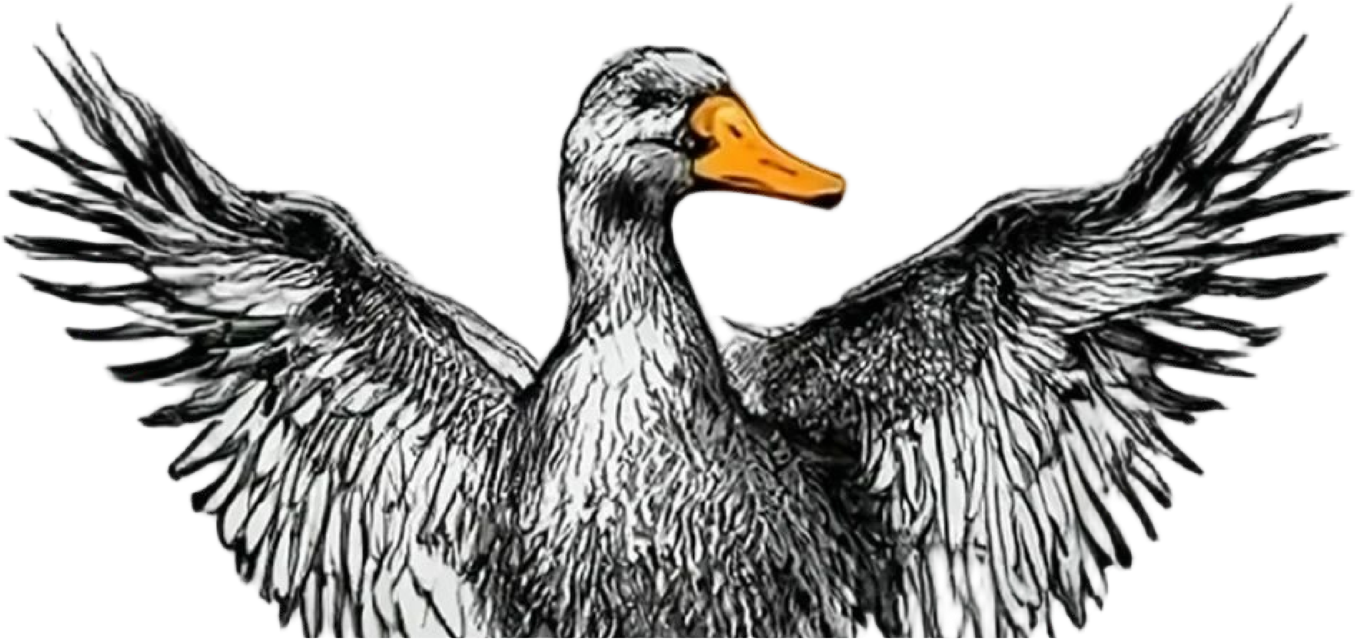


RECLAIM-DAQ

A framework for reclaiming the DAQ for computing

Carlos Abellan, Iaroslava Bezshyiko,
Maria Pilar Peco, Nicola Serra

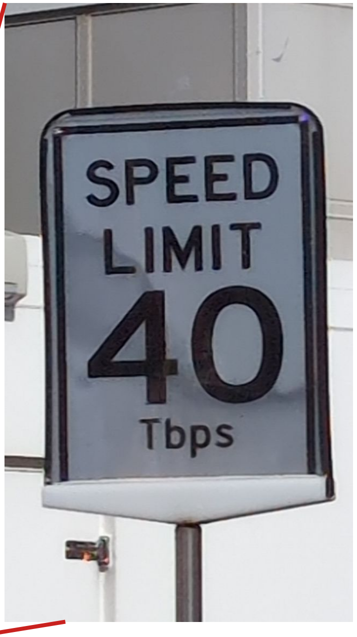
Universität
Zürich UZH

HAMLET
How to Apply Machine Learning to
Experimental & Theoretical
PHYSICS

August 19 - 21, 2024
Copenhagen, Denmark

What is a DAQ cluster?

- Large high energy physics (and related!) experiments usually produce huge amounts of data.



What is a DAQ cluster?

