Université Grenoble Alpes

A Versatile Methodology for Assessing the Electricity Consumption and Environmental Footprint of Machine Learning Training: from Supercomputers to Edge Devices

PhD Defense of Mathilde JAY

Supervised by Prof. Denis Trystram and Dr. Laurent Lefèvre

Teams: DataMove (LIG), Avalon (LIP), Edge Intelligence (MIAI)

October 15th, 2024

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'informa

Spécialité : Informatique

Unité de recherche : Laboratoire d'Informatique de Grenoble

Une méthodologie polyvalente pour évaluer la consommat électrique et l'empreinte environnementale de l'entraînemen l'apprentissage machine : des supercalculateurs aux équipem embarqués

A Versatile Methodology for Assessing the Electricity Consumptio and Environmental Footprint of Machine Learning Training: from Supercomputers to Edge Devices

Présentée par :

Mathilde JAY

Direction de thèse :

Denis TRYSTRAM PROFESSEUR DES UNIVERSITES, GRENOBLE INP - UGA	Directeur de thèse
Laurent LEFEVRE CHARGE DE RECHERCHE HDR, CENTRE INRIA DE LYON	Co-directeur de thèse

Rapporteurs :

AURELIE BUGEAU PROFESSEURE DES UNIVERSITES, UNIVERSITE DE BORDEAUX

ANNE-LAURE LIGOZAT PROFESSEURE DES UNIVERSITES, ENSIIE

Thèse soutenue publiquement le **15 octobre 2024**, devant le jury composé de :

DENIS TRYSTRAM, PROFESSEUR DES UNIVERSITES, GRENOBLE INP - UGA	Directeur de thèse
LAURENT LEFEVRE, CHARGE DE RECHERCHE HDR, CENTRE INRIA DE LYON	Co-directeur de thèse
AURELIE BUGEAU, PROFESSEURE DES UNIVERSITES, UNIVERSITE DE BORDEAUX	Rapporteure
ANNE-LAURE LIGOZAT, PROFESSEURE DES UNIVERSITES, ENSIIE	Rapporteure
EMMA STRUBELL, ASSISTANT PROFESSOR, CARNEGIE MELLON UNIVERSITY	Examinatrice
SYLVAIN BOUVERET, MAITRE DE CONFERENCES, GRENOBLE INP - UGA	Examinateur
CLAUDE LEPAPE, INGENIEUR DE RECHERCHE, SCHNEIDER ELECTRIC	Examinateur
CLAUDIA RONCANCIO, PROFESSEURE DES UNIVERSITES, GRENOBLE INP - UGA	Examinatrice

Invités :

BRUNO MONNET INGENIEUR, Hewlett Packard Enterprise

MIAI Grenoble Alpes · Inria · La Région Auvergne-Rhône-Alpes · LIG · LIP

Power measurement

A Versatile Methodology for assessing the Electricity Consumption and Environmental Footprint of Machine Learning Training: from Supercomputers to Edge Devices

Estimation from **environmental databases**

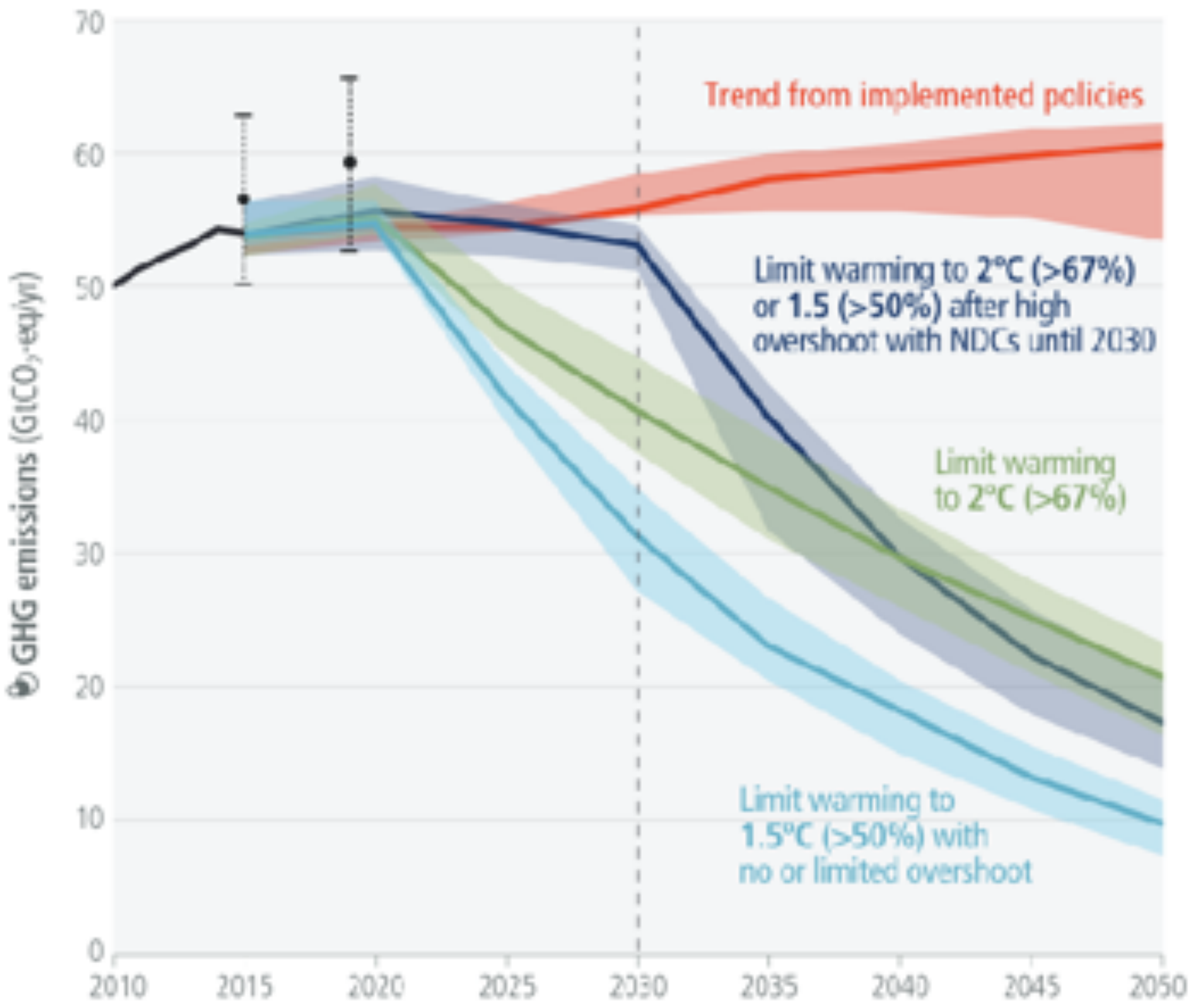
Contents

The impacts of Information and Communication Technologies

- Between **2.1% and 3.9%** of worldwide greenhouse gaz emissions in 2020 [Freitag2021]
- Emissions increased by ~ 5.5% every year (2015 to 2019) [ShiftProject2021]

- More than climate change, ICTs can contribute to
  - Freshwater change
  - Rare metal depletion
  - Primary energy consumption

[Benqassem2021]



Global GHG emissions of modeled pathways from IPCC 2023 Sixth Assessment Report.

GHG: GreenHouse Gaz , **IPCC**: Intergovernmental Panel on Climate Change, **ICT**: Information and Communication Technologies

Treaties and regulations to reduce emissions

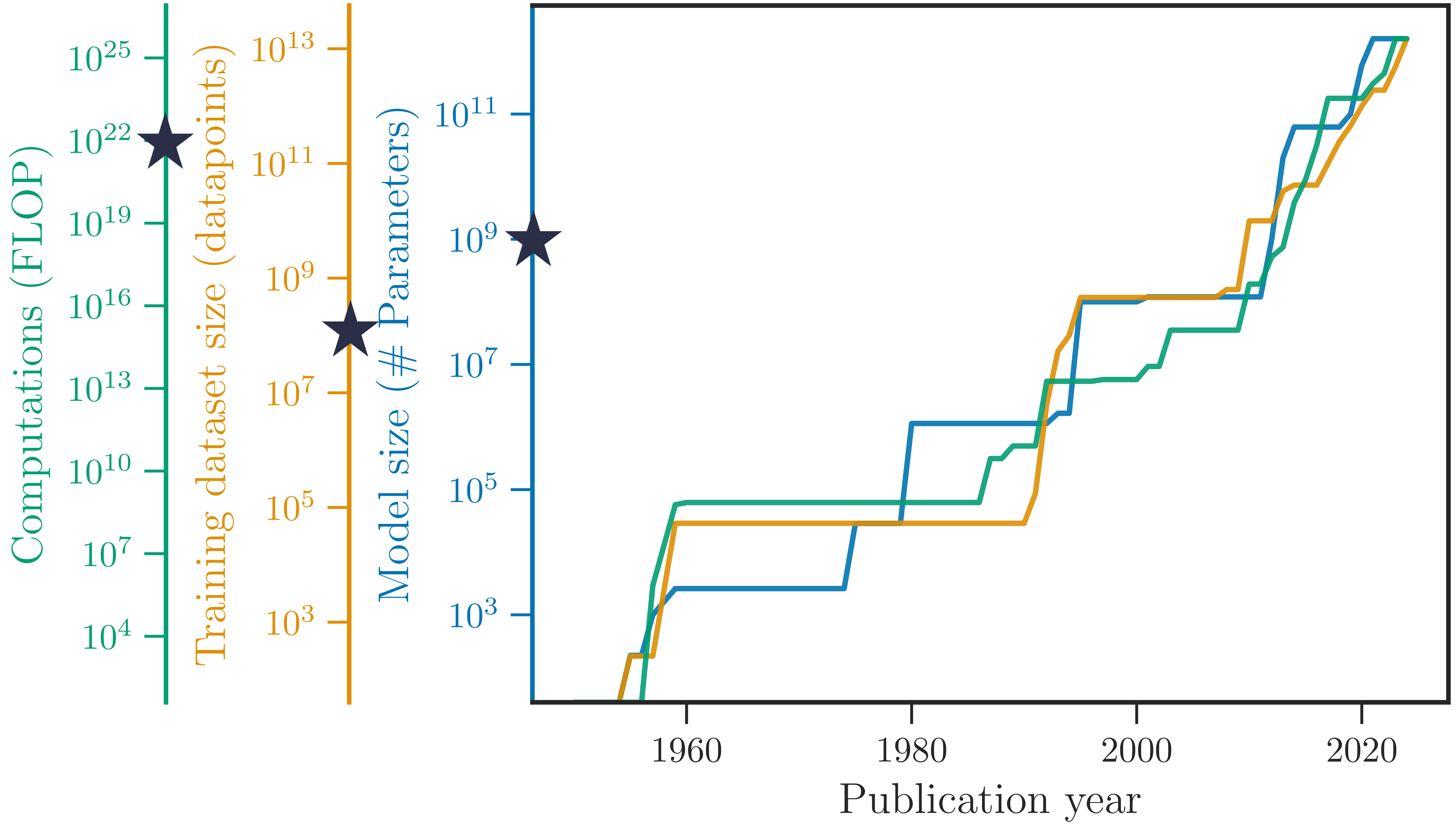
- Paris Agreement (2015)
- European Green Deal
  - Corporate Sustainability Reporting Directive (January 2023)
  - Energy Efficiency Directives

Need for the **companies** to assess the footprint of their digital services

Need for **policy makers** of assessment standards

Computational cost of Machine Learning

- Artificial Intelligence (AI)
  - Tasks that typically require human intelligence
- Machine Learning (ML)
  - AI that automatically learns from a set of data
- Deep Learning (DL)
  - ML that relies on deep neural network models



Evolutions in Machine Learning Model and their training, as the maximum value up to the given year. [EpochAI2024]

Stable Diffusion (2022): 256 A100 GPUs for 25 days

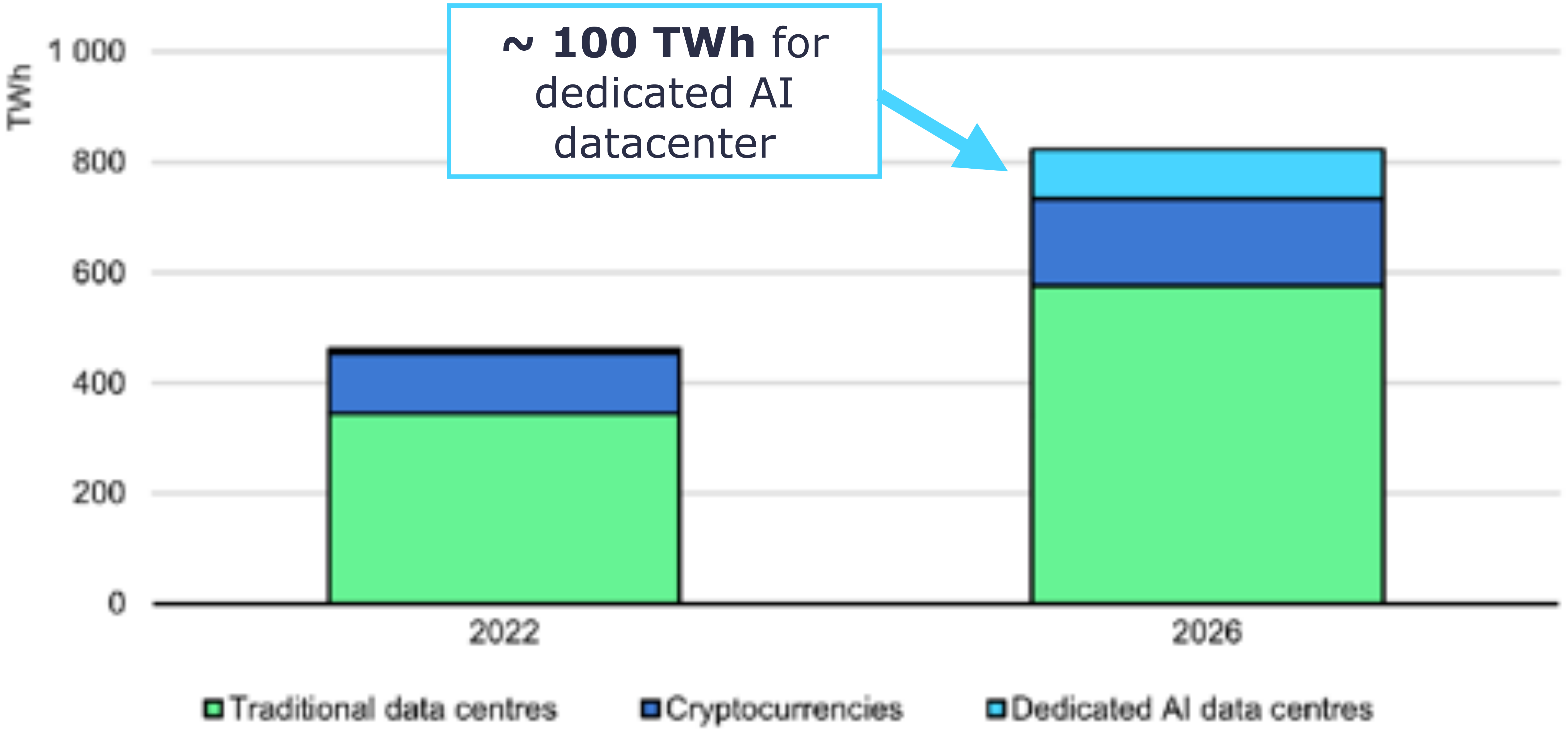
ICT: Information and Communication Technologies, **GPU**: Graphic Processing Units

The growing impact of Machine Learning

Machine Learning is one of main drivers of increase of **electricity demand** in the European Union between 2023 and 2026

[IEA2024]

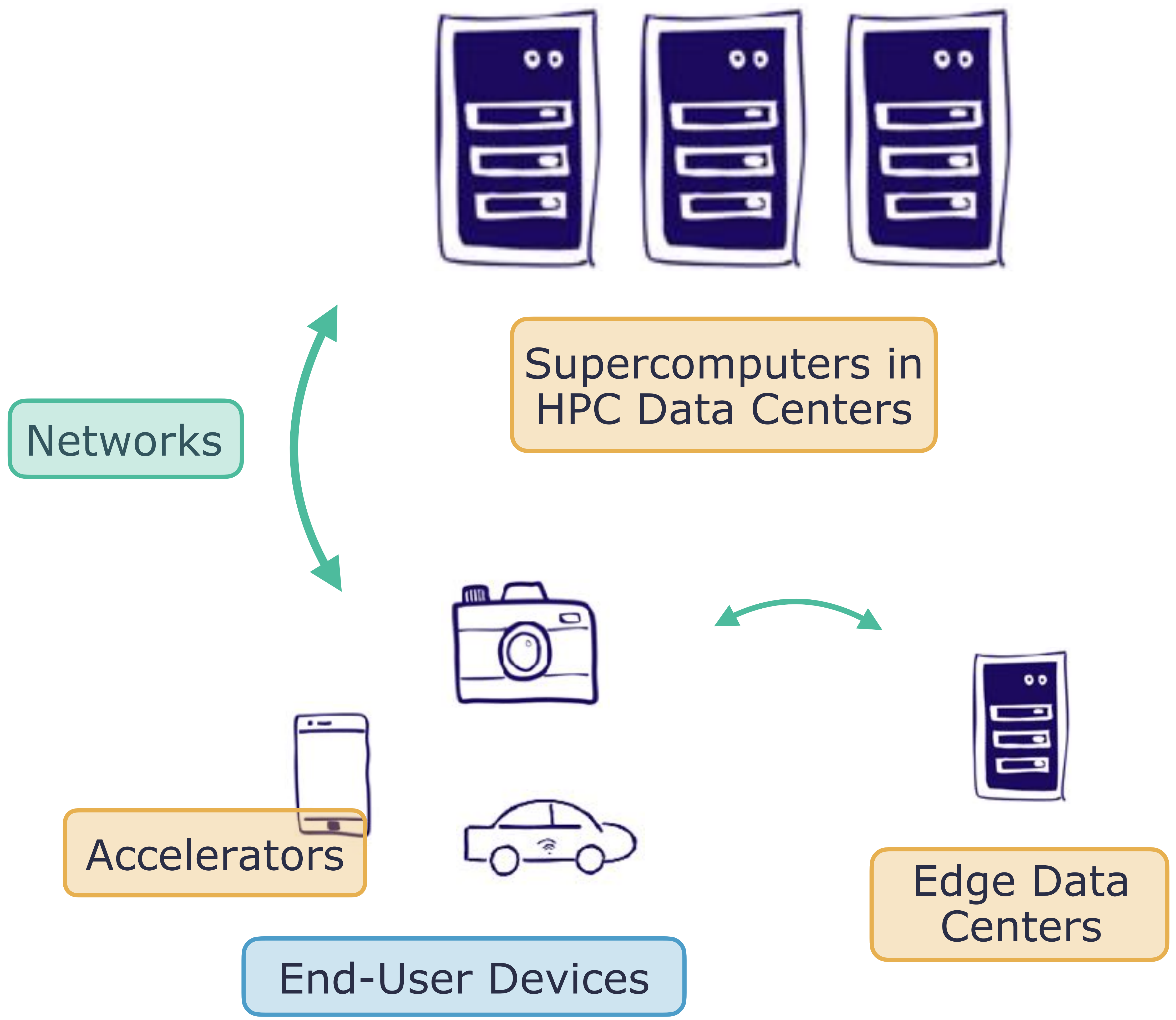
Need to assess footprint of machine learning services

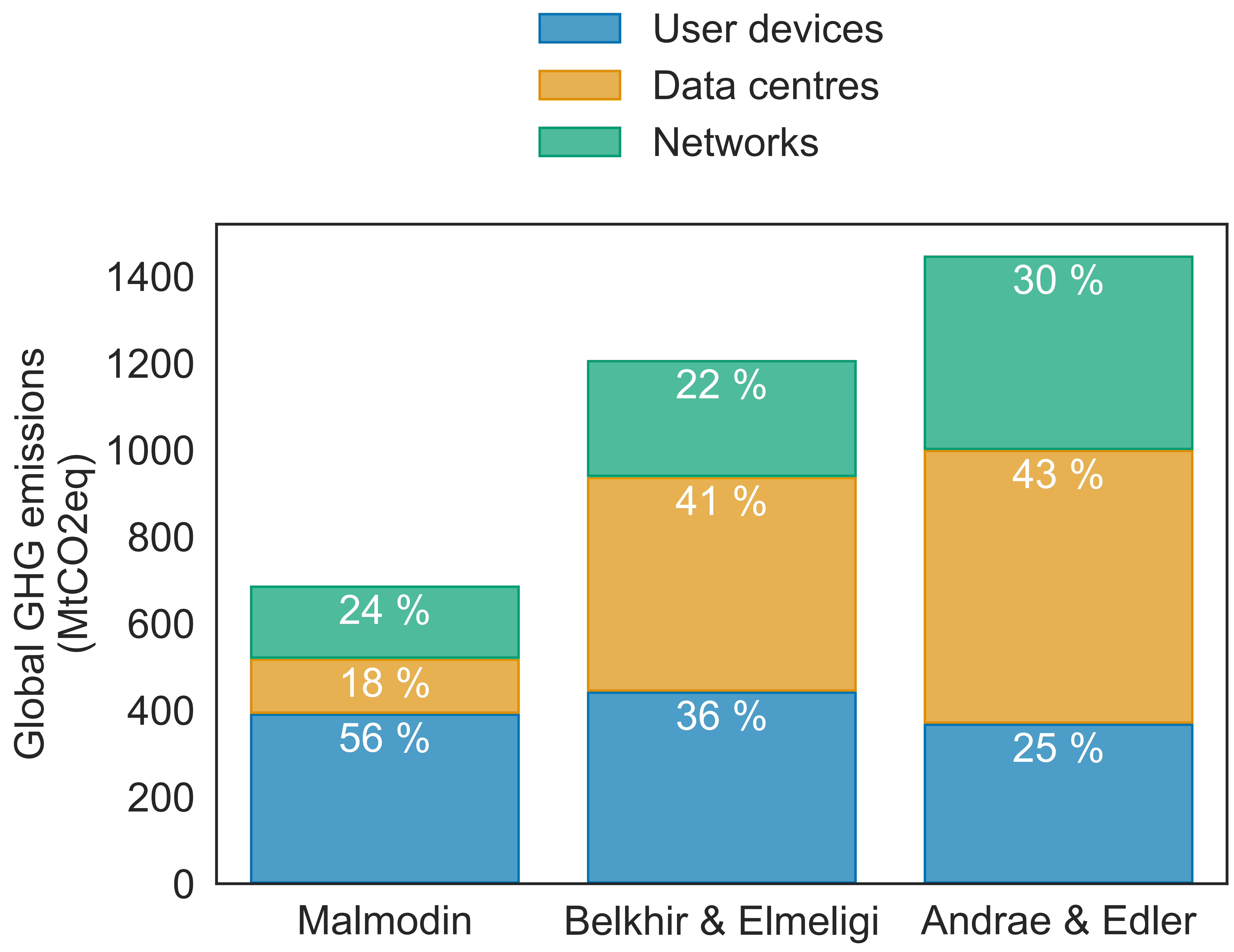


Estimated electricity demand from traditional data centres, dedicated AI data centres and cryptocurrencies in 2022 and 2026. [IEA2024]

IEA: International Energy Agency

Machine Learning services and materiality

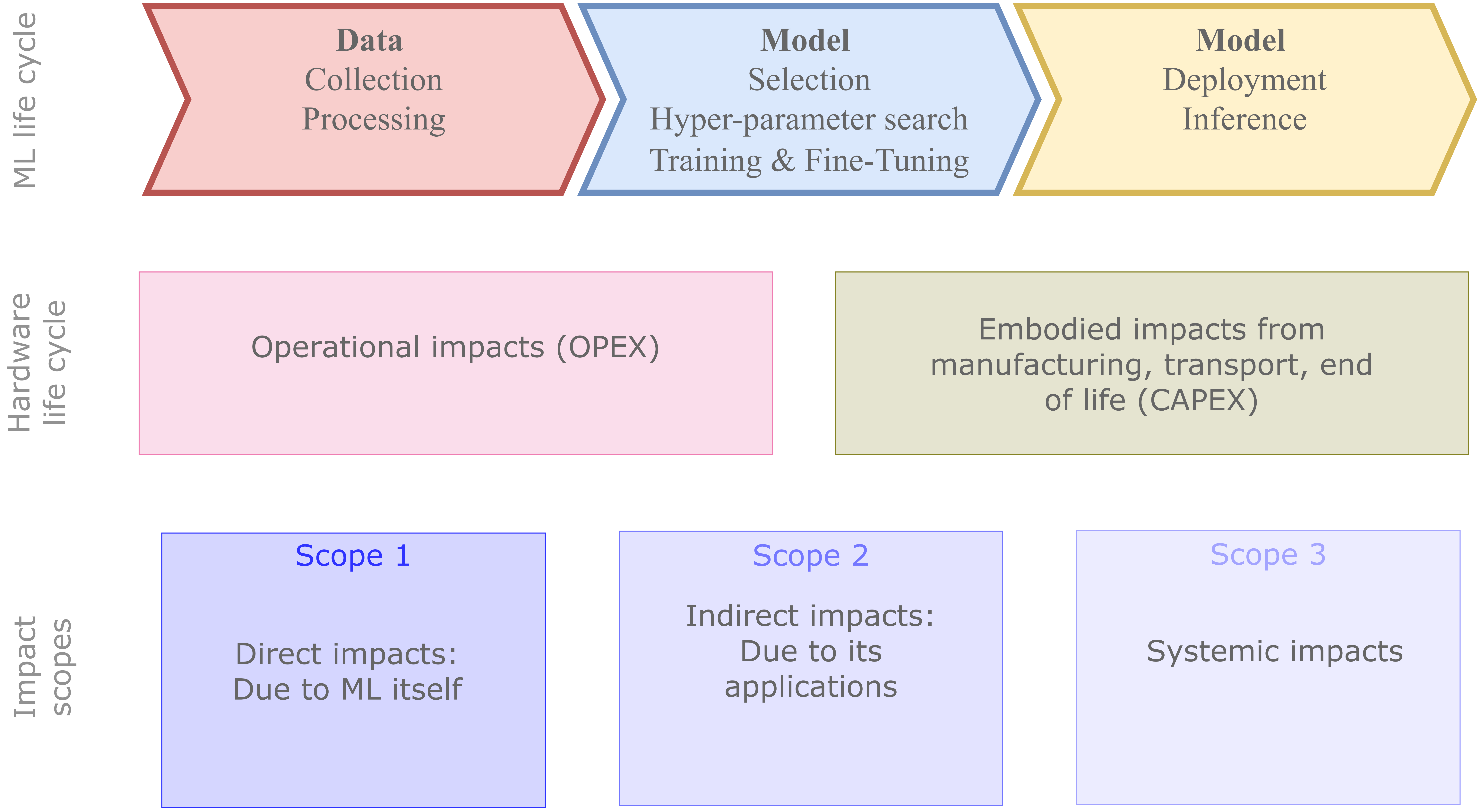




Estimates for global ICT’s carbon footprint in 2020 according to 3 studies [Freitag2021]

HPC: High Performance Computing

The multiple facets of an ML service



Research questions & objectives

How can we accurately **report** impacts of ML services?

How can those impacts be reduced?

Can **decentralizing** computations reduce the impacts?

