Synergy of Xena Vision and Supercomputing: A Transformative Duo

EUROHPC USER DAY 2023 Brussels 11.12.23

EuroHPC Joint Undertaking

Project “EU2022D10-008: Realtime Emergency Recognition via AI Powered Surveillance”

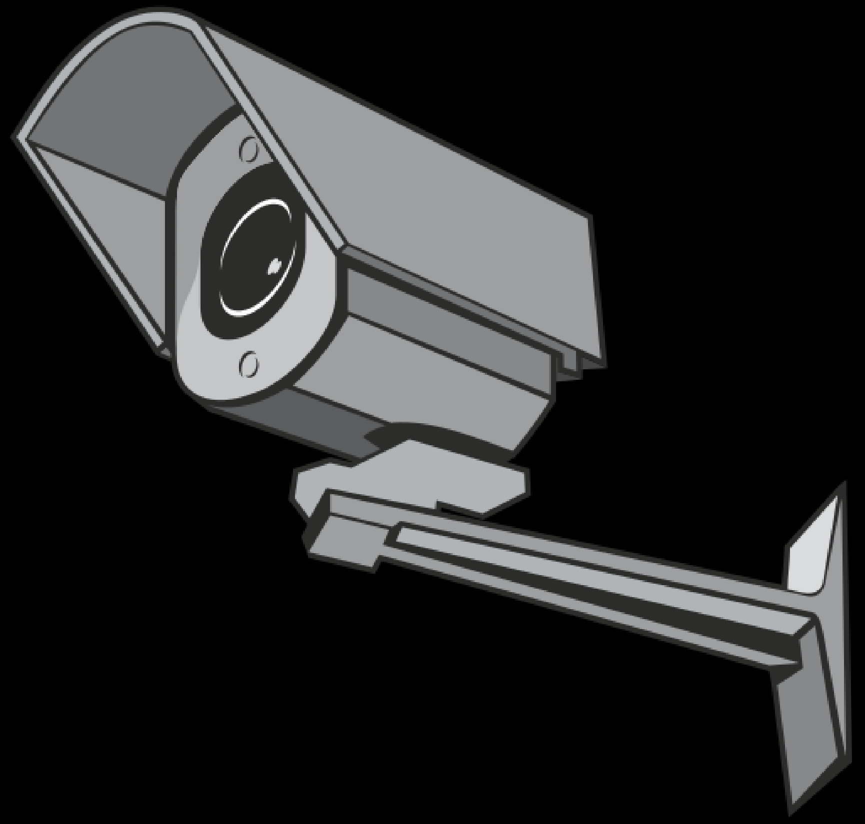
EuroHPC usedKAROLINA

Speaker: NAZLI TEMUR (XENA VISION)

EuroHPC
Joint Undertaking

XENA VISION

Content



XENA VISION

1. Send **Voice Message** through **Xena Crime-Hub** for 2 seconds

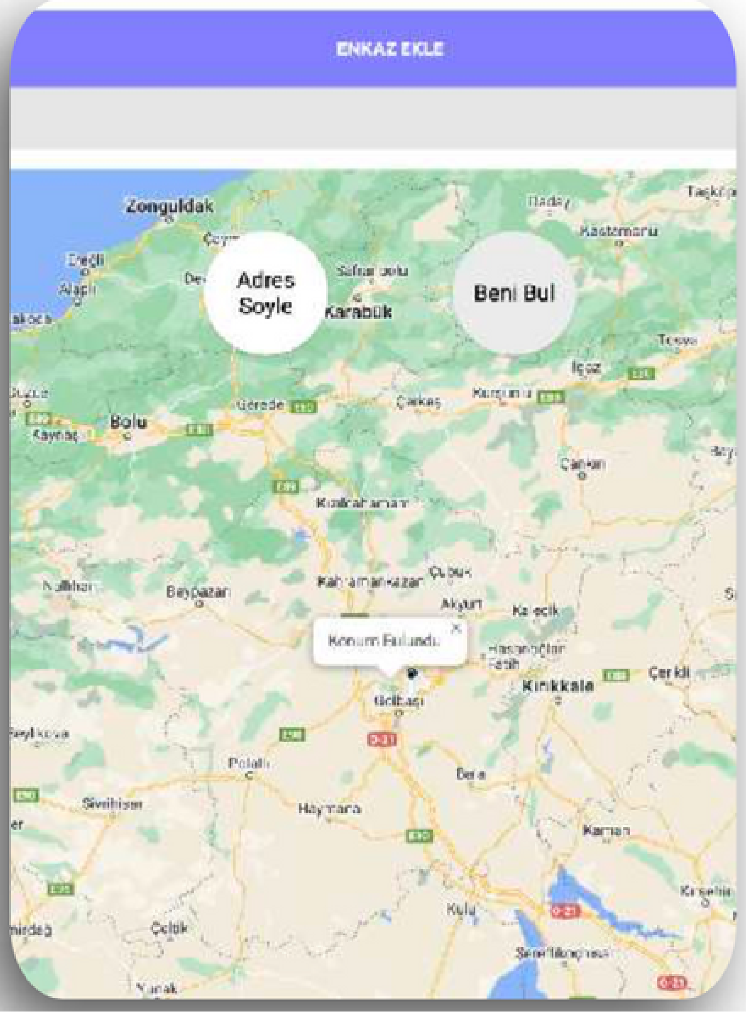
2. **Xena** Realtime **AI System** auto detect emergency in reported location CCTV in 2 seconds

3. Police accesses the realtime footage in the most advanced **Emergency Response & Control Software of Xena in 4 seconds**

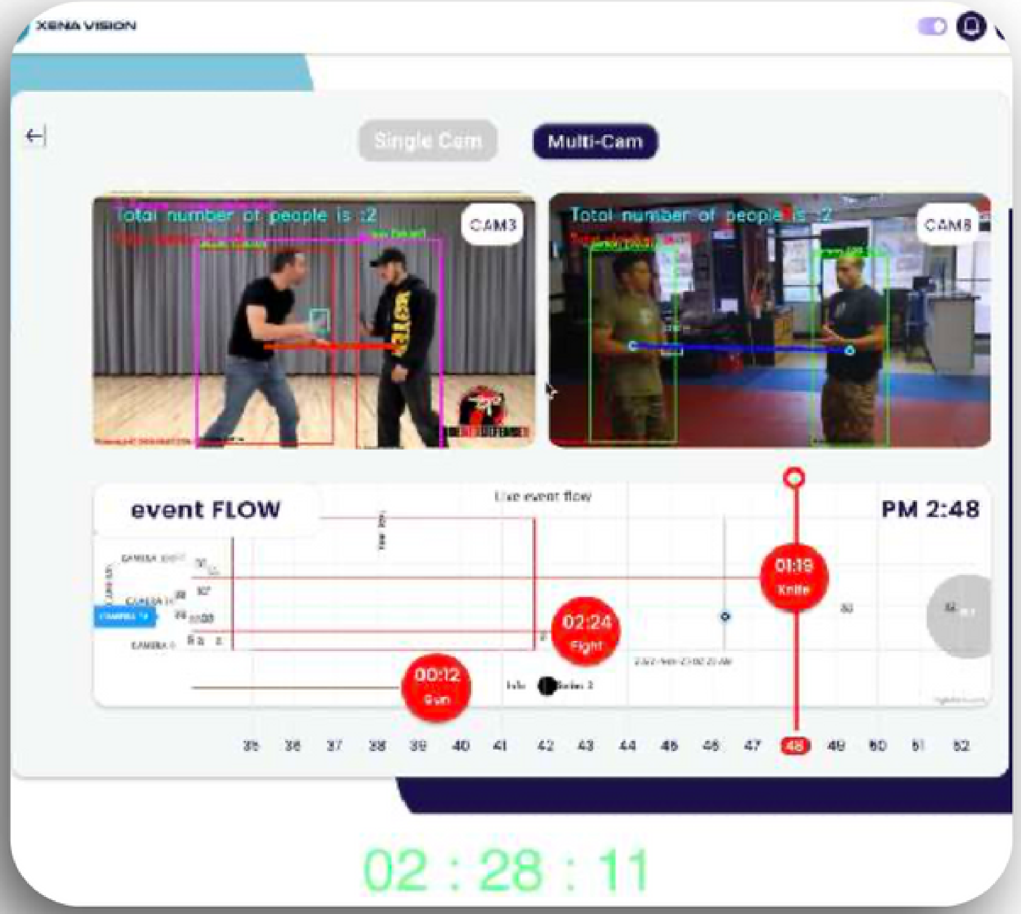
E2E Crime Prevention (~8 seconds)

1

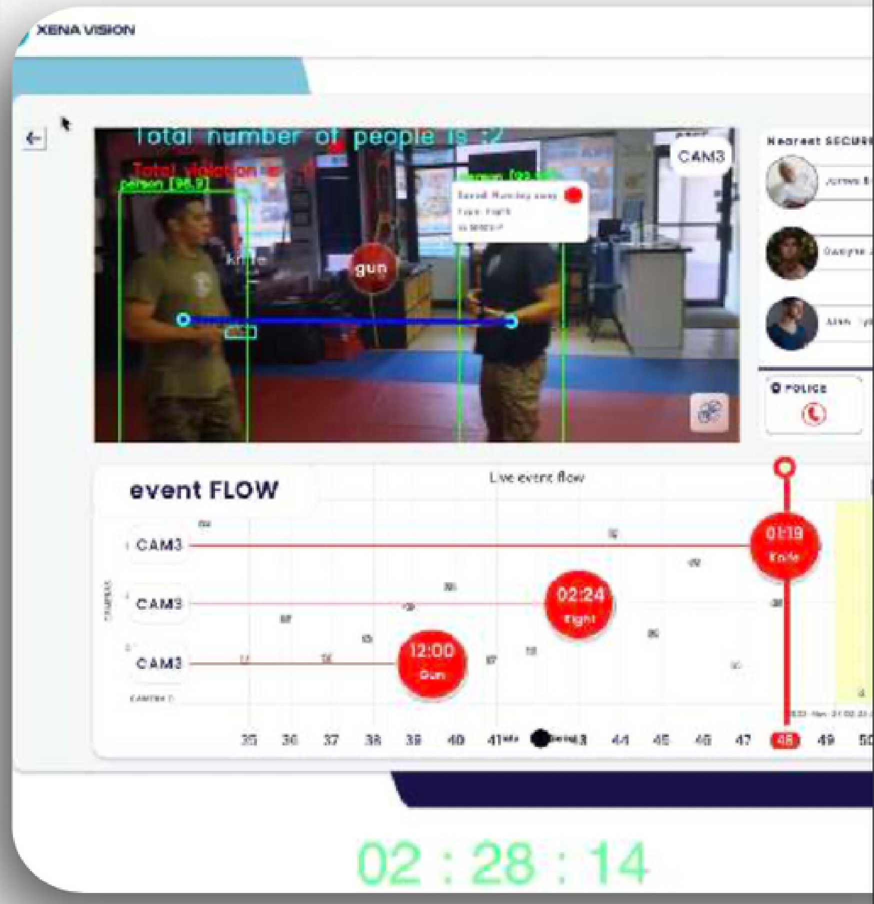
2



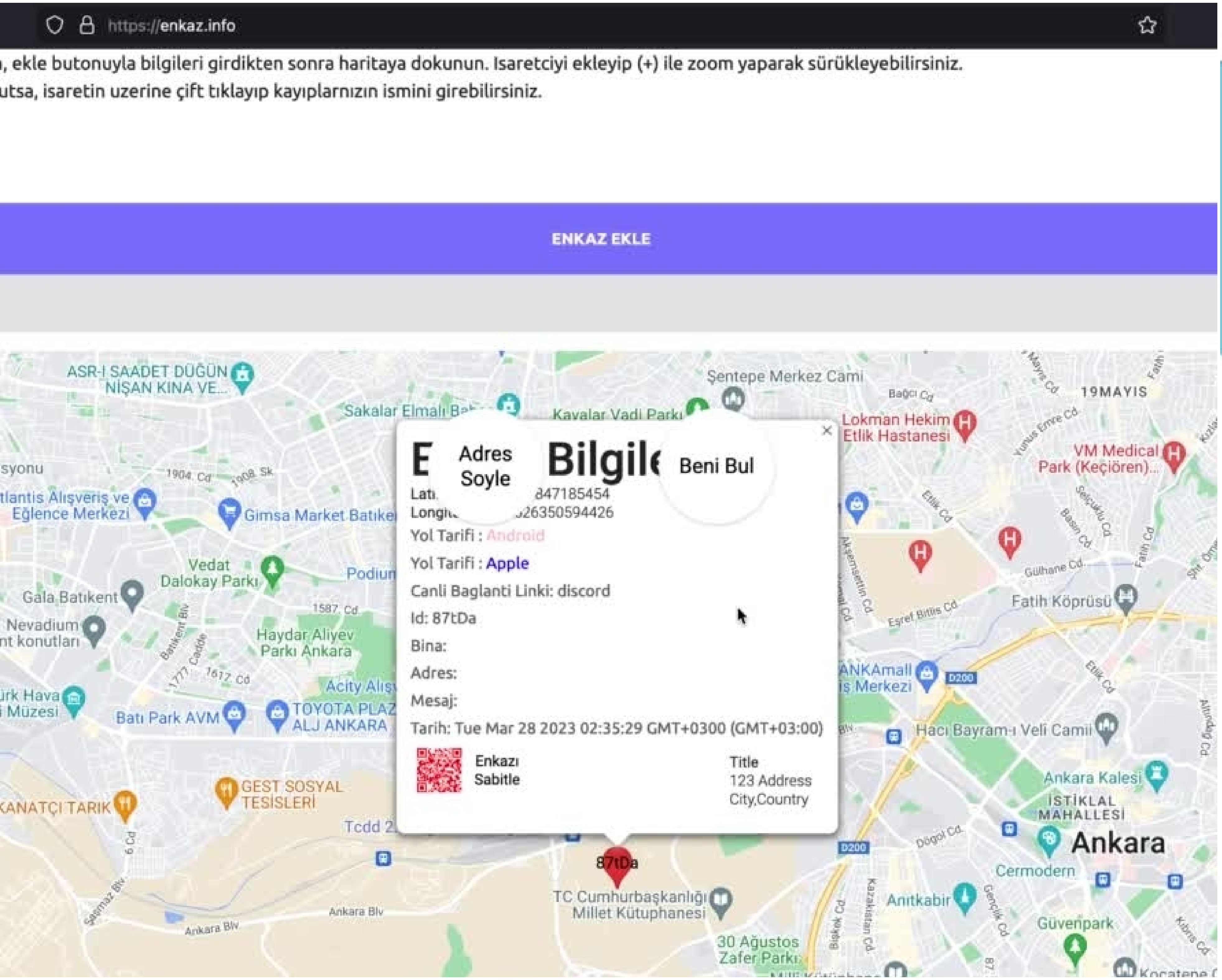
3

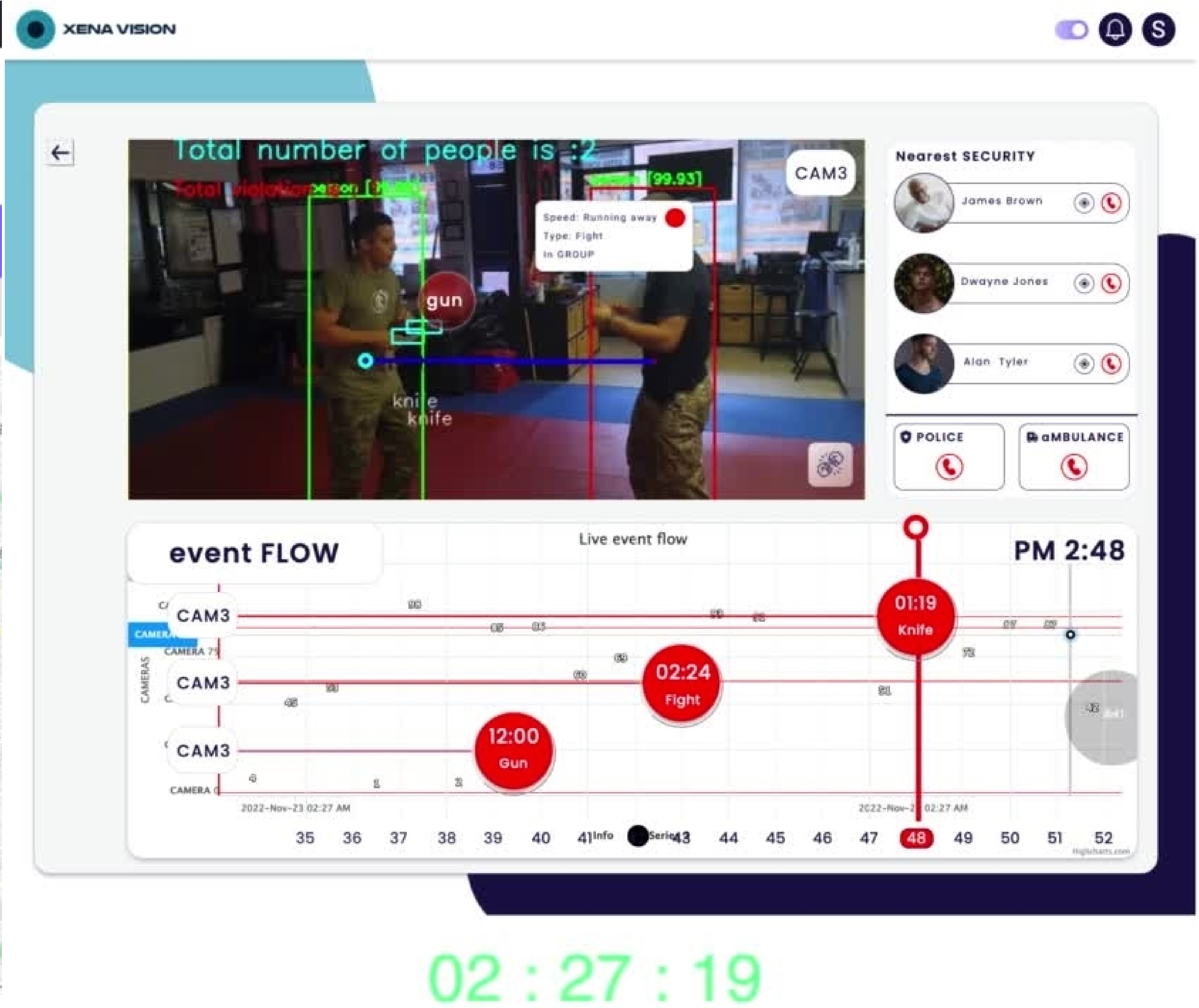


4



In Scientific Terms
EU2022D10-008: Realtime Emergency Recognition via AI Powered Surveillance





