BENCHMARKING HPC PERFORMANCE FOR STATE-OF-THE-ART AI WORKLOADS

Gianluca Mittone, *Iacopo Colonnelli, Robert Birke, Marco Aldinucci - PhD candidate - University of Turin, Computer Science Department, Italy*

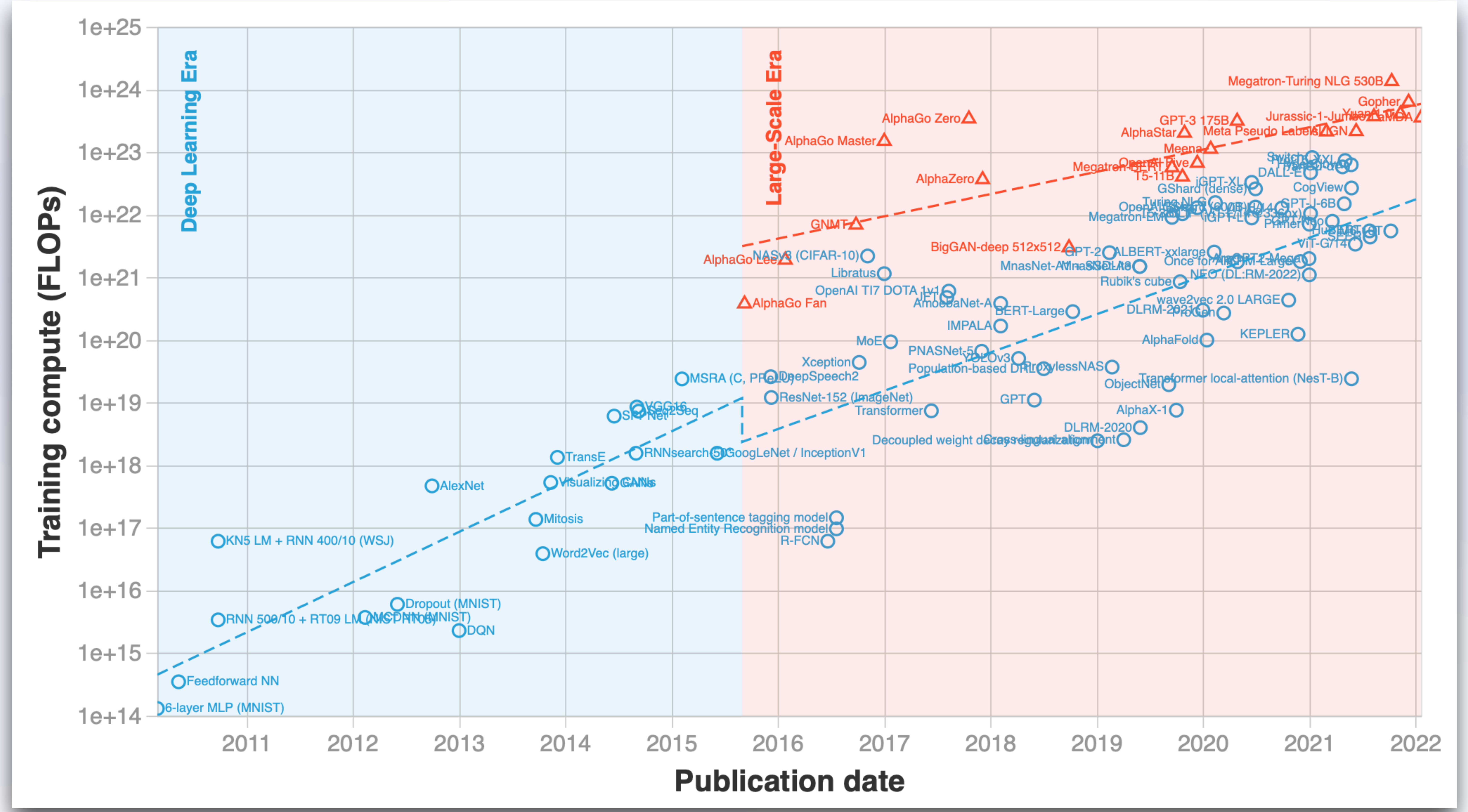
Pisa, Italy - September 18, 2024

- **THE DIGITAL DIVIDE**

- **PUBLIC COMPUTE SCENARIO**
- **LLMS AS AN HPC BENCHMARK**
- **FIRST EXPERIMENTAL RESULTS**
- **CONCLUSIONS**

UNIVERSITÀ DI TORINO

MODELS ARE GETTING LARGER...



The compelling **growth in resource requirements** of current ML models is reaching never-seen peaks, with a few Big Tech companies leading the state-of-the-art, with academia being increasingly impaired in competing due to a lack of resources

In 2012, **AlexNet** significantly impacted the ML community with its astonishing image recognition performance, obtained with 62M parameters trained on two GPUs. Ten years later, **GPT-3** accounts for 175B parameters trained on 1024 GPUs

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022, July). Compute trends across three eras of machine learning. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

… AND SO ARE DATACENTERS

"AI chips are often sold at high prices. Chip company Nvidia CEO Jensen Huang told CNBC earlier in March that the latest "Blackwell" B200 artificial intelligence chip will be priced between $30,000 and $40,000. […]

The report said the new project would be designed to work with chips from different suppliers."

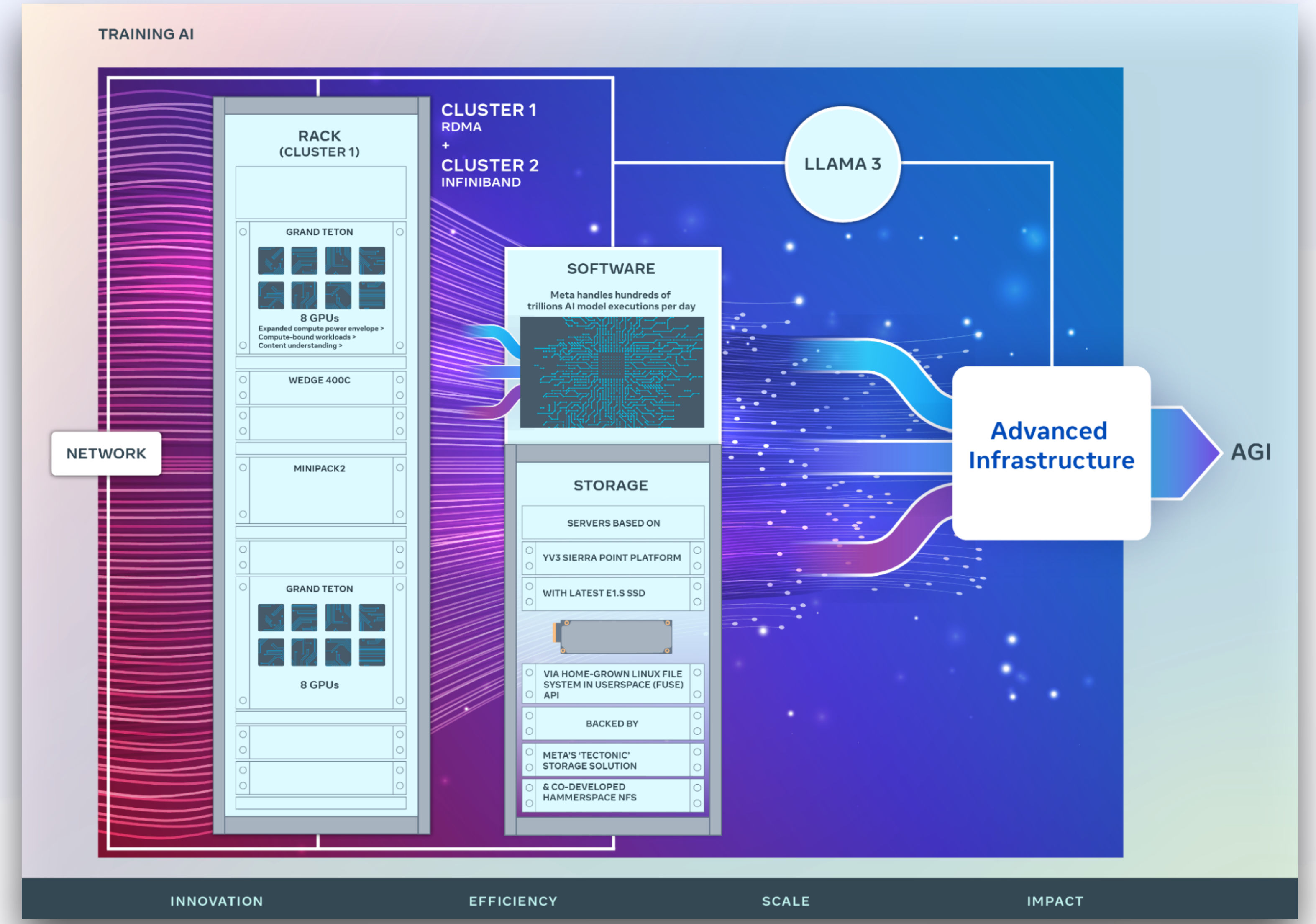
Technology | Data Privacy

Microsoft, OpenAI plan $100 billion data-center project, media report says

By **Reuters**

March 29, 2024 10:14 PM GMT+1 · Updated 6 months ago

Aa



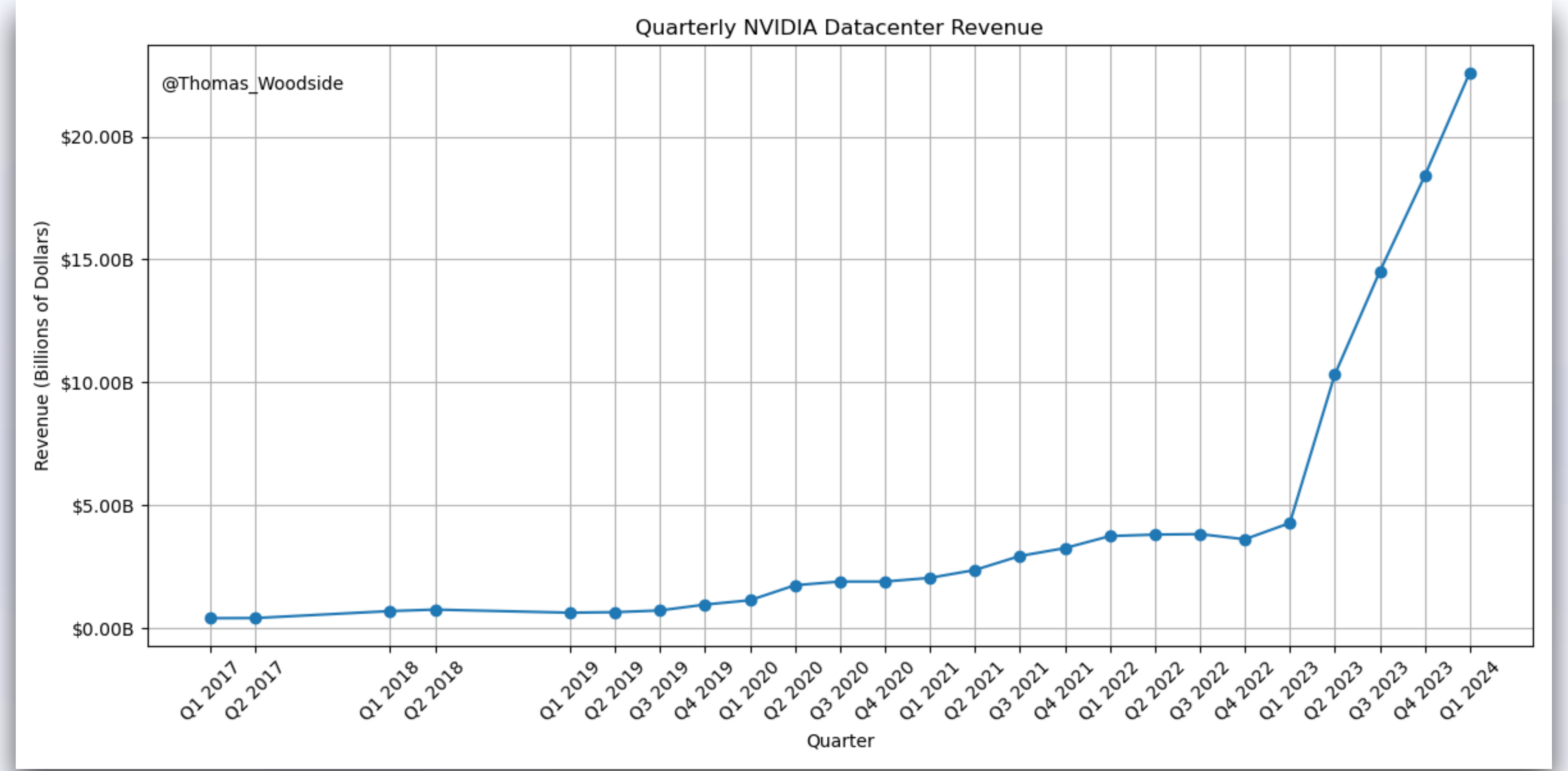
"By the end of 2024, we're aiming to continue to grow our infrastructure build-out that will include 350,000 NVIDIA H100 GPUs as part of a portfolio that will feature compute power equivalent to nearly 600,000 H100s."

Meta

https://www.reuters.com/technology/microsoft-openai-planning-100-billion-data-center-project-information-reports-2024-03-29/

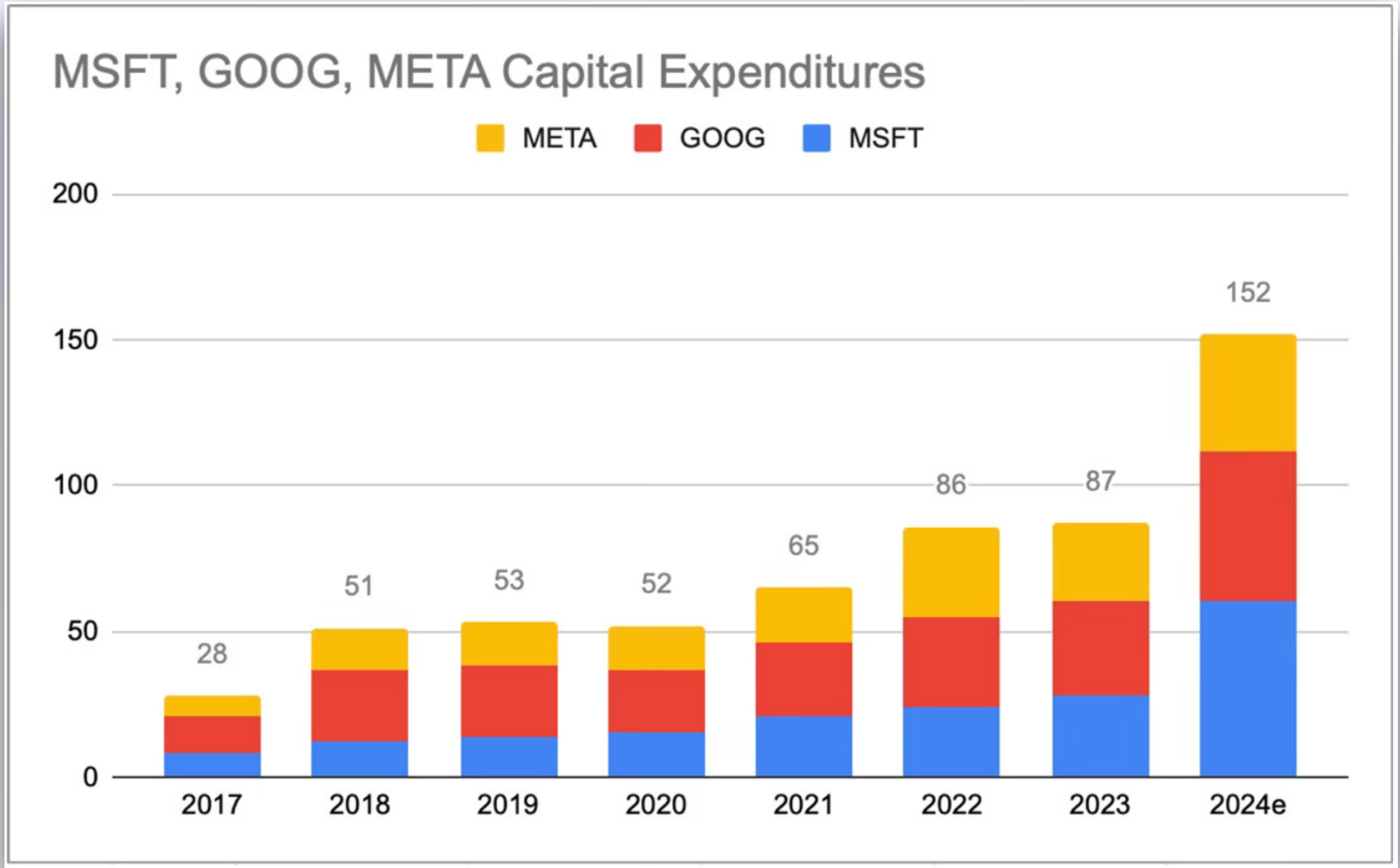
https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/

BIG TECH IS LEADING THE GAME



LLMs are driving this unprecedented growth in "Orders of Magnitude." **How can academia keep up with this?**

Year	Annual investment	AI accelerator shipments (in H100s-equivalent)	Power as % of US electricity production	Chips as % of current leading-edge TSMC wafer production
2024	~$150B	~5-10M	1-2%	5-10%
~2026	~$500B	~10s of millions	5%	~25%
~2028	~$2T	~100M	20%	~100%
~2030	~$8T	~100s of millions	100%	4x current capacity



Aschenbrenner, L. (2024, June) Situational Awareness: The Decade Ahead. https://situational-awareness.ai/

- THE DIGITAL DIVIDE
- **PUBLIC COMPUTE SCENARIO**
- LLMS AS AN HPC BENCHMARK
- FIRST EXPERIMENTAL RESULTS
- CONCLUSIONS

UNIVERSITÀ DI TORINO

THE EUROPEAN HPC JOINT UNDERTAKING (EUROHPC JU)

SYSTEM*	SITE (COUNTRY)	ARCHITECTURE	PARTITION	TOTAL RESOURCES**	FIXED ALLOCATION
MN5 MARENOSTRUM	BSC (ES)	Atos BullSequana XH3000	**MN5 ACC**	129 377	32 000
LEONARDO CINECA	CINECA (IT)	Atos BullSequana XH2000	**Leonardo Booster**	545 865	50 000
LUMI	CSC (FI)	HPE Cray EX	**LUMI-G**	351 455	35 000
MELUXINA HIGH PERFORMANCE COMPUTING IN LUXEMBOURG	LuxProvide (LU)	Atos BullSequana XH2000	**MeluXina GPU**	25 000	25 000
KAROLINA	IT4I VSB-TUO (CZ)	HPE Apollo 2000 Gen10 Plus and HPE Apollo 6500	**Karolina GPU**	7 500	7 500
VEGA HPC	IZUM Maribor (SI)	Atos BullSequana XH2000	**Vega GPU**	7 100	7 100

8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 32C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR, EVIDEN EuroHPC/BSC Spain	663,040	175.30	249.44	4,159
7	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,824,768	241.20	306.31	7,494
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107
89	**MeluXina - Accelerator Module** - BullSequana XH2000, AMD EPYC 7452 32C 2.35GHz, NVIDIA A100 40GB, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, EVIDEN LuxProvide Luxembourg	99,200	10.52	15.29	390
135	**Karolina, GPU partition** - Apollo 6500, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Infiniband HDR200, HPE IT4Innovations National Supercomputing Center, VSB-Technical University of Ostrava Czechia	71,424	6.75	9.08	311
226	**VEGA HPC CPU** - BullSequana XH2000, AMD EPYC 7H12 64C 2.6GHz, Mellanox InfiniBand HDR100, EVIDEN IZUM Slovenia	122,880	3.82	5.37	

https://eurohpc-ju.europa.eu/eurohpc-ju-access-call-ai-and-data-intensive-applications_en

TOP 500 The List.

