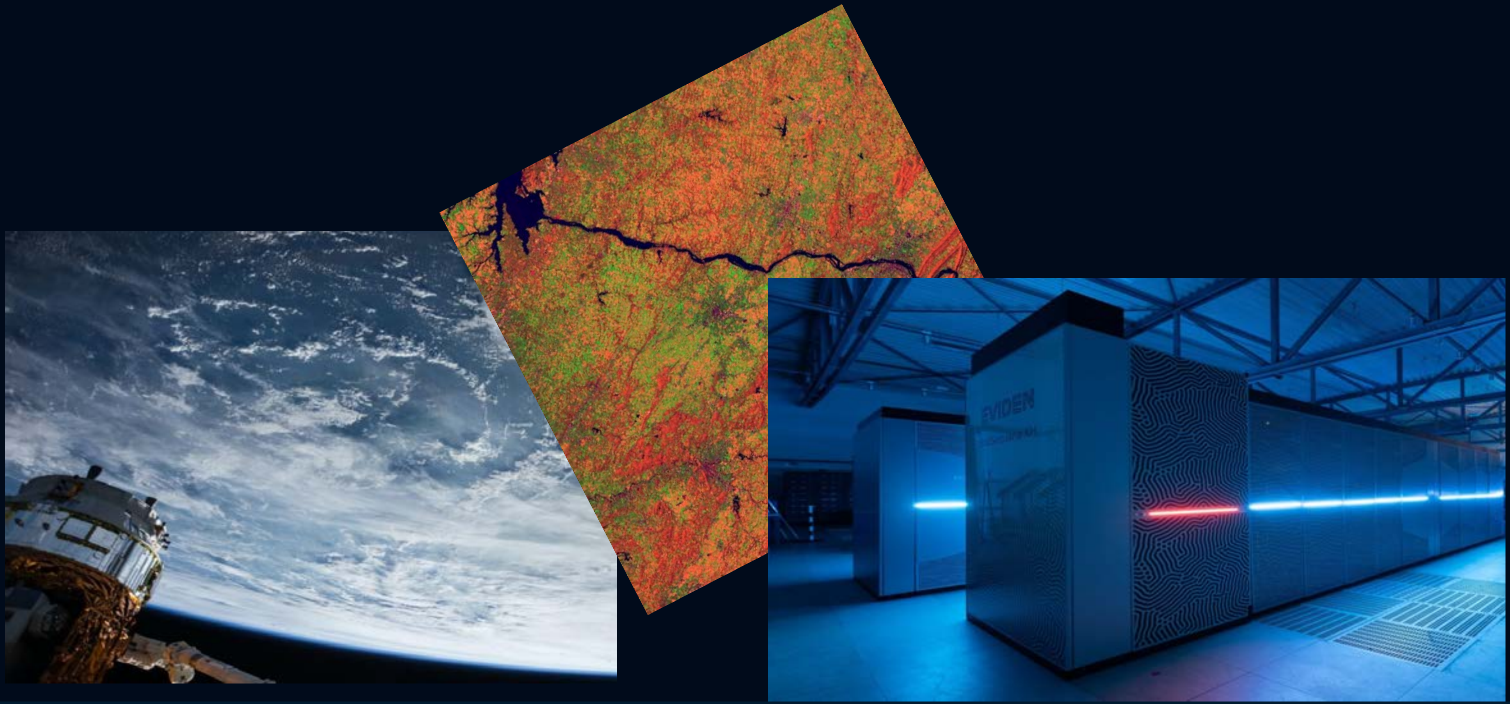


INTRODUCTION TO LARGE-SCALE AI WITH SUPERCOMPUTING

What's next for Infrastructure? | ESA Living Planet Symposium | 22 June 2025 | Vienna, Austria

GABRIELE CAVALLARO (WWW.GABRIELE-CAVALLARO.COM)
HEAD OF SIMULATION AND DATA LAB “AI AND ML FOR REMOTE SENSING”, JÜLICH SUPERCOMPUTING CENTRE (FORSCHUNGSZENTRUM JÜLICH)
ASSOCIATE PROFESSOR, FACULTY OF ELECTRICAL AND COMPUTER ENGINEERING (UNIVERSITY OF ICELAND)

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

UNIVERSITY OF ICELAND

OUTLINE

1. AI drives Computational Demands
2. Foundation Models
3. Supercomputing
4. Distributed Deep Learning

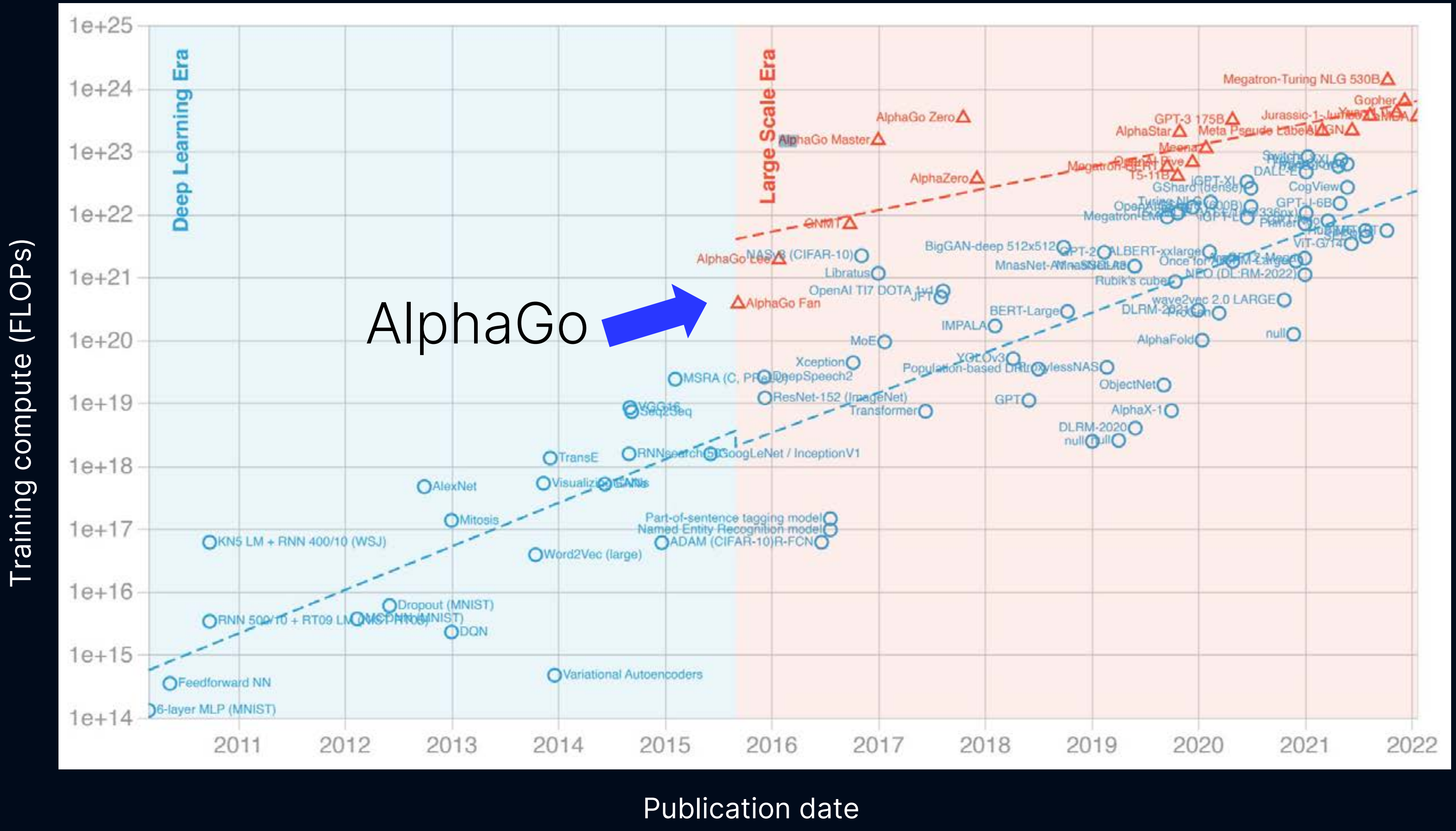
1

AI* DRIVES COMPUTATION DEMANDS

HPC and AI application growth in the past decade

*Large-Scale Deep Learning

THE ERA OF LARGE-SCALE DEEP LEARNING

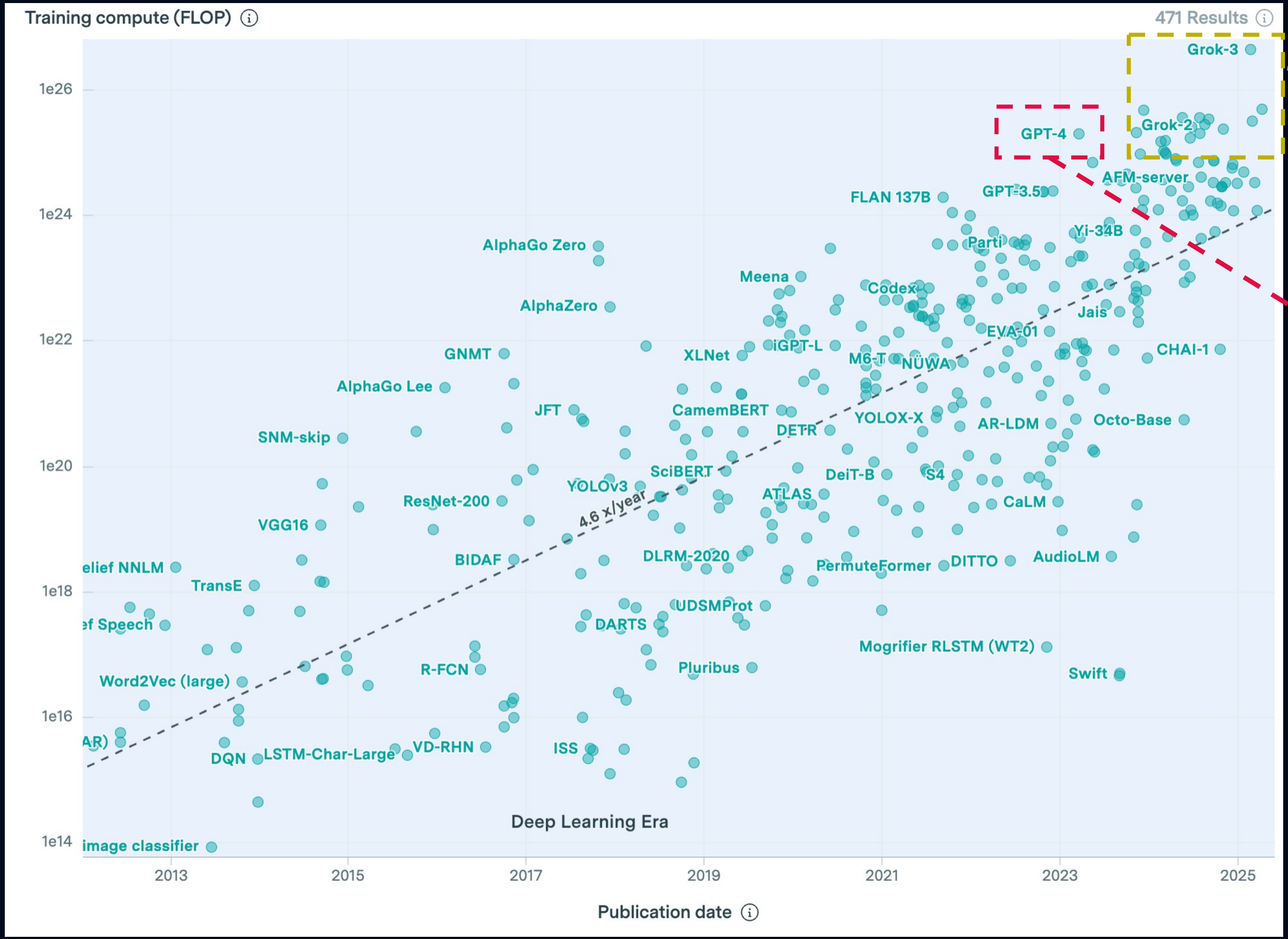


- 2015 marked the start of a new era in large-scale AI models
- Models like AlphaGo had much higher computational power than other models of that time
- OpenAI noted that compute capacity has been doubling approximately every 3.4 months

These advancements continue to push the demand for computing power

J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn and P. Villalobos, "Compute Trends Across Three Eras of Machine Learning," 2022 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2022, https://doi.org/10.1109/IJCNN55064.2022.9891914

NOTABLE AI MODELS



- 2025: Grok-x (speculation)
  - $20K GPU X 200,000 = $4B
  - 92 days of training
- 2022: GPT-4 (speculation)
  - $8K GPU × 25,000 = $200M
  - 90 days of training
- 2012: AlexNet
  - $500 GPU x 2 = $1K
  - 5 days of training