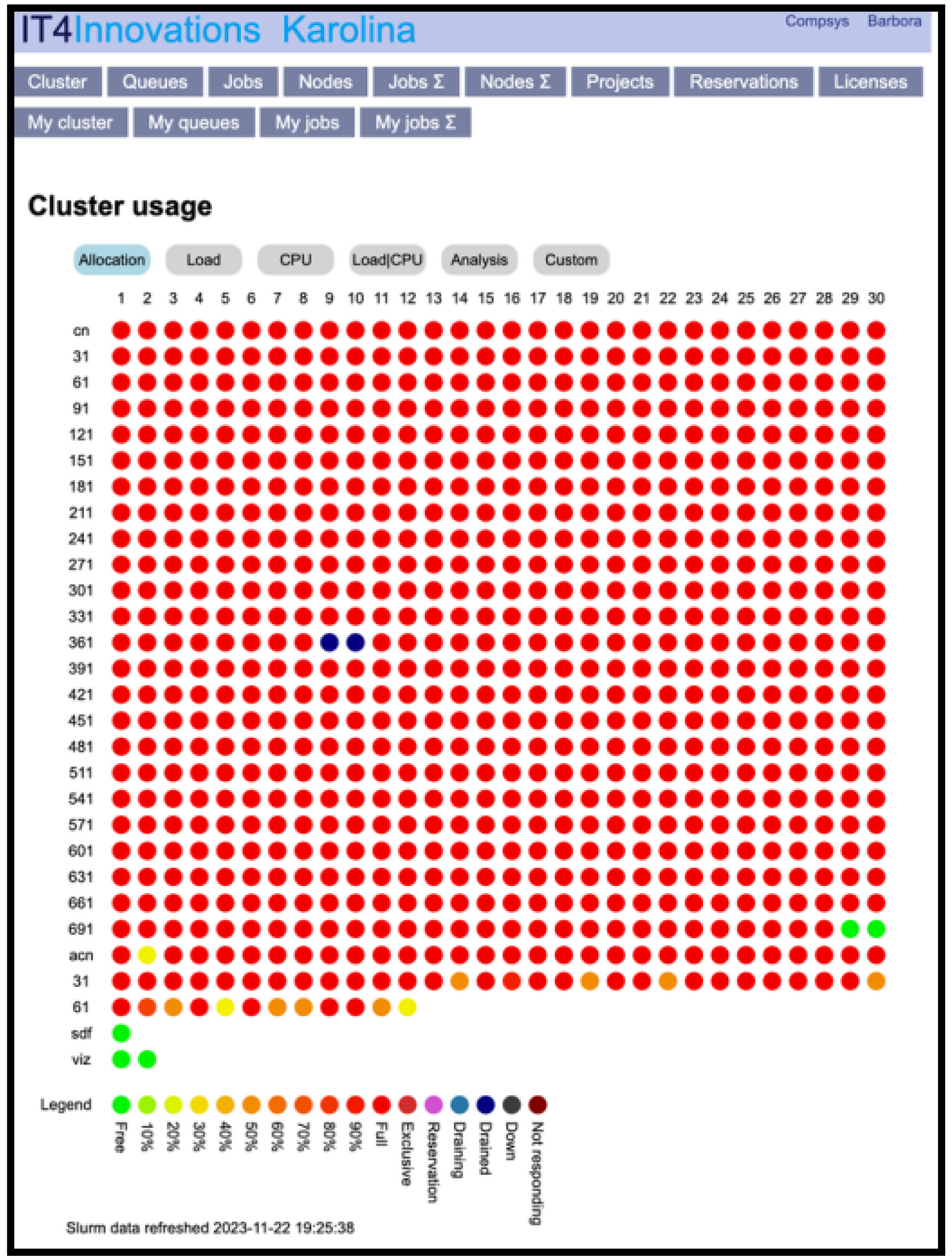
SUPER COMPUTING POWER

EUROHPC USER DAY 2023 Brussels 11.12.23

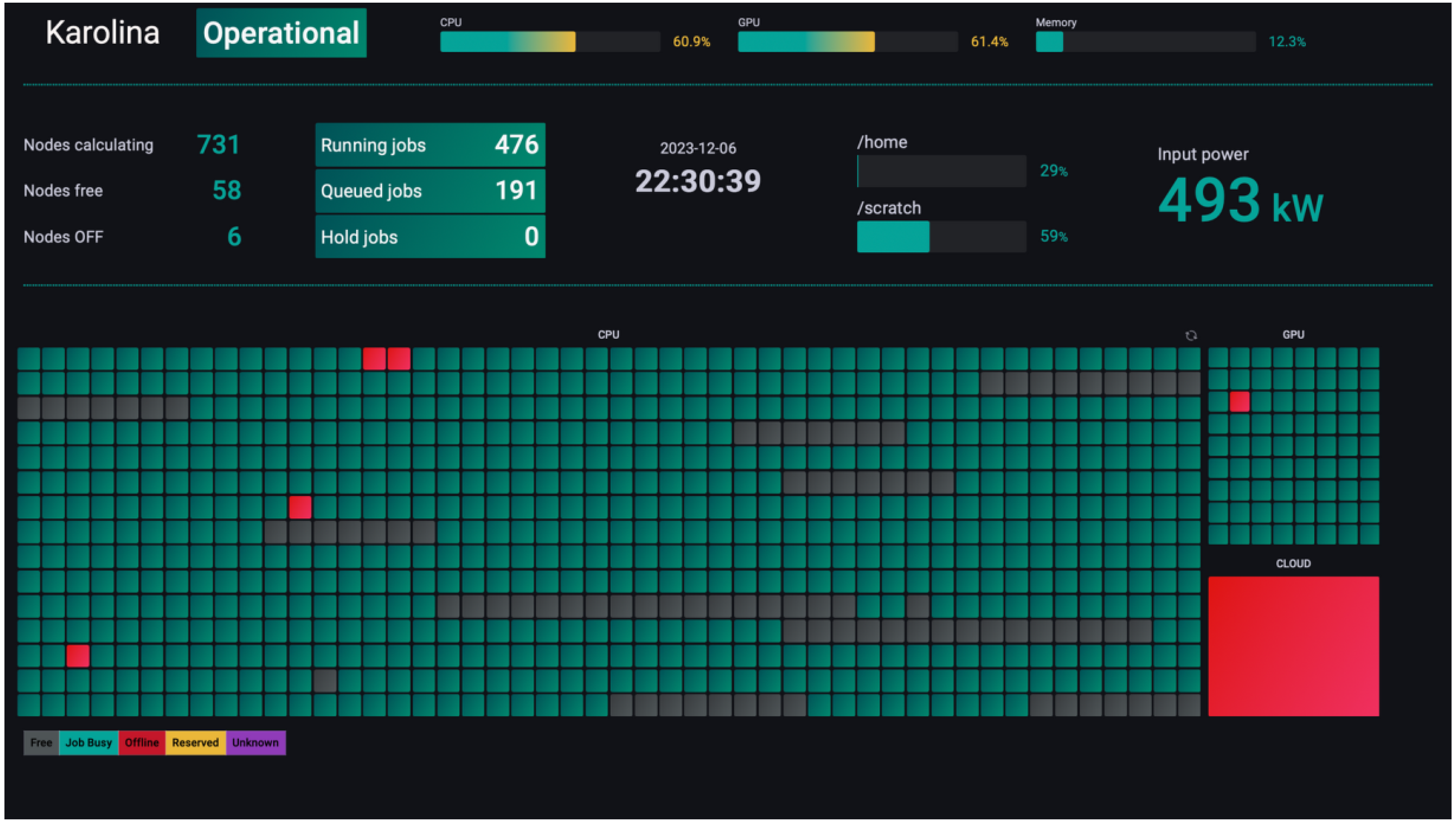
EuroHPC Joint Undertaking

In Scientific Terms
EU2022D10-008: Realtime Emergency Recognition via AI Powered Surveillance

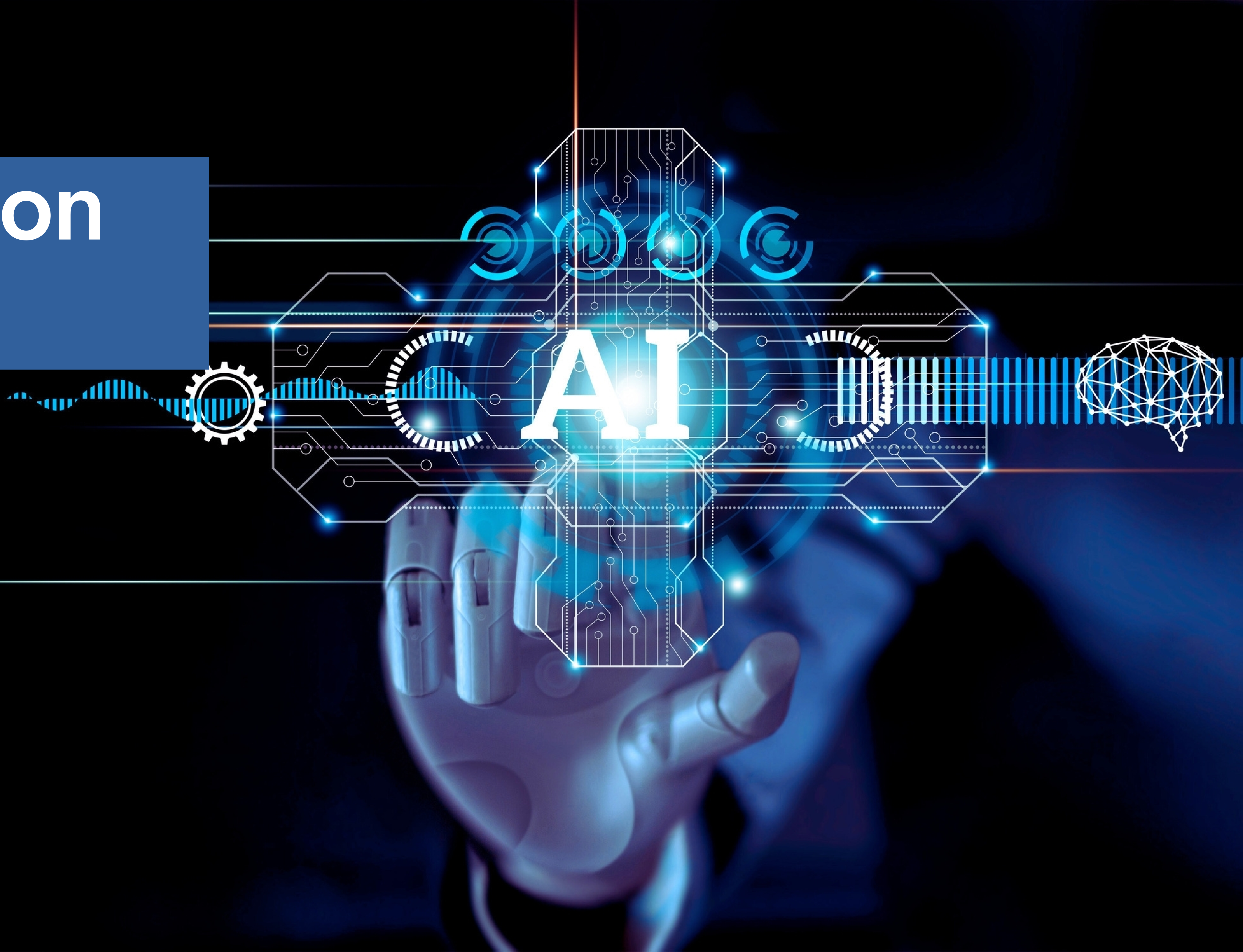
UTILIZATION



UTILIZATION



Integration Benefits

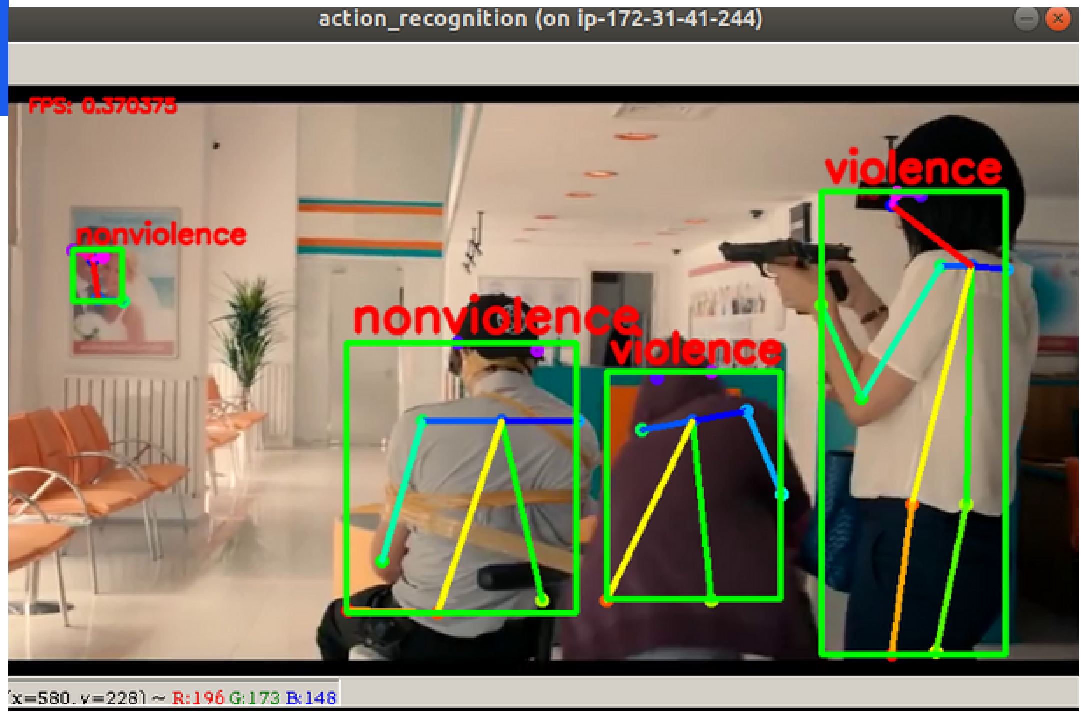


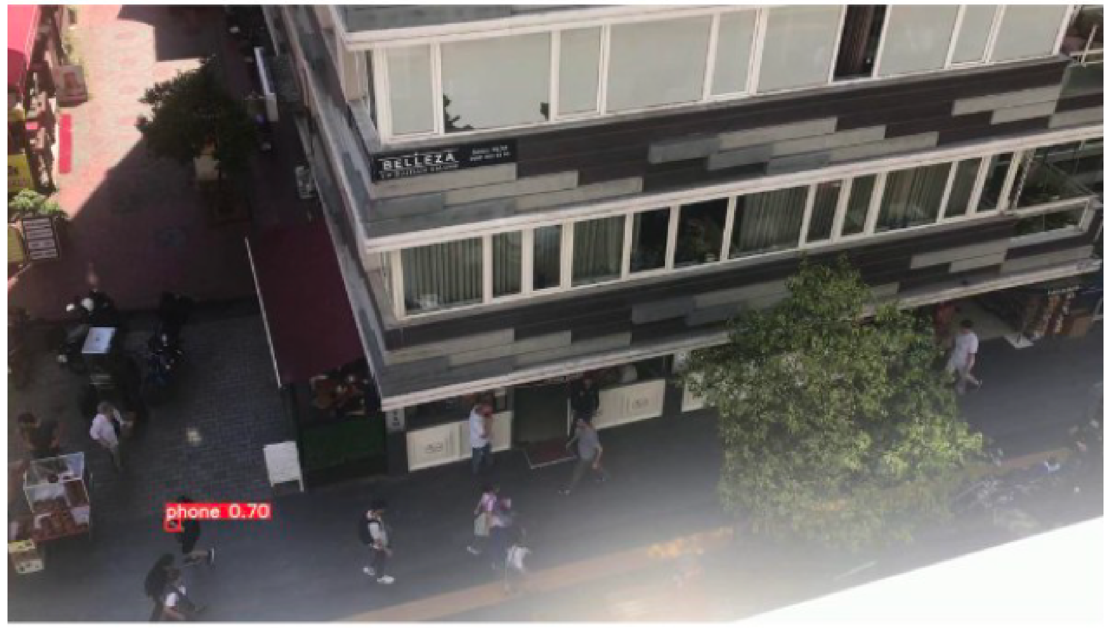
EUROHPC USER DAY 2023 Brussels 11.12.23

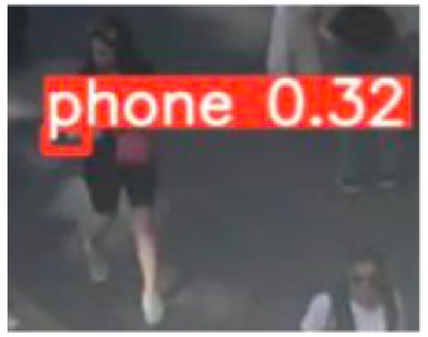
EuroHPC Joint Undertaking

EUROHPC USER DAY 2023 Brussels 11.12.23

EuroHPC Joint Undertaking







1- ARCHITECTURE FOCUSED MODEL GENERATION

Stage 2				Stage 1					
XNN	Training Set	Validation Set	Epoch	Pretrained Epoch	Pretrained	Backbone	Dataset	Training	Test Size
Face			15000			Resnet18	CrowdHuman		
Body	25831	6529	15000			Resnet18			
Knife	3535	962	15000		1000	Resnet18+Attention	Owned	4497	63813
Pose	xx		15000	500000	150000	VGG19+Openpose	Coco 2017		

2- DATASET FOCUSED MODEL GENERATION for KNIFE - LATE STAGE -EXTREME DATASET

Label	Precision	Recall	TP	FP	FN	TN	F1	Accuracy
428	0.919	0.825	353	31	75	0	0.869	0.769
425	0.927	0.776	330	26	95	0	0.844	0.731
427	0.846	0.770	329	60	98	0	0.806	0.675
425	0.903	0.772	328	35	97	0	0.832	0.712
853	0.894	0.788	672	80	181	0	0.837	0.720
1280	0.873	0.822	1052	153	228	0	0.846	0.734
1282	0.897	0.854	1095	126	187	0	0.874	0.777
1248	0.915	0.869	1085	101	163	0	0.891	0.804
1277	0.945	0.855	1092	64	185	0	0.897	0.814
3646	0.956	0.932	3398	156	248	0	0.943	0.893
127	0.887	0.803	102	13	25	0	0.842	0.728
528	0.920	0.829	438	38	90	0	0.872	0.773
453	0.563	0.471	213	166	240	0	0.512	0.344
475	0.686	0.732	348	159	127	0	0.708	0.548
1063	0.917	0.816	867	79	196	0	0.863	0.759

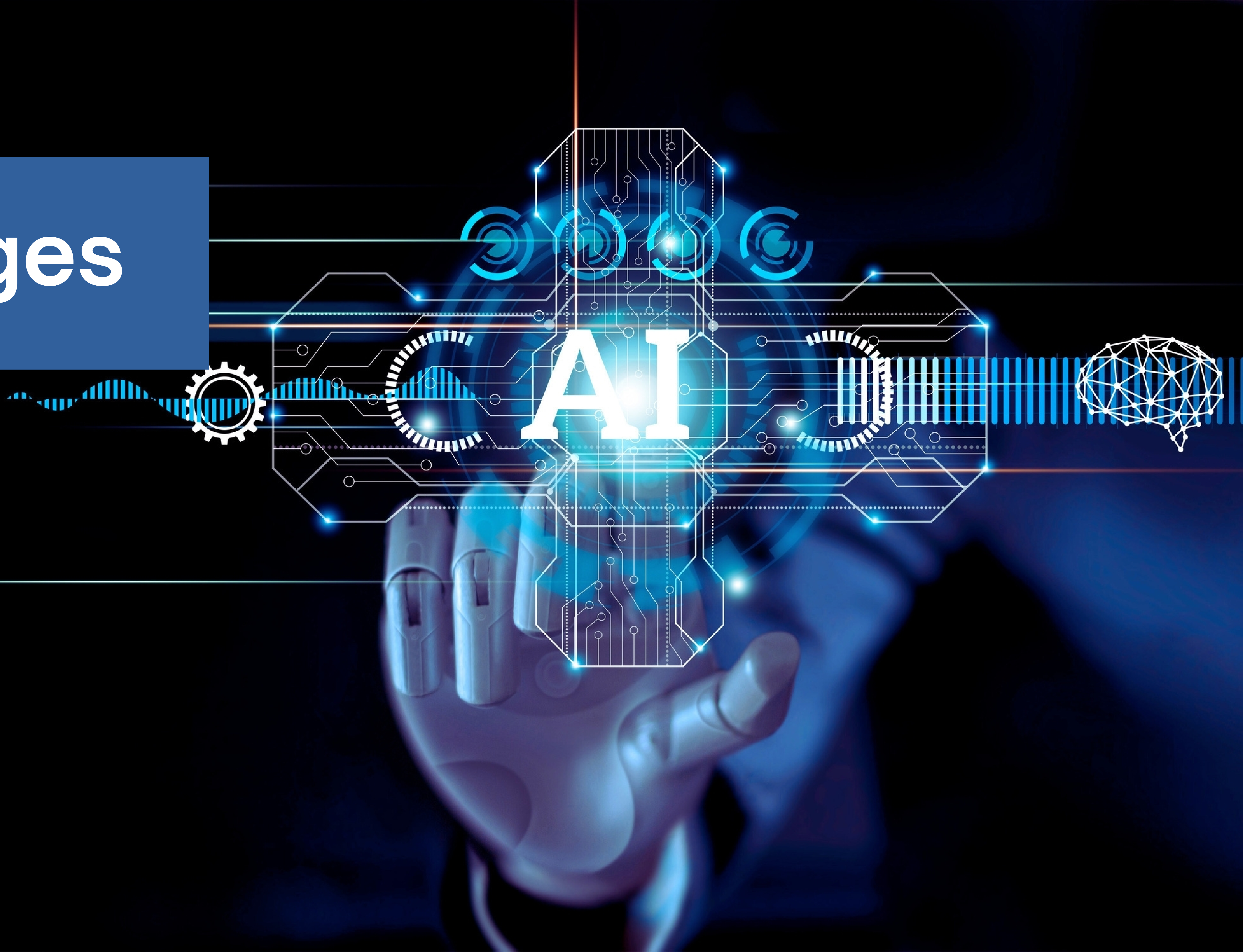
3 - EARLY STAGE OF GUN - LARGE DATASET DIRECT APPROACH

	Gunv2	Gunv3	Gunv4	Gunv5	Gunv6	Gunv7	Gunv8
Gun1 Video	1663/8530	2394/8530	2949/8530	2095/8530	897/8530	871/8530	5295/8530
Gun2 Video	414/750	512/750	205/750	190/750	19/750	144/750	209/750
Gun3 Video	19429/27193	11290/27193	21191/27193	6272/27193	9495/27193	12523/27193	13290/27193
Gun4 Video	227/1439	1188/1439	1185/1439	101/1439	1209/1439	1528/1439	538/1439
Gun5 Video	1444/4000	2002/4000	1717/4000	1166/4000	1371/4000	2113/4000	2027/4000
Gun6 Video	1879/4742	1569/4742	1260/4742	1317/4742	2549/4742	908/4742	1449/4742
Toplam Frame	**25056/39787**	**18955/39787**	**28507/39787**	**11141/39787**	**15540/39787**	**18087/39787**	**22808/39787**
Oran	**62.97**	**47.64**	**71.64**	**28**	**39.05**	**45.45**	**57.32**

	Gunv2	Gunv3	Gunv4	Gunv5	Gunv6	Gunv7	Gunv8
Kolay Dataset	2313/7806	2474/7806					

Gun Test Dataseti:
Ekstrem case true positive test frames= 40.048
Ekstrem case false positive test frames= 39.787
Airport false positive test frames= 334.263