Design a **methodology** to assess the impacts of training a ML model

Show its **versatility** by applying it on various models and infrastructures

Provide a better **understanding** of the impacts of ML training infrastructures

Compare a **Supercomputer** and an **Edge** node on model training

Contents

Existing literature on ML service impact assessments

		Alert on the carbon footprint of NLP training [Strubell2019]	Same on other tasks [Henderson2020, Patterson2020, Wu2022]	Federated Learning/ML at the Edge [Savazzi2021, Qiu2024, Patterson2024]	Addition of the embodied footprint [Ligozat2022, Dodge2022, Luccioni2023]	ML deployment [Patterson2022, Wu2022, Luccioni2023]	Our methodology
Impact Indicators	Electricity consumption						
	Carbon emission						
	Primary energy						✓
	Metal and mineral scarcity						✓
Hardware life cycle	OPEX						
	CAPEX				✓		✓
ML system scope	Supercomputer						
	Edge			✓			✓
ML life cycle	Training						
	Deployement					✓	✓

NLP: Natural Language Processing

Methodology

- Life Cycle Assessment (LCA)
  - ISO standards (14040 and 14044)
  - Standards from the International Telecommunication Union
  - Multi-criteria & multi phases

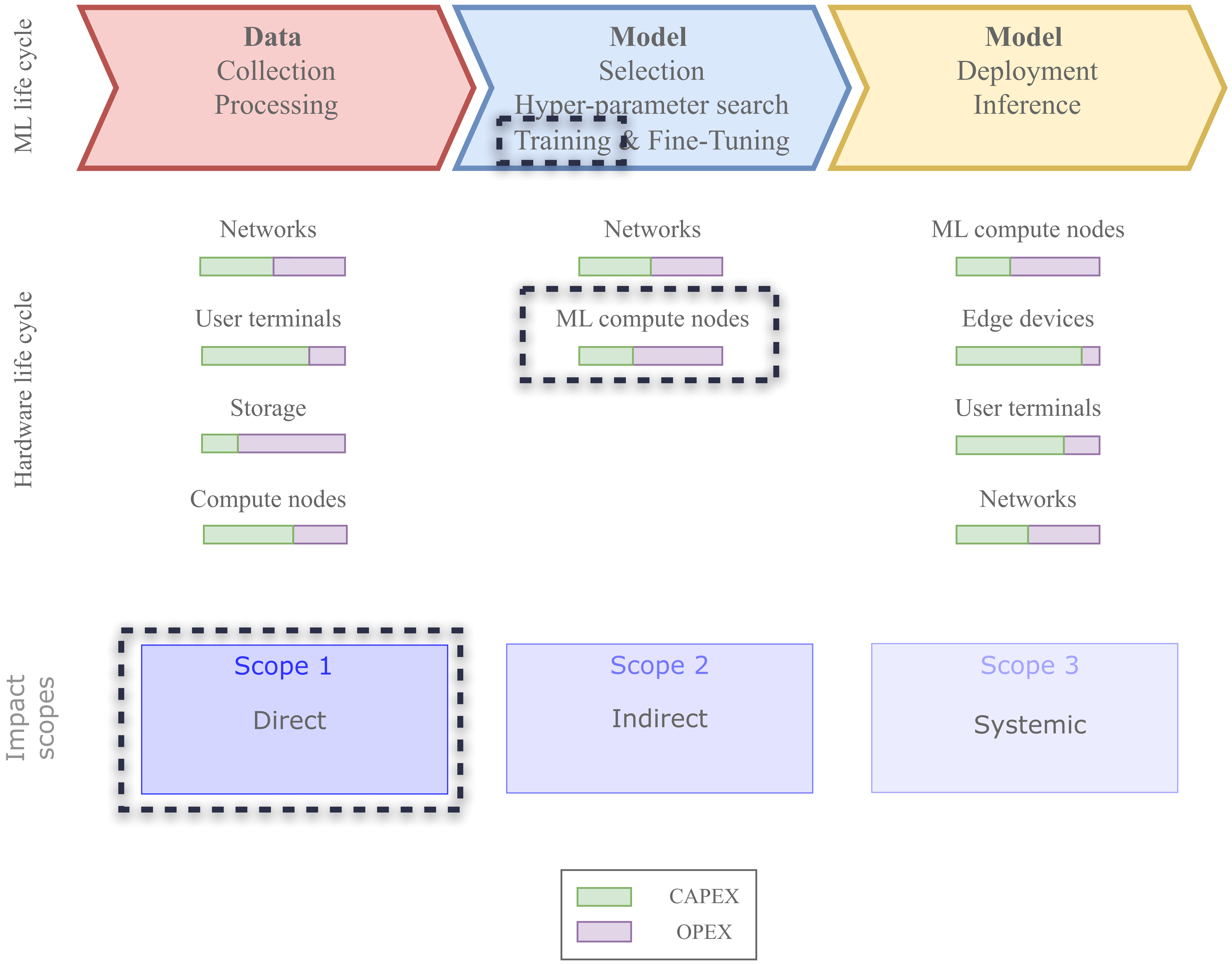
- We apply Attributional LCA to ML training

Methodology

- Life Cycle Assessment (LCA)
  - ISO standards (14040 and 14044)
  - Standards from the International Telecommunication Union
  - Multi-criteria & multi phases

- We apply Attributional LCA to ML training

- Scope



Methodology

- Life Cycle Assessment (LCA)
  - ISO standards (14040 and 14044)
  - Standards from the International Telecommunication Union
  - Multi-criteria & multi phases
- We apply Attributional LCA to ML training
- Scope
  - The CAPEX and OPEX impacts of the ML compute nodes during the training phase
- Functional Unit

Train the Model on the Dataset until the Quality Target is reached

Methodology

- Life Cycle Assessment (LCA)
  - ISO standards (14040 and 14044)
  - Standards from the International Telecommunication Union
  - Multi-criteria & multi phases
- We apply Attributional LCA to ML training
- Scope
  - The CAPEX and OPEX impacts of the ML compute nodes during the training phase
- Functional Unit
  - Train the model on the dataset until the quality target is reached
- Impacts
  - **Primary Energy** (PE) measured in mega joule (MJ)
  - **Global Warming Potential** (GWP) measured in equivalent CO_2 emission ($kg.CO_2.eq$)
  - **Abiotic Depletion Potential** (ADP) of minerals and metals measured in equivalent antimony (kg.Sb.eq)

Methodology

Operational phase (OPEX)

Embodied phase (CAPEX)

Evaluated Metrics

OPEX:

- Power measurement
  - Of each components
  - With a frequency high enough to capture evolution

CAPEX:

- Embodied impacts of each component of the ML compute node
- Time-based allocation

How

OPEX:

- Software-based power meters
- Electricity impact factors

CAPEX:

- LCA databases

Reproducibility

OPEX:

- Repeatability of experiments
- Fixed and controlled environments

CAPEX:

- Public databases

LCA: Life Cycle Assessment

Contents

Infrastructures

HPC

Apollo Node
From the
Champollion
cluster

EDGE

Nvidia Jetson
AGX Xavier

Hewlett Packard
Enterprise

NVIDIA
Grid'5000

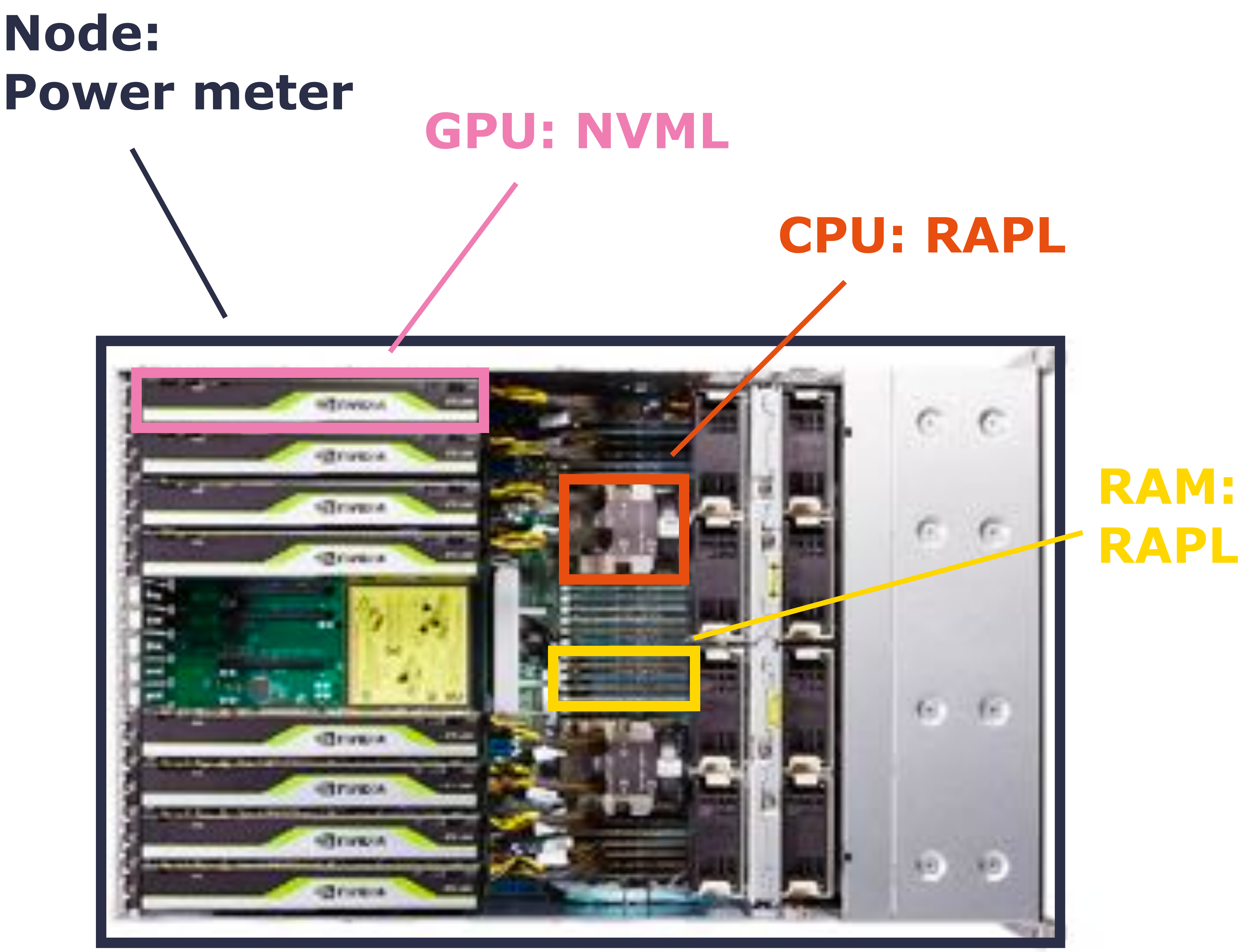
HPC: High Performance Computing

Infrastructures

	HPC: APOLLO	EDGE: JETSON
Node model	Apollo 6500 Gen10	Nvidia Jetson AGX Xavier
FL32 performance (FLOP/S)	$125 * 10^{12}$	$1.41 * 10^{12}$
GPU model	NVIDIA A100-SXM-80GB	NVIDIA GV10B, Volta architecture
Number of GPU per node	8	1
GPU TDP (W)	400	
CPU model	AMD EPYC 7763, 64 cores	Nvidia Carmel (Carmel), aarch64, 8 cores
Number of CPU per node	2	1
CPU TDP (W)	280	
Memory	1 TB	32 GB
Installation year	2022	2023
Available thought	HPE local network - slurm	Grid'5000 Estats cluster - OAR
Power meter	HPE iLO 5, RAPL, NVML	Jetson-stats
Node TDP (W)	3760 (GPUs + CPUs)	30
Energy efficiency (FLOP/S/W)	$3.3 * 10^{10}$	$2.3 * 10^{10}$

GPU: Graphic Processing Unit, **CPU**: Central Processing Unit, **TDP**: Thermal Design Power

Measuring the electricity consumption of a node



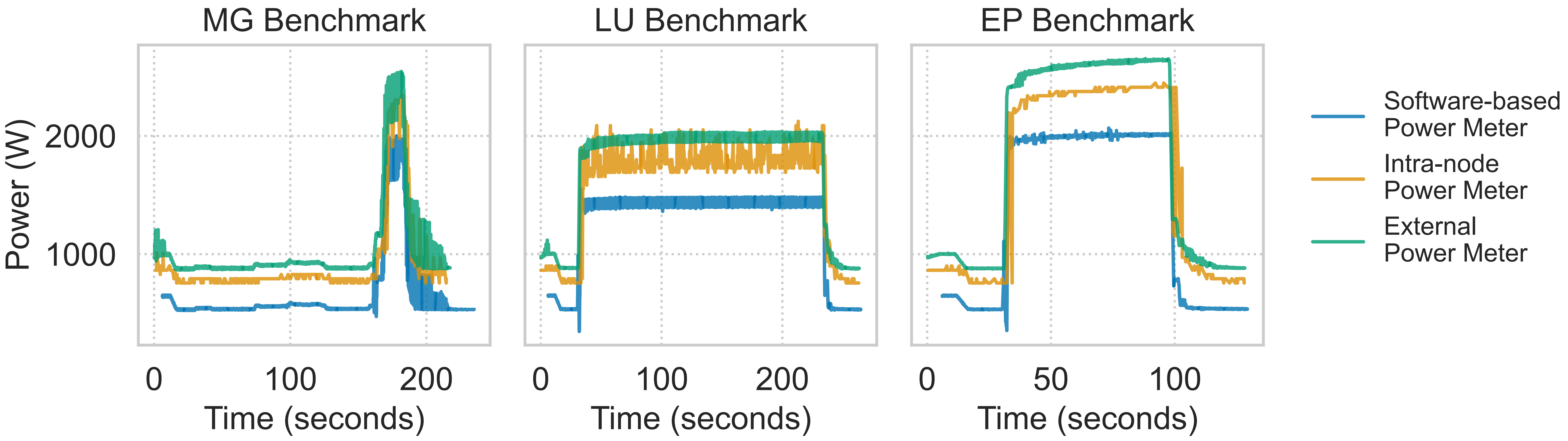
Example of a GPU node

- Power meters
  - Outside of node
  - Processor power management libraries
- Software-based power meters
  - Proven to be accurate
  - Available on most nodes
- Which one is better for our use case?

RAPL: Running Average Power Limit, **NVML**: Nvidia Management Library

A quantitative comparison

- Compared **software-based** power meters
- Evaluated them against **external** power meters and found a significant and non-constant offset



Comparison of power meters on 3 applications of the NAS benchmark on the Gemini cluster of Grid'5000.

And a qualitative comparison

- Not suitable for our use case because of lack of
  - Configurability (frequency) for node with GPUs
  - Versatility: power management librairies depend on infrastructures

Mathilde Jay, Vladimir Ostapenco, Laurent Lefèvre, Denis Trystram, Anne-Cécile Orgerie, Benjamin Fichel. An experimental comparison of software-based power meters: focus on CPU and GPU. CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing, May 2023, Bangalore, India

49 citations

2578 download

Open Research Objects

Reusable/Research Objects Reviewed

A software-based power meter for Apollo and Jetson

ALUMET

- **Time series** of GPU consumption as well as CPU and RAM
- **Configurable** acquisition frequency
- **Adaptive** to edge and HPC architectures
- **Modularity** (e.g. new data sources can be added by plugins)
- **Lightweight** (no significant slowdown of HPC benchmarks for RAPL+CSV)

alumet.dev
By Guillaume Raffin, BULL SAS, CNRS, INRIA, Grenoble INP-UGA.
Licensed under the EUPL-1.2 or later.
~ 12 300 lines
Written in Rust

Estimating the embodied impacts with Boavizta

Boavizta

Datavizta API: Multi-indicators/Multi-phase

- Aggregates data from various databases
  - ADEME carbon database
  - Manufacturer product carbon footprints
  - Semiconductor LCA
- For GPUs, need specifications
  - Die size
  - Printed Circuit Board area
  - Memory density and capacity

Total embodied or CAPEX impact of Jetson and Apollo

	GWP (kg CO2eq)	ADP (kg Sbeq)	PE (MJ)
Jetson	87	0.03	1 254
Apollo	3 858	0.28	49 660
	x 44	**x 10**	**x 39**

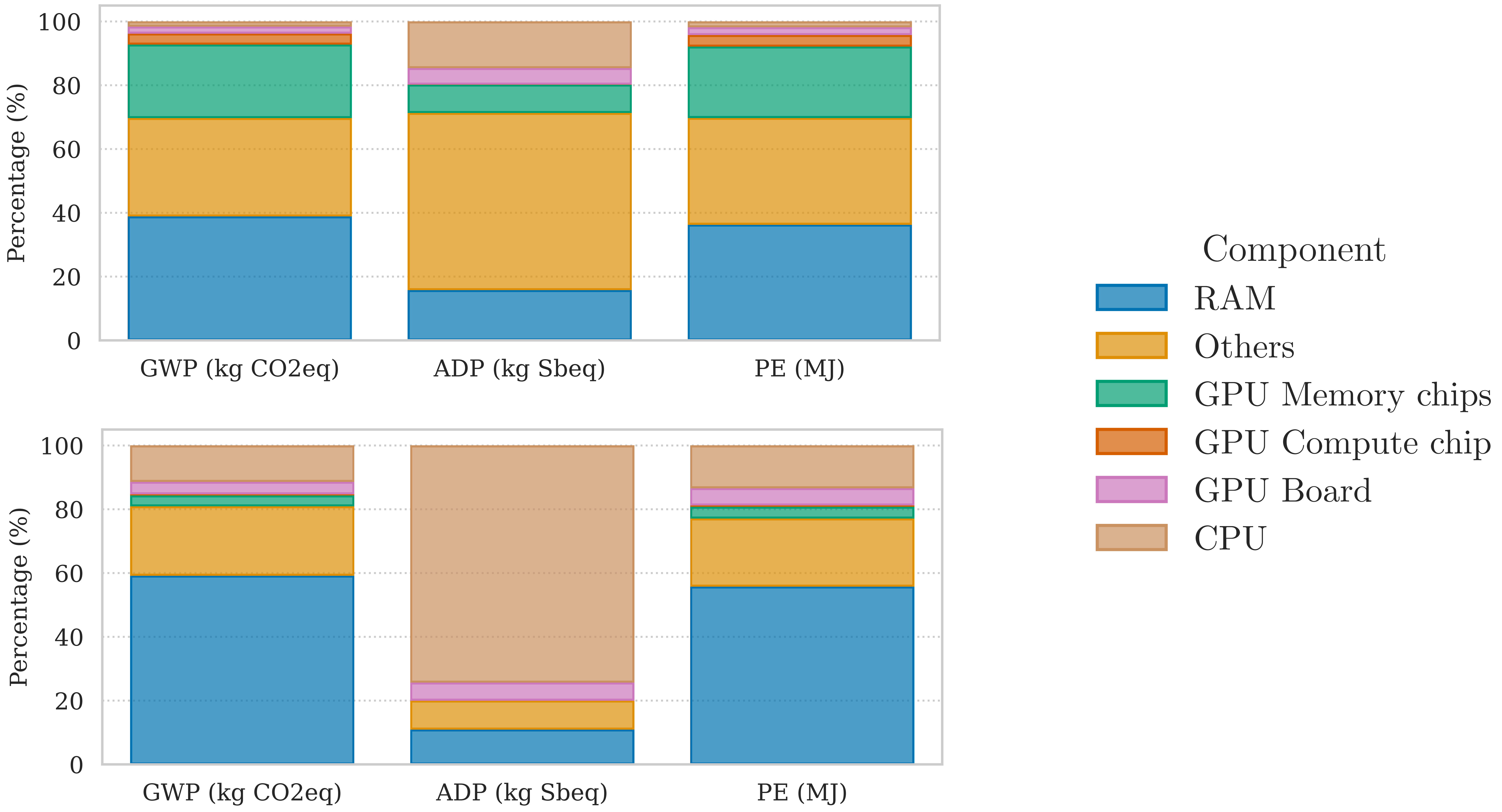
The embodied impact of one Apollo node is higher than the embodied impact of one Jetson node.

GWP: Global Warming Potential, **ADP**: Abiotic Depletion Potential, **PE**: Primary Energy

Memory dominates the embodied impacts

HPC: APOLLO

EDGE: JETSON



Share of each compute node component in the total embodied impacts.

GWP: Global Warming Potential, **ADP**: Abiotic Depletion Potential, **PE**: Primary Energy

Contents

MLPerf Benchmark

Area	Benchmark	Model	Dataset	Quality Target
Vision	Image classification	ResNet-50 v1.5	ImageNet	75.90% classification
Vision	Image segmentation	3D U-Net	KiTS19	0.908 Mean DICE score
Vision	Object detection	Mask R-CNN	COCO	0.377 Box min AP and 0.339 Mask min AP
Language	Speech recognition	RNN-T	LibriSpeech	0.058 Word Error Rate
Language	Natural Language Processing	BERT-large	Wikipedia 2020/01/01	0.72 Mask-LM accuracy
Commerce	Recommendation	DLRM	1TB Click Logs	0.8025 AUC

HPC: APOLLO

- Code provided by MLPerf
- Hyper-parameter search done by HPE

EDGE: JETSON

- Installation of software
- Memory limitations
- Long executions

Focus on ResNet-50

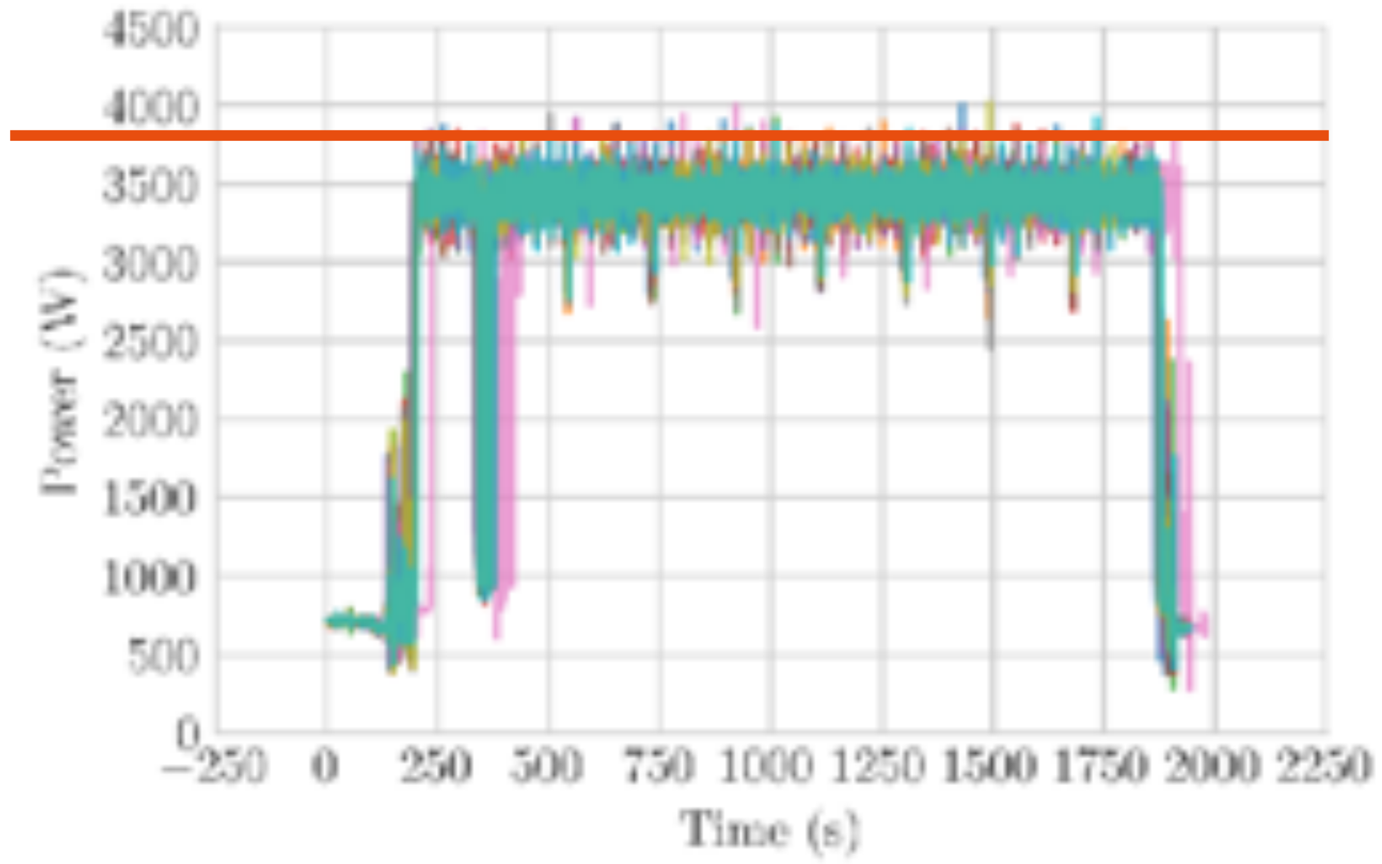
Power profiles can vary significantly

HPC: APOLLO

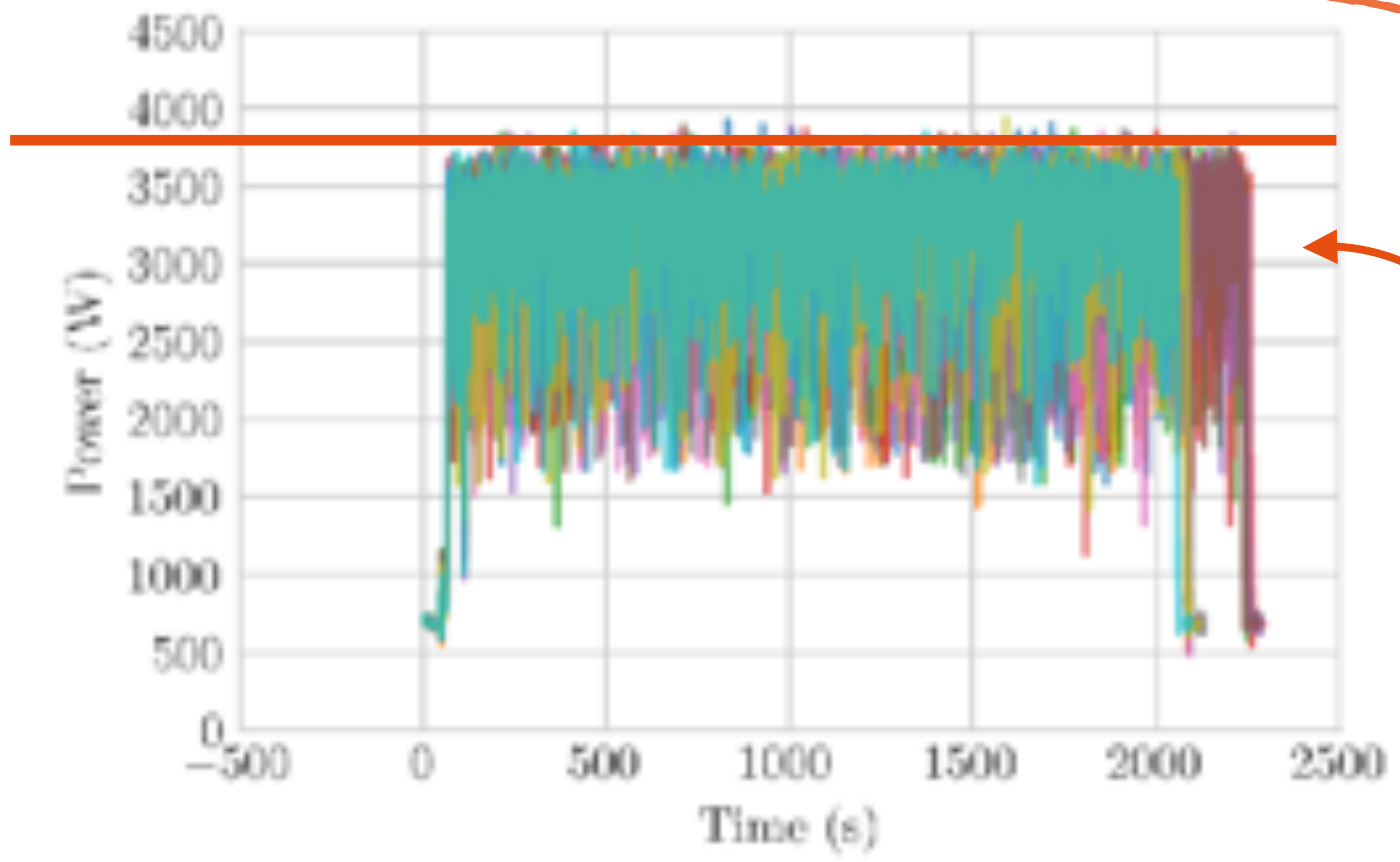
TDP = 3760 W

Power profile = Power consumed by the 8 GPUs and the 2 CPUs as measured by NVML and RAPL.