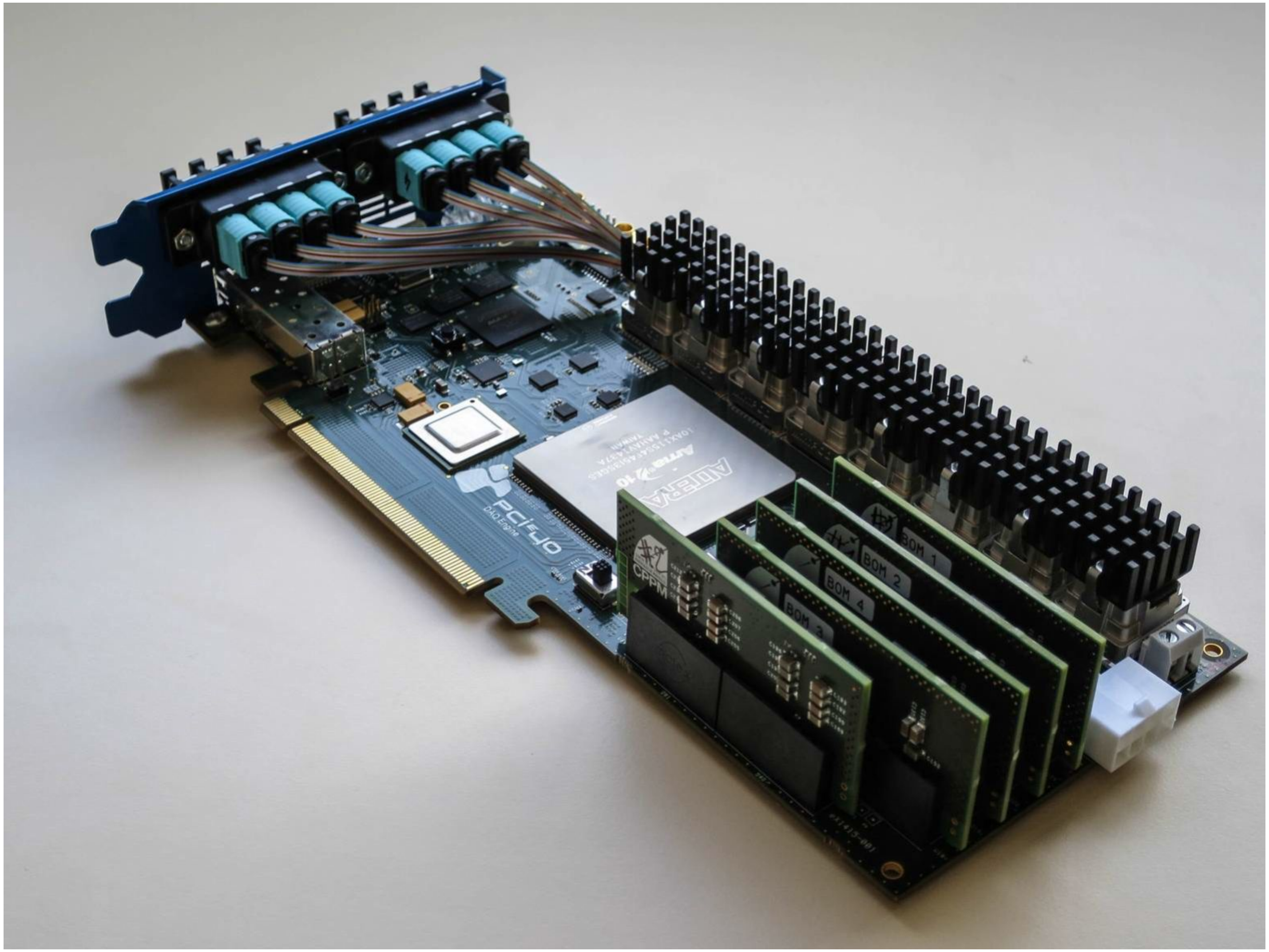
- A very powerful cluster is used: high performance servers, GPUs and FPGAs.

How performant is a DAQ cluster?

- The example of LHCb:
  - 184 SuperMicro Servers (20 for testing)
    - Each server features:
    - 2x AMD Epic 7502 CPUs (32 cores each)
    - 512 GB of RAM
    - 3 NVIDIA A5000 Ampere
    - 2 Infiniband ConnectX-5 network cards (2x100Gbps)
    - 3 TELL40 cards
  - We have 11.776 CPU cores, 555 GPUs and 555 FPGAs in this cluster
  - Several million € in computing power!

