DARE-ML: DEMOCRATIZED ACCESSIBLE RESOURCE ENVIRONMENT FOR MACHINE LEARNING IN THE SUPERCOM PLATFORM

M. MENDULA, C. LEONELLI, M. MIOZZO, P. DINI

SUSTAINABLE ARTIFICIAL INTELLIGENCE
RESEARCH UNIT

Centre Tecnològic de
Telecomunicacions de Catalunya

MOTIVATION

- The size of AI models is continuously growing over time
- Only a small subset of players is capable of:
  - Training
  - Fine-tuning
  - or just even Evaluating

newest (and deeper) DL models based on attention mechanism.

MOTIVATION

- The more resources the better, but can we do more?
- Can we question our way of allocating resources to researchers?
- Most of them do not have the a "hardware aware" background:
  - Math
  - Physics
  - Statistics

SUPERCOM EXPERIMENTAL SETUP

- Supercom nodes are classified according with tiers:

Node Tier	Component	Specification
Bronze	CPU	2x Xeon CascadeLake 6230 2.1 GHz, 40 cores
	GPU	4x NVIDIA RTX 2080 TI, 11 GB GDDR6
	RAM	192 GB
	HD	2 TB SSD
Silver	CPU	32x Intel(R) Xeon(R) Silver 4314, 2.40 GHz
	GPU	GeForce RTX 3090, 24 GB
	RAM	125 GB
	HD	8 TB SSD + 1 TB NVME
Silver	CPU	Xeon IceLake Silver 4314, 2.4 GHz, 16 cores
	GPU	NVIDIA RTX 3090 BLOWER, 24 GB GDDR6X
	RAM	128 GB
	HD	8 TB SSD
Gold	CPU	2x Xeon IceLake Platinum 8358, 32 cores, 2.6 GHz
	GPU	NVIDIA RTX 3090 BLOWER, 24 GB GDDR6X
	RAM	1024 GB
	HD	2 TB SSD, 3.84 TB NVME

DARE-ML

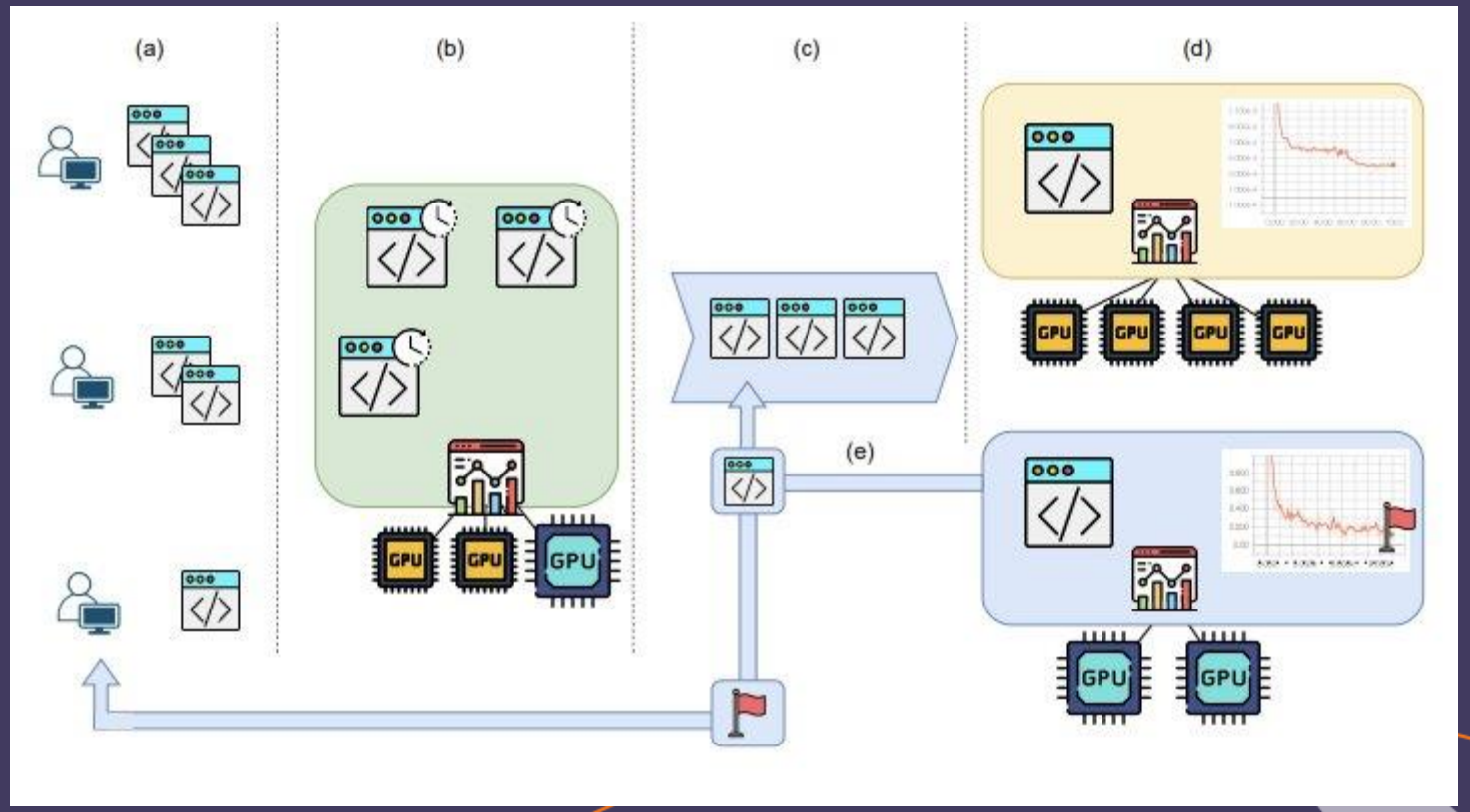
a) Users identify themself and getthe authorization to use theplatform

