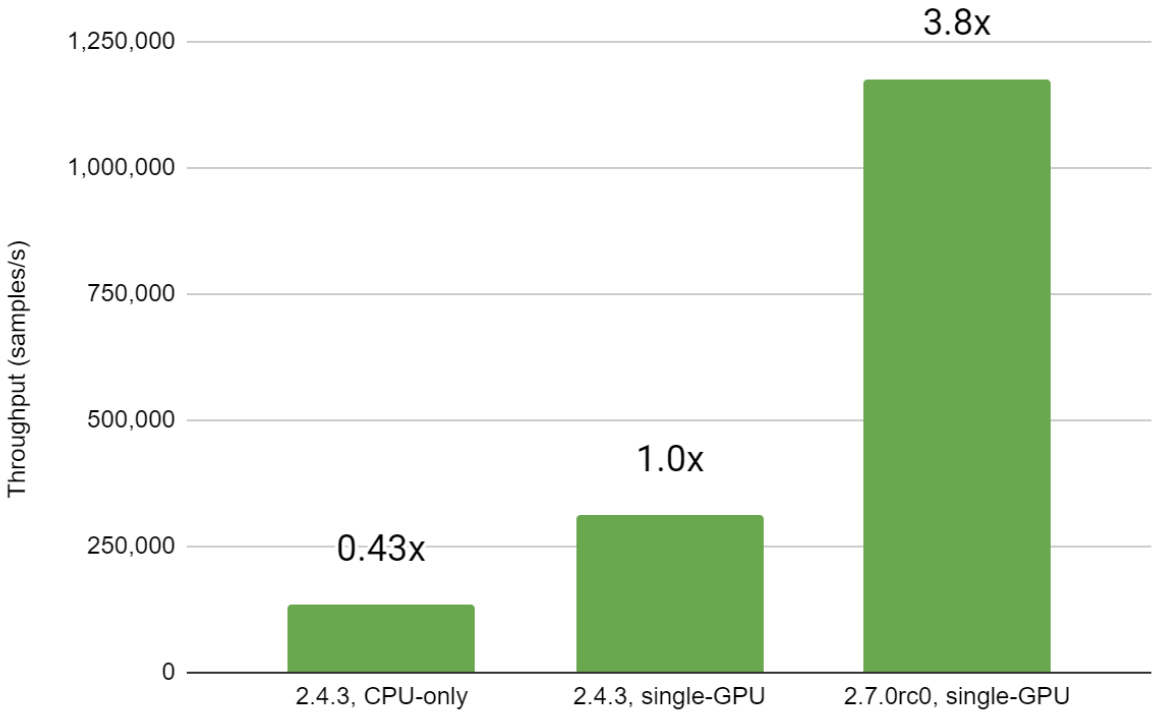


MŰEGYETEM 1782

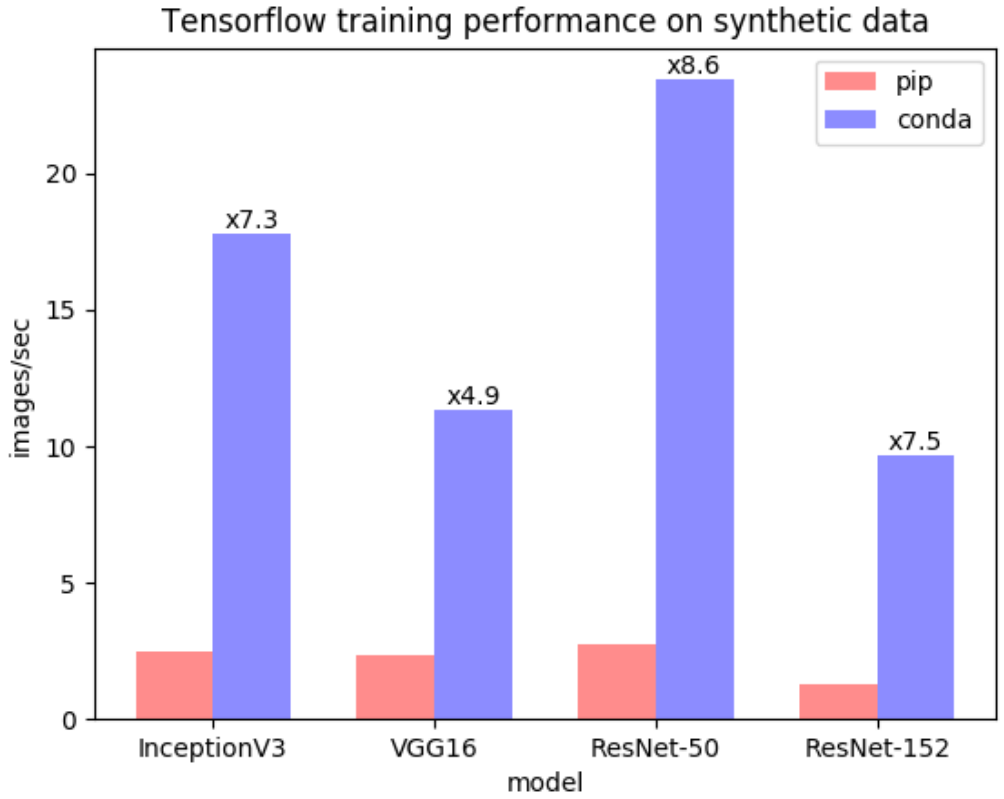
TMiT

Forrás: https://fullstackdeeplearning.com/course/2022/lecture-2-development-infrastructure-and-tooling/

Szoftver stack: kód minősége



Forrás: https://blog.tensorflow.org/2022/01/improved-tensorflow-27-operations-for.html



Forrás: https://towardsdatascience.com/stop-installing-tensorflow-using-pip-for-performance-sake-5854f9d9eb0c

AI önfejlesztés, tanulás

Harc a FOMO-val

- Könyvek
- NVIDIA Deep Learning Institute
- Coursera – Deeplearning.ai
- Saját projektek

BME

- Tárgyak, MSc specializáció
- ARtificial Intelligence Skills Alliance
  https://academy.aiskills.eu/

arisa

Gyires-Tóth Bálint (toth.b@tmit.bme.hu)

Összefoglalás

MI architektúrák ismerete

MI skálázhatósága

Hardver és szoftver architektúra fontossága

Köszönöm a figyelmet

https://www.linkedin.com/in/gyires-toth/