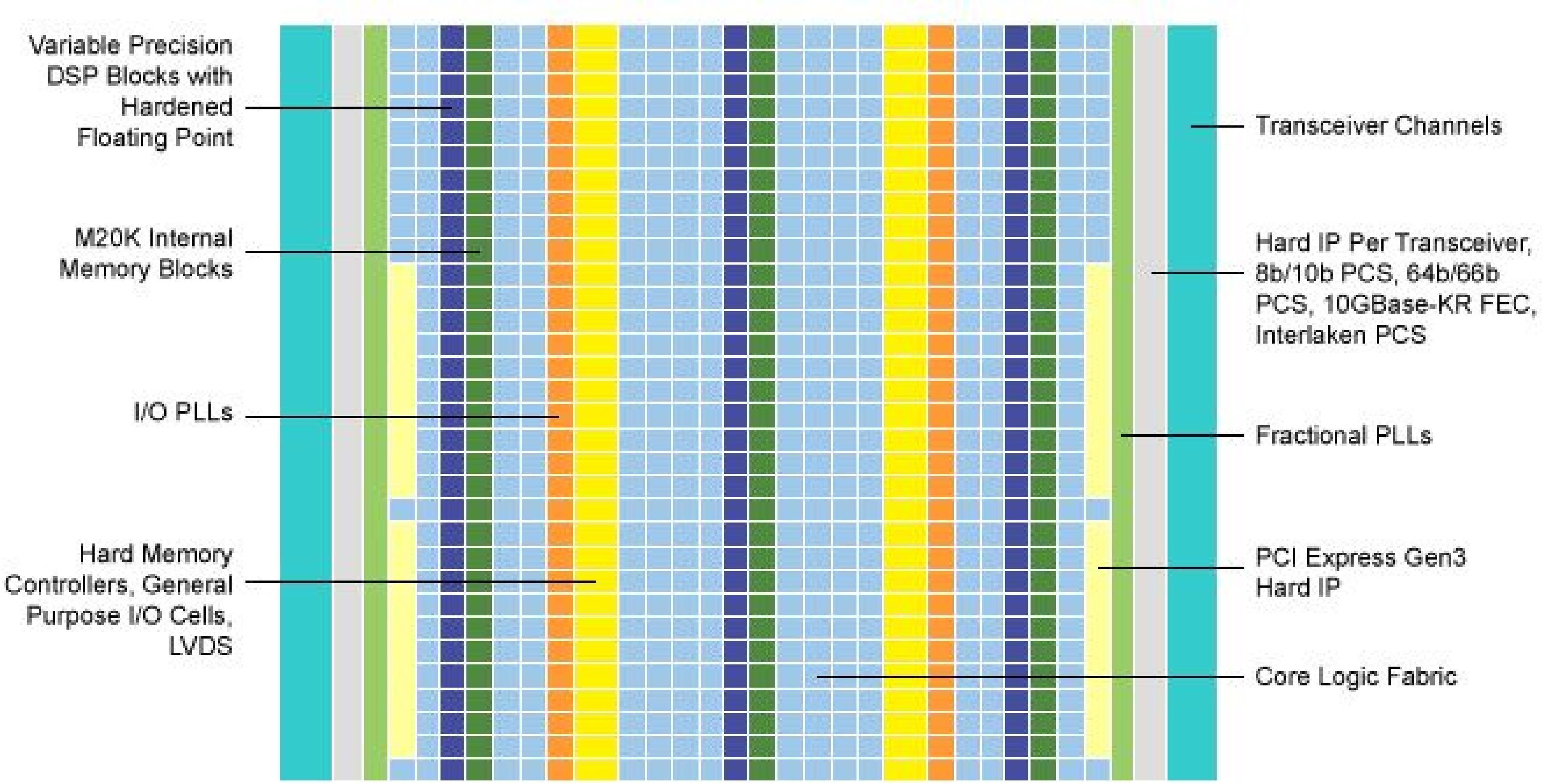
TELL40 cards?

- TELL40 cards are designed for DAQ applications:
  - For our purposes, it is a card with a large FPGA:
    - Intel Arria 10: 10AX115S3F45E2SG



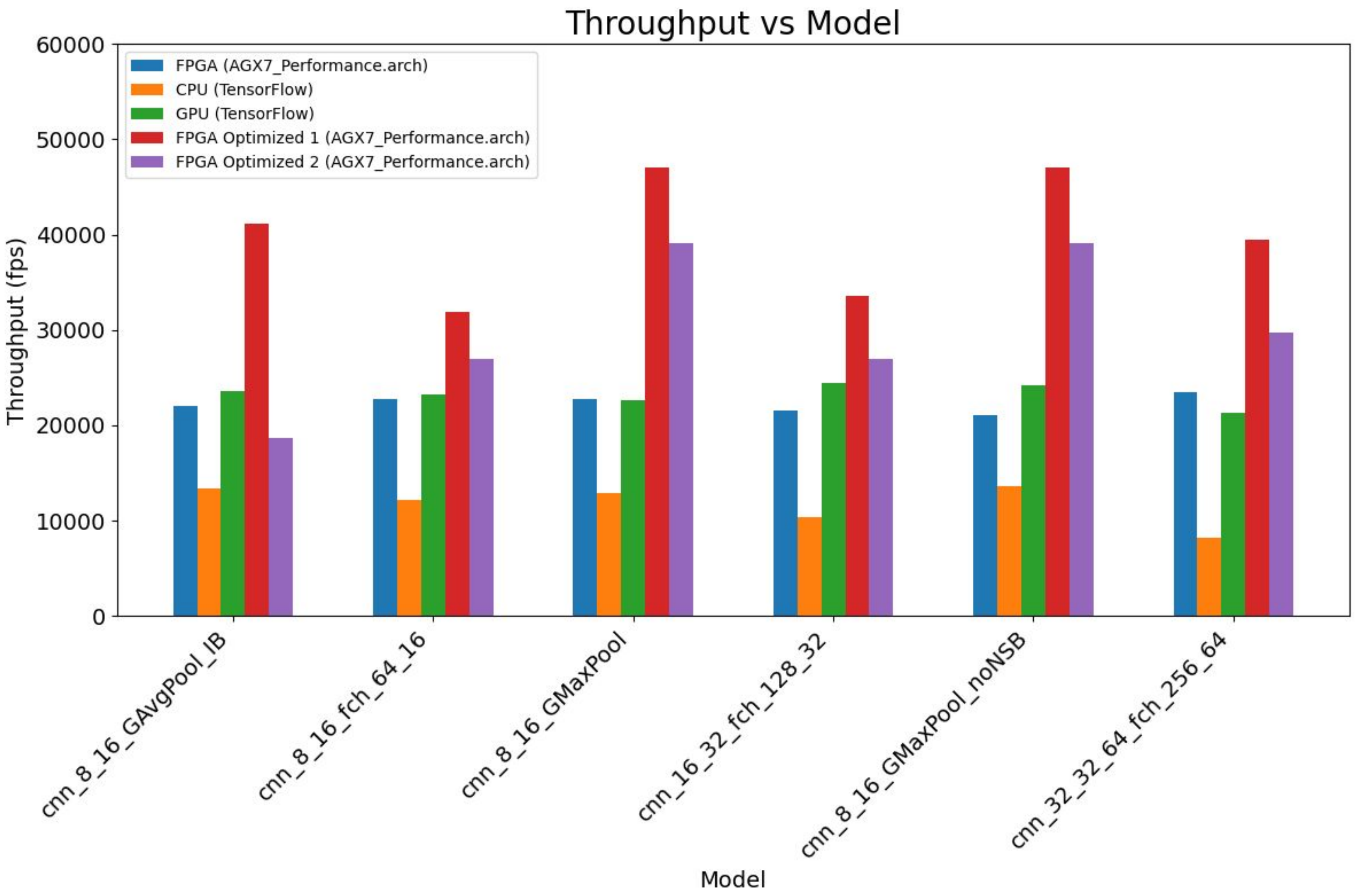
FPGA ?

