Task 1.2: Development framework towards fast inference of complex network architectures on LHC online systems

Maurizio Pierini (CERN)

Scope of the task

Task 1.2: work with existing expertise in the experiment collaboration on ongoing work on tools such as hls4ml to develop ML->FPGA model synthesis tools, addressing the needs of WP2 and WP3. The work will also focus on integrating modern ML tooling while maintaining the strict latency requirements set forth by LHC experiments' online selection system. ~~All task items are supposed to be co-developed by CERN researchers and external partners with qualified expertise on the topic.~~

Task 1.3: ~~leveraging external expertise on network architecture search (NAS) from selected academic and industrial partners to~~ develop the software infrastructure needed to enable hardware-aware neural network training workflows. This work will enable the development and deployment of hardware-optimal AI-based real-time algorithms at CERN, as described in WP2 and WP3. ~~All task items are supposed to be co-developed by CERN researchers and external partners with qualified expertise on the topic.~~

How to read this:

- The industrial partner is gone -> we are on our own
- "Such as hls4ml" means that we plan to work also with Conifer (for BDTs, but same spirit)

Milestones for Task 1.2

Year	Code	Milestone	Type
1	M1.1.2	MLonFPGA community workshop, to identify needs from ATLAS and CMS on WP 2 and 3 as inputs to WP 1.2 and 1.3	Event, report
2	M2.1.2	hls4ml software release 1 with open-access documentation	Software
4	M4.1.2	hls4ml and NNLO software release 2 with open-access documentation	Software
5	M5.1.2	hls4ml and NNLO software release 3 with open-access documentation	Software

Possible overlap with ATLAS milestones: ML development toolkit for AI algorithms on FPGA at fixed latency for the ATLAS global trigger

hls4ml

Tool developed to port NNs to FPGA for ultra-fast execution

- Started around 2018
- Developed as a standalone library, on top of various HLS backends
- Demonstrated workflow on various architectures
- Exploited network compression techniques after training and at training time (see Task 1.3)

Interfaced to various training packages

- Mostly TF, but also Pytorch
- ONNX
- Quantized networks via QKeras and QONNX

Work developed in the context of FasML

Conifer

Twin project to port BDTs to FPGA for ultra-fast execution

- Standalone library with full support for multiclass classifiers and regressions
- Work needed to extend its interface to various training packages

FastML

The work was imagined in the context of FasML Lab

- Created in 2018, together with hls4ml
- Gathers community interested in edge computing for Machine Learning
  - (A) Hls4ml & Conifer for ultra-low latency
  - (B) Distributed inference software (e.g., SONIC) for inference in heterogeneous computing
  - (C) Projects with industrial partners, e.g., QONNX with Xilinx
- Weekly meetings on Friday at 17.00
  - Topic rotates between (A), (B), and general
- Yearly workshop collecting inputs from HEP community and beyond
- US participants (MIT, FNAL, UCSD, etc.) received big grant to work on this topic (across HEP, Neuroscience, etc.)
  - They are working already on many topics relevant for these tasks
- Our work should happen within that context, serving the whole community

