AI RESEARCH AT CERN

Sofia Vallecorsa – CERN

May 29th, 2024

ML/DL AT CERN

Machine Learning since LEP years

- **Limited abstraction**
- **Preliminary feature engineering**
- Mostly for **classification & regression**

Then came **Deep Learning…**

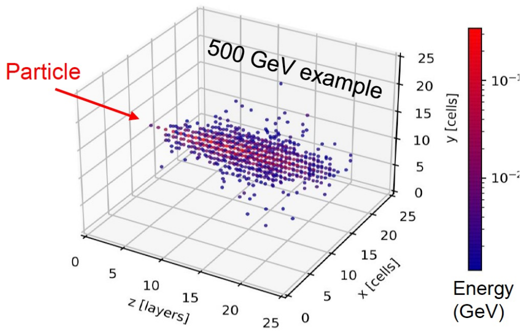
- **Increasingly abstract representations**
- **Complex computational graphs requiring larger infrastructure**
- A large range of **applications**

RESEARCH AREAS

- **Physics**
- **Computing infrastructure** optimization
- AI for **sustainability**
- **Sustainable AI**: workload optimisation wrt hw and computing models

PHYSICS APPLICATIONS

SIMULATION EXAMPLE: DETECTOR RESPONSE AS IMAGES



Rehm, Florian, et al. *arXiv:2105.08960* (2021).

Monte Carlo simulation of detector response is extremely **demanding in terms of computing resources**

→ 50 % of LHC Computing Grid resources today

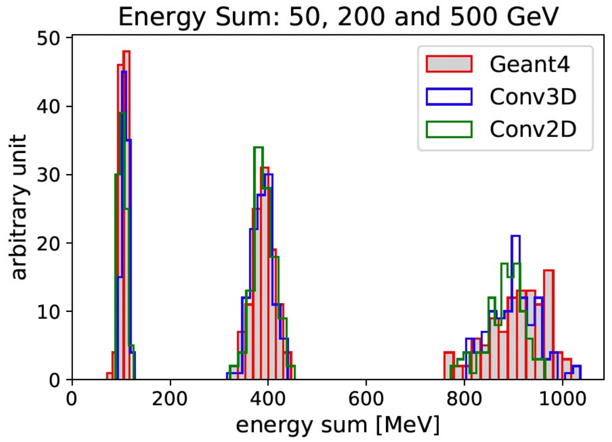
Interpret **detector output as images**

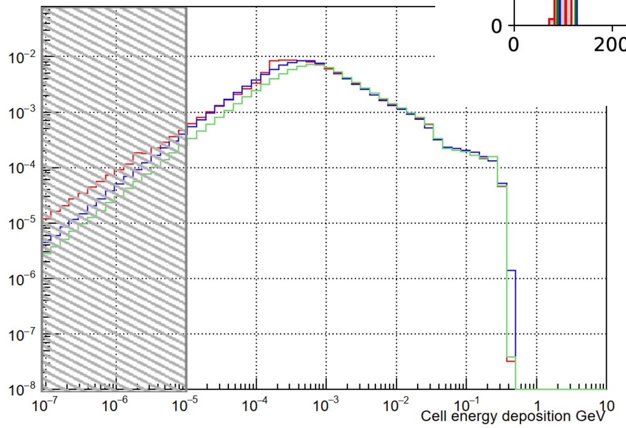
Sensors outputs become pixels in a image

Use **computer vision techniques** to interpret results

Replace Monte Carlo approach with **Generative Models**

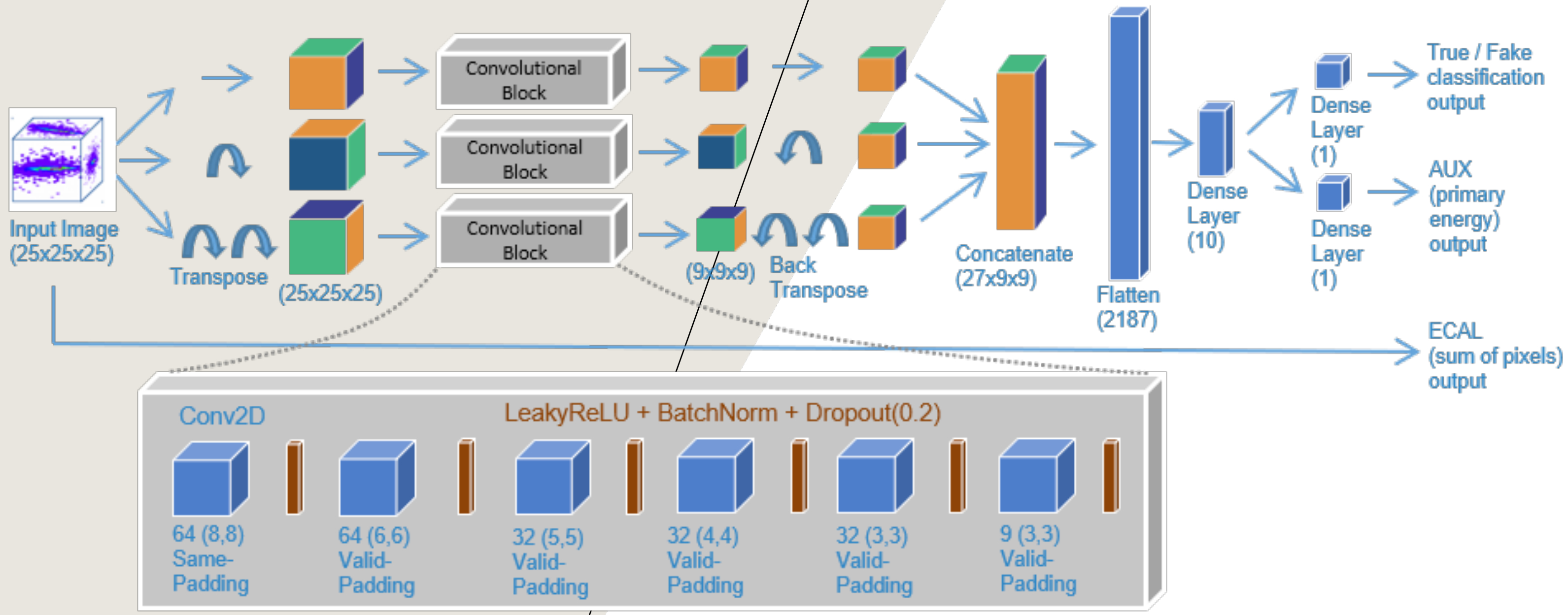
Interfacing DL to standard software is not trivial!

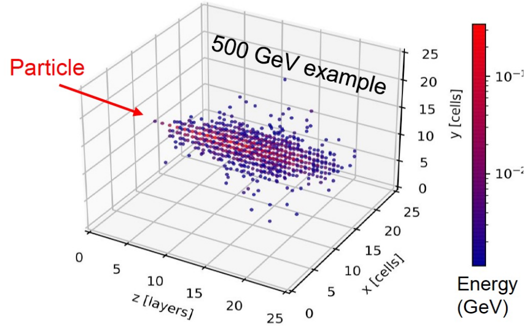


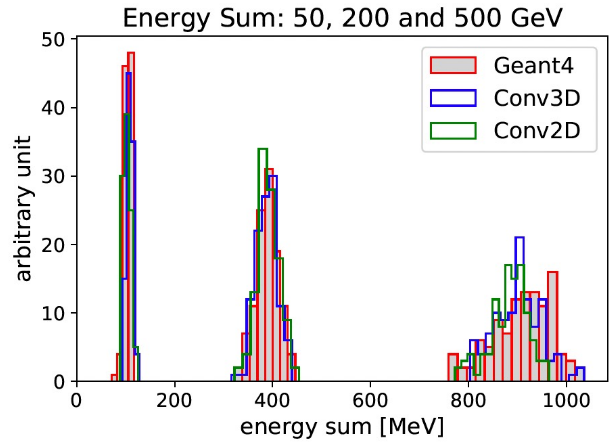


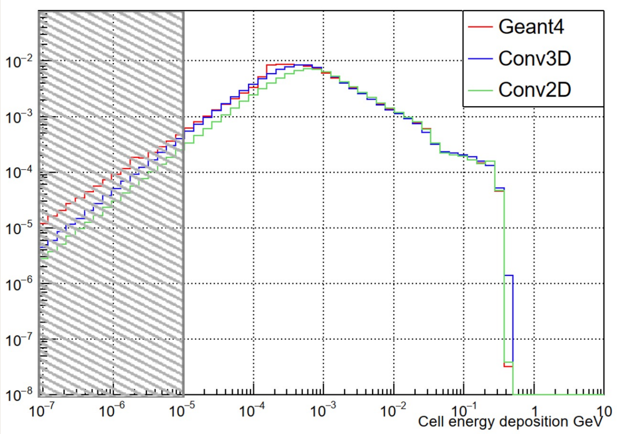
GAN-BASED SIMULATION

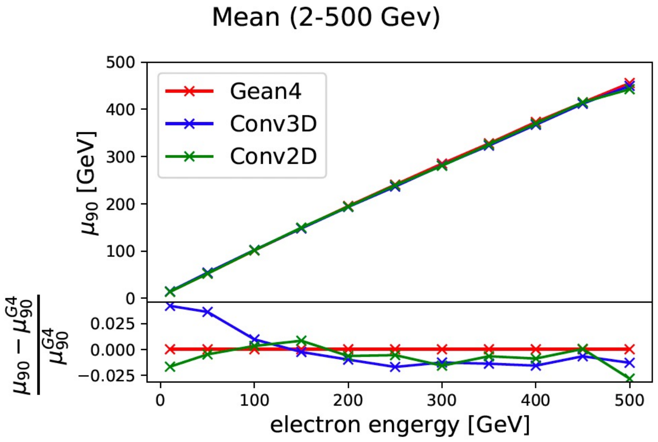
Rehm, Florian, et al. *arXiv:2105.08960* (2021).