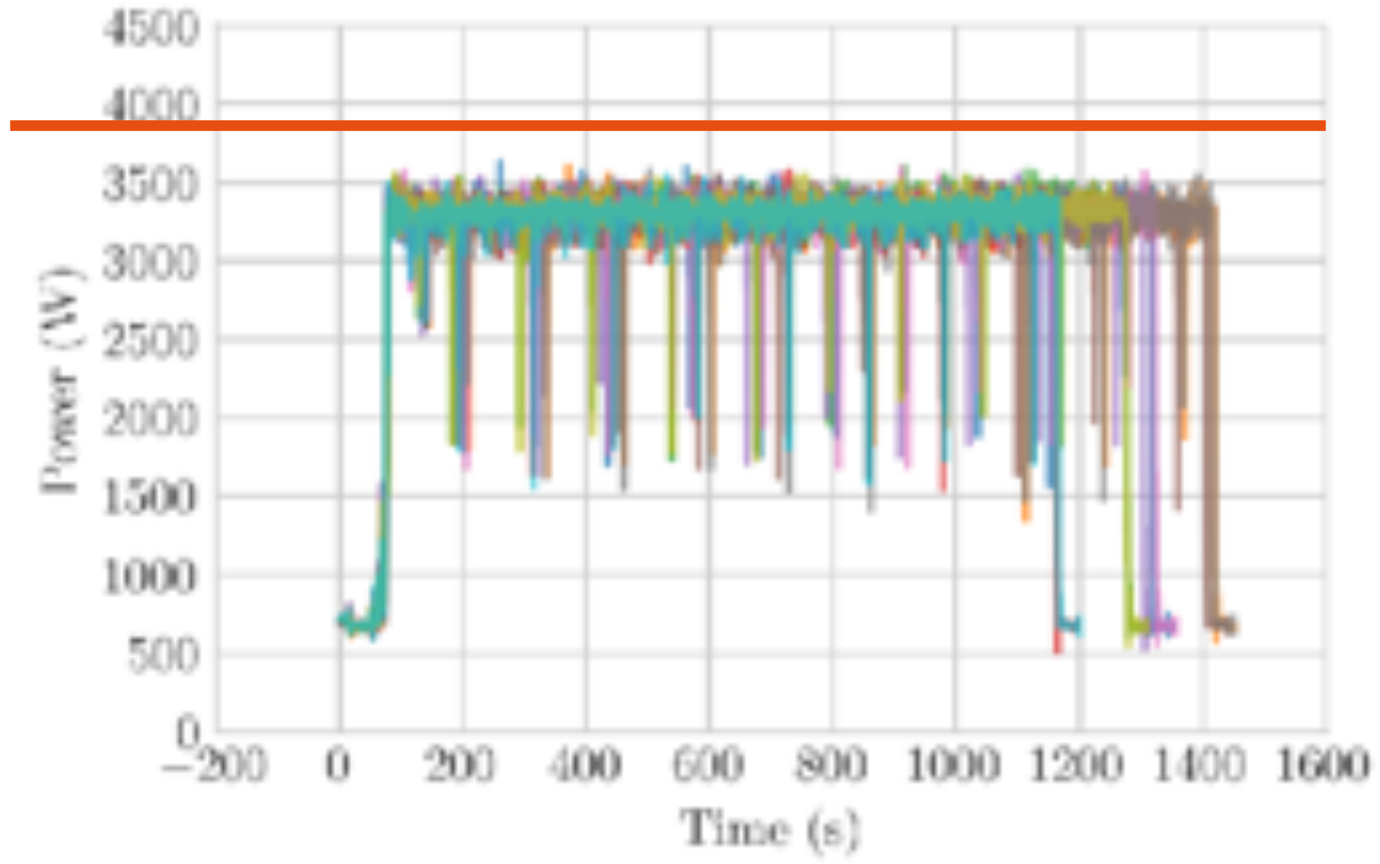
1 color = 1 repetition of the training



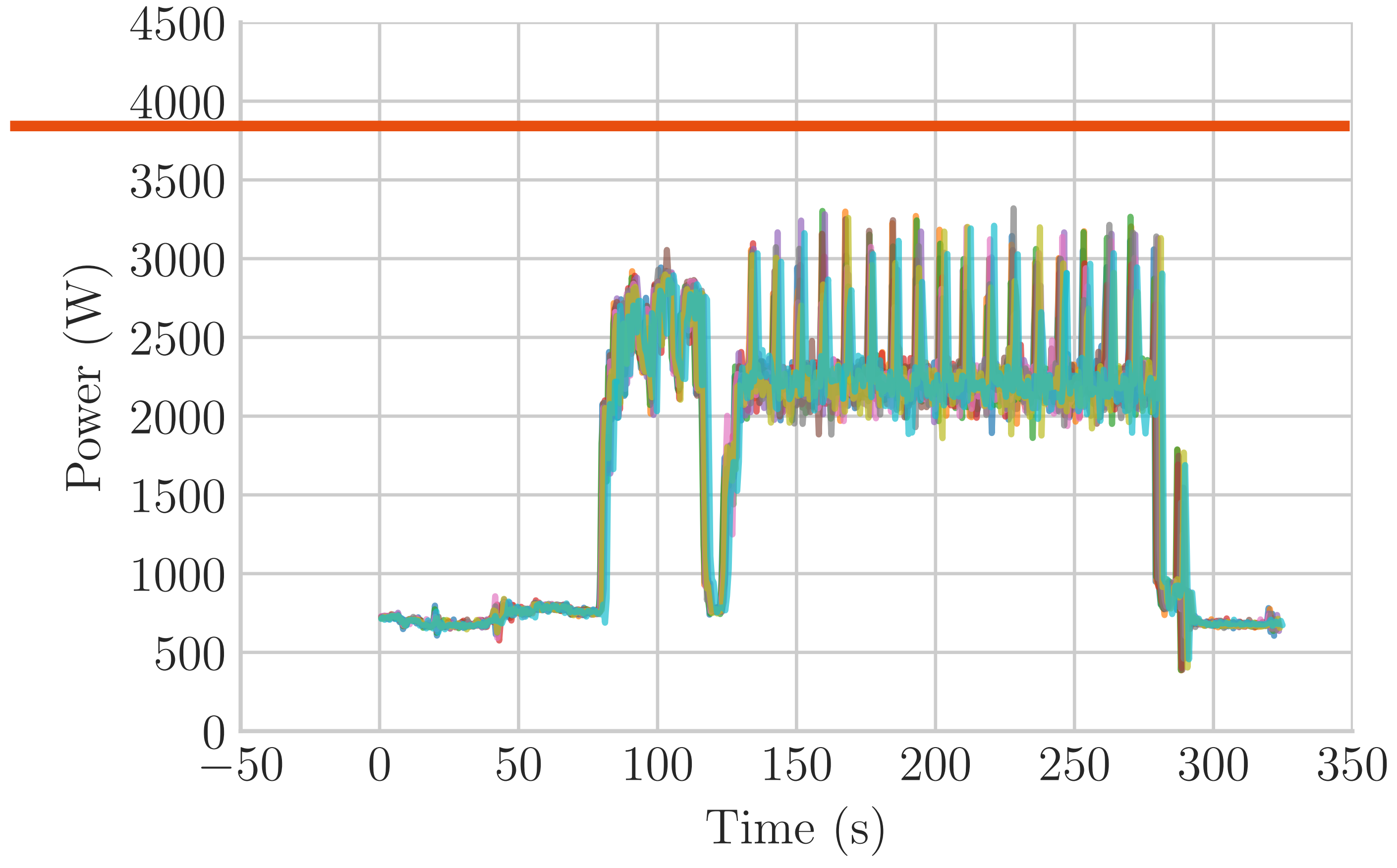
Power profile of ResNet-50 FU



Power profile of RNN-T FU



Power profile of Mask R-CNN FU

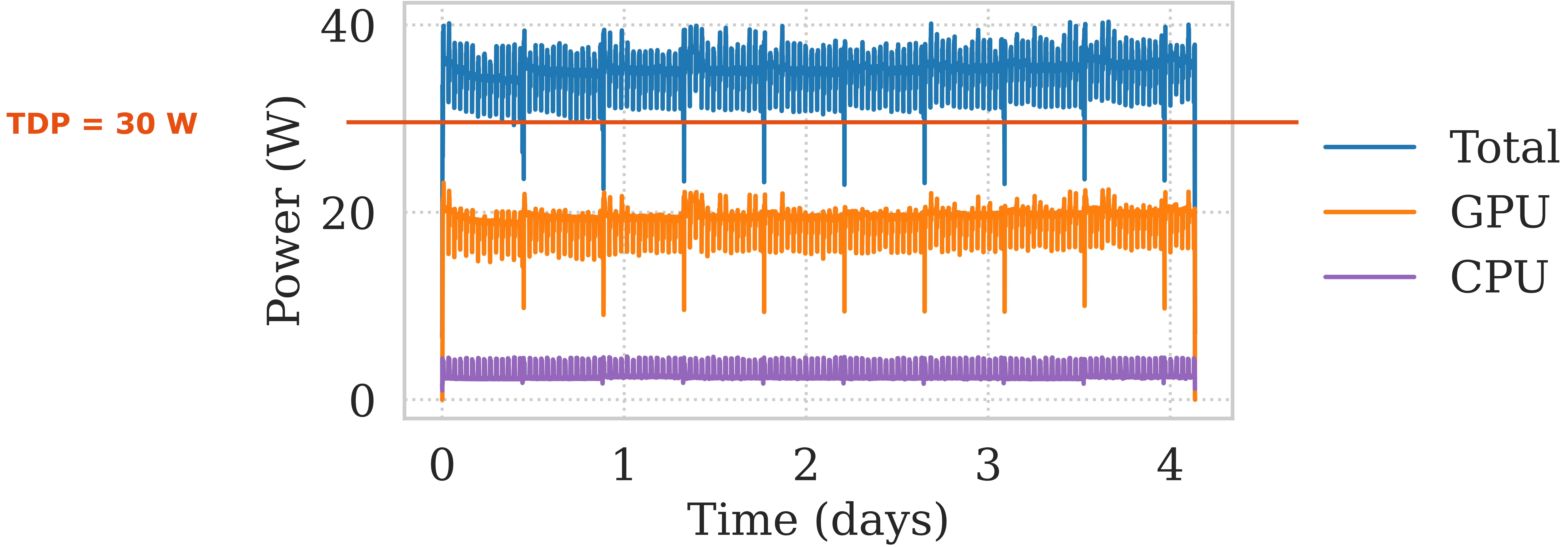


Power profile of DLRM FU

TDP: Thermal Design Power, **FU**: Functional unit

Power consumption can be higher than the TDP

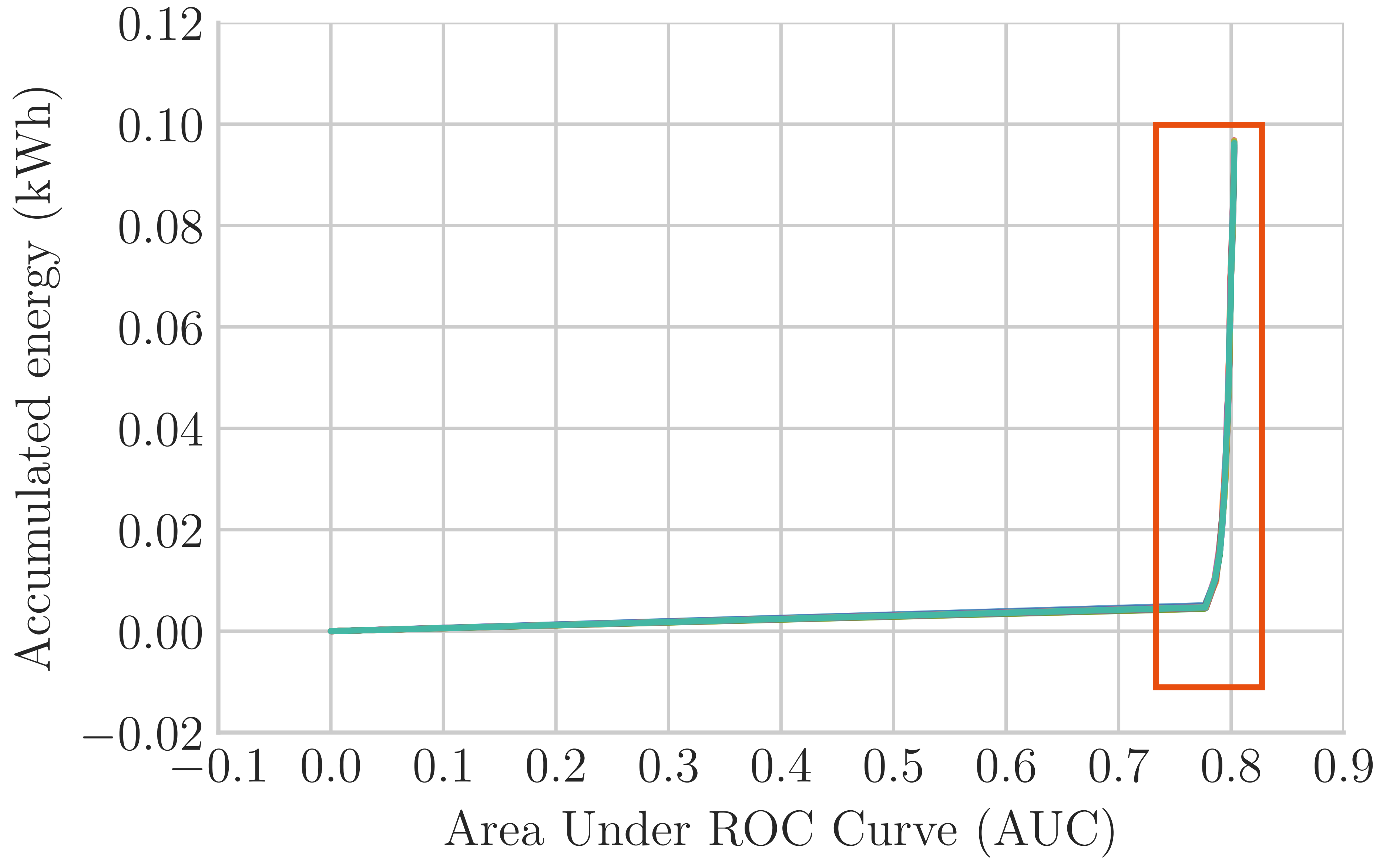
EDGE: JETSON



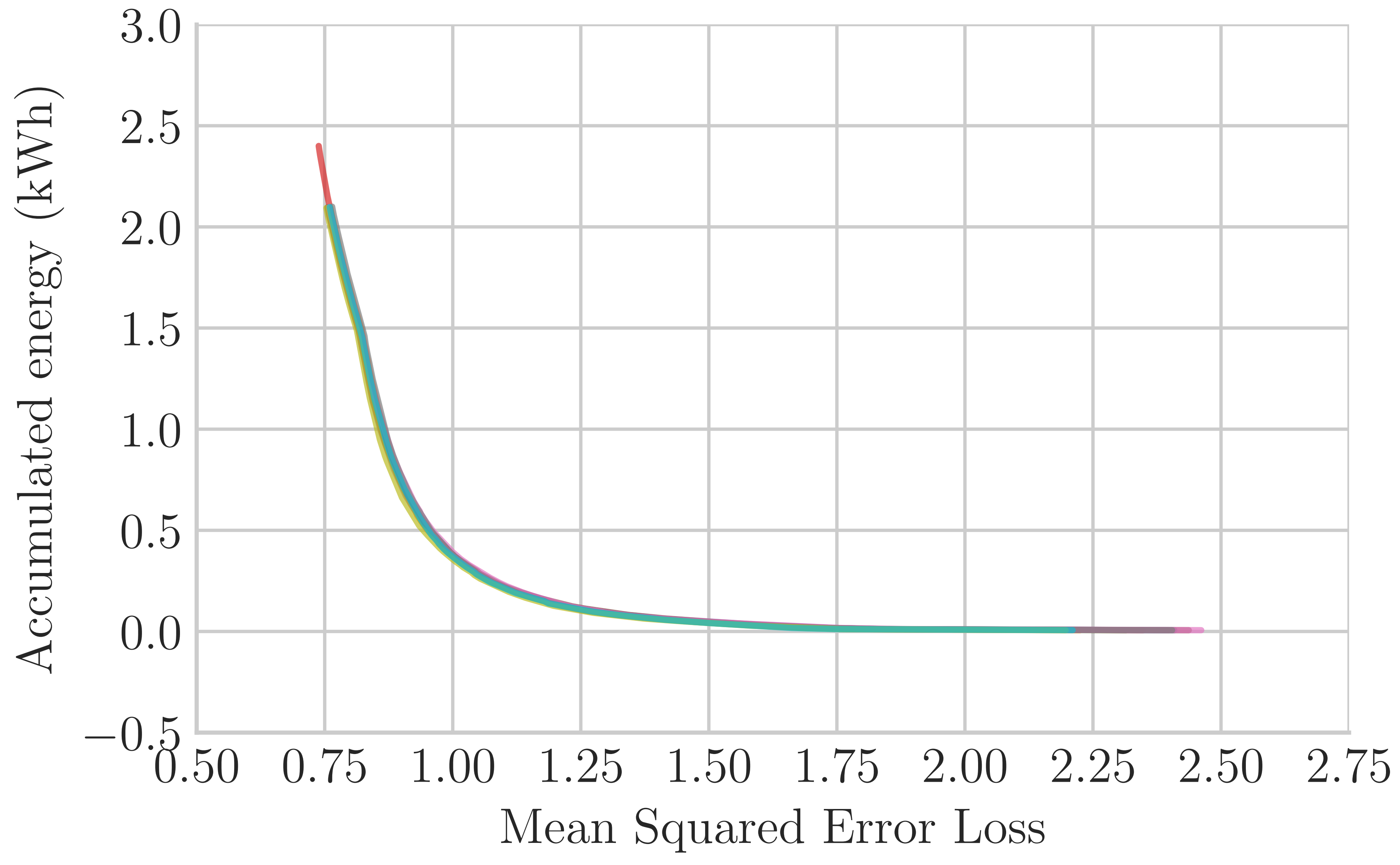
Power profile of ResNet-50 FU.

TDP: Thermal Design Power, **FU**: Functional unit

Most energy can be spent on the very last quality point



Energy required to reach each quality metric point for the DLRM FU



Energy required to reach each quality metric point for the Mask R-CNN FU

HPC: APOLLO

FU: Functional unit

The ResNet FU impacts are shared between phases

$$I_{Embodied,FU} = \frac{T_{FU}}{T_{node}} * I_{Embodied,Node}$$

$$I_{FU} = IF_{elec} * E_{FU} + I_{Embodied,FU}$$

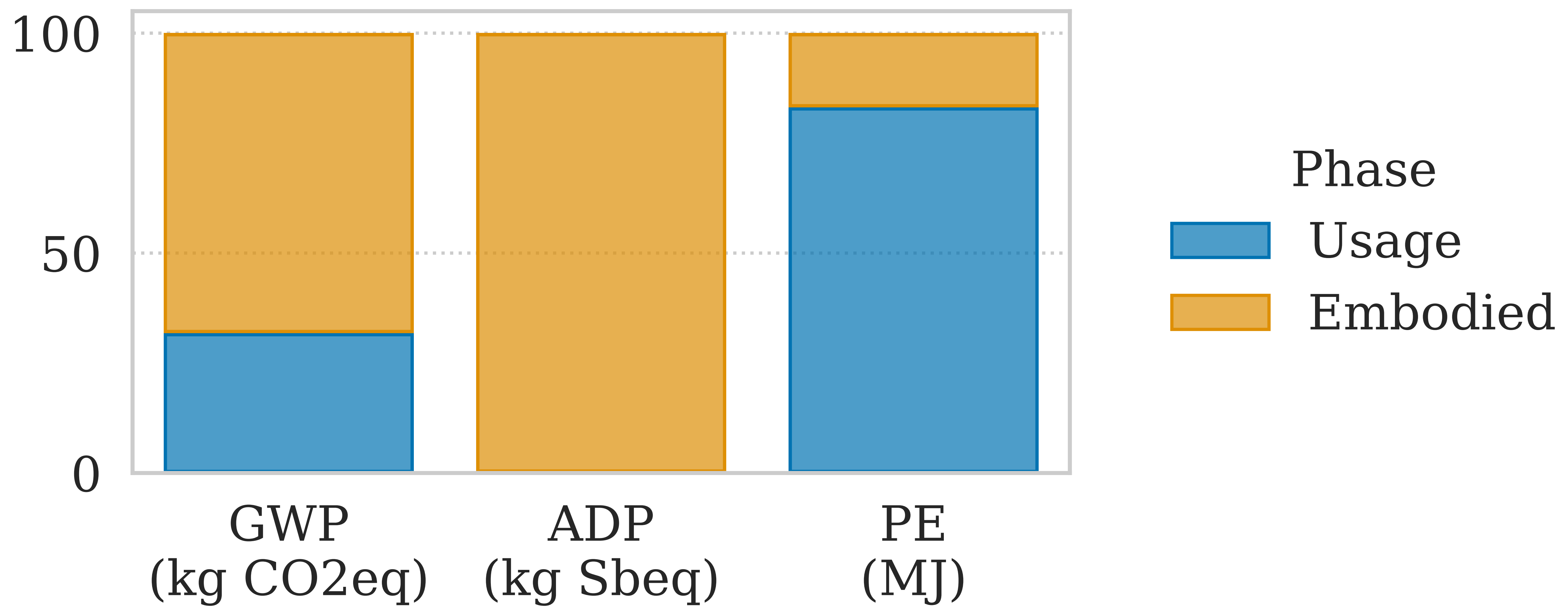
I : Environmental Impact (expressed in kg CO2 eq, kg Seb, and MJ)

T : Use time of the equipment

IF_{elec} : Electricity mix Impact Factor

E : Electricity consumption

EDGE: JETSON

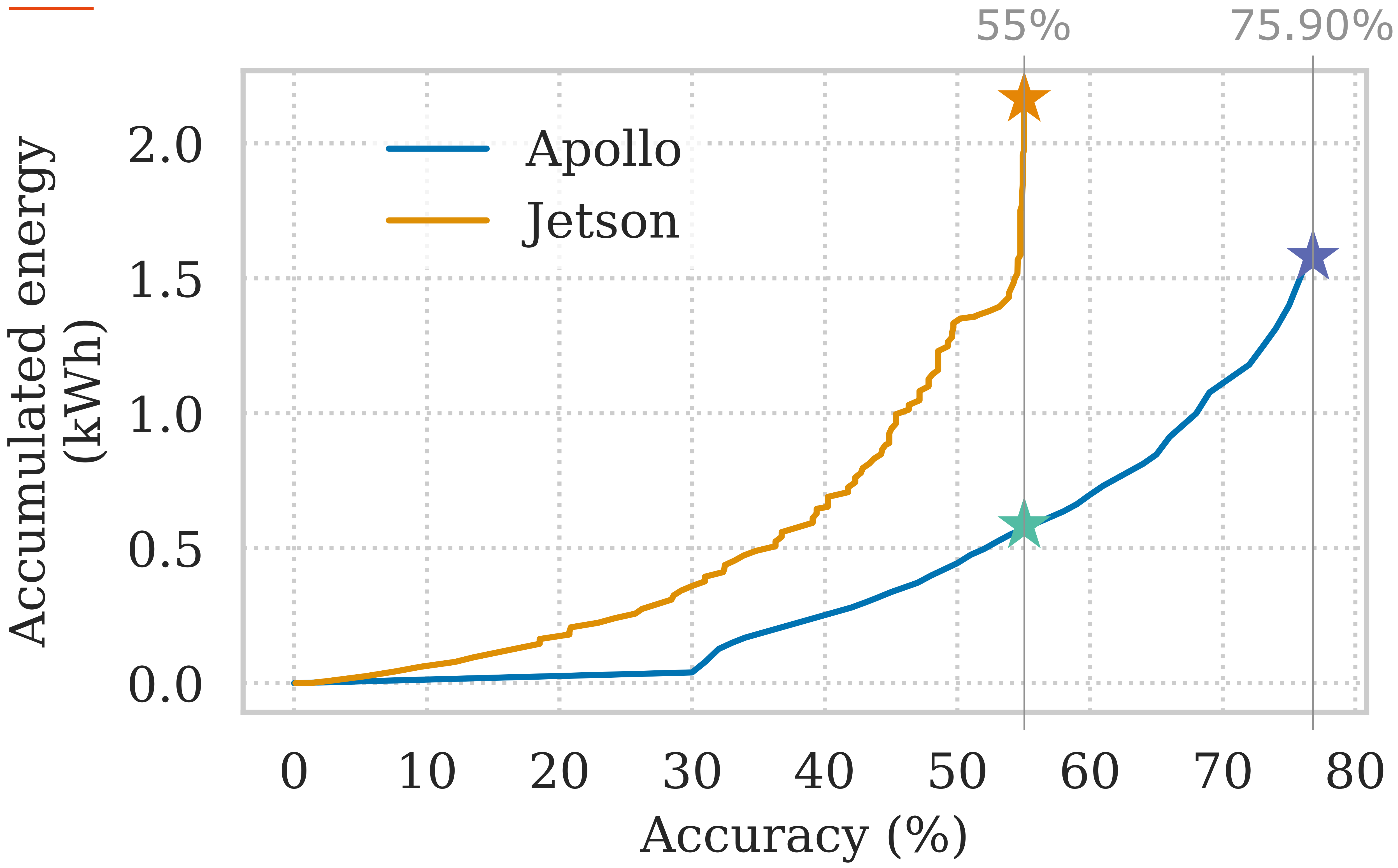


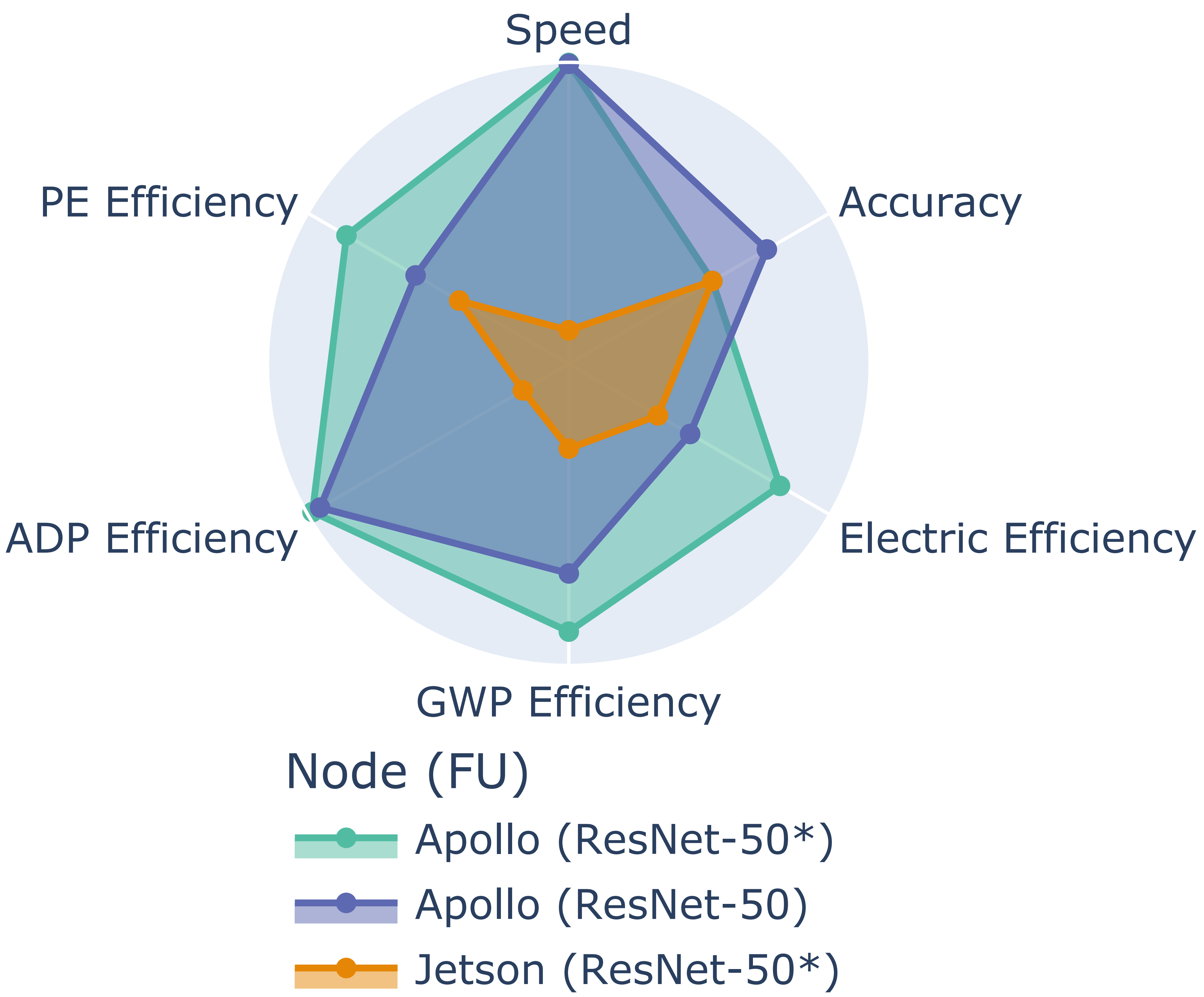
Share of usage and embodied phase in the total impacts of the ResNet-50 * FU on the **Jetson** node

GWP: Global Warming Potential, **ADP**: Abiotic Depletion Potential, **PE**: Primary Energy

Contents

Apollo outperforms Jetson on the performance and environmental criteria





ResNet-50 FU: *Train ResNet-50 on ImageNet until it achieves a **75.90%** classification score*

ResNet-50* FU: *Train ResNet-50 on ImageNet until it achieves a **55%** classification score*

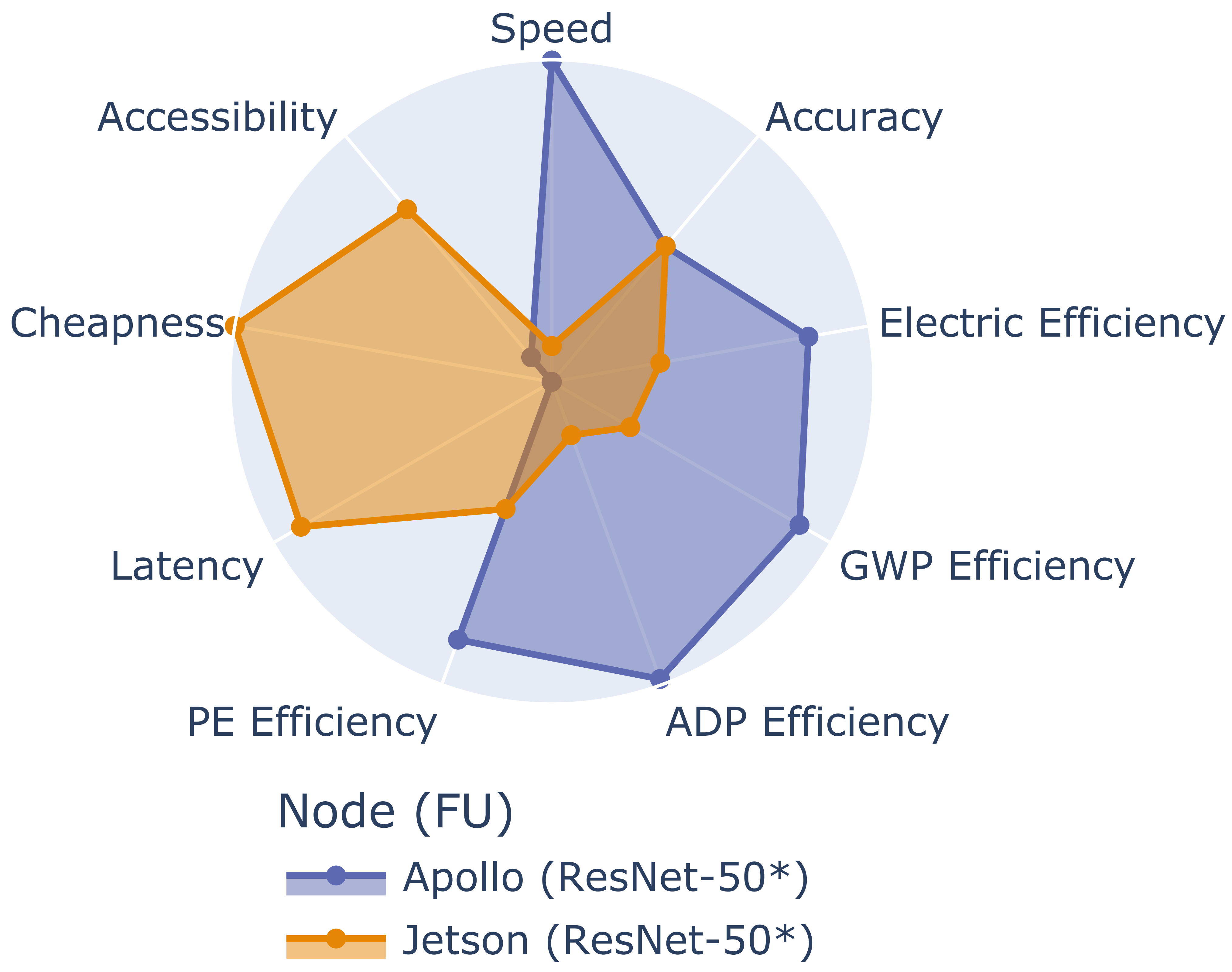
Multi-criteria comparison including performance metrics and environmental metrics

GWP: Globale Warming Potential, **ADP**: Abiotic Depletion Potential, **PE**: Primary Energy

Comparing infrastructures requires more criteria

- Jetson was designed for the edge and its associated constraints
  - Price
  - Latency

- Edge computing can be an opportunity for constraining computations



Multi-criteria comparison including performance, environmental, and qualitative metrics

Submitted & in review.

Contents

Advantages of our approach

- Versatility
  - Can do both **operational** and **embodied** assessments on most infrastructures
- Reproducibility
  - Rely on **open** databases and **open source** measuring tool
  - Code is available and reusable
- Insightfulness
  - Power profile, accuracy/power tradeoff, details of component embodied footprint provide a **better understanding** and the possibility to find **actionable reduction plans**
- Enables a **fair comparison** between infrastructures and ML models

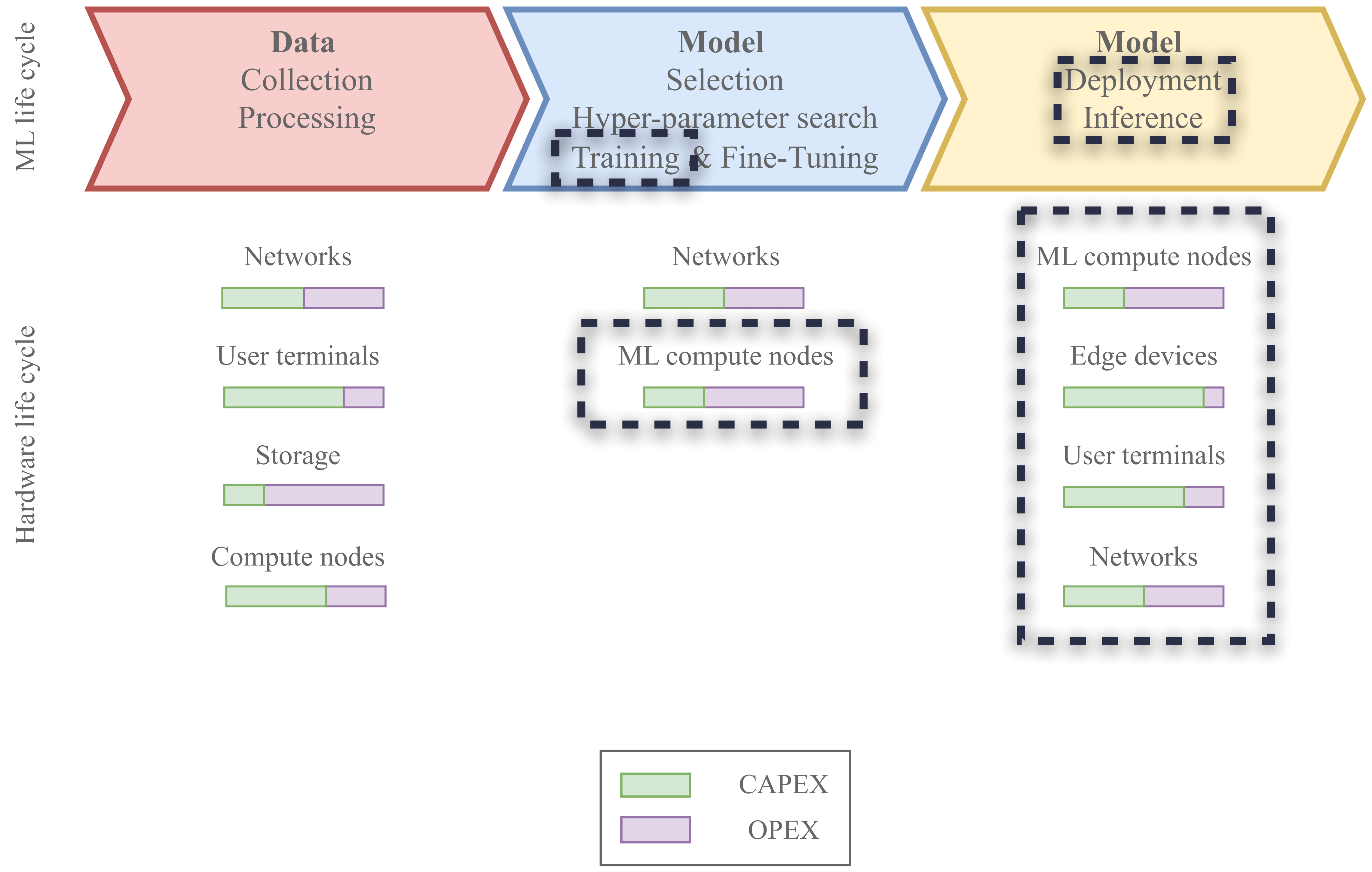
Limits

- Accuracy
  - Lack and **uncertainty** of LCA databases
  - **Offset** between software-based and external power meters not included
- Scope 2 and 3 outside of assessment
- Changes in learning
  - Focus on centralized training when other training approach are developed such as **Federated Learning**
  - Requires **replication** when trainings can last for days
- Other phases of ML life cycle
  - Focus on the training phase, omitting **deployment** and **data collection**

LCA: Life Cycle Assessment

Estimating the environmental impacts of a Generative AI service

Enlarge the scope

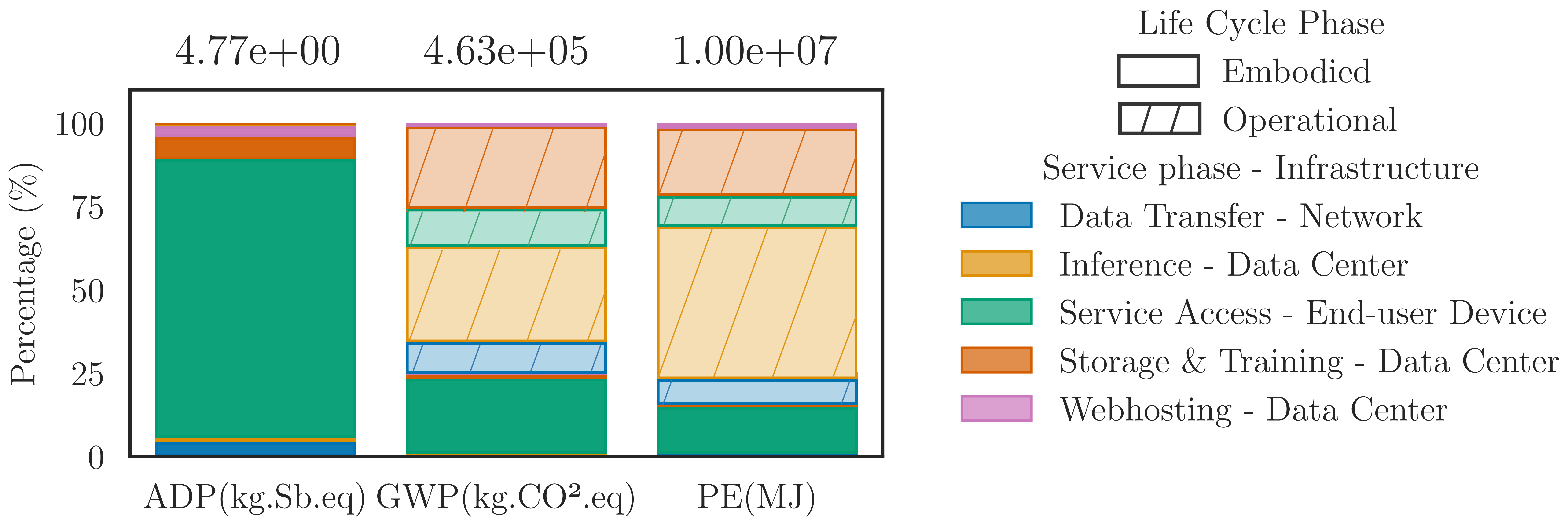


- Application: Stable Diffusion
  - Open source
  - **Deployed** online as a service
- Combinaison of
  - Measurements (Training, Inference)
  - Allocations from LCA databases
- Training to long to replicate
  - Measurement of a **fraction** of training
  - Linear **regression** from number of steps

Grid'5000

Each phase of the ML and hardware life cycles is significant

- The ADP impact is dominated by the embodied phase of hardware.
- The GWP and PE impacts come from the operational phase.
- The deployment causes most of the impacts.