- Field Programmable Gate Array:
  - A chip that contains generic digital electronics that can be configured externally: if you can design it, and it fits, the chip can do it. We use it to capture Gbps of data...


Variable Precision DSP Blocks with Hardened Floating Point
M20K Internal Memory Blocks
I/O PLLs
Hard Memory Controllers, General Purpose I/O Cells, LVDS
Transceiver Channels
Hard IP Per Transceiver, 8b/10b PCS, 64b/66b PCS, 10GBase-KR FEC, Interlaken PCS
Fractional PLLs
PCI Express Gen3 Hard IP
Core Logic Fabric

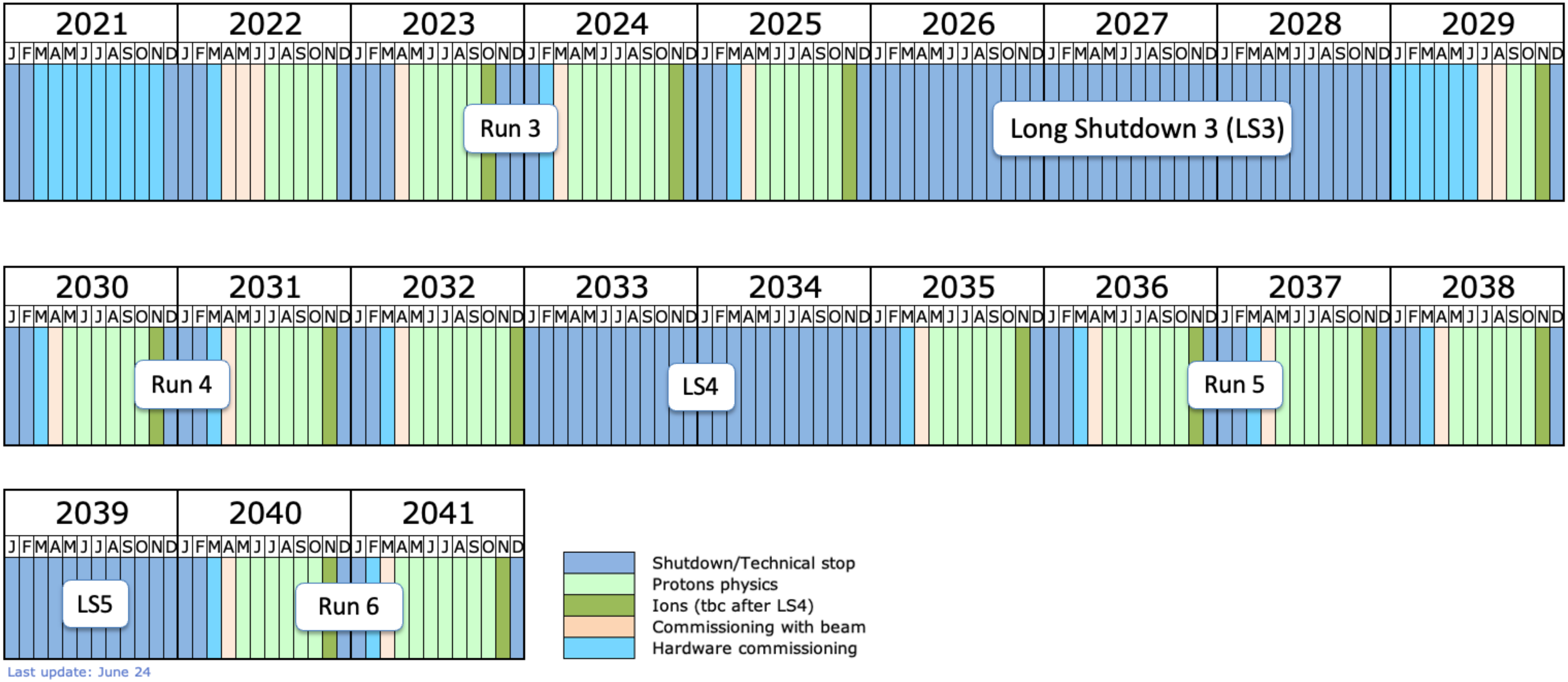
FPGA ?