Challenges

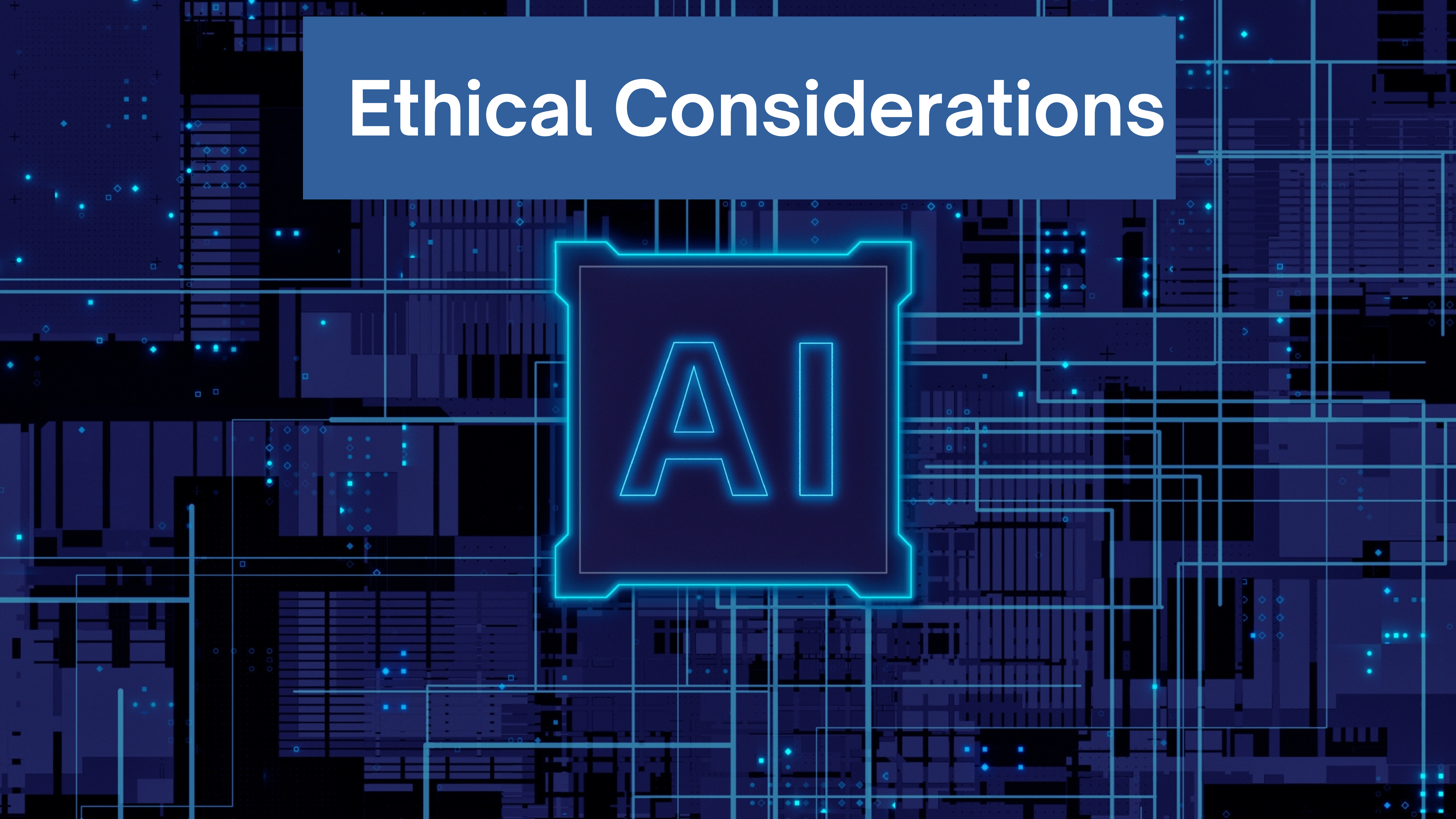


```
Thu Nov 23 16:45:36 2023
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 535.104.12             Driver Version: 535.104.12   CUDA Version: 12.2     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA A100-SXM4-40GB          Off | 00000000:07:00.0 Off |                    0 |
| N/A   29C    P0              56W / 400W |      4MiB / 40960MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA A100-SXM4-40GB          Off | 00000000:0B:00.0 Off |                    0 |
| N/A   27C    P0              51W / 400W |      4MiB / 40960MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   2  NVIDIA A100-SXM4-40GB          Off | 00000000:48:00.0 Off |                    0 |
| N/A   25C    P0              55W / 400W |      4MiB / 40960MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   3  NVIDIA A100-SXM4-40GB          Off | 00000000:4C:00.0 Off |                    0 |
| N/A   27C    P0              54W / 400W |      4MiB / 40960MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   4  NVIDIA A100-SXM4-40GB          Off | 00000000:88:00.0 Off |                    0 |
| N/A   25C    P0              51W / 400W |      4MiB / 40960MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   5  NVIDIA A100-SXM4-40GB          Off | 00000000:8B:00.0 Off |                    0 |
| N/A   29C    P0              55W / 400W |      4MiB / 40960MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   6  NVIDIA A100-SXM4-40GB          Off | 00000000:C8:00.0 Off |                    0 |
| N/A   26C    P0              55W / 400W |      4MiB / 40960MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   7  NVIDIA A100-SXM4-40GB          Off | 00000000:CB:00.0 Off |                    0 |
| N/A   26C    P0              50W / 400W |      4MiB / 40960MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|  No running processes found                                                           |
+---------------------------------------------------------------------------------------+
[it4i-xena-x14@acn54.karolina sample]$ ls
```

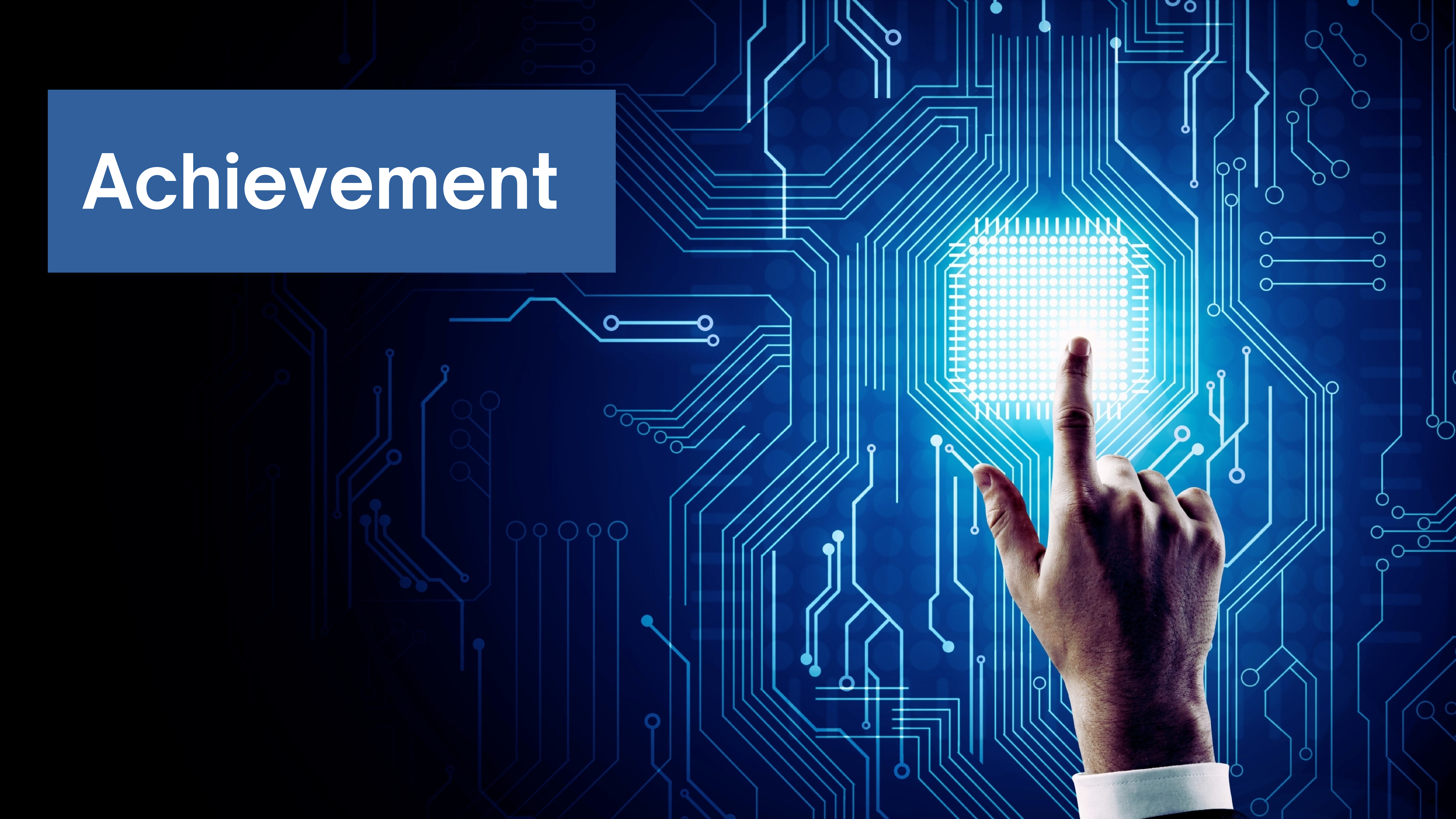
```
 _  __              _ _
| |/ /             | (_)
| ' / __ _ _ __ ___ | |_ _ __   __ _
|  < / _` | '__/ _ \| | | '_ \ / _` |
| . \ (_| | | | (_) | | | | | | (_| |
|_|\_\__,_|_|  \___/|_|_|_| |_|\__,_|

        ...running on Red Hat Enterprise Linux 7.x

: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8): No such file or directory
-xena-x14@login1.karolina ~]$ sudo -s

ust you have received the usual lecture from the local System
istrator. It usually boils down to these three things:

1) Respect the privacy of others.
2) Think before you type.
3) With great power comes great responsibility.

] password for it4i-xena-x14:

 Stopped                  sudo -s
```

```
Collecting torch==1.9.0+cu102




  Using cached https://download.pytorch.org/whl/cu102/torch-1.9.0%2Bcu102-cp36-cp36m-linux_x86_64.wh
Collecting torchvision==0.10.0+cu102
  Using cached https://download.pytorch.org/whl/cu102/torchvision-0.10.0%2Bcu102-cp36-cp36m-linux_x8
Requirement already satisfied: dataclasses; python_version < "3.7" in /usr/local/lib/python3.6/site-
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.6/site-packages (from tor
Collecting numpy (from torchvision==0.10.0+cu102)
  Downloading https://files.pythonhosted.org/packages/45/b2/6c7545bb7a38754d63048c7696804a0d94732812
    100% |################################| 13.4MB 113kB/s
Collecting pillow>=5.3.0 (from torchvision==0.10.0+cu102)
  Downloading https://files.pythonhosted.org/packages/7d/2a/2fc11b54e2742db06297f7fa7f420a0e3069fdcf
    100% |################################| 49.4MB 31kB/s
Installing collected packages: torch, numpy, pillow, torchvision
Exception:
Traceback (most recent call last):
  File "/usr/lib/python3.6/site-packages/pip/basecommand.py", line 215, in main
    status = self.run(options, args)
  File "/usr/lib/python3.6/site-packages/pip/commands/install.py", line 365, in run
    strip_file_prefix=options.strip_file_prefix,
  File "/usr/lib/python3.6/site-packages/pip/req/req_set.py", line 789, in install
    **kwargs
  File "/usr/lib/python3.6/site-packages/pip/req/req_install.py", line 854, in install
    strip_file_prefix=strip_file_prefix
  File "/usr/lib/python3.6/site-packages/pip/req/req_install.py", line 1069, in move_wheel_files
    strip_file_prefix=strip_file_prefix,
  File "/usr/lib/python3.6/site-packages/pip/wheel.py", line 345, in move_wheel_files
    clobber(source, lib_dir, True)
  File "/usr/lib/python3.6/site-packages/pip/wheel.py", line 316, in clobber
    ensure_dir(destdir)
  File "/usr/lib/python3.6/site-packages/pip/utils/__init__.py", line 83, in ensure_dir
    os.makedirs(path)
  File "/usr/lib64/python3.6/os.py", line 220, in makedirs
```

Ethical Considerations



Achievement



Conclusion


Thank You!

Contact Us!

XENA VISION

Contact

Scan the QR code

ODTU Technopart Innovation and Technology Center, Mustafa Kemal District,
Cankaya/Ankara
TURKIYE
nazli.temur@xena-vision.com

37 Richard Way SW #200, Calgary, AB
T3E 7M8, Canada
nazli@xenavision.com

Singapore
Coming Soon

EUBERT

A Language Model trained on EU Institutions documents

EUROHPC USER DAY 2023 Brussels 11.12.23

Project: EP LLM Fine Tuning

EuroHPC used: Meluxina

Speaker: Sébastien Campion *(European*

Plan

1. Context and needs
2. Technical description
3. Results, issues & limitations

Context & Needs

EU Institutions publish many documents of different type, report, brief, legal text, etc.

Each document could be described by keywords.

Keywords are chosen in a defined vocabulary called **EuroVoc** (~ 7000 terms).

EU publications

★★★★★ Rate this publication

Analysis of the requirements for the Open Source infrastructure of the Open Research Europe Publishing Platform

Open Research Europe (ORE) is the peer-reviewed open-access publishing platform of the European Commission. It follows the post-publication peer review model to promote scientific transparency and reuse. The Commission plans to develop an infrastructure to underpin ORE in the future that is based on open source software following the open-source code use and distribution model. The present analysis was commissioned to determine if open-source software (OSS) solutions can be used as a foundation for developing the new publishing platform and to document the necessary workflows and functionalities of the new platform. After conducting a thorough analysis, it has become evident that utilizing existing open-source software has its own advantages and disadvantages. Although some risks are associated with this approach, our research has identified a few mature existing solutions that could be further developed to support the future ORE platform.

View less

How to cite

Download and languages

Publication details

Published: 2023

Corporate author(s): Directorate-General for Research and Innovation (European Commission)

Personal author(s): Kouis, Dimitrios

Themes: Information technology and telecommunications

Subject: access to information , dissemination of information , information technology , innovation , open access publishing , open science , open source software , report , scientific research

PDF

ISSN	ISBN	DOI	Catalogue number
	978-92-68-05646-2	10.2777/13928	KI-03-23-367-EN-N

Released on EU Publications: 2023-11-13

Eurovoc Multilabel Classifier

Extreme Multilabel Classification

SOTA based on Deep Learning Network

pre trained language model

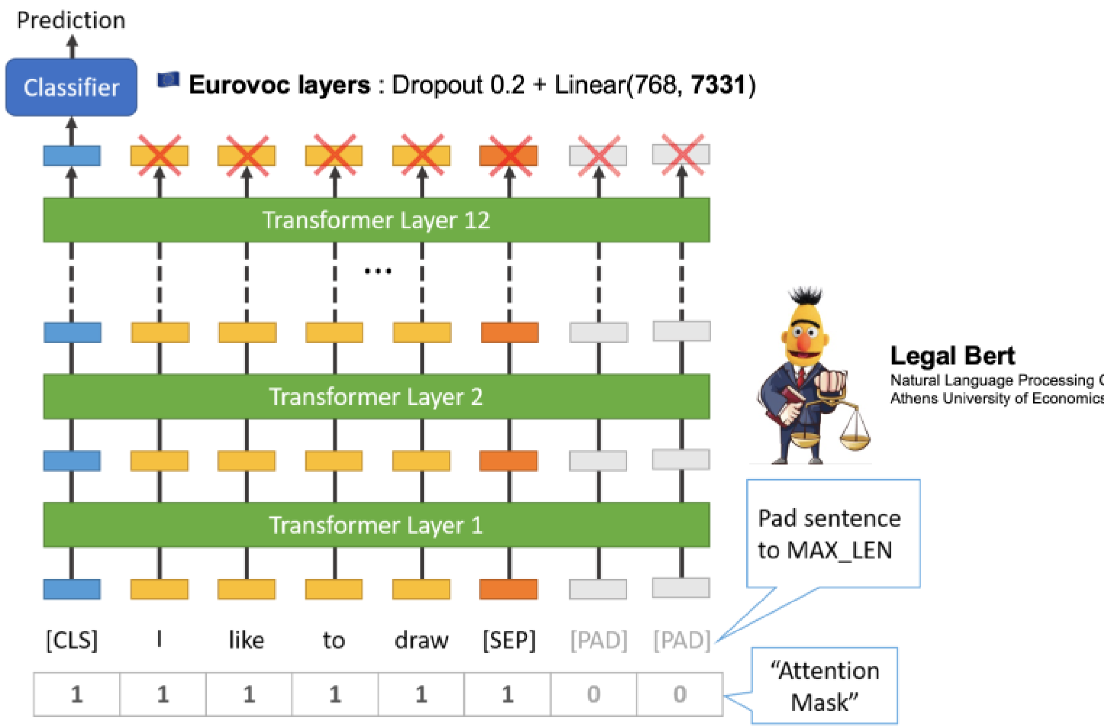
Legal Bert ⚖

+

a new classification layer

Inference on CPU 👍

new dataset has been set up



EuroVoc Dataset

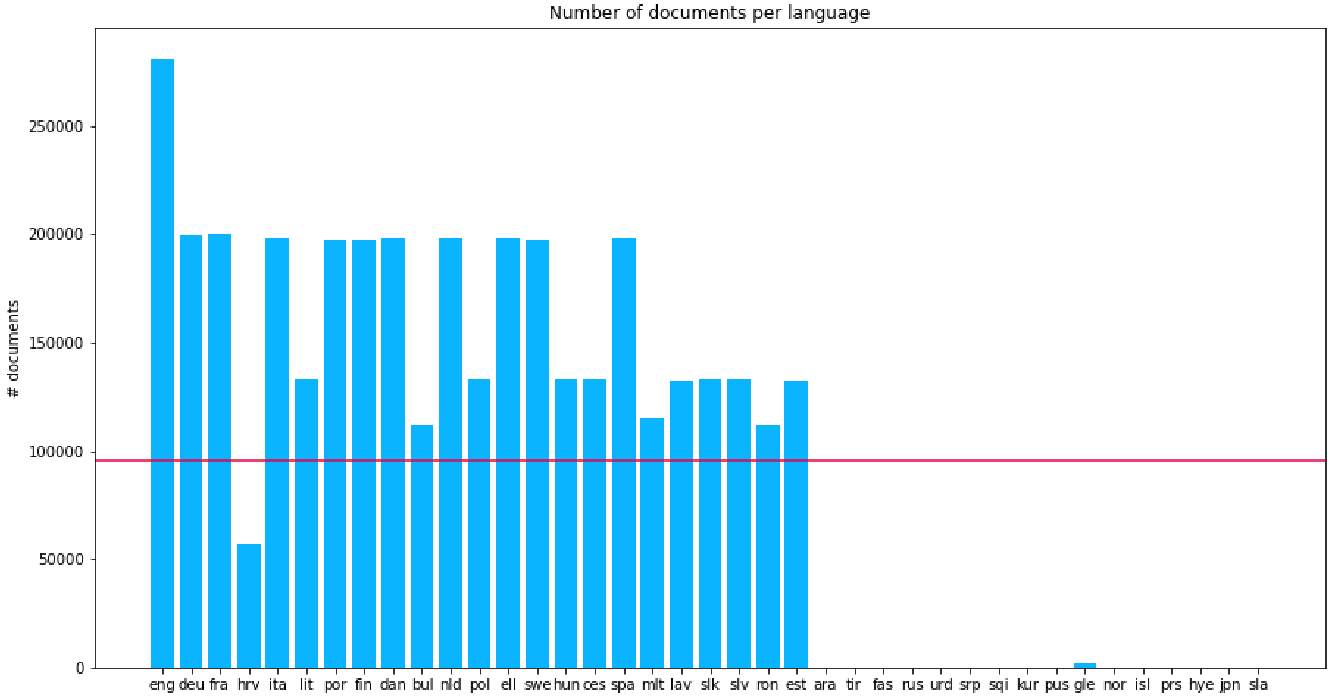
https://huggingface.co/datasets/EuropeanParliament/Eurovoc

~ 30 years

~ 3.7 Millions of documents

24 Languages

100GB



title string · *lengths* 58 498	date unknown	eurovoc_concepts sequence	url string · *lengths* 82 82	lang string · *classes* 23 values	formats sequence	text string · *lengths* 1.45k 211k
Propuesta de DECISIÓN DEL CONSEJO Y LA COMISIÓN…	"1996-03-29T00:00:00"	["EU relations", "Moldova", "accession…	http://publications.europa.eu/resource/cellar/b8f7a4b7-14f9-44a8-997d-10e5c3a33f58	spa	["pdf"]	COMISIÓN DE LAS COMUNIDADES EUROPEAS it ir" ir ir ir "ir ir <' • :, f' Bruselas, 29 03 1996 COM(%)…
Forslag til RÅDETS OG KOMMISSIONENS AFGØRELSE o…	"1996-03-29T00:00:00"	["EU relations", "Moldova", "accession…	http://publications.europa.eu/resource/cellar/b8f7a4b7-14f9-44a8-997d-10e5c3a33f58	dan	["pdf"]	<V s. KOMMISSIONEN FOR DE EUROPÆISKE FÆLLESSKABER i' /, ix Bruxelles, den 29. 03. 19% KOM(%) 132…
Πρόταση ΑΠΟΦΑΣΗΣ ΤΟΥ ΣΥΜΒΟΥΛΙΟΥ ΚΑΙ ΤΗΣ…	"1996-03-29T00:00:00"	["Ukraine", "accession to the…	http://publications.europa.eu/resource/cellar/b368c109-3812-4ab5-baa4-3c455b4ea4e9	ell	["pdf"]	jy it ΕΠΙΤΡΟΠΗ ΤΩΝ ΕΥΡΩΠΑΪΚΩΝ ΚΟΙΝΟΤΗΤΩΝ | { (ν ι ύ ; (λ λ ι : ;, ?. <><) V I <><)ί> ΙΌΜ(9ί>)Ι U…
Решение на Съвета от 29 март 1996 година за…	"1996-03-29T00:00:00"	["UNO", "forest conservation",…	http://publications.europa.eu/resource/cellar/13d2b871-3b38-42fc-a18e-6db5efae9e6a	bul	["html", "pdf",…	204 BG Официален вестник на Европейския съюз 11/т. 13 31996D0493 17. 8. 1996 ОФИЦИАЛЕН ВЕСТНИК НА…
Voorstel voor een BESLUIT VAN DE RAAD EN VAN DE…	"1996-03-29T00:00:00"	["EU relations", "Moldova", "accession…	http://publications.europa.eu/resource/cellar/b8f7a4b7-14f9-44a8-997d-10e5c3a33f58	nld	["pdf"]	COMMISSIE VAN DE EUROPESE GEMEENSCHAPPEN •ir •fr •it ^ 4* * it -îz Brussel, 29. 03. 19% COM(%) 132…
Proposta de DECISÃO DO CONSELHO E DA COMISSÃO…	"1996-03-29T00:00:00"	["EU relations", "Moldova", "accession…	http://publications.europa.eu/resource/cellar/b8f7a4b7-14f9-44a8-997d-10e5c3a33f58	por	["pdf"]	COMISSÃO DAS COMUNIDADES EUROPEIAS ft ft ^ ft it ft ft ft ft ft£ft Bruxelas, 29. 03. 1996 COM(%) 132…
1996 m. kovo 29 d. Tarybos direktyva 96/21/EB iš…	"1996-03-29T00:00:00"	["approximation of laws", "consumer…	http://publications.europa.eu/resource/cellar/bea89432-5000-4e57-9974-a088f586305b	lit	["html", "pdf",…	15/3 t. LT Europos Sąjungos oficialusis leidinys 53 31996L0021 L 88/5 EUROPOS BENDRIJŲ OFICIALUSIS…
Proposition de DECISION DU CONSEIL ET DE LA…	"1996-03-29T00:00:00"	["Ukraine", "accession to the…	http://publications.europa.eu/resource/cellar/b368c109-3812-4ab5-baa4-3c455b4ea4e9	fra	["pdf"]	! COMMISSION DES COMMUNAUTES EUROPEENNES A v',- A Bruxelles, le 29. 03. 19% COM(%)m final %AK)90…
Komisijas Regula (EK) Nr. 569/96 (1996. gada 29.…	"1996-03-29T00:00:00"	["aid to agriculture",…	http://publications.europa.eu/resource/cellar/40ef37df-702c-4359-b909-c889f82b5a0b	lav	["html", "pdf",…	03/19. sēj. LV Eiropas Savienības Oficiālais Vēstnesis 3 31996R0569 L 80/48 EIROPAS KOPIENU…
Proposition de décision du Conseil et de la…	"1996-03-29T00:00:00"	["EU relations", "Moldova", "accession…	http://publications.europa.eu/resource/cellar/b8f7a4b7-14f9-44a8-997d-10e5c3a33f58	fra	["pdf"]	COMMISSION DES COMMUNAUTES EUROPEENNES it it it * Bruxelles, le 29. 03. 1996 COM(%) 132 final 96/008…