Share (in percentage) of the life cycle phases of each digital infrastructure in the total impacts of the Stable Diffusion service for one year.

Adrien Berthelot, Eddy Caron, Mathilde Jay, Laurent Lefèvre. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. *CIRP LCE 2024 - 31st Conference on Life Cycle Engineering*, Jun 2024, Turin, Italy.

GWP: Global Warming Potential, **ADP**: Abiotic Depletion Potential, **PE**: Primary Energy

Contents

Conclusion

- Impacts of Machine Learning are booming when they should be reduced
- Need of reporting and controlling those impacts
- Methodology: LCA for training
  - 3 LCA impact indicators
  - Operational, ressource extraction, manufacturing & distribution phases
  - Insightfulness, Versatility, Reproducibility
- Multi-criteria comparison of an Edge Device and a Supercomputer
- A significant offset between power meters
- Quality target/energy tradeoff
- For bigger models, electricity consumption can be estimated
- Scope can be extended to other ML life cycle phases

Perspectives

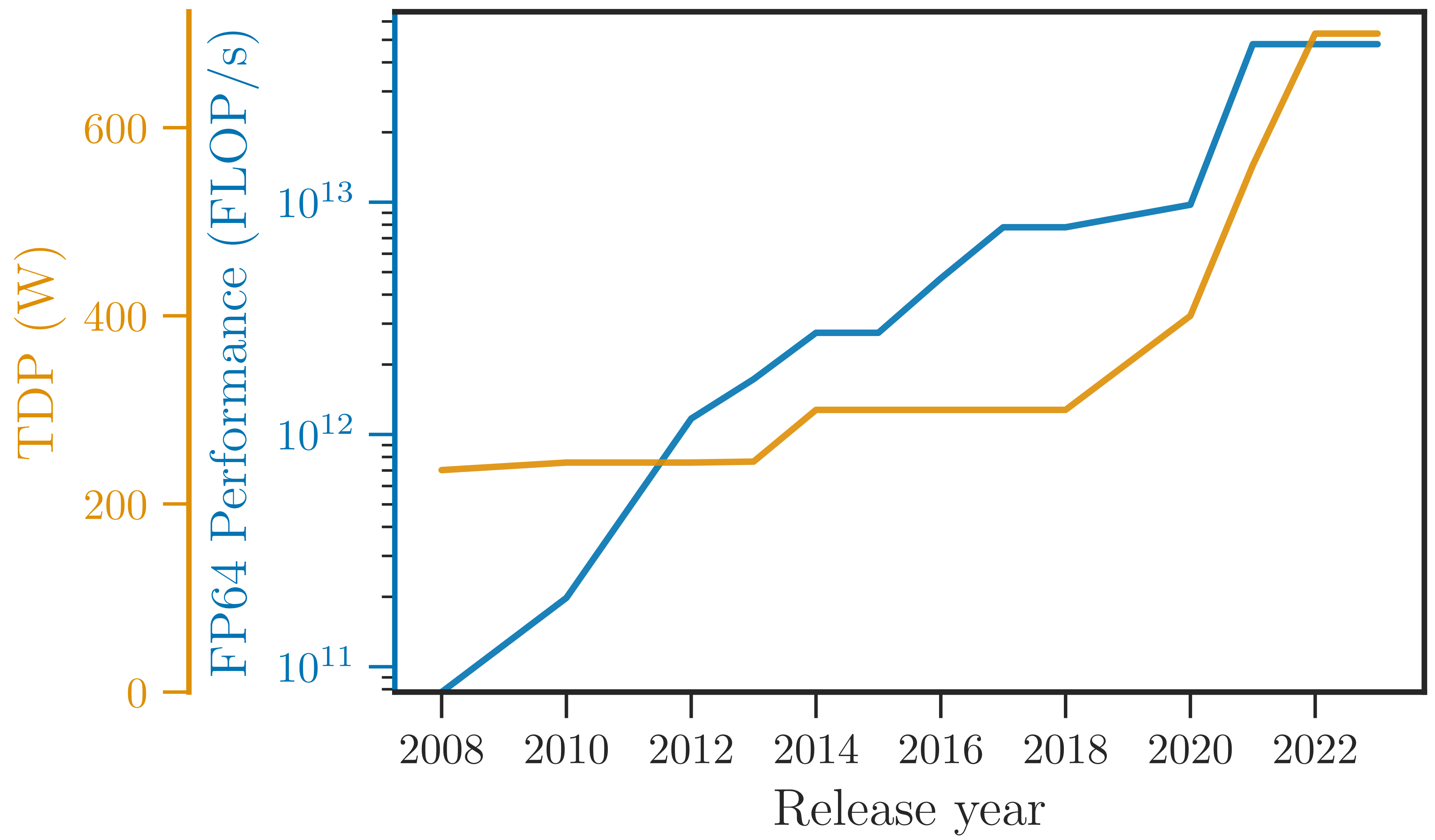
- Increasing the scope
  - Full node, storage, networks, cooling
  - Data collection
- Supercomputer-Edge scenario
  - Including network
- Consequential LCA
  - Encompasses indirect effects
  - But more complex to model
- Assessing the sustainability of an ML model
  - EU AI act: security, transparency, ethics

Appendix

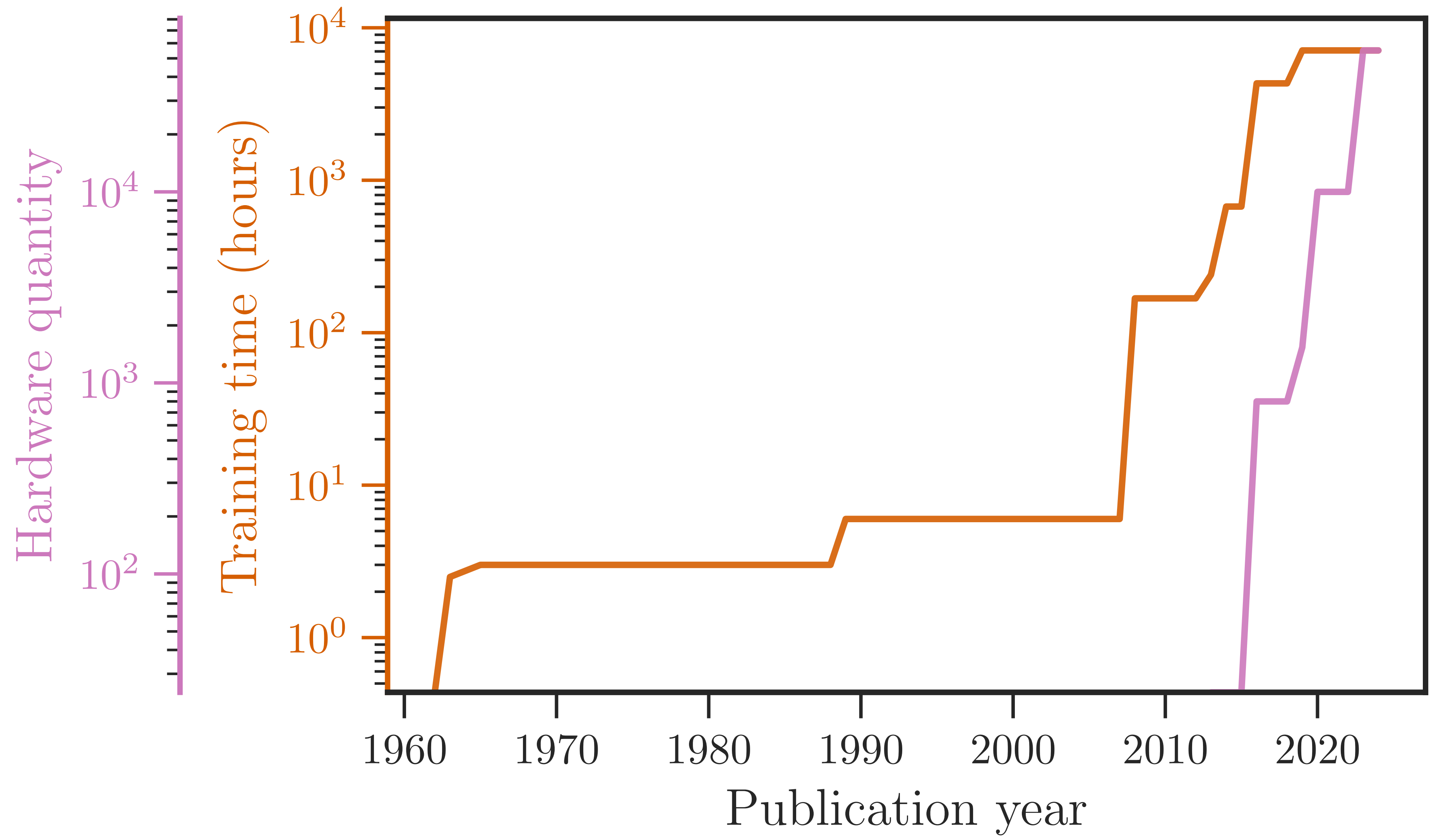
A versatile methodology for assessing
the electricity consumption and
environmental footprint of
machine learning training:
from supercomputers to edge devices

Context

Machine Learning and its materiality



Evolutions in Machine Learning specialized hardware (GPU, TPU) metrics, as the maximum value up to the given year [Hobbhahn2023].



Evolutions in Machine Learning Model and their training, as the maximum value up to the given year. [EpochAI2024]

TDP (Thermal Design Power): A hardware characteristic provided by the manufacturer that corresponds to the maximum amount of heat that can be generated by a component under a steady workload measured in watts.

Exponential increase in the electricity consumption of training Machine Learning models

Machine Learning services and its materiality

HPC data center | Networks | Edge data center | Networks | Edge Devices

Machine Learning and its materiality

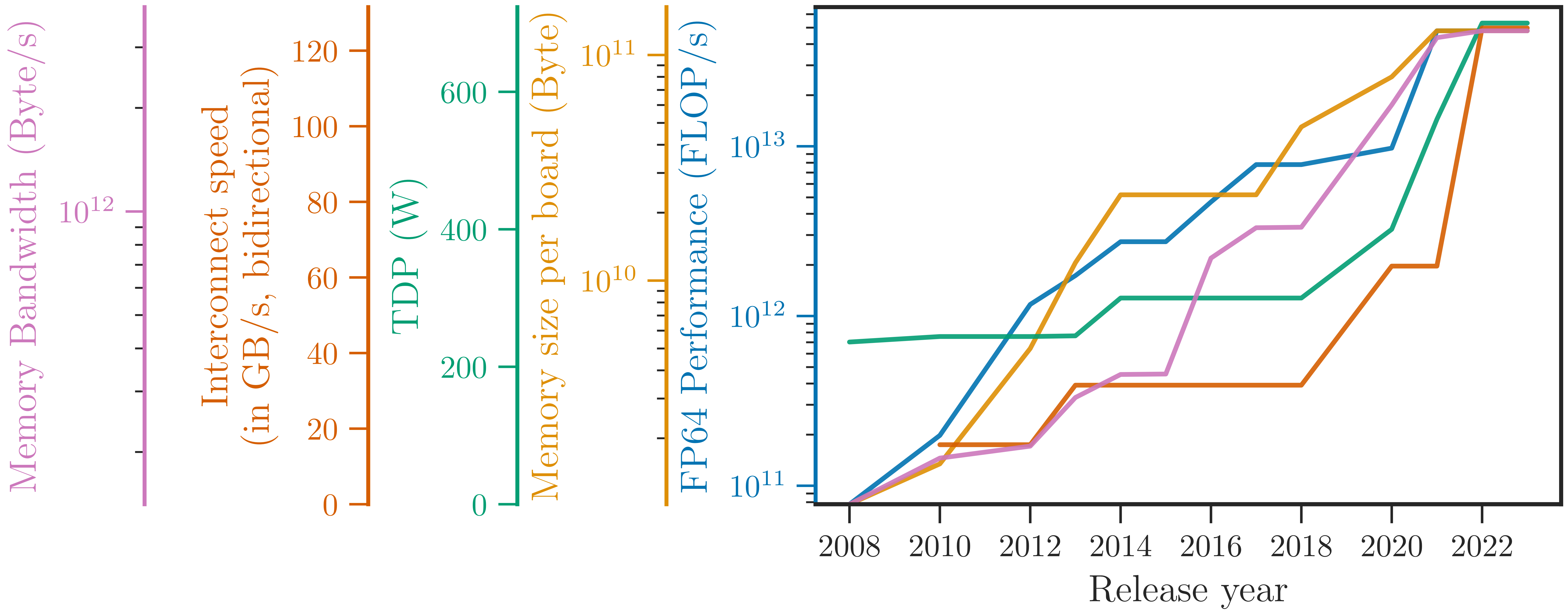
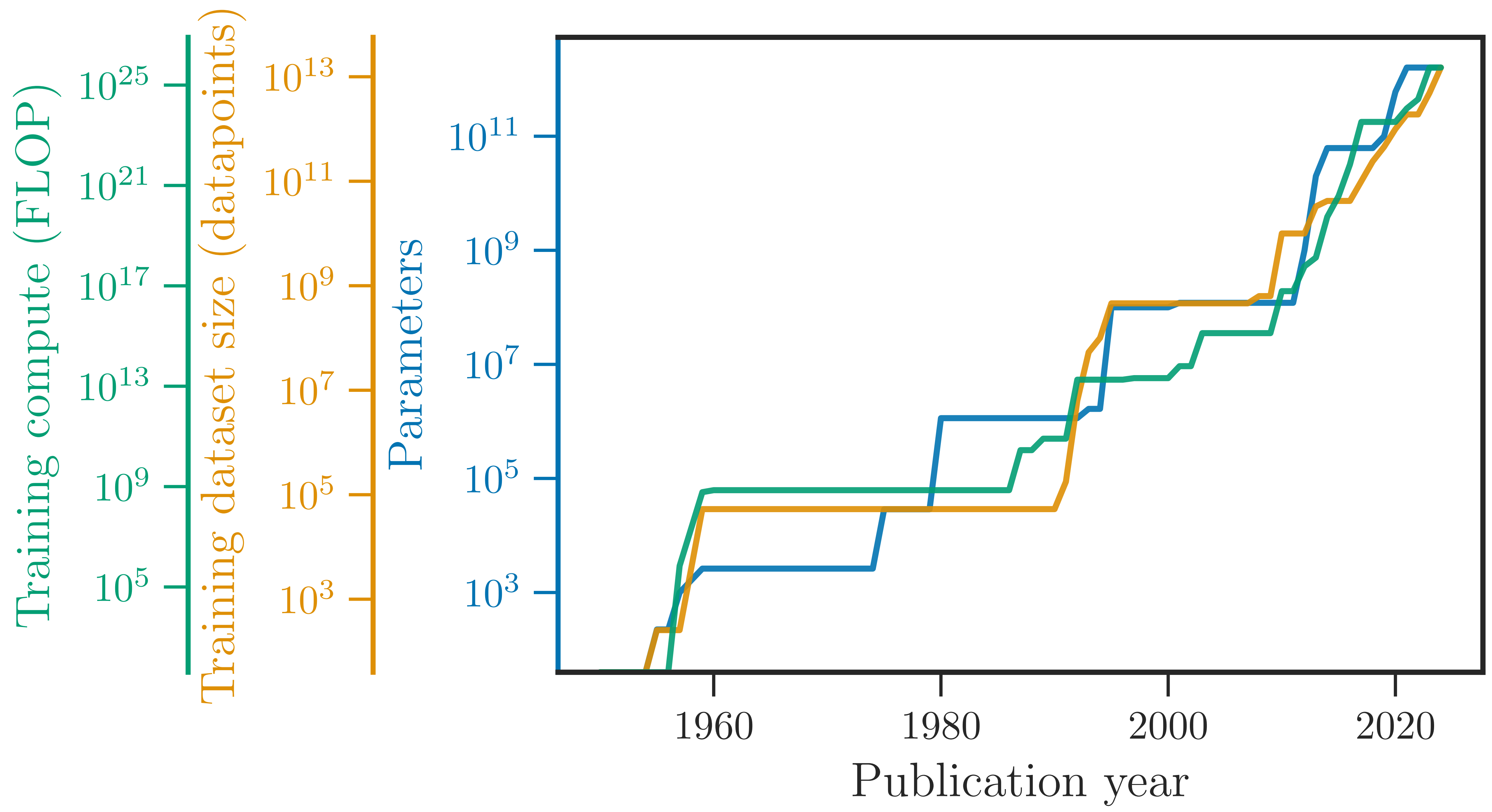


Figure: Evolutions in Machine Learning specialized hardware (GPU, TPU) metrics, as the maximum value up to the given year.

Machine Learning Boom



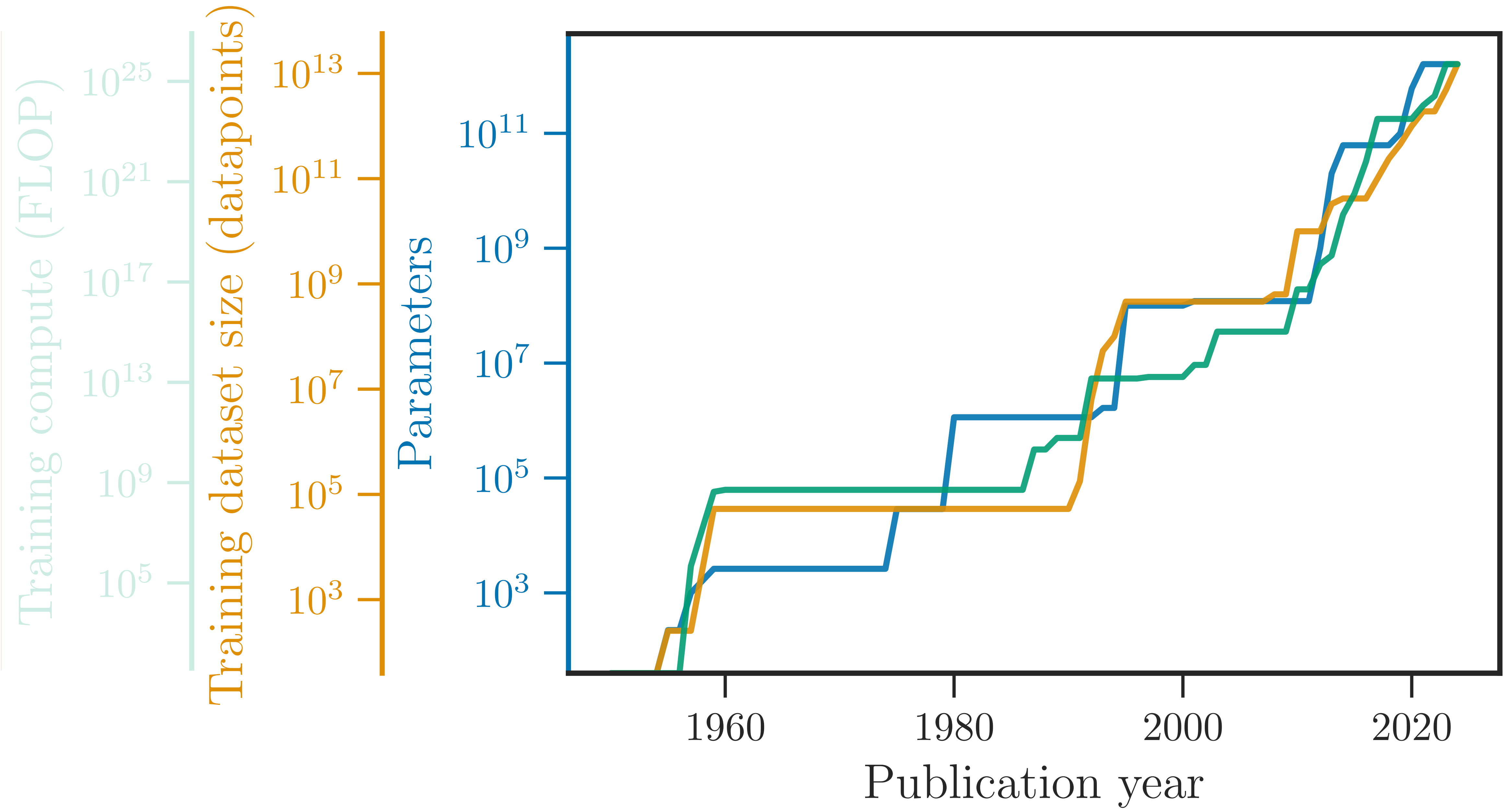
Evolutions in Machine Learning Model and their training, as the maximum value up to the given year. [EpochAI2024]

Parameters: Variables that are learned from the data during the training process. The number of parameters is a representation of the size of the model.

Training Dataset: Collection of data that the ML model is trained on.

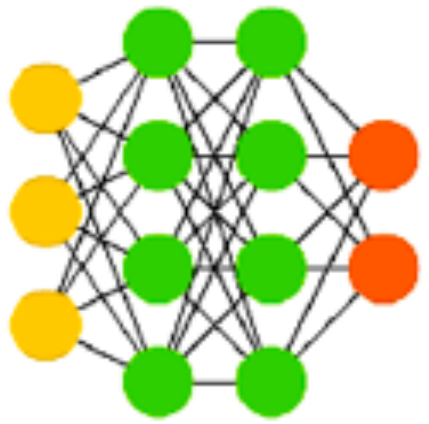
Training computing (FLOP): number of mathematical operations (+,-,*,/) performed to train the model.

Machine Learning Boom



Evolutions in Machine Learning Model and their training, as the maximum value up to the given year. [EpochAI2024]

Model



Parameters: Variables that are learned from the data during the training process. The number of parameters is a representation of the size of the model.

Training Dataset: Collection of data that the ML model is trained on.

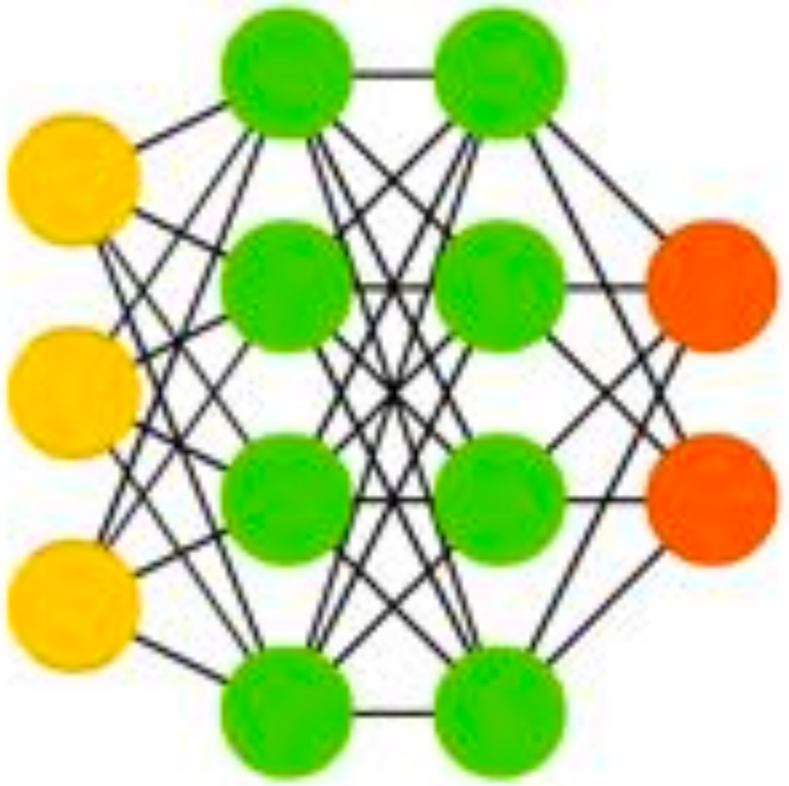
→ Pembroke Welsh Corgi with Cowboy Hat

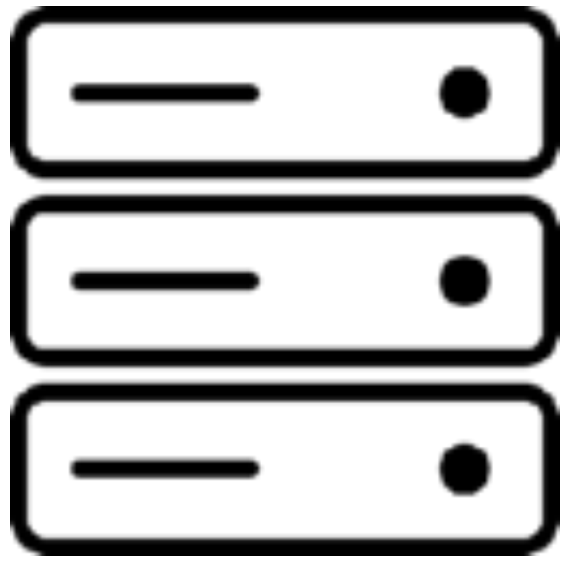
Quick definition of ML

Objective/Task	Data	Model	Learning

Generating image from text

Pembroke Welsh Corgi with Cowboy Hat





Machine Learning and its materiality

CPU
Central Processing Units

RAM
Random Access Memory

+

SSD
Solid State Drive

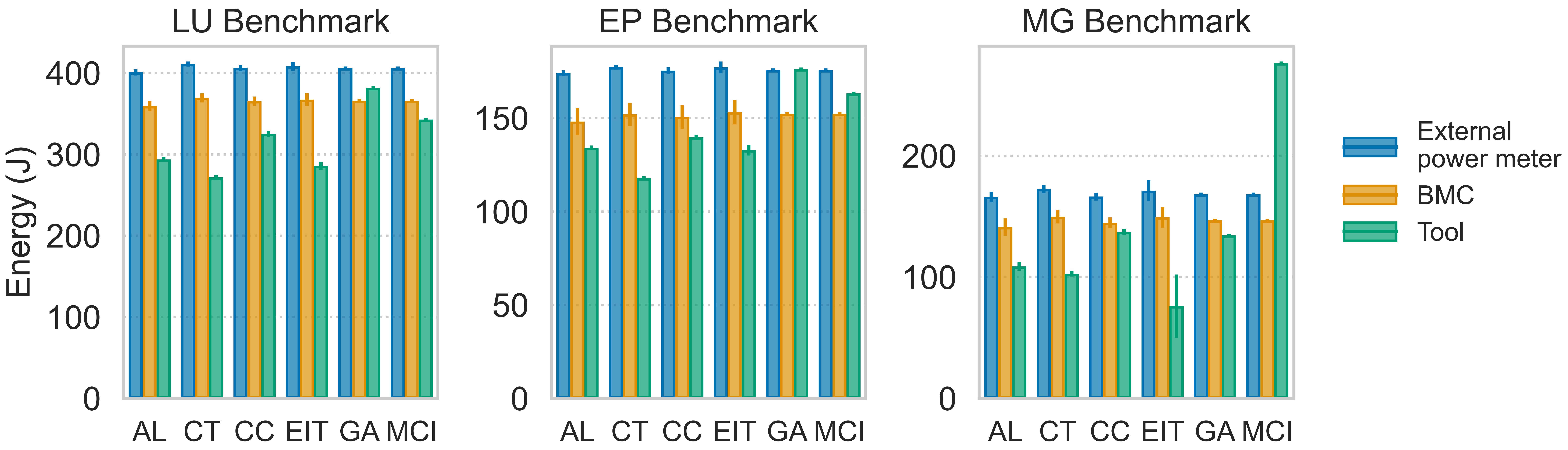
GPU
Graphic Processing Units

Example of a supercomputer in a Data Center

CCGRID

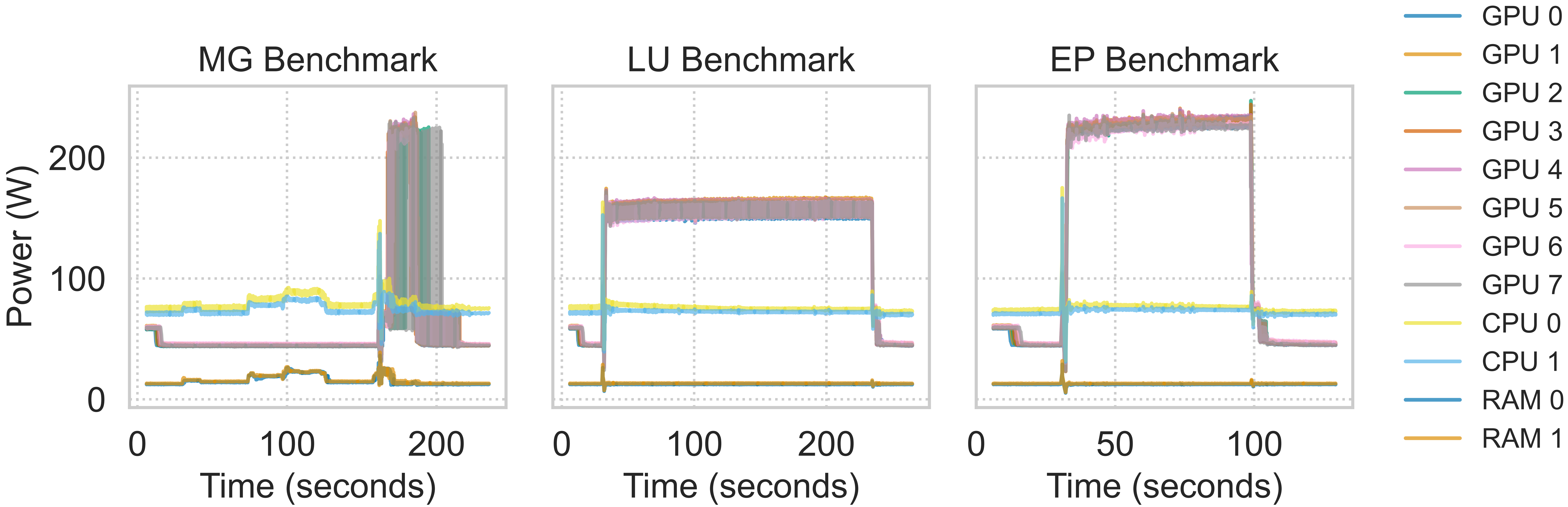
	External and intra-node devices		Power profiling software	Energy measurement software packages			Energy calculators	
	OmegaWatt	*BMC*	*Alumet*	*Code Carbon*	*Experiment Impact Tracker*	*Carbon Tracker*	*Green Algorithms*	*ML CO2 Impact*
Developement								
Origin			Eviden, LIG	MILA	University of Standford	University of Copenhagen	University of Cambridge	MILA
First (latest) release date			Mar. 2023 (May 2024)	Nov. 2020 (Jan. 2024)	Dec. 2019 (Jan. 2020)	Apr. 2020 (Sept. 2023)	Jul. 2020 (Apr. 2023)	Aug. 2019 (Jul. 2022)
Environment								
Hardware compatibility	Any	Any	Intel RAPL, Nvidia NVML	Any	Intel RAPL, Nvidia NVML	Intel RAPL, Nvidia NVML	Any	Any
Scope	Node	Node	CPU, DRAM, GPU	CPU, DRAM, GPU	CPU, DRAM, GPU, process	CPU, DRAM, GPU	Node	Node
Job management support			No	No	No	No		
Functional								
Hardware technology used			RAPL, NVML	RAPL, NVML, TDP	RAPL, NVML	RAPL, NVML	TDP	TDP
Software power model used					GPU, CPU and RAM usage-based			
Default sampling frequency (Hz)	1	0.2	2	1/15	1	0.1		
Online reporting	Yes	Yes	Yes	No	No	No	No	No
Power profiling	Yes	Yes	Yes	No	No	No	No	No
User-friendliness								
Availability of source code (License)			Yes (EUPL 1.2)	Yes (MIT)	Yes (MIT)	Yes (MIT)	Yes (CC-BY-4.0)	Yes (MIT)
Ease of use	Poor	Poor	Good	Good	Good	Good	Very good	Very good
Quality of documentation			Good	Fair	Fair	Good	Good	Fai
Configurability	Fair	Poor	Good	Poor	Quite good	Poor	Poor	Poor
Resulting data format	HTTP end-point	HTTP end-point	CSV	CSV	JSON, Code	File, Code	Web	Web, Latex
Data visualisation possibilities	Grafana (Kwollect)	Grafana (Kwollect)		Comet			Graphs on the web page	