- Can it be used for computing? Are they powerful?
- We have tested, it is at least comparable to a large GPU
  - See Iaroslava’s talk just after mine!



RECLAIM-DAQ

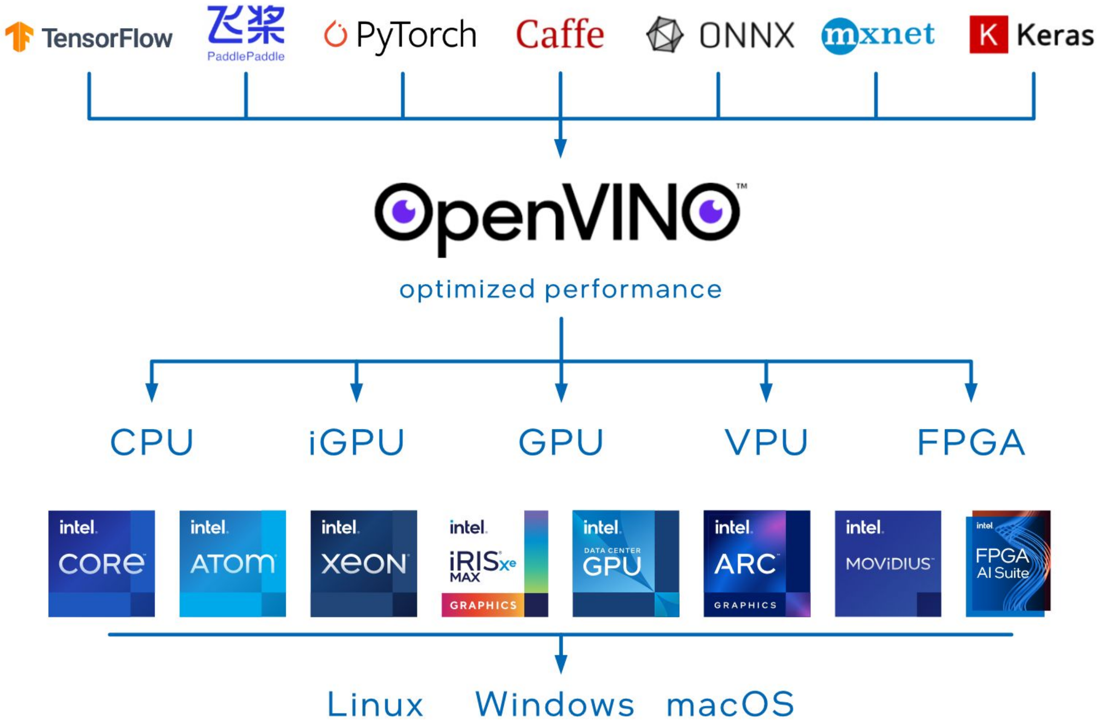
- RECLAIM-DAQ was born when we realized the LHCb DAQ cluster had a lot of very powerful computing resources that were underutilized.
  - **We want to profit this multimillion cluster to compute**