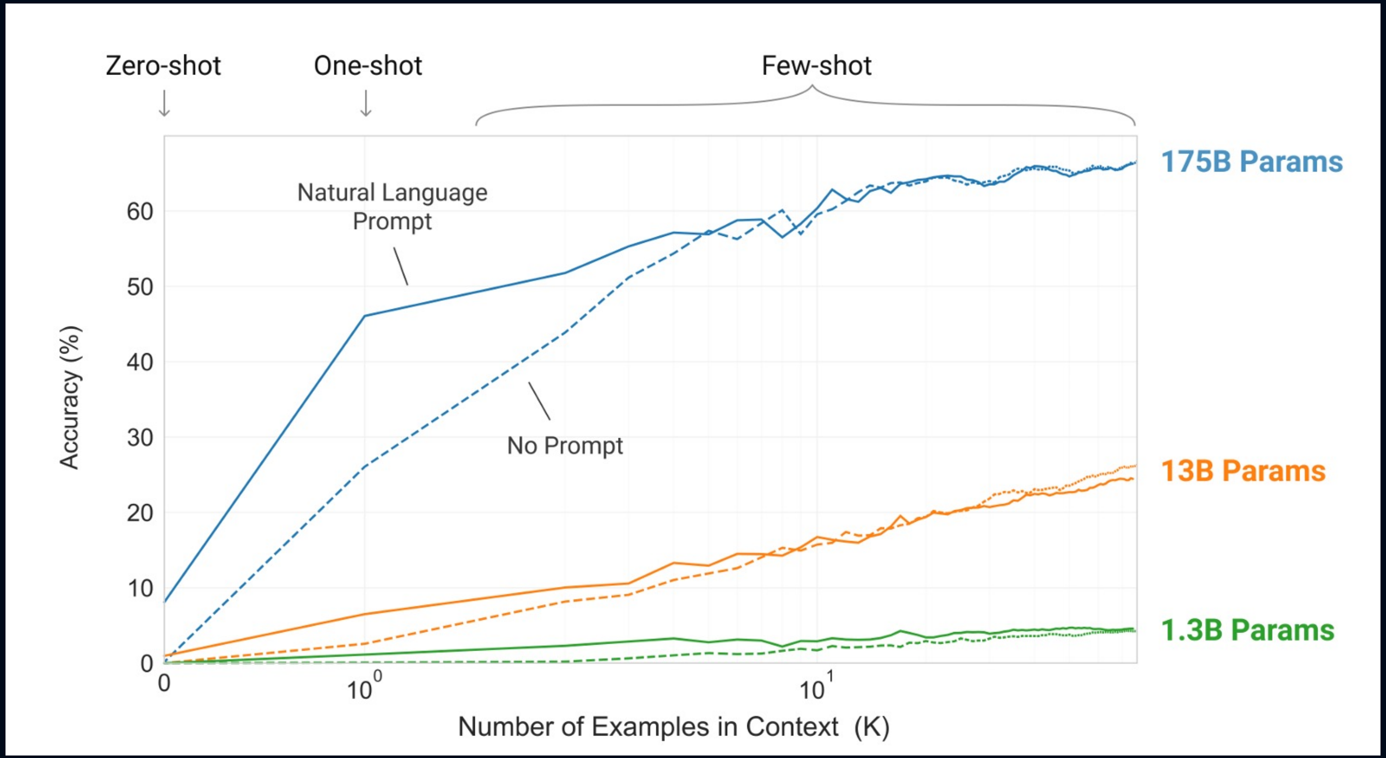
Rio Yokota (Institute of Science Tokyo), Scaling Laws in HPC and AI: Yesterday, Today and Tomorrow, Salishan Conference on High Speed Computing, 2025, https://salishan.ahsc-nm.org/program.html

EPOCH AI, "Notable AI Models", https://epoch.ai/data/notable-ai-models

SCALING UP MODELS

Driven by empirical observations and supported by theory

IN 2020, IT WAS SHOWN THAT SCALING MODELS COULD IMPROVE TASK-AGNOSTIC, FEW-SHOT PERFORMANCE



Task: give the model a word distorted by some combination of scrambling, addition, or deletion of characters, and askto recover the original word.

- Small models (=> 1 billion param.) couldn't understand tasks from prompts alone
- Large models demonstrate improved ability to learn a task from contextual information
  - Thanks in part to using more data during pretraining

Model Name	n_{params}
GPT-3 Small	125M
GPT-3 Medium	350M
GPT-3 Large	760M
GPT-3 XL	1.3B
GPT-3 2.7B	2.7B
GPT-3 6.7B	6.7B
GPT-3 13B	13.0B
GPT-3 175B or "GPT-3"	175.0B

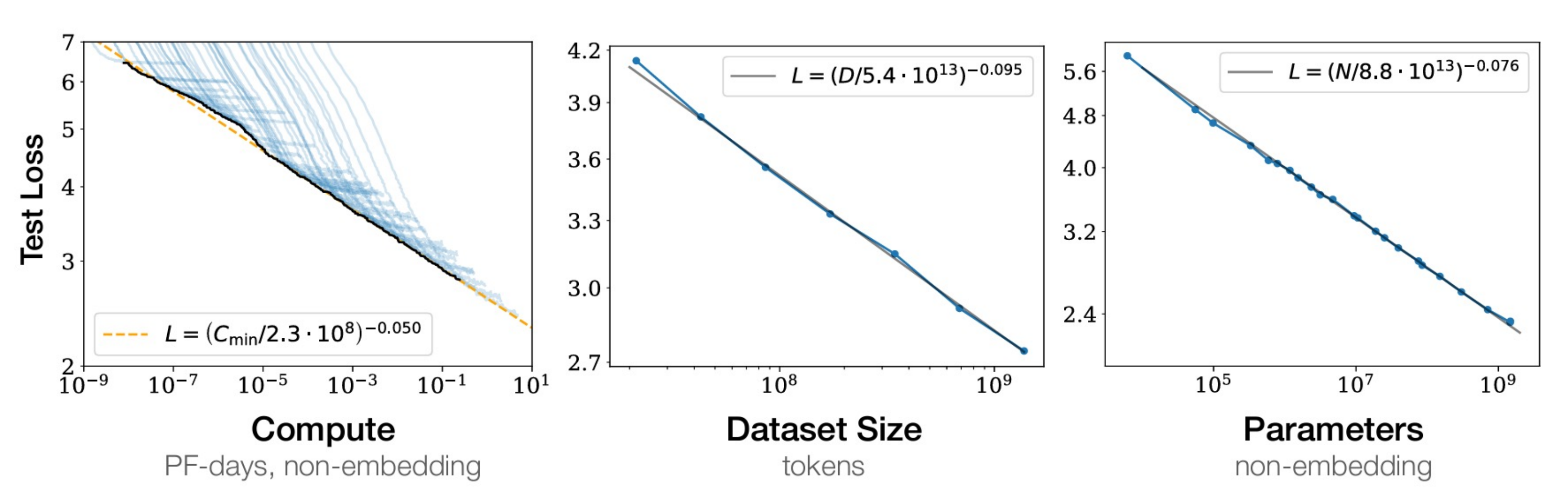
T. B. Brown, B. Mann, N. Ryder, "Language Models are Few-Shot Learners", 2020, https://doi.org/10.48550/arXiv.2005.14165

SCALING LAWS

- Refers to relations between functional properties of interest (i.e., test loss, performance metric for fine-tuning tasks) and properties of the architecture or optimization process (e.g., model size, data, or training compute)
- These laws can help inform the design and training of DL models, as well as provide insights into their underlying principles

SCALING LAWS

- **Predictable power laws**
  - Test loss drops as a power law with parameters, data, and compute
- **Quantified returns**
  - For text models, doubling parameters cuts loss by ≈ 5 %; doubling data by ≈ 7 – 10 %
  - Vision models gain even more per doubling.
- **Scalability across regimes**
  - The power-law exponents stay constant across ~7 orders of magnitude (results from small runs extrapolate accurately to trillion-parameter scale)



J. Kaplan, S. McCandlish, T. Henighan et al., "Scaling Laws for Neural Language Models," arXiv preprint, 2020, https://doi.org/10.48550/arXiv.2001.08361

2

FOUNDATION MODELS

WHAT MAKES A MODEL FOUNDATIONAL?

Train the model once (slowly) and use it many times ("quickly")

Generic scalable pre-training

Uses general data types:
text, images, or combined text-images

Simple loss functions
(e.g., predicting the next token
in autoregressive models)

Self-supervised training:
scalable without human-labeled data

Data-efficient transferability

Farm in California

Farm in China

After pre-training, models can handle a wide range of different tasks, even with minimal data (zero-shot or few-shot)

Scaling Laws

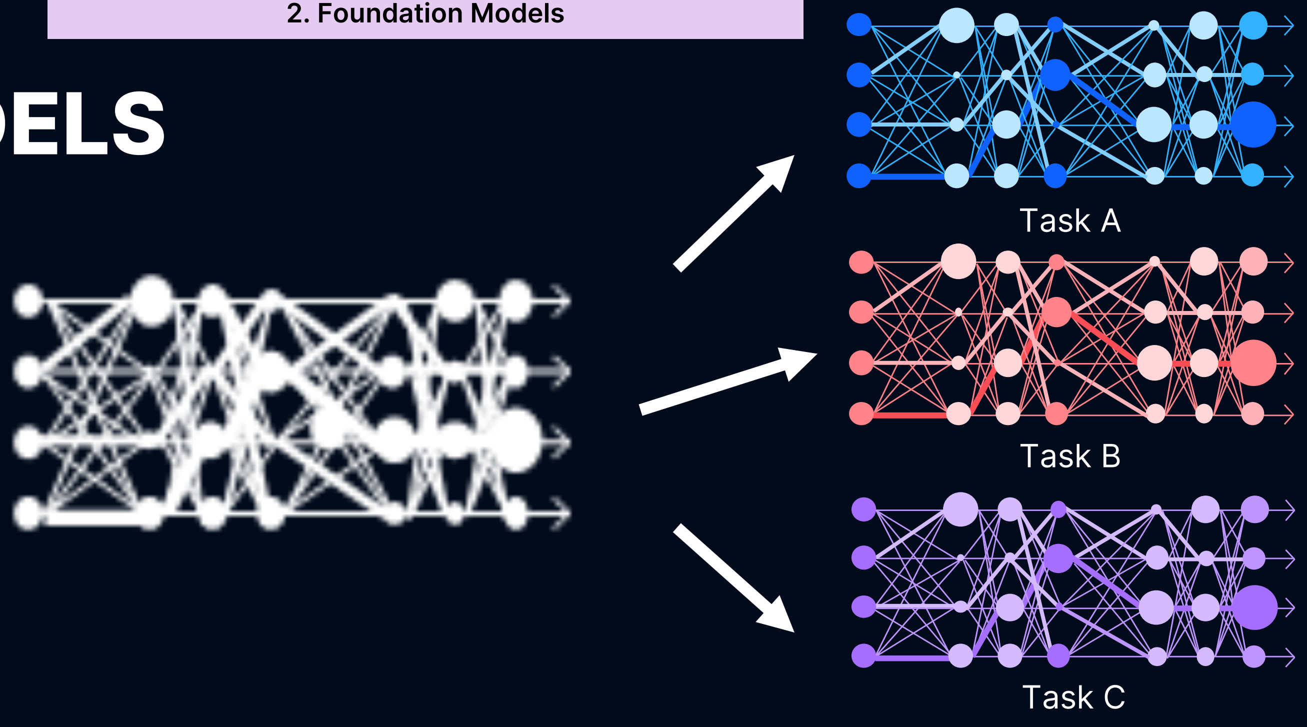
Performance improves as model size, data, and compute scale up. Larger scales lead to better transferability

Emergent abilities, like in-context learning and reasoning, become noticeable at larger scales

Adapted from Jenia Jitsev, Open foundation models: reproducible science of transferable learning, AUTOML23, https://youtu.be/R6olskkLmjA?si=6d42xDam-TreXyoh

FOUNDATION MODELS

"One model for all"
a universal model designed to
work across various tasks



Key Points

Focus on both new model ideas and top-quality data.

Pre-trained models can be used for many tasks while saving energy.

Models learn patterns from a lot of data without needing labels.

Can be quickly customized for specific tasks, even with limited labeled data.

Applications

Language: Writing and understanding text (e.g., GPT, BERT, T5)

Vision: Recognizing images and extracting features (e.g., ResNet, ViT, DINO)

Combined Language and Vision: Image interpretation and generation (e.g., CLIP, Stable Diffusion).

Key Factors for Success

Training with large amounts of varied data (e.g., over 100 million examples).

Building large-scale models with billions of parameters.

Using massive training samples for long-term learning (e.g., over 300 billion tokens).

Adapted from Jenia Jitsev, Open foundation models: reproducible science of transferable learning, AUTOML23, https://youtu.be/R6olskkLmjA?si=6d42xDam-TreXyoh

GEOSPATIAL AI FOUNDATION MODELS FOR EO

2021

- Beginning of foundation Models developed for EO and Earth science applications

Lacoste, Alexandre et al.. "Toward Foundation Models for Earth Monitoring: Proposal for a Climate Change Benchmark"NeurIPS /CCAI workshop 2021

2022

- Language model based on Earth science literature

Ramachandran, Rahul et al. "Language model for Earth science: Exploring potential downstream Applications as well as current challenges". IGARSS 2022.

Koirala, Prasanna. "Transforming Language Understanding in the Earth Sciences". Medium 2022. (BERT-E)

2023

- SSL4EO-L, a large model trained on LandSat imagery (Microsoft TorchGeo team)

Stewart, Adam et al. "SSL4EO-L: Datasets and Foundation Models for Landsat Imagery". Arxiv 2023.

- ClimaX, foundation model designed for weather and climate science ClimaX

Nguyen, Tung et al. "ClimaX: A foundation model for weather and climate". Arxiv 2023.

- ClimSim, dataset for AI emulators of atmospheric storms, clouds, turbulence, rainfall, radiation

Yu, Sungduk et al. "ClimSim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators". Arxiv 2023.

- PRESTO, pretrained transformer focussing on time-series

Tseng, Gabriel et al. "Lightweight, Pre-trained Transformers for Remote Sensing Timeseries" (PRESTO). Arxiv 2023.

- Prithvi, generalist geospatial AI foundation model (image classification, object detection)

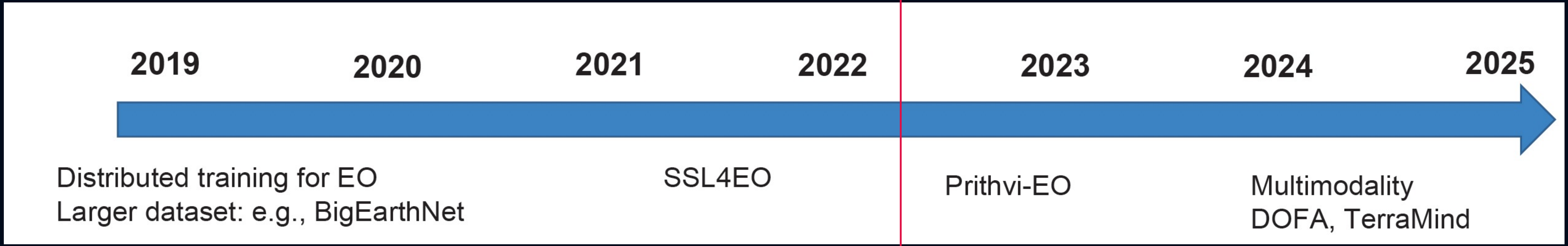
Jakubik, J., Roy, S., Phillips, C. E., et al. "Foundation Models for Generalist Geospatial Artificial Intelligence." Preprint Available on arXiv:2310.18660, 2023.

... and EarthPT, SkySense, VITO-DINO, EarthGPT, SkySenseGPT, Clay, Hypersigma, CORSA, ...