PRACE

THE TRILLION PARAMETER CONSORTIUM

"[...] given the scale of the effort to prepare datasets for training and the scale of cycles that need to be allocated to build and train a model, it became clear that while the community could develop a number of smaller models independently, and compete for cycles, a broader "**AI for Science**" community must work together if we are to create models that are at the scale of the largest private models."

https://tpc.dev/

The founding partners of TPC are from the following organizations (listed in organizational alphabetical order):

- Agency for Science, Technology and Research (A*STAR)
- Amazon Web Services, Inc (AWS)
- AI Singapore
- Allen Institute For AI
- AMD
- Argonne National Laboratory
- Australian National University
- Barcelona Supercomputing Center
- Brookhaven National Laboratory
- CalTech
- CEA
- LAION
- Lawrence Berkeley National Laboratory
- Lawrence Livermore National Laboratory
- Leibniz Supercomputing Centre
- Los Alamos National Laboratory
- Max Planck Computing & Data Facility (MPCDF)
- Microsoft
- National Center for Supercomputing Applications
- National Energy Technology Laboratory
- Sandia National Laboratories
- Seoul National University
- SLAC National Accelerator Laboratory
- Sony Research
- Stanford University
- STFC Rutherford Appleton Laboratory, UKRI
- Stonybrook University
- SURF
- Texas Advanced Computing Cent
- Thomas Jefferson National Accelerator Facility

AN AI-ORIENTED BENCHMARK: MLPERF

Benchmark	Performance metrics	Application domain	Data volume	Comments
HPL, HPL-AI	Flops, Flops/Watt	Random dense system of linear equations	Variable	Used in Top500 and Green500 to rank supercomputers. Problem size scaled to optimize the performance for machine size. HPL measures performance at double precision, HPL-AI measures performance in mixed precision
HPCAI500	Valid Flops, Valid Flops/Watt	Image classification, Weather analytics	150 GB & 1.65 TB	Convolution and GEMM layers measure the performance in valid Flops which impose penalty based on failure to meet target accuracy. Limited to Microbenchmarks, Object Detection and Image Classification tasks with microscopic view of common deep learning models (Faster-RCNN, ResNet)
Deep500	Throughput, Time to solution	any machine learning application	150 GB	Provides infrastructure to help evaluate different framework implementations and multiple levels of operators. Challenging to integrate into scientific applications. Evaluated with ImageNet dataset.
MLPerf HPC	Time to train	Cosmology and weather analytics	5.1 TB & 8.8 TB	Targets representative scientific machine learning applications with massive datasets. Provision of two types submissions, closed and open enable novel optimizations. Time to solution metric and focused timing captures holistic model performance

"The MLPerf Training benchmark suite measures how fast systems can train models to a **target quality** metric. Current and previous results can be reviewed through the results dashboard below."

"The **strong scaling** metric measures the wall clock time required to train a model on the specified dataset to achieve the specified quality target […] The **weak scaling** metric benchmark measures the throughput for a supercomputing system training multiple models concurrently on the specified dataset to achieve the specified quality target."

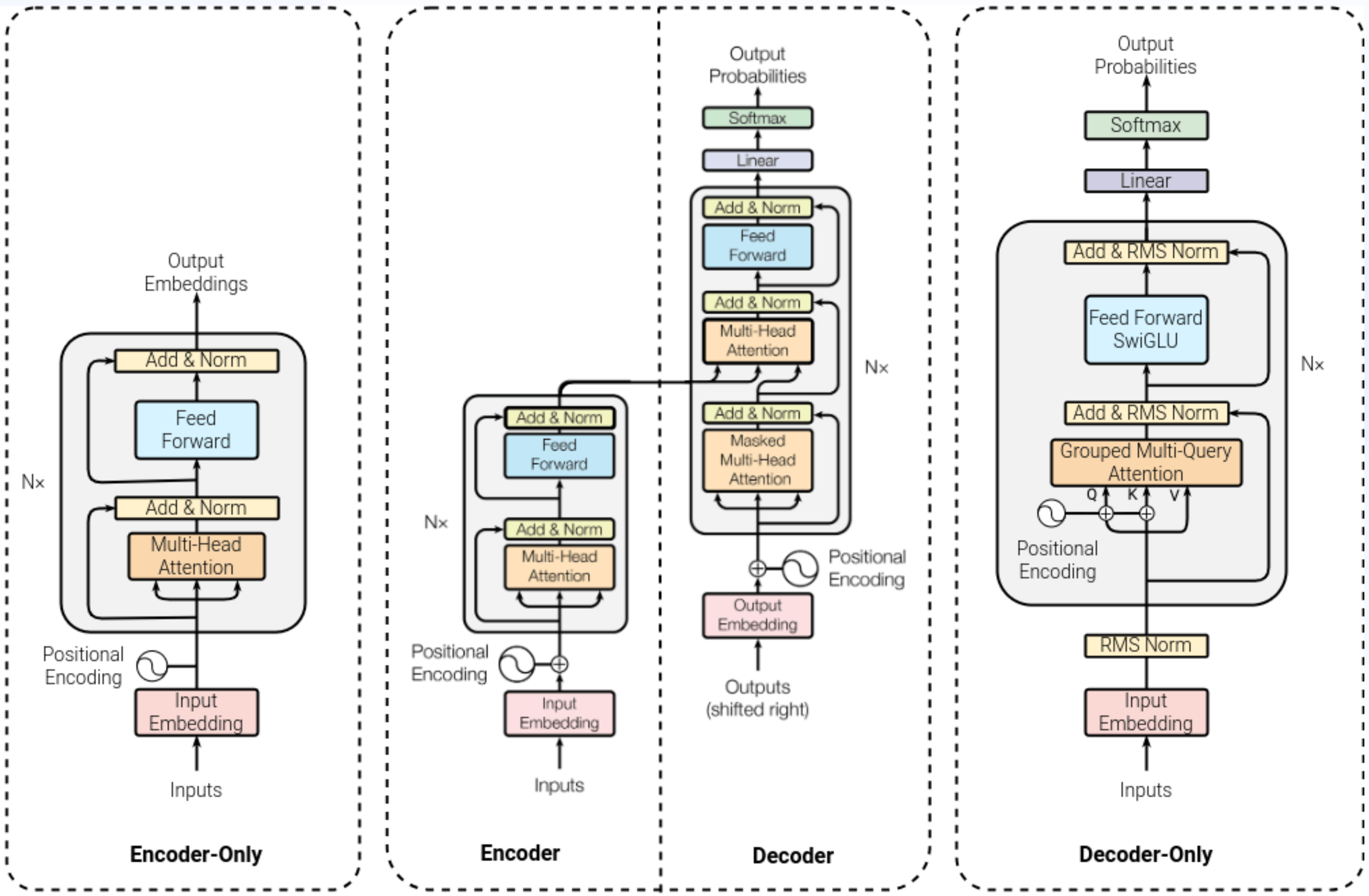
Organization	System Name	Host Processor Model Name	Host Processors ..	Accelerator Model Name	Accelerators Per Node	Software
ANL	theta_gpu_n128_pt1.7.1	AMD EPYC 7742 64-Core Pr..	32	NVIDIA A100-SXM4-40GB	128	PyTorch 1.7.1
ANL	theta_gpu_n128_pt1.9.0	AMD EPYC 7742 64-Core Pr..	32	NVIDIA A100-SXM4-40GB	128	PyTorch 1.9.0
CSCS	piz_daint_gpu_n1024_pt1.9.0	Intel(R) Xeon(R) E5-2690 v3 ..	1024	NVIDIA P100-PCIE-16GB	1024	PyTorch 1.9.0
CSCS	piz_daint_gpu_n128_pt1.9.0	Intel(R) Xeon(R) E5-2690 v3 ..	128	NVIDIA P100-PCIE-16GB	128	PyTorch 1.9.0
CSCS	piz_daint_gpu_n256_pt1.8.0	Intel(R) Xeon(R) E5-2690 v3 ..	256	NVIDIA P100-PCIE-16GB	256	PyTorch 1.8.0
Fujitsu/RIKEN	fugaku_A64FX_tensorflow	FUJITSU Processor A64FX	512	NULL	0	TensorFlow 2.2.0 + Mesh TensorFlow
HelmholtzAI	horeka_gpu_n512_pytorch1.10	Intel Xeon Platinum 8368	256	NVIDIA A100-PCIE-40GB	512	PyTorch 1.10
HelmholtzAI	juwelsbooster_gpu_n1024_mxnet..	AMD EPYC 7402	512	NVIDIA A100-SXM4-40GB	1024	MXNet 1.9
HelmholtzAI	juwelsbooster_gpu_n1024_pytorc..	AMD EPYC 7402	512	NVIDIA A100-SXM4-40GB	1024	PyTorch 1.10
HelmholtzAI	juwelsbooster_gpu_n2048_pytorc..	AMD EPYC 7402	1024	NVIDIA A100-SXM4-40GB	2048	PyTorch 1.10
HelmholtzAI	juwelsbooster_gpu_n512_mxnet1.9	AMD EPYC 7402	256	NVIDIA A100-SXM4-40GB	512	MXNet 1.9
LBNL	perlmutter_128x4_ngc21.08_pytor..	AMD EPYC 7763	128	NVIDIA A100-SXM4-40GB	512	PyTorch NVIDIA Release 21.08
LBNL	perlmutter_256x4_ngc21.09_mxnet	AMD EPYC 7763	256	NVIDIA A100-SXM4-40GB	1024	MXNet NVIDIA Release 21.09
LBNL	perlmutter_512x4_ngc21.09_pytor..	AMD EPYC 7763	512	NVIDIA A100-SXM4-40GB	2048	PyTorch NVIDIA Release 21.09

Farrell, S., Emani, M., Balma, J., Drescher, L., Drozd, A., Fink, A., … & Yin, J. (2021, November). MLPerf™ HPC: A holistic benchmark suite for scientific machine learning on HPC systems. In 2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC) (pp. 33-45). IEEE.

- THE DIGITAL DIVIDE
- PUBLIC COMPUTE SCENARIO
- **LLMS AS AN HPC BENCHMARK**
- FIRST EXPERIMENTAL RESULTS
- CONCLUSIONS

UNIVERSITÀ DI TORINO

ANATOMY OF A (MONOLITHYC) LLM



BERT
(Devlin et al., 2018)

Original Transformer
(Vaswani et al., 2017)

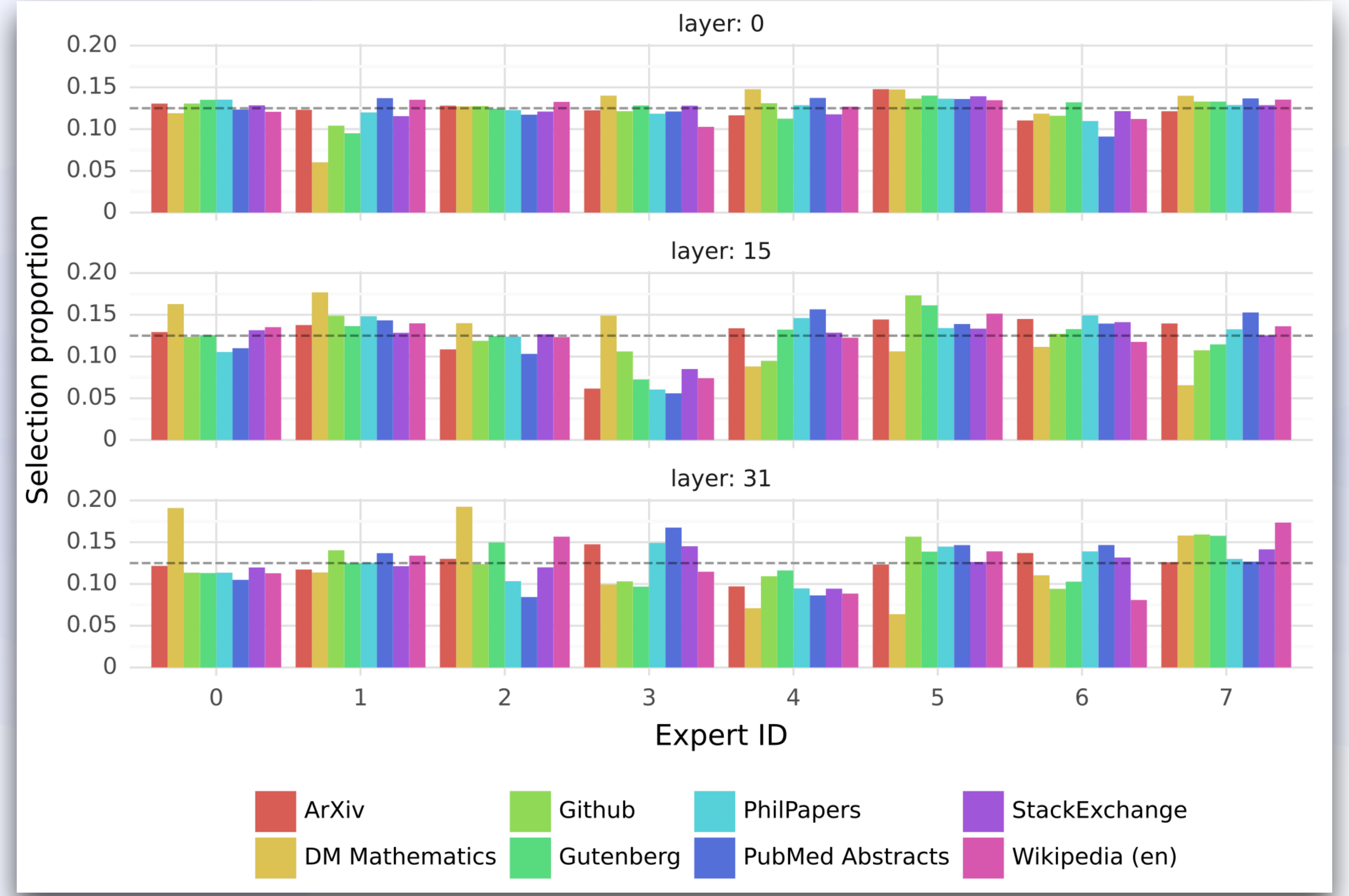
LLaMA
(Touvron et al., 2023)

Different structures, same heart: **multiple parallel attention heads** followed by a **feed-forward NN**. We focus on **decoder-only** architectures.

1. The human language inputs are **tokenised**;
2. Tokens are updated through **positional encoding**;
3. **Multi-head attention layers** update tokens' value based on the other tokens present in the sequence;
4. A **feed-forward NN** updates tokens' value, extracting high-level abstractions of them (**FFNs constitute 2/3 of the parameters!**);
5. **Output probabilities for** the next token prediction are produced.

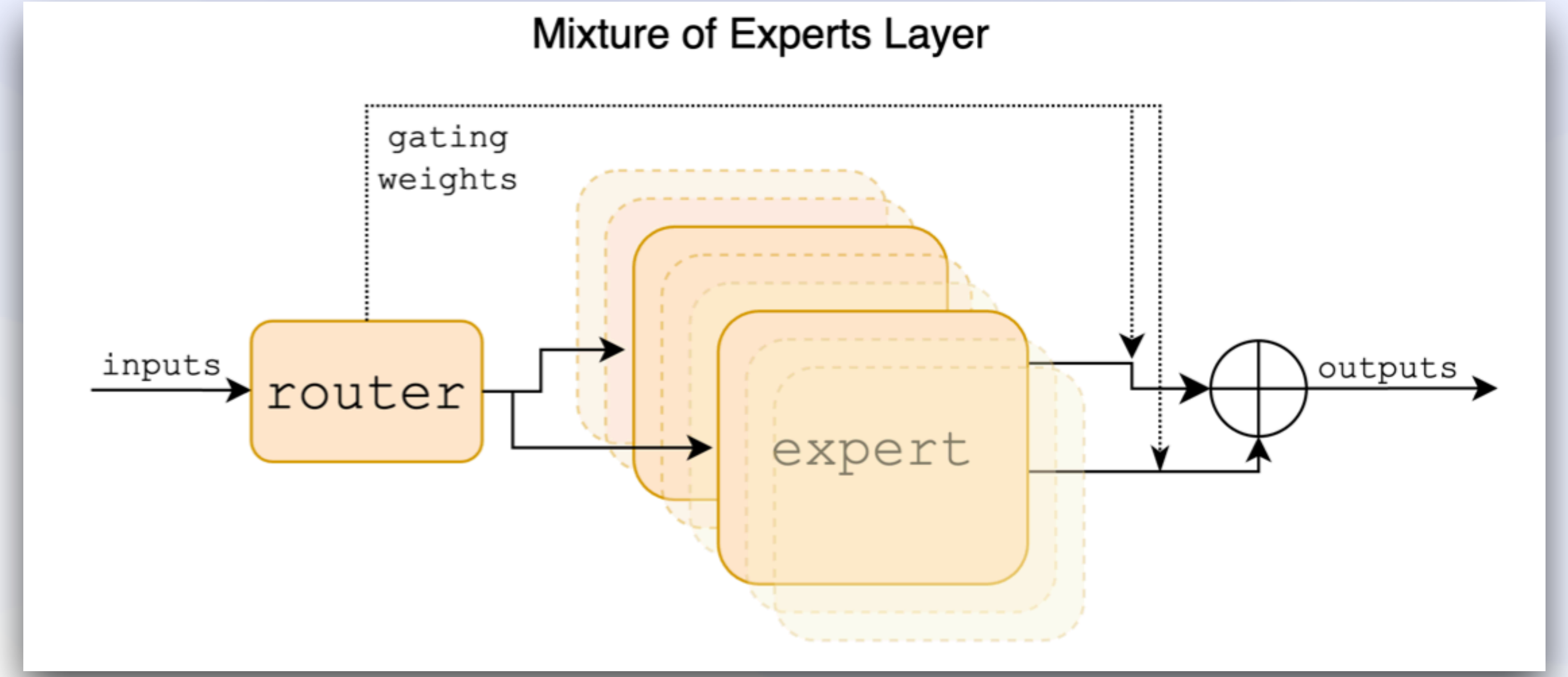
https://www.luminis.eu/blog/llm-series-part-1-a-comprehensive-introduction-to-large-language-models/

MIXTURE-OF-EXPERT LLMS



In a **Mixture-of-Expert** (MoE) LLM, the FFN is replaced by a set of (smaller) FFNs, which activation is handled by a **router** (most commonly an NN layer)

This design **decouples** the total number of parameters of an LLM and the number of parameters effectively needed to process a token, increasing training throughput and lowering inference costs



The experts are **not really "experts"**: the data distribution seen by them observed experimentally starts to be marginally significant in the vary last layers

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., ... & Sayed, W. E. (2024). Mixtral of experts. arXiv preprint arXiv:2401.04088.