Developments in experiments

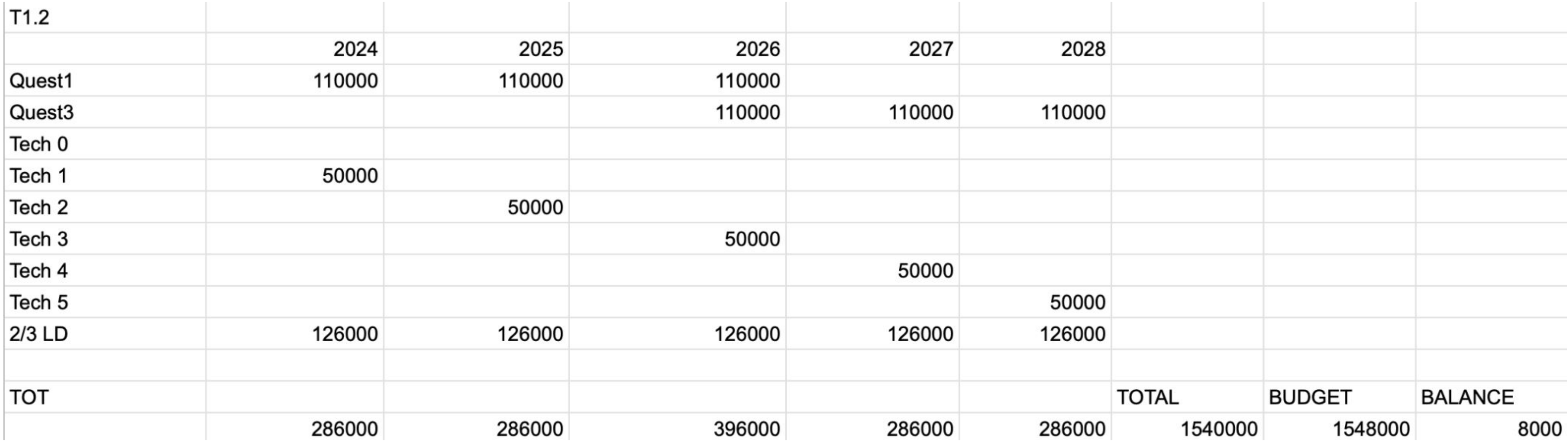
To our knowledge

- CMS is operating already a few triggers using hls4ml
- NA62 developed one trigger at some point
- CMS and ATLAS are looking at hls4ml for their upgrades
  - At least in CMS there are more than 10 projects based on hls4ml
- Community of developers across experimental collaborations and outside LHC
  - DUNE
  - NA62
  - Industry
- Development beyond trigger
  - E.g., people looking at ASICs

Budget and spending plans

We have ~ 1.5M CHF for personnel

- 70% of LD position, shared with Task 1.3
- Two 3-year quests
- 5x12 months technical students



T1.2								
	2024	2025	2026	2027	2028			
Quest1	110000	110000	110000					
Quest3			110000	110000	110000			
Tech 0								
Tech 1	50000							
Tech 2		50000						
Tech 3			50000					
Tech 4				50000				
Tech 5					50000			
2/3 LD	126000	126000	126000	126000	126000			
TOT						TOTAL	BUDGET	BALANCE
	286000	286000	396000	286000	286000	1540000	1548000	8000

Task 1.3: Hardware-aware AI optimization

Maurizio Pierini (CERN)

Scope of the task

Task 1.2: work with existing expertise in the experiment collaboration on ongoing work on tools such as hls4ml to develop ML->FPGA model synthesis tools, addressing the needs of WP2 and WP3. The work will also focus on integrating modern ML tooling while maintaining the strict latency requirements set forth by LHC experiments' online selection system. ~~All task items are supposed to be co-developed by CERN researchers and external partners with qualified expertise on the topic.~~

Task 1.3: ~~leveraging external expertise on network architecture search (NAS) from selected academic and industrial partners to~~ develop the software infrastructure needed to enable hardware-aware neural network training workflows. This work will enable the development and deployment of hardware-optimal AI-based real-time algorithms at CERN, as described in WP2 and WP3. ~~All task items are supposed to be co-developed by CERN researchers and external partners with qualified expertise on the topic.~~

How to read this:

- The industrial partner is gone -> we are on our own
- "Such as hls4ml" means that we plan to work also with Conifer (for BDTs, but same spirit)

Milestones for Task 1.3

Year	Code	Milestone	Type
1	M1.1.2	MLonFPGA community workshop, to identify needs from ATLAS and CMS on WP 2 and 3 as inputs to WP 1.2 and 1.3	Event, report
3	M3.1.2	NNLO software release 1 with open-access documentation	Software
4	M4.1.2	hls4ml and NNLO software release 2 with open-access documentation	Software
5	M5.1.2	hls4ml and NNLO software release 3 with open-access documentation	Software

Possible overlap with ATLAS milestones: ML development toolkit for AI algorithms on FPGA at fixed latency for the ATLAS global trigger