GEOSPATIAL FOUNDATION MODELS FOR EO

Beginning of FMs in EO

Benchmarking: GEO-Bench, PANGAEA



A. Lacoste, N. Lehmann, P. Rodriguez,"GEO-Bench: Toward Foundation Models for Earth Monitoring", 2023, https://doi.org/10.48550/arXiv.2306.03831 Focus to learn more

V. Marsocci, Y. Jia, G. Le Bellier, et al., "PANGAEA: A Global and Inclusive Benchmark for Geospatial Foundation Models", 2025, https://doi.org/10.48550/arXiv.2412.04204 Focus to learn more

PRITHVI-EO-2.0 AND TERRAMIND

A community effort

arXiv:2412.02732v2 [cs.CV] 3 Feb 2025

Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications

Daniela Szwarcman[1,†], Sujit Roy[2,3,†,‡], Paolo Fraccaro[1,†,‡], Þorsteinn Elí Gíslason[4], Benedikt Blumenstiel[1], Rinki Ghosal[3], Pedro Henrique de Oliveira[1], Joao Lucas de Sousa Almeida[1], Rocco Sedona[5], Yanghui Kang[6], Srija Chakraborty[12], Sizhe Wang[7], Carlos Gomes[1], Ankur Kumar[3], Myscon Truong[8], Denys Godwin[9], Hyunho Lee[7], Chia-Yu Hsu[7], Ata Akbari Asanjan[12], Besart Mujeci[12], Disha Shidham[12], Trevor Keenan[11], Paulo Arevalo[10], Wenwen Li[7], Hamed Alemohammad[9], Pontus Olofsson[2], Christopher Hain[2], Robert Kennedy[8], Bianca Zadrozny[1], David Bell[12], Gabriele Cavallaro[4,5], Campbell Watson[1], Manil Maskey[2], Rahul Ramachandran[2], Juan Bernabe Moreno[1]

[1]IBM Research (UK and Ireland, Brazil, Zurich, and USA).
[2]NASA Marshall Space Flight Center, Huntsville, AL, USA.
[3]Earth System Science Center, The University of Alabama in Huntsville, AL, USA.
[4]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland.
[5]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany.
[6]Department of Biological Systems Engineering, Virginia Tech, Blacksburg, USA.
[7]School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA.
[8]College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA.
[9]Center for Geospatial Analytics, Clark University, Worcester, MA, USA.
[10]Department of Earth and Environment, Boston University, Boston, MA, USA.
[11]Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, USA.
[12]Earth from Space Institute, Universities Space Research Association, USA.

Abstract

This technical report presents Prithvi-EO-2.0, a new geospatial foundation model that offers significant improvements over its predecessor, Prithvi-EO-1.0. Trained on 4.2M global time series samples from NASA's Harmonized Landsat and Sentinel-2 data archive at 30m resolution, the new 300M and 600M parameter models incorporate temporal and location embeddings for enhanced performance across various geospatial tasks. Through extensive benchmarking with GEO-Bench, the 600M version outperforms the previous Prithvi-EO model by 8% across a range of tasks. It also outperforms six other geospatial foundation models when benchmarked on remote sensing tasks from different domains and resolutions (i.e. from 0.1m to 15m). The results demonstrate the versatility of the model in both classical earth observation and high-resolution applications. Early involvement of end-users and subject matter experts (SMEs) are among the key factors that contributed to the project's success. In particular, SME involvement allowed for constant feedback on model and dataset design, as well as successful customization for diverse SME-led applications in disaster response, land use and crop mapping, and ecosystem dynamics monitoring. Prithvi-EO-2.0 is available on Hugging Face and IBM TerraTorch, with additional resources on GitHub. The project exemplifies the Trusted Open Science approach embraced by all involved organizations.

†Equal contribution, ‡ Corresponding authors:paolo.fraccaro@ibm.com, sujit.roy@nasa.gov
*https://huggingface.co/ibm-nasa-geospatial/Prithvi-EO-2.0
*https://github.com/NASA-IMPACT/Prithvi-EO-2.0

1

arXiv:2504.11171v1 [cs.CV] 15 Apr 2025

TerraMind: Large-Scale Generative Multimodality for Earth Observation

Johannes Jakubik[1,*] Felix Yang[1,2*] Benedikt Blumenstiel[1,*] Erik Scheurer[3] Rocco Sedona[3] Stefano Maurogiovanni[3,6] Jente Bosmans[4] Nikolaos Dionelis[4] Valerio Marsocci[4] Niklas Kopp[1] Rahul Ramachandran[5] Paolo Fraccaro[1,†] Thomas Brunschwiler[1,†] Gabriele Cavallaro[3,6,†] Juan Bernabe-Moreno[1,†] Nicolas Longépé[4,†]

[1]IBM Research – Europe
[2]ETH Zurich
[3]Forschungszentrum Jülich
[4]European Space Agency Φ-Lab
[5]NASA IMPACT
[6]University of Iceland
johannes.jakubik1@ibm.com

Abstract

We present TerraMind, the first any-to-any generative, multimodal foundation model for Earth observation (EO). Unlike other multimodal models, TerraMind is pretrained on dual-scale representations combining both token-level and pixel-level data across modalities. On a token level, TerraMind encodes high-level contextual information to learn cross-modal relationships, while on a pixel level, TerraMind leverages fine-grained representations to capture critical spatial nuances. We pretrained TerraMind on nine geospatial modalities of a global, large-scale dataset. In this paper, we demonstrate that (i) TerraMind's dual-scale early fusion approach unlocks a range of zero-shot and few-shot applications for Earth observation, (ii) TerraMind introduces "thinking in modalities" (TiM)—the capability of generating additional artificial data during finetuning and inference to improve the model output—and (iii) TerraMind achieves beyond state-of-the-art performance in community-standard benchmarks for EO like PANGAEA. The pretraining dataset, the model weights, and our code will be open-sourced under a permissive license.

1. Introduction

Earth observation (EO) increasingly benefits from multimodality because of the important integration of complementary information from different data sources. This becomes particularly relevant as data from specific satellite missions can be unavailable for a specific time or location due to low revisiting times or weather phenomena like cloud coverage. Vice versa, for the computer vision domain, EO data is an important playground for the development of new approaches as there is significant publicly available data of very high quality and complexity. The available modalities range from sensors of different satellite missions to relevant complementary information like digital elevation.

In this work, we introduce TerraMind, the first large-scale, any-to-any generative, multimodal model for EO. With TerraMind, we introduce a dual-scale pretraining on pixel-level and token-level for generative multi-modality extending previous approaches like 4M [52] and demonstrate benefits over training primarily on tokens. TerraMind encodes high-level contextual information in tokens to enable correlation learning and scaling, while, additionally capturing important fine-grained representations using pixel-level inputs. During the pretraining, TerraMind predicts masked target tokens so that our pretraining objective boils down to a cross-modal patch classification problem that results in high-quality latent spaces. TerraMind is pretrained on a custom global-scale geospatial dataset named TerraMesh with 9 million samples that have been aligned spatiotemporally and across modalities [7]. In addition to radar and optical satellite imagery from the Copernicus Sentinel-1 (S-1) and Sentinel-2 (S-2) missions, our dataset contains task-specific modalities such as land use/land cover (LULC) and normalized difference vegetation index (NDVI) maps, metadata like digital elevation models (DEM) and geographic coordinates, and natural language in the form of captions. To the best of our knowledge, TerraMind represents the first truly generative, multimodal

* Equal contribution
† Equal supervision

1

D. Szwarcman, S. Roy, P. Fraccaro, Þ. E. Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. de Sousa Almeida, R. Sedona, Y. Kang, S. Chakraborty, S. Wang, C. Gomes, A. Kumar, M. Truong, D. Godwin, H. Lee, C.-Y. Hsu, A. Akbari Asanjan, B. Mujeci, D. Shidham, T. Keenan, P. Arevalo, W. Li, H. Alemohammad, P. Olofsson, C. Hain, R. Kennedy, B. Zadrozny, D. Bell, G. Cavallaro, C. Watson, M. Maskey, R. Ramachandran, and J. B. Moreno, "Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications," in arXiv preprint, vol. 2412.02732, 2025, https://arxiv.org/abs/2412.02732

J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, R.Ramachandran, P. Fraccaro, T. Brunschwiler, G.Cavallaro, J. Bernabe-Moreno, N. Longépé, "TerraMind: Large-Scale Generative Multimodality for Earth Observation", in arXiv:2504.11171, 2025, https://arxiv.org/abs/2504.11171

FAST-EO PROJECT

FOSTERING ADVANCEMENTS IN FOUNDATION MODELS VIA UNSUPERVISED AND SELF-SUPERVISED LEARNING FOR DOWNSTREAM TASKS IN EARTH OBSERVATION

esa Φ-lab

DLR IBM JÜLICH Forschungszentrum KP LABS Mission Complete FAST-EO

https://www.fast-eo.eu/

FAST-EO project is funded by the the European Space Agency (ESA) Phi-Lab under the contract No. 4000143501/23/I-DT

TERRAMIND: FOUNDATION MODEL FOR EO

TerraMind: Large-Scale Generative Multimodality for Earth Observation

Johannes Jakubik[1,*] Felix Yang[1,2*] Benedikt Blumenstiel[1,*] Erik Scheurer[3]
Rocco Sedona[3] Stefano Maurogiovanni[3,6] Jente Bosmans[4] Nikolaos Dionelis[4]
Valerio Marsocci[4] Niklas Kopp[1] Rahul Ramachandran[5] Paolo Fraccaro[1,†]
Thomas Brunschwiler[1,†] Gabriele Cavallaro[3,6,†] Juan Bernabe-Moreno[1,†] Nicolas Longépé[4,†]
[1]IBM Research – Europe
[2]ETH Zurich
[3]Forschungszentrum Jülich
[4]European Space Agency Φ-Lab
[5]NASA IMPACT
[6]University of Iceland
johannes.jakubik1@ibm.com

arXiv:2504.11171v1 [cs.CV] 15 Apr 2025

Abstract

We present TerraMind, the first any-to-any generative, multimodal foundation model for Earth observation (EO). Unlike other multimodal models, TerraMind is pretrained on dual-scale representations combining both token-level and pixel-level data across modalities. On a token level, TerraMind encodes high-level contextual information to learn cross-modal relationships, while on a pixel level, TerraMind leverages fine-grained representations to capture critical spatial nuances. We pretrained TerraMind on nine geospatial modalities of a global, large-scale dataset. In this paper, we demonstrate that (i) TerraMind's dual-scale early fusion approach unlocks a range of zero-shot and few-shot applications for Earth observation, (ii) TerraMind introduces "thinking in modalities" (TiM)—the capability of generating additional artificial data during finetuning and inference to improve the model output—and (iii) TerraMind achieves beyond state-of-the-art performance in community-standard benchmarks for EO like PANGAEA. The pretraining dataset, the model weights, and our code will be open-sourced under a permissive license.

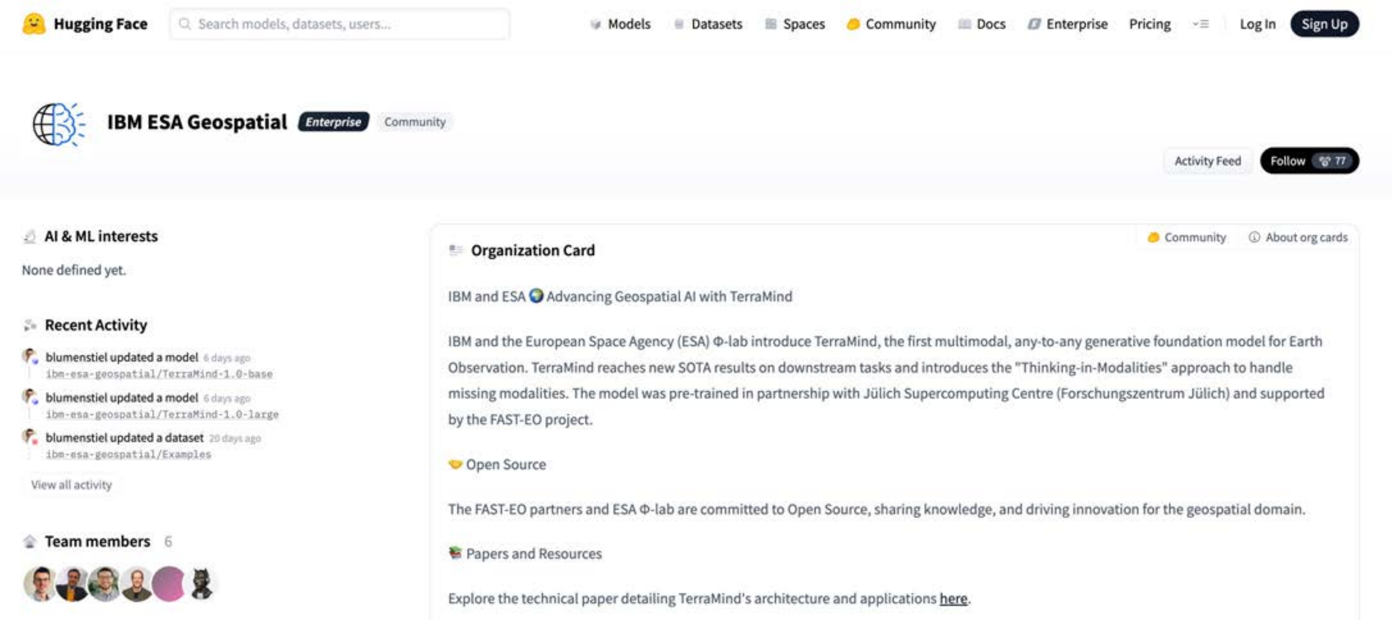
1. Introduction

Earth observation (EO) increasingly benefits from multimodality because of the important integration of complementary information from different data sources. This becomes particularly relevant as data from specific satellite missions can be unavailable for a specific time or location due to low revisiting times or weather phenomena like cloud coverage. Vice versa, for the computer vision domain, EO data is an important playground for the development of new approaches as there is significant publicly available data of very high quality and complexity. The available modalities range from sensors of different satellite missions to relevant complementary information like digital elevation.

