Work Package 1

WP1 Intro

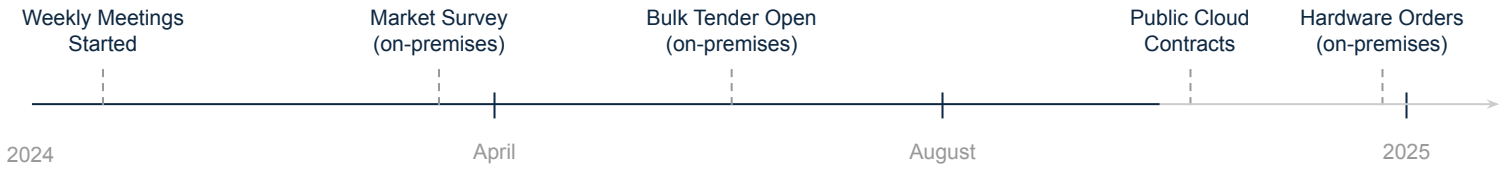
Participation of ALICE, ATLAS, CMS, LHCb, EP-SFT, IT, TH

Wide variety of topics; goal: provide common, foundational R&D, also to other WPs

- WP leads: Axel Naumann, Michelangelo Mangano
- Mailing list

1.1

Procurement & Platforms



Hardware specification, procurement and provisioning (Feb 2024, ongoing)

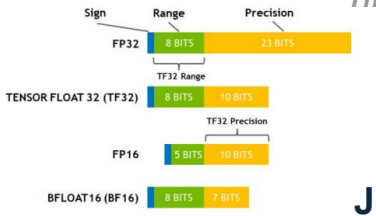
8 different hardware flavors identified (H100 NVL + SXM, L40S, MI300X + RadeonPro, HighMem, …)

Cloud and on-premises procurement, access to initial seeding resources (on premises + Oracle cloud)

Identify and collect reference use cases, hardware benchmarking (April 24, ongoing)

Initial results for A100 vs H100 vs MI300X, ongoing survey for new use cases

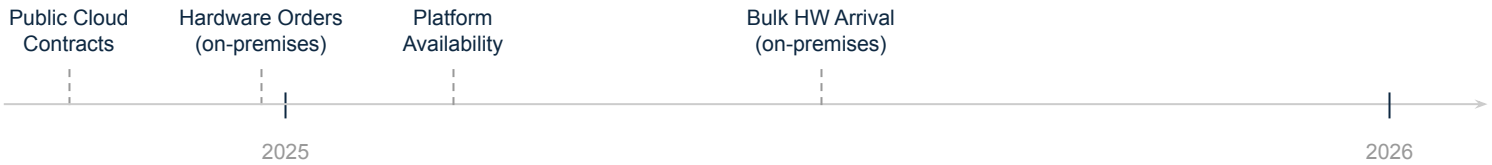
Thanks Diana Gaponcic!



Who: **Hannes Hansen, Raulian-Ionut Chiorescu, Ricardo Rocha**

Join us ! NGT #Task1.1 , ngt-wp1-task1-1@cern.ch , weekly meetings Fridays 2pm

Procurement & Platforms



Platform for efficient access to resources, MLOps (June 24, ongoing) , iterative process

Efficient GPU usage and sharing, integration with external resources, entry-points (ssh, notebooks, …)

Support for pipelines, distributed training, hyper-parameter tuning, model serving and management

Provisioning of on-premises hardware (Mid 25)

Benchmarking and end to end validation of use cases (2025)

Across all available hardware (on-premises and external)

Who: **Hannes Hansen, Raulian-Ionut Chiorescu, Ricardo Rocha**

Join us ! NGT #Task1.1 , ngt-wp1-task1-1@cern.ch , weekly meetings Fridays 2pm

1.2

Fast inference of NNs on FPGAs*

ngt-wp1-t2
NGT/Task 1.2
FastML/ngt-wp12

*but not limited to!

hls4ml - a source-to-source compiler creating HW-optimal HLS designs for FPGAs (docs | git)

- ~~Widely,~~ “widely” and wildly used in CERN + large community of users @ FastML
- Currently supporting common NN architectures - MLP, CNN, RNN

What about custom NNs, emerging architectures and HW-efficient NNs?

- Main focus of Task 1.2 throughout the 5-year project
- Creating next generation HW-aware inference engine
- Start by creating a more expressive code-generation core (2025)
  - Successful prototype made in the past
  - Output optimized inference kernels for target HW platform

Team: Sebastian Dittmeier, Maurizio Pierini, Vladimir Loncar, + QUEST [to start Oct. 1] + Student [TBD]

But how do I train HW-efficient NNs? Glad you asked… ➜

1.3

Hardware-aware AI optimization tools

ngt-wp1-t3
NGT/Task 1.3
FastML/ngt-wp13

Numerous ways to make efficient models

- Pruning, quantization, distillation, low-rank approximation, symbolic regression, neural-architecture search…

How do we know which method to use for a given task+hardware constraints?