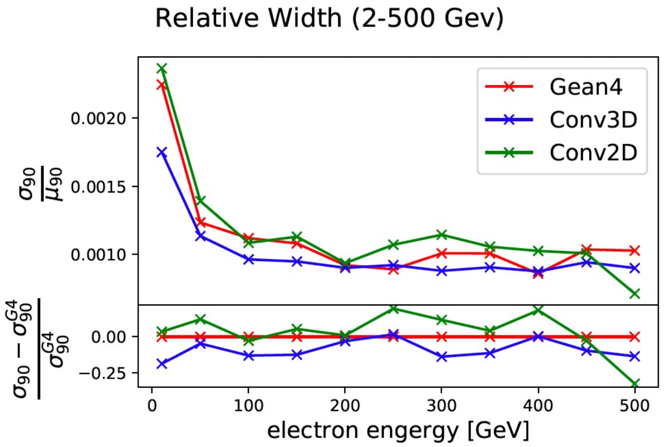






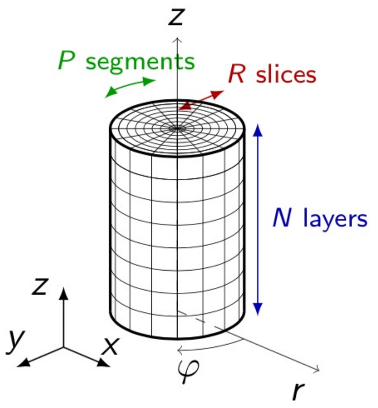


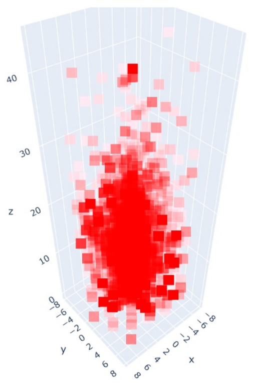




Renato Cardoso, et al.
CHEP 2024

DIFFUSION + TRANSFORMERS



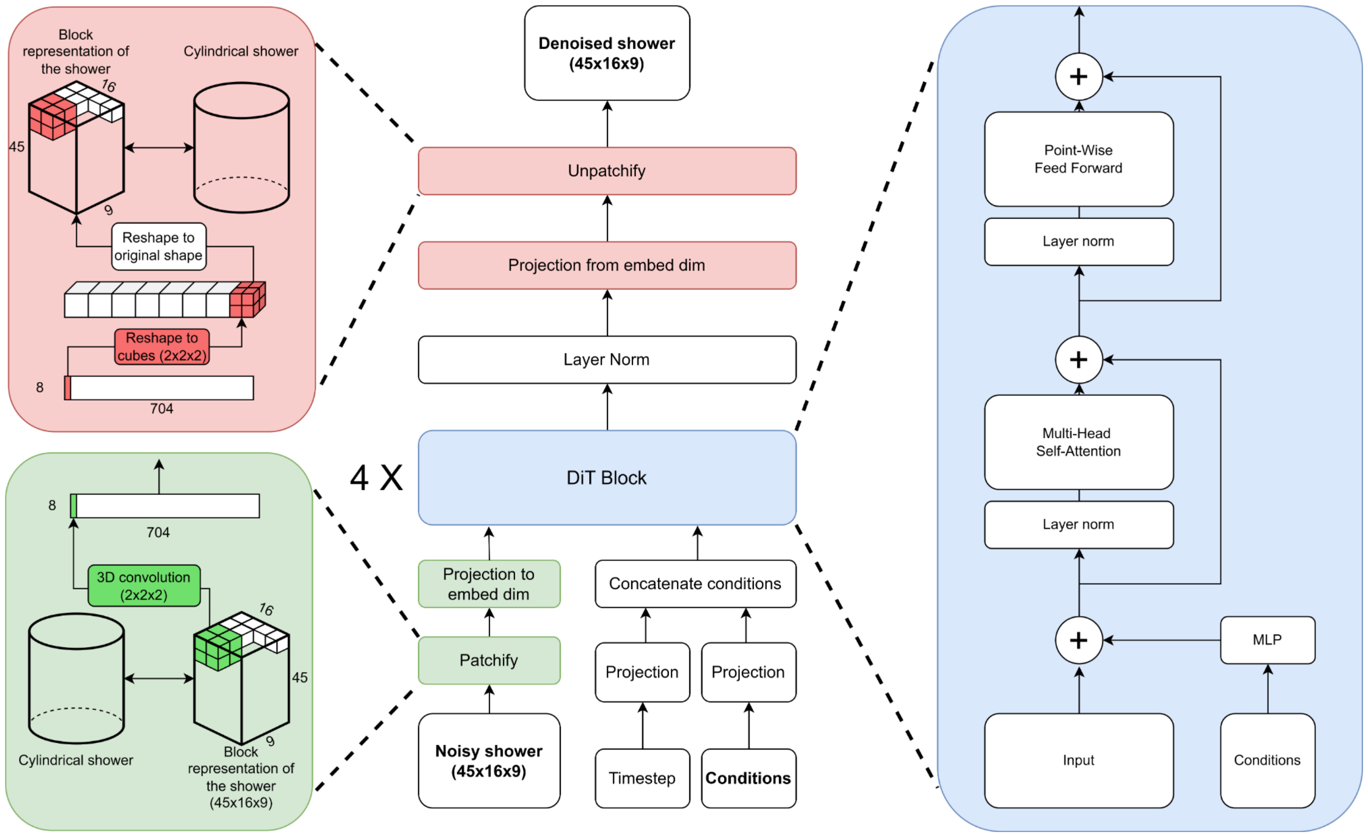


Change to cylindrical geometry (more realistic)

Match SOA diffusion models to transformers to ensure:

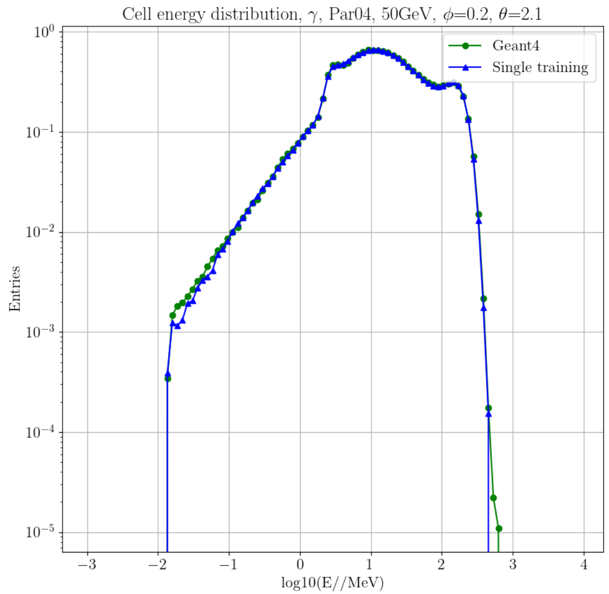
High quality images

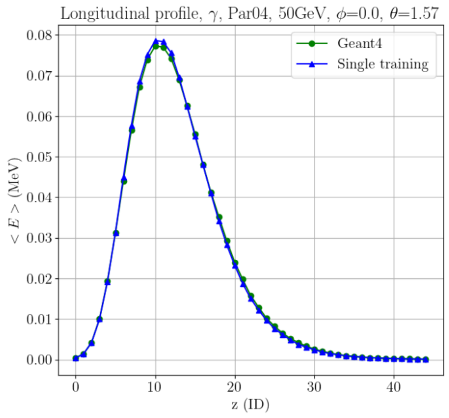
Generalizable results

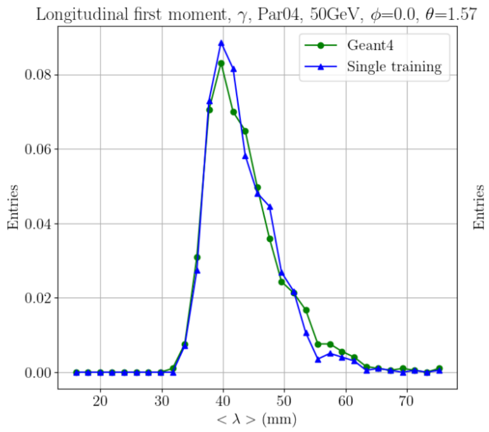


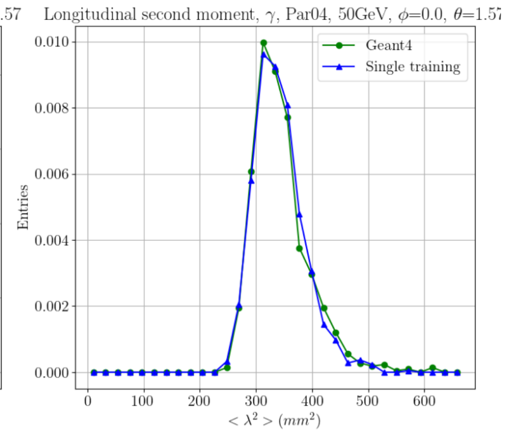
RESULTS

Pixel distributions are correct over a range spanning 5 orders of magnitude



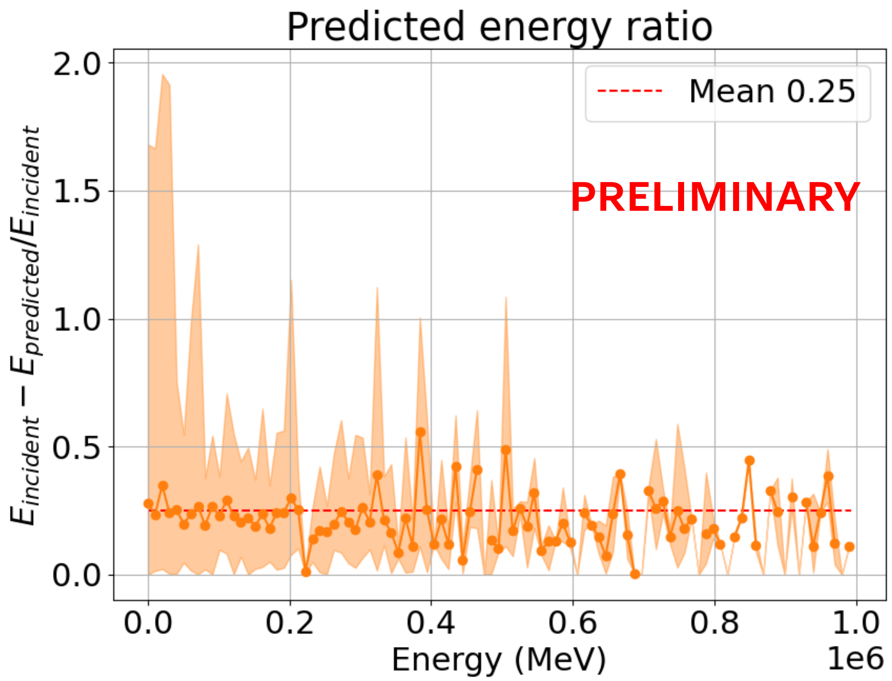






Adaptability to Multi-Tasking:

From image generation to regression

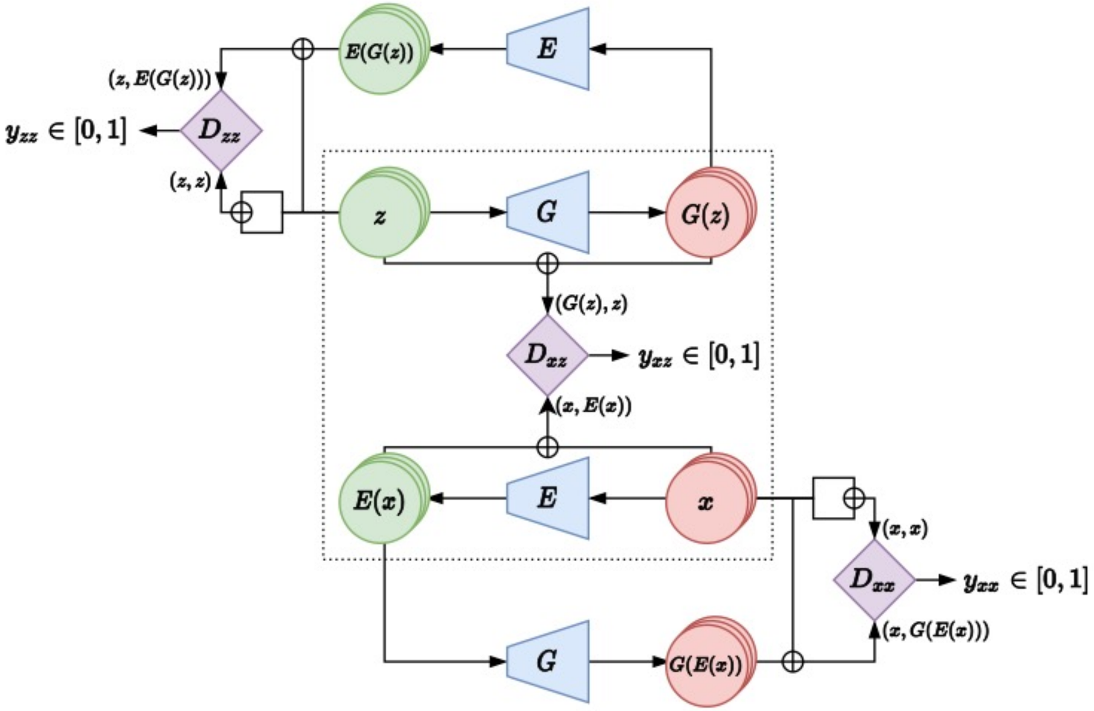


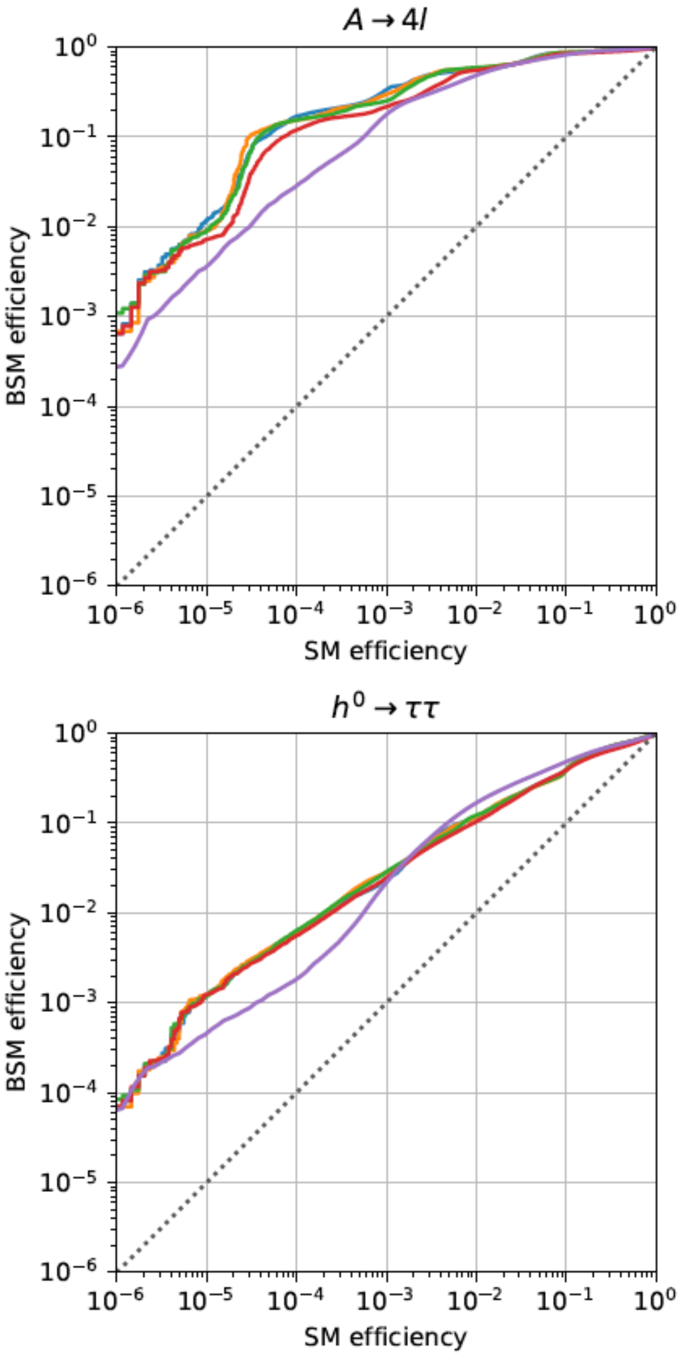
arxiv: 2005.01598

PHYSICS MINING AS ANOMALY DETECTION

- Classical strategy uses very **loose selection**
  - 1M Standard Model ("known physics") events per day
- Train **Variational AD models** on known physics

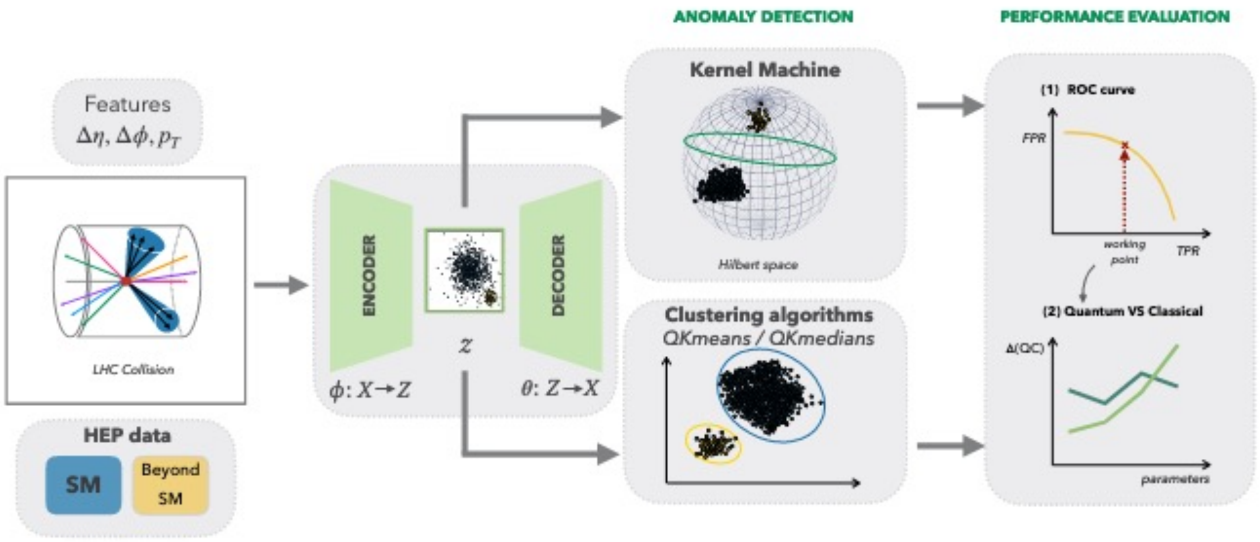
Real time application to store only "**anomalies**"