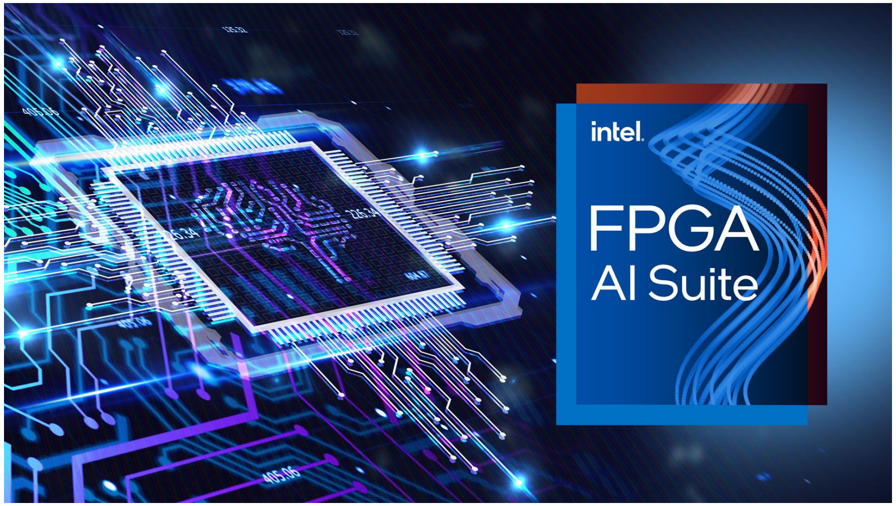


https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm

RECLAIM-DAQ

- Running on CPUs isn’t too complicated
- Running on GPUs takes more work, but isn’t cutting edge these days…
- Can we also profit the FPGAs?
- UZH has been an Intel Partner for the last few years, and we have had support from them to work on AI in Intel FPGAs.



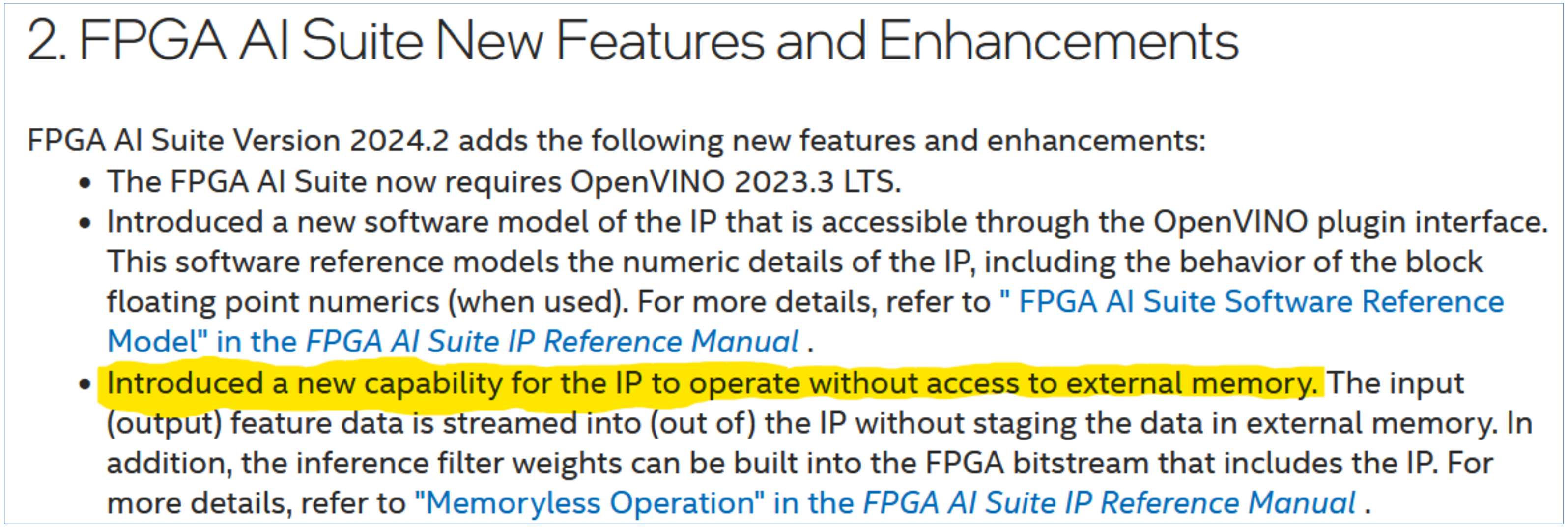


- Our group works with DNN for physics

Clear path… let’s try to use the FPGAs for computing DNNs

RECLAIM-DAQ

- Can we profit the FPGAs at LHCb?
  - The main problem was that TELL40 cards do not have external DDR memory that Intel AI Suite gives for granted.
  - For a while we discussed with Intel about ways around this.
  - Eventually decided it would take unreasonable effort for UZH to do it.
  - Since version 2024.2 (released last month) this is not a problem any more: https://www.intel.com/content/www/us/en/docs/programmable/772497/2024-2/new-features-and-enhancements.html



