Report: 2nd ML Infrastructure Workshop

Ricardo Rocha, Sofia Vallecorsa

https://indico.cern.ch/event/1358625/

January 26th, 2024

Introduction

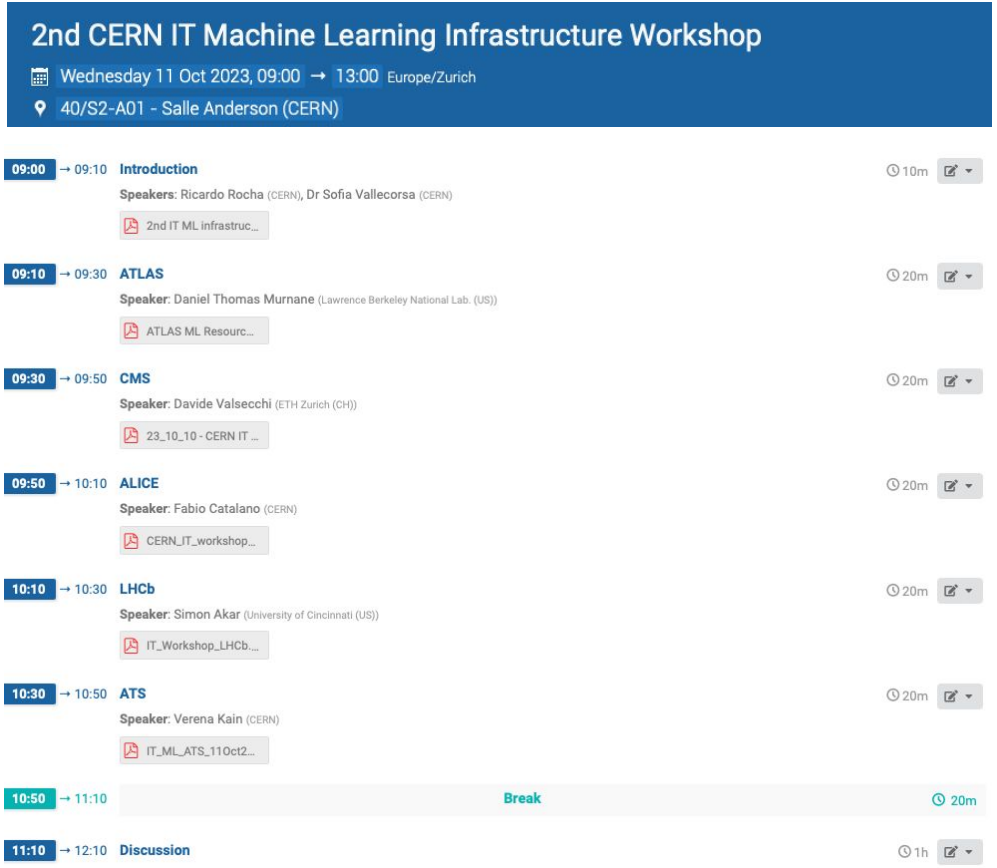
Important steps in the overall initiative to define **ML dedicated infrastructure** for CERN

Input from **experiments and ATS** to our request for information on “on-going and planned AI-research, infrastructure in use today and needs”

> **Infrastructure**: hardware, platforms, access and documentation

NB: This is official feedback, requested by IT via Engagement channels and sent by experiments according to their internal review

https://indico.cern.ch/event/1298990/



Status: AI is here to stay

ATLAS:

- Most simulation is still classical (but **Fast ML based on GAN is in production**)
- **Tagging is fully ML**, tracking classical, trigger mostly classical.
- Analysis is mostly classical or simple ML models
- **Expect 50% of ATLAS algorithms accelerated by GPU-based ML by 2030s**

ALICE:

- Multiple ML workloads with different data, training, deployment patterns
- **So far, smaller scale and simpler models than in ATLAS and CMS**

CMS:

- Multiple ML-based reconstruction **already in production**
- **Advanced use cases, highly customized**
- **Moving toward larger models (transformer - based)**
- **Extensive work** at the level of ML optimisation, frameworks (ML fully integrated in CMSSW),
- **At least 30% of CMS algorithms are ML-based today**

LHCb:

- Main use cases for online operations and trigger
- **Requirements at the analysis level are lower, given the data is simpler and luminosity lower than at ATLAS or CMS**

ATS:

- **Automation of the accelerators infrastructure** is the main scope for ML research
- In addition: **accelerator design and AI assistants (LLMs)**

Status: AI is here to stay

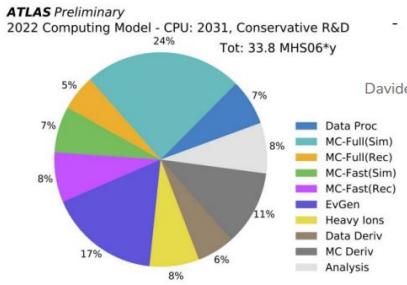
ATLAS:

- Most simulation is still classical (but **Fast ML**

Hardware: GPUs

Production Inference

- Consider that the bulk of MC simulation and reconstruction could be targeted by GPU-based ML solutions, and some part of analysis and derivations could be accelerated with GPU ML
- We could estimate that O(50%) of the ATLAS computing model could be accelerated by GPU-based ML by 2031



than in ATLAS and CMS

CMS:

1. ML Models in production - footprint

CMS uses many ML models running in production in the trigger and reconstruction sequence.
Until now all the models have been run in *direct inference* on CPU.