PARALLEL AND DISTRIBUTED LLM TRAINING

LLMs' pre-training can be distributed and **parallelised** according to a wide variety of strategies, calibrating **memory occupation**, **load balancing**, and **communication**:

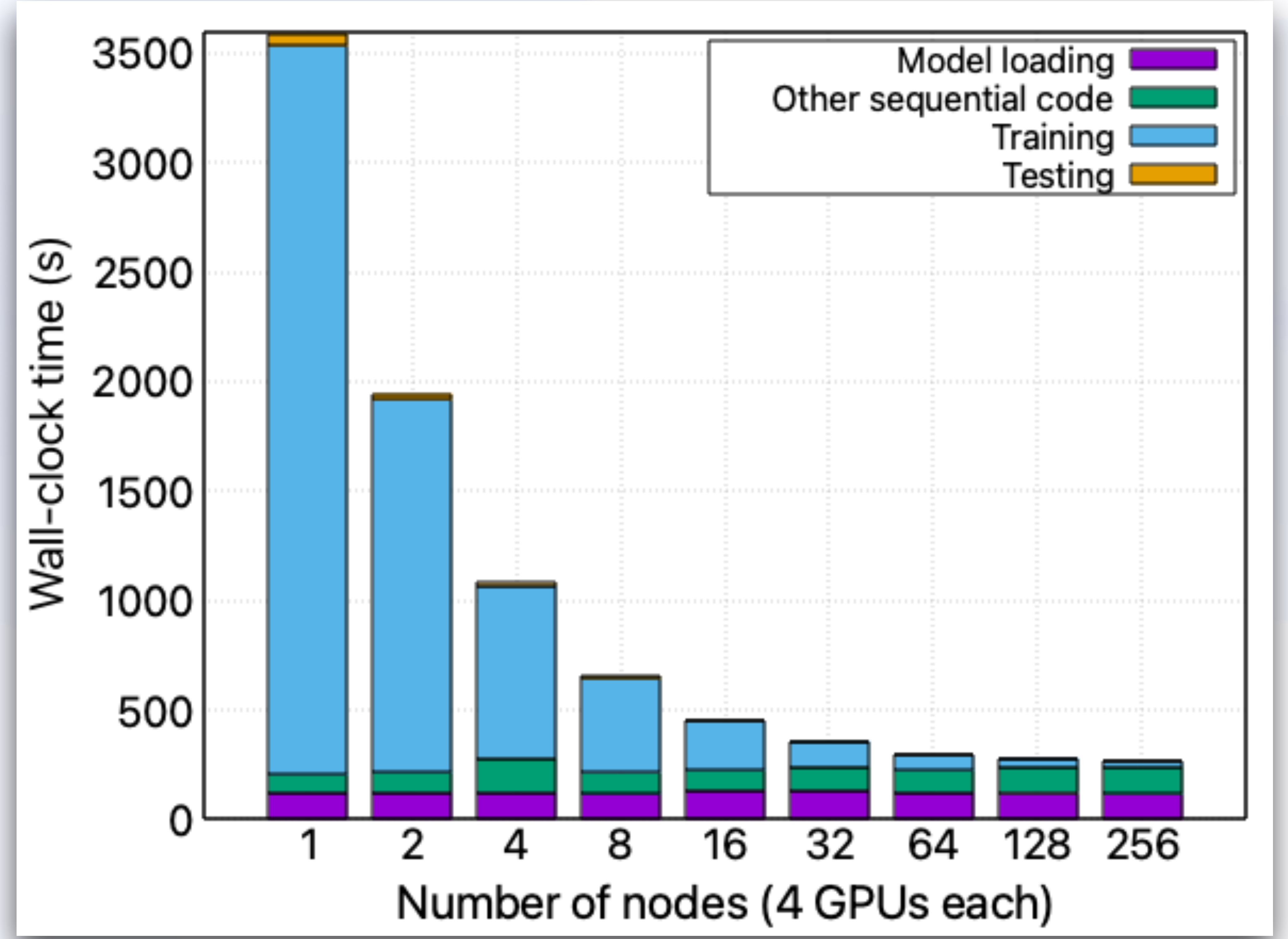
- **Data Parallelism**: multiple copies of the same model processing different data (impacts global batch size, requires synchronisations);
    - **Distributed Data Parallelism**: Data Parallel on multiple nodes (reduces memory occupation);
    - **Fully-Sharded Data Parallelism**: a Distributed Data Parallel approach sharding model's parameters, gradients, and optimiser (reduces memory occupation even more but requires increased communications);
- **Model Parallelism**: the model is partitioned and distributed on multiple computing elements (distributes memory and computing);
- **Pipeline Parallelism**: subdivides a mini-batch into micro-batches and interleaves their processing in a pipeline fashion through the model (optimises computation but requires careful tuning);
- **Tensor Parallelism**: the model's tensors are subdivided, and operations on mini-batches are run in parallel (increased communication);
- **Sequence Parallelism**:mini-batches are subdivided, and tensor operations are run in parallel (increased communication);
- **Expert Parallelism**: a mix of data parallelism and model parallelism in which an MoE model is trained in a Data Parallel fashion apart from the experts' layers, that are in common between all the instances (Model Parallel, balances load but requires more communications).

Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C. C., Xu, M., ... & Li, S. (2023). Pytorch fsdp: experiences on scaling fully sharded data parallel. arXiv preprint arXiv:2304.11277.

- THE DIGITAL DIVIDE
- PUBLIC COMPUTE SCENARIO
- LLMS AS AN HPC BENCHMARK
- **FIRST EXPERIMENTAL RESULTS**
- CONCLUSIONS

UNIVERSITÀ DI TORINO

TRAINING TIME ANALYSIS (FSDP)



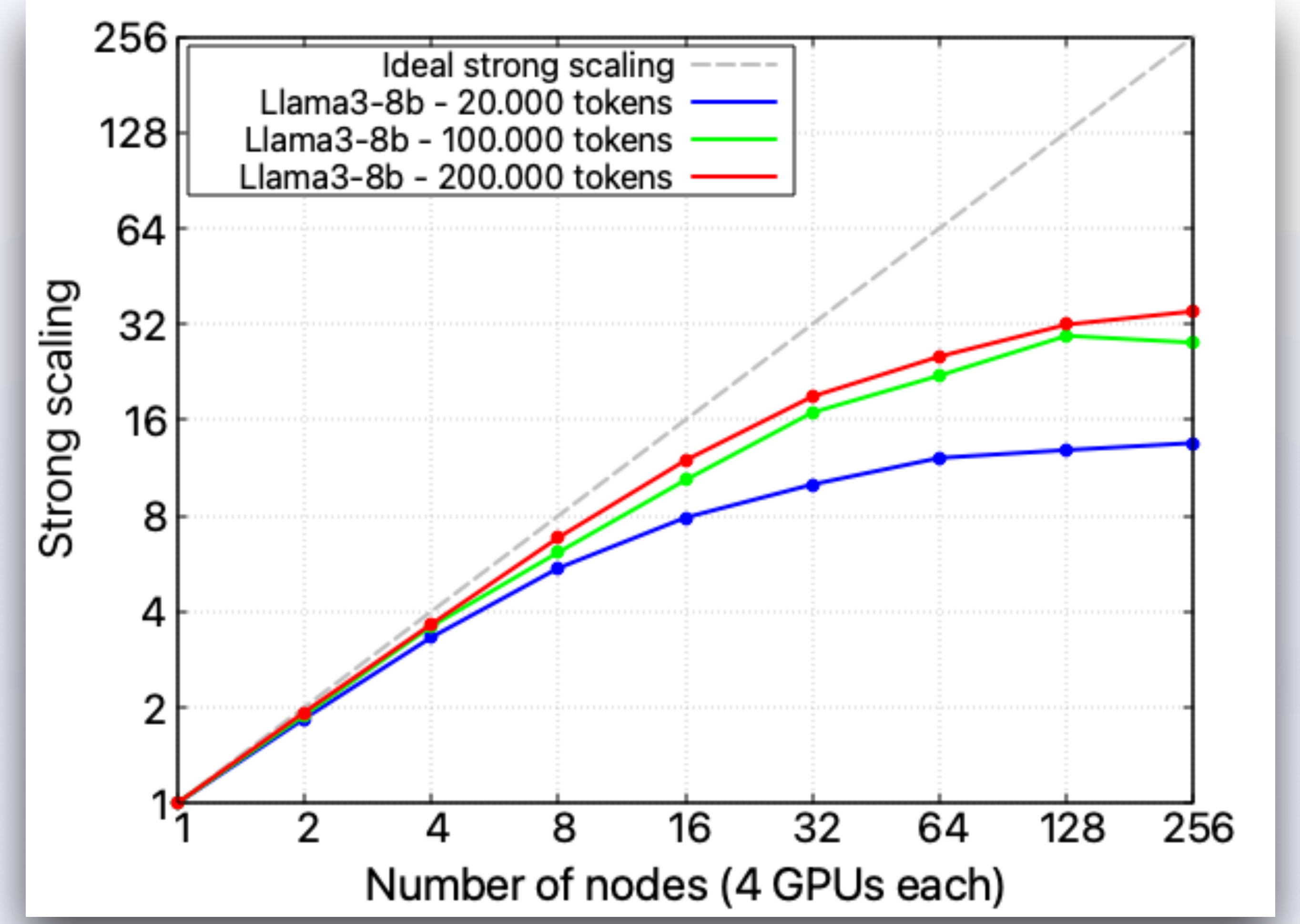
LLaMA-3 (8B version) execution time subdivided in its main components. The training is done on 20,000 training samples of 2,048 tokens each on the Leonardo HPC.

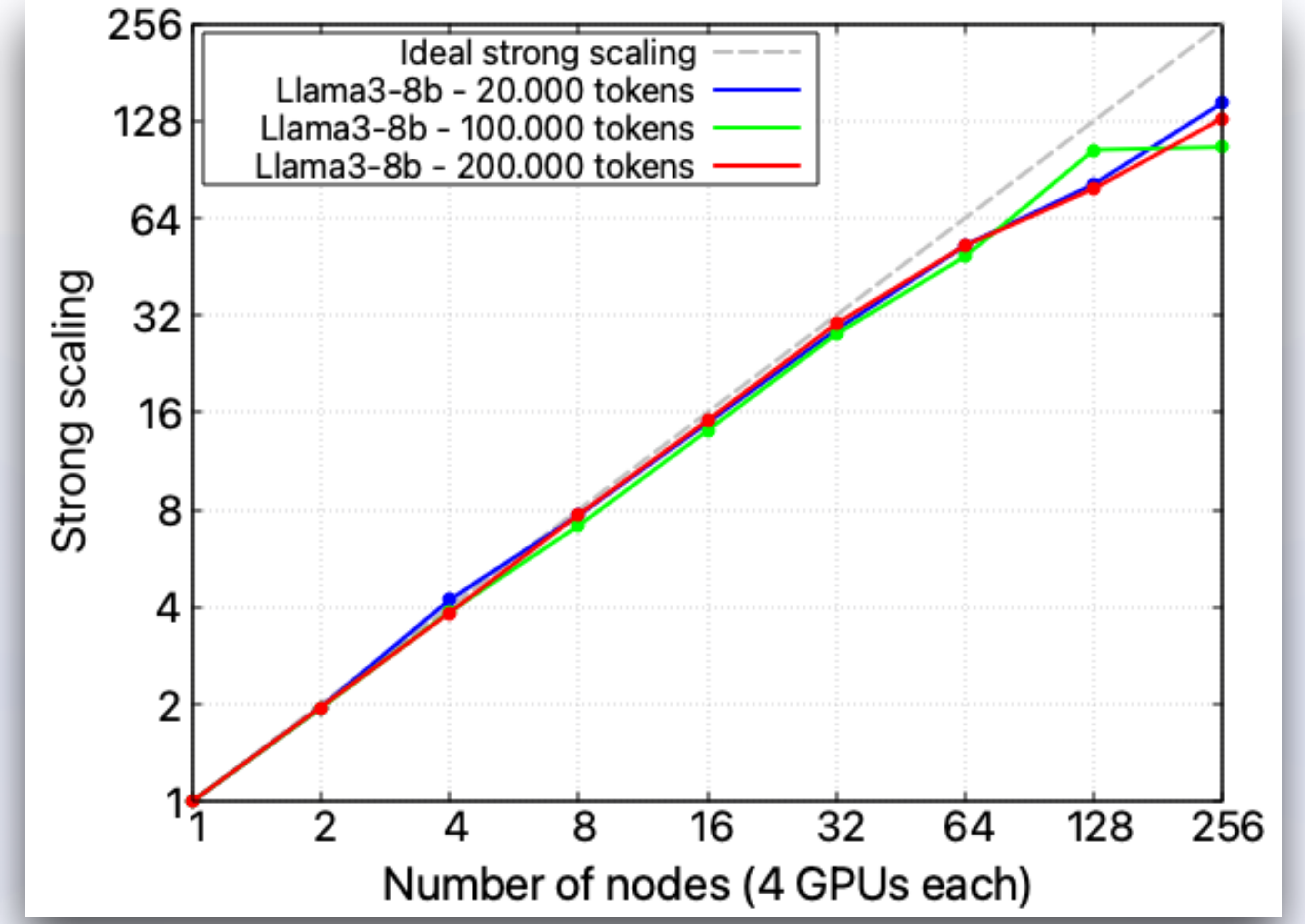
The first thing to be noticed is the code's relatively **poor scalability** performance: the **setup overhead** becomes predominant starting from 16 nodes (64 GPUs). The training itself, on the other hand, seems to scale reasonably well

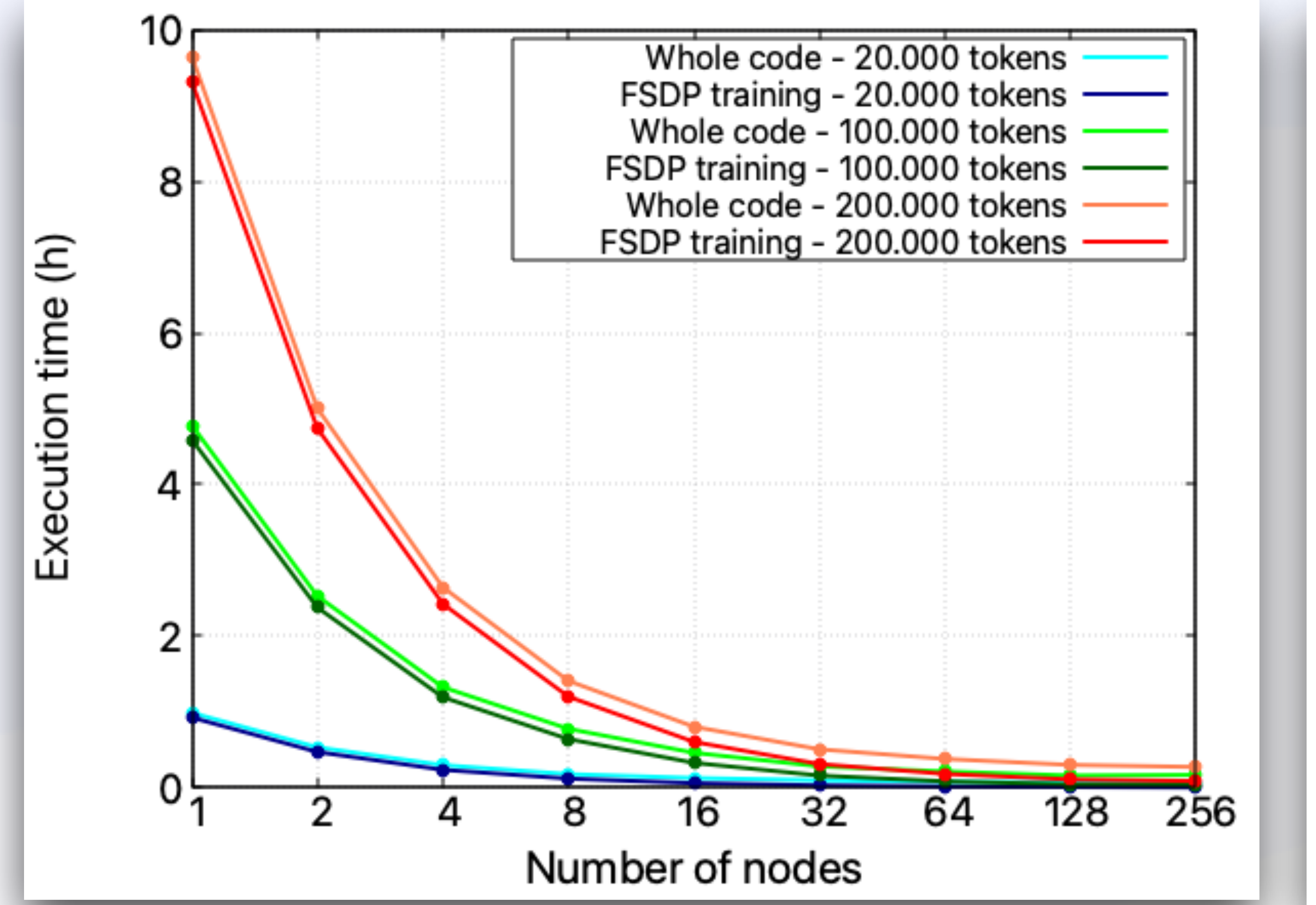
# Nodes	Model loading (s)	Distributed setup (s)	Training (s)	Testing (s)
1	120.6	87.788	3325.4	54.4
2	123.6	95.3375	1700	27
4	120.6	158.44	788.4	13
8	121.8	94.5164	432.6	6
16	131.8	95.7248	223.4	3
32	131.4	109.98325	115	1
64	122	109.5124	63.2	~0
128	118.6	119.1875	40.8	~0
256	117.2	124.8	22.8	~0

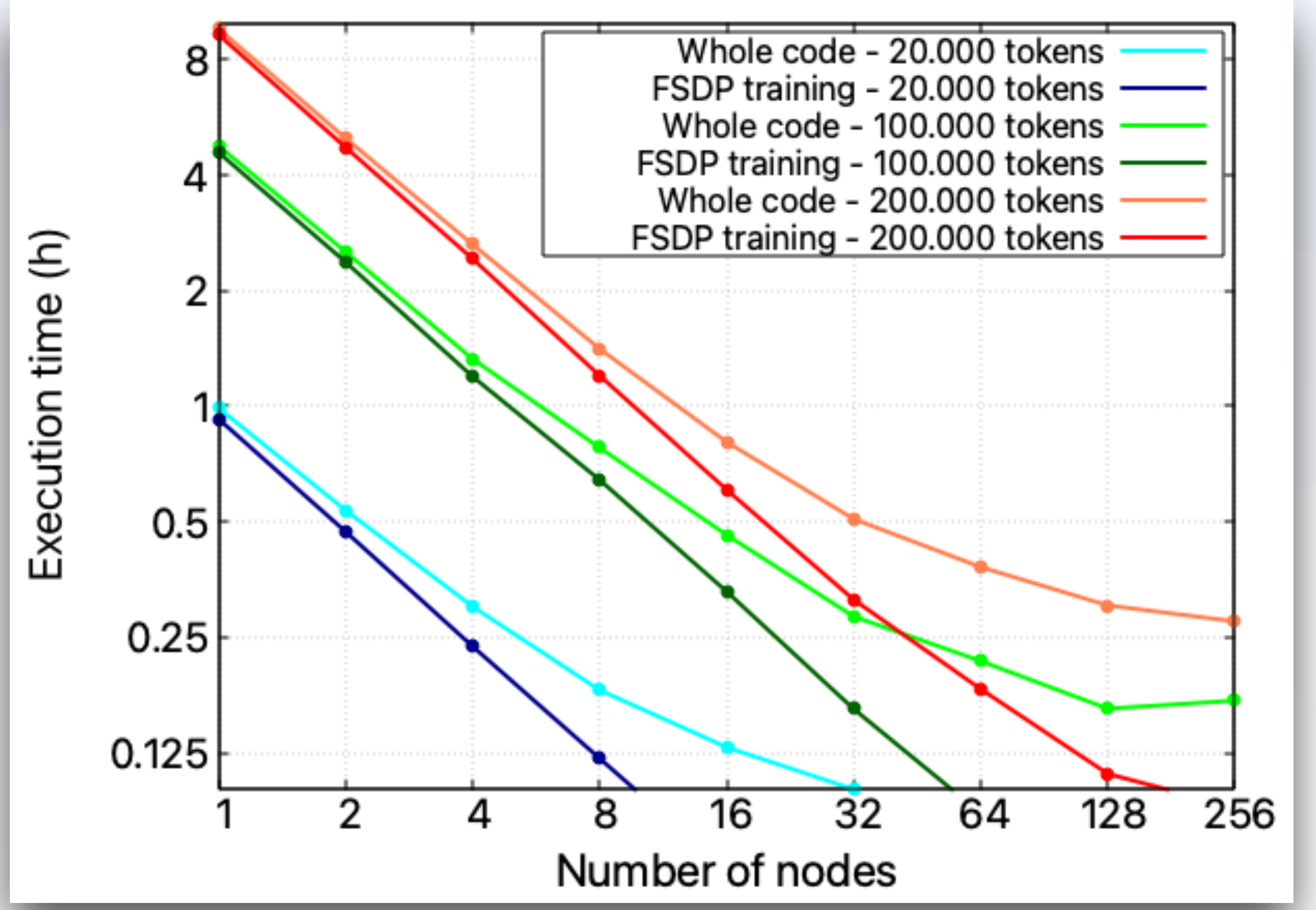
Colonnelli, I., Birke, R., Malenza, G., Mittone, G., Mulone, A., & Aldinucci, M. (2024). Cross-Facility Federated Learning - Part II. Presented at the ELISE Wrap-Up Conference & ELLIS Community Event.