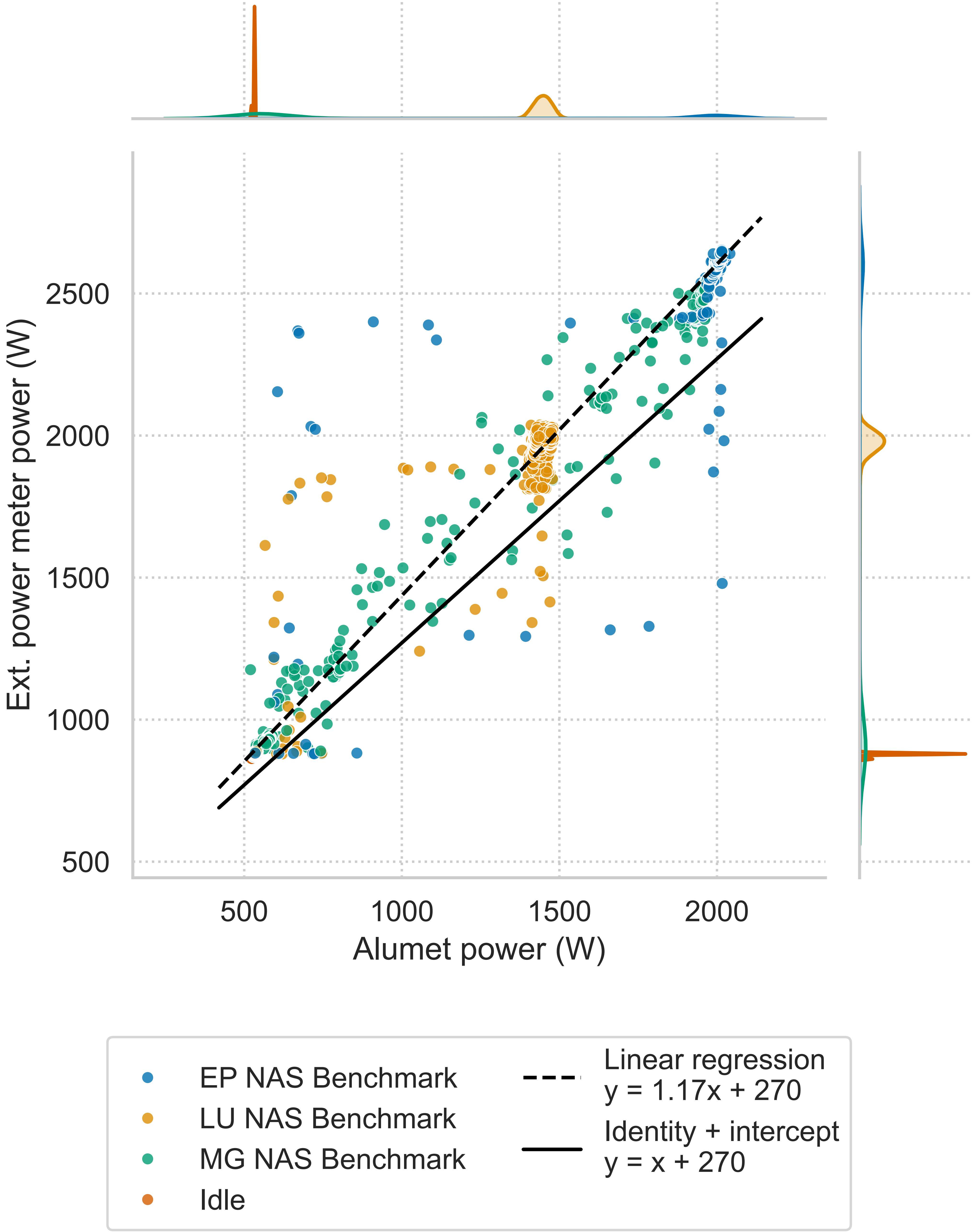
Comparing tools



AL: ALUMET, CT: Carbon Tracker, CC: Code Carbon,
EIT: Expriment Impact Tracker, GA: Green Algorithm, MCI: ML CO2 Impact

Per component consumption





Ext. power meter power (W)
2500
2000
1500
1000
500
500
1000
1500
2000
Alumet power (W)
EP NAS Benchmark
LU NAS Benchmark
MG NAS Benchmark
Idle
Linear regression
y = 1.17x + 270
Identity + intercept
y = x + 270

Electricity

Infrastructures

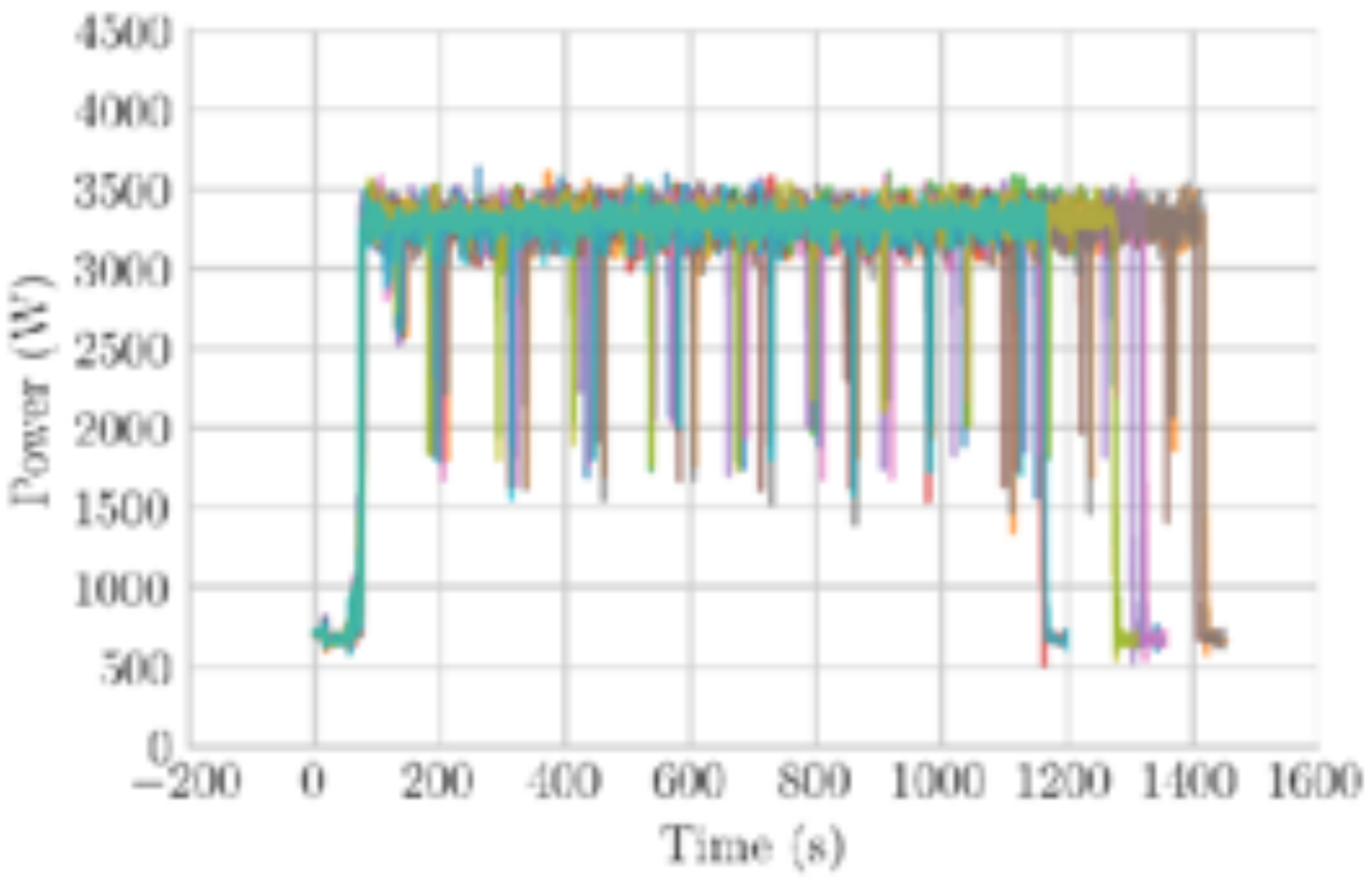
	HPC: APOLLO	EDGE: JETSON
Node model	Apollo 6500 Gen10	Nvidia Jetson AGX Xavier
Number of nodes	20	12
GPU model	NVIDIA A100-SXM-80GB	NVIDIA GV10B, Volta architecture
Number of GPU per node	8	1
GPU TDP (W)	400	
CPU model	AMD EPYC 7763 64-Core Processor	Nvidia Carmel (Carmel), aarch64, 8 cores
Number of CPU per node	2	1
CPU TDP (W)	280	
Node TDP (W)	3760	30
Memory	1 TB	32 GB
Switch model	Mellanox HDR Infiniband	
Number of switch	8	
Switch power consumption (W)	375	
Installation year	2022	2023
Available thought	HPE local network - slurm	Grid'5000 Estats cluster - OAR
Power meter	HPE iLO 5 + RAPL + NVML	Jetson-stats

Machine Learning Framework for Distributed Platforms

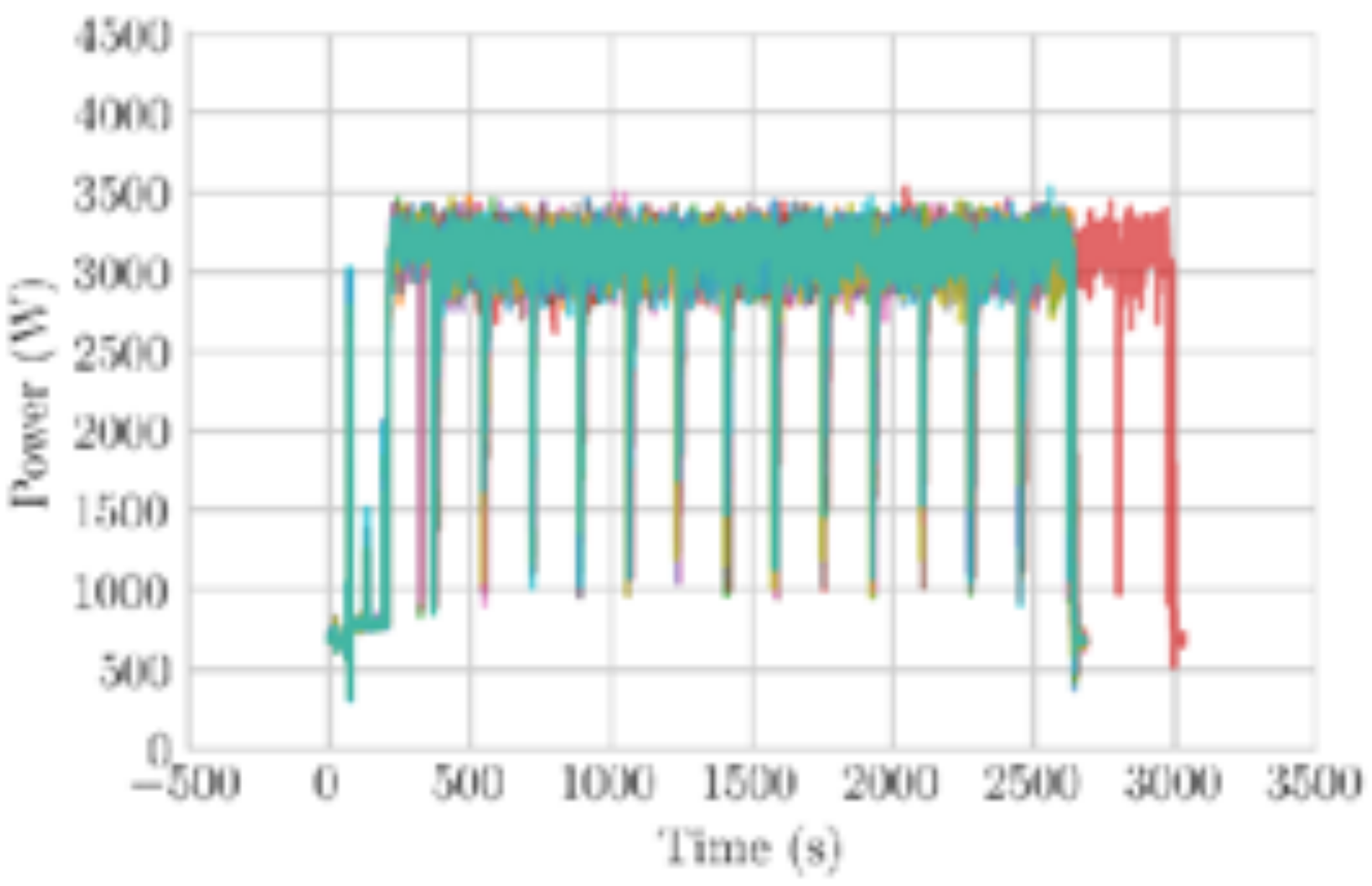
Model	Apollo	Jetson
UNet	MXNet / Horovod	
RNN-T	PyTorch	
ResNet	MXNet / Horovod	PyTorch
MaskRCNN	PyTorch	
BERT	PyTorch	
DLRM	HugeCTR	