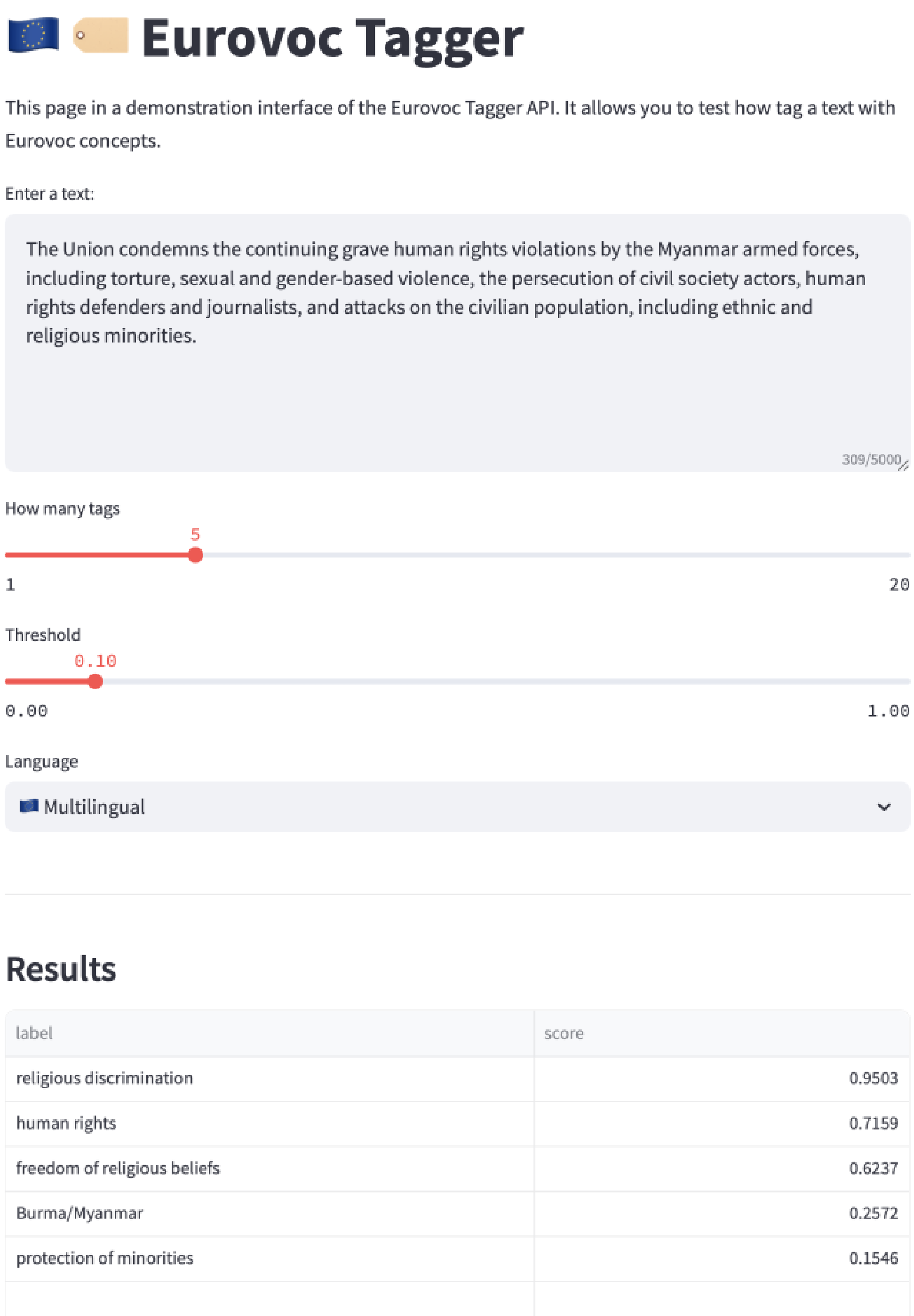
Eurovoc Multilabel Classifier

🤗 https://huggingface.co/EuropeanParliament

A first version, but new needs are emerging:

What about the European Parliament's archives in French?

What about the national documents sent each day in their own language?



💡 Replace the pretrained model

Legal Bert GB + EuroVoc ➡ EUBERT EU + EuroVoc

EUBERT - Technical description

- Masked Language Model or pretrained model
- A dedicated tokenizer for 24 languages
- < 100 Millions of parameters
- Architecture based on RoBERTa
- Licence EUPL
- 16 GPU/days to train it
- 3 epochs

https://huggingface.co/EuropeanParliament/EUBERT

Mask token: <mask>

The transition to a climate neutral, sustainable, energy and resource-efficient, circular and fair economy is key to ensuring the long-term competitiveness of the economy of the union and the well-being of its peoples. In 2016, the Union concluded the Paris Agreement2. Article 2(1), point (c), of the Paris Agreement sets out the objective of strengthening the response to climate change by, among other means, making finance flows consistent with a pathway towards low greenhouse gas [MASK] and climate resilient development.

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

emissions	0.765
emission	0.184
mitigation	0.009
reduction	0.004
production	0.003

</> JSON Output

Maximize

Implementations

- EuroVoc MultiLabels & Multilingual Classifier
- EUBERT Embeddings v1 for semantic search engine

EuroVoc Evaluation

Metric	EuroVoc EU based on EUBERT (strat 1/9)	Large-Scale Multi-Label Text Classification on EU Legislation GB
Micro F1	0.8345	0.732
NDCG@3	0.8819	
NDCG@5	0.8689	0.823
NDCG@10	0.8780	

EuroVoc Evaluation

Eurovoc Classifier 645 docs from september (never seen before)

Work still in progress.

Metrics	poc	Legal BERT	EUBERT	PyEuroVoc
NDCG@3	0.5239	0.7071	0.8059	0.5013
NDCG@5	0.4583	0.6353	0.7445	0.4325
NDCG@10	0.4253	0.5863	0.6939	0.3891

Limitations & Issues

What about data quality ? a lot of text are extracted from PDF

Why 3 epochs ? Overfitting ? considering the scaling law

Is 100 millions the best size ?

Conclusion

Train a model with 100 millions of parameters is too expensive without dedicated accelerator such as GPU.

Access to GPUs for public administrations is difficult (calls for tender, hardware configuration maintenance, restrictions linked to cloud computing, etc.)

EuroHPC Development Access solves this problem of access to computing resources and produces tangible results.

Is the future going to be light models adapted to the business domain or to LLMs with multiple capacity ?

perhaps both

Bibliography

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2019. Extreme Multi-Label Legal Text Classification: A Case Study in EU Legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota. Association for Computational Linguistics.

I. Chalkidis, M. Fergadiotis, P. Malakasiotis and I. Androutsopoulos, "Large-Scale Multi-Label Text Classification on EU Legislation". Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, (short papers), 2019 ()

Andrei-Marius Avram, Vasile Pais, and Dan Ioan Tufis. 2021. PyEuroVoc: A Tool for Multilingual Legal Document Classification with EuroVoc Descriptors. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), pages 92–101, Held Online. INCOMA Ltd..

SHAHEEN, Zein, WOHLGENANNT, Gerhard, et FILTZ, Erwin. Large scale legal text classification using transformer models. arXiv preprint arXiv:2010.12871, 2020.

Training acoustic models at the National Library of Sweden

EUROHPC USER DAY 2023 Brussels 11.12.23

EuroHPC Joint Undertaking

Project: “*Speech recognition for Swedish using Wav2vec2.0 and Whisper*”

EuroHPC used: Leonardo

Speaker: Leonora VESTERBACKA

The National Library of Sweden

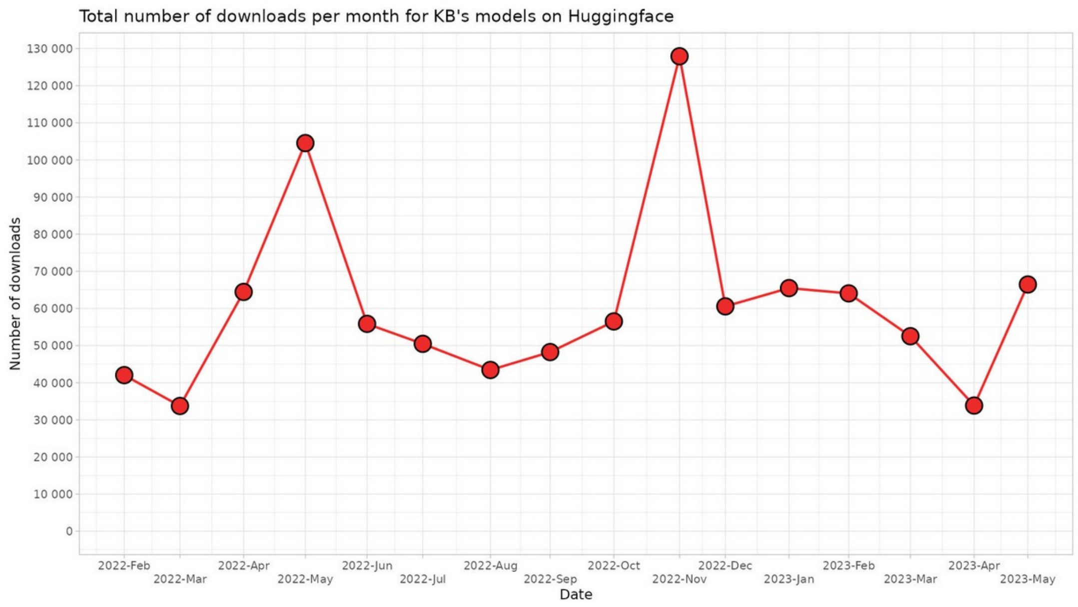
- **collects, preserves and gives access** to nearly everything published in Sweden
- legal deposit act from 1661 required all printers to deliver one copy to the library
- a censorship law that now helps preserve Sweden's cultural heritage
  - includes sound and moving images from 1979
    - **TV/radio/podcasts**
- collections currently hold 18 million items

National Library of Sweden

https://huggingface.co/KBLab

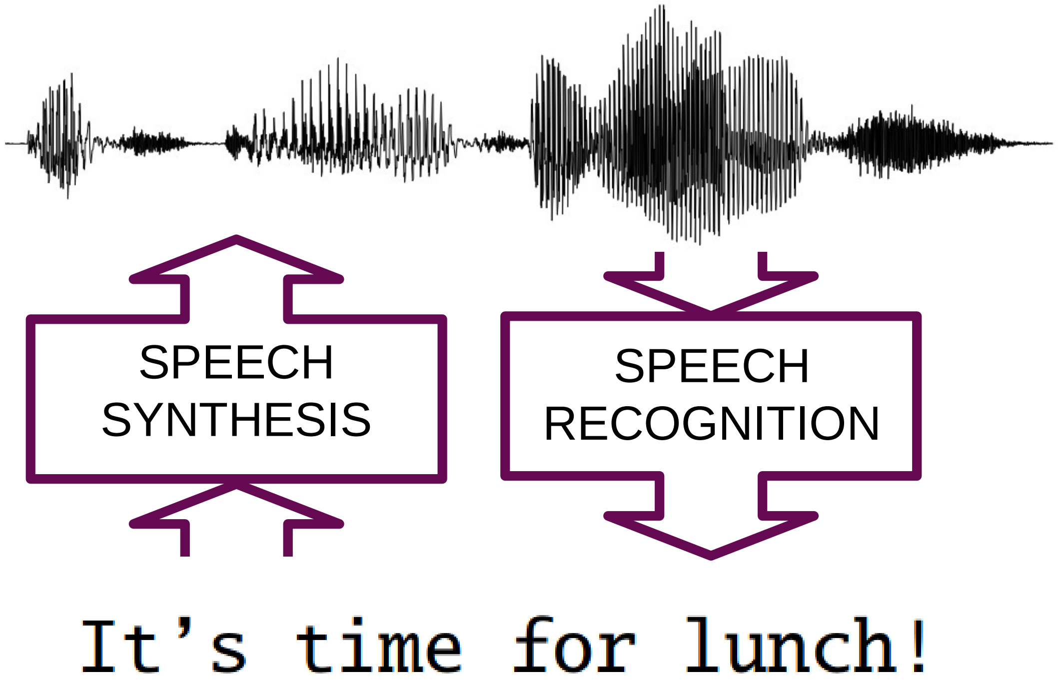
KBlab

- a data and AI lab at the National Library of Sweden (KB)
- started in 2019
- enable large scale quantitative research
- trained laguage models on the librarys unique datasets
  - frequently used by private and public sector
- models published on Huggingface
  - BERT, BART, wav2vec, NER model, sentence-BERT and many more...



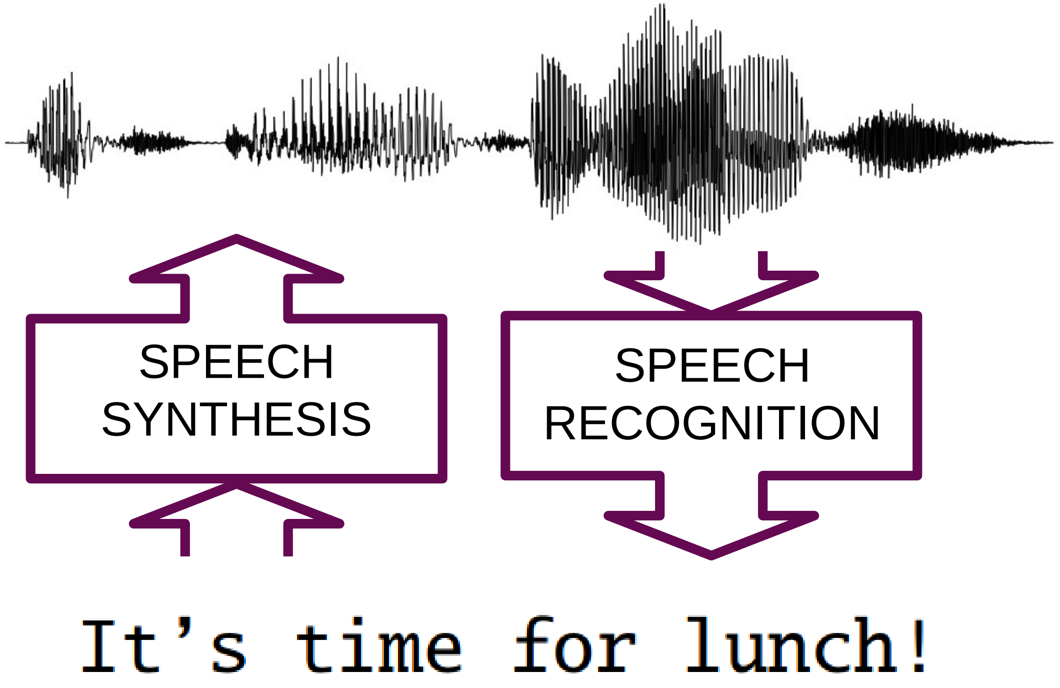
Speech synthesis & speech recognition

- map waveform to string of words
- speech synthesis / text-to-speech (TTS)
- automatic speech recognition (ASR)
- accessibility adaption, transcriptions, automatic captions



Speech synthesis & speech recognition

- map waveform to string of words
- speech synthesis / text-to-speech (TTS)
- automatic speech recognition (ASR)
- accessibility adaption, transcriptions, automatic captions



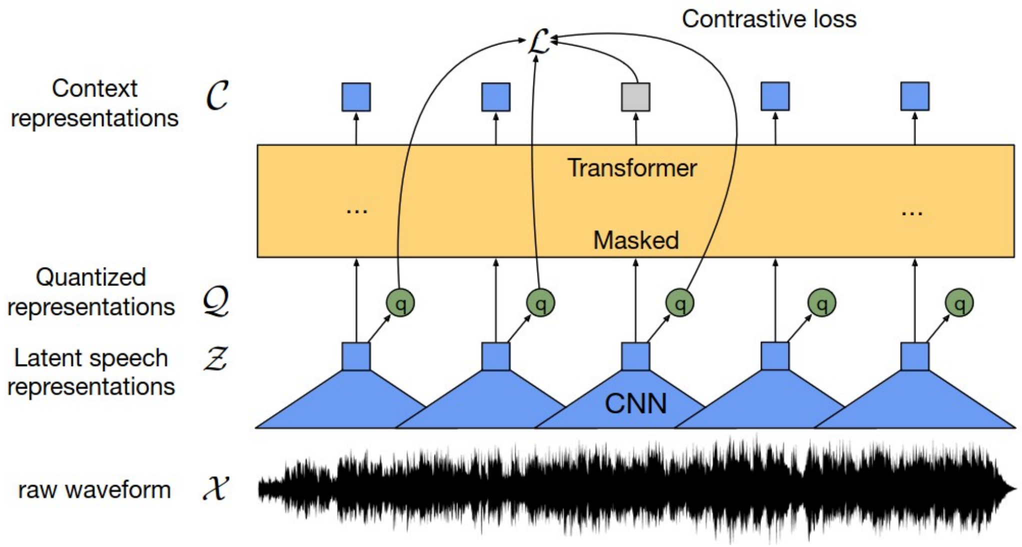
- the development in the field of ASR is driven by a few tech companies
  - with limited access to good training data for smaller languages such as Swedish
    - with even less representation w.r.t. dialects
- however there is a huge demand for high quality swedish ASR models
- **This is why the National Library of Sweden is training acoustic models**

KUNGL. BIBLIOTEKET
National Library of Sweden

https://arxiv.org/pdf/2006.11477.pdf

Wav2vec2.0

- Training code and models released by Meta in 2020
- Transformer – ideal for HPC
- Similar to BERT the model is trained by **predicting speech units for masked parts of the audio**

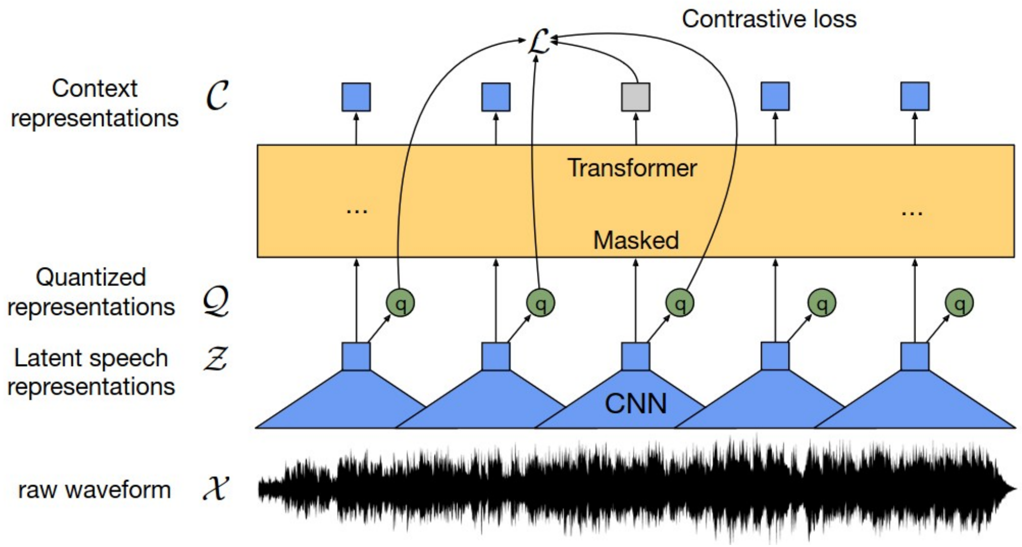


National Library of Sweden

https://arxiv.org/pdf/2006.11477.pdf

Wav2vec2.0

- Training code and models released by Meta in 2020
- Transformer – ideal for HPC
- Similar to BERT the model is trained by **predicting speech units for masked parts of the audio**

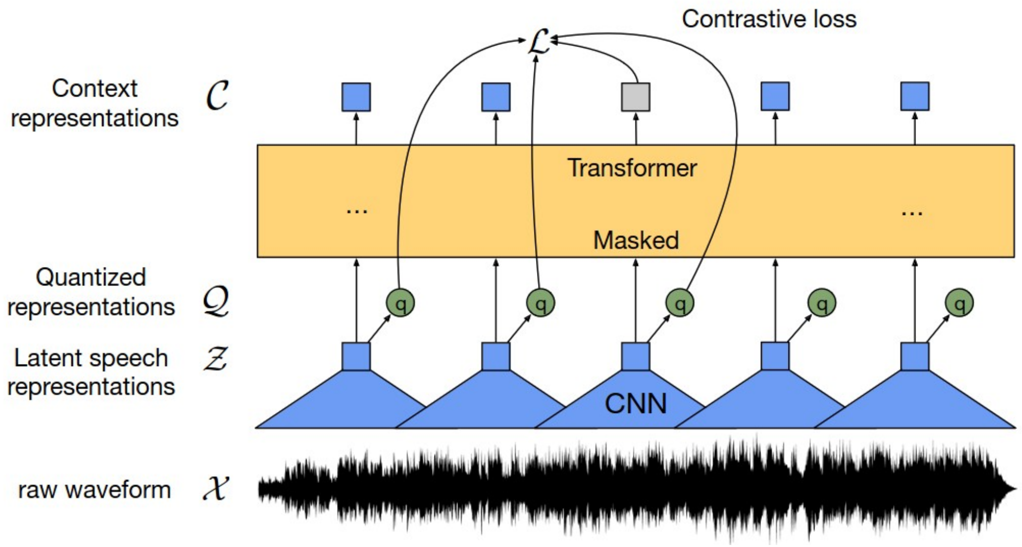


Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
10 min labeled						
Discrete BERT [4]	LS-960	4-gram	15.7	24.1	16.3	25.2
BASE	LS-960	4-gram	8.9	15.7	9.1	15.6
		Transf.	6.6	13.2	6.9	12.9
LARGE	LS-960	Transf.	6.6	10.6	6.8	10.8
	LV-60k	Transf.	4.6	7.9	4.8	8.2

https://arxiv.org/pdf/2006.11477.pdf

Wav2vec2.0

WER: word-error-rate, an evaluation metric



- Training code and models released by Meta in 2020
- Transformer – ideal for HPC
- Similar to BERT the model is trained by **predicting speech units for masked parts of the audio**