TRAINING TIME ANALYSIS (FSDP)







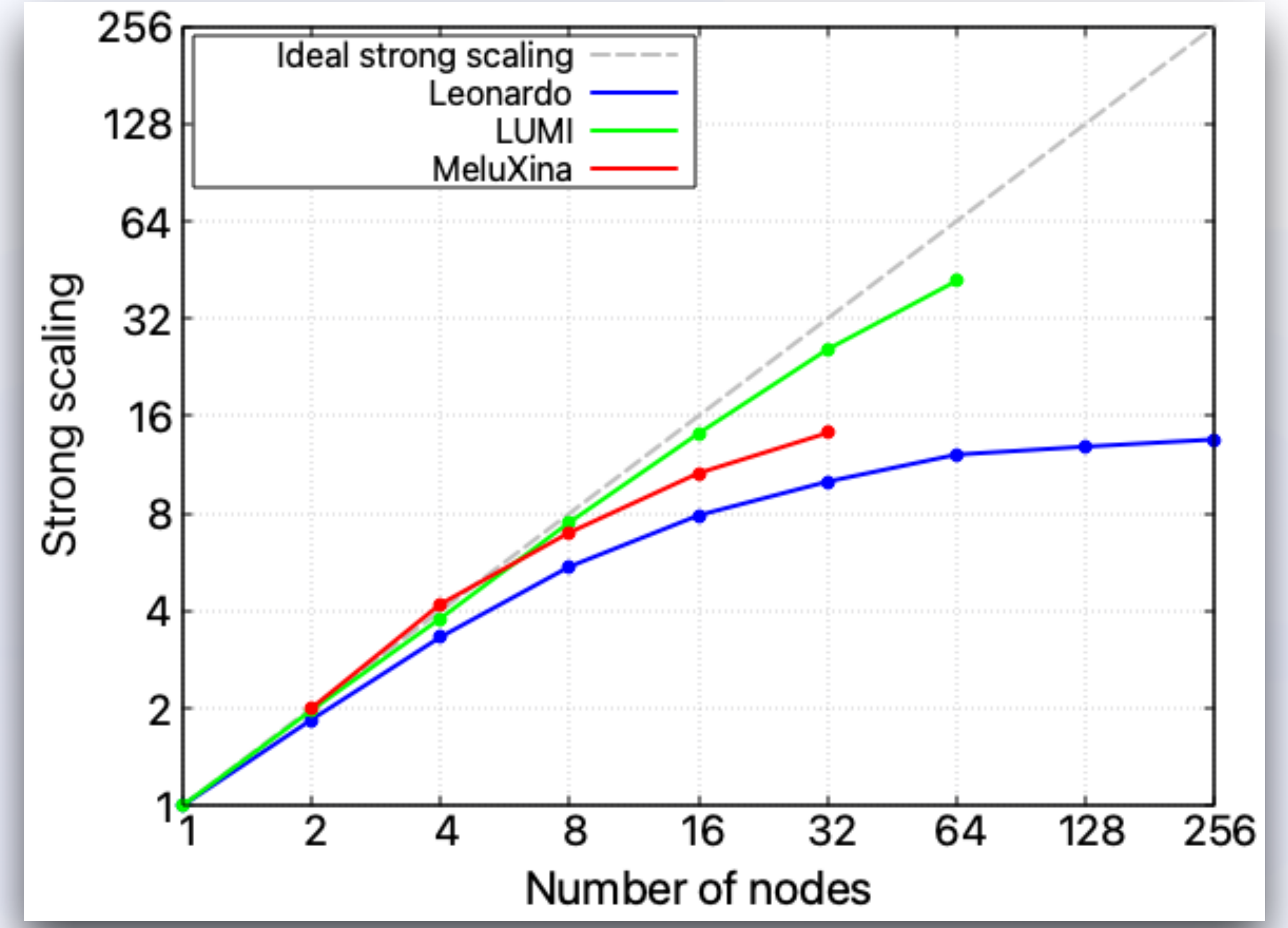


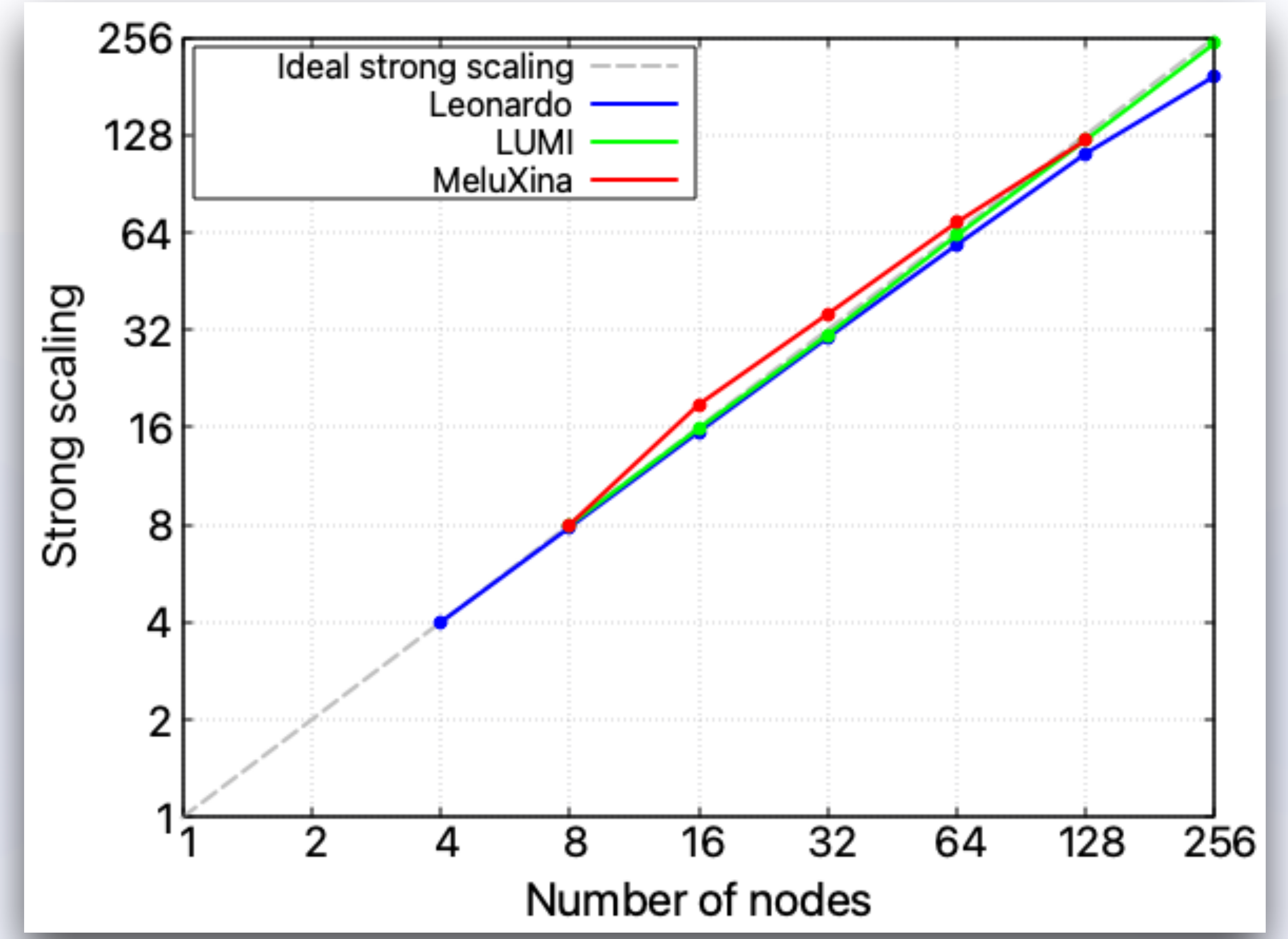
LLaMA-3 (8B version) scaling performance on Leonardo. Both the whole deployment code and the FSDP code are analysed. The training is done on 20,000 train- ing samples of 2,048 tokens each on the Leonardo HPC.

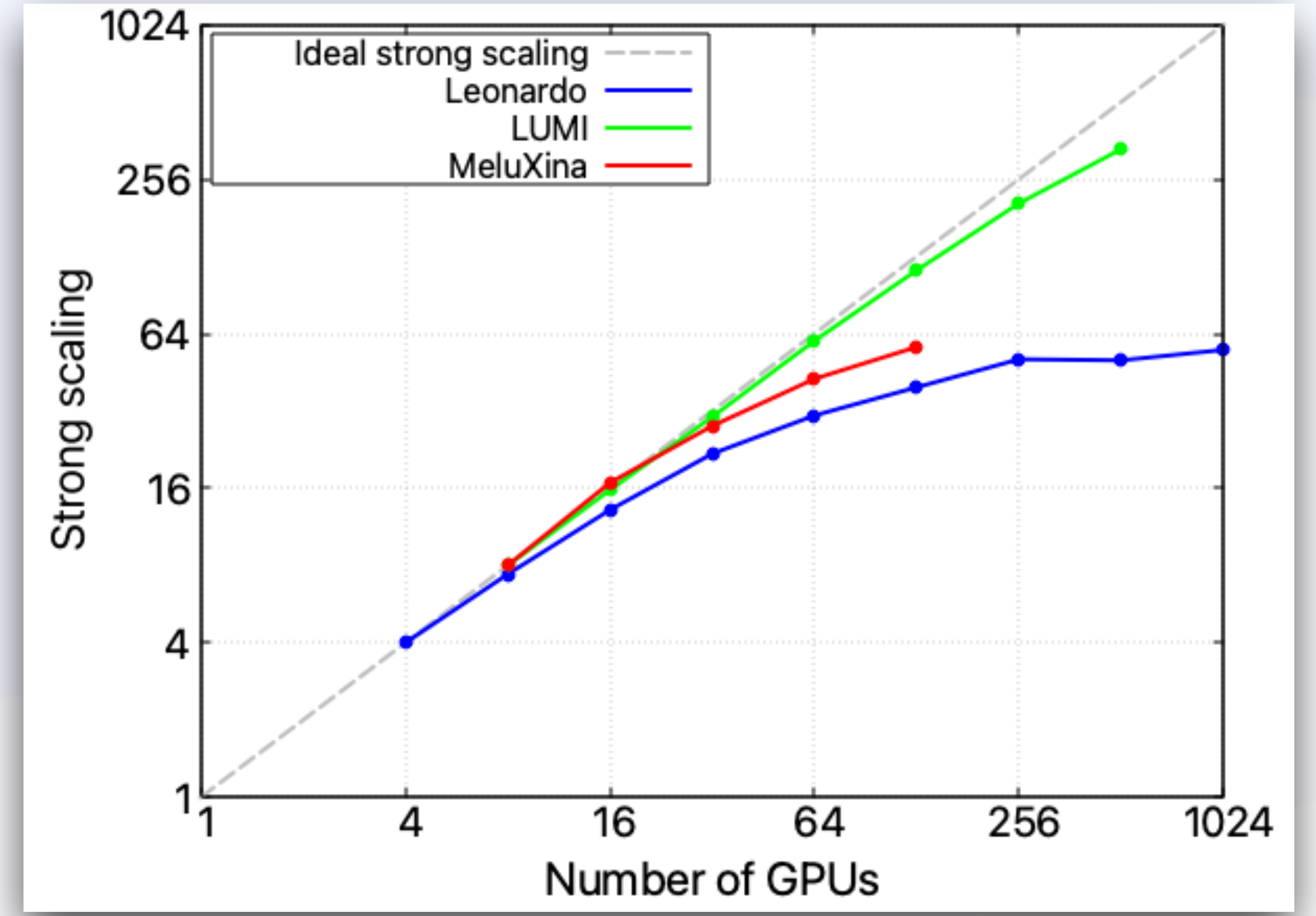
These performance issues do not seem to be correlated with the **problem's size** (i.e., the size of the training dataset) or related to the FSDP distributed training technique. These statements are confirmed by the fact that increasing the training dataset size does **not change the code's scaling behaviour** and that isolating the performance of the FSDP code section ensures its nice scalability performance up to 128 nodes.

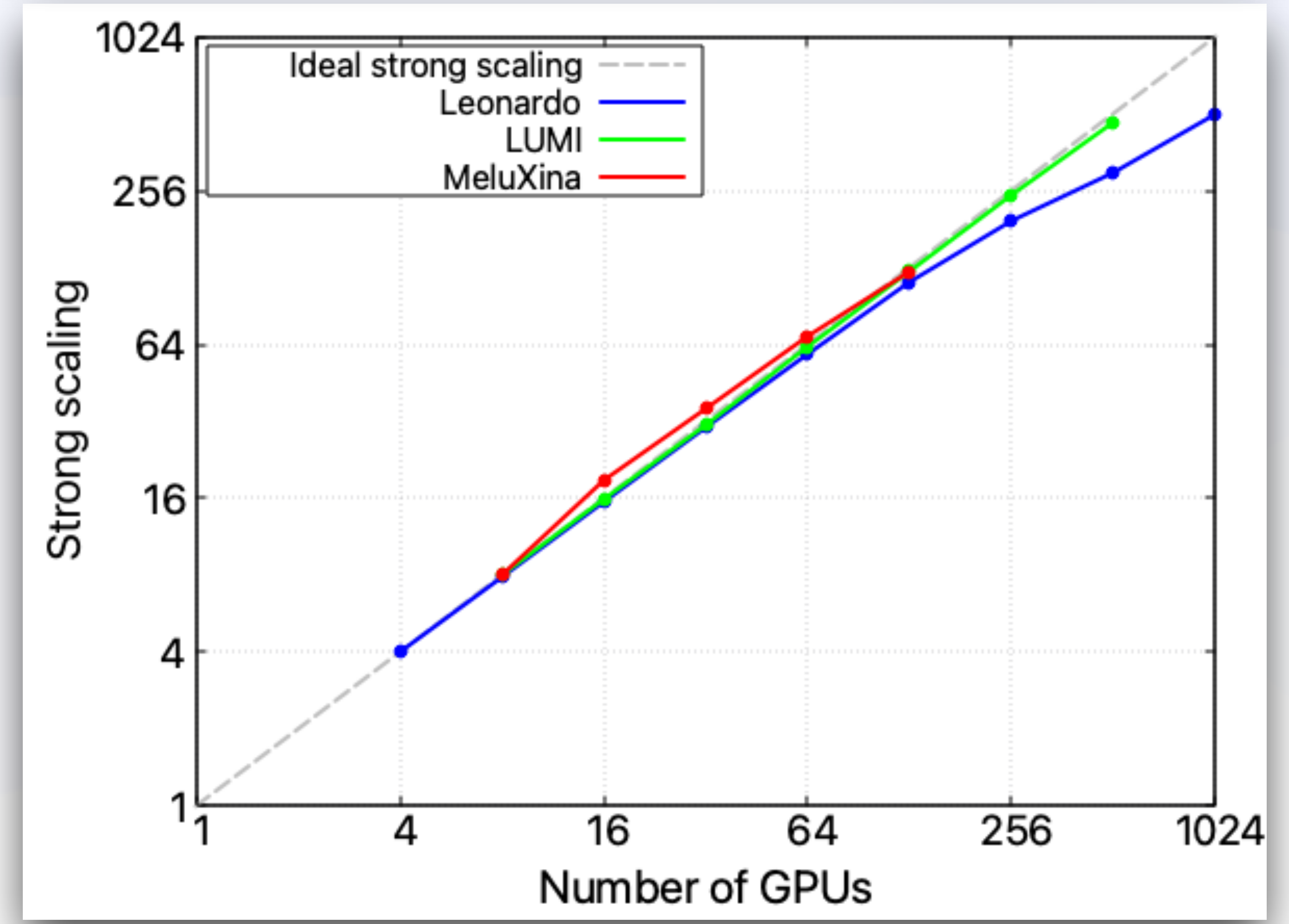
Colonnelli, I., Birke, R., Malenza, G., Mittone, G., Mulone, A., & Aldinucci, M. (2024). Cross-Facility Federated Learning - Part II. Presented at the ELISE Wrap-Up Conference & ELLIS Community Event.

DIFFERENT HPCS, DIFFERENT SCALING









Comparison between LLaMA-3 (8B version) scaling performance on Leonardo, LUMI and MeluXina. Both the whole deployment code and the FSDP code are analysed. The training is done on a different number of tokens on each HPC infrastructure to accommodate the different computing power.