2. FPGA AI Suite New Features and Enhancements

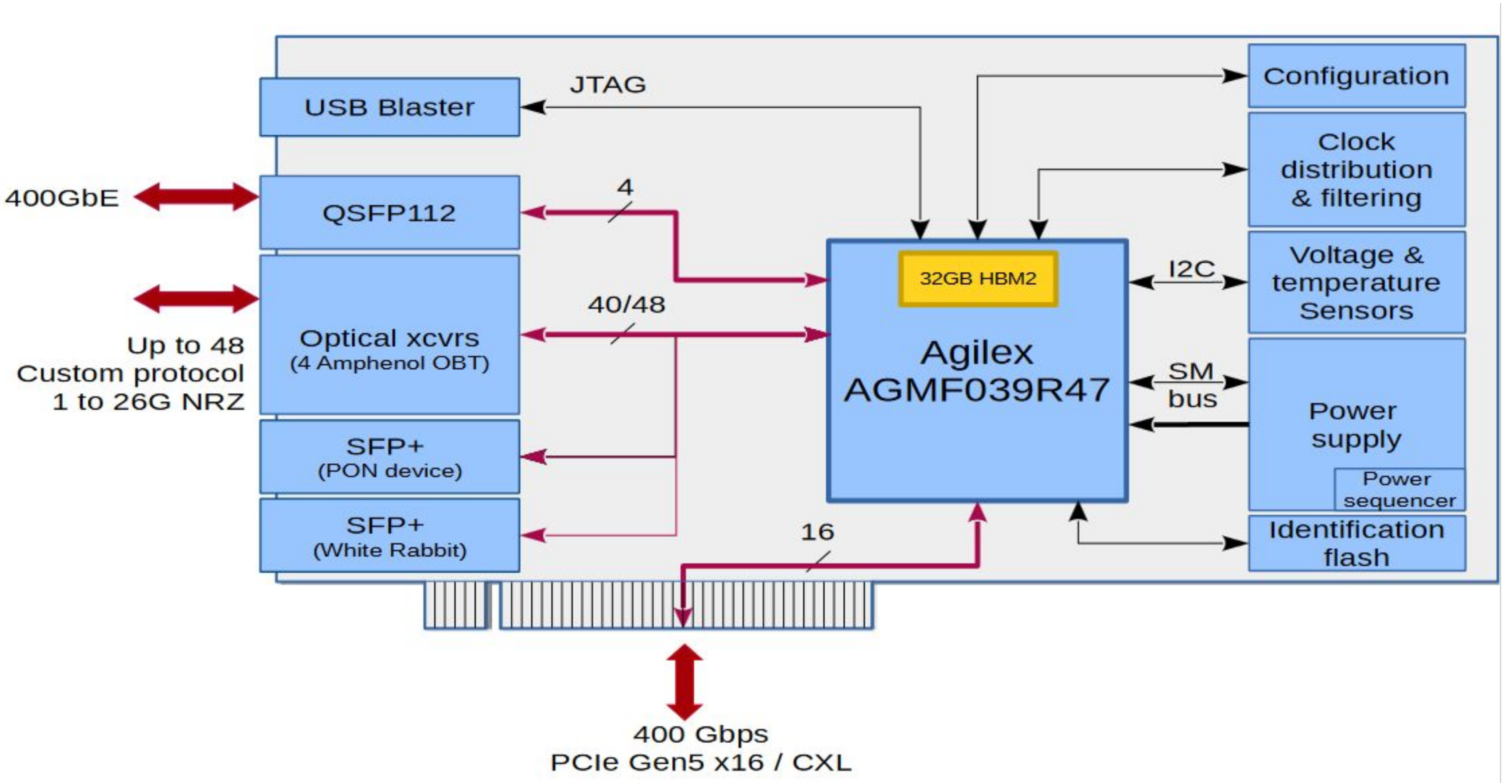
FPGA AI Suite Version 2024.2 adds the following new features and enhancements:

- The FPGA AI Suite now requires OpenVINO 2023.3 LTS.
- Introduced a new software model of the IP that is accessible through the OpenVINO plugin interface. This software reference models the numeric details of the IP, including the behavior of the block floating point numerics (when used). For more details, refer to " FPGA AI Suite Software Reference Model" in the *FPGA AI Suite IP Reference Manual* .
- Introduced a new capability for the IP to operate without access to external memory. The input (output) feature data is streamed into (out of) the IP without staging the data in external memory. In addition, the inference filter weights can be built into the FPGA bitstream that includes the IP. For more details, refer to "Memoryless Operation" in the *FPGA AI Suite IP Reference Manual* .

  - We are starting to do our own tests, to see how it performs...