In this work, we introduce TerraMind, the first large-scale, any-to-any generative, multimodal model for EO. With TerraMind, we introduce a dual-scale pretraining on pixel-level and token-level for generative multi-modality extending previous approaches like 4M [52] and demonstrate benefits over training primarily on tokens. TerraMind encodes high-level contextual information in tokens to enable correlation learning and scaling, while, additionally capturing important fine-grained representations using pixel-level inputs. During the pretraining, TerraMind predicts masked target tokens so that our pretraining objective boils down to a cross-modal patch classification problem that results in high-quality latent spaces. TerraMind is pretrained on a custom global-scale geospatial dataset named TerraMesh with 9 million samples that have been aligned spatiotemporally and across modalities [7]. In addition to radar and optical satellite imagery from the Copernicus Sentinel-1 (S-1) and Sentinel-2 (S-2) missions, our dataset contains task-specific modalities such as land use/land cover (LULC) and normalized difference vegetation index (NDVI) maps, metadata like digital elevation models (DEM) and geographic coordinates, and natural language in the form of captions. To the best of our knowledge, TerraMind represents the first truly generative, multimodal

* Equal contribution
† Equal supervision

1



https://huggingface.co/ibm-esa-geospatial

J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, R.Ramachandran, P. Fraccaro, T. Brunschwiler, G.Cavallaro, J. Bernabe-Moreno, N. Longépé, "TerraMind: Large-Scale Generative Multimodality for Earth Observation", in arXiv:2504.11171, 2025, https://arxiv.org/abs/2504.11171

HOW TO CREATE A FOUNDATION MODEL?

Simplified workflow

1) Gather data at scale

2) Train foundation model one time and evaluate

3) Fine-tune model for multiple downstream tasks

4) Inference (operational)

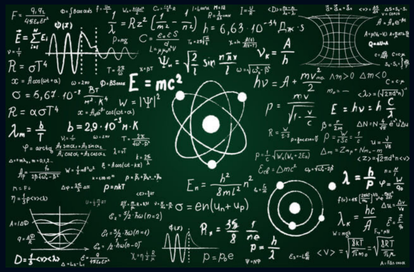
3

SUPERCOMPUTING

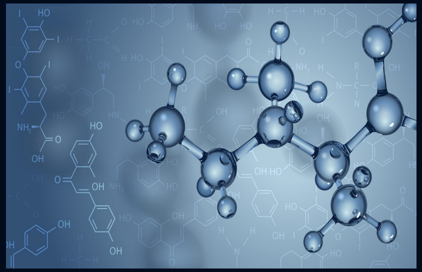
WHAT ARE SUPERCOMPUTERS USED FOR*?

*Also Beyond AI 😅

ENGINE OF SCIENTIFIC PROGRESS.
TACKLE PRESSING SOCIETAL PROBLEMS



Physics



Chemistry

Medicine

Climate

Society

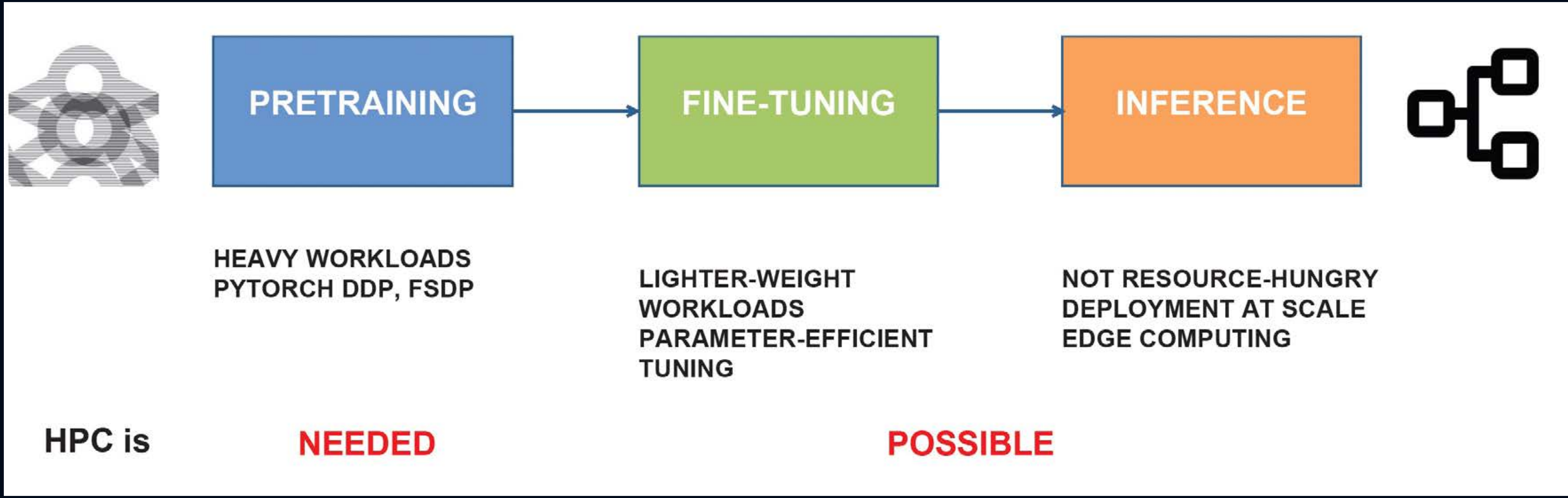
Used for data-intensive and computation-heavy scientific and engineering applications

COMMON TRENDS IN SCIENTIFIC COMMUNITIES

- More **Digital Twins**
- More **AI for science**
  - Foundation models
  - Data assimilation and interpretation
- **Tighter link to experiments** and data sources
  - Streaming data
  - Jointly analysing data from different sources
- **Challenges around** dealing with and moving **large data volumes**
  - Rather than about compute power

Estela Suarez, "HPC Advancing Science: Highlights 2024, at ACM/IEEE Supercomputing Conference, 2024, https://youtu.be/aiykGz7Y8o0?si=whl8CsCOUssYDbGP

SUPERCOMPUTING FOR FOUNDATION MODELS

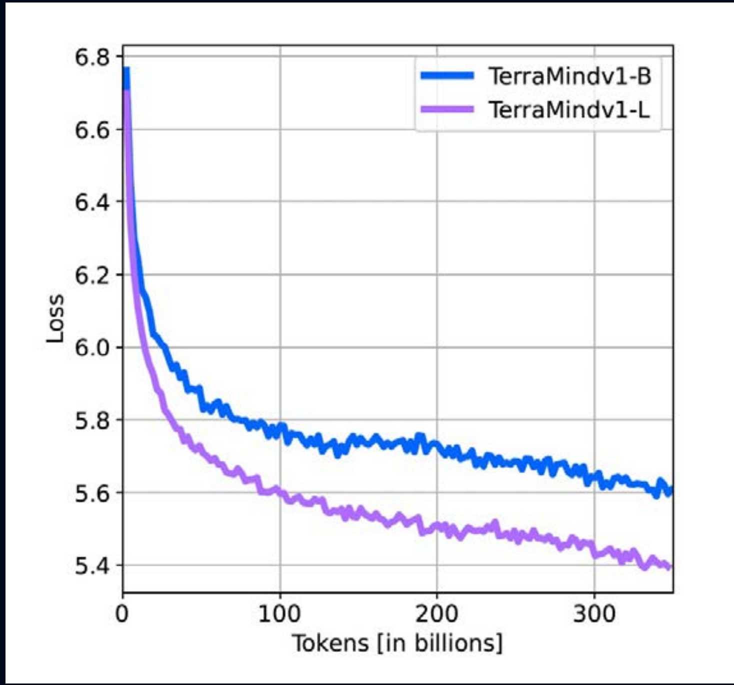


Credit: Rocco Sedona (Jülich Supercomputing Centre, Germany)

TERRAMIND WAS PRE-TRAINED ON JUWELS

Ranking in November 2020
(TOP500 (7 World, 1 Europe), Green500 (1 in TOP100) TOP10 AI (4)

- TerraMindv1-B (500B tokens)
  - 6 days on 32 NVIDIA A100 GPUs
- TerraMindv1-L (1 trillion tokens)
  - 10 days on 32 NVIDIA A100 GPUs

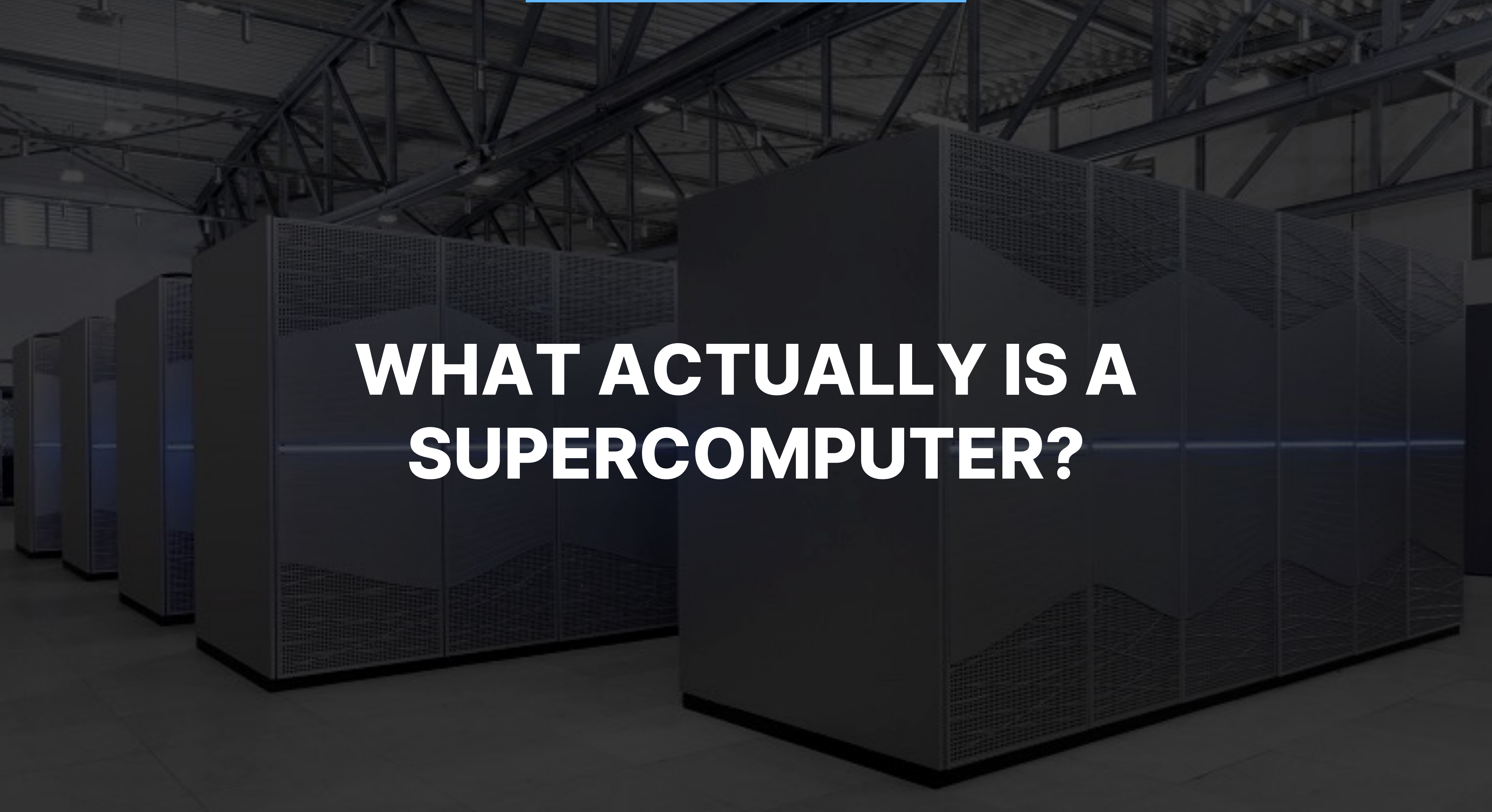


Scaling behavior comparing v1-B and v1-L models for the first 350B tokens on the validation loss of optical S-2 L2A data.

Overall, TerraMind-L outperforms TerraMind-B after approximately 10% of the training schedule of the large model.

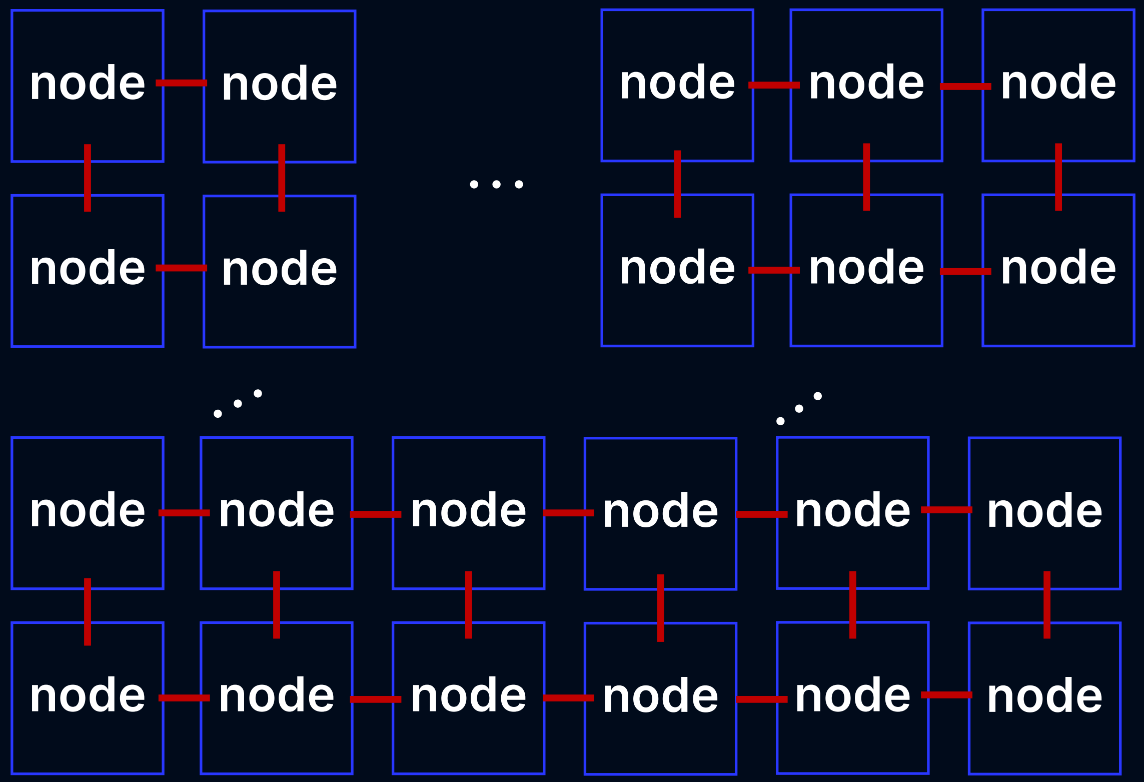
Julich Supercomputing Centre, "JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Julich Supercomputing Centre," Journal of large-scale research facilities, vol. 7, no. A138, 2021.