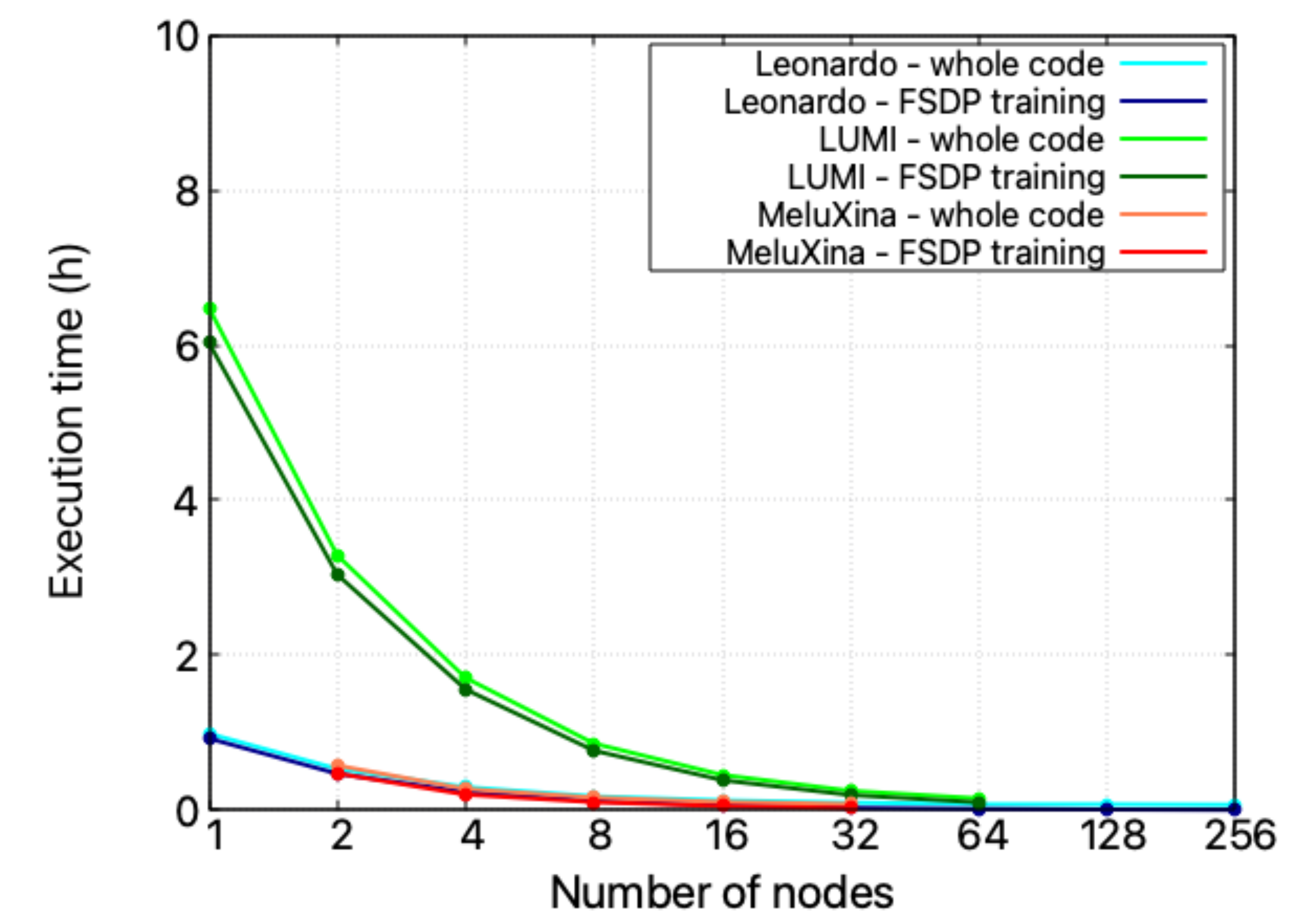
When investigating this performance issue and comparing all three available HPC infrastructures, it appears clear that **the problem is unrelated to the hardware itself**: also on LUMI and MeluXina the whole code's scalability performance starts to spoil after 16 nodes, while the FSDP component scales reasonably well on all infrastructures. A single instance of LLaMA-3 8 billion can fit into a single Leonardo or LUMI node but requires two nodes on MeluXina.

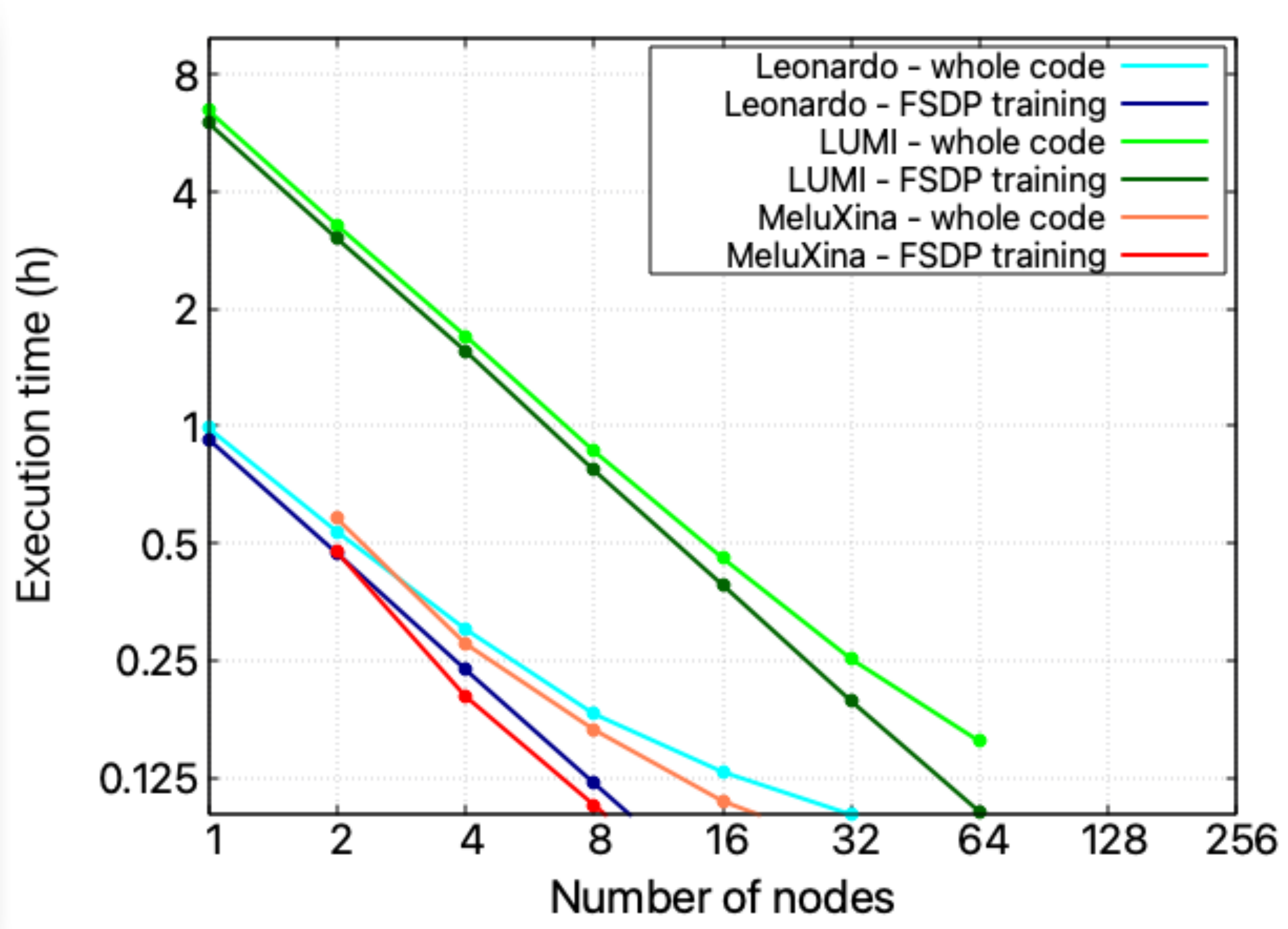
Colonnelli, I., Birke, R., Malenza, G., Mittone, G., Mulone, A., & Aldinucci, M. (2024). Cross-Facility Federated Learning - Part II. Presented at the ELISE Wrap-Up Conference & ELLIS Community Event.

DIFFERENT HPCS, DIFFERENT SCALING

	Leonardo			LUMI			MeluXina	
# Nodes	Queue time (min:sec)	Execution time (min:sec)	# Nodes	Queue time (min:sec)	Execution time (min:sec)	# Nodes	Queue time (min:sec)	Execution time (min:sec)
1	51:08	49:49	1	13:35	50:50	2	00:01	34:33
2	03:03	27:05	2	05:05	26:36	4	11:28	16:46
4	04:35	15:18	4	07:39	14:37	8	50:58	10:36
8	04:29	09:37	8	09:28	08:27	16	41:38	07:26
16	38:11	06:09	16	03:42	05:25	32	00:18	06:09

LLaMA-3 (8B version) queuing and execution times on Leonardo, LUMI and MeluXina. The training is done on a different number of tokens on each HPC infrastructure to accommodate the different computing power.





These numbers do not reflect the Top500 ranking!

Comparison between LLaMA-3 (8B version) scaling performance on Leonar- do, LUMI and MeluXina. Both the whole deployment code and the FSDP code are anal- ysed. The training is done on 16,384 training samples of 2048 tokens each on each HPC infrastructure.

Colonnelli, I., Birke, R., Malenza, G., Mittone, G., Mulone, A., & Aldinucci, M. (2024). Cross-Facility Federated Learning - Part II. Presented at the ELISE Wrap-Up Conference & ELLIS Community Event.

- THE DIGITAL DIVIDE
- PUBLIC COMPUTE SCENARIO
- LLMS AS AN HPC BENCHMARK
- FIRST EXPERIMENTAL RESULTS

- **CONCLUSIONS**

UNIVERSITÀ DI TORINO

BENCHMARKING HPC PERFORMANCE FOR STATE-OF-THE-ART AI WORKLOADS

We have tested a state-of-the-art AI workload on different HPCs and found that the current tools used to assess HPC computing power are unreliable indicators of their performance on such types of workloads.

- LLM training workflows scale differently on different HPC facilities;
- This is mainly due to overhead handling (model loading, PyTorch distributed setup);
- FSDP-training scales well up to 128 nodes on all HPC facilities but with very different compute times;

Future works will investigate:

- Reduce model loading time by using high-end storage and I/O optimisation techniques (e.g., GPUDirect storage);
- Investigate computing and communication bottlenecks at large scales;
- Investigate strategies to avoid PyTorch cold restarts on all nodes (caching, faster setup algorithms);
- **Sum up all these considerations into a single number!**

BENCHMARKING HPC PERFORMANCE FOR STATE-OF-THE-ART AI WORKLOADS

Gianluca Mittone, Iacopo Colonnelli, Robert Birke, Marco Aldinucci - PhD candidate - University of Turin, Computer Science Department, Italy

[Parallel Computing group [ALPHA]]

UNIVERSITÀ DI TORINO

ITA DATA 2024

EXTRA

EXPERIMENTAL SETUP

MODEL: ResNet-18

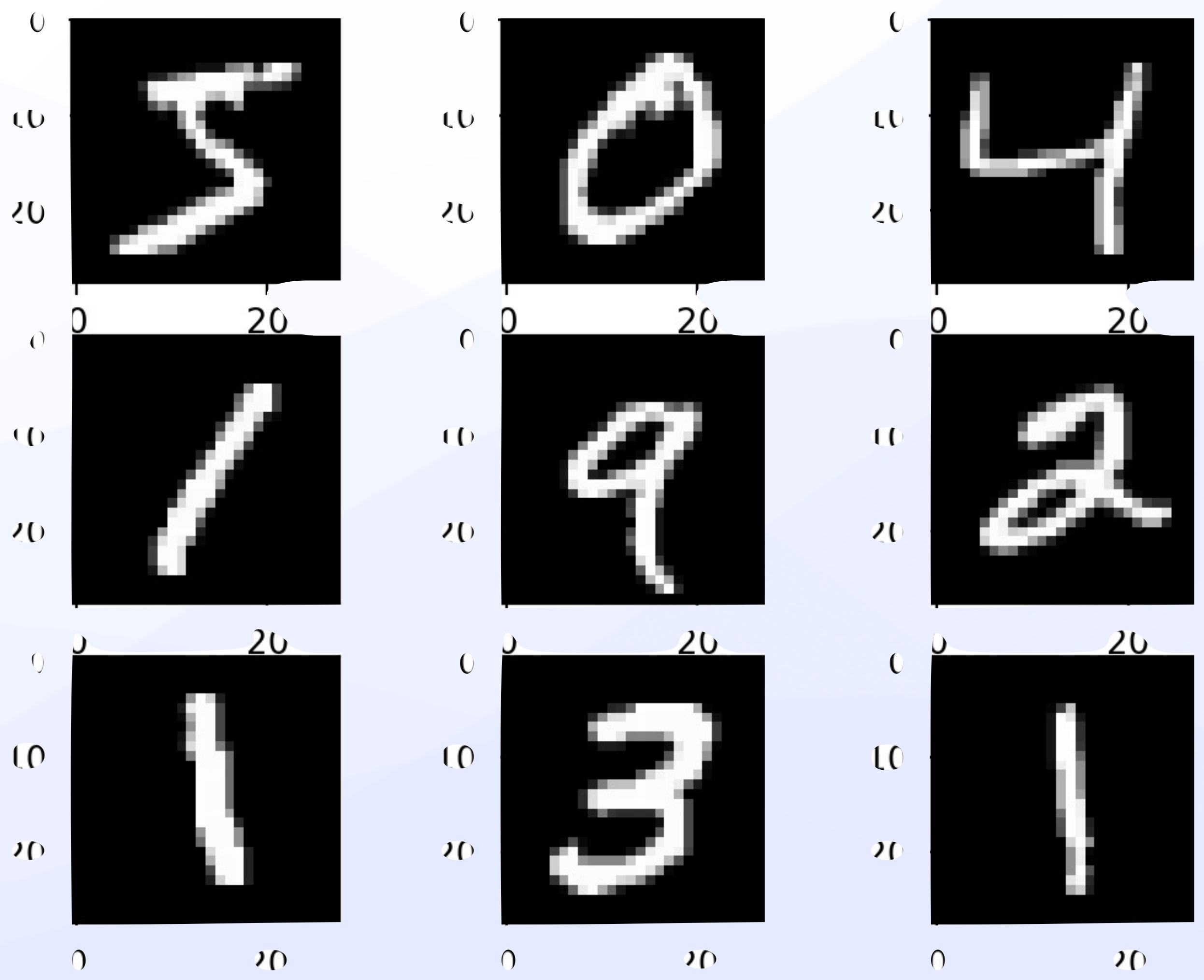
- Standard Convolutional Neural Network
- ~11 million trainable parameters

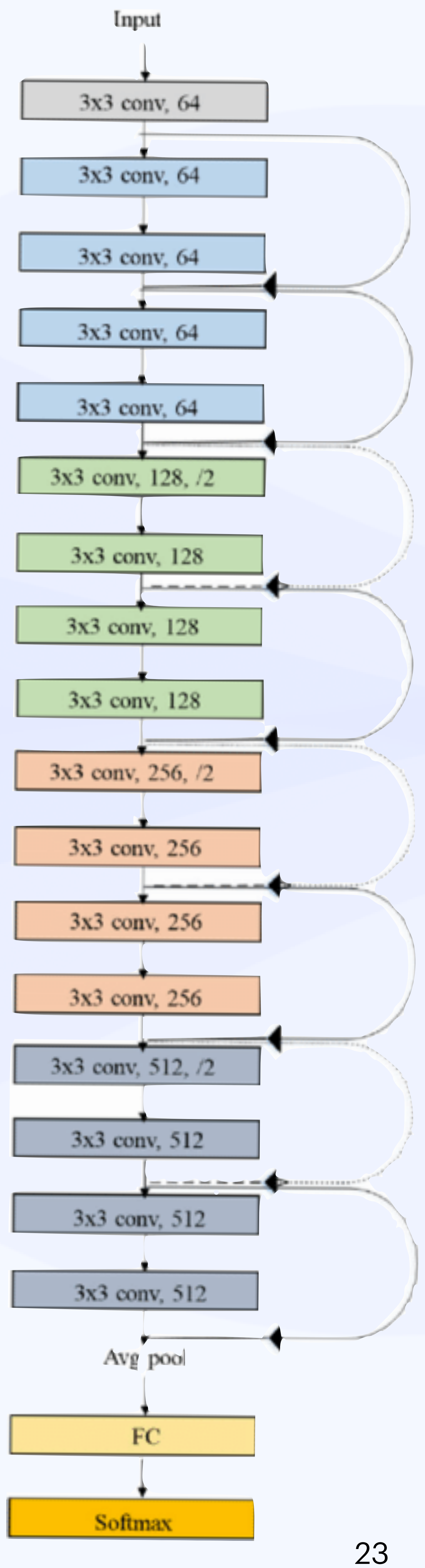
DATASET: MNIST

- Standard benchmarking dataset
- 60.000 train/10.000 test images
- 28x28 pixel

COMPUTING: C3S

- OmniPath network
- ~2 x Intel Xeon CPU E5-2697 v4 per node





He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep residual learning for image recognition". In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Deng, L. (2012). "The mnist database of handwritten digit images for machine learning research". IEEE Signal Processing Magazine, 29(6), 141–142.

Aldinucci, M., Rabellino, S., Pironti, M., Spiga, F., Viviani, P., Drocco, M., ... & Galeazzi, F. (2018, May). "HPC4AI: an ai-on-demand federated platform endeavour". In Proceedings of the 15th ACM International Conference on Computing Frontiers (pp. 279-286).