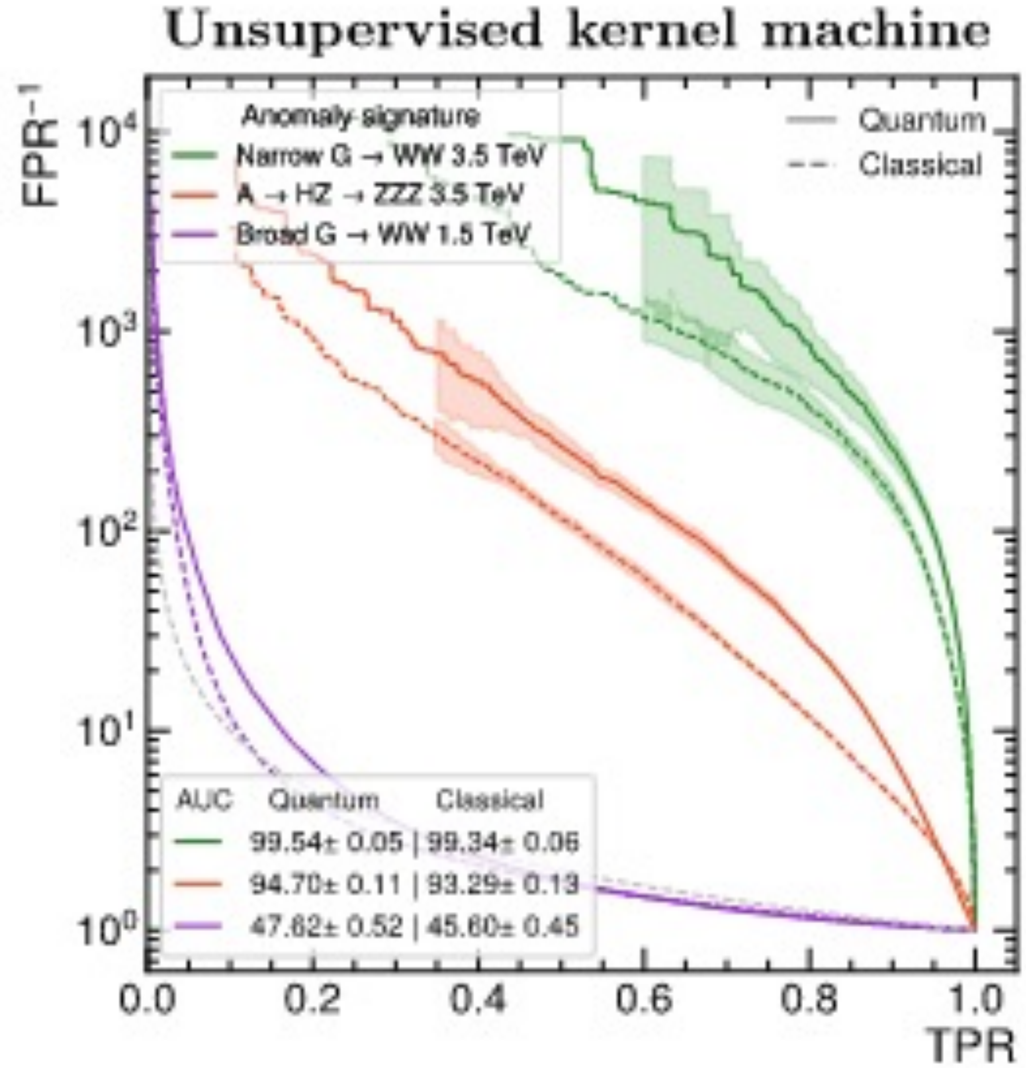




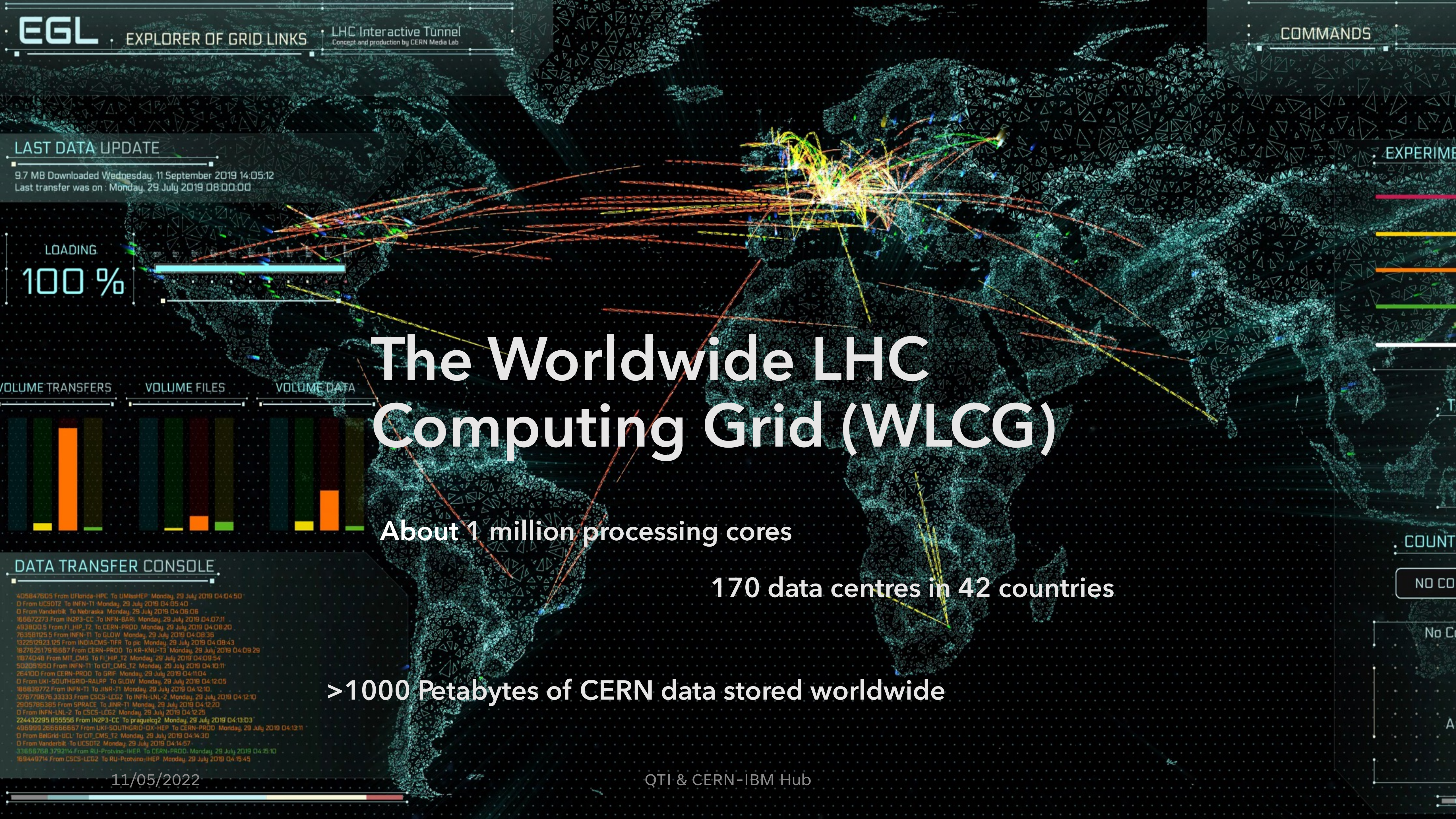
arxiv:2301.10780

DATA COMPRESSION + UNSUPRVISED SVM





INFRASTRUCTURE OPTIMIZATION



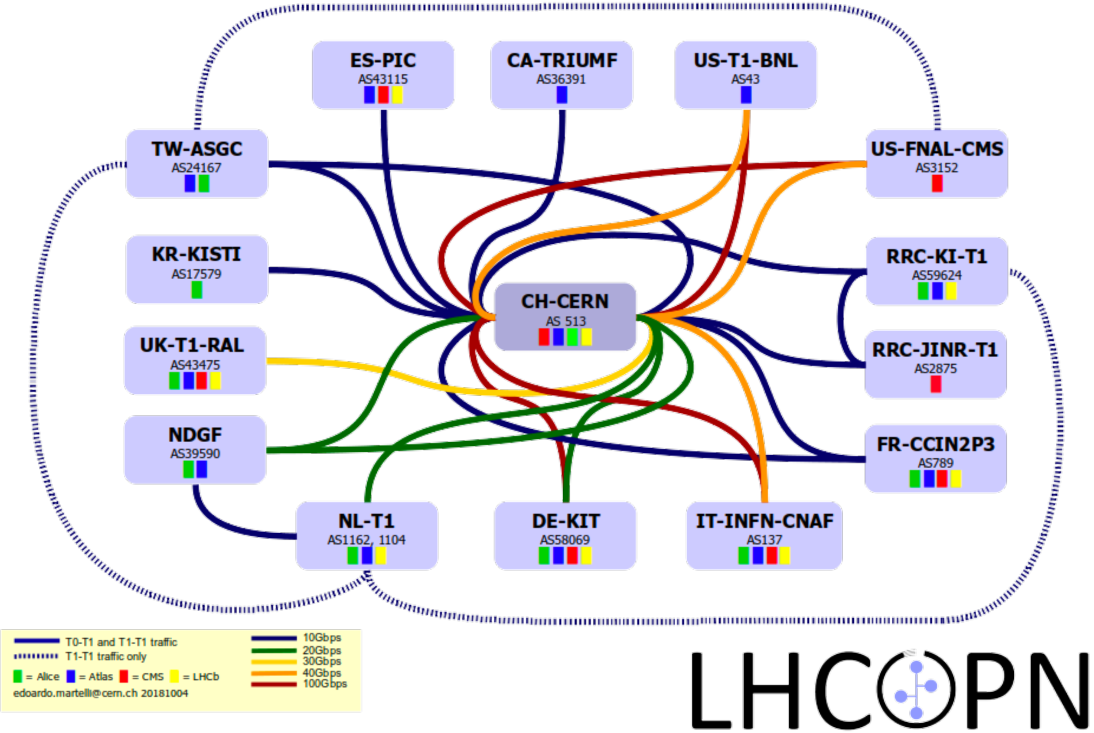
The Worldwide LHC Computing Grid (WLCG)