S. Kesselheim, A. Herten, K. Krajsek, J. Ebert, J. Jitsev, M. Cherti, M. Langguth, B. Gong, S. Stadtler, A. Mozaffari, G. Cavallaro, R. Sedona, A. Schug, A. Strube, R. Kamath, M. G. Schultz, M. Riedel, and T. Lippert, "JUWELS Booster – A Supercomputer for Large-Scale AI Research," in High Performance Computing (H. Jagode, H. Anzt, H. Ltaief, and P. Luszczek, eds.), (Cham), pp. 453–468, Springer International Publishing, 2021.



WHAT ACTUALLY IS A SUPERCOMPUTER?

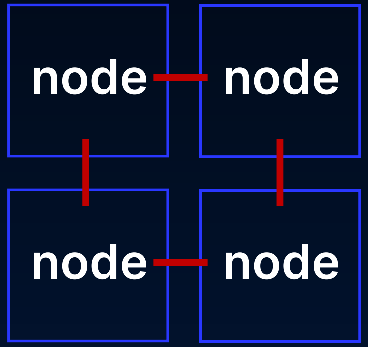
HIGH-PERFORMANCE COMPUTING SYSTEMS

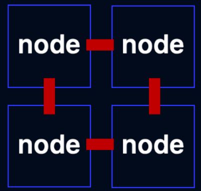


- High number of **compute nodes**
- Vast amounts of memory
- **High-speed interconnects**

HPC == tightly coupled parallel workloads

WHAT IS INSIDE COMPUTE NODES?

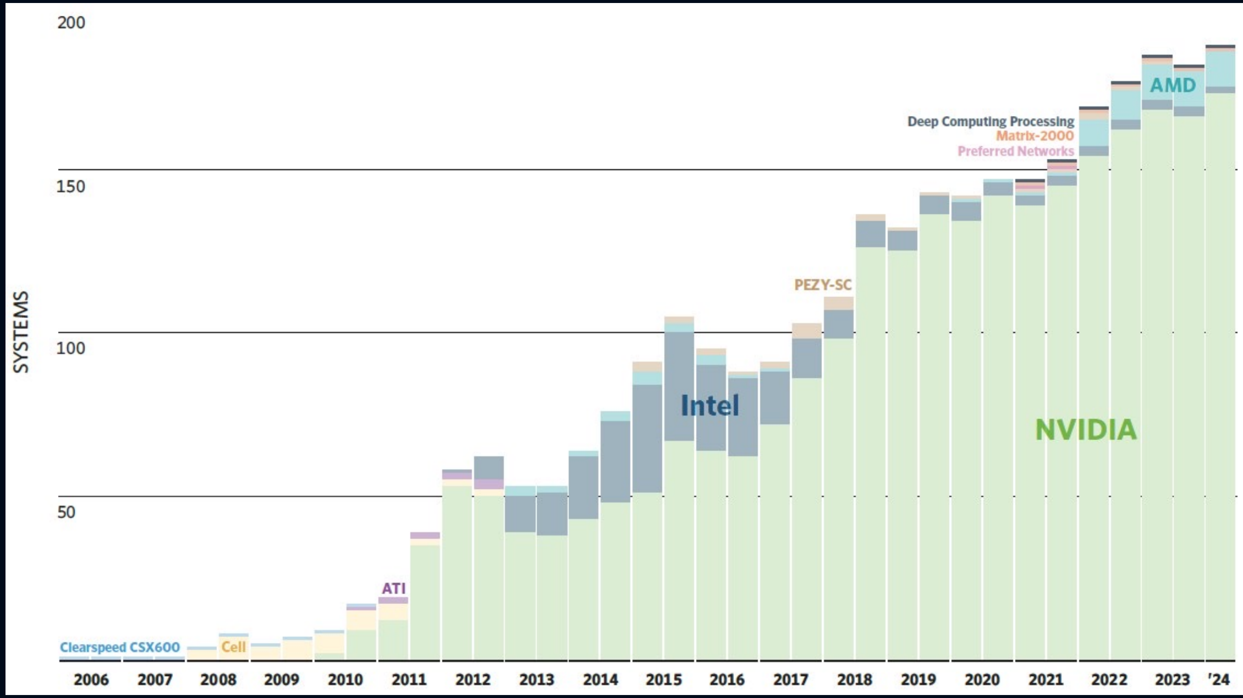




HPC SYSTEMS ARE ACCELERATED

- **Last decade**
  - More and more GPUs installed in HPC machines
  - More and more HPC machines with GPUs
  - More and more GPUs in each system

- **Future**
  - GPUs selected as technology for enabling Exascale
  - Even larger GPU machines with larger GPUs

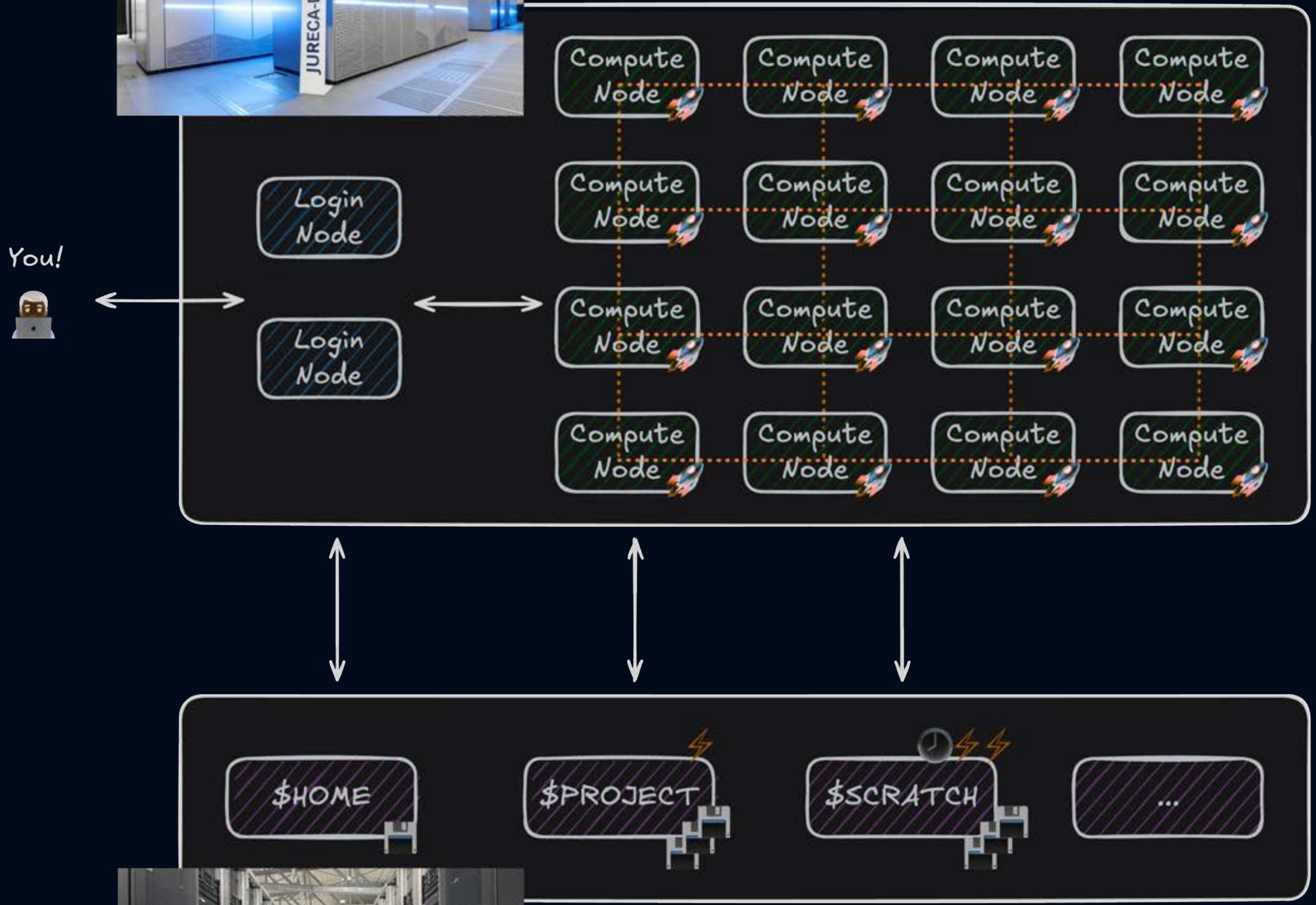


Timothy Prickett Morgan, "Top500 Supers: This Is Peak Nvidia For Accelerated Supercomputers", https://www.nextplatform.com/2024/05/13/top500-supers-this-is-peak-nvidia-for-accelerated-supercomputers

Credits: Andreas Herten (Jülich Supercomputing Centre, Forschungszentrum Jülich)

3. Supercomputing

JURECA
Copyright - Forschungszentrum Jülich
https://www.fz-juelich.de/en/ias/jsc/systems/supercomputers/jureca

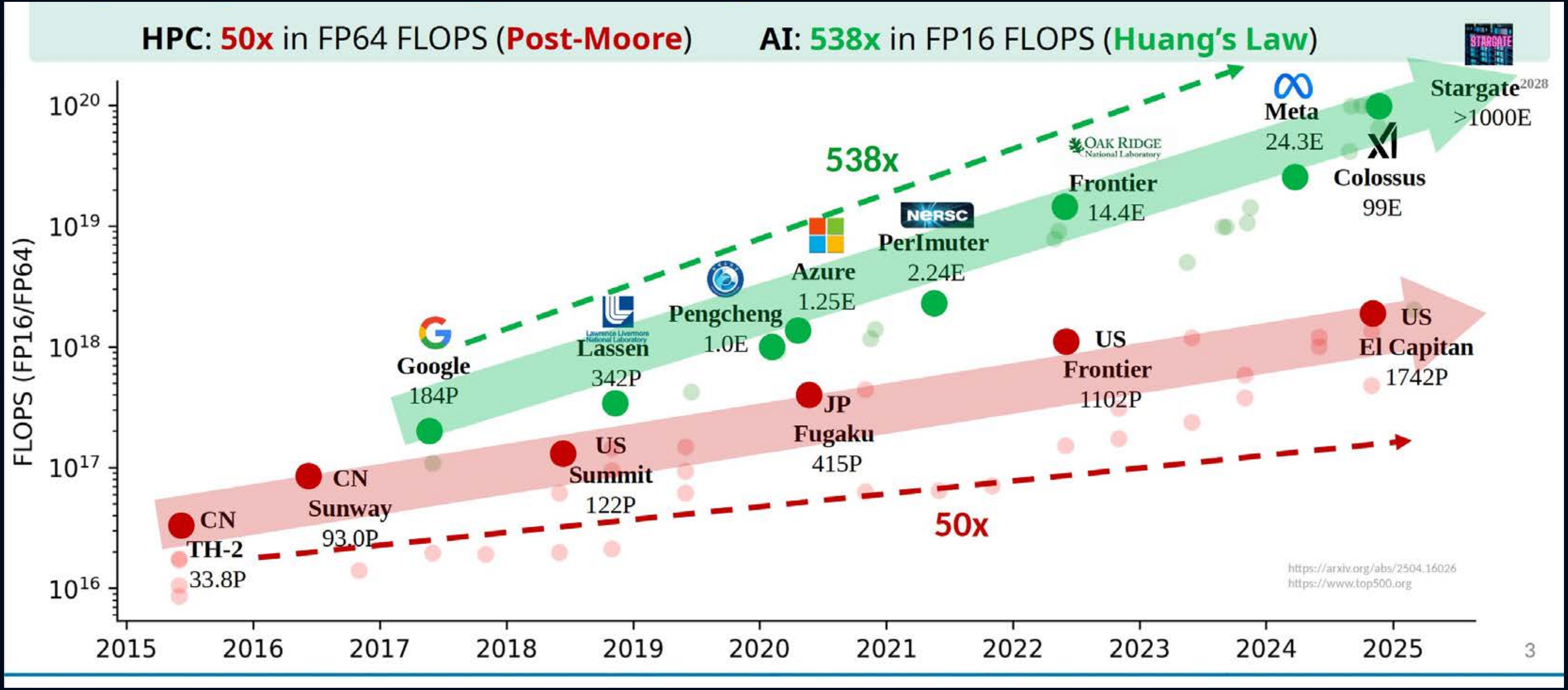


JUST
Copyright - Forschungszentrum Jülich
https://www.fz-juelich.de/en/ias/jsc/systems/storage-systems/just

Credit: Stefano Maurogiovanni (Jülich Supercomputing Centre, Germany)

PETA TO EXA-SCALE SYSTEM

System performance growth in the past decade



Yutong Lu, "Closing Keynote: Tackling Fragmentation in Exascale Supercomputing and Beyond", ISC High Performance 2025, https://isc-hpc.com/
https://arxiv.org/abs/2504.16026
https://www.top500.org

EUROPE IS INVESTING IN AI LEADERSHIP

To support scientific advances, infrastructure development, and wider technology adoption.