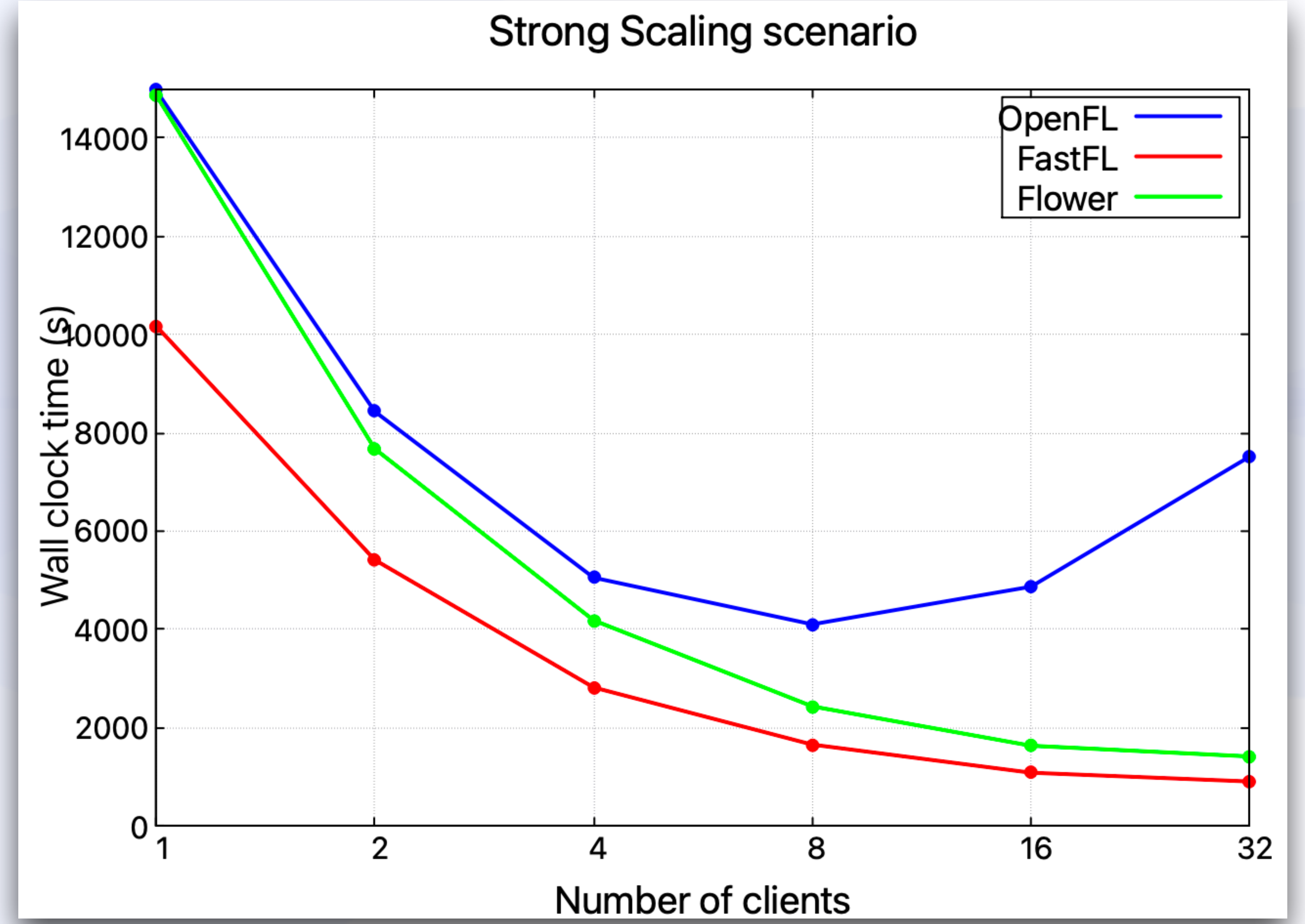
MEASURED WALL CLOCK TIME

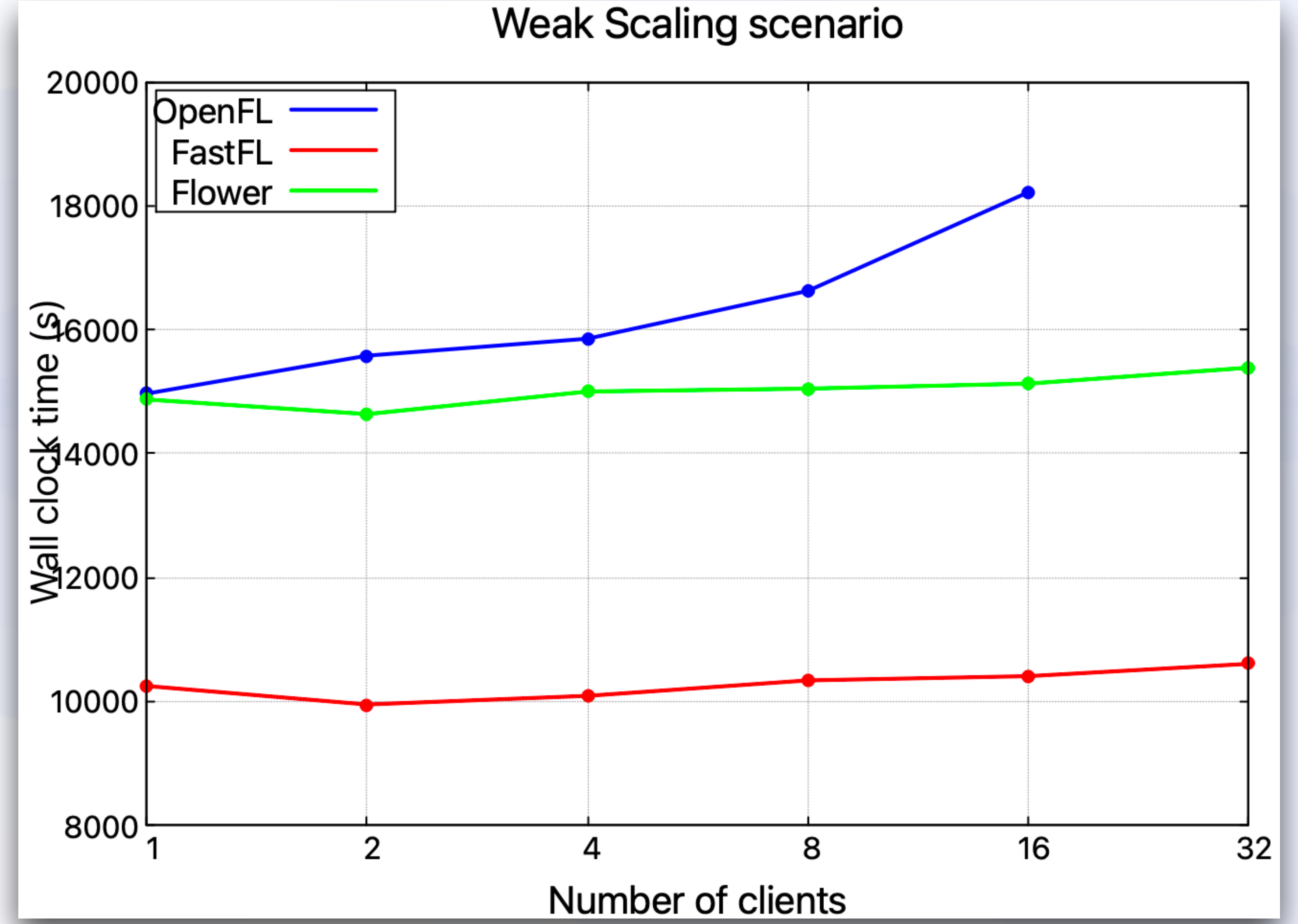
Strong Scaling	1	2	4	8	16	32
OpenFL [2]	14967	8433	5051	4104	4870	7517
Flower [3]	14872	7672	4184	2435	1633	1415
FastFL [4]	10175	5414	2821	1656	1085	905
Weak Scaling	**1**	**2**	**4**	**8**	**16**	**32**
OpenFL	14967	15578	15853	16624	18216	—
Flower	14872	14636	14999	15046	15128	15385
FastFL	10249	9951	10090	10340	10407	10607

- **OpenFL** and **Flower** display different scaling behaviors despite being built with the same technologies
- **Flower** outperforms **OpenFL** in both scenarios.
- **FastFL** is comparable to **Flower**
- **OpenFL** exceeded the maximum runtime for this benchmark in the 32 clients weak scaling scenarios (> 6 hours)

Mittone, G., Fonio, S. (2023). "Benchmarking Federated Learning Scalability".
In Proceedings of the 2nd Italian Conference on Big Data and Data Science (ITADATA 2023). CEUR.

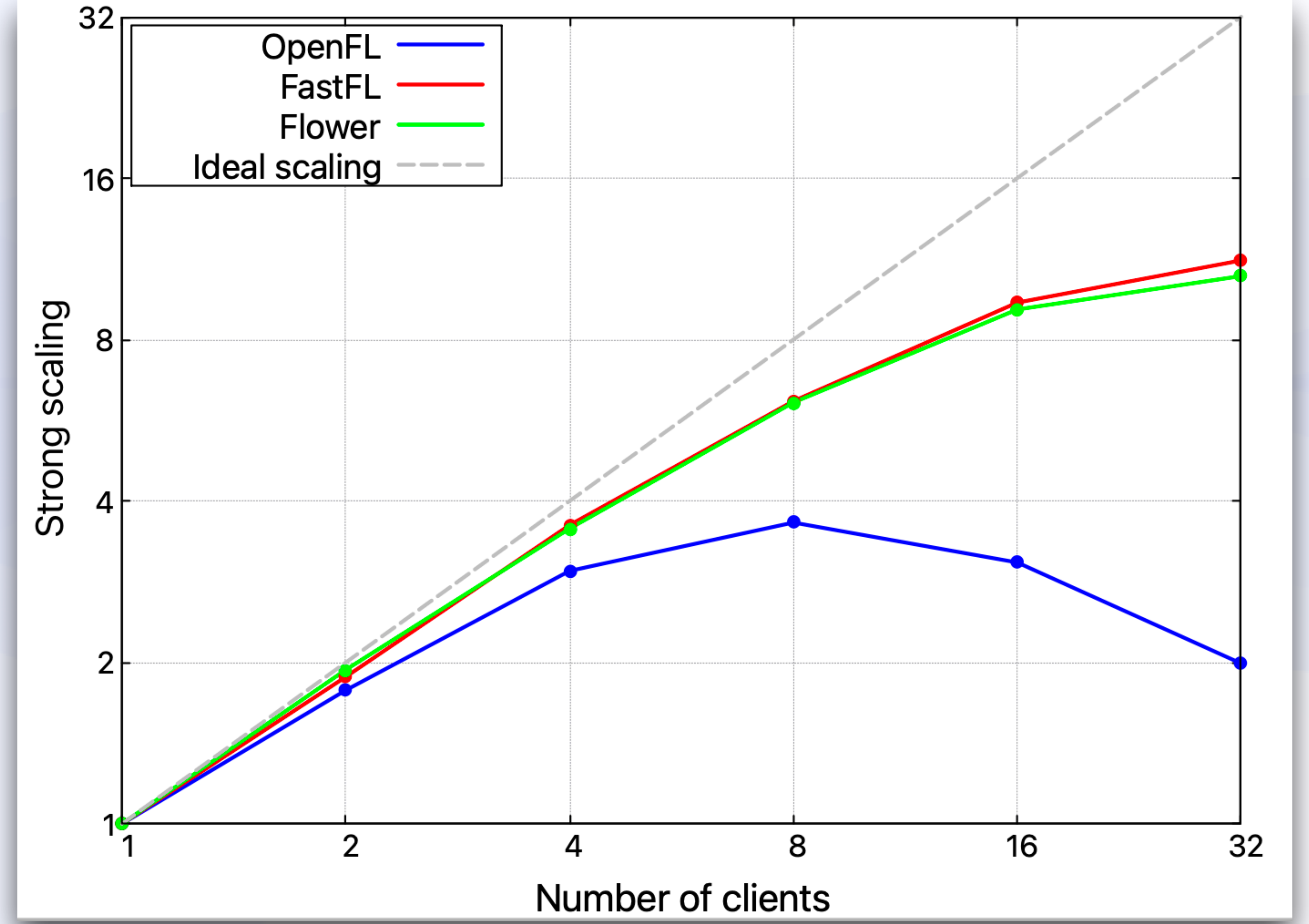
VISUALIZING WALL CLOCK TIME

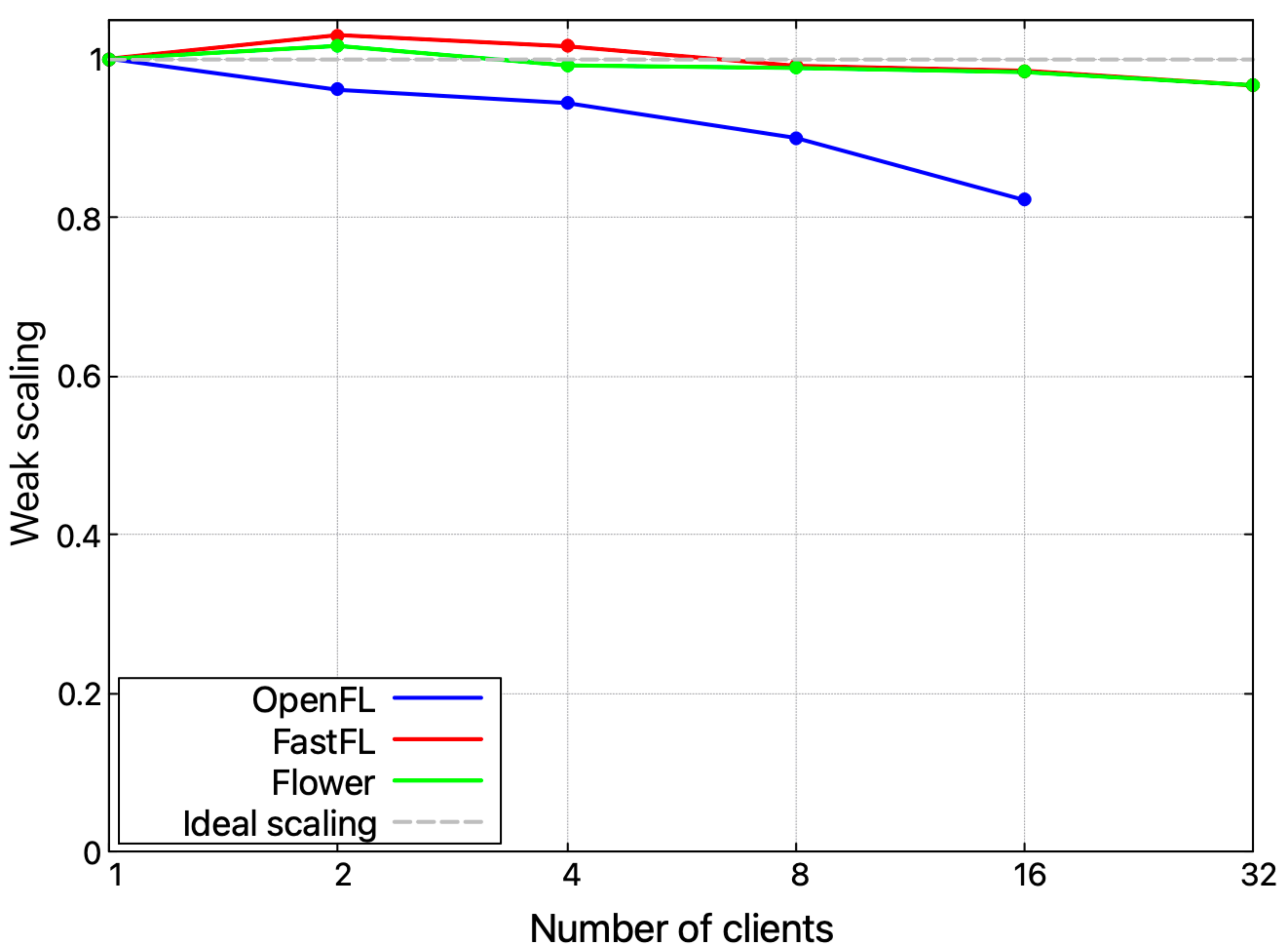




Mittone, G., Fonio, S. (2023). *"Benchmarking Federated Learning Scalability".*
In Proceedings of the 2nd Italian Conference on Big Data and Data Science (ITADATA 2023). CEUR.