About 1 million processing cores

170 data centres in 42 countries

>1000 Petabytes of CERN data stored worldwide

NETWORK TRAFFIC PREDICTION

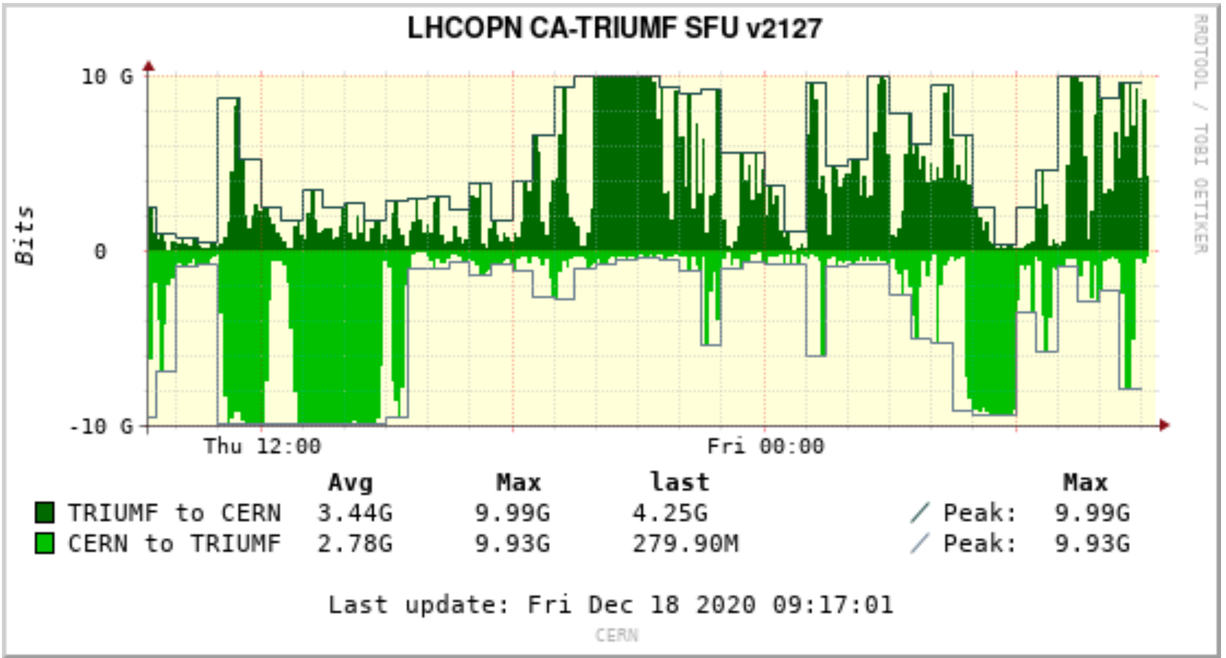


LHCOPN (Large Hadron Collider Optical Private Network) topology

Network traffic on CERN – TRIUMF link

Predict saturation (can occur in both directions)

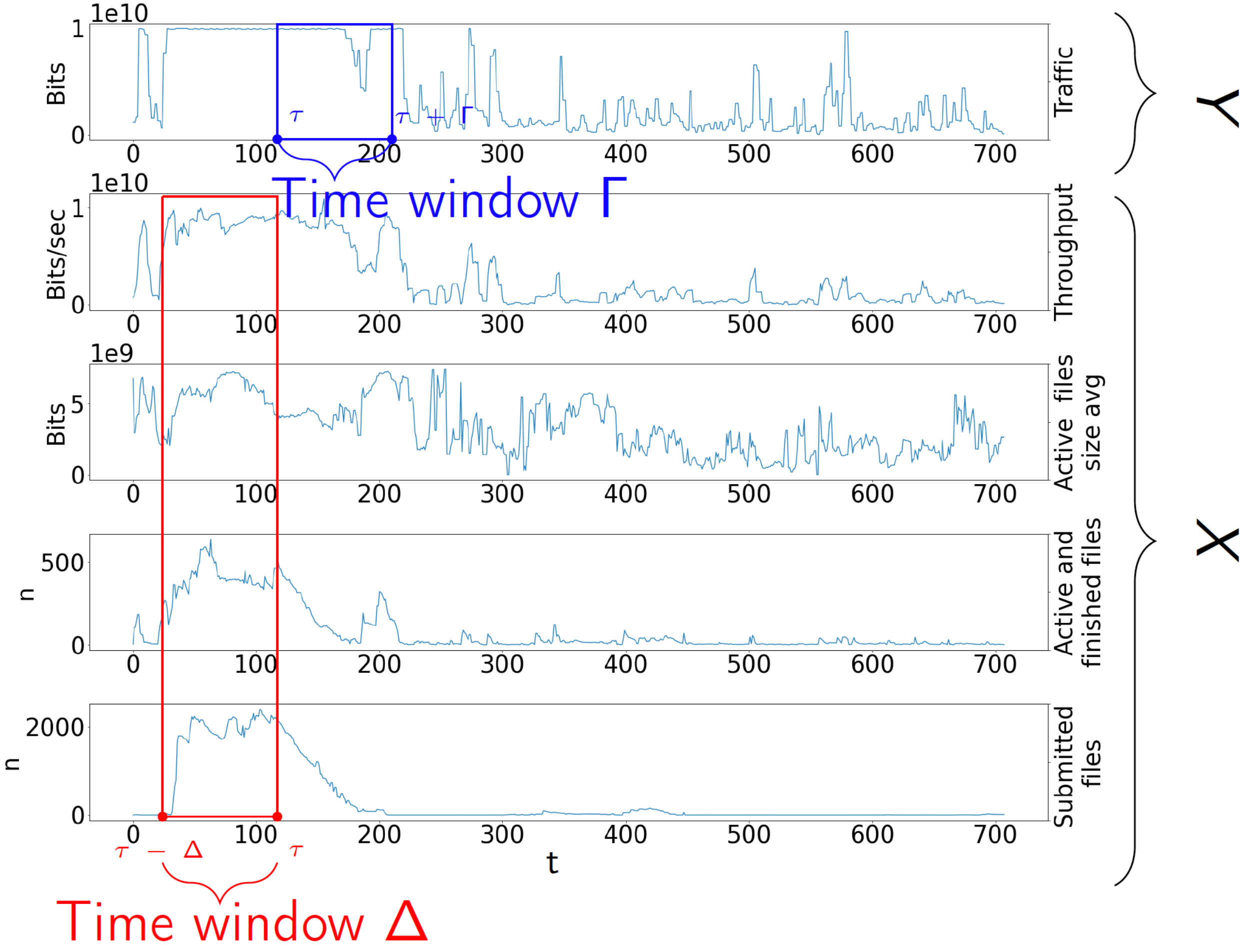
Optimise transfer: automatically modify network devices configuration (SDNC)