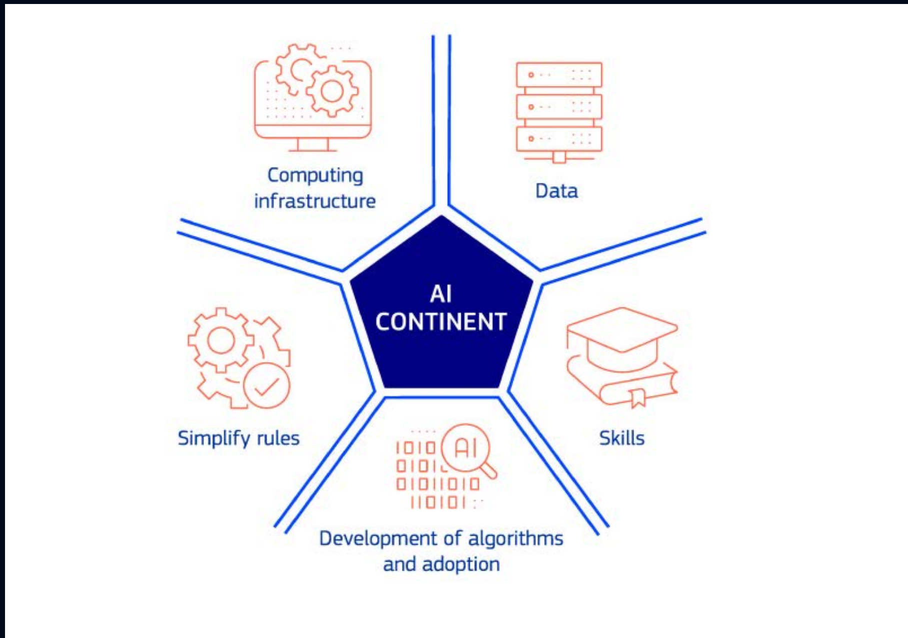
AI CONTINENT ACTION PLAN


PRESS RELEASE | Feb 11, 2025 | Paris | 3 min read

EU launches InvestAI initiative to mobilise €200 billion of investment in artificial intelligence*

5 Pillars for Europe to become the AI Continent

1. Building a large-scale AI computing infrastructure
2. Increasing access to high-quality data
3. Promoting AI in strategic sectors
4. Strengthening AI skills and talents
5. Simplifying the implementation of the AI act



European Commision, "Shaping Europe's leadership in artificial intelligence with the AI continent action plan", https://commission.europa.eu/topics/eu-competitiveness/ai-continent_en
European Commision, "AI Continent Action Plan COM(2025)165", 2025, https://commission.europa.eu/topics/eu-competitiveness/ai-continent_en
AI Continent Action Plan, https://digital-strategy.ec.europa.eu/en/factpages/ai-continent-action-plan

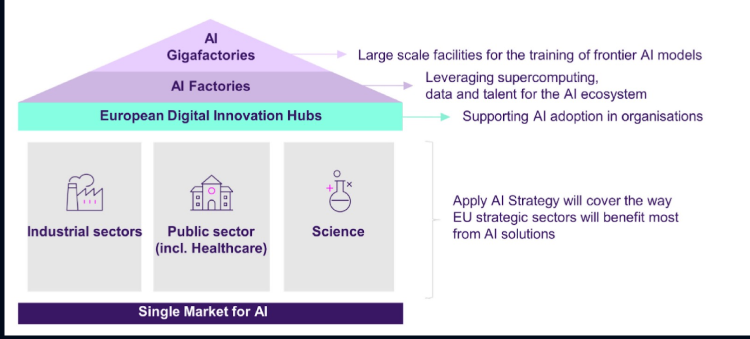
HPC POWER IN THE EU IS PUBLICLY ACCESSIBLE

In 2024 Europe hosts 30% of the world's top ten supercomputers

- European network of cutting-edge supercomputers deployed by the European High-Performance Computing Joint Undertaking (**EuroHPC JU**)
  - EuroHPC was launched in 2018 and co-funded by the EU, Member States, and private actors.

- TOP500 (November 2024, global ranking)
  - LUMI (#8)
  - Leonardo (#9)
  - MareNostrum 5 ACC (#11)

European High Performance Computing Joint Undertaking (EuroHPC JU), "EuroHPC Supercomputers", https://eurohpc-ju.europa.eu/supercomputers_en



AI CONTINENT ACTION PLAN

Area: Computing infrastructure

AI Factories

- Objective: train and finetune AI models
- Budget: €10 billion from 2021 to 2027
- At least 13 operational AI factories by 2026

AI Gigafactories

- Objective: train and develop complex AI models
- 4x more powerful than AI Factories
- €20 billion mobilised by InvestA!
- Deploy up to 5 Gigafactories

Cloud and AI development Act

- Objective: boost research in highly sustainable infrastructure
- Encourage investments
- Triple the EU's data centre capacity in the next 5-7 years

European Commision, “Shaping Europe’s leadership in artificial intelligence with the AI continent action plan”, https://commission.europa.eu/topics/eu-competitiveness/ai-continent_en

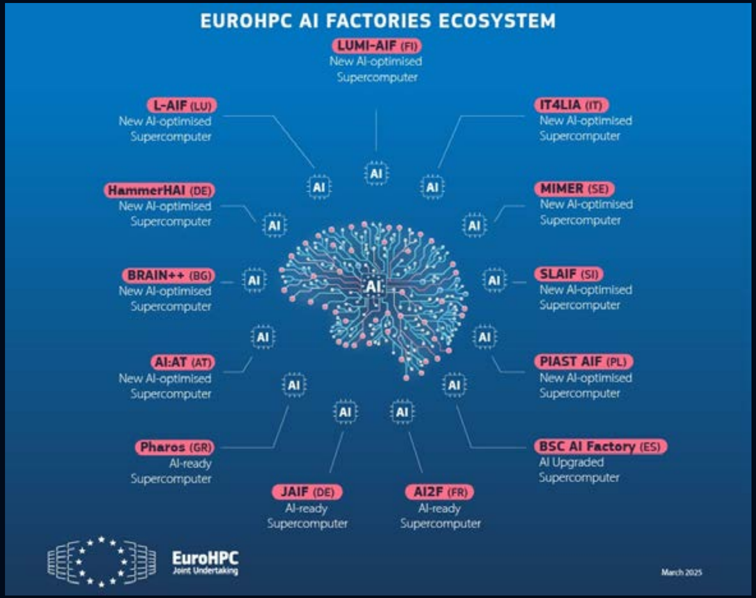
European Commision, “AI Continent Action Plan COM(2025)165”, 2025, https://commission.europa.eu/topics/eu-competitiveness/ai-continent_en

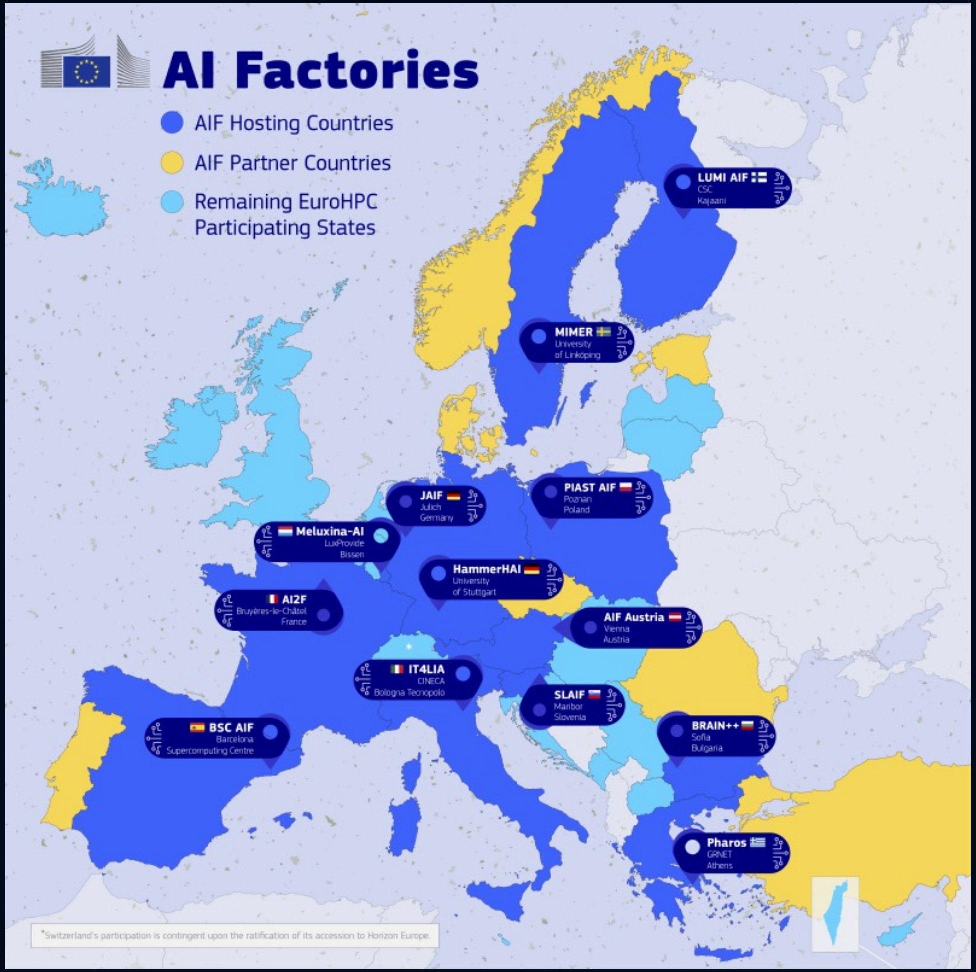
AI Continent Action Plan, https://digital-strategy.ec.europa.eu/en/factpages/ai-continent-action-plan

EUROHPC AI FACTORIES

To triple the current EuroHPC AI computing capacity

- Facilitate access to AI provided by HPC facilities
- Dynamic ecosystems that foster innovation, collaboration, and development in the field of AI
- Support startups, industry, and researchers to develop cutting-edge AI models and applications.



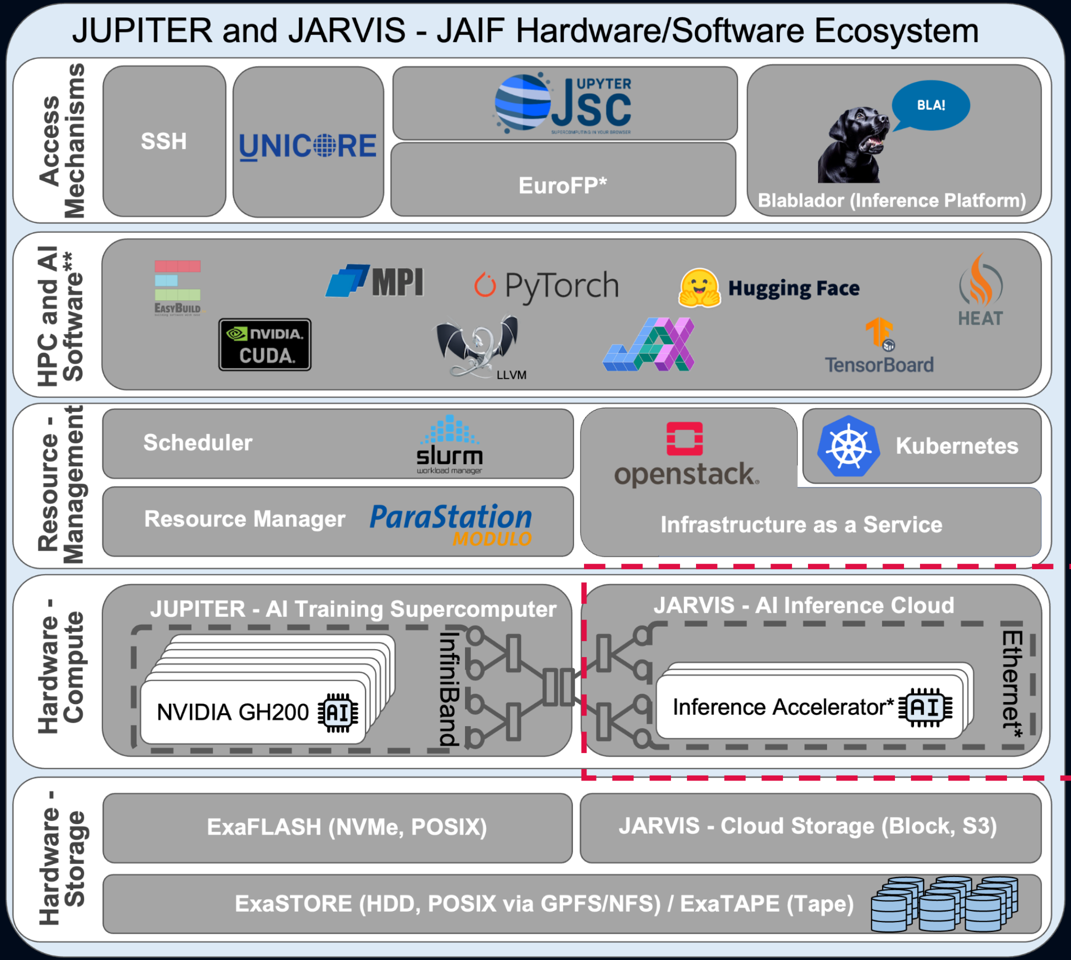


THE JUPITER AI FACTORY (JAIF)

Modulard JUPITER - Hybrid Training/Inference AI System

Consortium

Contact: jaif@fz-juelich.de



*Depending on procurements and available functionality

**This is a subset of the available software

JUPITER inference system **JARVIS** (JUPITER Advanced Research Vehicle for Inference Services), a cloud-based AI inference platform

Jülich Supercomputing Centre, "Europe's AI Booster: JUPITER AI Factory Brings Exascale Power to Business and Science", https://www.fz-juelich.de/en/news/archive/press-release/2025/europes-ai-booster-jupiter-ai-factory