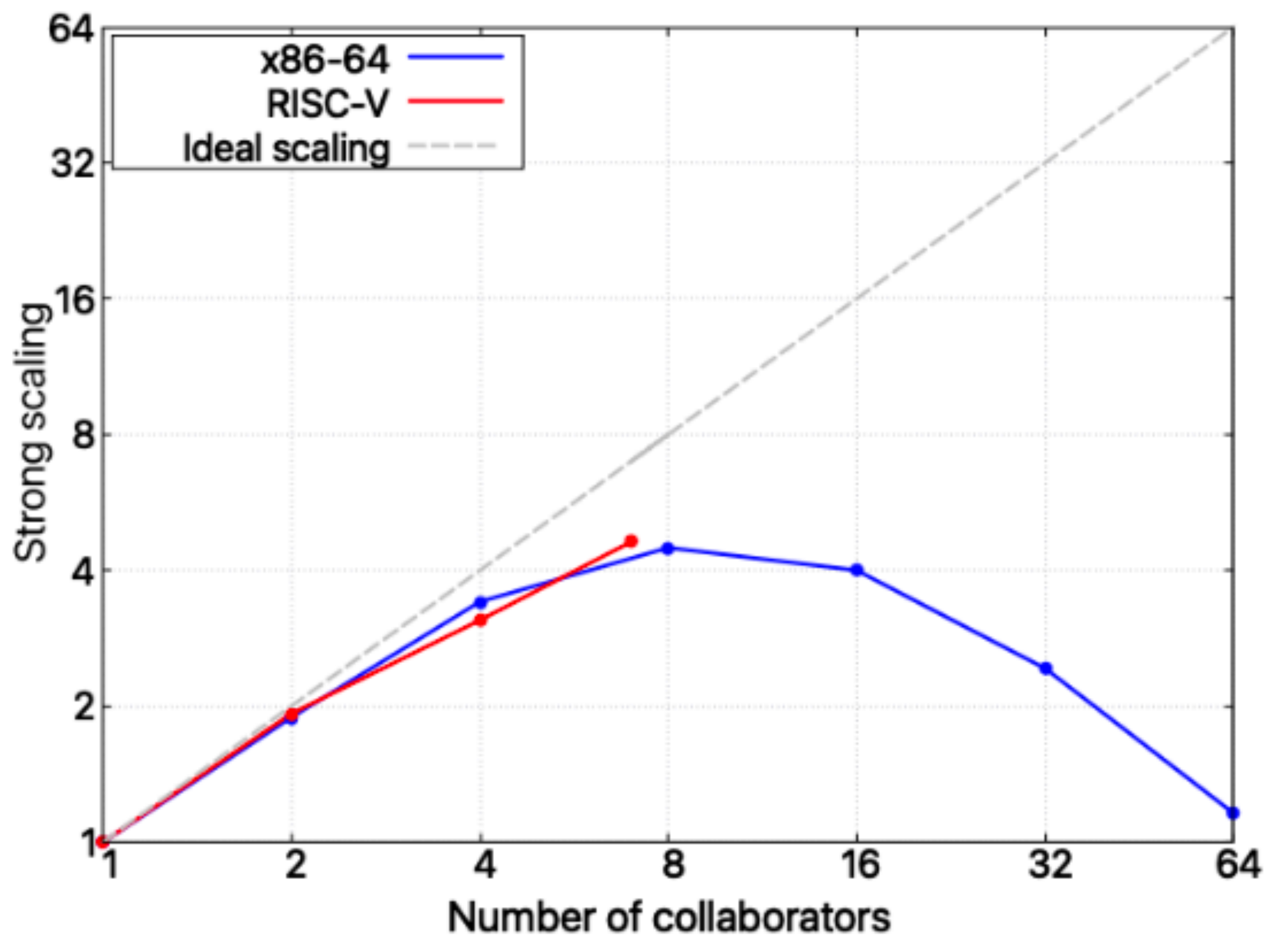
VISUALIZING SCALING PERFORMANCE



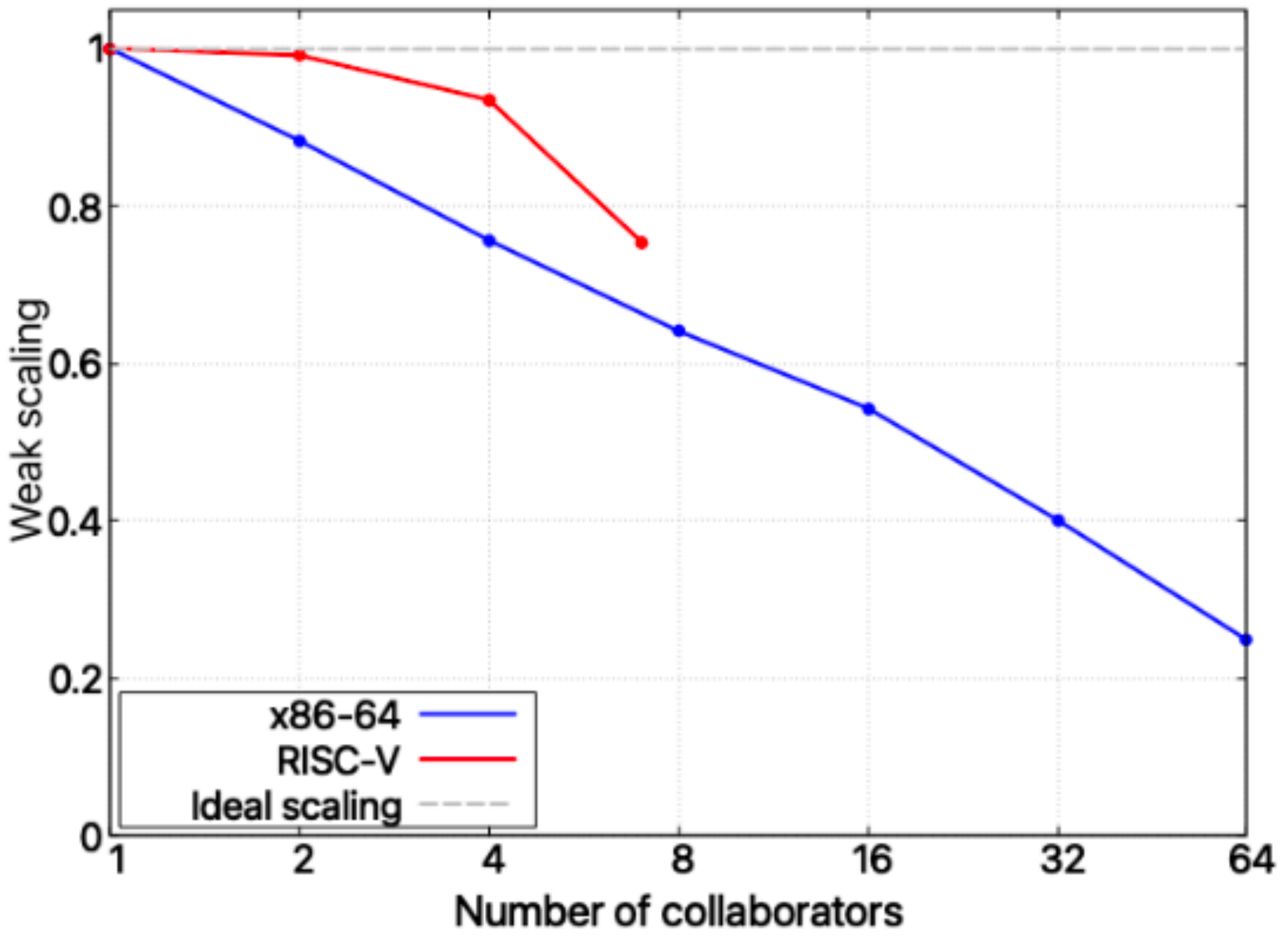


Mittone, G., Fonio, S. (2023). "Benchmarking Federated Learning Scalability".
In Proceedings of the 2nd Italian Conference on Big Data and Data Science (ITADATA 2023). CEUR.

OPENFL SCALABILITY IS STILL LACKING



(a) Strong scaling



(b) Weak scaling

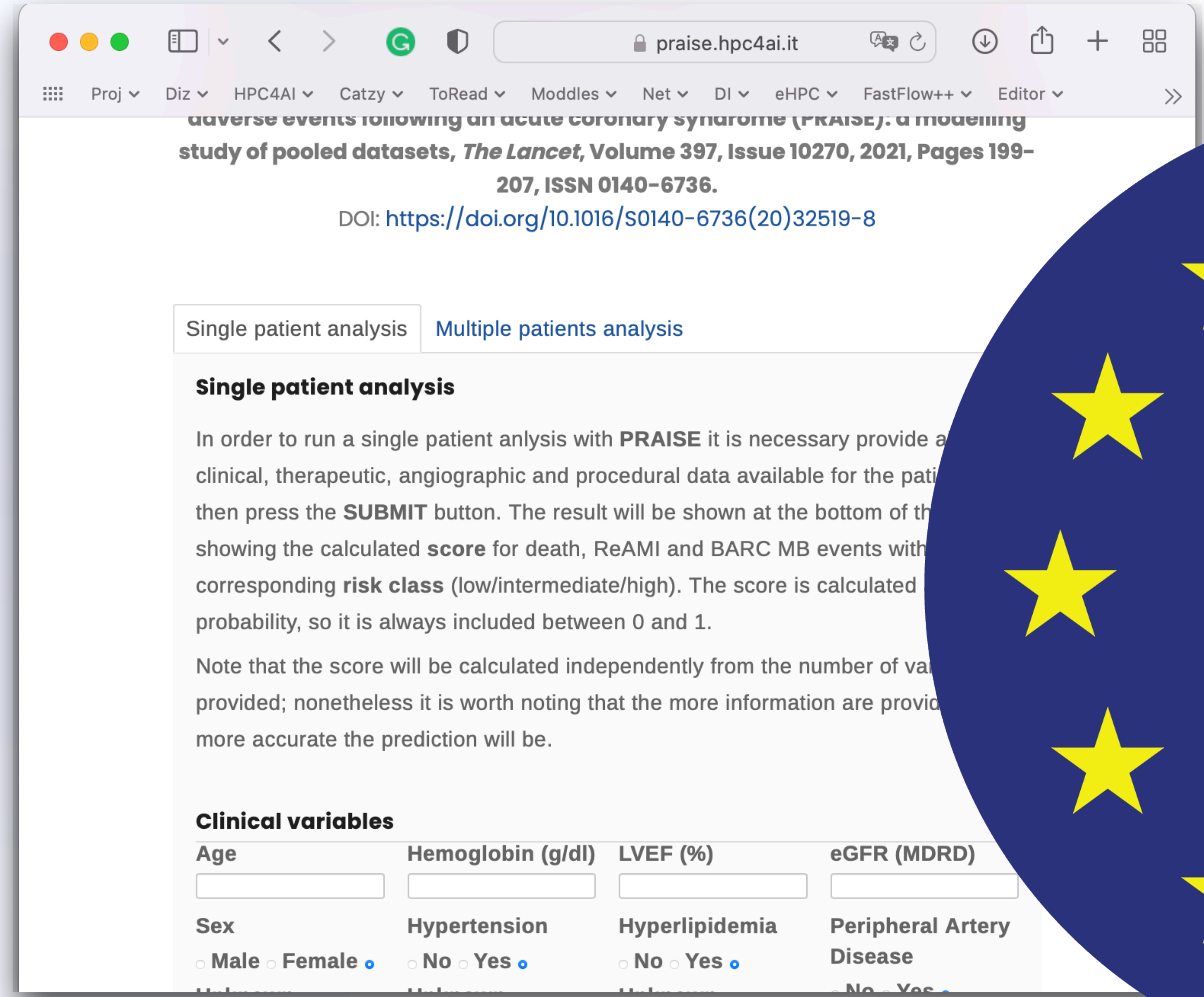
X86-64:
TWO 18-CORE INTEL® XEON E5-2697 V4 @2.30 GHZ AND 126 GB OF RAM PER NODE AND 100GB/S INTEL® OMNIPATH

RISC-V:
U740 SOC FROM SIFIVE INTEGRATING FOUR U74 RV64GCB CORES @ 1.2 GHZ, 16GB RAM AND A 1 GB/S INTERCONNECTION NETWORK

DATASET: FORESTCOVER

Mittone, G., Riviera, W., Colonnelli, I., Birke, R., Aldinucci, M. (2023). "Model-Agnostic Federated Learning". Euro-Par 2023: Parallel Processing. Euro-Par 2023. Lecture Notes in Computer Science, vol 14100. Springer.

MACHINE LEARNING... AND PRIVACY



GDPR


Articles

THE LANCET

Machine learning-based prediction of adverse events ...llowing an acute coronary syndrome (PRAISE): a modelling ...of pooled datasets

...vidio De Filippo, Guglielmo Gallone, Gianluca Mittone, Marco Agostino Deriu, Mario Iannaccone, Albert Ariza-Solé, ...gio Manzano-Fernández, Giorgio Quadri, Tim Kinnaird, Gianluca Campo, Jose Paulo Simao Henriques, James M Hughes, ...uez, Marco Aldinucci, Umberto Morbiducci, Giuseppe Patti, Sergio Raposeiras-Roubin, Emad Abu-Assi, ...n behalf of the PRAISE study group

...of current prediction tools for ischaemic and bleeding events after an acute coronary syndrome ...t for individualised patient management strategies. We developed a machine learning-based ...predict all-cause death, recurrent acute myocardial infarction, and major bleeding after ACS.

...e learning models for the prediction of 1-year post-discharge all-cause death, myocardial ...ding (defined as Bleeding Academic Research Consortium type 3 or 5) were trained on a ...ients with ACS (split into a training cohort [80%] and internal validation cohort [20%]) ...RENAMI registries, which included patients across several continents. 25 clinical features ...harge were used to inform the models. The best-performing model for each study outcome ...tested in an external validation cohort of 3444 patients with ACS pooled from a randomised ...ee prospective registries. Model performance was assessed according to a range of learning ...under the receiver operating characteristic curve (AUC).

...SE score showed an AUC of 0·82 (95% CI 0·78–0·85) in the internal validation cohort and ...in the external validation cohort for 1-year all-cause death; an AUC of 0·74 (0·70–0·78) in the ...on cohort and 0·81 (0·76–0·85) in the external validation cohort for 1-year myocardial infarction; and ...70 (0·66–0·75) in the internal validation cohort and 0·86 (0·82–0·89) in the external validation cohort ...major bleeding.

...erpretation A machine learning-based approach for the identification of predictors of events after an ACS is feasible and effective. The PRAISE score showed accurate discriminative capabilities for the prediction of all-cause death, myocardial infarction, and major bleeding, and might be useful to guide clinical decision making.

Funding None.

Copyright © 2021 Elsevier Ltd. All rights reserved.

Introduction

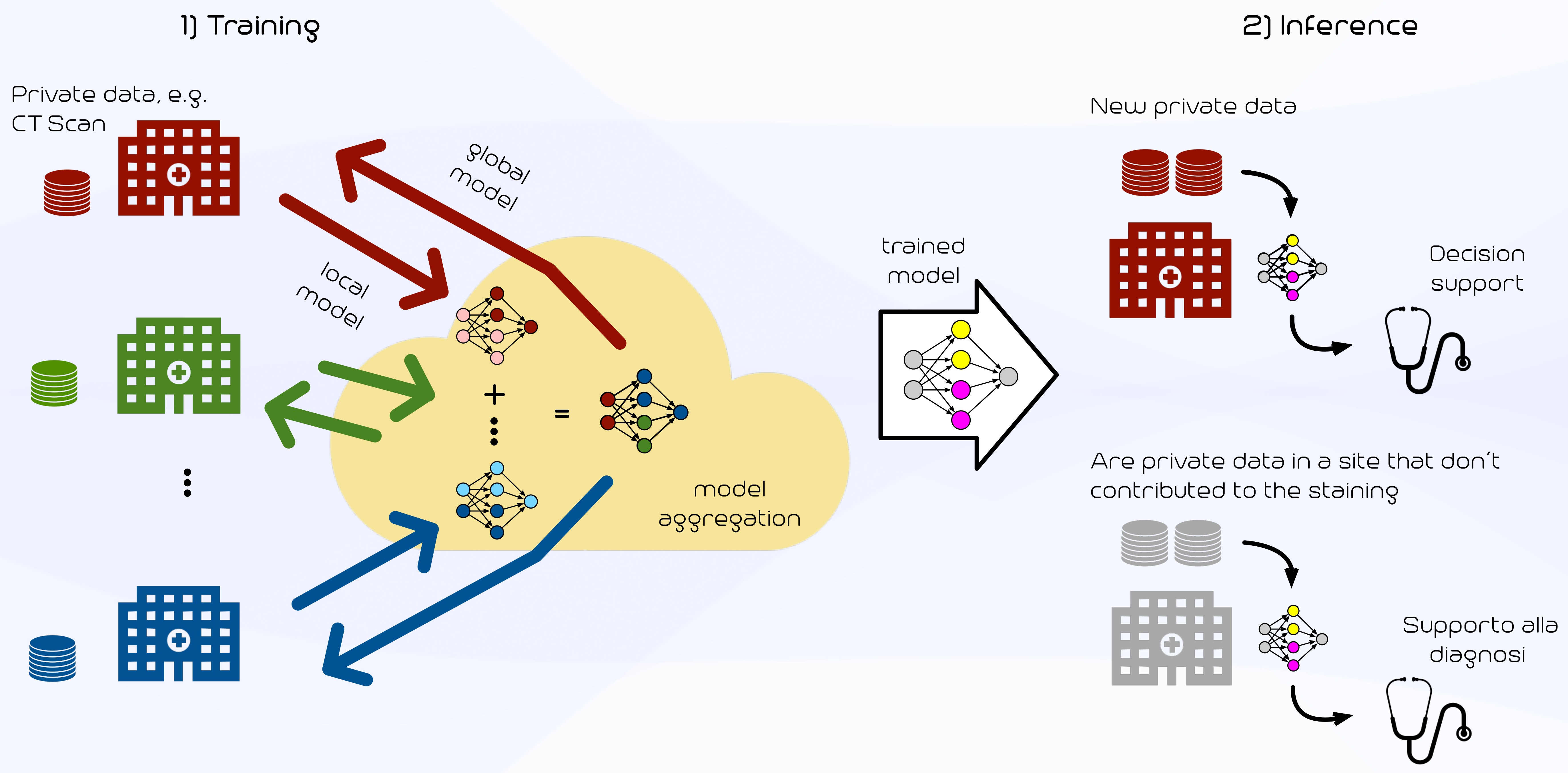
Patients with acute coronary syndrome (ACS) are at high risk for ischaemic and bleeding events, with both being drivers of adverse prognosis.[1] Careful evaluation of these risks plays a fundamental role in the clinical management of each patient, with important implications regarding the choice of optimal medical therapy for secondary prevention.[2–6]

To this aim, several predictive tools have been developed to estimate ischaemic and bleeding risks following an related to their derivation from unselected percutaneous coronary intervention populations encompassing patients with stable presentation. Moreover, machine learning methods might be able to overcome some of the limitations of current analytical approaches to risk prediction by applying computer algorithms to large datasets with numerous, multidimensional variables, capturing high-dimensional, non-linear relationships among clinical features to make data-driven outcome predictions.[14] The effectiveness of this approach has been shown in several

Lancet 2021; 397: 199–207
See Comment page 172
Division of Cardiology, Cardiovascular and Thoracic Department, Città della Salute e della Scienza, Turin, Italy (F D'Ascenzo MD, O De Filippo MD, G Gallone MD, Prof G M De Ferrari MD); Cardiology, Department of Medical Sciences (F D'Ascenzo, O De Filippo, G Gallone, Prof G M De Ferrari) and Department of Computer Science (G Mittone MSc, Prof M Aldinucci PhD), University of Turin, Turin, Italy; Department of Cardiology, University Hospital Álvaro Cunqueiro, Vigo, Spain (S Raposeiras-Roubin MD, E Abu-Assi MD); Cardiology Department, University Hospital of Wales, Cardiff, UK (T Kinnaird MD); Department of Cardiology, University Hospital de Bellvitge, Barcelona, Spain (A Ariza-Solé MD); Kerckhoff Heart and Thorax Center, Frankfurt, Germany (Prof C Liebetrau MD); Department of Cardiology, University Hospital Virgen Arrtixaca, Murcia, Spain (S Manzano-Fernández MD); Department of Cardiology, S G Bosco Hospital, Turin, Italy (M Iannaccone MD); University of Amsterdam, Academic Medical Center, Amsterdam, Netherlands (J P Simao Henriques MD); Catheterization Laboratory, Maggiore della Carità Hospital, Novara, Italy (Prof G Patti MD);

D'Ascenzo, F., De Filippo, O., Gallone, G., Mittone, G., et al. (2021). "Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets". The Lancet, 397(10270), 199-207.

VISUALIZING FEDERATED LEARNING



Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
- initialize w_0
- **for** each round $t = 1, 2, \dots$ **do**
  - $m \leftarrow \max(C \cdot K, 1)$
  - $S_t \leftarrow$ (random set of m clients)
  - **for** each client $k \in S_t$ **in parallel do**
    - $w^k_{t+1} \leftarrow \text{ClientUpdate}(k, w_t)$
  - $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w^k_{t+1}$

ClientUpdate(k, w): *// Run on client k*
- $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
- **for** each local epoch i from 1 to E **do**
  - **for** batch $b \in \mathcal{B}$ **do**
    - $w \leftarrow w - \eta \nabla \ell(w; b)$
- return w to server

Cross-device: many unreliable clients (mobile or IoT devices)

Cross-silo: a few reliable clients (companies and/or data centers)

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). "Communication-efficient learning of deep networks from decentralized data". In Artificial intelligence and statistics (pp. 1273-1282). PMLR.

HPC4AI - EPITO

RISC-V: The Free and Open RISC Instruction Set Architecture

- 2 **RISC-V** (RV64) compute cluster - U740 SoC from SiFive with four U74 RV64GCB application cores, 1.2 GHz and 16GB of DDR4, 1 TB node-local NVME storage
- 4 Intel 2 sockets Xeon Gold 6230 CPU (40-threads@2.10GHz), 1536GB RAM, and 2 x NVidia V100 GPU
- 4 NVidia/ARM-dev kits, each including 1 socket Ampere-Altra Q80-30 (80-core@3GHz), 512GB RAM, 2 x NVidia BlueField-2 DPU, and 2 x NVidia A100 GPU.

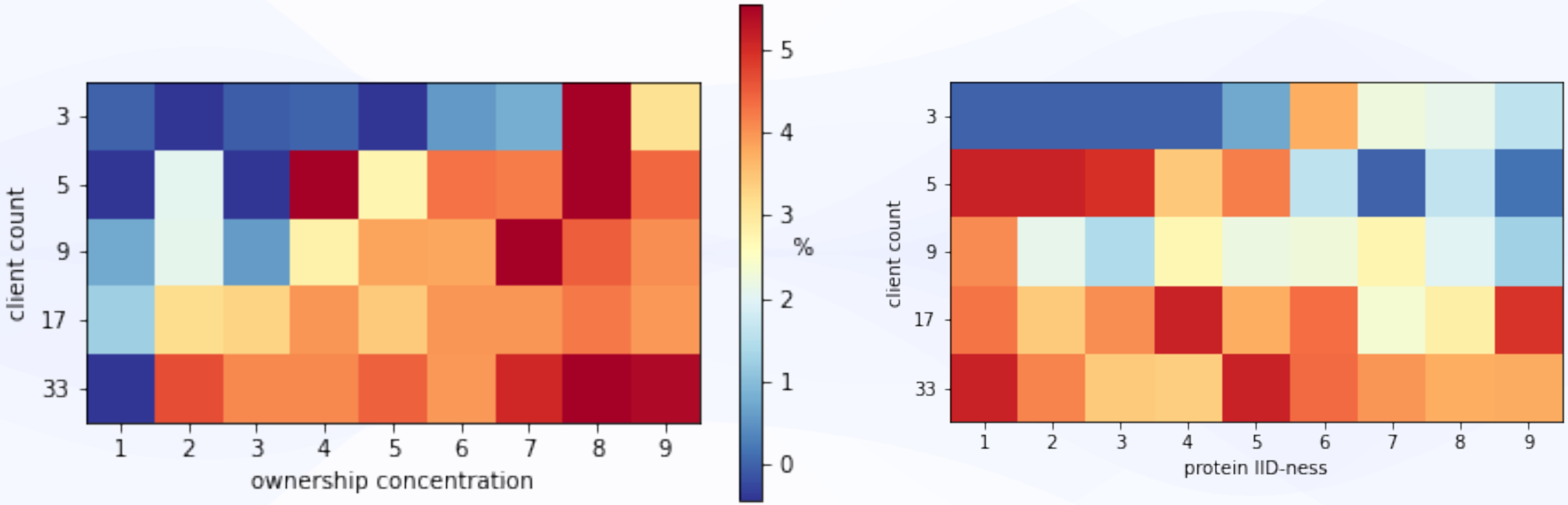
MONTE CIMONE

RISC-V

RISC-V: The Free and Open RISC Instruction Set Architecture

- 8-node **RISC-V** 4-core@1.2GHz (U740 Sifive SoC) HPC compute cluster integrating processors, main memory, non-volatile storage, and interconnect.