b) They get access to a subset of SUPERCOM resources to"profile" their workloads(sessions)

c) Jobs are scheduled in a Slurmfashion according to a FIFOqueue

d) Jobs are actually excuted andmonitored during training

e) When a target accuracy isachived the model goes back tothe user



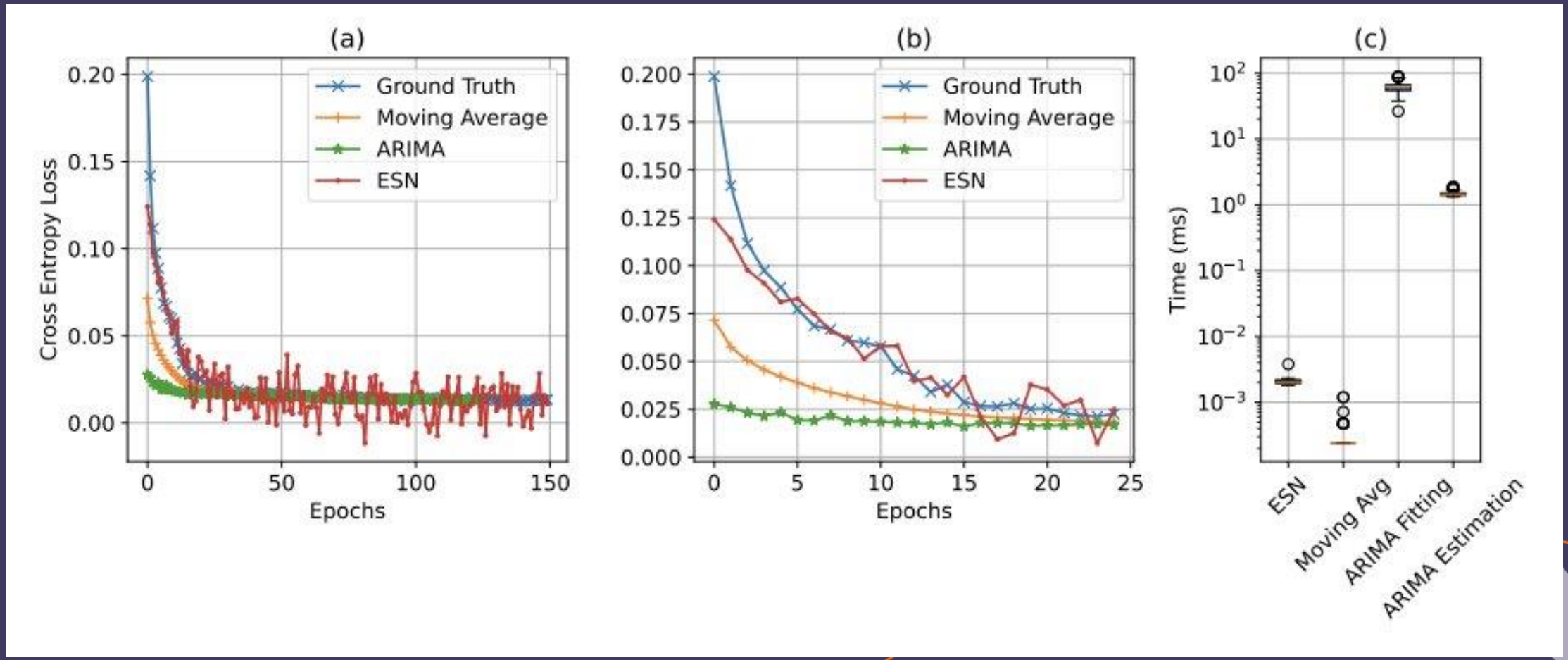
FINETUNING TASK

- Three different LLM are selected, each representing an incremental level of complexity interms of trainable parameters and virtual memory requirements:

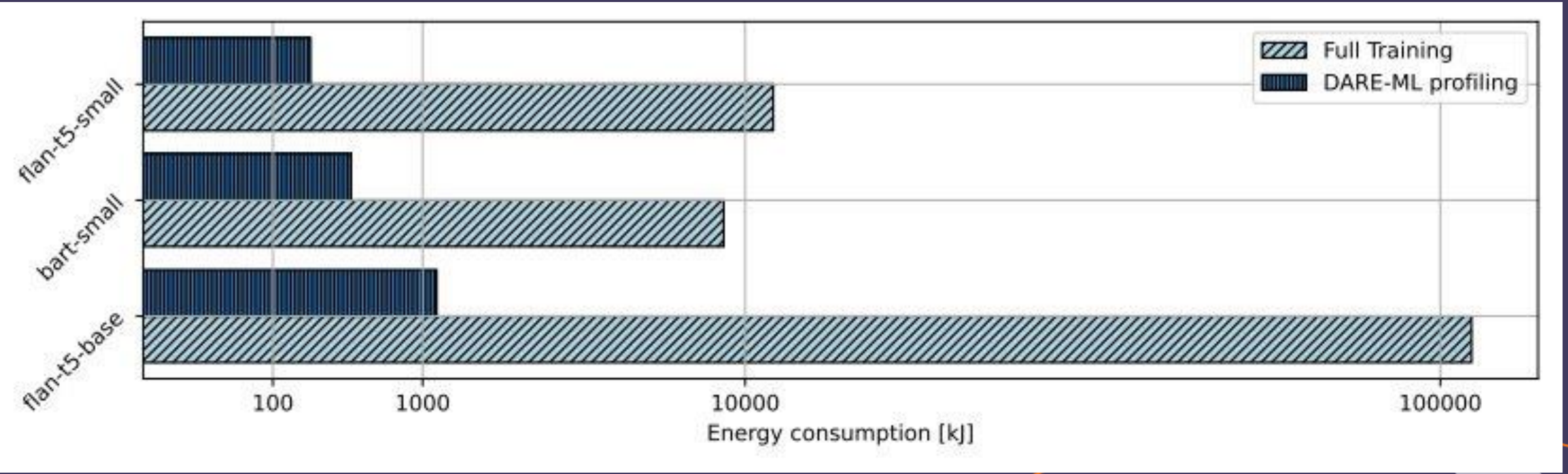
- lucadiliello/bart-small, 70.5M, **7GB**;
- google/flan-t5-small with 77M, **11GB**;
- google/flan-t5-base with 248M, **24GB**.

- We finetuned those models on **DialogSum**, a well-known dialog summarization dataset using LoRA+.
- Cross-entropy was chosen as the optimal loss function for all models tested. The learning rate is set to 2×10−4 , and the LoRA dropout rate to 0.1 across all models.
- Batch size is configured at 4 for the two smaller models and reduced to 2 for the mostmemory-intensive model.