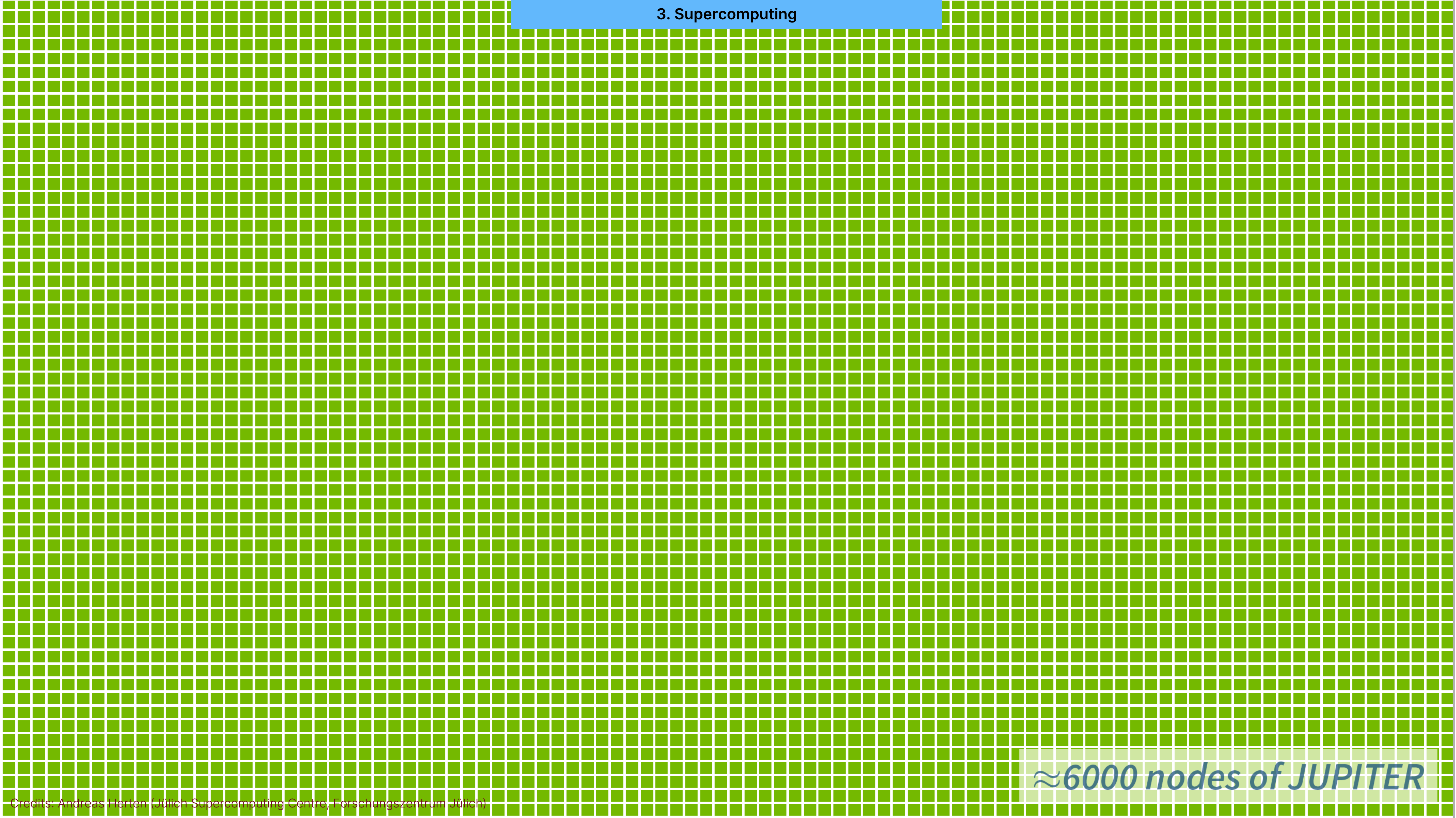
Credits: Mathis Bode, Jülich Supercomputing Centre (Forschungszentrum Jülich)

JUPITER: FIRST EUROPEAN EXASCALE SYSTEM

JU Pioneer for Innovative and Transformative Exascale Research (JUPITER)

Performance of 1 million smartphones (a stack as tall as Mount Everest)

Jülich Supercomputing Centre (Forschungszentrum Jülich), "Supercomputer JUPITER: The Journey to Exascale", https://youtu.be/5602Il8y7YE?si=398SLbXPnFAZKCRv

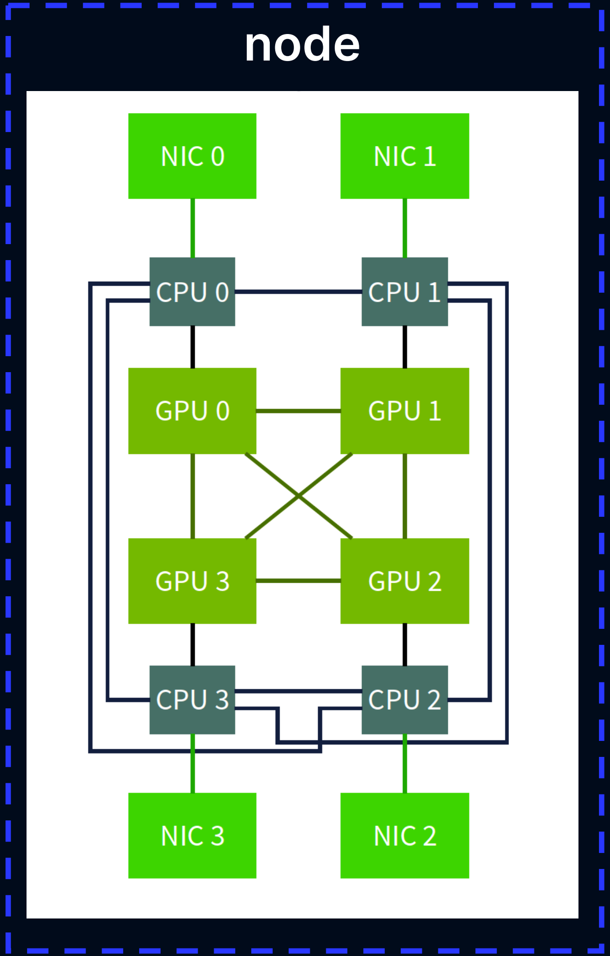


Credits: Andreas Herten (Jülich Supercomputing Centre, Forschungszentrum Jülich)

JUPITER BOOSTER

~6000 nodes, ~24 000 GPUs, 224 000 network devices

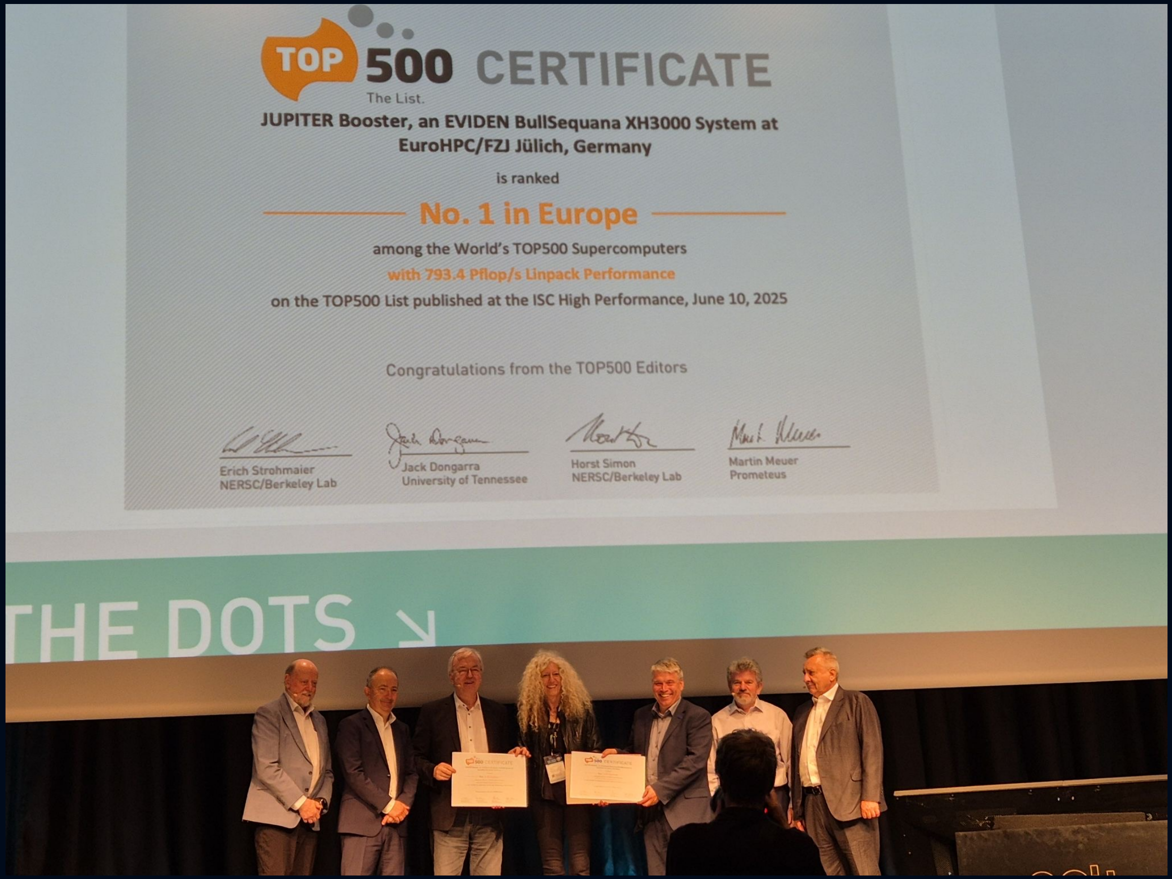
- GPU : 4 × NVIDIA H100 Grace-Hopper flavor 96 GB memory per GPU
- CPU: 4 × NVIDIA Grace, 4 × 72 cores; 4 × 120 GB LPDDR5X memory
- Network : 4 × NVIDIA Mellanox InfiniBand NDR200, 4 × 25GB/s



Jülich Supercomputing Centre (Forschungszentrum Jülich), "JUPITER | The Arrival of Exascale in Europe", https://www.fz-juelich.de/en/ias/jsc/jupiter

JUPITER RANKS 4TH ON THE TOP500 LIST AND IS EUROPE'S FASTEST SUPERCOMPUTER

June 2025



ISC High Performance 2025, https://isc-hpc.com/

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN EuroHPC/FZJ Germany	4,801,344	793.40	930.00	13,088

June 2025, https://top500.org/lists/top500/2025/06/

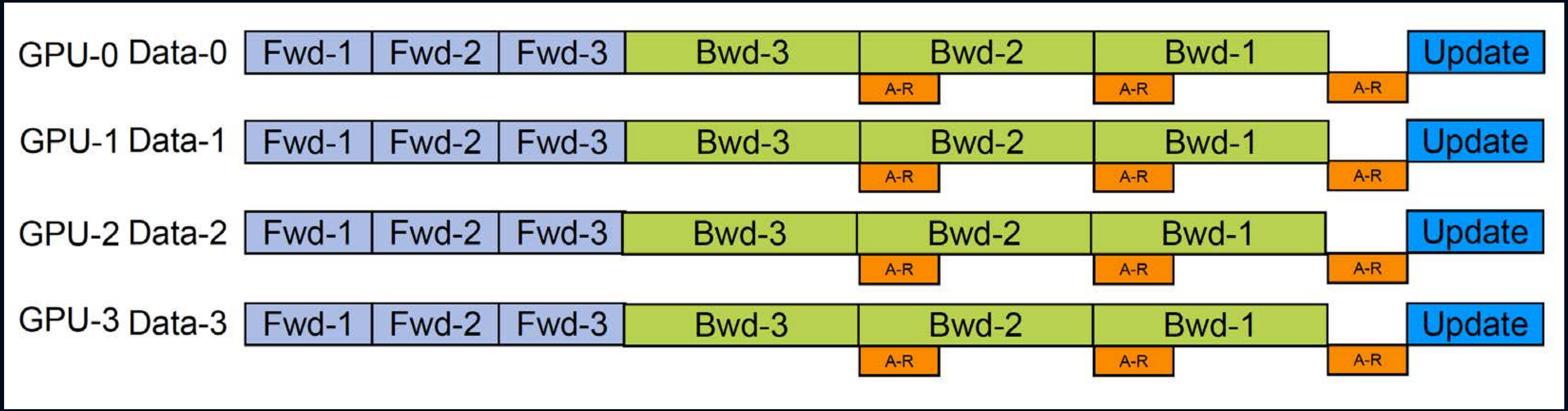
4

DISTRIBUTED DEEP LEARNING

DATA-PARALLEL TRAINING FLOW

- Each GPU processes a different data shard through forward and backward passes in parallel
- After backward pass, gradients are averaged across GPUs (All-Reduce) before synchronizing model updates

Replica of the same model on all GPUs



Fwd-1: Forward pass for mini-batch 1 (model on GPU is processing the first batch of data to compute predictions and loss)

Bwd-1: Backward pass for mini-batch 1 (GPU computes gradients for the first batch by backpropagating the loss)

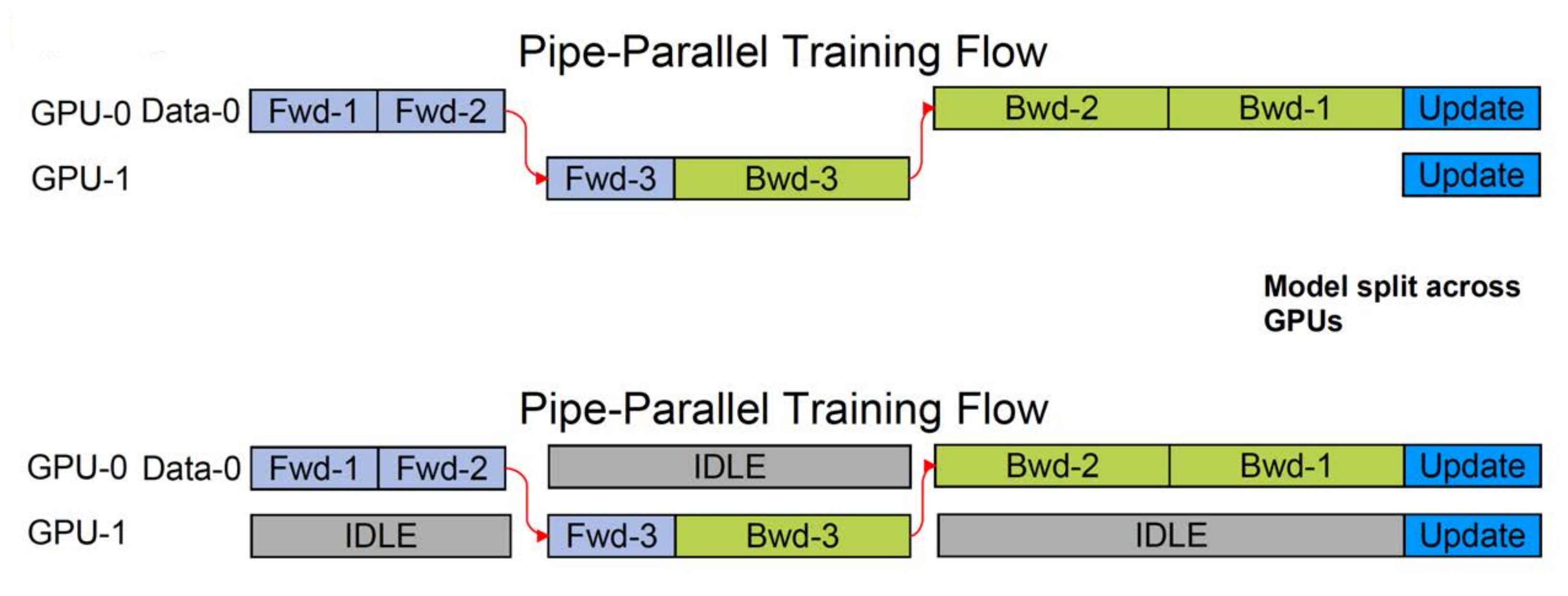
A-R: All-Reduce operation (collective communication step where all GPUs average their gradients (ensure that each model replica ends up with the same gradients for the weight update)

Shen Li, "Getting Started with Distributed Data Parallel", https://docs.pytorch.org/tutorials/intermediate/ddp_tutorial.html

MODEL PARALLELISM

Pipelining

- Model parallelism splits a single model across multiple GPUs
  - Each GPU handling a portion of the model's layers during forward and backward passes.
- Pipeline parallelism enables overlapping computation across GPUs
  - Can suffer from idle time if not carefully scheduled, as shown in the lower example.