Power profile of each FU

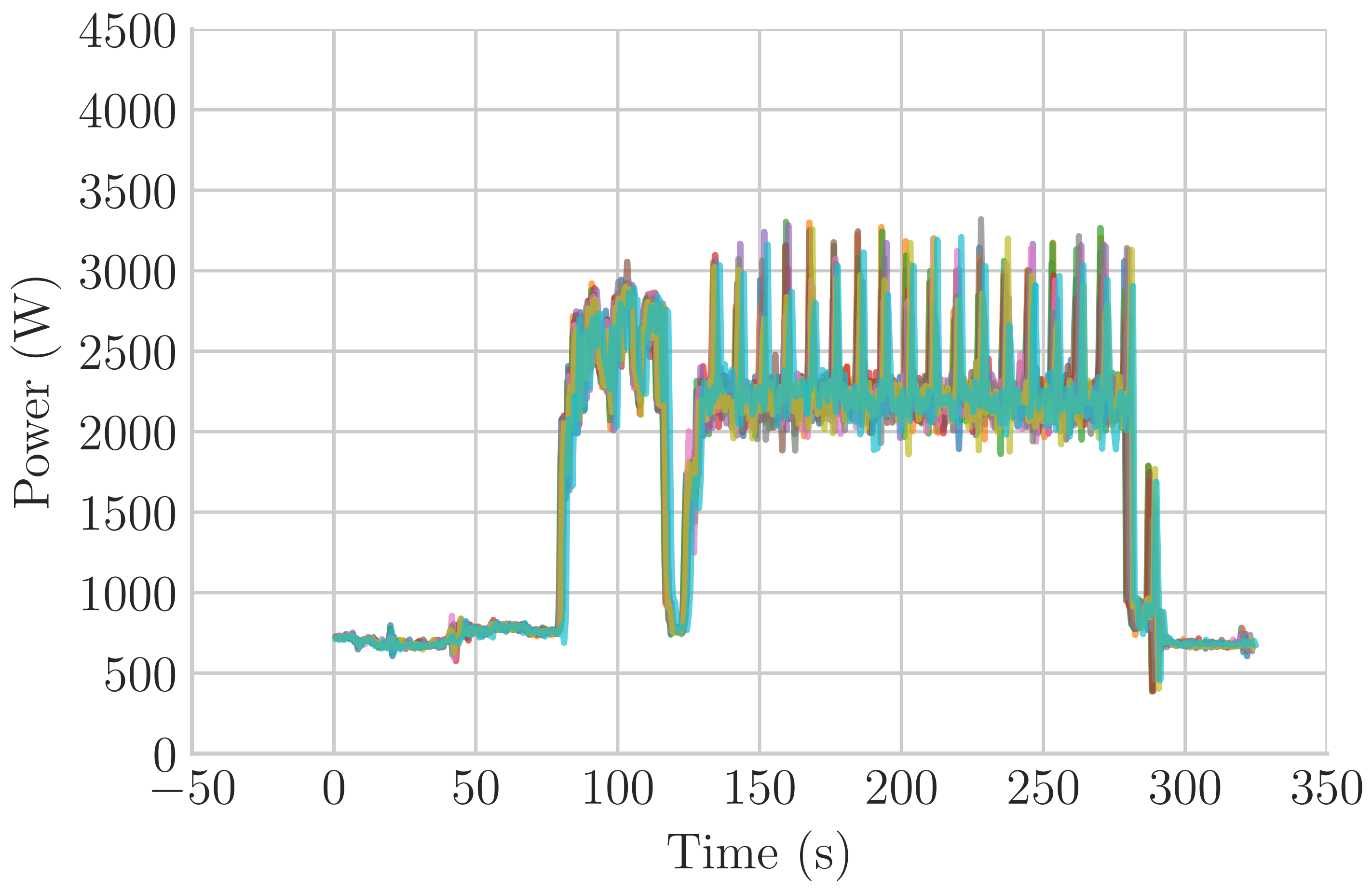
BERT FU



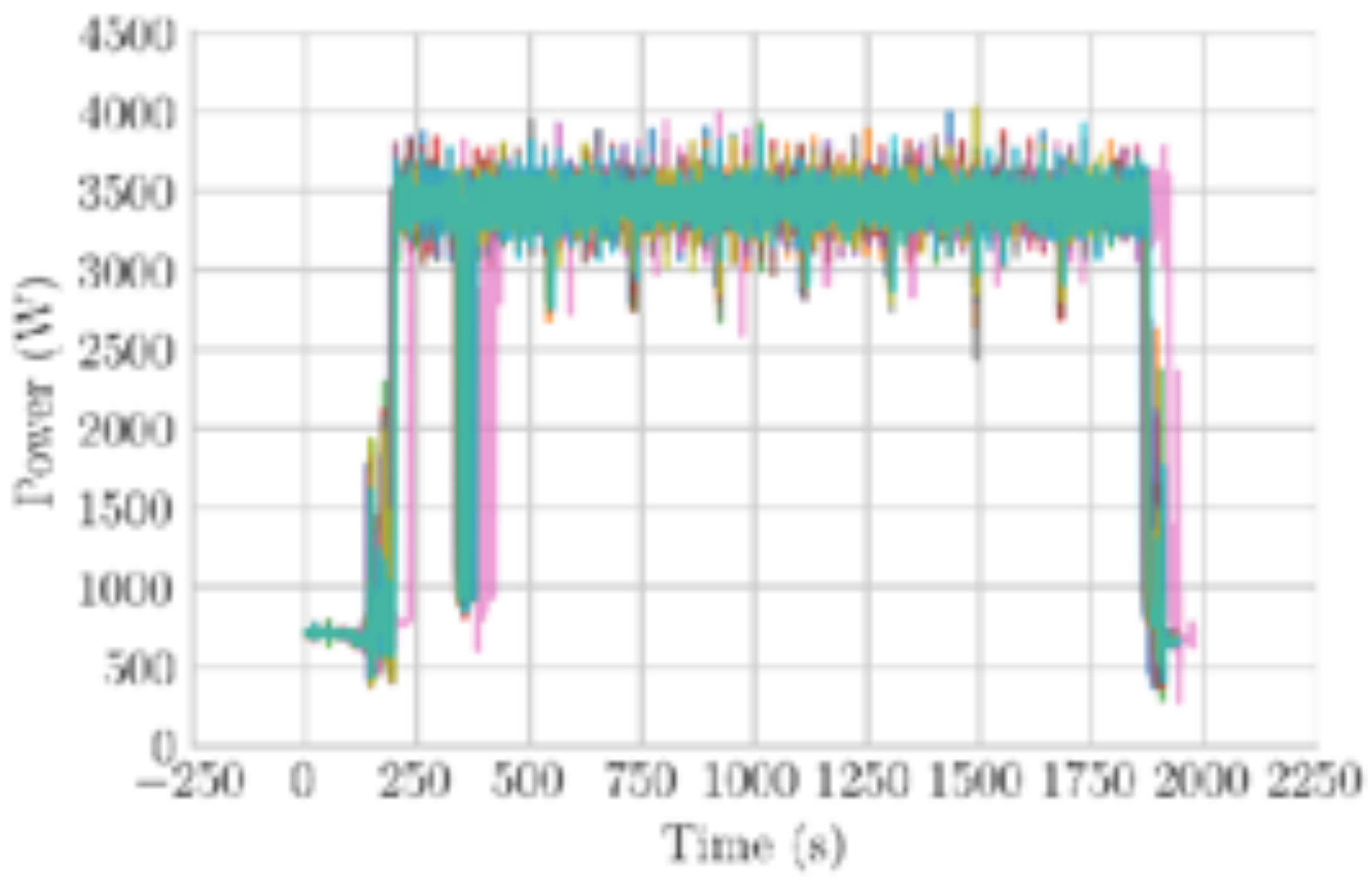
Mask R-CNN FU



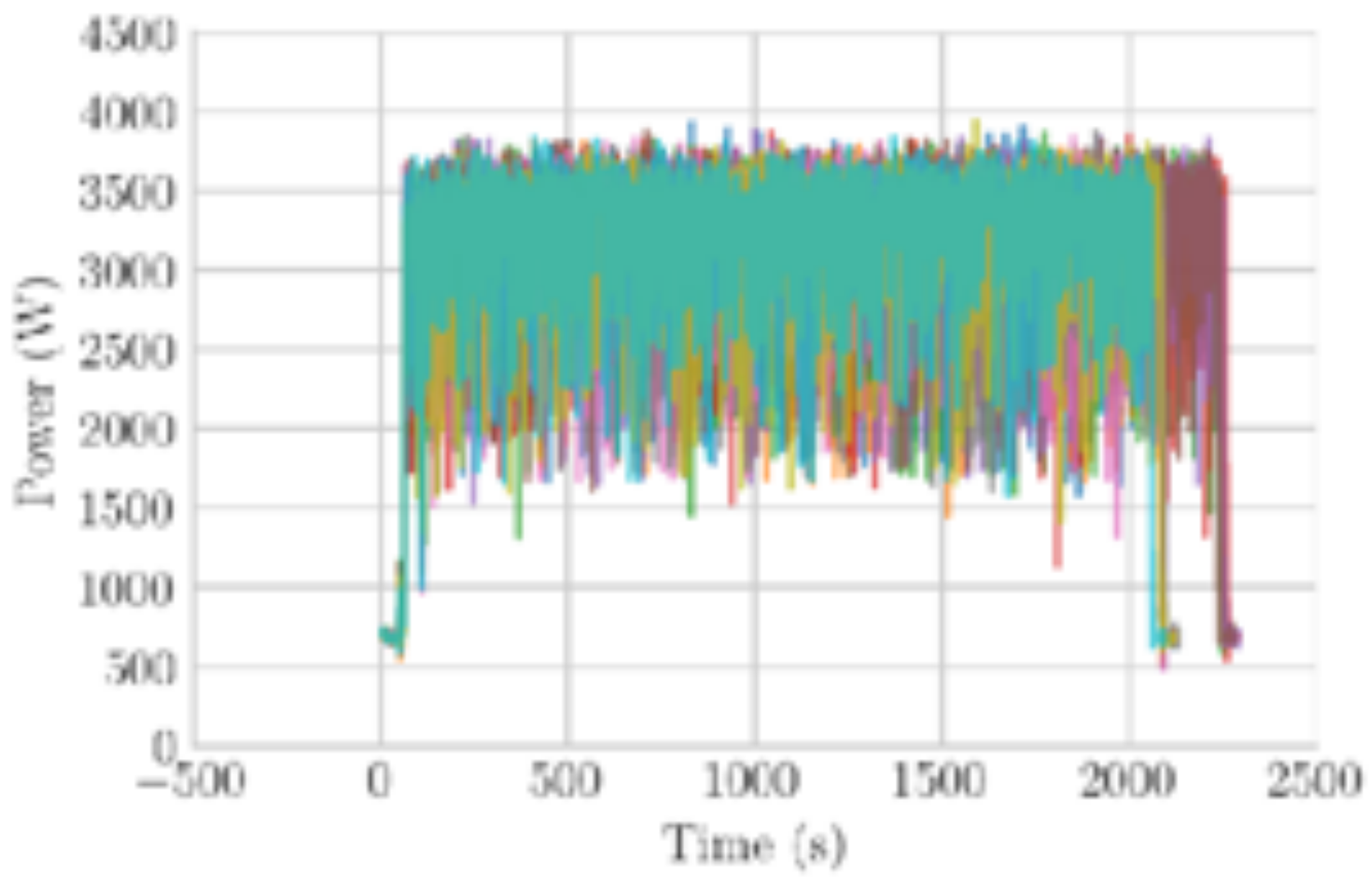
DLRM FU



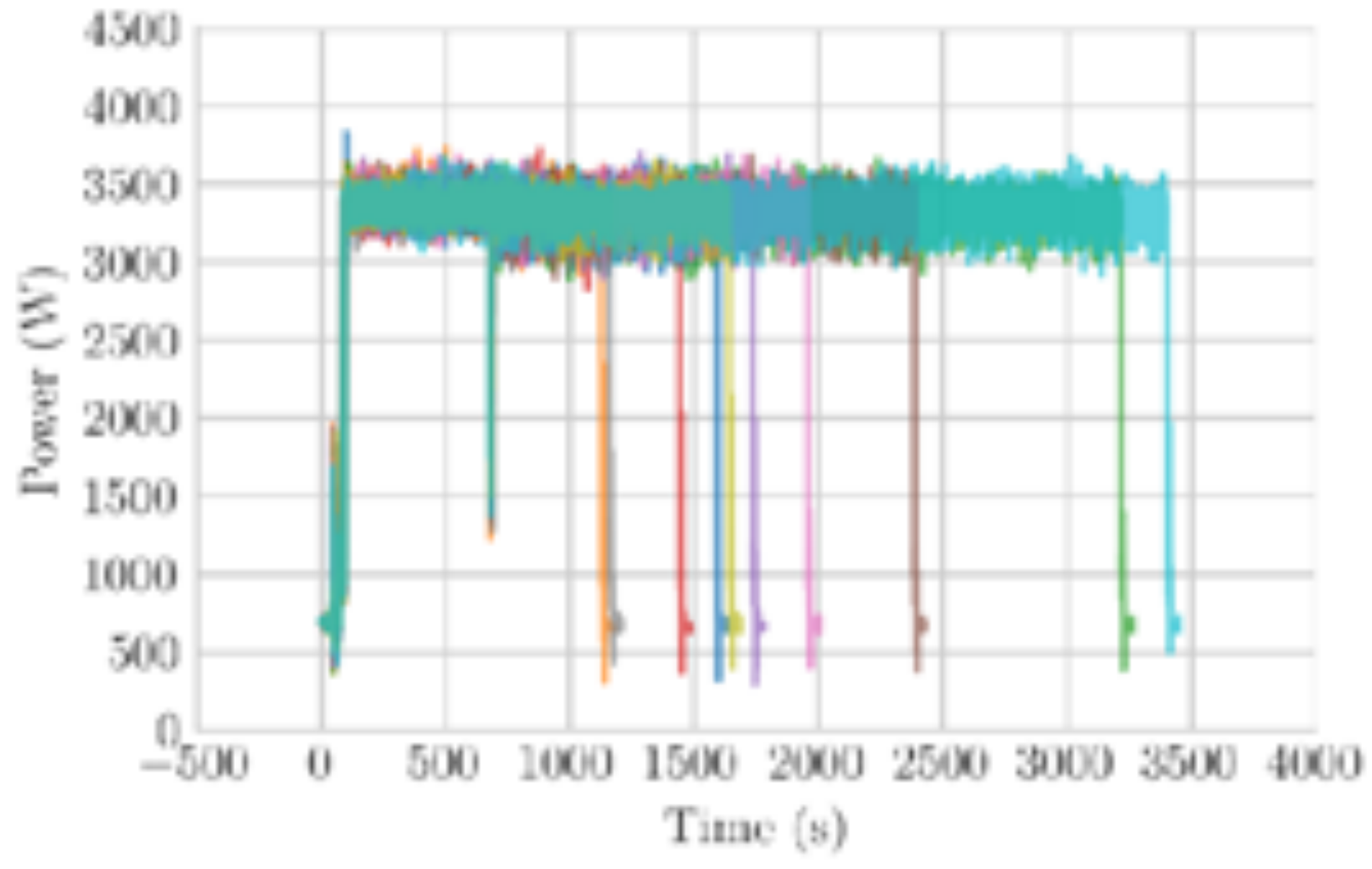
ResNet FU



RNN-T FU

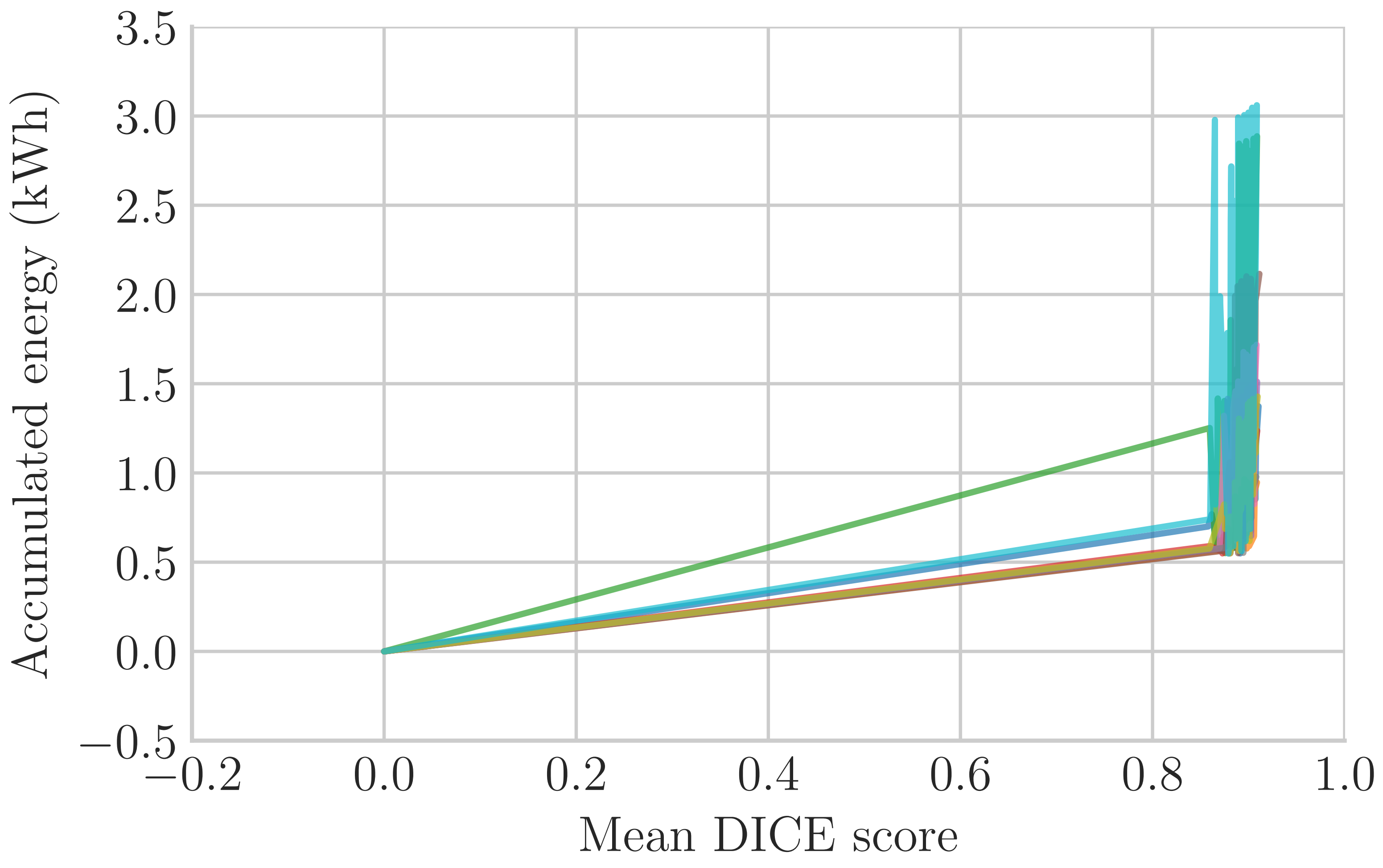


UNet FU

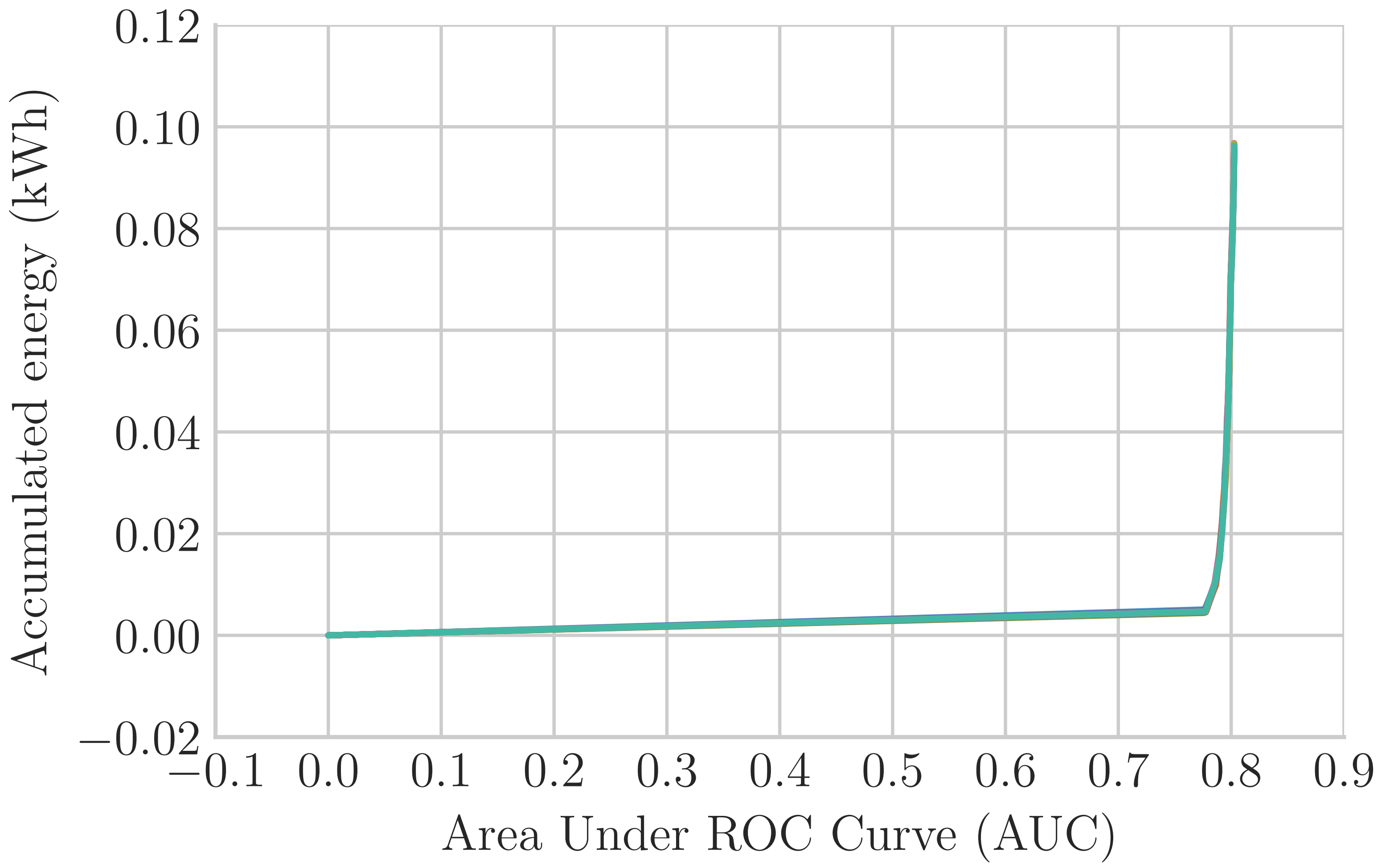


Accuracy/Energy tradeoff on Apollo

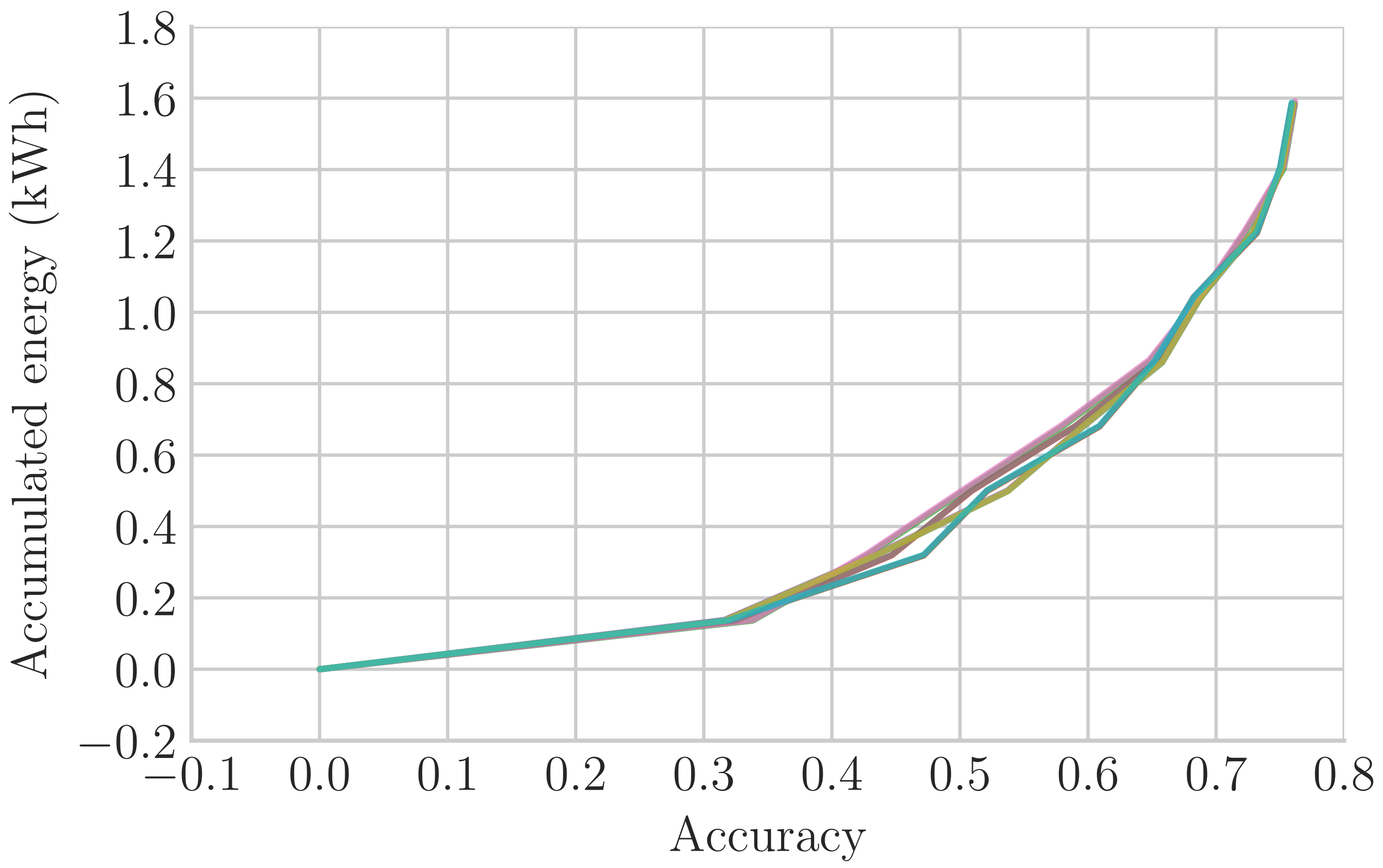
U-Net-FU

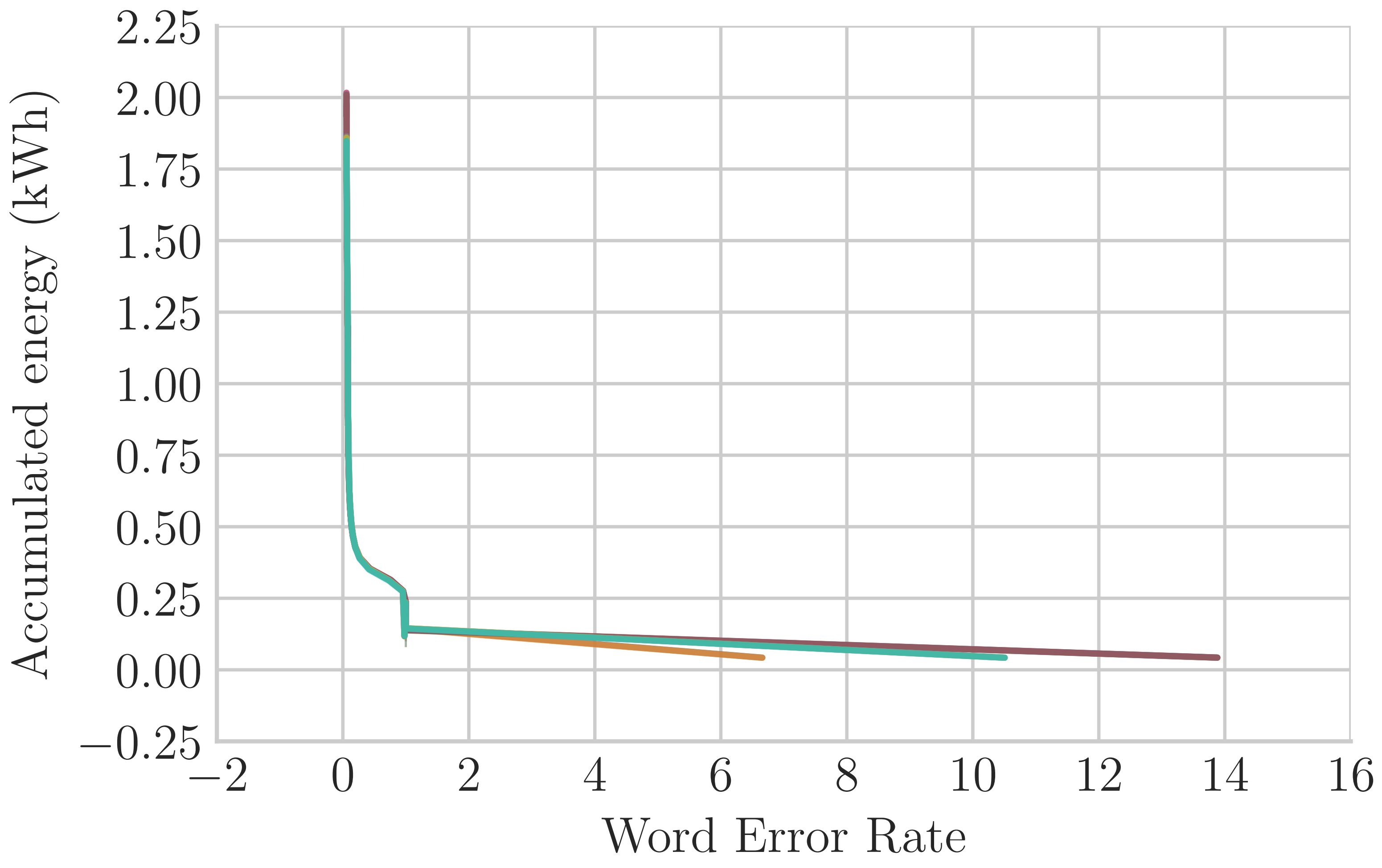


DLRM FU

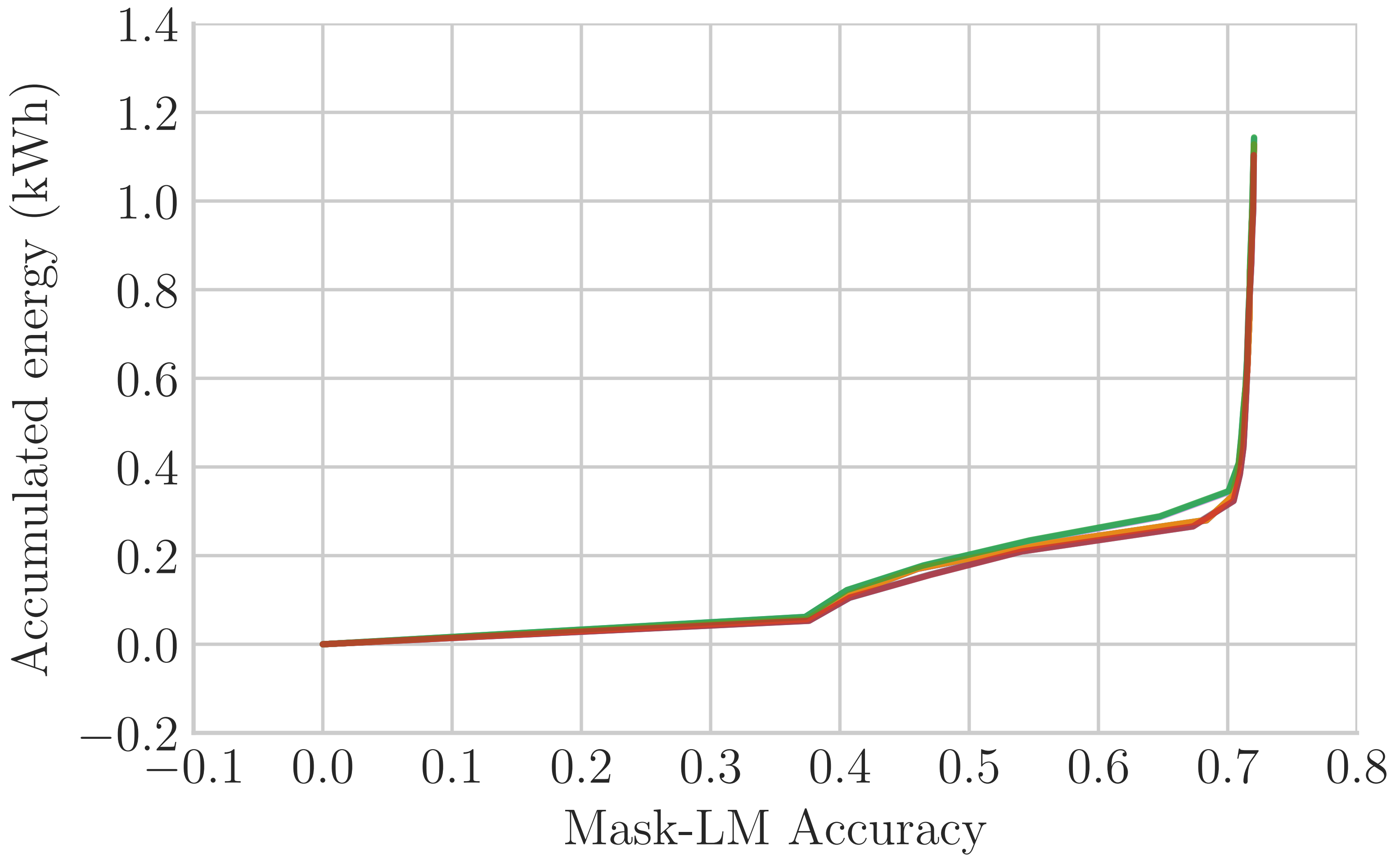


ResNet-50 FU

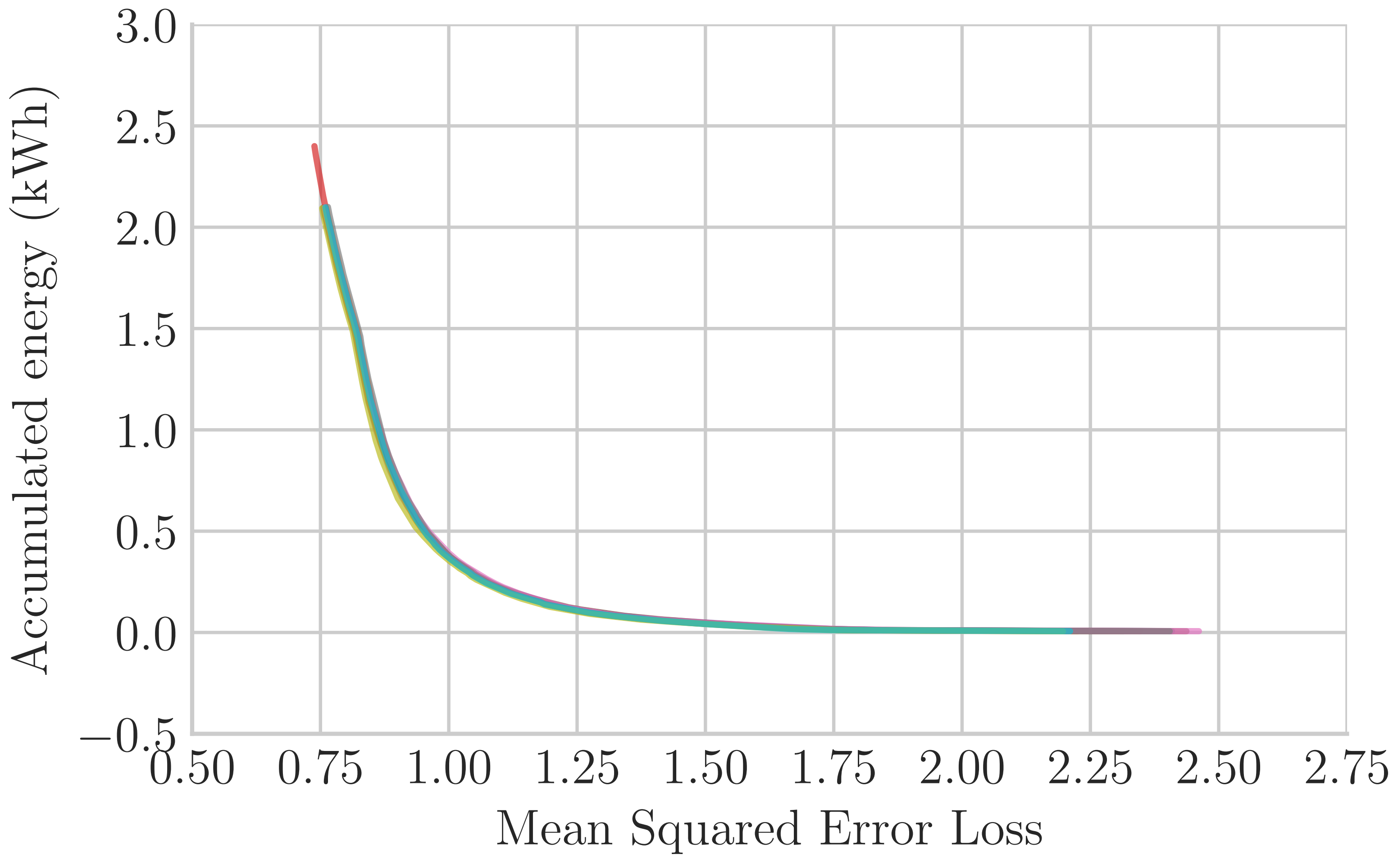




RNN-T



BERT FU



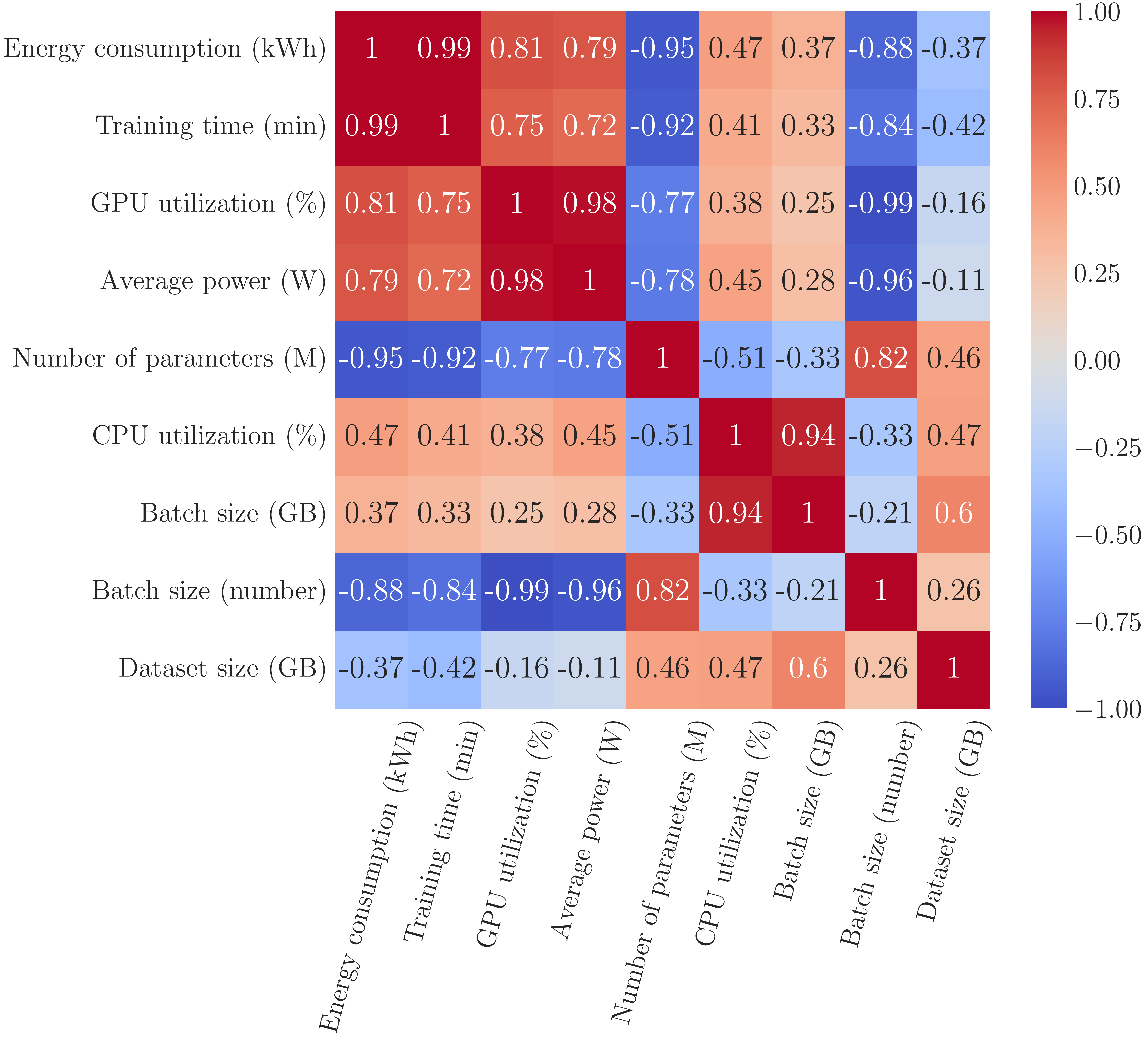
Mask R-CNN

Performances of each FU

	Energy (kWh)	Time (min)	Utilization (%)	
			GPU	CPU
FU				
ResNet-50	1.61 ± 0	29.42 ±0.19	94.84 ± 0.71	28.57 ± 0.72
3D U-Net	1.73 ± 0.7	31.72 ± 12.6	96.94 ± 0.89	8.2 ± 0.07
Mask R-CNN	2.16 ± 0.09	43.6 ± 1.69	89.87 ± 0.2	8.84 ± 0.02
RNN-T	1.97 ± 0.11	36.12 ± 2.21	95.46 ± 0.28	68.47 ± 0.89
BERT-large	1.13 ± 0.01	20.83 ± 0.25	96.88 ± 0.11	6.88 ± 0.01
DLRM	0.14 ± 0	4.18 ± 0.02	57 ± 0.46	5.36 ±0.03

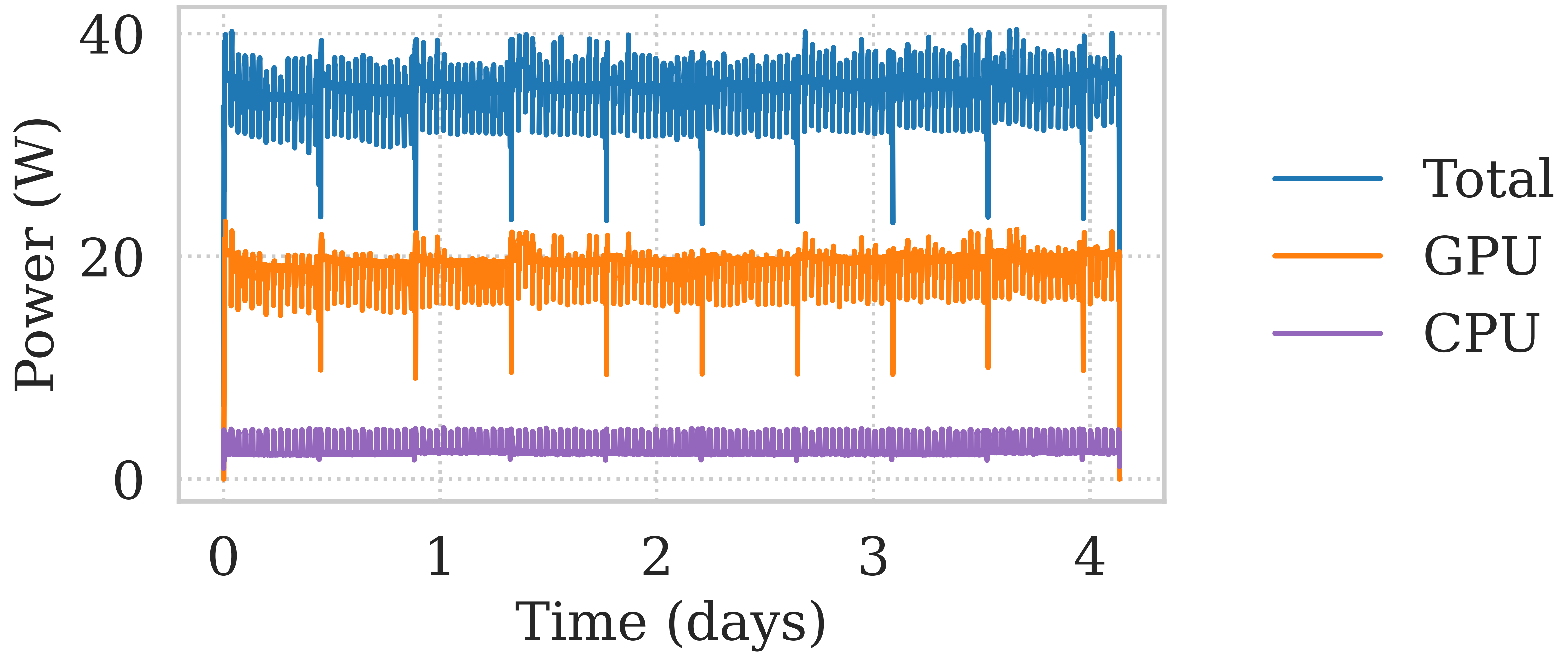
Characteristics of the MLPerf models and datasets

Model	Parameter number (M)	Dataset size		Batch size	
		Sample	GB	Sample	In GB
ResNet-50 v1.5	25.6	1.28e+6	167	408	53
3D U-Net	19	6.72e+4	40	56	229
Mask R-CNN	25.6	4.00e+4	20	96	48
RNN-T	29.8	2.78e+5	500	1536	2763
BERT-large	345	3.00e+6	400	384	51
DLRM	540	3.78e+9	342	55296	5