- We aim to develop a collection of proven methods and set of recommendations for end-users not requiring extensive knowledge of efficient ML and HW specifics

Started prototyping a common framework

- Also looking into existing solutions
- Started work on evaluating (un)structured pruning methods for efficient implementations on FPGAs
- Goal for 2025: Create an end-to-end loop for training efficient NNs for deployment on hardware triggers (e.g., ATLAS L0 and CMS L1)

Team: Roope Niemi, Michael Kagan, Maurizio Pierini, Vladimir Loncar + Student [TBD]

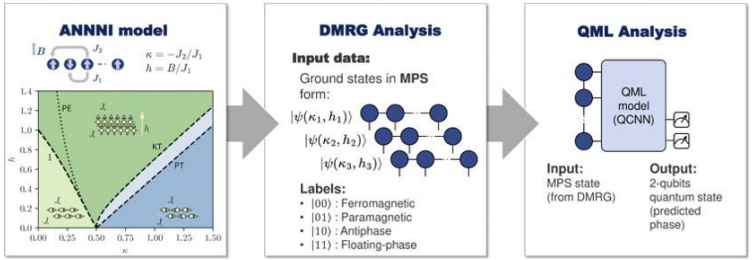
1.4

Tensor Networks for Quantum Systems

This task will develop and apply quantum-inspired methodology, in particular Tensor Network algorithms, to simulate quantum many-body problems unreachable by classic approaches and benchmark future applications of quantum hardware on low-entangled systems to O(100) qubits, progressing towards the development of a software stack for quantum machine learning model design, simulation, and deployment.

- *First results (preprints) on Tensor Networks and Quantum Machine Learning analysis with O(100) sites.*
- *Ongoing projects on real-time dynamics of High-Energy Physics*
- *Collaboration with CERN-TH and CERN-QTI*

Summary of the workflow



First NGT workshop on Tensor Networks and Quantum Machine Learning: https://indico.cern.ch/event/1455226/
The topics covered will include:

- *Exploring the use of (Quantum) Machine Learning algorithms within tensor network wavefunctions.*
- *Analysing the application of GPU technology for tensor network simulations in high-energy physics.*
- *Investigating new strategies for enhancing quantum machine learning using tensor networks.*

Team: Enrique Rico Ortega (NGT LD, TH), **Stefano Carrazza** (NGT SASS, TH), **Sofia Vallecorsa** (IT - QTI)

1.5

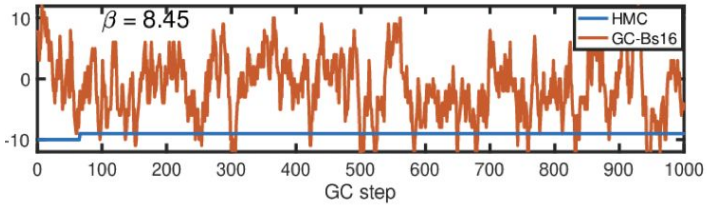
New computing strategies for data modeling: LQFT

Code modernization on parallel architectures and utilising AI: aligned with WP2/3

- Development of SW and algos to best exploit next-gen architectures in LQFT simulations on extreme-scaling low-latency/high-bandwidth accelerator-based clusters

Milestones (M12)

- benchmarking support with LQFT codes guiding hardware procurement and commissioning for HPC hardware (see task 1.1)



Deliverables (M12):

- Organization of several community Workshops
  - First workshop “NGT - Algorithms for lattice QCD” scheduled for 9-11 December
  - https://indico.cern.ch/e/NGT_algorithms_for_latticeQCD_Dec24
  - Variance reduction (day 1), novel update for MCMC simulations (day 2) and adaptation to novel hardware (day 3) … Day 1-2 LQFT focused, day 3 of interest to NGT at large
- Develop LQFT benchmarking software tailored to hardware infrastructure procured under 1.1.
  - Preliminary numbers from A100 and H100 collected
  - On-going, Early Access to Alps, a GH200 cluster at CSCS, obtained and benchmark numbers collected, currently steps to obtain early access for Jupiter (JSC)
- Share expertise on parallelism and accelerator-based algorithms with TH/IT/CMS/ATLAS

Team: Jacob Finkenrath (NGT LD in TH), **Andreas Juettner** (TH)

Indico: https://indico.cern.ch/category/18029/

New computing strategies for data modeling: MC evt gen acceleration

Adapt computing strategies and algorithms of MC event generators to new hardware infrastructure:

- Meet the HL-LHC event-generation computing needs
- Fully exploit global HPC facilities and resources

Activity will build on existing first results and on community-wide efforts, reviewed at kick-off workshop on MC acceleration. Example:

Process	Matrix elm type	Total
$gg \to t\bar{t}gg$	Fortran	116.35(3)s
	C++ AVX2	29.92(6)s 3.89(1)×
	CUDA Tesla A100	~~7.88(2)s~~ 14.77(4)×
$gg \to t\bar{t}ggg$	Fortran	2233.1(6) s
	C++ AVX2	687.5(9) s 3.25(1)×
	CUDA Tesla A100	~~27.57(2)s~~ 81.0(1) ×

Examples for GPU speedup over single-threaded CPU execution

Next steps:

- Develop and adapt GPU implementations for more parts of the MC ev , phase-space sampling, NLO, etc
- Work on sustaining the scalability of the applications in a highly parallel environment
  - Improve I/O via binary file formats such as HDF5
  - Increase the "density" by combining multiple calculations in one GPU kernel
- Extend and make the work usable for more event generator packages

Team: **Daniele Massaro** (NGT Quest, IT), **Stefan Roiser** (IT), **Enrico Bothmann** (IT), **Zenny Wettersten** (IT)

Indico: https://indico.cern.ch/category/18029/

1.6

New physics scenarios and SM properties as trigger benchmarks

Identify BSM scenarios and signatures to be used as benchmarks for the assessment of new-generation triggers performance, aligned with WP2/3, in close collaboration with the experiments

Work plan

- Hire of dedicated fellow and formal start of work in 2025
- Preliminary discussions led by **Joe Davighi** and **Matthew McCullough** (TH) with ATLAS and CMS reps
- Target models well-motivated for BSM (not just “simplified models” for the signature), in particular related to
  - Dark Matter
  - Flavour puzzle
  - Neutrino masses

First ideas (see also https://www.overleaf.com/read/zfntggvczkvn#064769)**:**

- LLP derived leptons
  - >1 displaced lepton; same or different DVs; prompt + displaced leptons; late leptons
  - Natural in dark sectors e.g. $A \to l^+ l^-$ (same DV) from dark mediator with $m_A \approx 2m_l$
  - Charged leptonic LLPs in neutrino models?
- Slow Stuff
  - Ultra-slow particles; out-of-time decays with a peaked (in time) signal
  - Again natural in dark sectors; e.g. s-channel $pp \to A \to \phi\phi$, mass $m_A \approx 2m_\phi$
  - Time-dependent interactions (set by cosmologically varying BSM scalar field)
- High Multiplicity, Low-Energy, Heavy-Flavour
  - Lots (e.g. n > 5) of soft b-jets; lots (n > 2) of soft taus
  - Simple benchmark flavour models predict light, flavour non-universal $Z' \to bb, \tau\tau$
  - Odd # of τs ~ LFV & ν masses?

E-group: ngt-1-6-coord@cern.ch

Indico: https://indico.cern.ch/category/18030/

1.7

Common SW dev for heterogeneous architectures

“Everything heterogenous / accelerated that’s not ML”:

- Efficient scheduling across CPU / GPU
- Efficient data structures for heterogeneous software
- Accelerated HEP standard library
- Efficient ML inference interfaces / kernels
- Novel programming languages

Team: Andrea Bocci* (CMS), Jolly Chen (EP), Marco Clemencic (LHCb), Attila Krasznahorkay* (ATLAS), Daniele Massaro (IT), Lorenzo Moneta (EP), Axel Naumann* (EP), Felice Pantaleo (CMS), Oliver Rietmann (ALICE+ATLAS), David Rohr (ALICE), Stefan Roiser (IT), Arkadijs Slobodkins (ATLAS+CMS)

*: task leads

✉ Mailing list Mattermost Indico

1.7: What we were up to in 2024

- Test-drove Julia for CMS reco code (CPU-only). Fast & easy; tooling? Talk
- Using C++ reflection proposals P2996 and P3294 prototype implemented in EDG to transform AoS ⇔ SoA (⇔ AoSoA); talk (to be repeated in 1.7) code
- Implemented demo for scheduling code
- Input to hardware specification which is rather different from ML / Lattice QCD

…but most people started too late to share progress already now! Ongoing:

- Porting MadGraph on GPU; implies profiling and testing designs and optimizations
- Investigating SoA for ALICE's O2 code

1.7: Plans for 2025

Super-exciting yet very diverse:

- Collect first accelerated library; benchmark; also look at “weird” architectures (thanks, Openlab!)
- Investigate at least Mojo, next to Julia; design case for testing accelerator use
- Test drive coroutines for scheduling
- Propose first abstraction kernel and memory layout for ML inference
- Investigate options for (ex-) member functions with AoS ⇔ SoA
- Implement EDM + reco prototypes for reflection and scheduling
- Benchmark O2 with SoA, design such that reflection-based implementation can be used
- Possibly look at other generators to be ported to GPU

If you have experience or opinions:

Please come to our meetings! Or follow us + visit when you care!
✉ Mailing list Mattermost Indico


CERN
NexTGen
Next Generation Triggers