PROFILING: FORECASTING MODEL PERFORMANCE



PROFILING: PROFILING VS FULL TRAINING



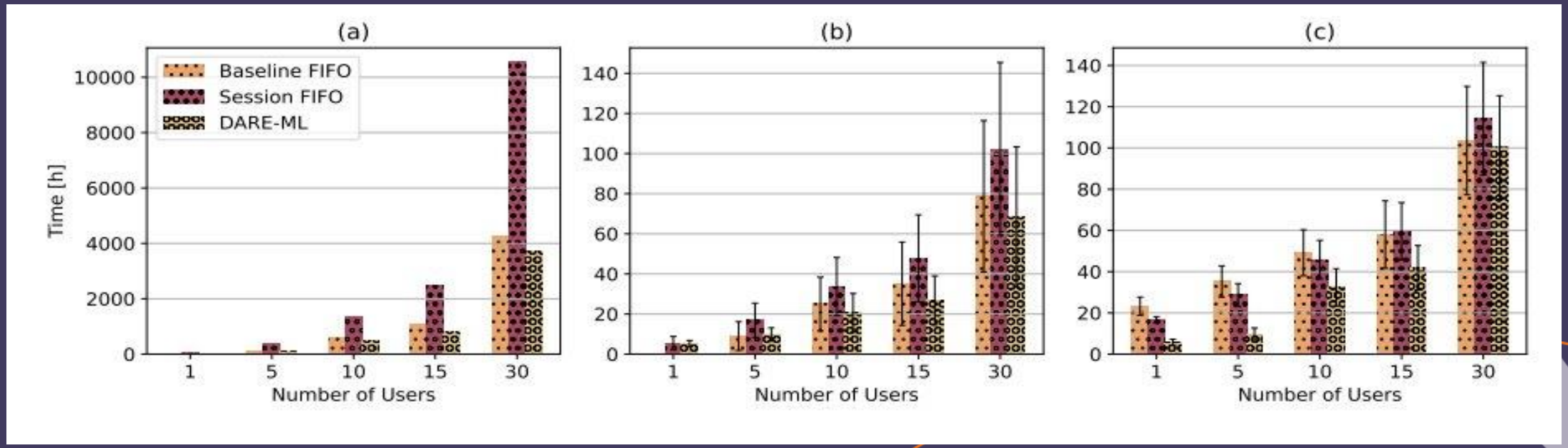
BALINES

- **Session FIFO:** Time limited sessions are dedicated to jobs, afterthe max time duration for each job, this is suspended and the nextjob execution is schedules

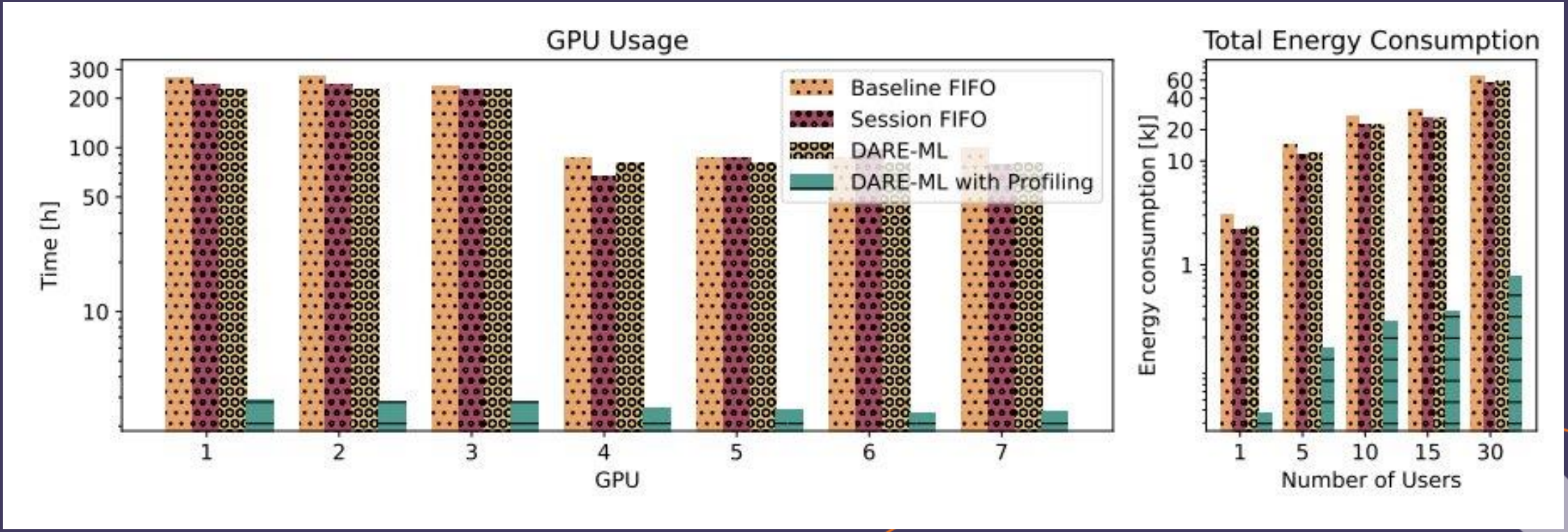
- **Baseline FIFO:** Jobs are executed until completion guaranteeingthe same order of arrival.