10 min LibriSpeech fine-tuning

Low WER despite 10 min labeled data

Model	Unlabeled data	LM	dev		test	
			clean	other	clean	other
10 min labeled						
Discrete BERT [4]	LS-960	4-gram	15.7	24.1	16.3	25.2
BASE	LS-960	4-gram	8.9	15.7	9.1	15.6
		Transf.	6.6	13.2	6.9	12.9
LARGE	LS-960	Transf.	6.6	10.6	6.8	10.8
	LV-60k	Transf.	4.6	7.9	4.8	8.2

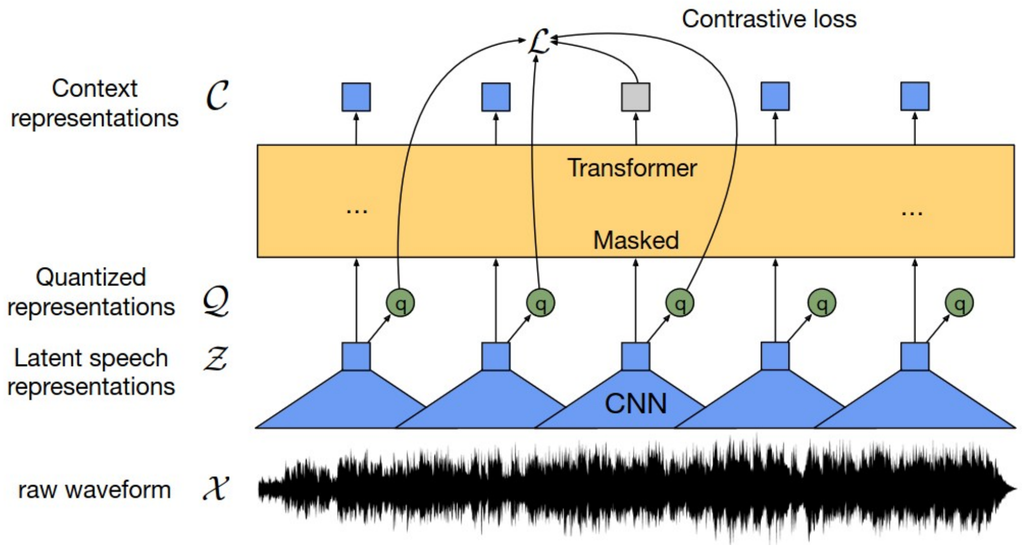
National Library of Sweden

https://arxiv.org/pdf/2006.11477.pdf

WER: word-error-rate, an evaluation metric

Wav2vec2.0

- Training code and models released by Meta in 2020
- Transformer – ideal for HPC
- Similar to BERT the model is trained by **predicting speech units for masked parts of the audio**



10 min LibriSpeech fine-tuning

Drawback:
Need a separate LM

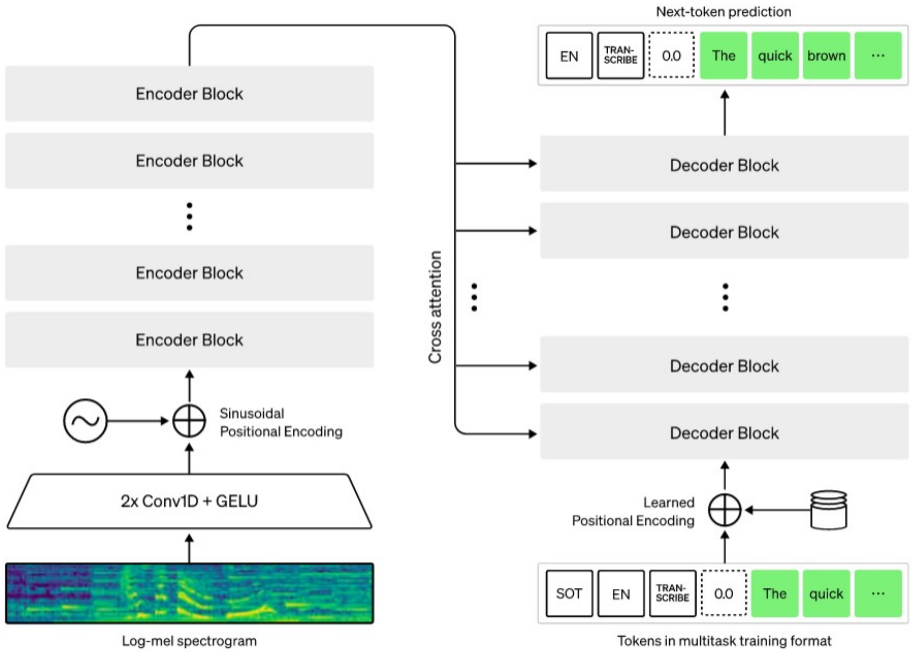
Low WER despite 10 min labeled data

Model	Unlabeled data	LM	dev clean	dev other	test clean	test other
10 min labeled						
Discrete BERT [4]	LS-960	4-gram	15.7	24.1	16.3	25.2
BASE	LS-960	4-gram	8.9	15.7	9.1	15.6
		Transf.	6.6	13.2	6.9	12.9
LARGE	LS-960	Transf.	6.6	10.6	6.8	10.8
	LV-60k	Transf.	4.6	7.9	4.8	8.2

https://arxiv.org/pdf/2212.04356.pdf

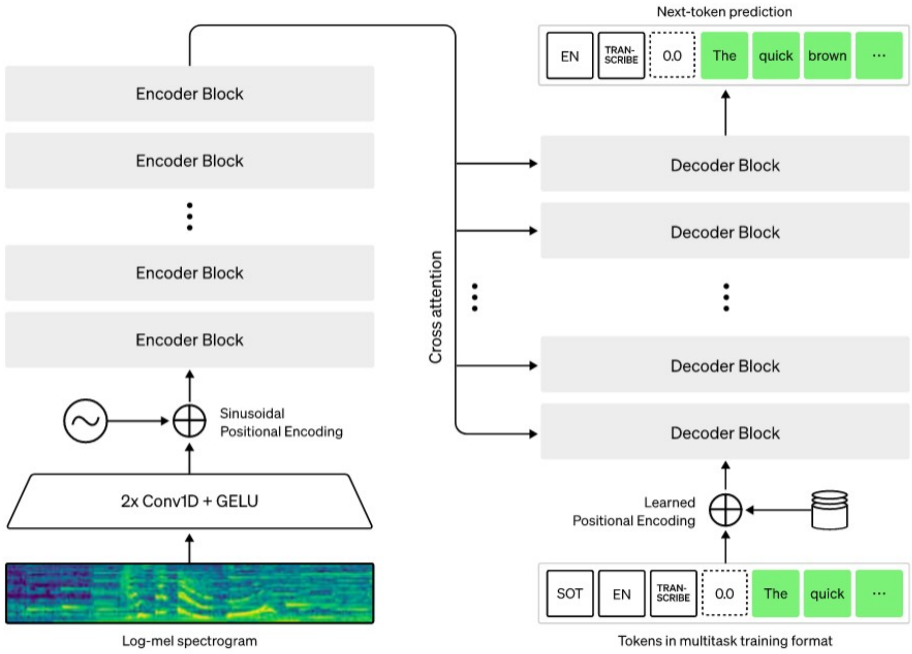
Whisper

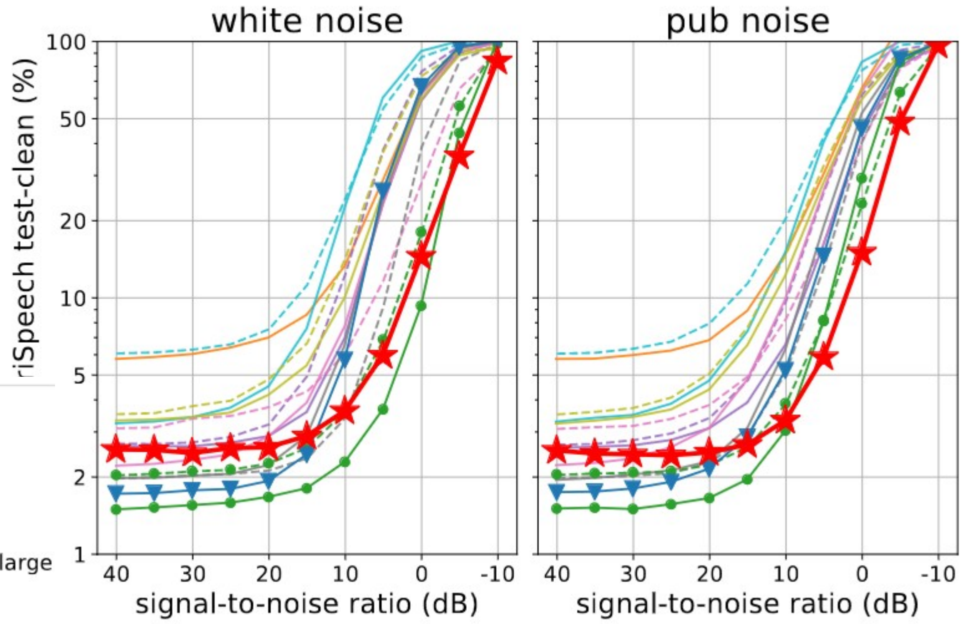
- Fine-tuning code and models released by OpenAI in 2021
  - End-to-end approach
    - **Single model for the whole speech recognition pipeline**
  - Encoder-decoder Transformer
  - **supervised** training required
    - however only **weakly** supervised
    - **relaxed requirements** on gold-standard transcripts

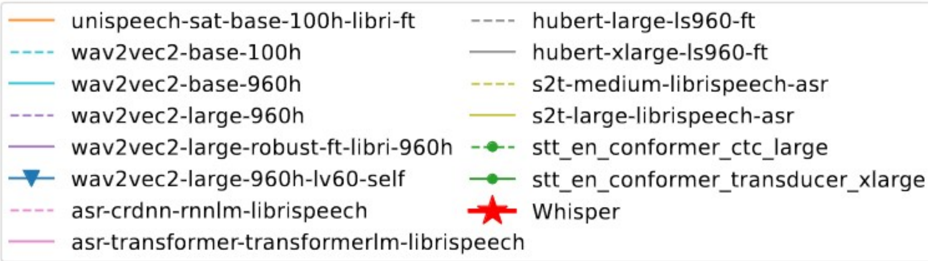


Whisper

- Fine-tuning code and models released by OpenAI in 2021
  - End-to-end approach
    - **Single model for the whole speech recognition pipeline**
  - Encoder-decoder Transformer
  - **supervised** training required
    - however only **weakly** supervised
    - **relaxed requirements** on gold-standard transcripts
  - Trained on **680 000 h** transcribed audio from the web
    - Does not beat other models when evaluating on LibriSpeech
    - However more robust when evaluating on other datasets







https://arxiv.org/pdf/2205.03026.pdf

VoxRex: a Swedish wav2vec2.0

- In 2021 the lab trained a **Wav2vec2.0 on Swedish** data
  - pretraining on **10 000 h local radio** (unlabeled)
    - local radio provides a large variation of Swedish dialects
  - fine-tuning with gold standard labeled datasets
- At the time of release it it **outperformed Metas** equivalent model
- **Used widely** by the public and private sector:
  - Transcription of meetings, audio archives, hearings, etc.

Hearing voices at the National Library -
a speech corpus and acoustic model for the Swedish language

Martin Malmsten, Chris Haffenden, Love Börjeson
KBLab, National Library of Sweden
Humlegården, Stockholm
www.kb.se/kb-labb
{martin.malmsten, chris.haffenden, love.borjeson}@kb.se

Abstract

This paper details our work in developing new acoustic models for automated speech recognition (ASR) at KBLab, the infrastructure for data-driven research at the National Library of Sweden (KB). We evaluate different approaches for a viable speech-to-text pipeline for audiovisual resources in Swedish, using the wav2vec 2.0 architecture in combination with speech corpuses created from KB's collections. These approaches include pretraining an acoustic model for Swedish from the ground up, and fine-tuning existing monolingual and multilingual models. The collections-based corpuses we use have been sampled from millions of hours of speech, with a conscious attempt to balance regional dialects to produce a more representative, and thus more democratic, model. The acoustic model this enabled, "VoxRex", outperforms existing models for Swedish ASR. We also evaluate combining this model with various pretrained language models, which further enhanced performance. We conclude by highlighting the potential of such technology for cultural heritage institutions with vast collections of previously unlabelled audiovisual data. Our models are released for further exploration and research here: https://huggingface.co/KBLab.

KUNGL. BIBLIOTEKET
National Library of Sweden

Swedish acoustic models @Leonardo

- This project have been awarded development access to Leonardo BOOSTER
- 3.500 node hours

KB-Whisper @ Leonardo

- Continued pre-training with transcribed Swedish (30 000 h)
- Ongoing work to collect and preprocess transcribed audio from archives
  - transcribed audio from
    - parliament debates
    - TV with subtitles from archives
    - youtube
    - dialects from The Institute for Language and Folklore
  - Test training code on Leonardo

Dataset size	English WER (↓)	Multilingual WER (↓)	X→En BLEU (↑)
3405	30.5	92.4	0.2
6811	19.6	72.7	1.7
13621	14.4	56.6	7.9
27243	12.3	45.0	13.9
54486	10.9	36.4	19.2
681070	**9.9**	**29.2**	**24.8**

KUNGL. BIBLIOTEKET
National Library of Sweden

Wav2vec2.0 @ Leonardo

- Upgrade of Wav2vec2.0 trained on Swedish New!
  - P4 radio: 10 000 h, 100 000 h, 1 000 000 h …
  - Augmented sounds:
    - Noise, various environments, phone, car, subway etc.
- Fine-tuning with transcribed material collected for Whisper training
  - NST + Commonvoice (12 h)
  - Parliament debates (5000 h)
  - Subtitles Youtube (9700 h)
  - Subtitles from the TV from our archives
- Successfully test and optimized training code for Wav2vec2.0
  - benchmark training times on the specific hardware setup

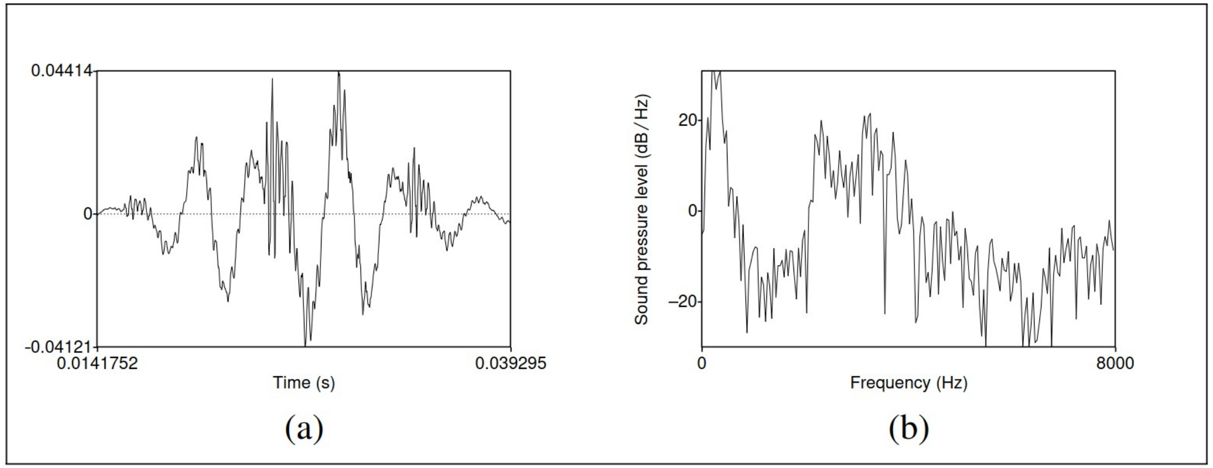
Thank you for listening!

We acknowledge EuroHPC JU for awarding this project access to Leonardo

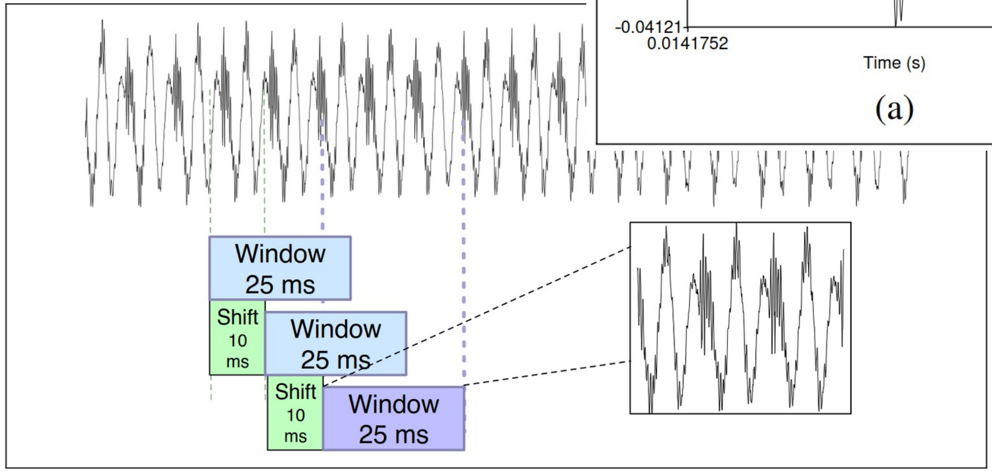
Feature extraction for ASR

- 1. analog to digital conversion by **sampling** and **quantization**
- 2. extract features from **window** of speech that characterizes a particular phoneme
- 3. extract the amplitude for each frequency using **fast Fourier Transform** (FFT)
- 4. model human perceptual property of log-like sensitivity using **mel filter banks**

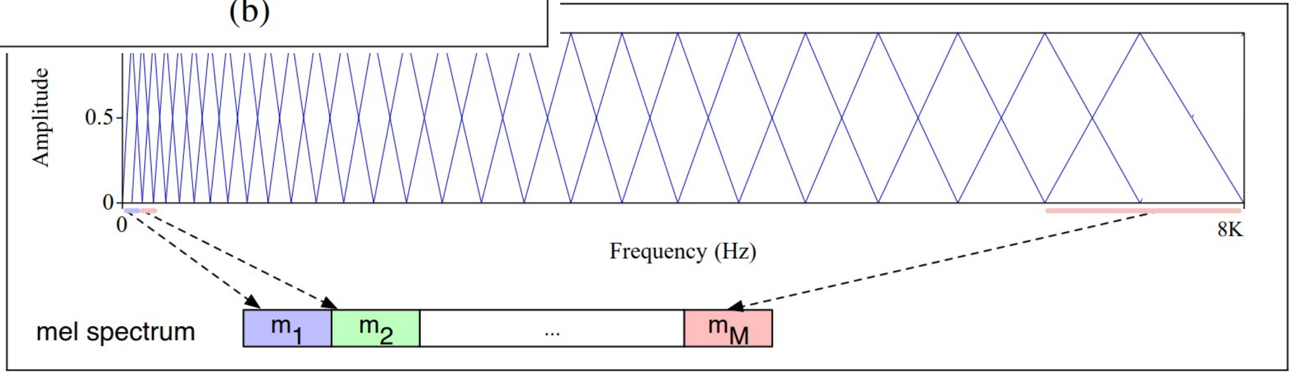
3. FFT



2. windowing



4. Mel filter banks

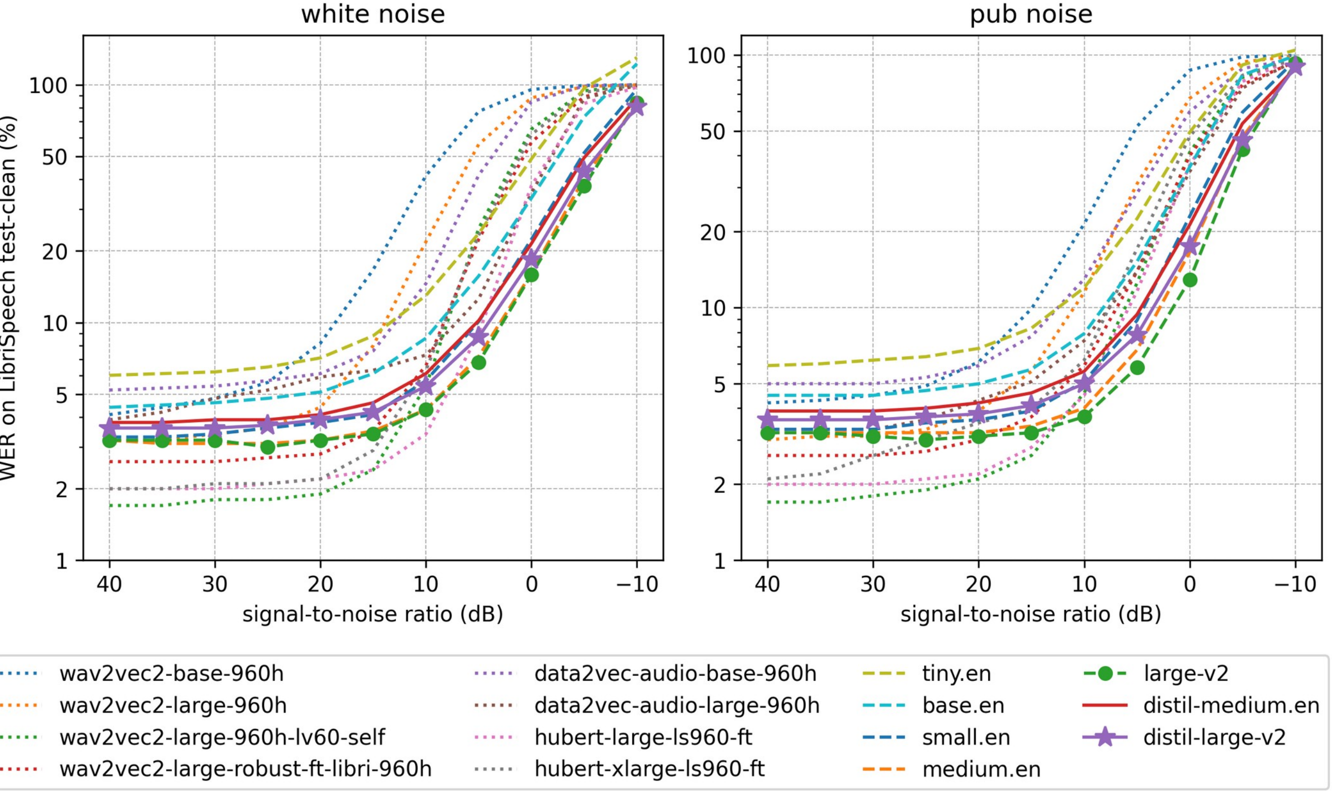


Backup

- Fast inference

Model	**Params / M**	**Short Form** Rel. Latency	**Short Form** Avg. WER	**Long Form** Rel. Latency	**Long Form** Avg. WER
tiny.en	**39**	6.1	18.9	5.4	18.9
base.en	74	4.9	14.3	4.3	15.7
small.en	244	2.6	10.8	2.2	14.7
medium.en	769	1.4	9.5	1.3	12.3
large-v2	1550	1.0	**9.1**	1.0	11.7
distil-medium.en	394	**6.8**	11.1	**8.5**	12.4
distil-large-v2	756	5.8	10.1	5.8	**11.6**