Work plans

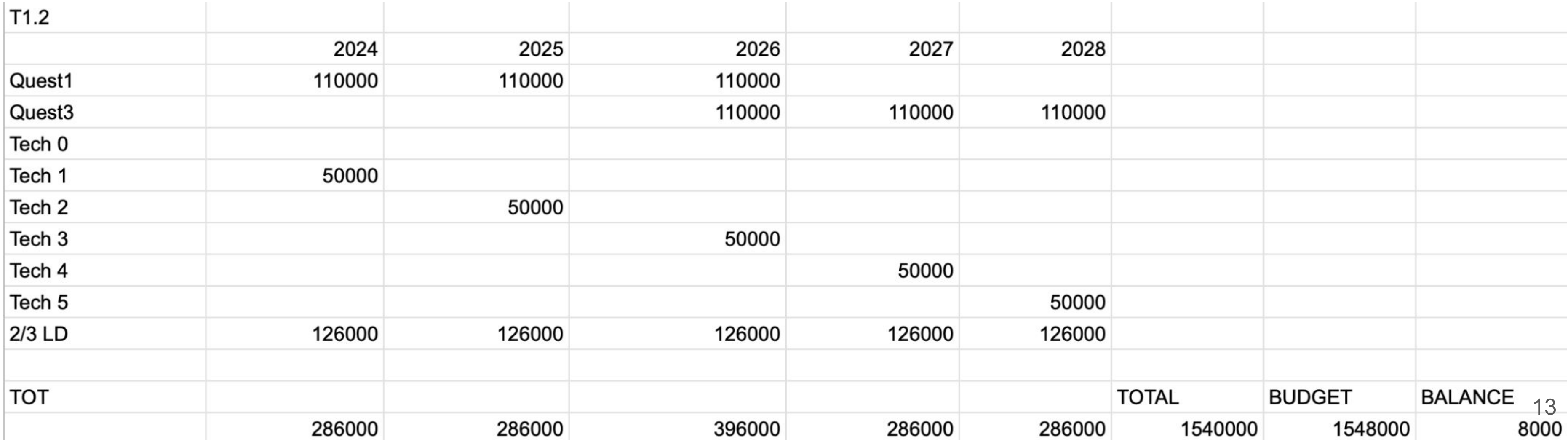
We plan to investigate compression techniques

- We start on work already done on quantization and pruning at training time
- We plan to investigate heterogenous quantization
  - Implement FIT algorithm, developed in KT project
  - Look at Knowledge Distillation
  - Integrate Semantic Regression
- Investigate architecture-specific pruning strategies
  - And their implementation on specific hardware (xTalk to 1.2)
- Integrate this workflow in an end-to-end training and optimization pipeline of existing codes (NNLO project in KT)
- Study ways to model resource usage on specific hardware
  - Integrate the model in the loss function at training time?

Budget and spending plans

We have ~ 850K CHF for personnel

- 30% of LD position, shared with Task 1.3
- Two 2-year quests
- 2x 12-months + 1 9-months technical students



T1.2								
	2024	2025	2026	2027	2028			
Quest1	110000	110000	110000					
Quest3			110000	110000	110000			
Tech 0								
Tech 1	50000							
Tech 2		50000						
Tech 3			50000					
Tech 4				50000				
Tech 5					50000			
2/3 LD	126000	126000	126000	126000	126000			
TOT						TOTAL	BUDGET	BALANCE
	286000	286000	396000	286000	286000	1540000	1548000	8000

The full team

The idea is to have a strong cross talk across 1.2 and 1.3

The to-be-hired LD will take the lead

Each task will have two deputies

- MP and Sebastian Dittmeier for Task 1.2
- MP and Michael Kagan for Task 1.3

And contacts

- Gloria Corti for LHCb
- David Rohr for ALICE
- Lorenzo Moneta for SFT
- Vladimir Loncar for FastML

Contacts

- e-group / mailing list for your task
    - ngt-wp1-task2@cern.ch
    - ngt-wp1-task3@cern.ch

- Indico category for your meetings (*):
    - Task 1.2: https://indico.cern.ch/category/17884/
    - Task 1.3: https://indico.cern.ch/category/17885/
- We have two channels (Task 1.2 and Task 1.3) under NGT Mattermost
- We will also use the FastML slack channel. It depends on how many non-NGT people from FastML will join the project

- email addresses of task leader and deputies:
    - maurizio.pierini@cern.ch
    - sebastian.dittmeier@cern.ch
    - michael.aaron.kagan@cern.ch

(*) so far meeting agendas are close. We would love to open it to all NGT participants, if we had a fully inclusive e-group