RESULTS: SCALING THE NUMBER OF USERS

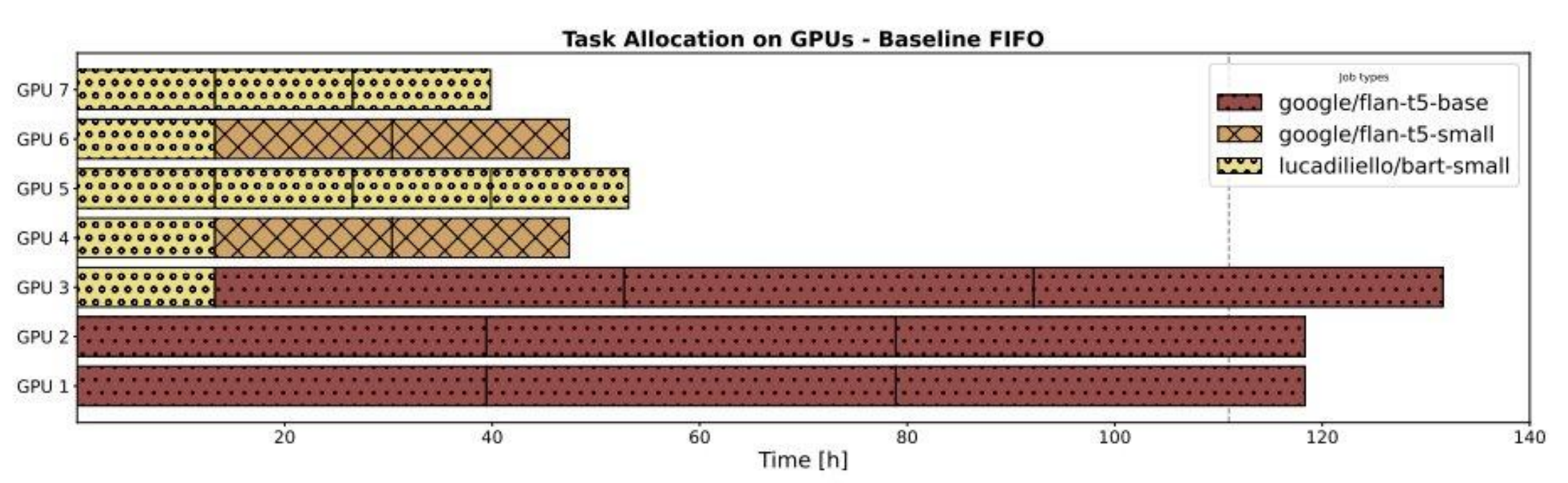
DARE-ML scheduler achieves the lowest total waiting time (a), minimizes average waiting time per user (b), and yields the shortest Avg. JCT (c) across all user counts.



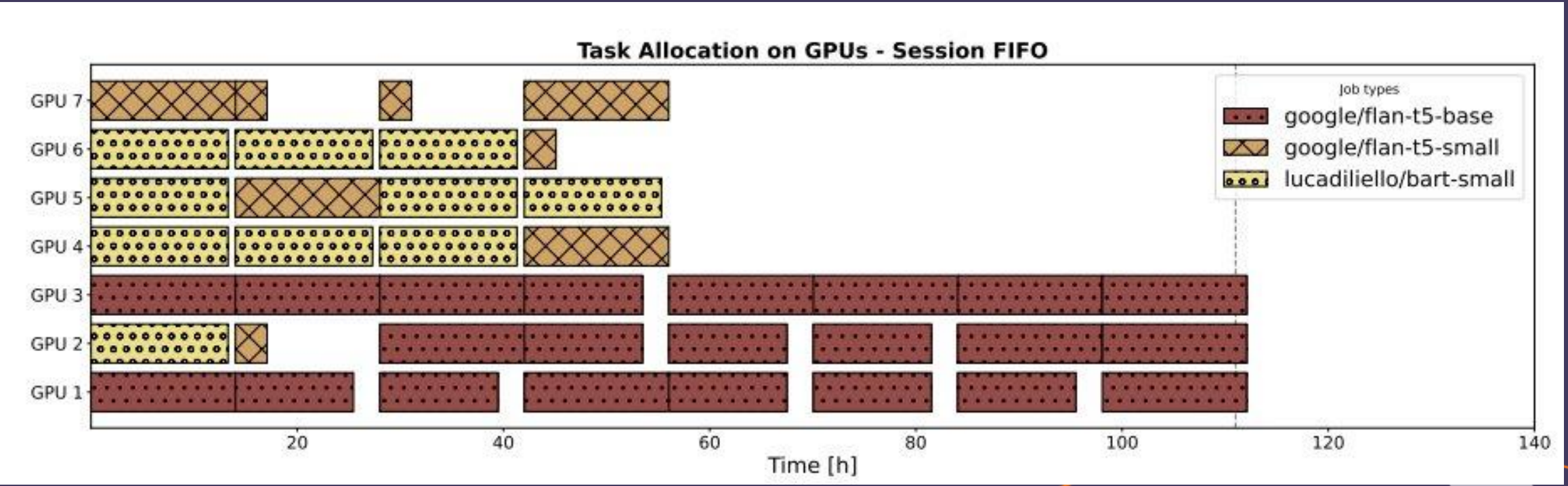
RESULTS: GPU USAGE AND TOTAL ENERGY CONSUMPTION