- Robustness to noise



- Robustness to hallucinations

Model	**5-Dup.**	**IER**	**SER**	**DER**	**WER**
wav2vec2-large-960h	**7971**	4.8	18.9	4.6	28.3
tiny.en	23313	5.1	8.9	4.8	18.9
base.en	22719	4.3	6.6	4.8	15.7
small.en	26377	3.3	5.0	6.5	14.7
medium.en	23549	3.5	4.2	4.6	12.3
large-v2	23792	3.3	**3.9**	4.5	11.7
distil-medium.en	18918	2.5	5.6	4.4	12.4
distil-large-v2	18503	**2.1**	5.3	**4.2**	**11.6**

Knowledge distillation with large teacher ensembles for efficient and high quality bilingual and multilingual neural machine translation

EUROHPC USER DAY 2023 Brussels 11.12.23

Project: KD with large teacher models
EuroHPC used: MeluXina (LuxProvide)
Speaker: Csaba Oravecz (DGT – EC)

Overview

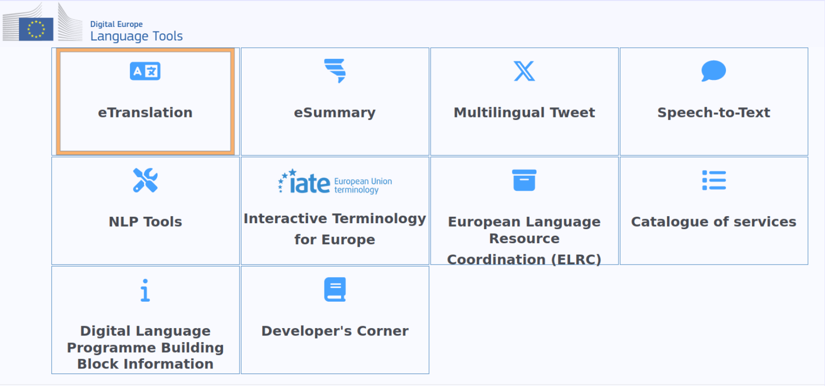
European Commission

Overview

European Commission

Background

eTranslation*



*https://language-tools.ec.europa.eu/

Background

eTranslation*

- European Commission's machine translation (MT) service
- flagship AI project under the Digital Europe programme
- provides secure access to neural machine translation between all 26 official languages of the EU and the EEA
- leverages the European Institutions' high-quality internal translation data (Euramis translation memories)
- $>$ 100 million pages translated yearly

*https://language-tools.ec.europa.eu/

European Commission

Background

Quality MT services

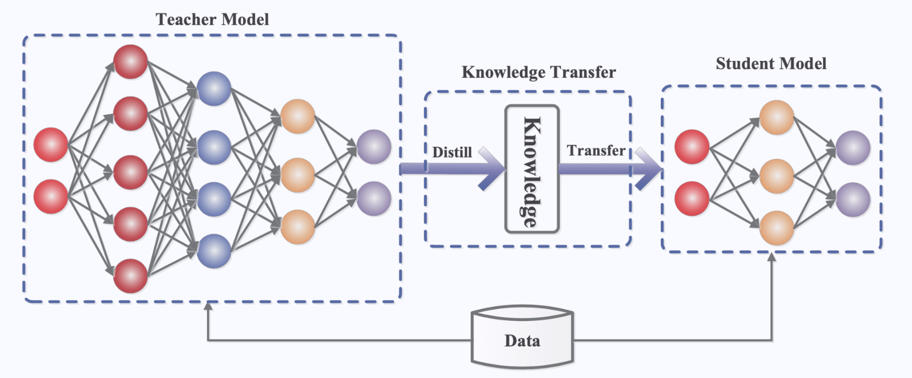
- require substantial computational power and a continuous search for the right balance between use of available resources and best possible performance of models
- tendency: more complex model architectures have better performance
  $\rightarrow$ increase the size of the models
- big models need substantial compute to train, could be inefficient in production use
- plan:
  - use HPC* resources to train deep, powerful MT models
  - resolve the resource-performance dilemma with knowledge distillation

*https://eurohpc-ju.europa.eu

European Commission

KD

Knowledge Distillation



*Source: https://arxiv.org/abs/2006.05525

Objectives

- explore the potential of deeper models to maximize the use of the information in high quality training data
- investigate the scalability of trainings in the HPC environment
- create models of improved quality that could benefit the eTranslation service in general
- set up an efficient production workflow with extended functionalites

European Commission

Overview

European Commission

Workflow

Workflow

1. train very strong teacher (ensemble) models

Workflow

1. train very strong teacher (ensemble) models
2. decode training data with teacher models

Workflow

1. train very strong teacher (ensemble) models
2. decode training data with teacher models
3. optimize student models, select best architecture

Workflow

1. train very strong teacher (ensemble) models
2. decode training data with teacher models
3. optimize student models, select best architecture
4. train student models on teacher decoded data set

European Commission

Workflow

1. train very strong teacher (ensemble) models $\Rightarrow$ HPC
2. decode training data with teacher models
3. optimize student models, select best architecture
4. train student models on teacher decoded data set

Training ecosystem

- model training: MarianNMT v11.0, v12.0
- GPU communication: NCCL 2.8.3, CUDA 12.1
- model evaluation: Sacrebleu 2.3.1, COMET v1.0, v2.0
- software environment packaged into a Singularity container
- most efficient setup: one model per node (full precision) training
- intermediate checkpoints at 10k updates
- long trainings to get insights into model convergence

Teacher models

Language pairs

- EU formal language:
  {Danish, Dutch, German, Finnish, Hungarian, Swedish} → English
  English → {German, Finnish, Hungarian}
- General language (combined):
  English → {German, Hungarian}

Data sets – Euramis, ELRC, ParaCrawl, Opus

	Da→En	De↔En	Fi↔En	Hu↔En	Nl→En	Sv→En
Euramis	22.7M	33.3M	25.2M	23.7M	26.4M	25.6M
All	–	498.8M	–	114.1M	–	–

European Commission

Teacher models

Architecture

- standard big Transformer (≈ 630M parameters):
  - 6 encoder/decoder layers
  - 16 heads
  - embedding size: 1024
  - FFN layer size: 4096

European Commission

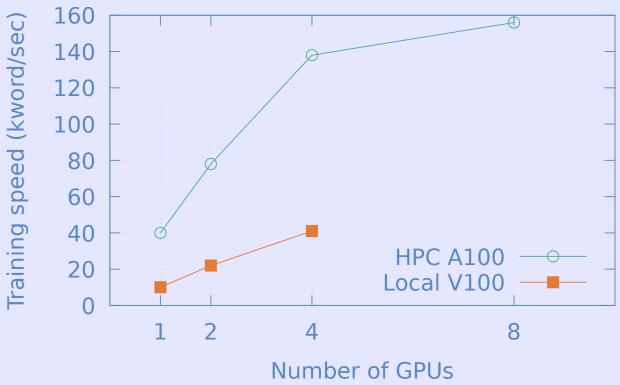
Student models

Find the optimal architecture and data set

- wide scale of experiments but only on one language pair (En$\rightarrow$Fi), outcome:
- training data: teacher output most similar to gold target (measured with sentence level smoothed BLEU)
- architecture: best trade-off between model quality and efficiency (≈ 58M parameters)
  - 12 encoder, 1 decoder layers
  - 8 heads
  - embedding size: 512
  - FFN layer size: 2048
- multilingual models:
  - {Danish, Dutch, German, Swedish} $\rightarrow$ English
  - {Finnish, Hungarian} $\rightarrow$ English

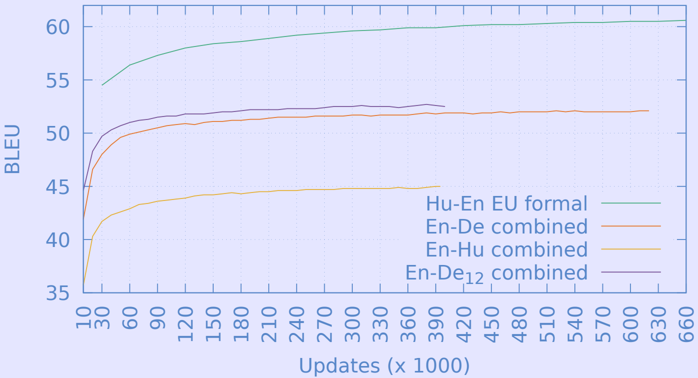
HPC power

Average speed of teacher model trainings



The economical use of power

Convergence of models during training

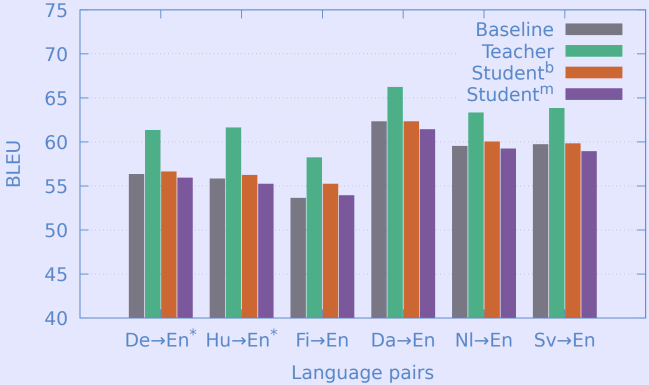


Overview

European Commission

Evaluation of model architectures

BLEU scores on standard in-house test set (10k segments)



^{b} bilingual student models; ^{m} multilingual student models; * four member teacher ensembles

European Commission

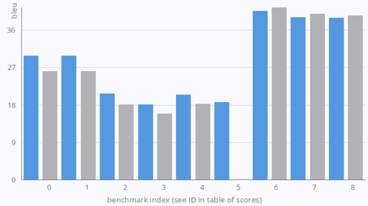
Evaluation of general models

OPUS-MT Dashboard*

[restart] [share link] select language: eng hun [swap] [compare scores] [compare models] [map] [release history] [uploads]

OPUS-MT Dashboard

- **Language pair:** eng-hun
- **Models:** [all models] [OPUS-MT] [external] [compare]
- **Benchmark:** all benchmarks [average score]
- **Evaluation metric:** bleu [spbleu][chrf][chrf++][comet]
- **Chart Type:** [standard][diff]



- blue = OPUS-MT / Tatoeba-MT models, grey = external models

Model Scores (comparing between OPUS-MT and external models)

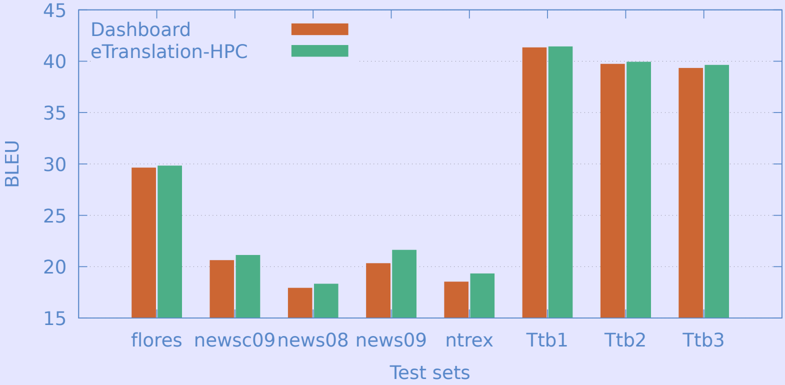
ID	Benchmark (bleu)	Output	OPUS-MT	bleu	external	bleu	Diff
0	flores101-devtest	compare	eng-hun/opus..2022-02-25	29.6	facebook/m2m100_1.2B	25.9	3.7
1	flores200-devtest	compare	eng-hun/opus..2022-02-25	29.6	facebook/m2m100_1.2B	25.9	3.7
2	newssyscomb2009	compare	eng-hun/opus..2022-02-25	20.6	facebook/m2m100_1.2B	17.9	2.7
3	newstest2008	compare	eng-hun/opus..2022-02-25	17.9	facebook/nll..illed-1.3B	15.7	2.2
4	newstest2009	compare	eng-hun/opus..2022-02-25	20.3	NYTK/transla..-128-en-hu	18.1	2.2
5	ntrex128	show	eng-hun/opus..2022-02-25	18.5			18.5
6	tatoeba-test-v2020-07-28	compare	eng-hun/opus..2022-02-25	40.4	NYTK/transla..-128-en-hu	41.3	-0.9
7	tatoeba-test-v2021-03-30	compare	eng-hun/opus..2022-02-25	38.9	NYTK/transla..-128-en-hu	39.7	-0.8
8	tatoeba-test-v2021-08-07	compare	eng-hun/opus..2022-02-25	38.7	NYTK/transla..-128-en-hu	39.3	-0.6
		average		**28.3**		**28.0**	**0.3**

*https://opus.nlpl.eu/dashboard

European Commission

Model comparison on public test sets

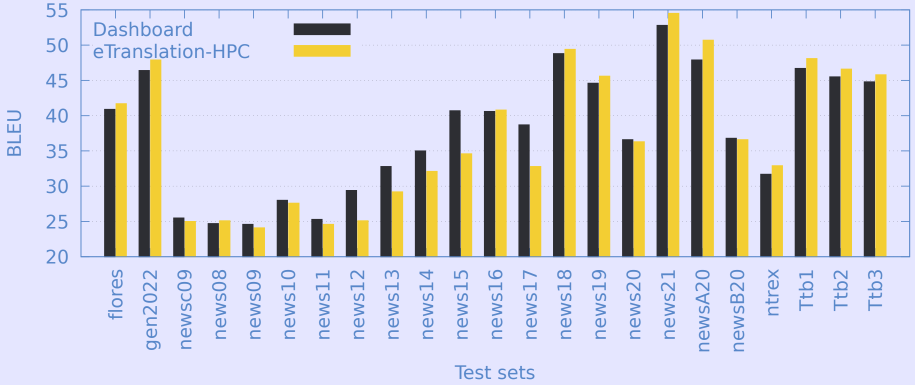
Best En→Hu Dashboard models vs. eTranslation HPC models



European Commission

Model comparison on public test sets

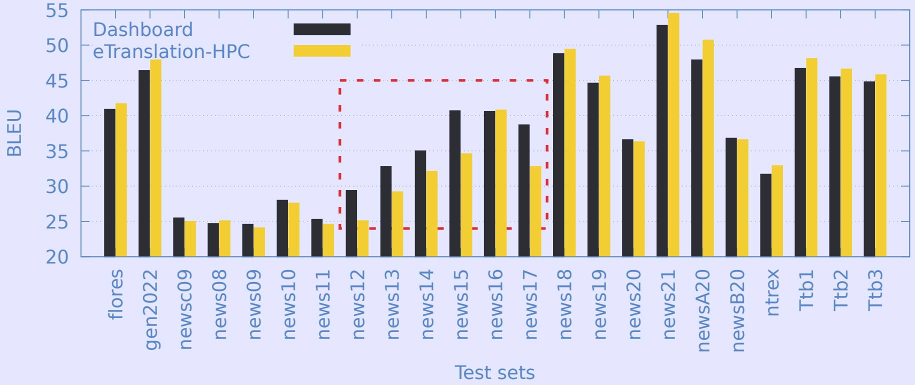
Best En→De Dashboard models vs. eTranslation HPC models



European Commission

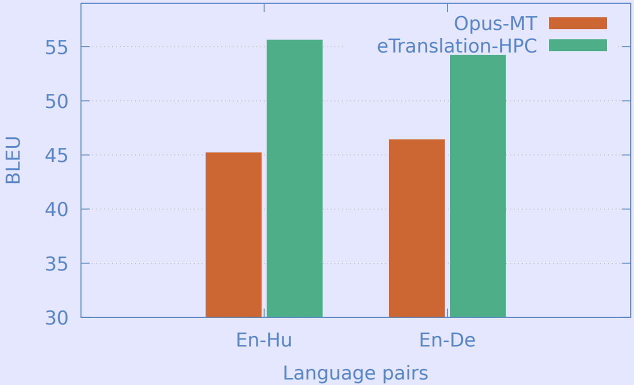
Model comparison on public test sets

Best En→De Dashboard models vs. eTranslation HPC models



The strength of the general models

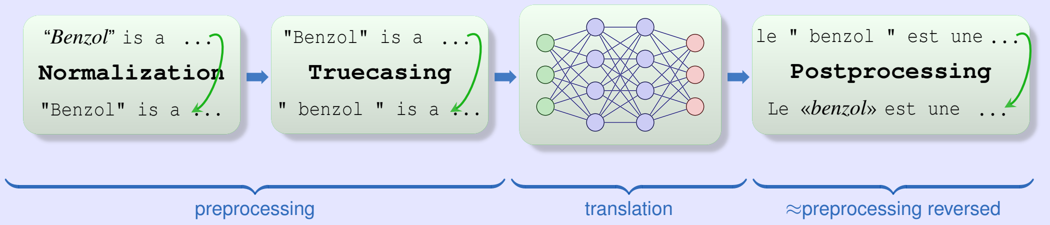
Model comparison on EU formal language test set (10k segments)



Overview

MT pipeline

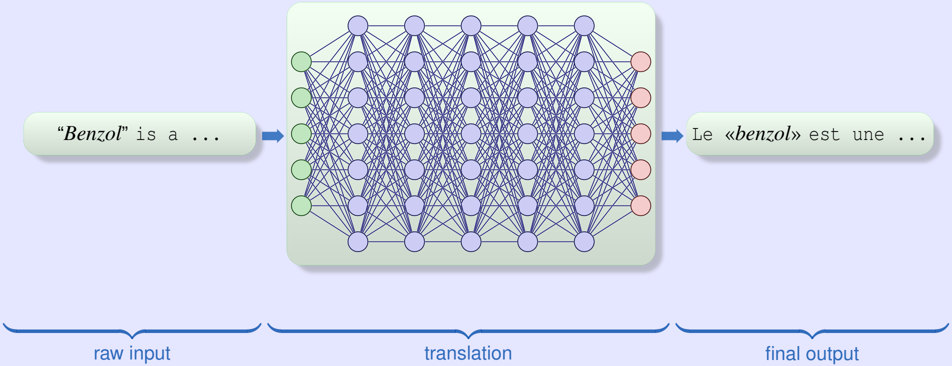
Modules galore



Code to display the neural network is adapted from `https://tikz.net/neural_networks/`

MT pipeline simplified

One model to rule them all



Code to display the neural network is adapted from `https://tikz.net/neural_networks/`

Towards better eTranslation services

Deep teacher models

- directly deployable when quality is primary over translation speed
- resource need (GPU memory) for inference 40% higher but still manageable

Compact student models

- when latency, efficiency and costs are critical factors
- some quality can be sacrificed for efficiency

European Commission

Overview

European Commission

Results and lessons learnt

- with more compute better and very competitive models can be built
- established workflow of building high quality MT models
- smooth integration of models into the eTranslation service pipeline
  - deep models for regular tasks
  - efficient student models under fast response conditions
- for large scale deployment additional costs can be substantial (especially for high-resurce languages) but trade-off is possible

🤗 all teacher models will be open sourced

Acknowledgement

We acknowledge the support of EuroHPC Joint Undertaking in awarding us access to MeluXina at LuxProvide, Luxembourg

European Commission

Cross-Facility Federated Learning

University of Turin – Parallel Computing Group: Iacopo Colonnelli, Robert Birke, Giulio Malenza, Gianluca Mittone, Alberto Mulone, Marco Aldinucci