- **Time spent in evaluating ML models** is reaching 2% in the reconstruction sequence (RECO), and 10% in the ntuplization step (MiniAOD)

Few examples:

- **Tracking** (~1% of RECO time, TensorFlow): larger number of evaluations but with ~low complexity (DNNs)
- **Jets taggers** (~3% of MiniAOD time, ONNX):
  - Many variants of jet taggers and mass regression: AK4, AK8, AK15 jets
  - Medium size input: ~20 Jets, ~30 constituents, ~10 features x event
  - Complex operations (graphs and transformers)
- **DeepTau** (~3% of MiniAOD time, TensorFlow): complex CNN with ~large input: R&D ongoing to upgrade to graph networks
- **DeepMET** (~2% of MiniAOD time, TF): DNN with large inputs but single evaluation
- **Regression** and **classification** (XGboost, TF): very simple, BDTs or DNNs, usually needed in local reconstruction

'S:

Automation of the accelerators infrastructure is the main scope for ML research

- In addition: **accelerator design and AI assistants (LLMs)**

Constraints & Requirements

ATLAS:

- Data access to allow for **different data formats**
- Expect **data storage** to be O(10) TB
- Need **flexibility in software tooling** (conda on lxplus?)
- I**nteractive GPU access** is important for testing jobs and environment
- Need access to **inference hardware** for efficiency: custom precision, pruned models, etc..
- Explore **high speed interconnection at national labs**

ALICE (relatively light models, i.e. P.ID):

- Energy based: **simple models trained for 300 hours per year** of data taking → expect scale to a 1000 hour
- Tracking: **light model trained once** (stable detector) → 30 minutes on A6000 (INFN Torino cluster)
- Combined: **1h training time** per model on GTX 1660
- **Data format is an issue** (TMVA ONNX embedding needs further development)

CMS:

- >50% training takes **more than 1 day**
- **Can't do hyperparameter optimization for lack of resources**
- **Memory:** Use C++ compiler to load trained models in CMSSW (x5-x10 reduction)
- **Inference is limited by single event calls** → use indirect inference with batching

LHCb:

- GNN based full event interpretation: **several days of training – currently impossible at CERN**
- Need ability to: **maintain fast sim models**, retrain, do **management and inference** of large models
- Need v**ersioning tools** for models, data and hyperparameters storage for models and data (**most can be very large**)
- **Need robust flexible pipelines**

ATS:

- Main constraints coming from running on **Technical Network**
- **Missing GPUs** in K8s cluster on TN and the online data facility

Status and Feedback

Opportunistic use of resources within or outside CERN

Investigations on HLT1 cluster (LHCb), EPN cluster equipped with 8 AMD GPUs (ALICE), US National Labs, external institutes (ALICE, CMS, ATLAS, LHCb)

Interactive access is limited

“Interactive lightweight access would be most useful in Alice (SWAN like) but it is strongly limited by resource availability at the moment”

Missing infrastructure for hyperparameter scans and AI architecture search

= a problem of scale but mostly **missing interconnect among nodes**

Status and Feedback

Missing infrastructure to run efficient training of complex models

= a problem of scale but mostly **missing interconnect among nodes**

Missing infrastructure for optimised inference

Tests on **commercial clouds welcome** but concerns about costs

“(...) within ATLAS we exercised with GCP, data ingress is not a problem, we connected one of their sites in Europe within WLCG. Big benefit is the scalability and modern hardware. Drawback is that it still seems expensive compared to the compute our funding agencies provide. **We plan to release a public document to report details**” **(and dedicated high speed links exist)**

Need complete and transparent communication about costs

Wishlist

Filling the gaps highlighted to build a “robust flexible pipeline”

- A scalable interactive access for initial R&D
- Access to a multi-GPU setup with (fast) internode communication
- Specialized hardware for optimised inference

Wishlist

And more specifically:

- Tools to **track full lifecycles**
- **Versioning of models** in production, possibility to trigger re-training if needed
- **Continuous training** for Level 1 trigger
- **Resources beyond a ML training facility.** ML in custom environments (e.g. HW L1 triggers)
- **Integration of ML models in experiments software frameworks**. Further development needed for using ONNX Runtime, comparison between custom in-house and industry solutions
- Storage of **large models** and versioning of data in production
- **Energy profiling tools** (highlighted as a nice to have during the discussion)

Action items

1. **Better advertising of available tools and resources in IT (and how to use them)**

   Request for a single entrypoint documenting access to GPUs and ML capable services, with recommendations

2. **Clarify reported bottlenecks on accessing storage**

   Unclear which backends were being referred, discussion on patterns accessing from public cloud

3. **AutoML / Hyper-parameter optimization seen as a crucial aspect**

   Integrated in ml.cern.ch, request for multi-GPU support in Batch (essential for upcoming years)

   Evaluate a central ML model repository

4. **Better coordination of access to accelerator resources**

   Sharing of GPUs between services, hurdles integrating online experiment resources (ALICE EPN, LHCb HLT1)

5. **Help with profiling and benchmarking of ML workloads**

   Deployment optimization requires expertise, IT can help. Opportunity to collaborate with industry partners

6. **Tackle needs for dedicated architectures and licensed software (including drivers)**

Next steps

We have a series of requirements from the community

Address short term action items in IT, based on the input collected (previous slide)

Ongoing items in our IT ML Infrastructure initiative

- Survey state of the art and best practices beyond CERN
- Series of seminars and workshops in the coming months, to be announced
- Continue engaging with different CERN communities, dedicated initiatives already in place

The workshop happened in December, experiments are already asking us about the next steps!