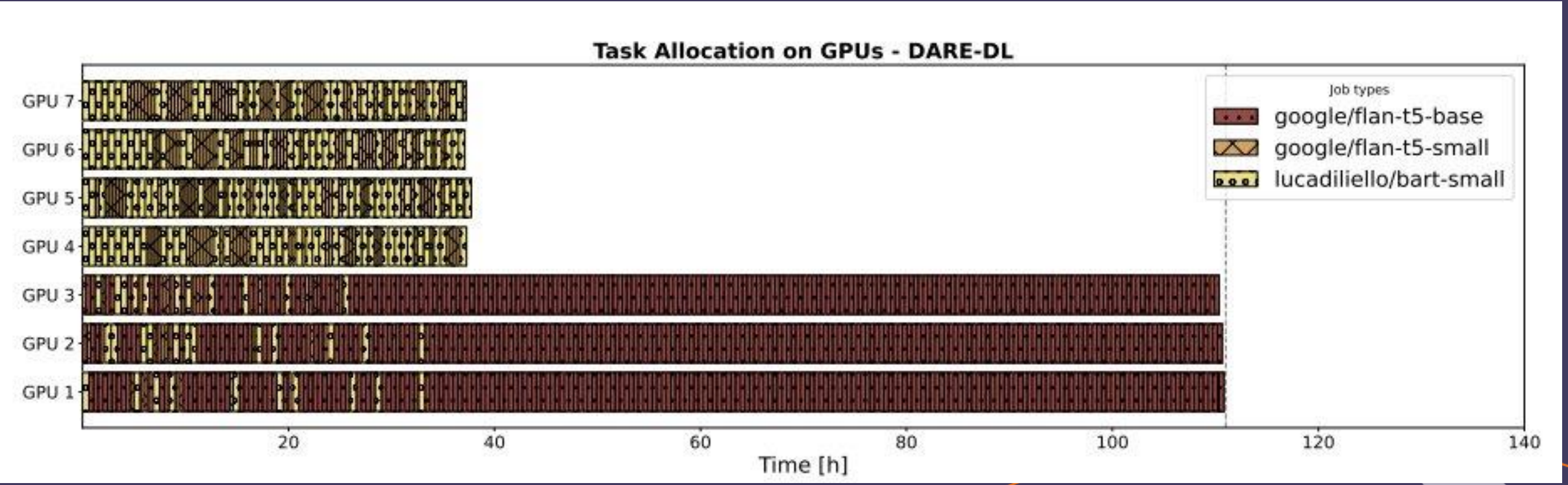
RESULTS: VISUAL SCHEDULING OF BASELINE FIFO



RESULTS: VISUAL SCHEDULING OF SESSION FIFO



RESULTS: VISUAL SCHEDULING OF BASELINE OF DARE-ML



CONCLUSIONS

- Model profiling using training reduces resource use, lowering average JCT and wait times by
  - 15% when retraining to overfitting,
  - and up to 80x with loss-aware interruptions.

- In addition, in high-demand cases, energy consumption drops byup to 83x.

THANK YOU FOR YOUR ATTENTION!

Dr. Matteo Mendula

Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)

mmendula@cttc.es

ctt c

Advanced research for everyday life

tecnio catalonia
ACCIÓ Generalitat de Catalunya
Fitxa # 31/12/2023

AENOR
GESTIÓ R+D+I
UNE 166002

HR EXCELLENCE IN RESEARCH

Place here author photograph