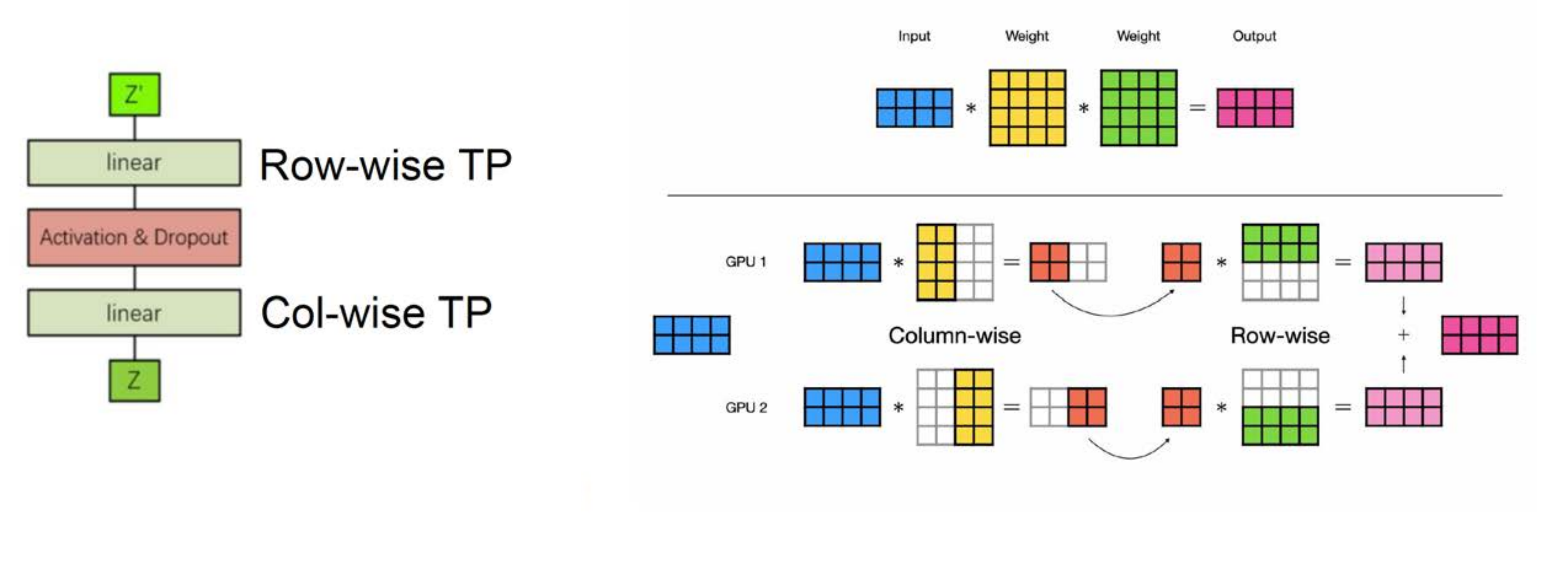


Shen Li, "Getting Started with Distributed Data Parallel", https://docs.pytorch.org/tutorials/intermediate/ddp_tutorial.html

MODEL PARALLELISM

Tensor Parallelism

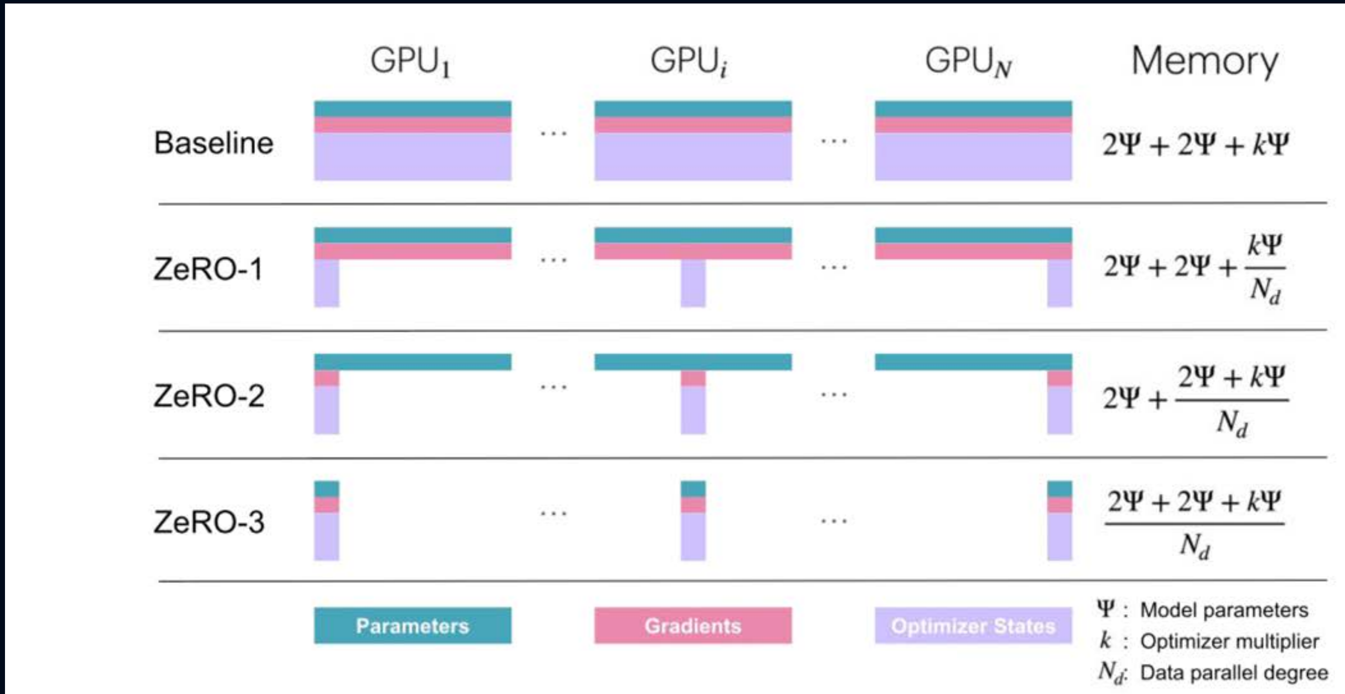
- Tensor parallelism splits individual tensor operations (like matrix multiplications) across GPUs
  - Enable finer-grained parallelism within a single layer
- Row-wise and column-wise strategies partition tensors differently
  - Requiring intermediate communication steps to assemble the final output



Tensor parallelism, https://lightning.ai/docs/pytorch/stable/advanced/model_parallel/tp.html

ZERO REDUNDANCY OPTIMIZER

- ZeRO progressively partitions model states (parameters, gradients, and optimizer states) across GPUs to reduce memory redundancy and scale training efficiently.



Activations are not distributed!

- PyTorch Fully Sharded Data Parallel (FSDP) implements ZeRO-3
  - Highest memory efficiency by fully sharding all states
  - But activations remain undivided and must still fit in memory

The Ultra-Scale Playbook: Training LLMs on GPU Clusters, https://huggingface.co/spaces/nanotron/ultrascale-playbook?section=high-level_overview#ultrascale-playbook?section=zero-1:_partitioning_optimizer_states

TOOLS FOR DISTRIBUTED DEEP LEARNING

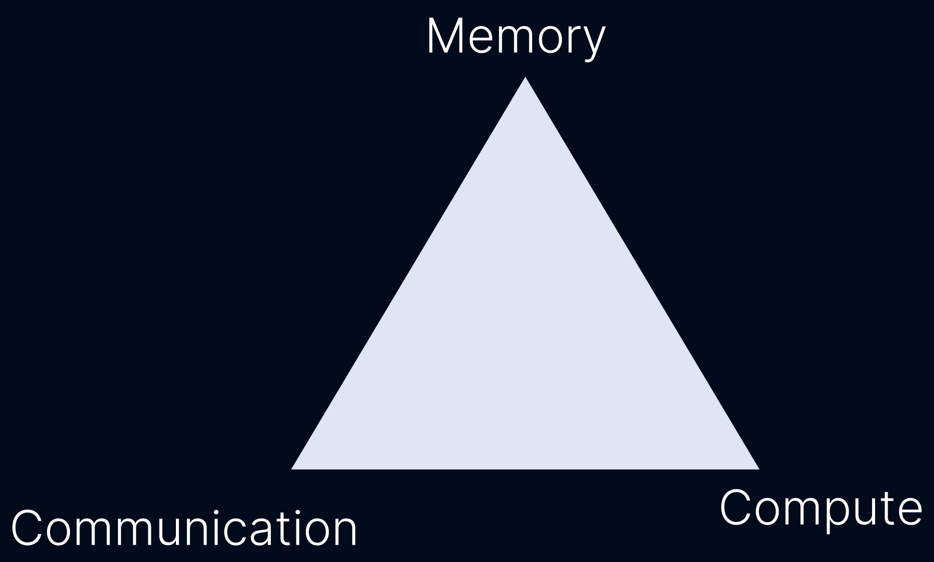
- 2013- DistBelief (Google) first large-scale parameter-server architecture
- 2015 -TensorFlow PS & SyncReplicas standardised PS + all-reduce hybrids
- 2016- MXNet KVStore flexible data/model parallel key-value store
- 2017 - Horovod (Uber) ring-allreduce over MPI/NCCL; drop-in for TF & Keras
- 2018- PyTorch DDP gradient bucketing; NCCL backend becomes default
- 2019 - Megatron-LM (NVIDIA) tensor & pipeline model-parallelism for > 10B-param LMs
- 2020 - DeepSpeed (Microsoft) ZeRO optimizer sharding, 8-bit optimisers
- 2021 - PyTorch FSDP fully-sharded data-parallel; near-linear scaling
- 2022- Alpa / Colossal-AI automatic 3-D parallel + memory optimisation
- 2023- Ray Train 2.x elastic autoscaling clusters; heterogeneous resources
- 2024 - VLLM / Paginated KV-cache inference-oriented distributed serving

TOOLS FOR DISTRIBUTED DEEP LEARNING

Is that all?

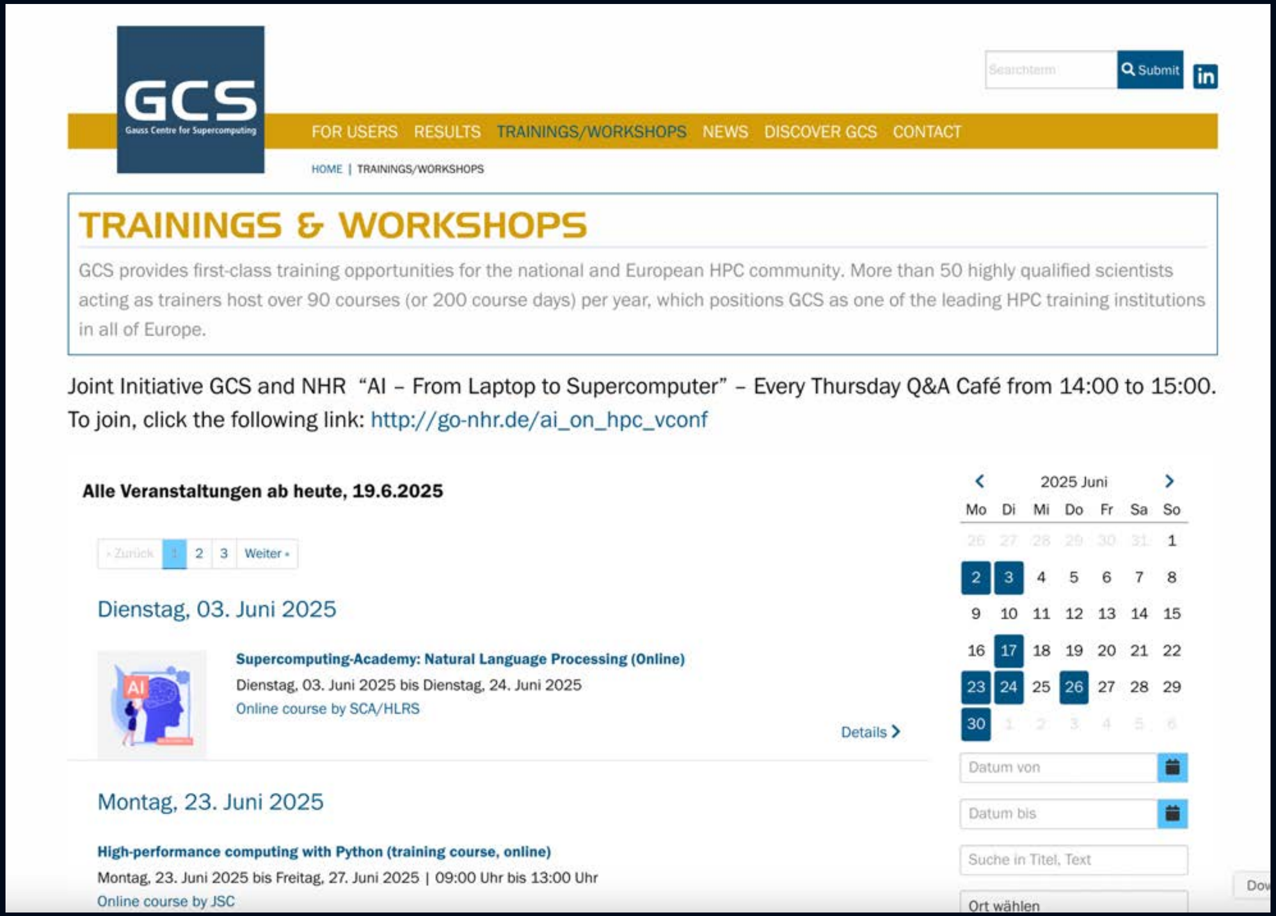
- Parallelism techniques
  - (expert parallelism, ...)
- Flash Attention
- Kernel Fusion
- Hyperparameter Optimization
- Inference & deployment

- Memory ⇄ Compute ⇄ Communication trade-off
- Every scaling trick pushes stress along triangle



Credit: Rocco Sedona (Jülich Supercomputing Centre, Germany)

LEARN MORE



https://www.gauss-centre.eu/trainingsworkshops


THANK YOU FOR YOUR ATTENTION