LEARNING DRUG-TARGET INTERACTION



Unbalance in data quantity distribution

Unbalance in protein information distribution

Svoboda, F., <u>Mittone, G.</u>, Lane, N. D., Lio', P.. "A Federated Learning Benchmark for Drug-Target Interaction"
Accepted at the "Machine Learning in Structural Biology" workshop, NeuriIPS 2022

TRADITIONAL NON-SGD ML APPROACHES

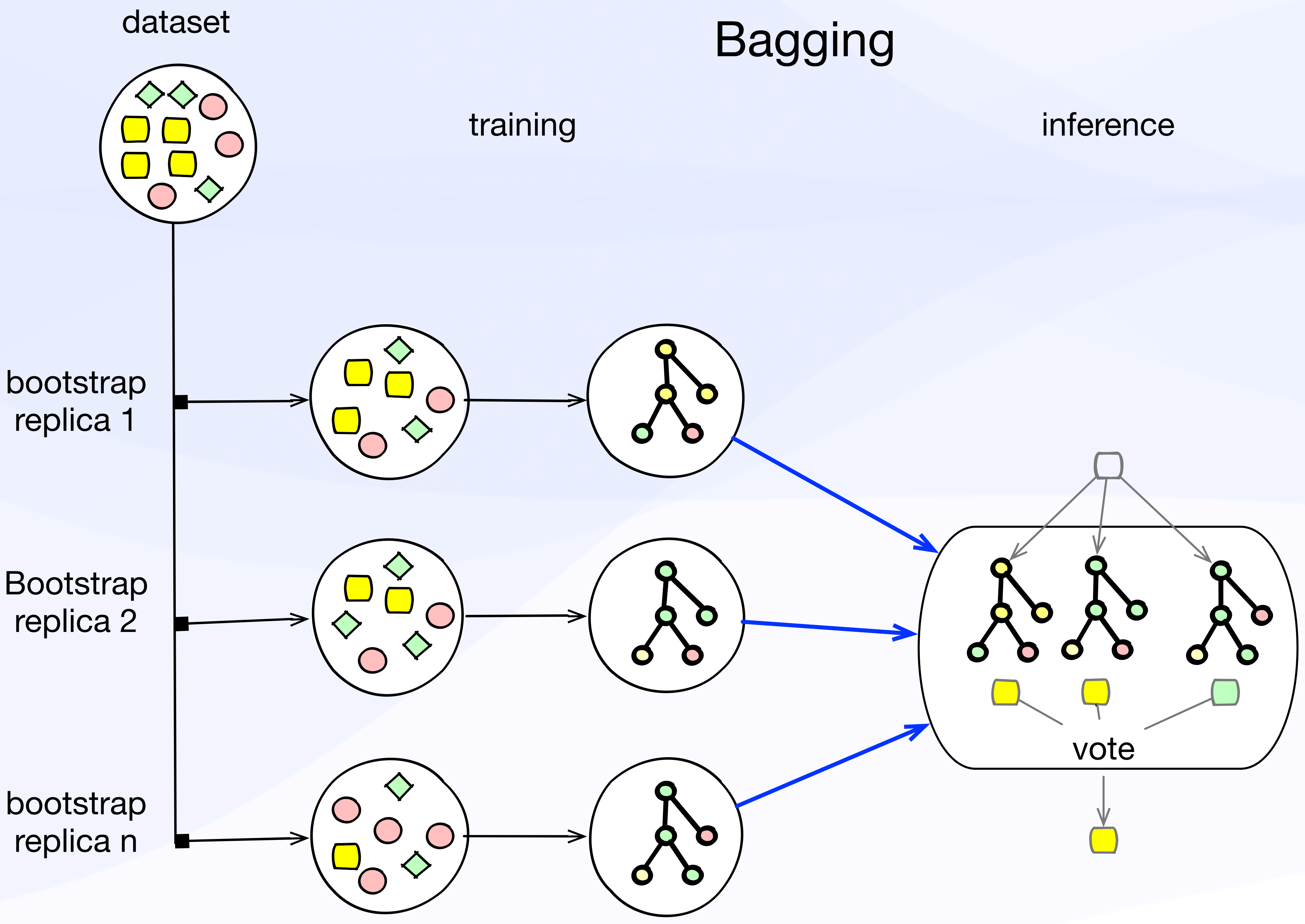
Parallel

Training

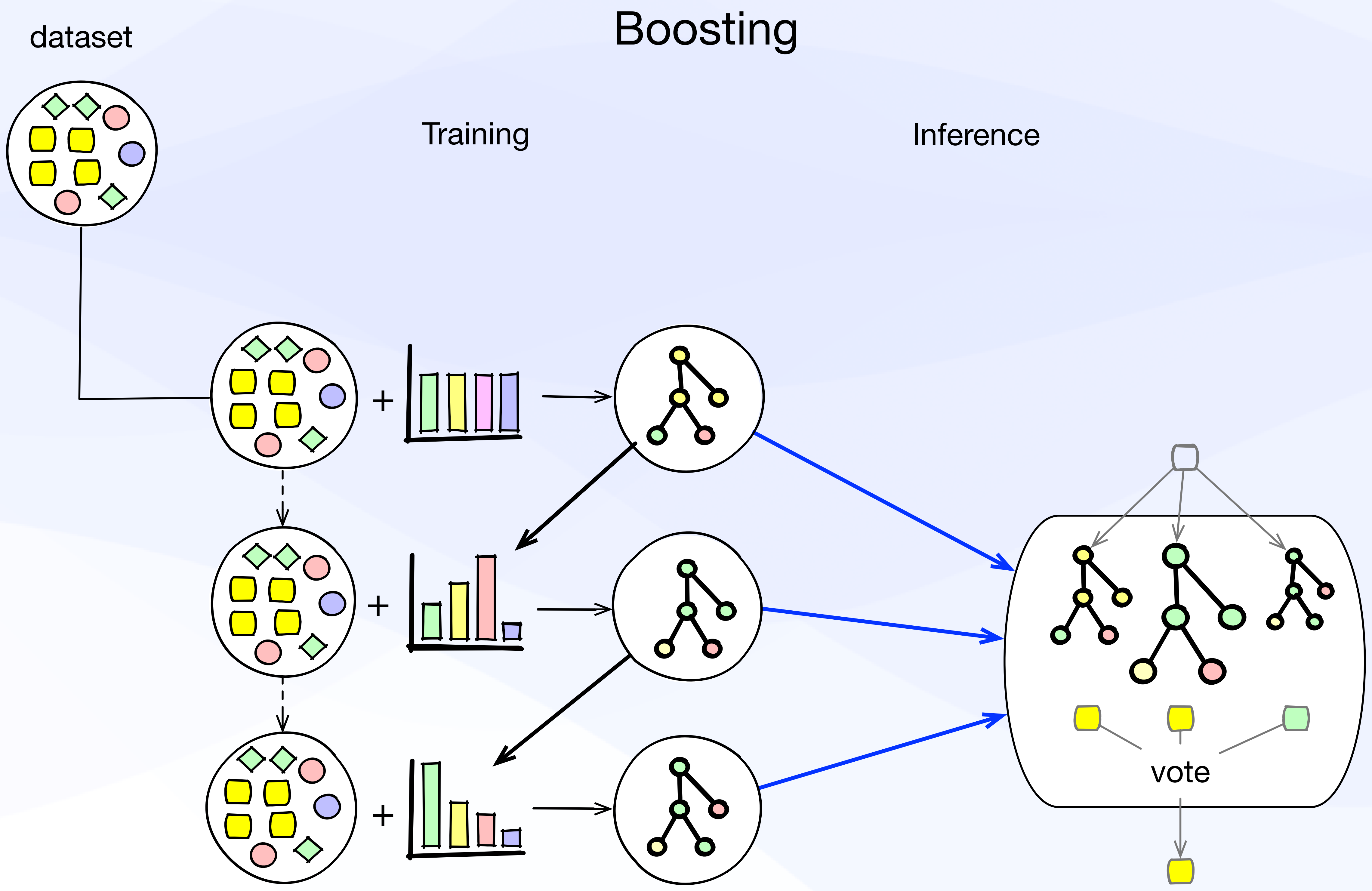
Pros: embarrassingly parallel training

Inference

Cons: horrible accuracy for non iid distributions of examples among silos



Sequential



… AND COMPARABLE TO DNNS

TABLE I: Prediction performance of the FNN model. Values reported are the average ± stdev of 5 runs. The first run in the strong scaling setting is equivalent to the non-federated case.

Clients	Accuracy	F1 Score	F2 Score	Precision	Recall
Strong scaling setting					
1	.39 ± .47	.14 ± .08	.22 ± .04	.17 ± .09	.72 ± .39
2	.56 ± .47	.19 ± .09	.26 ± .06	.15 ± .09	.61 ± .36
4	.88 ± .01	.23 ± .01	.30 ± .01	.17 ± .01	.39 ± .02
8	.72 ± .38	.20 ± .06	.27 ± .04	.16 ± .06	.48 ± .29
16	.90 ± .01	.24 ± .01	.29 ± .01	.12 ± .01	.35 ± .02
Weak scaling setting					
1	.56 ± .47	.16 ± .07	.22 ± .03	.12 ± .07	.56 ± .40
2	.69 ± .37	.17 ± .05	.25 ± .04	.12 ± .06	.49 ± .30
4	.72 ± .38	.20 ± .07	.27 ± .05	.15 ± .06	.49 ± .29
8	.90 ± .04	.18 ± .10	.24 ± .13	.13 ± .08	.30 ± .17
16	.55 ± .46	.17 ± .08	.26 ± .06	.11 ± .06	.63 ± .34

TABLE II: Prediction performance of `AdaBoost.F`. Values reported are the average ± stdev of 5 runs. The first run in the strong scaling setting is equivalent to the non-federated case.

Clients	Accuracy	F1 Score	F2 Score	Precision	Recall
Strong scaling setting					
1	.95 ± .00	.19 ± .07	.15 ± .06	.35 ± .10	.13 ± .05
2	.95 ± .00	.23 ± .03	.19 ± .03	.36 ± .04	.17 ± .03
4	.94 ± .00	.19 ± .02	.16 ± .02	.26 ± .04	.15 ± .02
8	.94 ± .00	.20 ± .04	.17 ± .03	.28 ± .06	.16 ± .03
16	.94 ± .00	.19 ± .03	.17 ± .03	.25 ± .04	.16 ± .03
Weak scaling setting					
1	.95 ± .00	.09 ± .02	.06 ± .01	.33 ± .05	.05 ± .01
2	.95 ± .00	.10 ± .02	.07 ± .01	.45 ± .05	.05 ± .01
4	.95 ± .00	.15 ± .04	.12 ± .04	.32 ± .06	.10 ± .10
8	.95 ± .00	.17 ± .02	.14 ± .01	.28 ± .04	.13 ± .01
16	.94 ± .00	.20 ± .03	.18 ± .02	.27 ± .04	.16 ± .02

olonnelli, I., D'Ascenzo, F., et al. "Pooling critical datasets with Federated Learning". Accepted at PDP 2023.

OVERVIEW OF THE OBTAINED RESULTS

	master + 2 workers		master + 4 workers		master + 7 workers	
	Time (s)	Energy/worker (J): Δ (Tot)	Time (s)	Energy/worker (J): Δ (Tot)	Time (s)	Energy/worker (J): Δ (Tot)
Intel	23.84	973 (1992)	**23.56**	1011 (2069)	**24.38**	1049 (2146)
ARM	**23.33**	**133** (**483**)	25.66	**146** (**531**)	25.86	**148** (**535**)
RISC-V	674.47	269 (2562)	673.70	269 (2560)	687.03	274 (2610)
Intel-ARM	29.50	NA	29.55	NA	33.34	NA

(a) **MNIST Master-Worker training results:** These performance metrics have been taken on a set of 20 federation rounds made up of 5 training epochs each (total 100 epochs); each client was assigned 1/8 of the entire dataset.

	2 peers		4 peers		8 peers	
	Time (s)	Energy/peer (J): Δ (Tot)	Time (s)	Energy/peer (J): Δ (Tot)	Time (s)	Energy/peer (J): Δ (Tot)
Intel	**23.15**	2082 (4261)	**24.05**	2162 (4422)	24,95	2210 (4522)
ARM	24.39	**169** (**535**)	24.90	**173** (**546**)	**26.65**	**185** (**585**)
RISC-V	819.35	409 (3195)	815.55	407 (3180)	933.62	466 (3641)
Intel-ARM	45.20	NA	39.13	NA	50.88	NA

(b) **MNIST Peer-to-Peer training results:** These performance metrics have been taken on a set of 20 federation rounds made up of 5 training epochs each (total 100 epochs); each client was assigned 1/8 of the entire dataset.

	root + 2 leaves		root + 4 leaves		root + 7 leaves	
	Time (s)	Energy/leaf (J): Δ (Tot)	Time (s)	Energy/leaf (J): Δ (Tot)	Time (s)	Energy/leaf (J): Δ (Tot)
Intel	**19,76**	1520 (2389)	**19,38**	1491 (2343)	**19,01**	1462 (2298)
ARM	37.16	**291** (**848**)	39.88	**312** (**910**)	43.15	**338** (**985**)
RISC-V	1201.51	841 (4926)	1205.77	844 (4943)	1212.77	848 (4972)
Intel-ARM	35.65	NA	35.65	NA	36.10	NA

(c) **YOLO Tree-based inference results:** These performance metrics have been obtained by assigning each leaf a video with 148 frames.

enting with Emerging ARM and RISC-V Systems for Decentralised Machine Learning",

Submitted at ACM CF 2023.

POWER CONSUMPTION ANALYSIS

INTEL VS ARM VS RISC-V

	Energy/FLOP (CPU only)	Avg CPU power (idle)	TDP (per socket)	Avg system power (idle)
Intel	5 nJ	44 W	125 W	190 W
ARM	1 nJ	15 W	250 W	290 W
RISC-V	12 nJ	3.4 W	5 W	5 W

TABLE III: **Comparison of the different systems for power employed by CPU and overall systems.** The Intel system has two sockets, whereas the others have one socket.

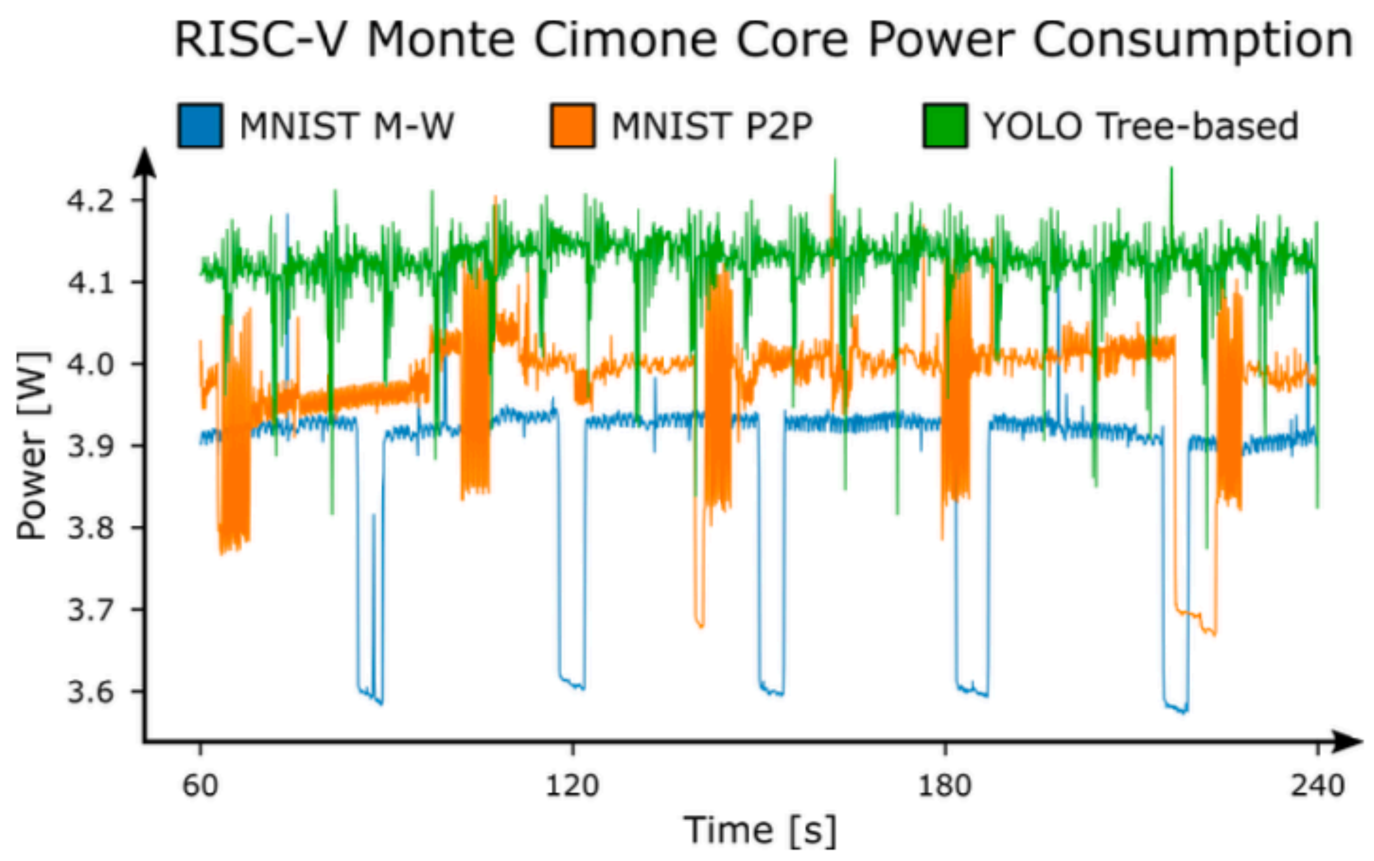


Fig. 3: **RISC-V power consumption traces** Three minutes power consumption on the Monte Cimone RISC-V system.

100kHz measurement

enting with Emerging ARM and RISC-V Systems for Decentralised Machine Learning", Submitted at ACM CF 2023.