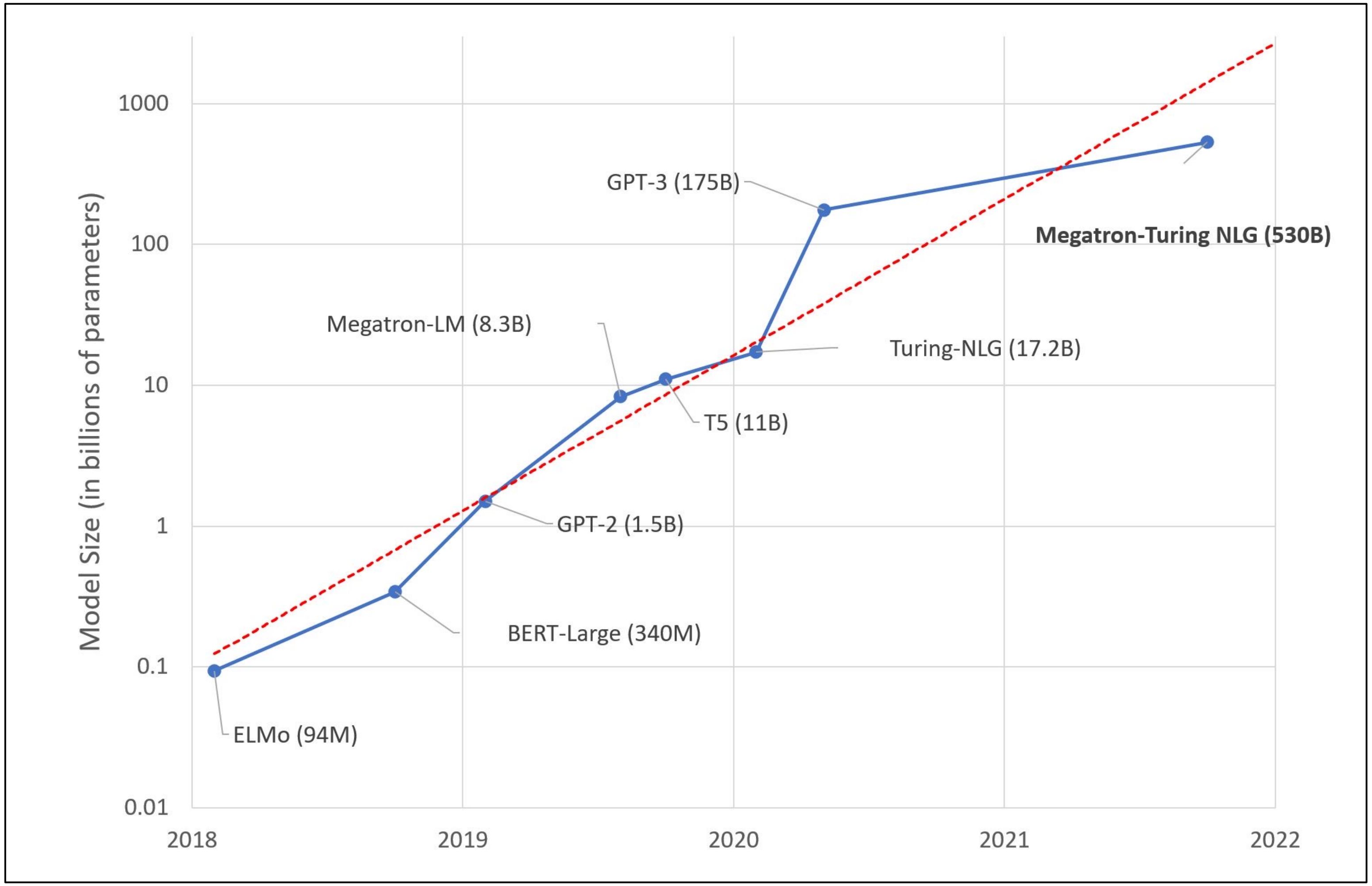
University of Turin – Content Centered Computing Group: Valerio Basile, Marco Antonio Stranisci, Viviana Patti

Delft University of Technology: Jeroen Galjaard, Lydia Y. Chen

CINECA Supercomputing Center: Sanzio Bassini, Massimiliano Guarrasi, Gabriella Scipione

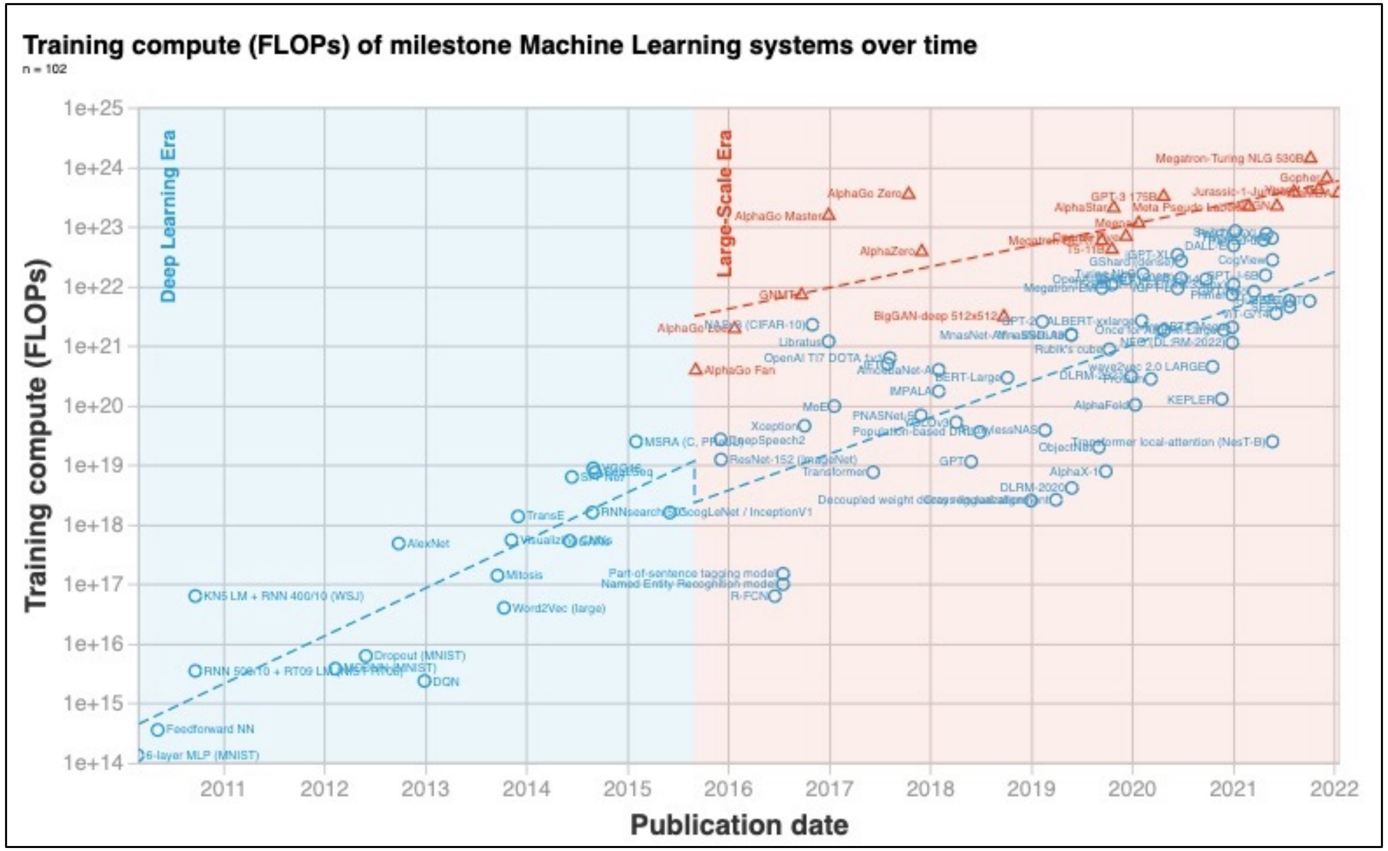
IT4I Supercomputing Center: Jan Martinovic, Vit Vondrák

Large Language Models: A New Moore's Law?



Julien Simon. https://huggingface.co/blog/large-language-models. 2021

Democratize AI → Democratize HPC access



Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, Pablo Villalobos. Compute trends across three eras of Machine Learning. *arXiv Preprint*, arXiv: 2202.05924, 2022.

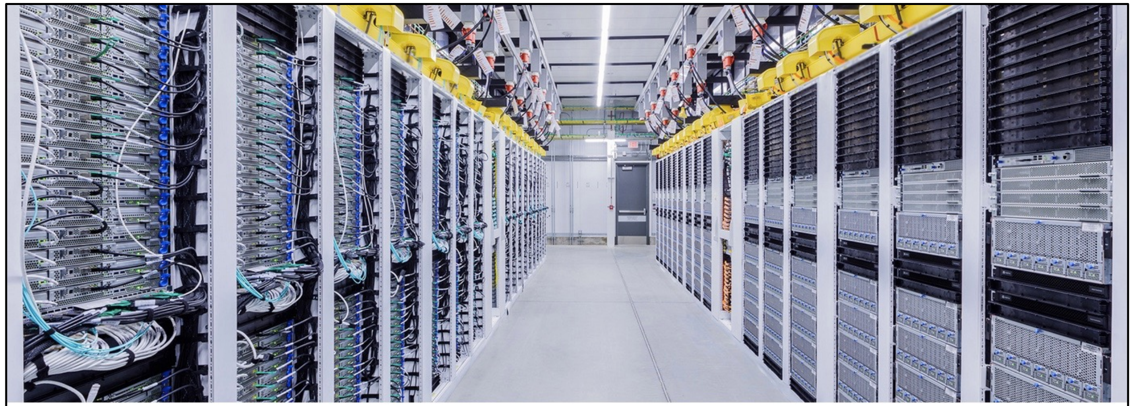
Exclusive: ChatGPT-owner OpenAI is exploring making its own AI chips

By **Anna Tong**, **Max A. Cherney**, **Christopher Bing** and **Stephen Nellis**

October 6, 2023 12:59 PM GMT+2 · Updated 2 months ago

Aa

Anna Tong, Max A. Cherney, Cristopher Bing, and Stephen Nellis. Exclusive: ChatGPT-owner OpenAI is exploring making its own AI chips. *Reuters*. 2023



AI

How Microsoft's bet on Azure unlocked an AI revolution

John Roach. How Microsoft's bet on Azure unlocked an AI revolution. *Microsoft blog*. 2023

Search

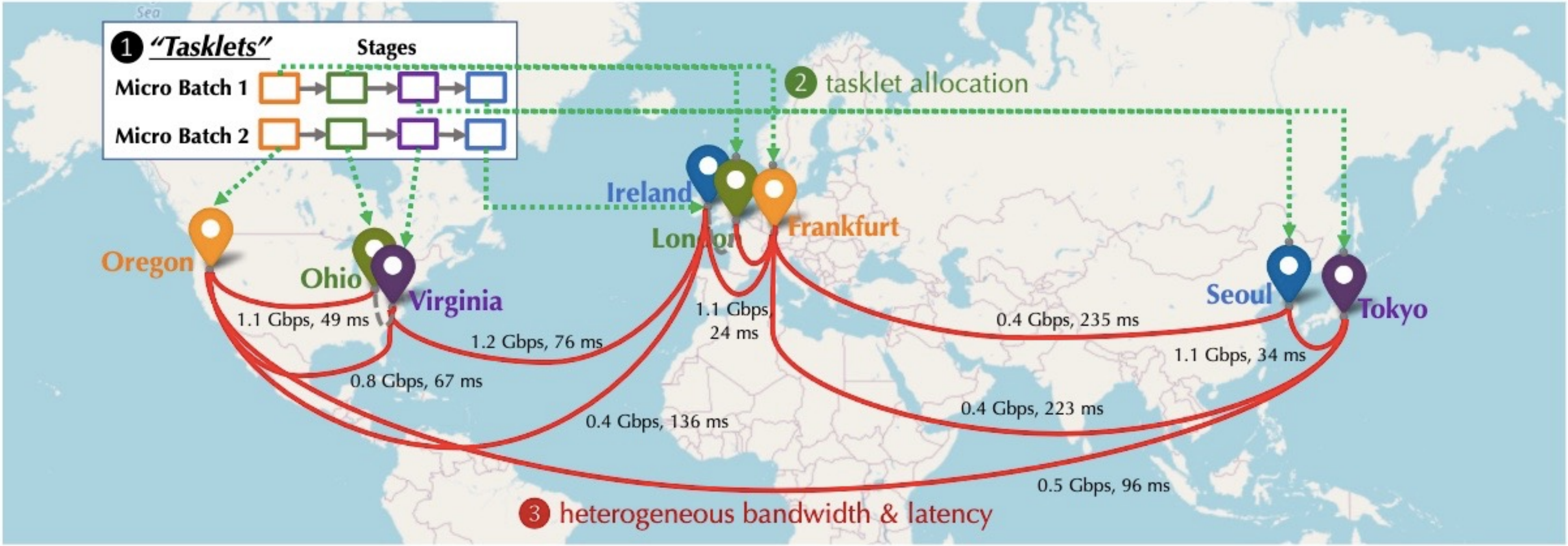
TRILLION PARAMETER CONSORTIUM (TPC)

Home *About the TPC* *Participating Organizations* *Posts*

The overarching focus of the consortium is to bring together groups interested in building, training, and using large-scale models with those who are building and operating large-scale computing systems. The target community encompasses (a) those working on AI methods development, natural language processing/multimodal approaches and architectures, full stack implementations, scalable libraries and frameworks, AI workflows, data aggregation, cleaning and organization, training runtimes, model evaluation, downstream adaptation, alignment, etc.; (b) those that design and build hardware and software systems; and (c) those that will ultimately use the resulting AI systems to attack a range of problems in science, engineering, medicine, and other domains.

https://tpc.dev

Cross-Facility Distributed Training



Binhang Yuan, Yongjun He, Jared Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy S Liang, Christopher Re, and Ce Zhang. Decentralized training of foundation models in heterogeneous environments. *Advances in Neural Information Processing Systems*, 35:25464–25477, 2022.

Cross-Facility Distributed Training

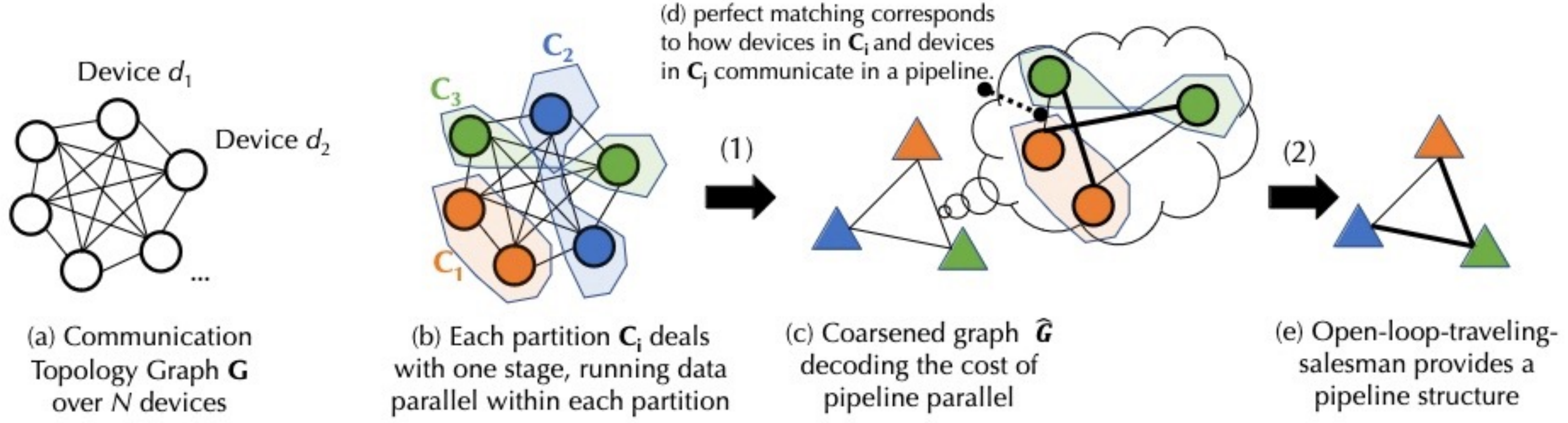


Figure 2: (a) Communication graph $\mathbf{G}$; and (b, c, d, e) an illustration of the cost model given $\mathbf{G}$.

Binhang Yuan, Yongjun He, Jared Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy S Liang, Christopher Re, and Ce Zhang. Decentralized training of foundation models in heterogeneous environments. *Advances in Neural Information Processing Systems*, 35:25464–25477, 2022.

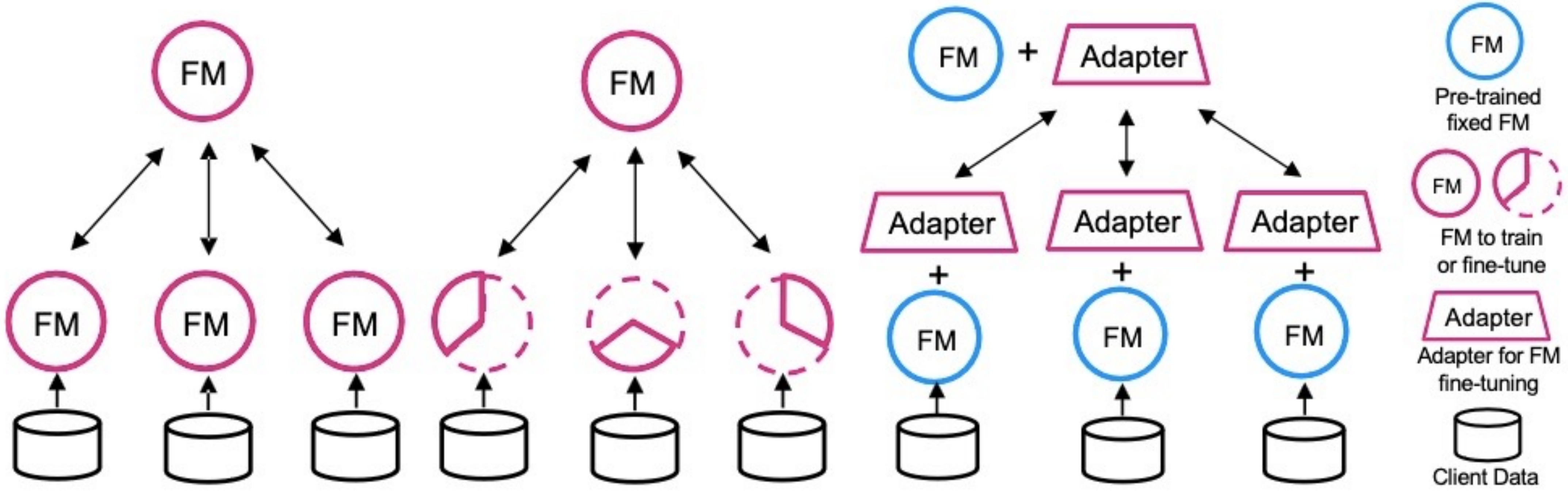
Cross-Facility Federated Learning (XFFL)

Motivation	Challenges	Opportunities and Future Directions
• Shortage of Large-scale High-quality Legalized Data • High Computation Resource Demand • FM Dominance by Affluent Companies • Continuous Data Growth • Data Privacy and Control • Enhancing User Experience through Local Deployment of FMs	• Large Model Incurs High Memory, Communication, and Computation • Challenges in Data Privacy and System Security • IP and Copyright Issues • Complex Incentive Mechanisms for Collaboration	• Integrating FL into the Lifespan of FMs • Memory, Communication, and Computation Reduction • Designing FL System and Benchmark for FM • Improving FM with Decentralized Data • Advancing Trustworthy FL for FM • Exploring Incentive Mechanism in FL for FM

Figure 1: Motivations, challenges, and future directions of Federated Learning for Foundation Model.

Weiming Zhuang, Chen Chen, and Lingjuan Lyu. When Foundation Model Meets Federated Learning: Motivations, Challenges, and Future Directions. *arXiv Preprint*, arXiv:2306.15546, 2023.

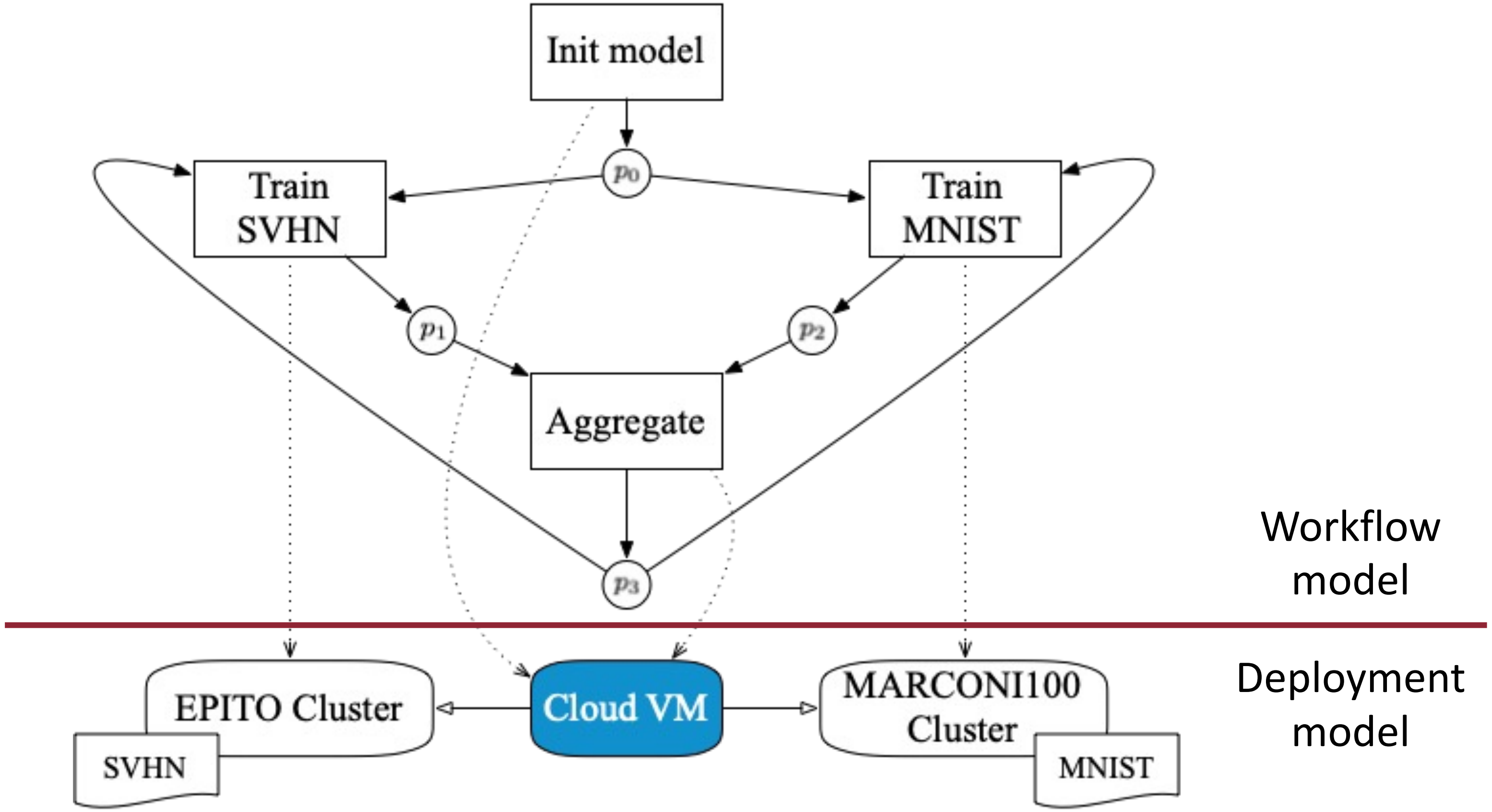
Cross-Facility Federated Learning (XFFL)



(a) Train or fine-tune FM (b) Parallel FM training (c) Fine-tune FM with adapter

Weiming Zhuang, Chen Chen, and Lingjuan Lyu. When Foundation Model Meets Federated Learning: Motivations, Challenges, and Future Directions. *arXiv Preprint*, arXiv:2306.15546, 2023.