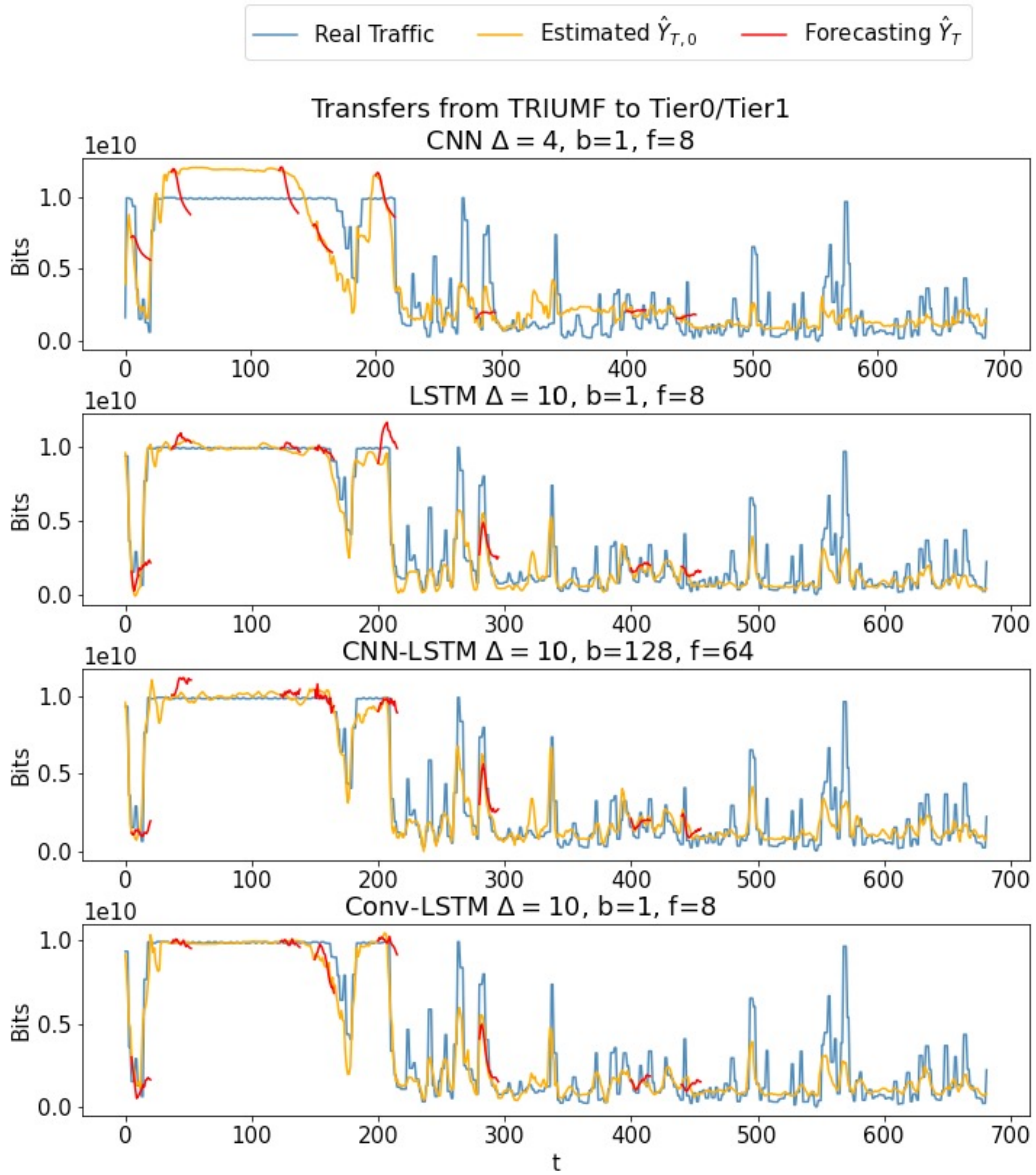


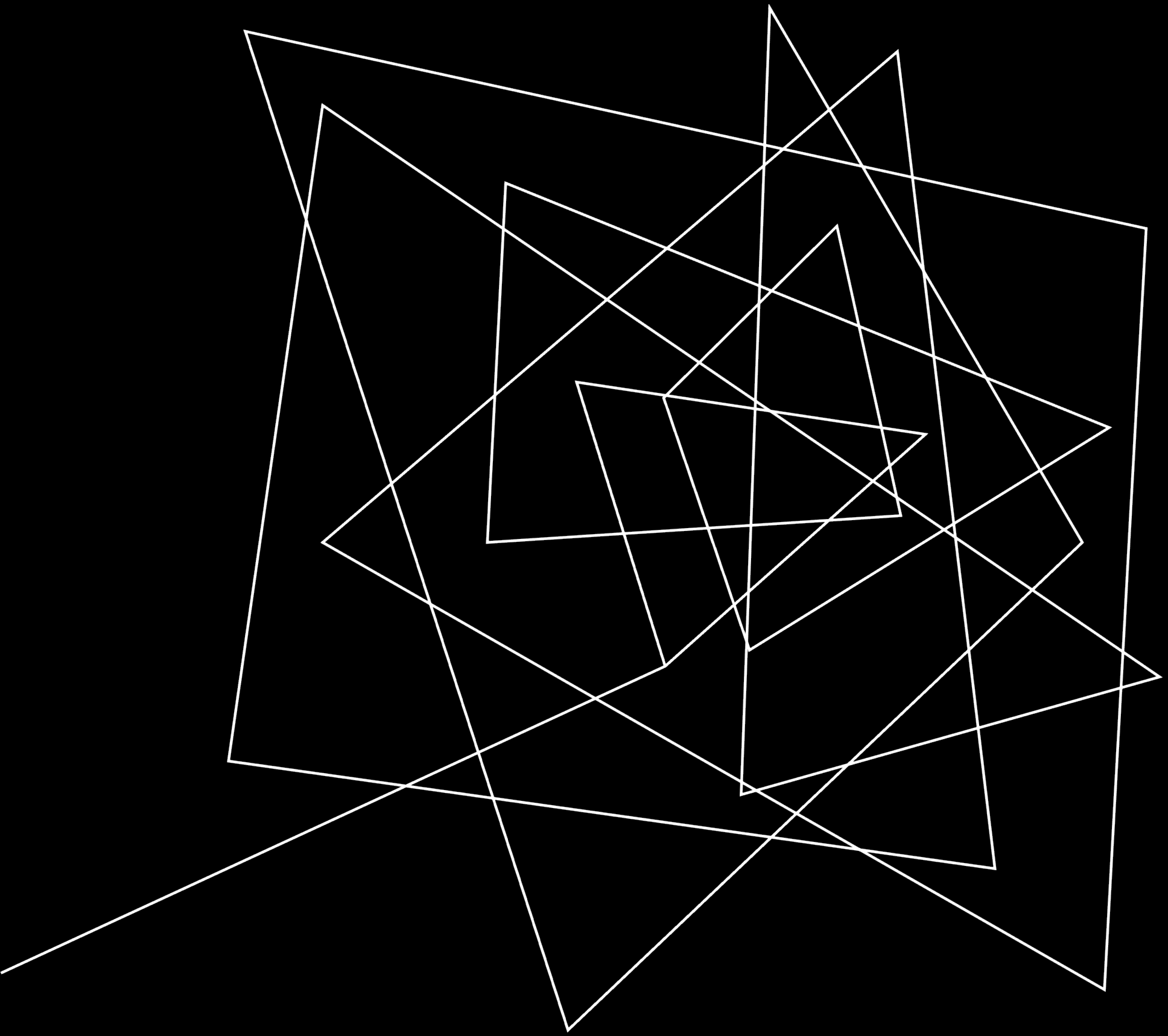
PERFORMANCE

Joanna Waczynska, vCHEP2021, Grid21
arxiv: 2107.02496

Compare CNN, LSTM and hybrid architectures





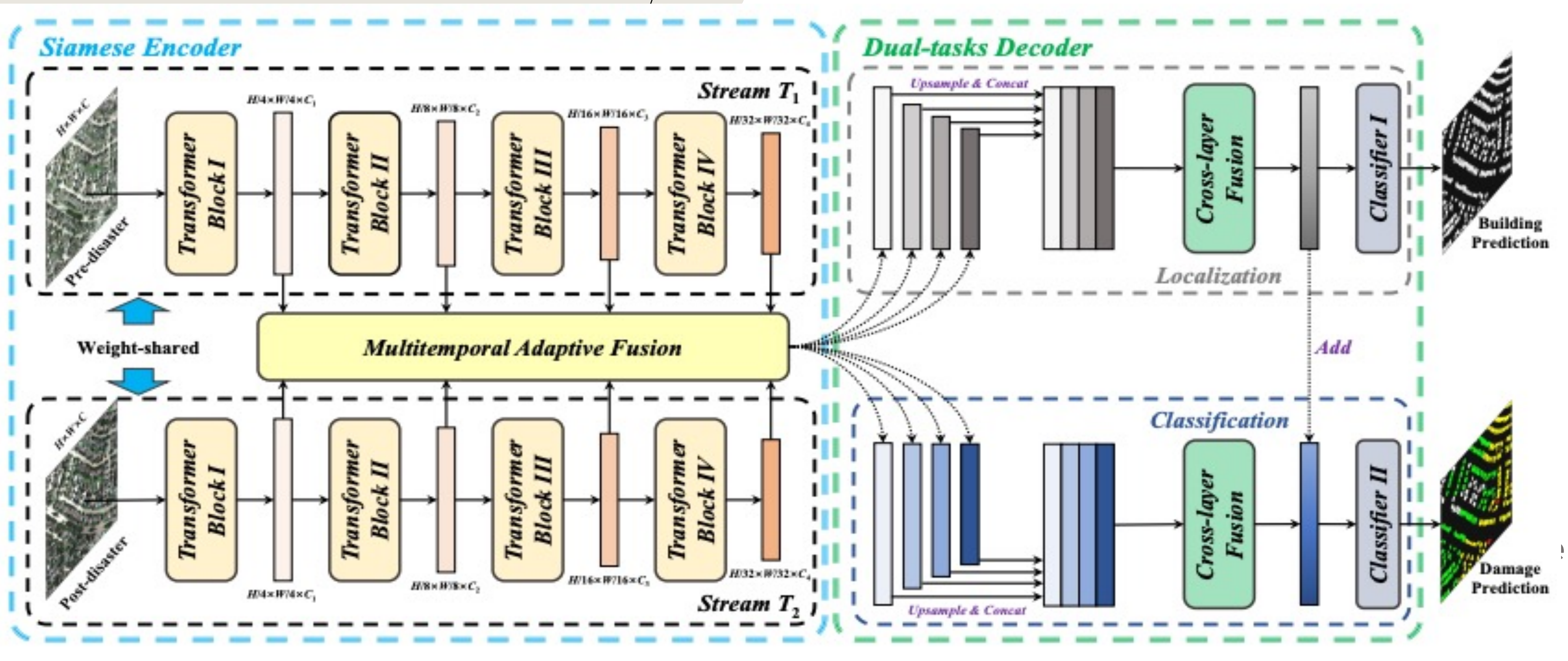


AI FOR SUSTAINABILITY

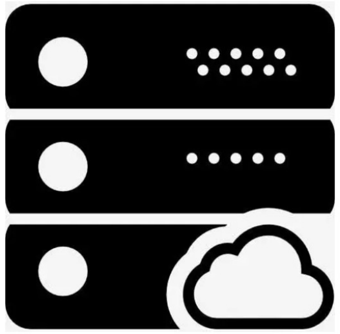
BUILDING DAMAGE DETECTION

arxiv:2201.10953

Training



Public dataset

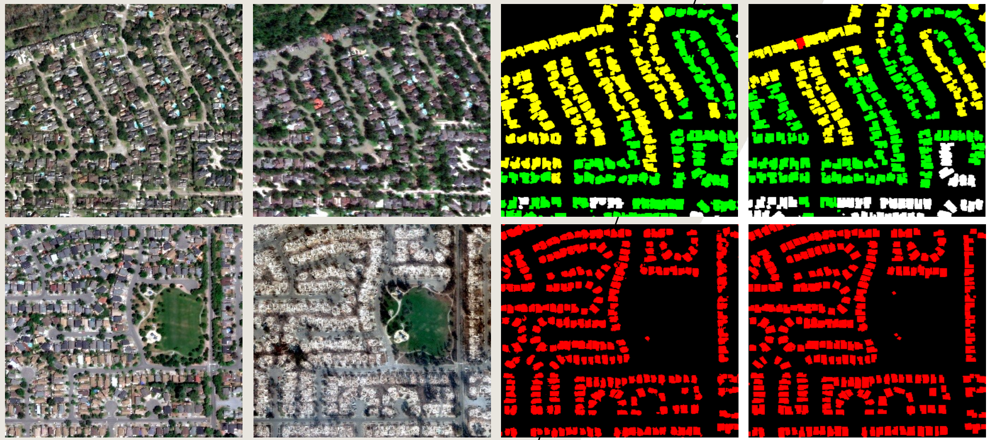


Deploying

UN Server

PERFORMANCE

xBD dataset (https://xview2.org/dataset) from **Maxar Open Data Program**



T_1 (Pre-damage) | T_2 (Post-damage) | Prediction | Reference

Method	F_1^{oa}	F_1^{loc}	F_1^{dam}	Damage F_1 per class			
				No	Minor	Major	Destroyed
xView2 Baseline	26.54	80.47	3.42	66.31	14.35	0.94	46.57
Siamese-UNet	71.68	85.92	65.58	86.74	50.02	64.43	71.68
MaskRCNN	74.10	83.60	70.02	**90.60**	49.30	72.20	**83.70**
ChangeOS	75.50	85.69	71.14	89.11	53.11	72.44	80.79
DamFormer	**77.02**	**86.86**	**72.81**	89.86	**56.78**	**72.56**	80.51