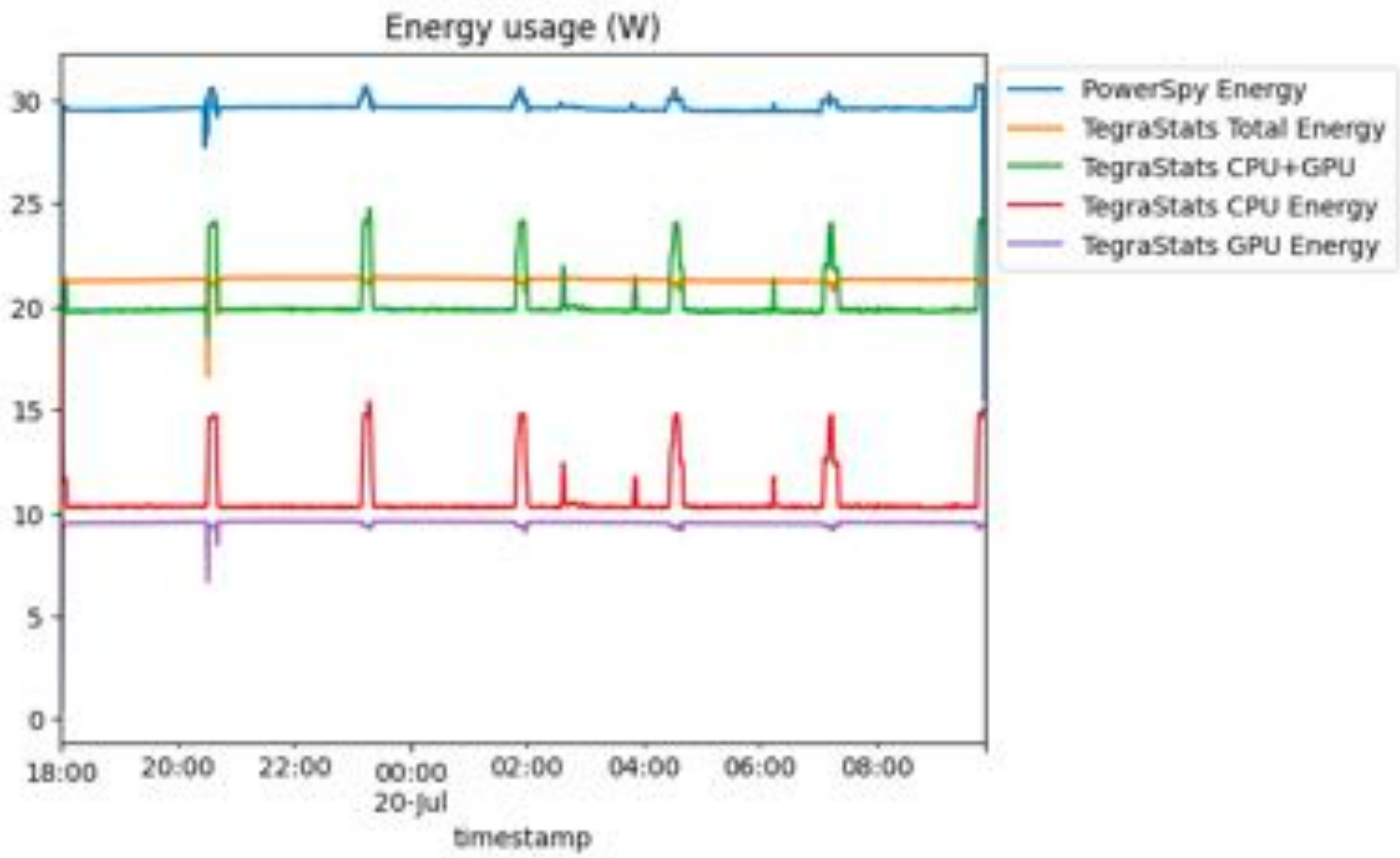


Correlation heatmap

Jetson: Incoherence of electricity measurements

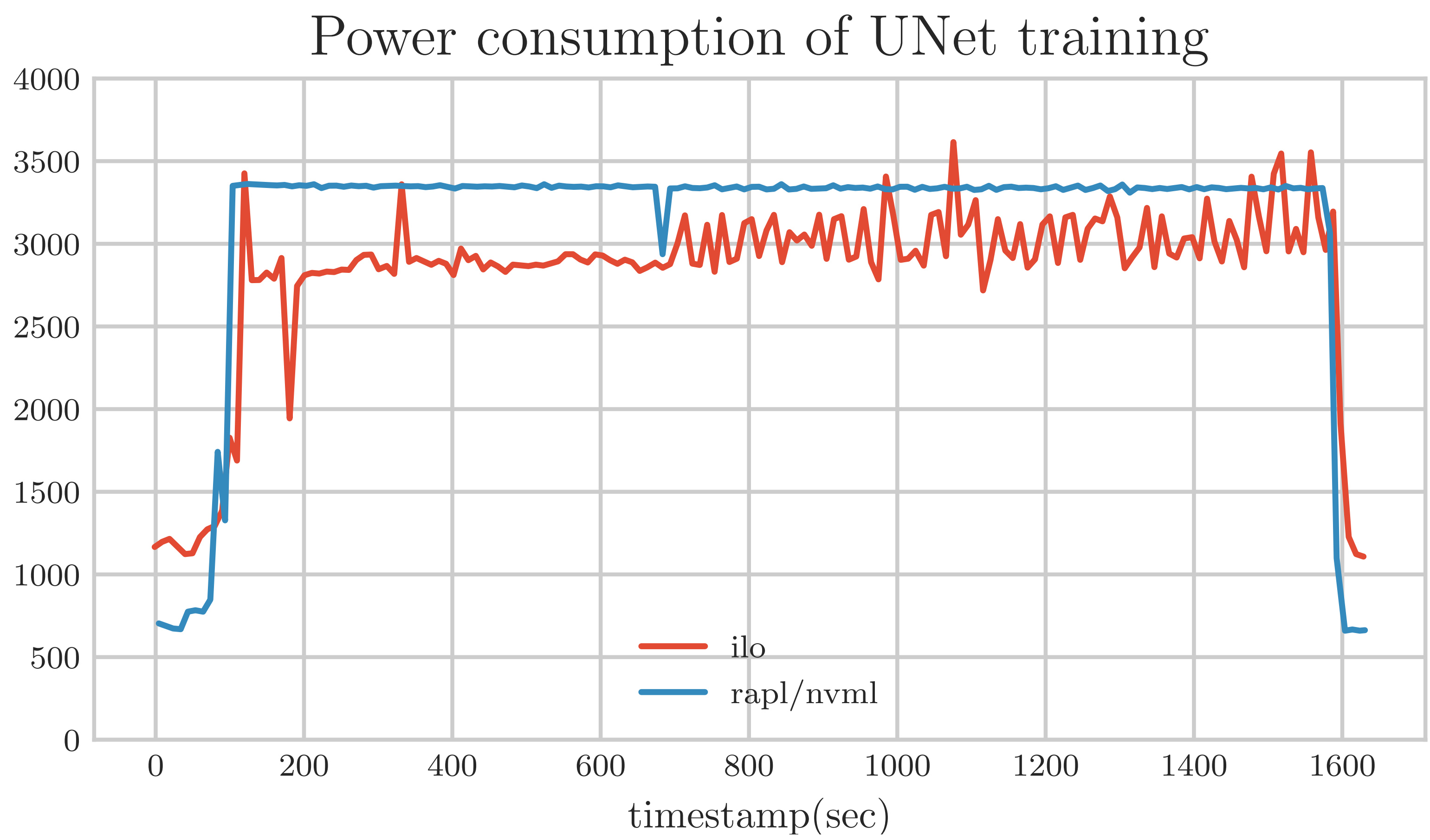


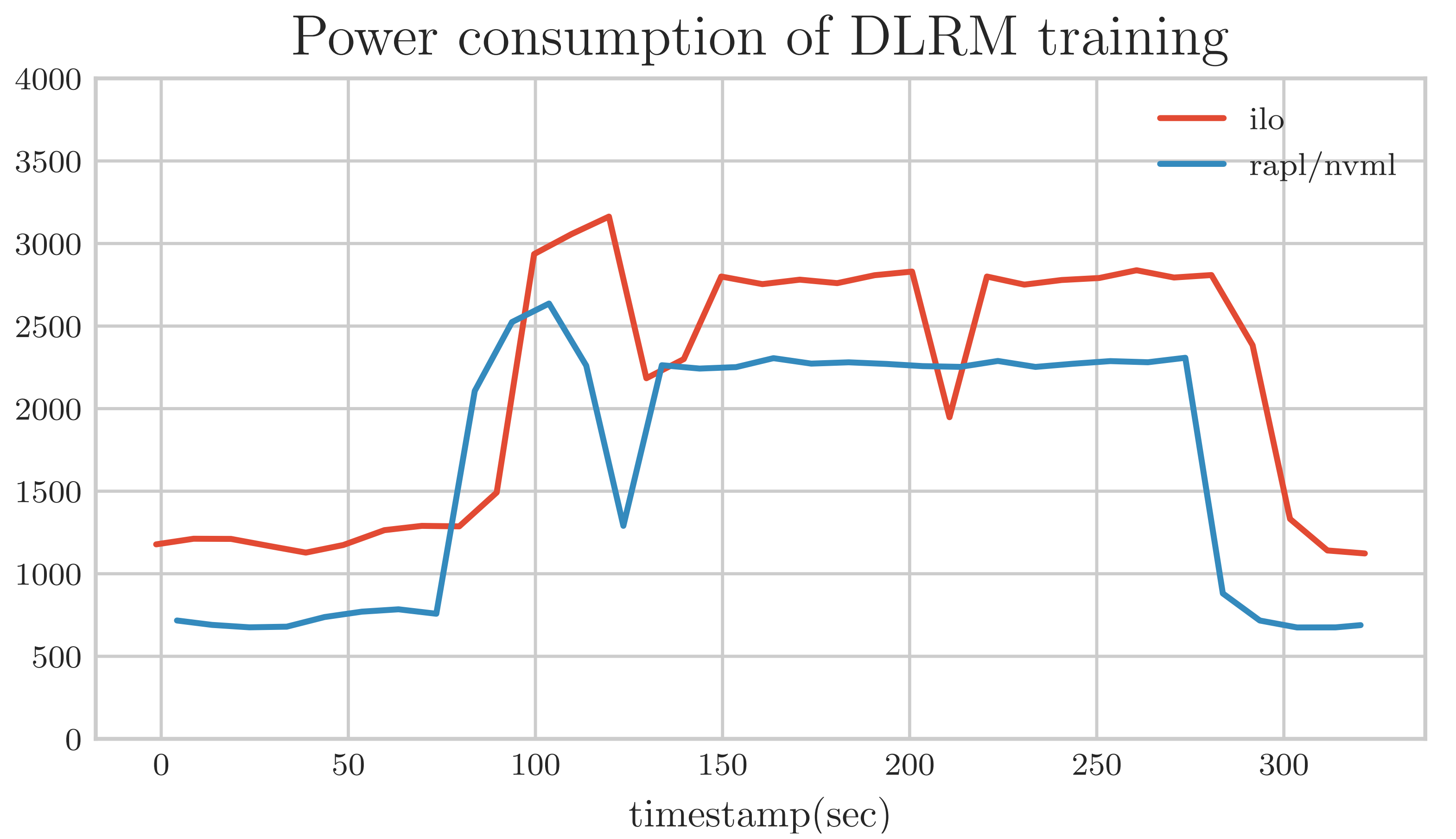
Power profile of ResNet-50 FU on Jetson.

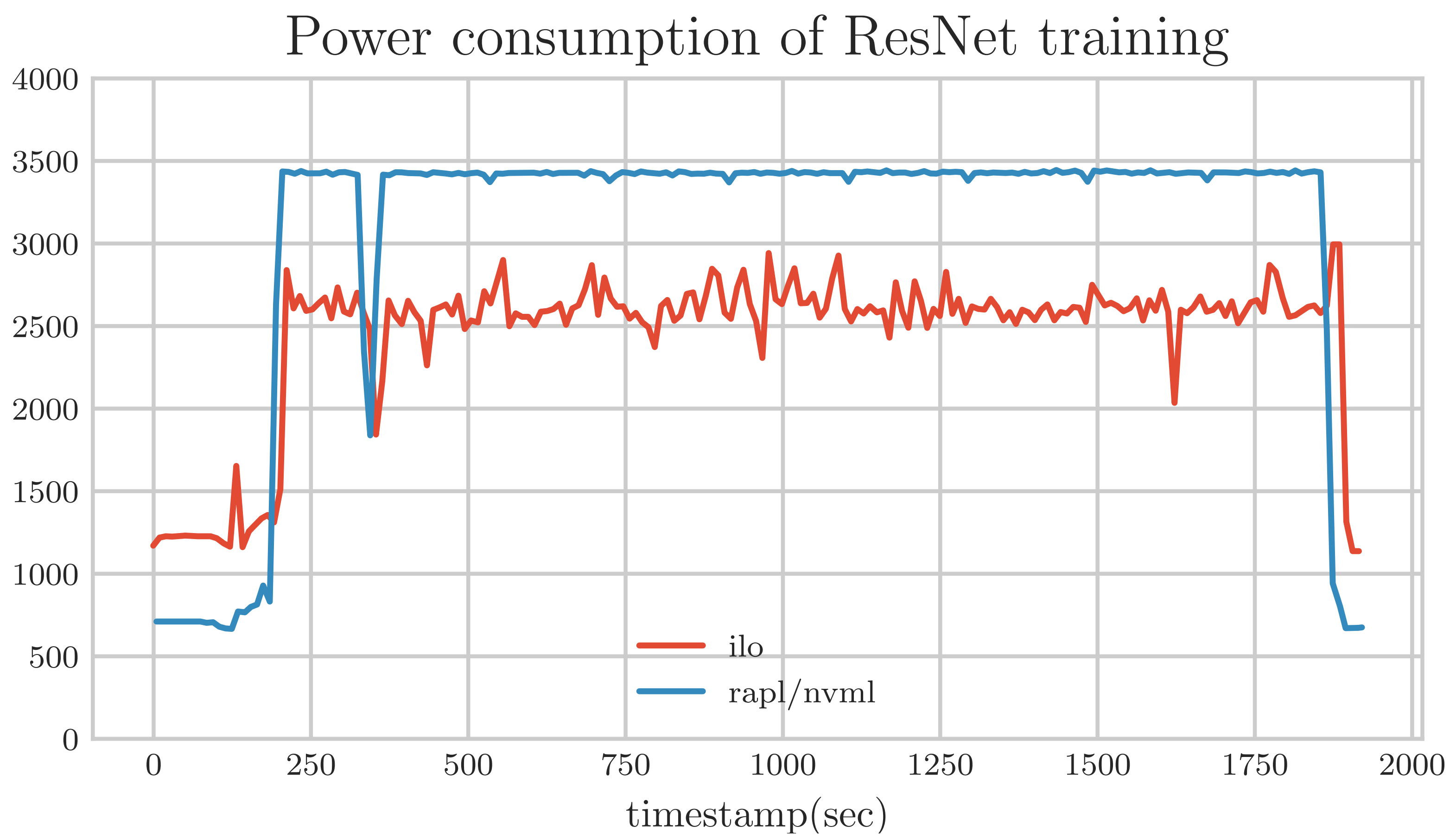


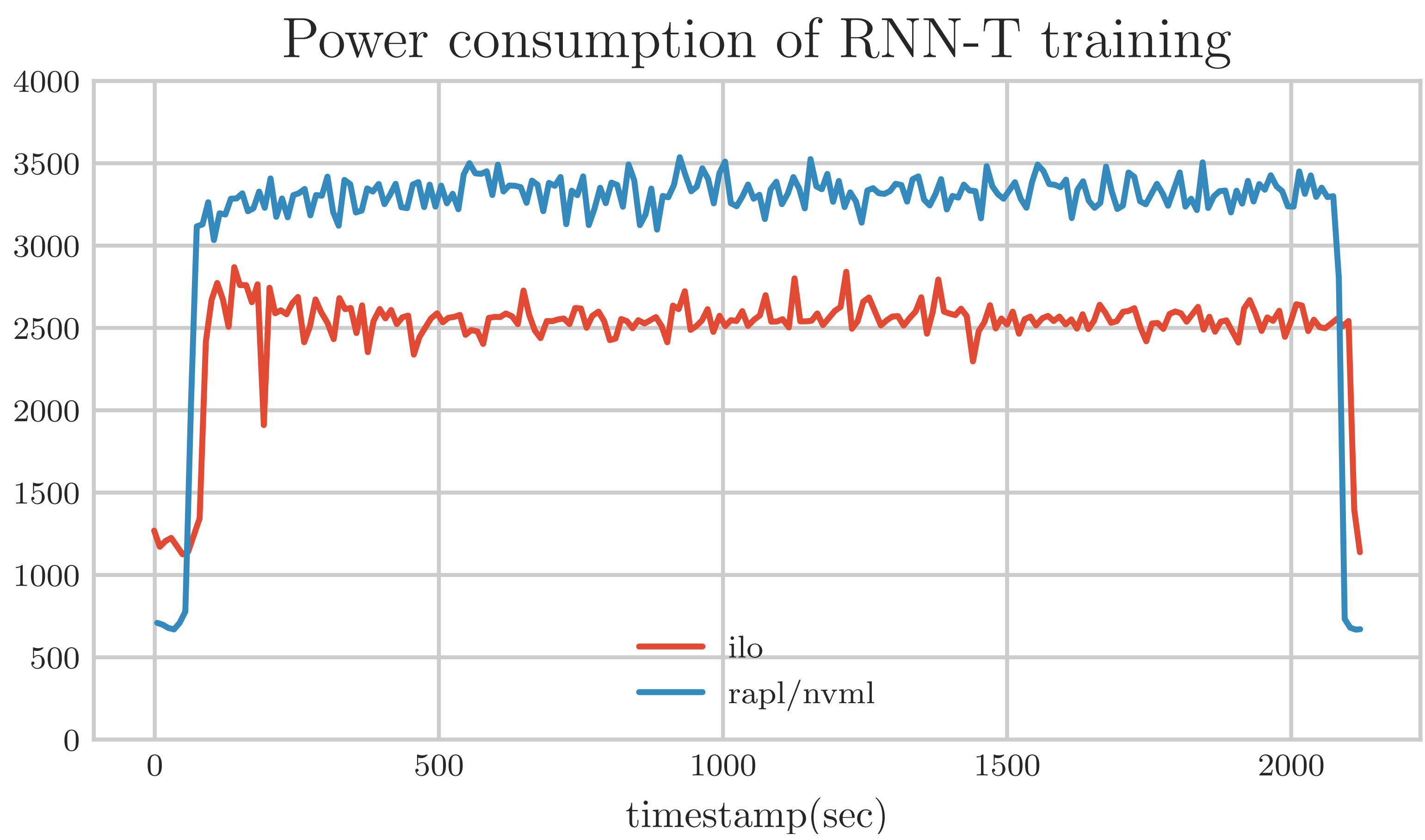
Comparing power meter and software-based power meter on Jetson (development kit, 16 Go)

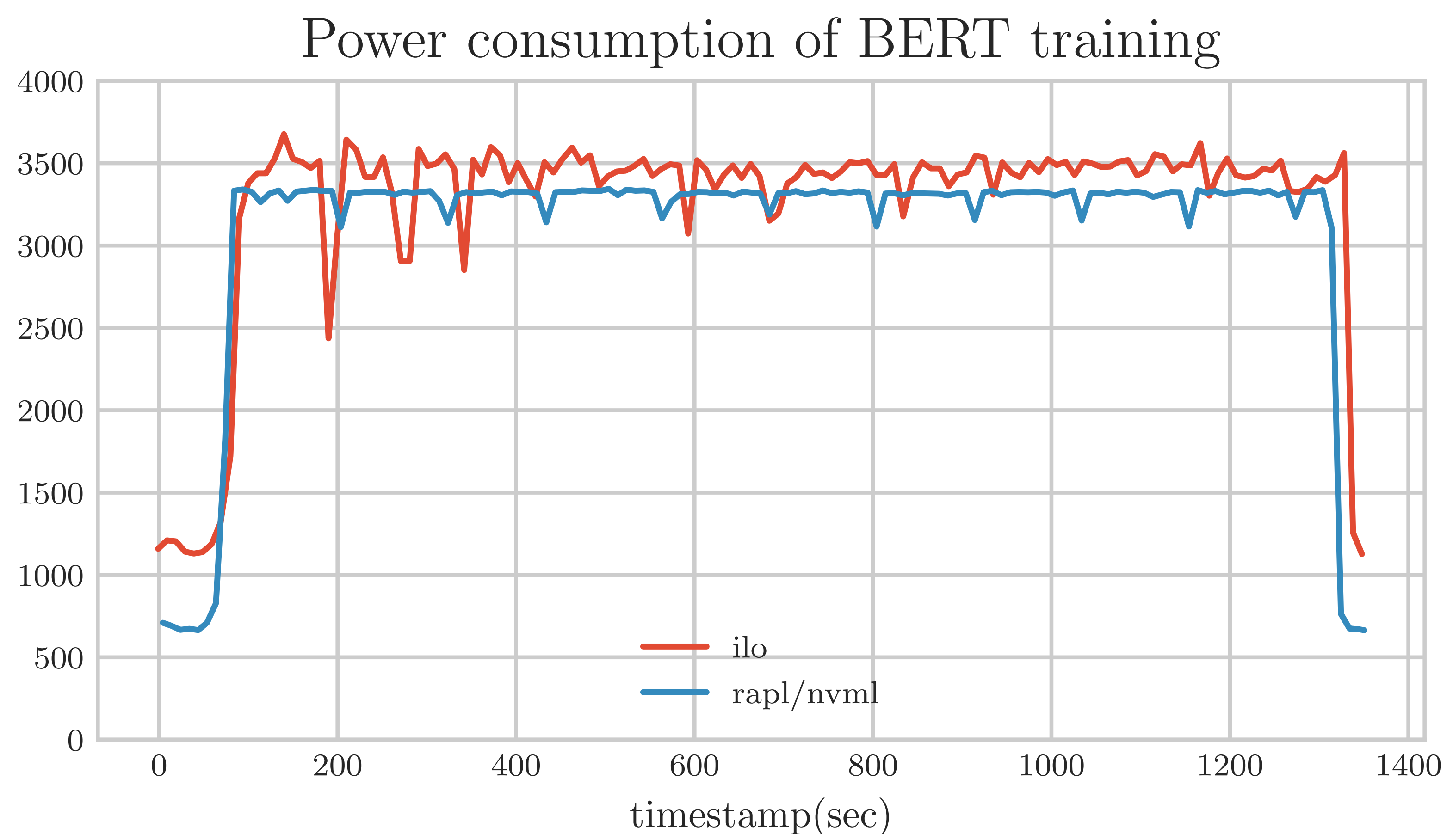
Apollo: Incoherence of electricity measurements

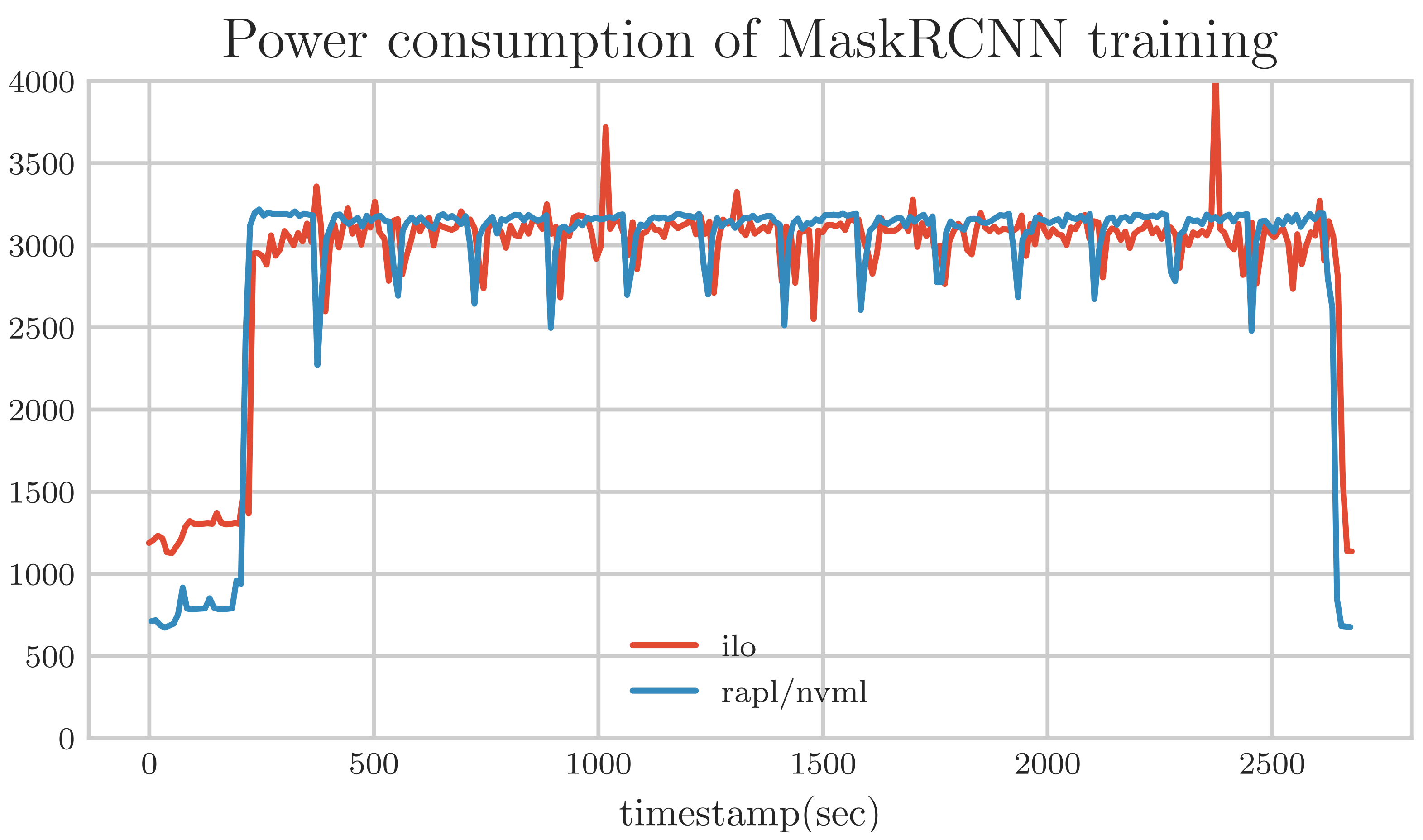




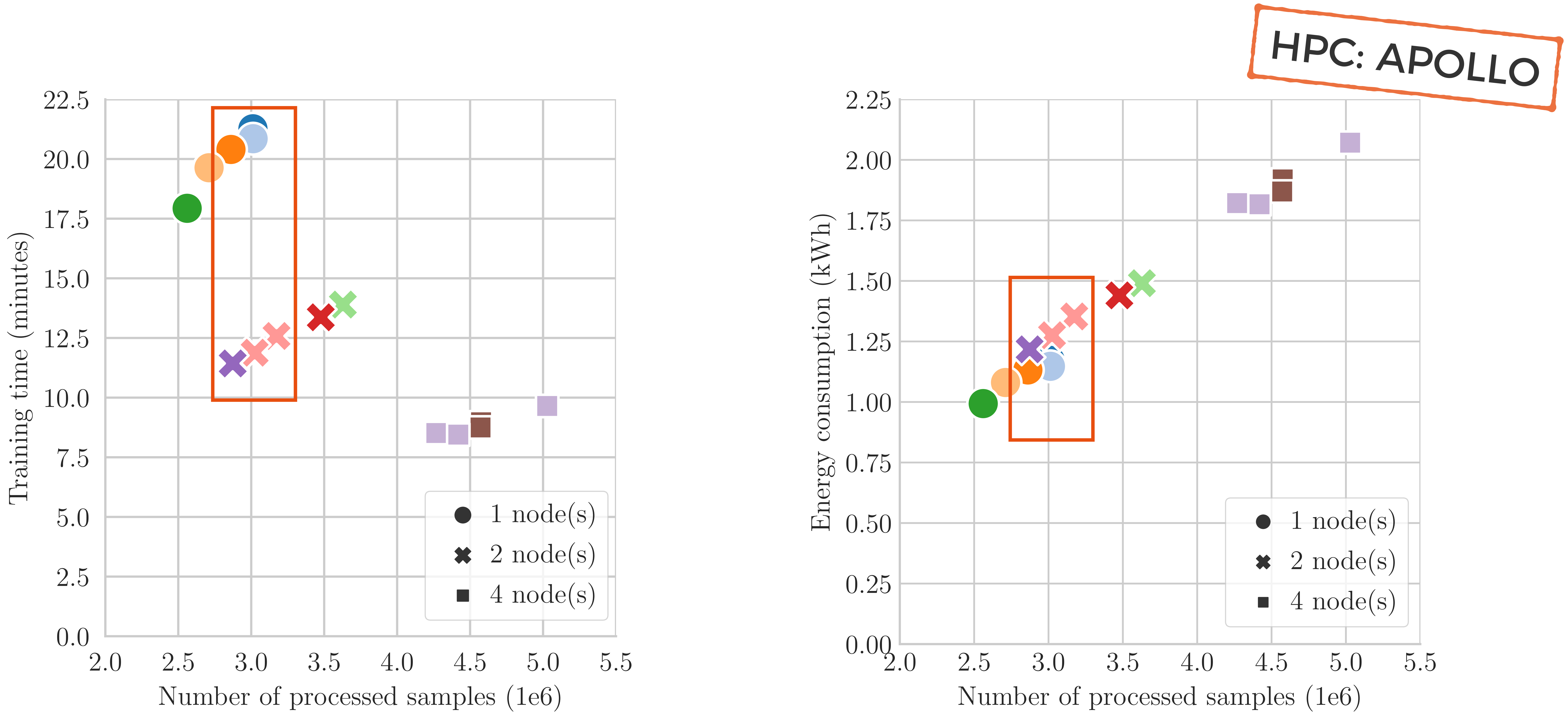








Increasing the number of nodes can reduce the training time but can also increase the energy consumed



1 color = 1 set of nodes

Impact of sizing up the number of Apollo node of Champollion
On the BERT FU

Embodied

Estimating the embodied impact

Of GPUs

Notations	
I	: Environmental Impact (expressed in kg CO2 eq, kg Seb, and MJ)
s	: Surface or area (of the GPU die or of the board)
c	: Memory capacity
d	: Memory density
w	: Node weight

Boavizta

$$I_{compute}(s_{die}) = s_{die} * I_{compute,manufactoring} + I_{compute,transport}$$

$$I_{memory}(c,d) = \frac{c}{d} * I_{memory,manufactoring} + I_{memory,transport}$$

$$I_{board}(s_{PCB}) = s_{PCB} * I_{board}$$

$$I_{GPU,capex} = I_{compute}(s_{die}) + I_{memory}(c,d) + I_{board}(s_{PCB}) + I_{HeatSink} + I_{PCIEConnector}$$

GWP: Global Warming Potential, ADP: Abiotic Depletion Potential, PE: Primary Energy

Estimating the embodied impact

Notations		
I	:	Environmental Impact (expressed in kg CO2 eq, kg Seb, and MJ)
s	:	Surface or area (of the GPU die or of the board)
c	:	Memory capacity
d	:	Memory density
w	:	Node weight

Boavizta

Of the CPUs & RAM: Directly from database

Of other components:

$$I_{case} = \frac{w}{\bar{w}} * \bar{I}_{case}$$

GWP: Global Warming Potential, ADP: Abiotic Depletion Potential, PE: Primary Energy

Estimating the impacts of the electricity consumption

$$I_{training,opex} = PUE * IF_{elec} * E_{training}$$

Notations		
I	:	Environmental Impact (expressed in kg CO2 eq, kg Seb, and MJ)
I_{capex}	:	Embodied Impact (manufacture, transport, and end of life)
I_{opex}	:	Operational Impact (usage)
IF_{elec}	:	Electricity mix Impact Factor
E	:	Electricity consumption
AUR	:	Active Utilization Rate
PUE	:	Power Usage Effectiveness of the data center

In this presentation, PUE = 1

GWP: Global Warming Potential, ADP: Abiotic Depletion Potential, PE: Primary Energy

Allocation on the ResNet-50 FU

Notations		
I	:	Environmental Impact (expressed in kg CO2 eq, kg Seb, and MJ)
I_{capex}	:	Embodied Impact (manufacture, transport, and end of life)
I_{opex}	:	Operational Impact (usage)
AUR	:	Active Utilization Rate
T	:	Use time of the equipment

$$I_{training,capex} = \frac{T_{training}}{AUR * T_{lifetime}} * (I_{GPU,capex} + I_{CPU,capex} + I_{RAM,capex} + I_{Other,capex})$$