Federated Learning as a Workflow



COMMON WORKFLOW LANGUAGE

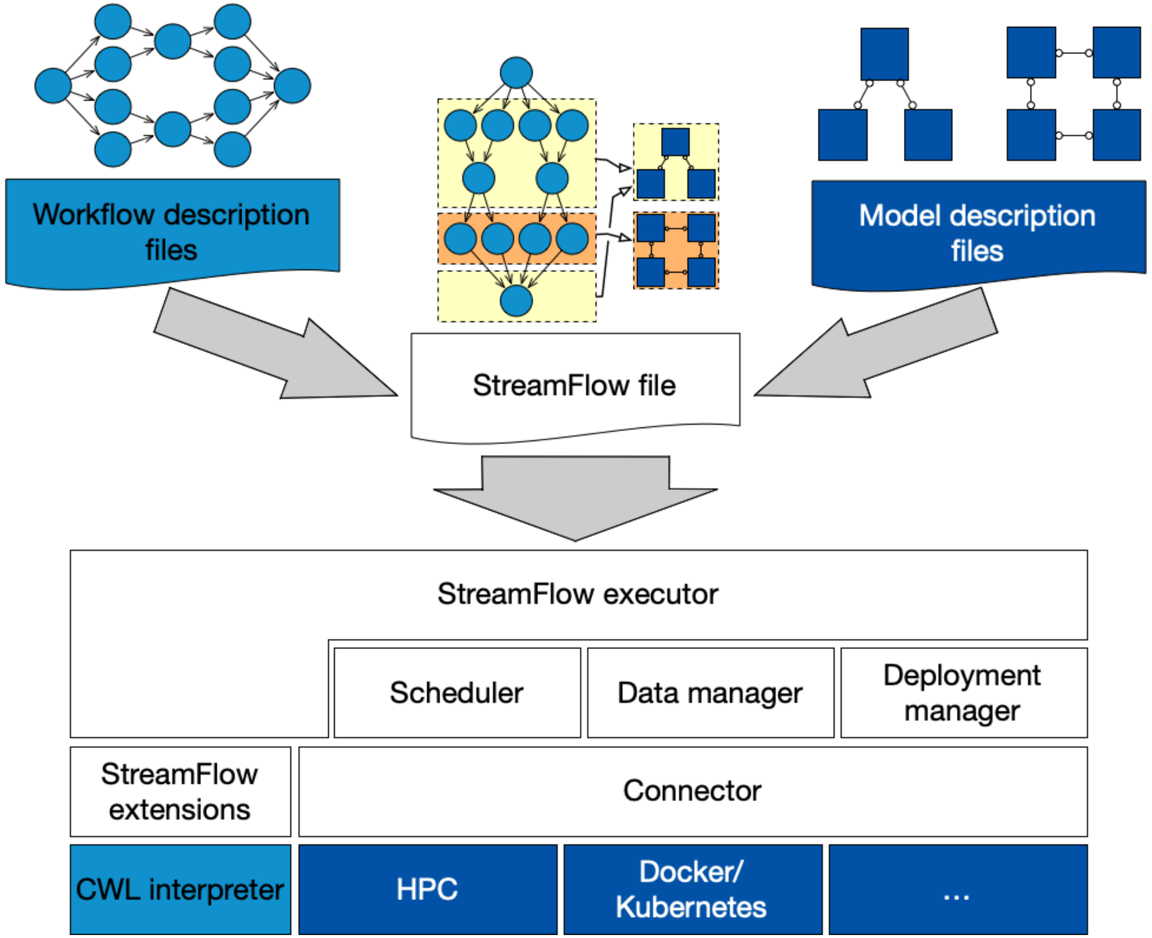
Iacopo Colonnelli, Bruno Casella, Gianluca Mittone, Yasir Arfat, Barbara Cantalupo, Roberto Esposito, Alberto Riccardo Martinelli, Doriana Medić, and Marco Aldinucci. Federated learning meets HPC and cloud. *Astrophysics and Space Science Proceedings*, 60:193–199, 2022.

Portable Federations with StreamFlow

COMMON WORKFLOW LANGUAGE

StreamFlow

https://streamflow.di.unito.it



docker

slurm workload manager

apptainer.org

HyperQueue

flux

Iacopo Colonnelli, Barbara Cantalupo, Ivan Merelli, and Marco Aldinucci. StreamFlow: cross-breeding cloud with HPC. *IEEE Transactions on Emerging Topics in Computing*, 9(4):1723–1737, 2021.

XFFL at scale

Training Llama2-7B on the EuroHPC network

XFFL at Scale: Llama2-7B on EuroHPC

Task: train Llama2-7B for Italian and Czech using a **prompt-tuning approach** for an open-ended generation task:

- **Feed a template** “scrivi un seguente documento/Napište dokument:: {{text}}” with all the Italian and Czech documents included in the multilingual version of C4;
- **Compute the perplexity** between the generated text and the document passed on the template.

	Training Data	Params	Context Length	GQA	Tokens	LR
Llama 1	*See Touvron et al. (2023)*	7B	2k	✗	1.0T	3.0×10^{-4}
		13B	2k	✗	1.0T	3.0×10^{-4}
		33B	2k	✗	1.4T	1.5×10^{-4}
		65B	2k	✗	1.4T	1.5×10^{-4}
Llama 2	*A new mix of publicly available online data*	7B	4k	✗	2.0T	3.0×10^{-4}
		13B	4k	✗	2.0T	3.0×10^{-4}
		34B	4k	✓		
		70B	4k	✓		

Size on disk: 13GB

		Time (GPU hours)	Power Consumption (W)	Carbon Emitted (tCO_2eq)
Llama 2	7B	184320	400	31.22
	13B	368640	400	62.44
	34B	1038336	350	153.90
	70B	1720320	400	291.42
Total		3311616		539.00

Hugo Touvron et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv Preprint*, arXiv:2307.09288, 2023.

XFFL at Scale: Llama2-7B on EuroHPC

CINECA

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

Custom BullSequana X2135 “Da Vinci” blades:

- 1 x CPU Intel Xeon 8358 32 core, 2.6 GHz
- 512 (8 x 64) GB RAM DDR4 3200 MHz
- 4 x GPU NVidia A100 SXM6 64GB HBM2
- 2 x Card NVidia HDR 2×100 Gb/s

HPE Apollo 6500 Gen10 blades:

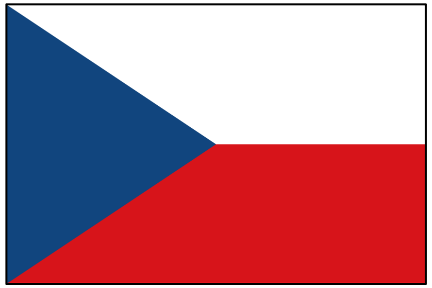
- 2 x CPU AMD EPYC 7763, 64 core, 2.45 GHz
- 1024 GB RAM DDR4 3200 MHz
- 8 x GPU NVidia A100 40GB HBM2
- 4 x InfiniBand 200 Gb/s

XFFL at Scale: Llama2-7B on EuroHPC

https://huggingface.co/datasets/gsarti/clean_mc4_it

Cleaned Italian mC4 Corpus:
- Size: 102GB
- Documents: 10M
- Tokens: 20G

Gabriele Sarti and Malvina Nissim. IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation. *arXiv Preprint*, arXiv:2203.03759, 2022.

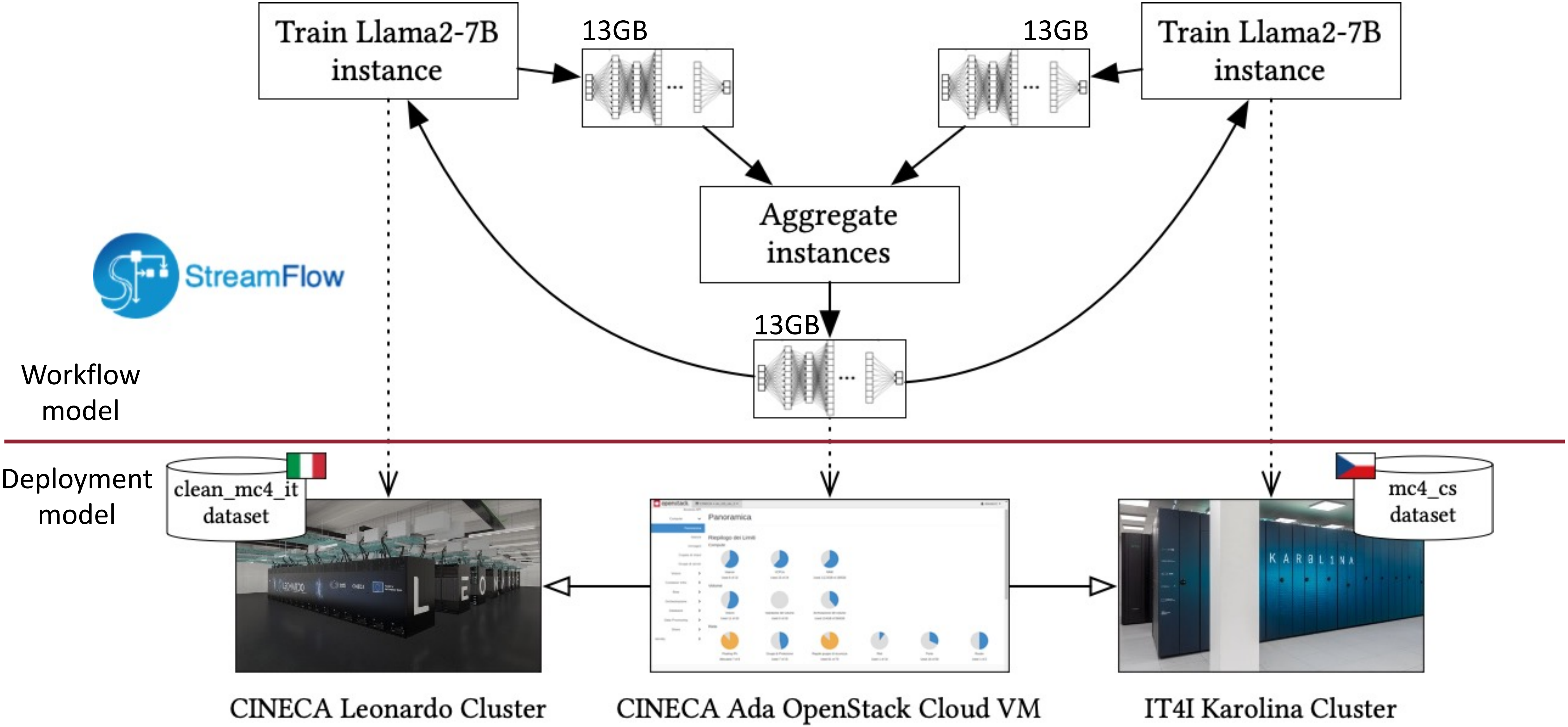


https://huggingface.co/datasets/mc4/viewer/cs

Subset of the Czech mC4 Corpus:
- Size: 169GB
- Documents: 10M
- Tokens: 20G

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498, 2021.

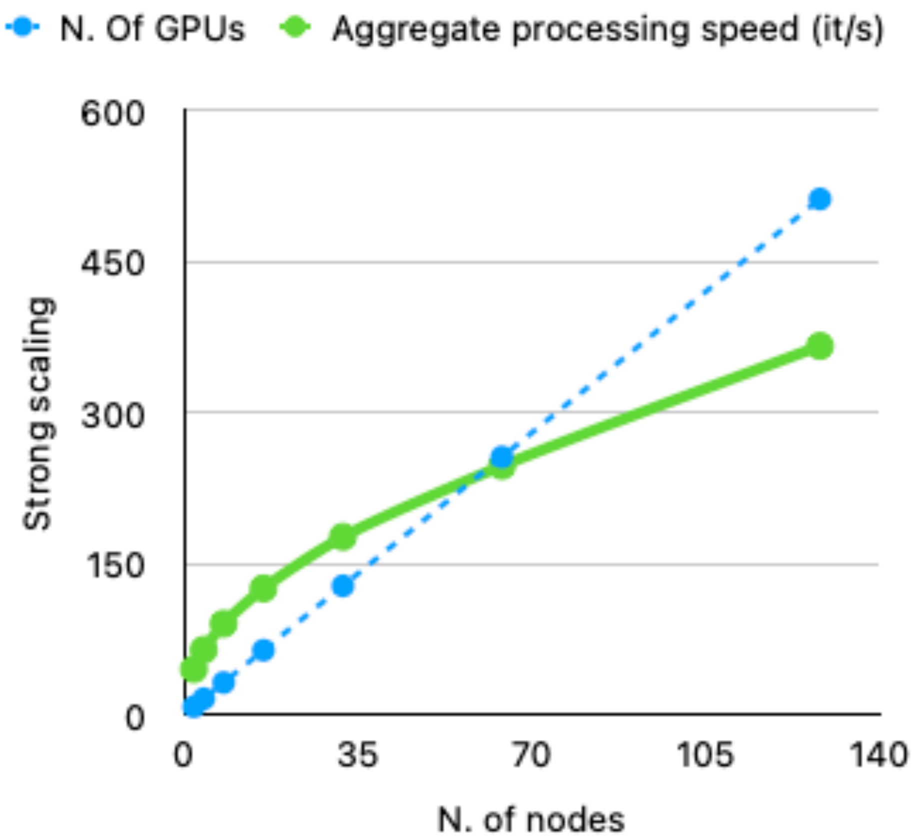
XFFL at Scale: Llama2-7B on EuroHPC



XFFL at Scale: Llama2-7B on EuroHPC

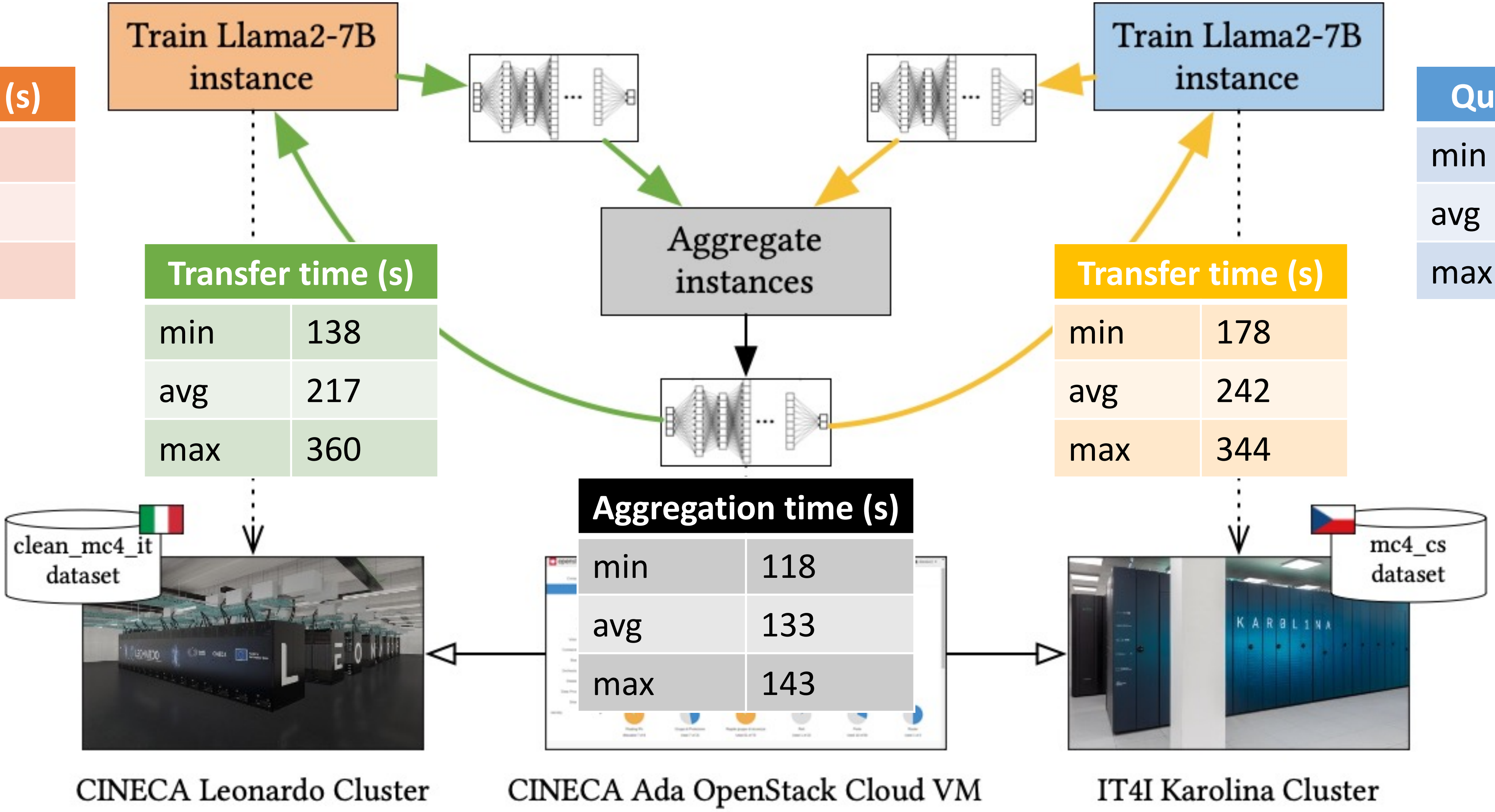
LLaMA-2 7B training on Leonardo@CINECA

N. of nodes	N. Of GPUs	Loading time (s)	Dataset processing speed per node (it/s)	Aggregate processing speed (it/s)	Tot Execution time (hours)	Node speedup
2	8	34	22,64	45,28	774	2
4	16	34	16,12	64,48	385	4
8	32	34	11,3	90,4	193	8
16	64	34	7,84	125,44	98	15,8
32	128	38	5,52	176,64	49	31,6
64	256	90	3,86	247,04	25	61,9
128	512	120	2,86	366,08	14	110,6
clean_mc4_it - Training Set Length = 4085342, Validation Set Length = 13252						



Estimated time to train Llama2-7B with clean_mc4_it on Leonardo

XFFL at Scale: Llama2-7B on EuroHPC



Queue time (s)	
min	0
avg	34
max	164

Transfer time (s)	
min	138
avg	217
max	360

Aggregation time (s)	
min	118
avg	133
max	143

Transfer time (s)	
min	178
avg	242
max	344

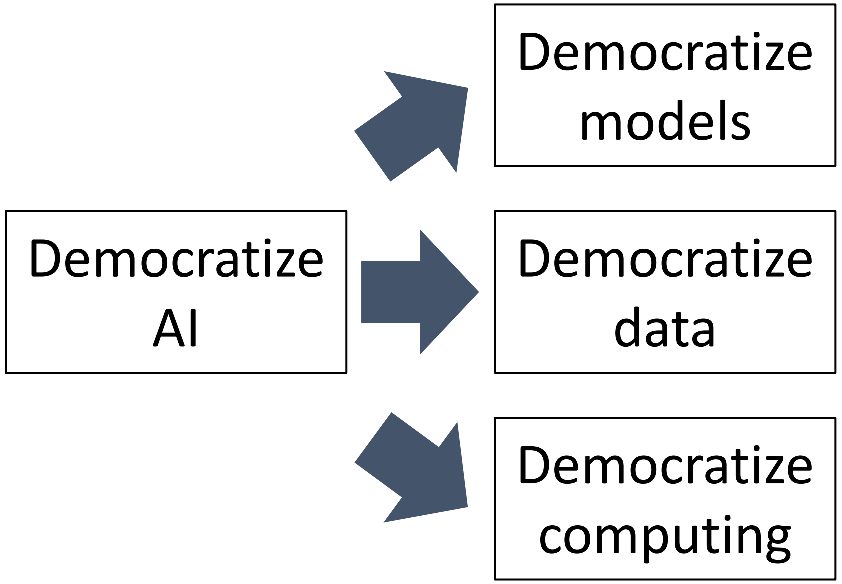
Queue time (s)	
min	0
avg	175
max	699

Measured overhead for a small Federated Learning setting (8 GPUs)

Conclusion

What to do now?

The Chain of Democratization



The Chain of Democratization

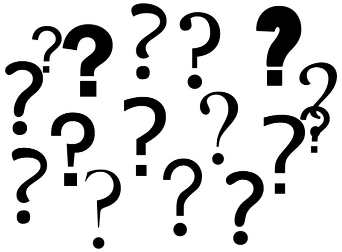
Democratize AI

Democratize models

Democratize data

Hugging Face

Democratize computing



The Chain of Democratization

Democratize AI

Democratize models

Democratize data

Hugging Face

Democratize computing

Research collaborations

Portable HPC federations

The Chain of Democratization

Democratize AI → Democratize models

Democratize AI → Democratize data

Democratize AI → Democratize computing

Democratize computing → Research collaborations

Democratize computing → Portable HPC federations

Meta

Hugging Face

TRILLION PARAMETER CONSORTIUM (TPC)

FENIX RESEARCH INFRASTRUCTURE

The Chain of Democratization

Democratize AI → Democratize models

Democratize AI → Democratize data

Democratize AI → Democratize computing

Democratize computing → Research collaborations

Democratize computing → Portable HPC federations

Meta

Hugging Face

TRILLION PARAMETER CONSORTIUM (TPC)

FENIX RESEARCH INFRASTRUCTURE

StreamFlow

LEXIS THE PLATFORM

EuroHPC Portable Federations: What’s next?

- Experiment with **larger-scale workloads** (Llama-70B, whole mc4 dataset, …)
- Experiment with **larger federations** (more data centres, geographically distributed, …)
- Experiment the portable federation approach with **different workloads** (ensembles of large-scale simulations, hybrid quantum/classical computing, …)

Web page: https://hpc4ai.unito.it/hpc-federation

Contact me: iacopo.colonnelli@unito.it

HPC federation website

Thank you!

Any question?