Use modified F_1 score combining localization, damage assessment scores*

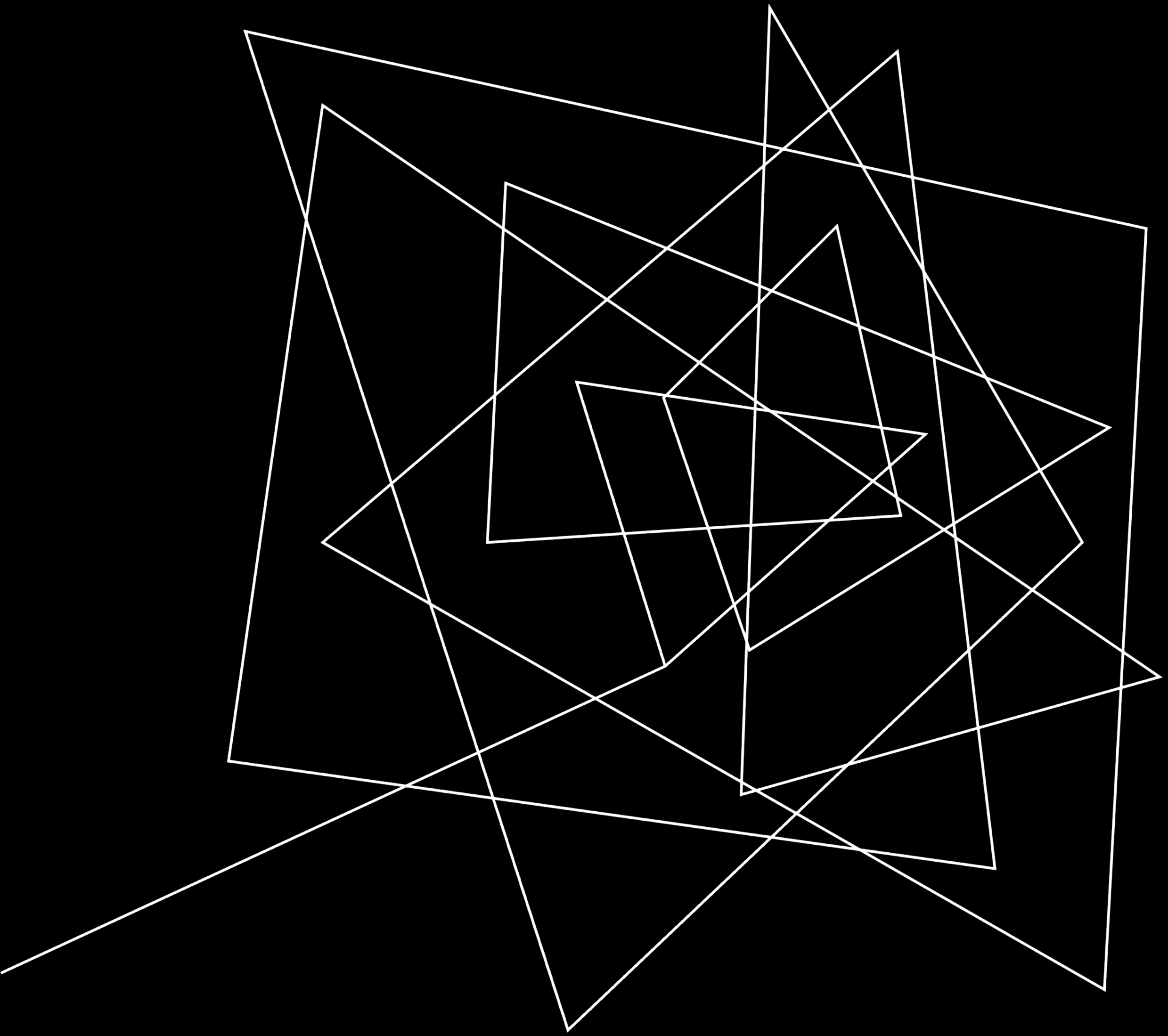
*suggested in the “CV for Building damage assessment challenge” (https://www.xview2.org/)

Transfer learning

Pre-damage

Post-damage

Natural & Man-made damages



SUSTAINABLE AI ?

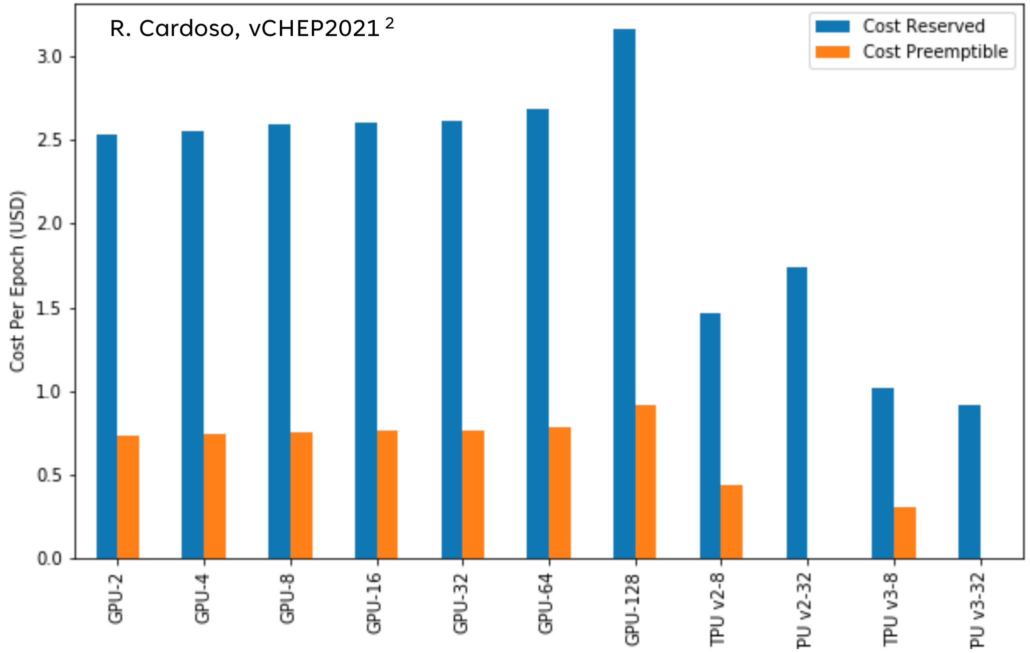
SUSTAINABLE AI ?

ML/DL inference can be more energy efficient than classical algorithms

Training Energy cost can be very high

Contribute to AI community efforts to design best practices[1]

- **Efficient ML architectures**
- **Processors and systems optimized for ML training**, versus general-purpose processors
- **Centralised computing ? (Cloud vs on prem)**
- **Efficient training strategies** (Self-supervision, few-short learning, pre-training)
- **New hardware** ? (neuromorphic, quantum)

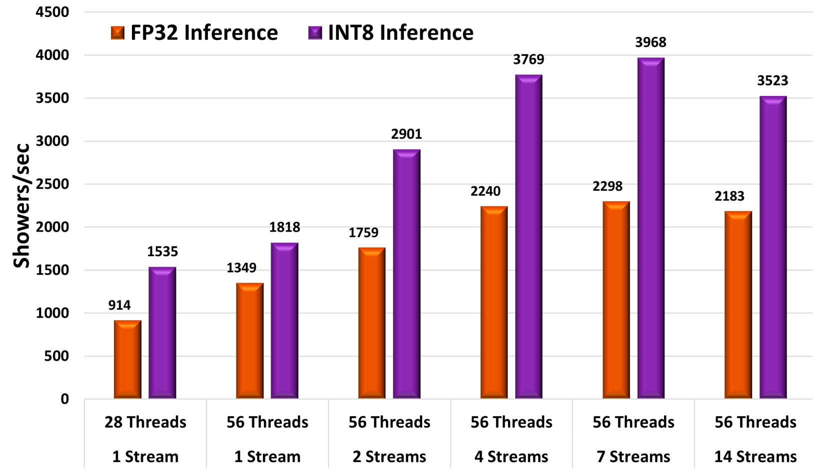


1 Patterson, David, et al. "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink." (2022).
2 Cardoso, Renato, et al. "Accelerating GAN training using highly parallel hardware on public cloud." EPJ Web of Conferences. Vol. 251. EDP Sciences, 2021.