Hypothesis

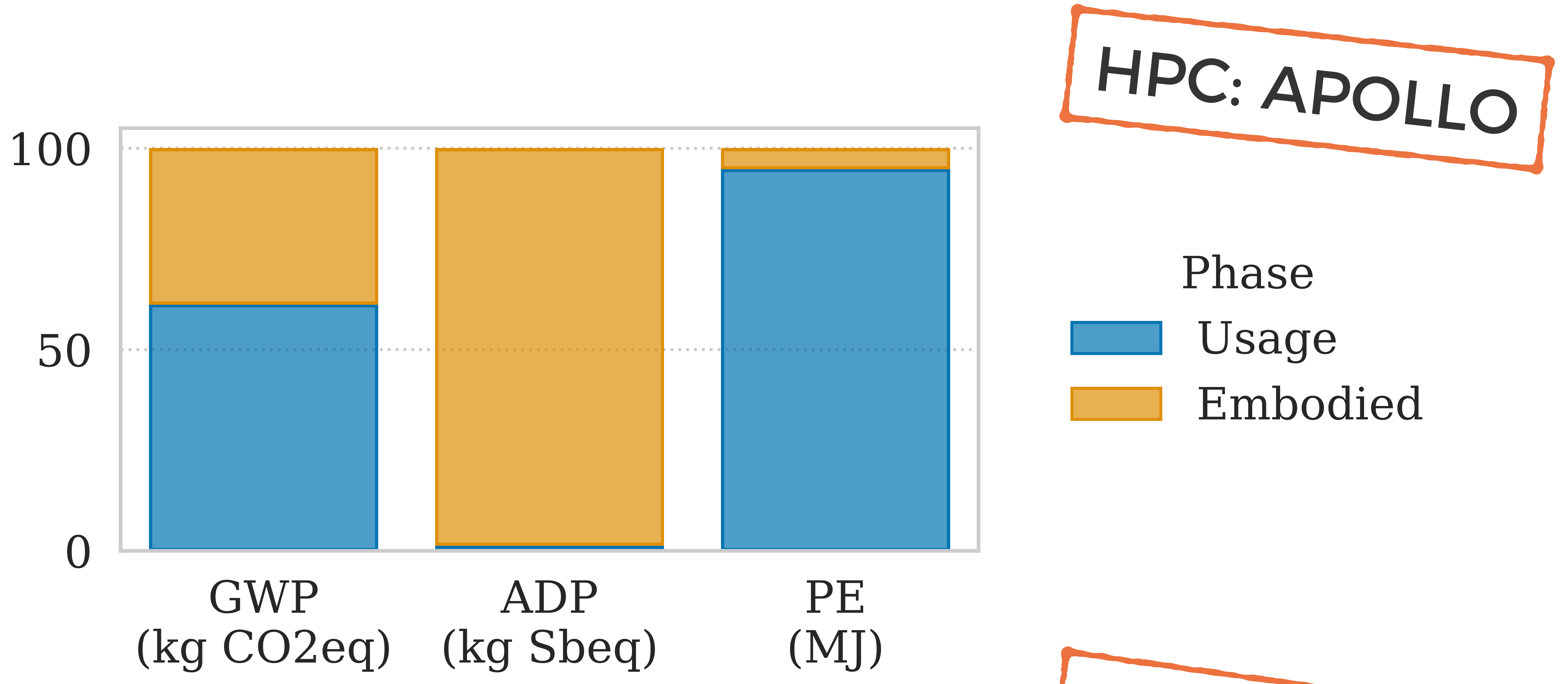
- AUR = Active Utilization Rate : 50%
- Lifetime : 4 years

GWP: Global Warming Potential, ADP: Abiotic Depletion Potential, PE: Primary Energy

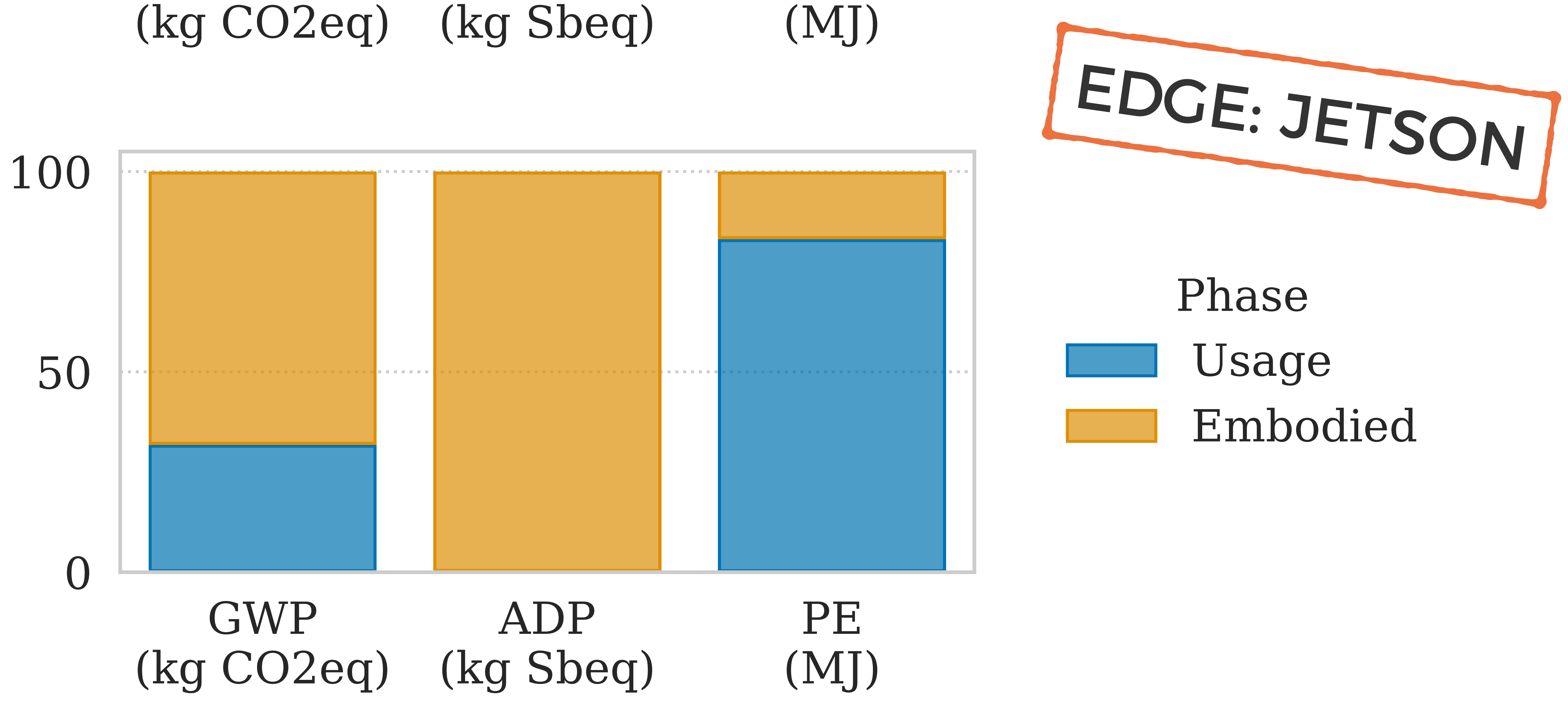
Total

Allocation on the ResNet-50 and ResNet-50* FU

	Usage	Embodied
GWP (kg CO2eq)	1.31E-01	8.30E-02
ADP (kg Sbeq)	7.80E-08	6.00E-06
PE (MJ)	1.96E+01	1.07E+00



	Usage	Embodied
GWP (kg CO2eq)	1.60E-01	3.43E-01
ADP (kg Sbeq)	9.59E-08	1.07E-04
PE (MJ)	2.41E+01	4.89E+00



Share of usage and embodied phase in the total impacts of the ResNet-50 * FU on the **Jetson** node

GWP: Global Warming Potential, ADP: Abiotic Depletion Potential, PE: Primary Energy

FU impacts on Apollo

GWP (kg CO2eq)	2.14E-01	2.30E-01	2.98E-01	2.62E-01	1.51E-01	2.28E-02
ADP (kg Sbeq)	6.08E-06	6.55E-06	9.00E-06	7.47E-06	4.30E-06	8.59E-07
PE (MJ)	2.07E+01	2.22E+01	2.79E+01	2.54E+01	1.46E+01	1.81E+00

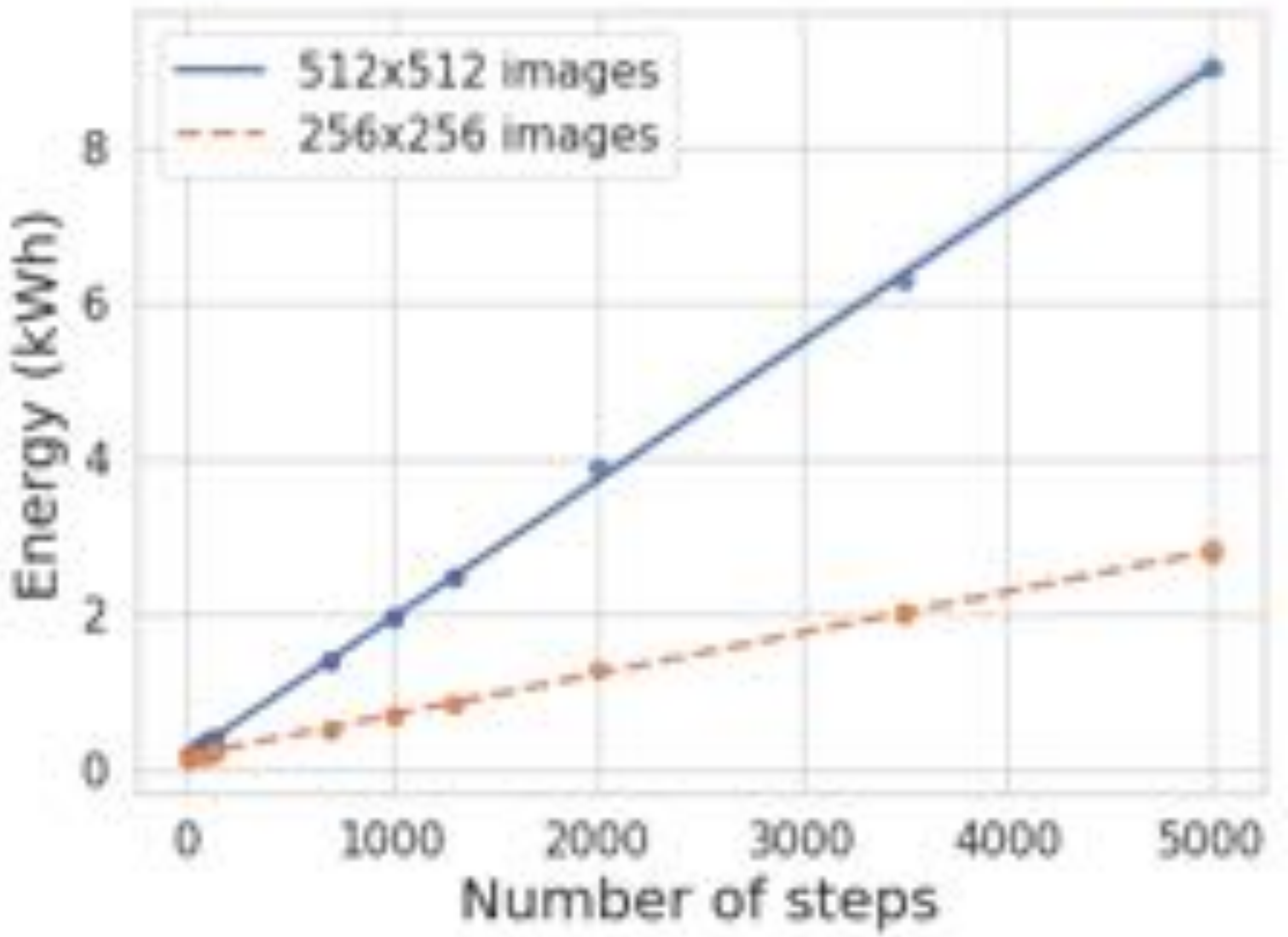
Summary of ResNet-50 and ResNet-50* FU performances

	Jetson (ResNet-50*)	Apollo (ResNet-50)	Apollo (ResNet-50*)	Min	Max
Electricity c (kWh)	1.974036108	1.605368	0.5723368444	3	0
Duration (hours)	88.86541726	0.4903240667	0.1876559211	100	0
Accuracy	0.55	0.759	0.55	0	1
GWP (kg CO2eq)	5.03E-01	2.14E-01	7.83E-02	0.70	0
ADP (kg Sbeq)	1.07E-04	6.08E-06	2.33E-06	0.00	0
PE (MJ)	2.90E+01	2.07E+01	7.39E+00	50.00	0
Personnalisation	0.9	0.3	0.3	0	1
Latency	0.9	0	0	0	1
Price	1	0.001	0.001	0	1
Accessibility	0.7	0.1	0.1	0	1

Stable Diffusion

Cluster	Champollion	Estats	Gemini	Sirius
Node model	Apollo 6500 Gen10	Nvidia Jetson AGX Xavier	Nvidia DGX-1	Nvidia DGX A100
Number of nodes	20	12	2	1
GPU model	NVIDIA A100-SXM-80GB	NVIDIA GV10B, Volta architecture	Nvidia Tesla V100-SXM2-32GB	Nvidia A100-SXM4-40GB
Number of GPU per node	8	1	8	8
GPU TDP (W)	400		400	280
CPU model	AMD EPYC 7763 64-Core Processor	Nvidia Carmel (Carmel), aarch64, 8 cores	Intel Xeon E5-2698 v4 (Broadwell, 64 cores/CPU)	AMD EPYC 7742 (Zen 2, 64 cores/CPU)
Number of CPU per node	2	1	2	2
CPU TDP (W)	280		135	225
Memory	1 TB	32 GB	1 TB	512 GiB
Switch model	Mellanox HDR Infiniband			
Number of switch	8			
Switch power consumption (W)	375			
Installation year	2022	2023	2019	2021
Available thought	HPE local network - slurm	Grid'5000 - OAR	Grid'5000 - OAR	Grid'5000 - OAR
Power meter	HPE iLO 5 + RAPL + NVML	Jetson-stats	RAPL/NVML + OmegaWatt+ BMC	RAPL/NVML + OmegaWatt+ BMC

Linear regression from smaller number of steps



Estimation the electricity consumption of training Stable Diffusion

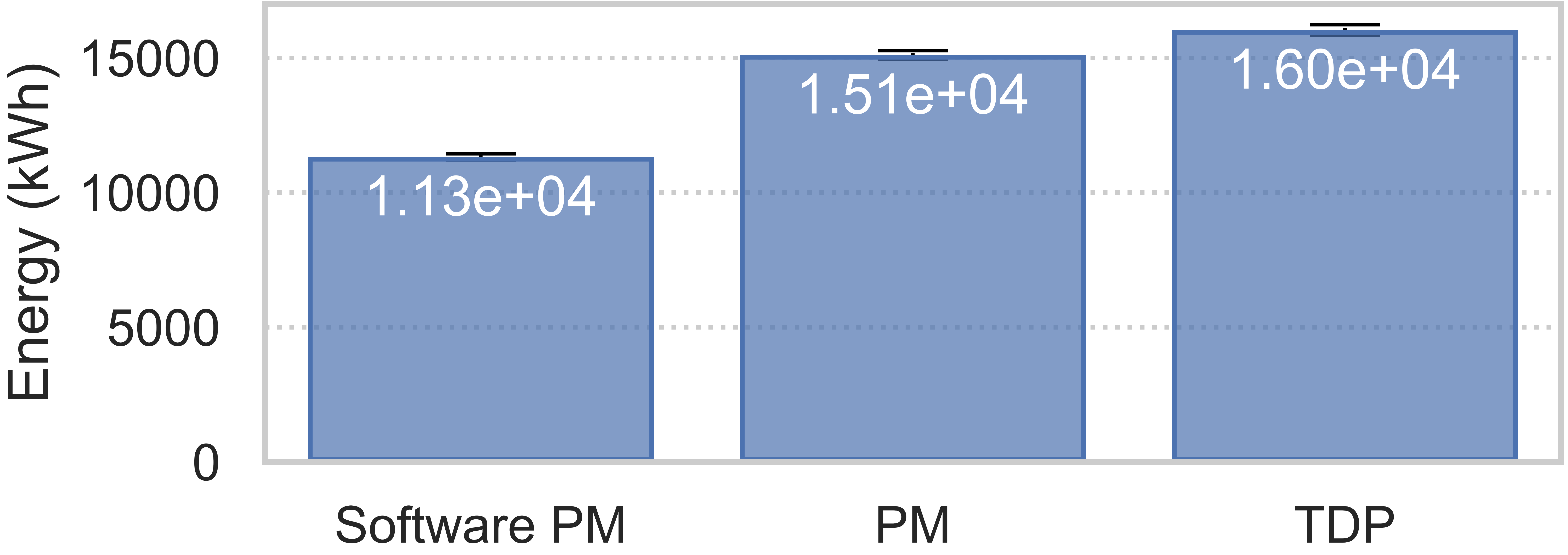
For 256x256 images $$\text{Energy (kWh)} = 5.26e^{-04} \times N + 2.01e^{-02}$$

For 512x512 images $$\text{Energy (kWh)} = 1.78e^{-03} \times N + 1.64e^{-02}$$

Where N is the number of training steps.

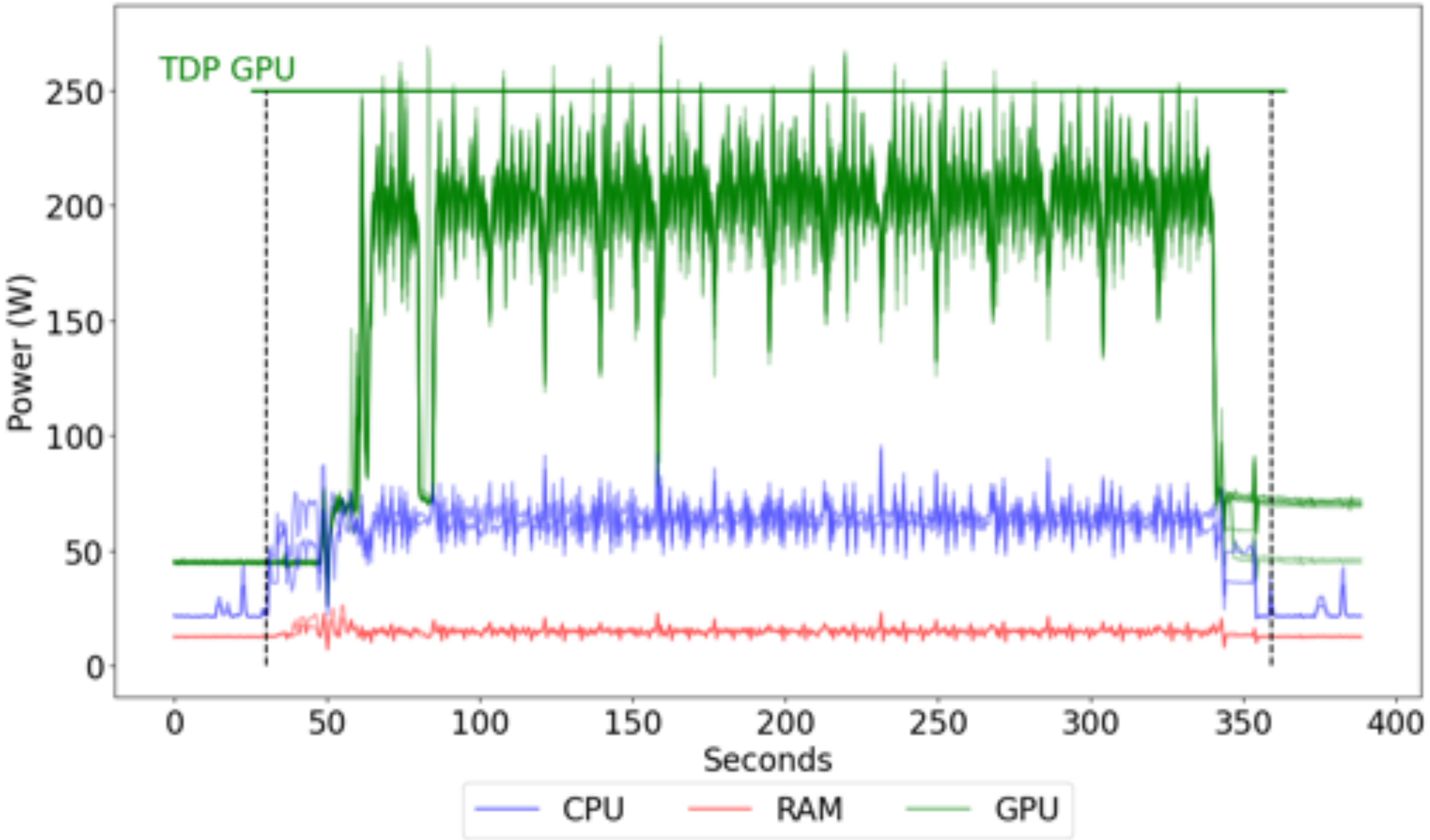
Version	Image size	# steps	Estimated energy (kWh) 1 node	32 nodes
v1-1	256	$2.37e^{+05}$	$4.70e^{+02}$	$1.50e^{+04}$
	512	$1.94e^{+03}$		
v1-4	512	$2.25e^{+05}$	$4.01e^{+02}$	$1.28e^{+04}$
v1-5	512	$5.95e^{+05}$	$1.06e^{+03}$	$3.39e^{+04}$

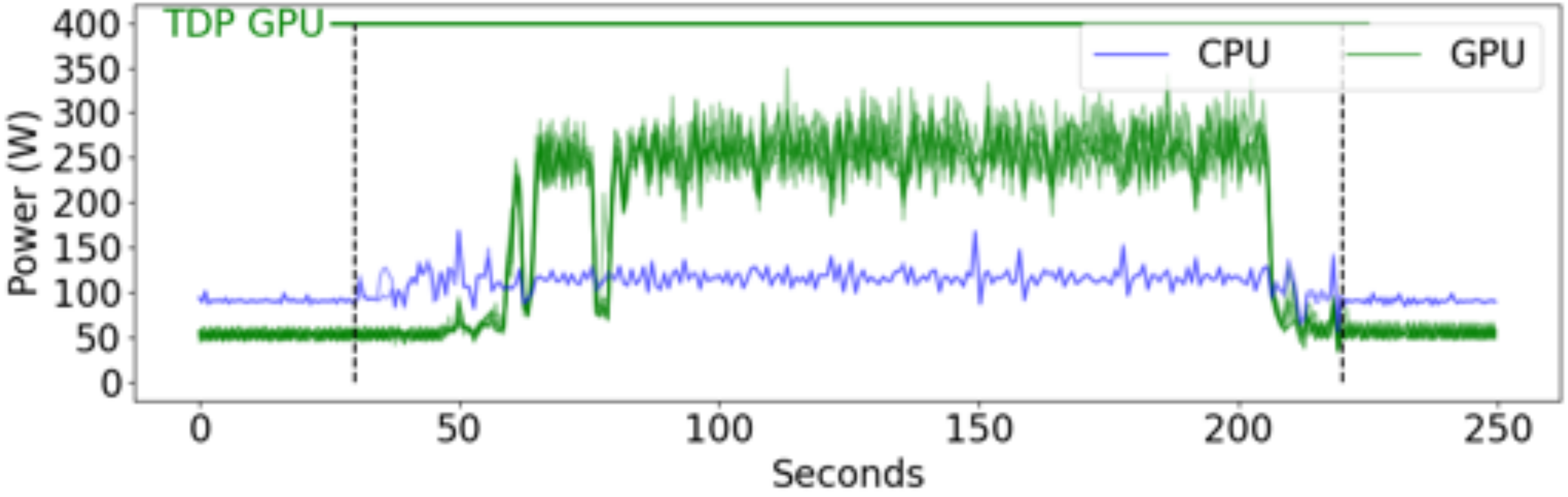
Impact of the energy measurement method



Estimation of the electricity cost of training Stable Diffusion with different measurement method
PM: Power Meter, **TDP**: Thermal Design Power

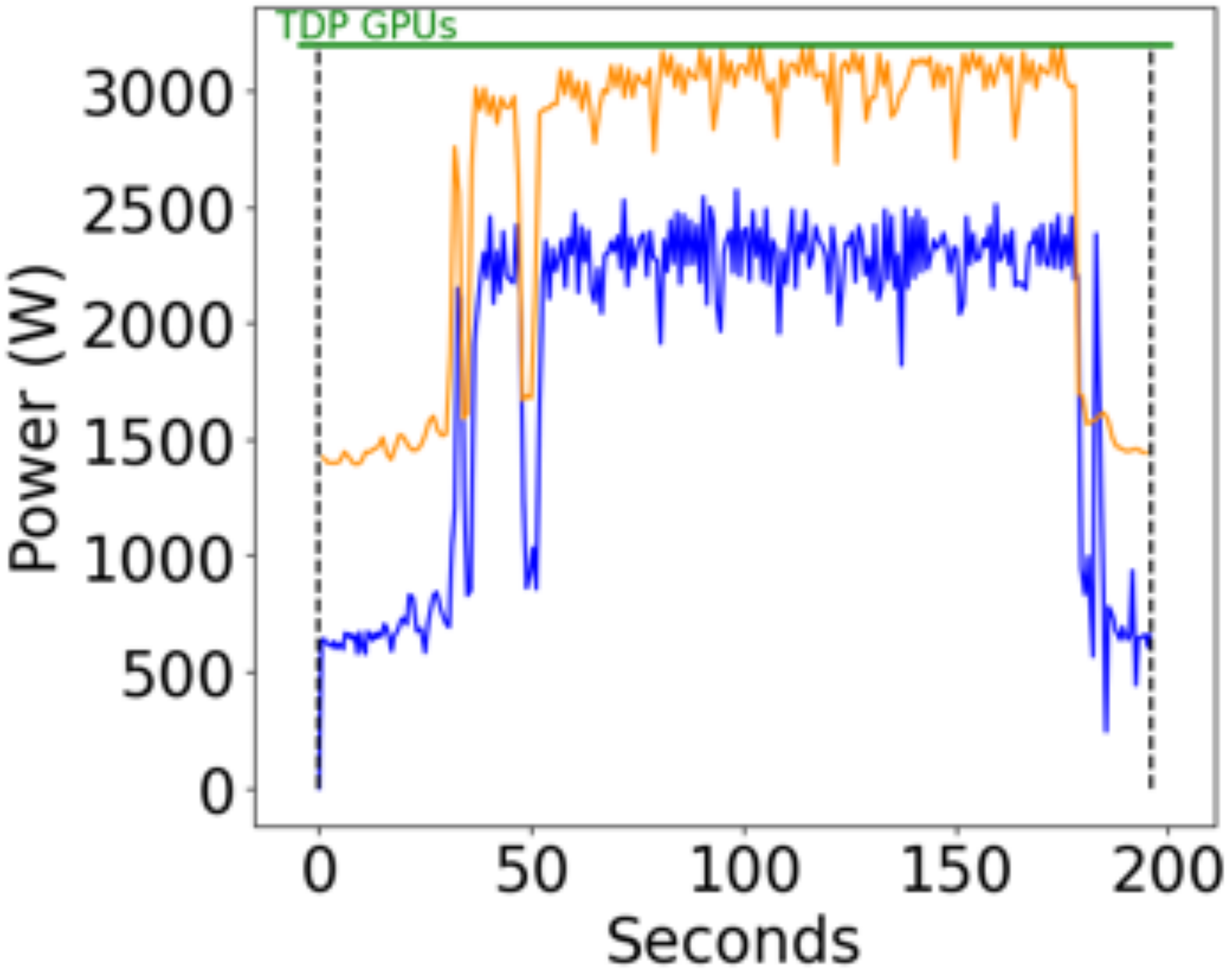
Training Stable Diffusion on the Gemini cluster



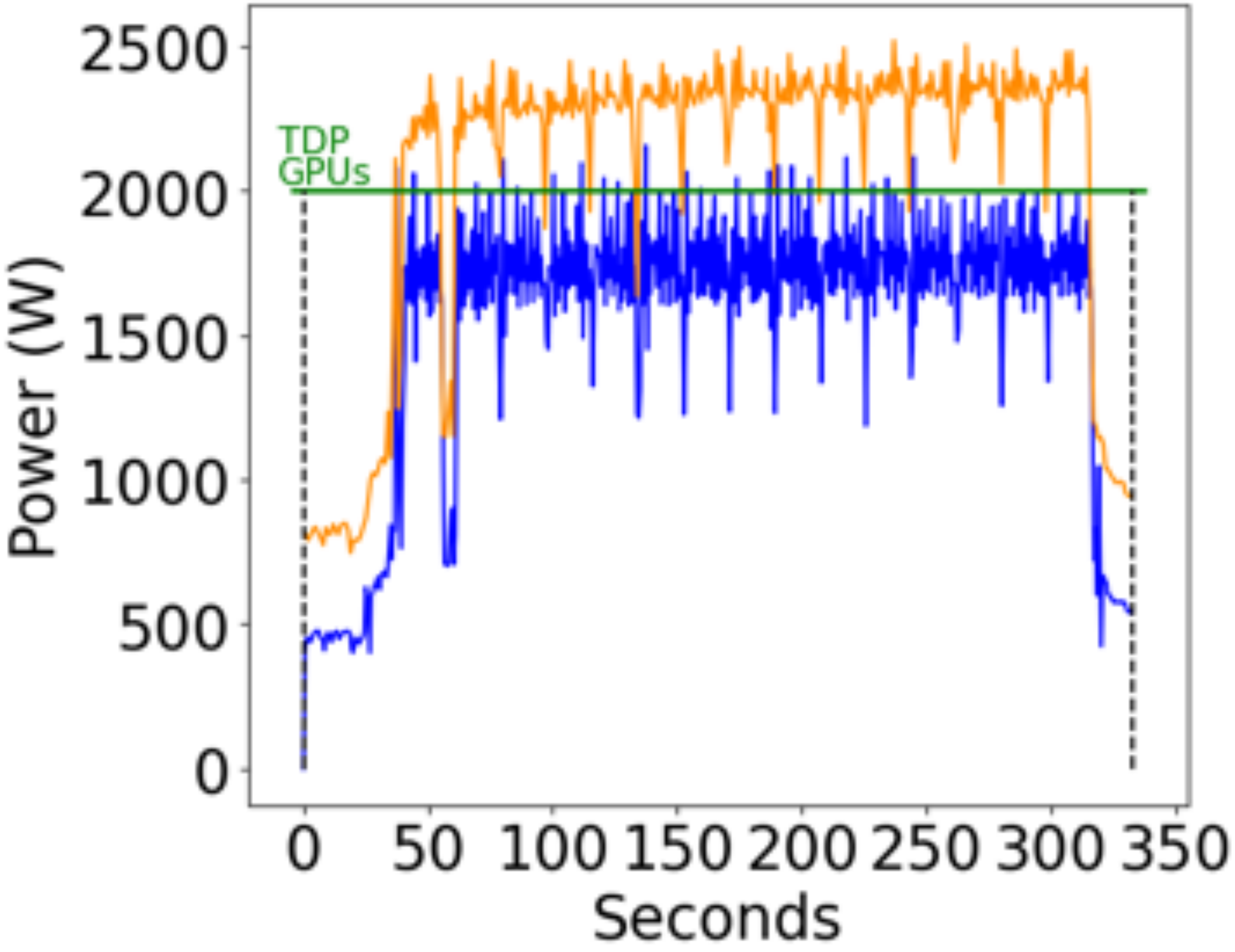


Grid'5000

Training Stable Diffusion on the Sirius cluster



Training Stable Diffusion on the Sirius cluster



Training Stable Diffusion on the Gemini cluster

Grid'5000

Reproducibility

Reproducibility

- Fixed environment
  - One node, and always the same
- Controlled environment
  - Fixed frequencies
  - Fixed power cap or mode
  - Minimalist software stack
  - Empty cache before each experiments
- Make sure the initial conditions are the same
  - Idle period between experiments

- Breach to reproducibility
  - Temperature in the Jetson cluster varies when I used all of them and that influenced the power draw
  - Didn’t fix the set of nodes when working on Champollion at first, thus we noticed a big difference in energy consumption