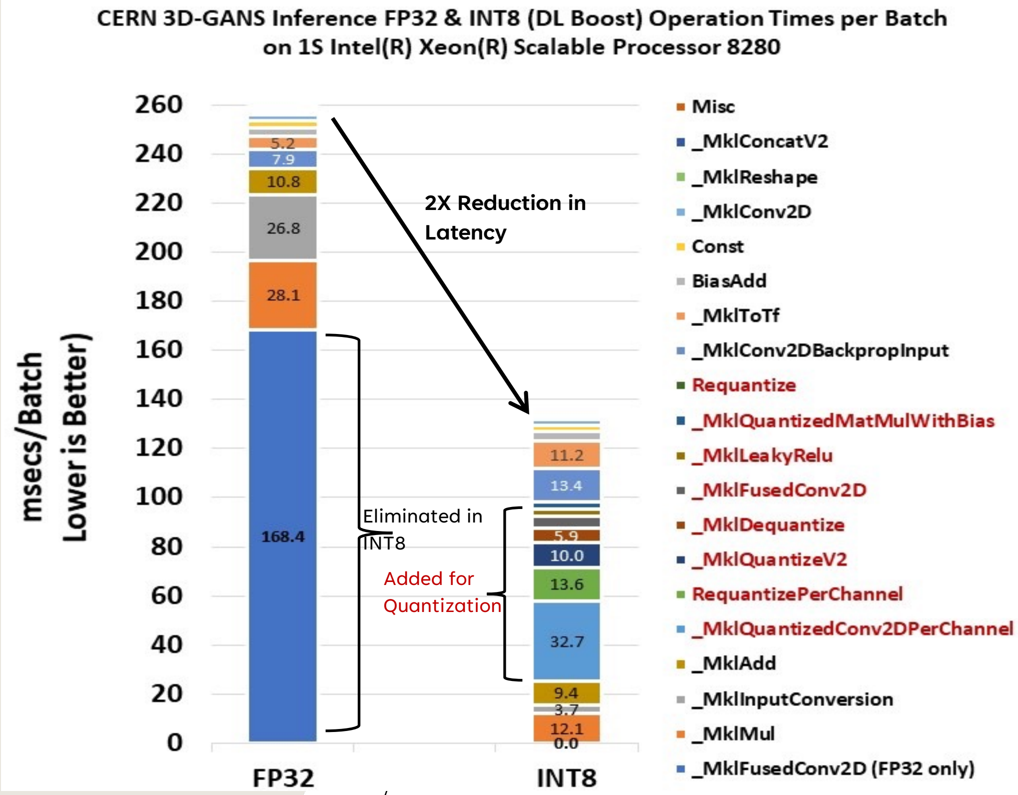
F. Rehm, ICPRAM2021
in collaboration with Intel

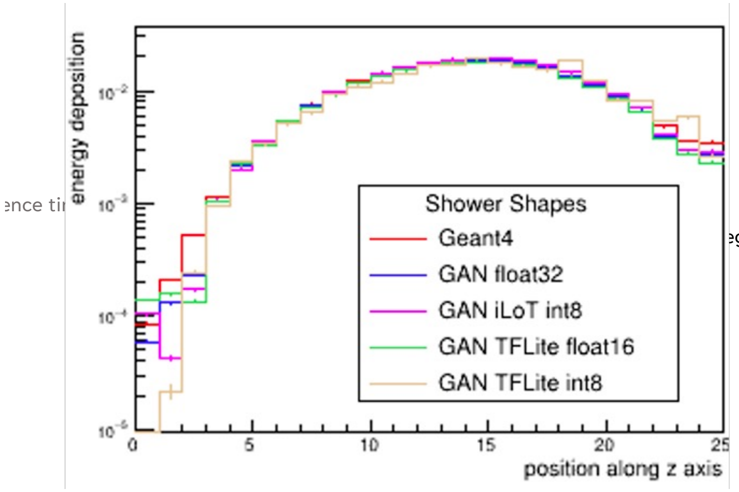
FASTER THEN MONTE CARLO?

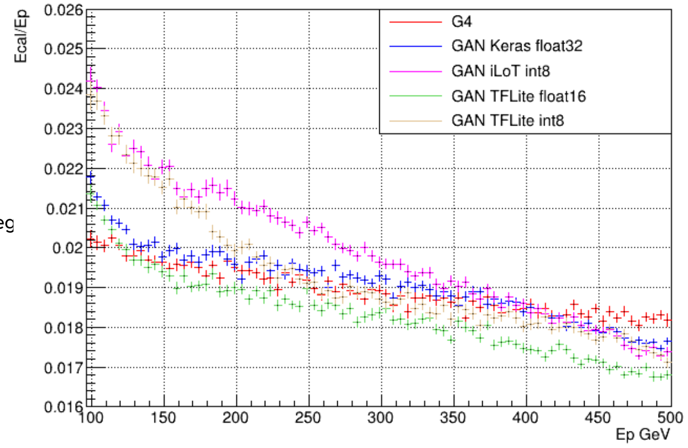
Post training quantization (INT 8) using Intel DLBoost and iLoT tool



FP32: 3DGAN is **38000x faster** than Monte Carlo
INT8: quantized 3DGAN is **68000x faster** than Monte Carlo





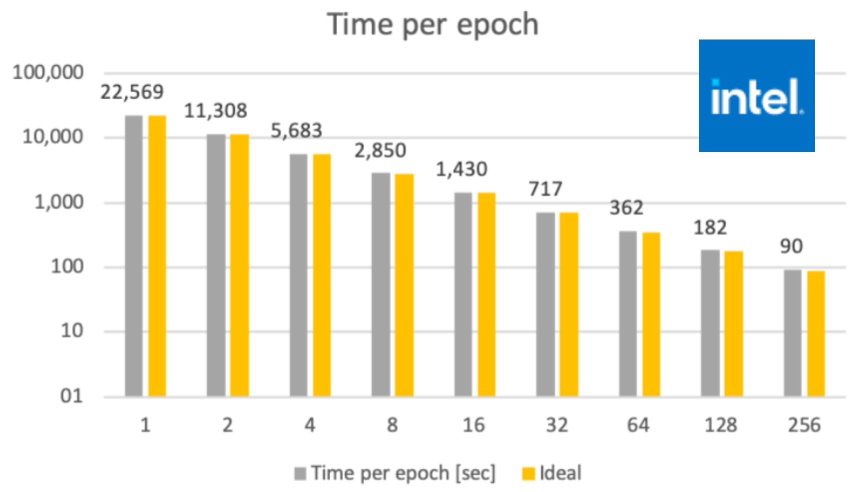


OPTIMIZED TRAINING

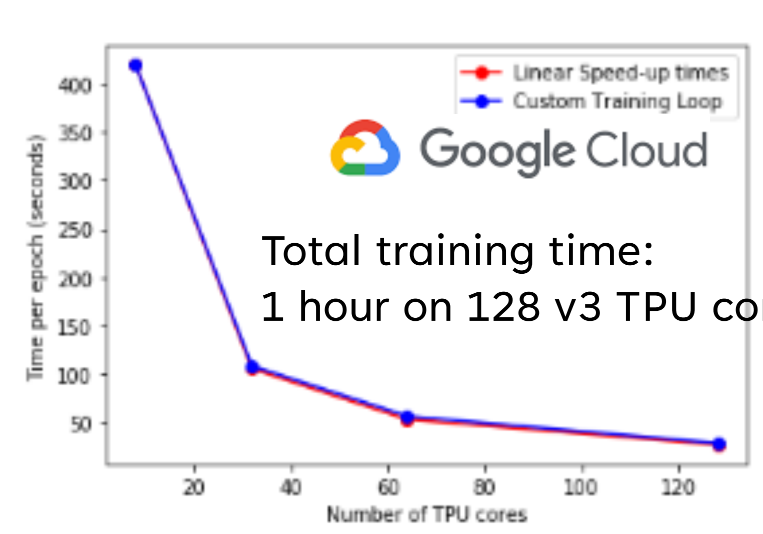
Training 3DGAN (3M parameters) takes ~7 days on a GPU

- **Distributed training** is essential
- Keep **physics under control**
- **Optimise costs**

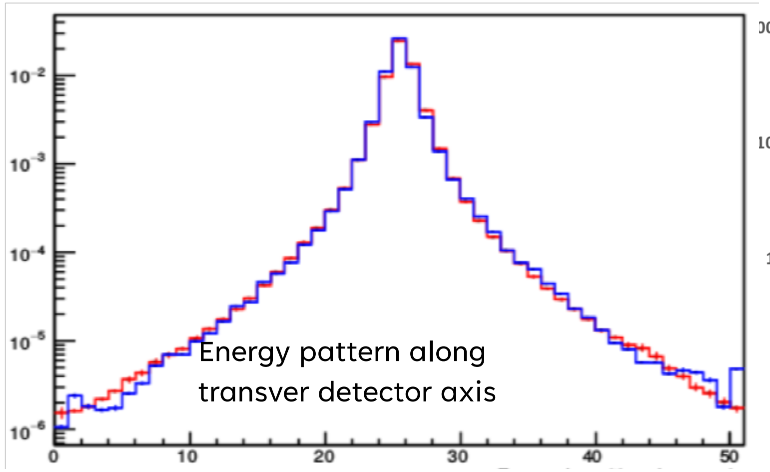
Total training time: 3 hours on 256 Intel Xeons



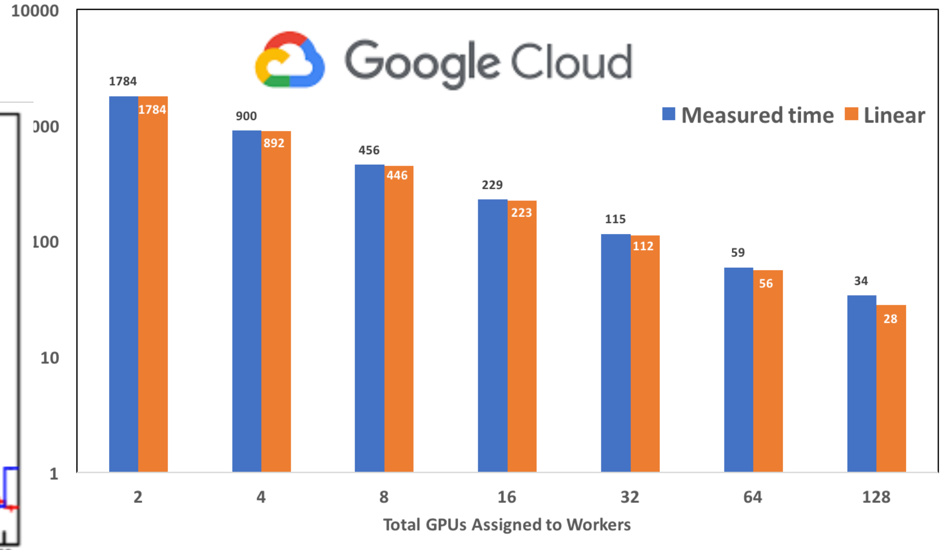
Total training time:
1 hour on 128 V100 GPUs



Total training time:
1 hour on 128 v3 TPU cores



Energy pattern along transver detector axis



THANK YOU

Sofia Vallecorsa

Sofia.Vallecorsa@cern.ch