RECLAIM-DAQ

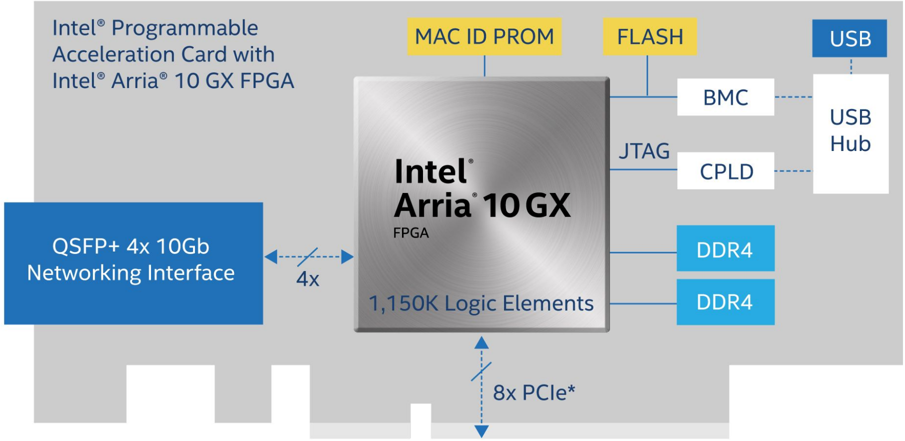
- Can we profit the FPGAs at LHCb?
  - For the future, we contacted CPPM asking if they could provide us DDR memory in the TELL400 upgrade they are preparing.
    - This card will have a more performant Intel Agilex 7 FPGA
  - They came back offering something much better: HBM2 memory embedded in the new FPGAs!



PCIe400 synoptic

RECLAIM-DAQ

- Can we profit the FPGAs at LHCb?
  - Yes, we can.
- Can we proof it?
  - We started with an Intel PAC board


Intel® Programmable Acceleration Card with Intel® Arria® 10 GX FPGA
MAC ID PROM
FLASH
USB
BMC
USB Hub
JTAG
CPLD
Intel® Arria® 10 GX FPGA
1,150K Logic Elements
QSFP+ 4x 10Gb Networking Interface
4x
DDR4
DDR4
8x PCIe*

  - Has an FPGA of the same family of the TELL40, and is well supported by Intel.

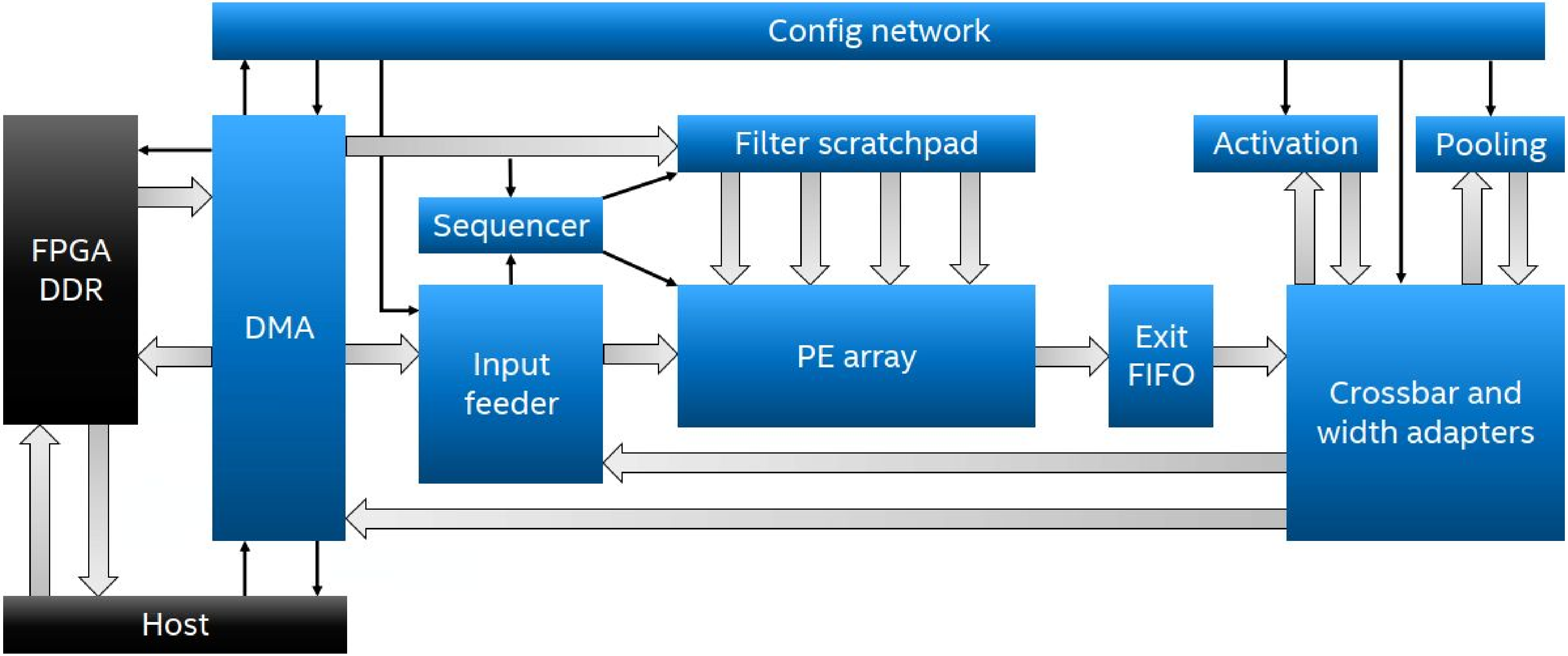
RECLAIM-DAQ

- Can we proof it?
  - Had the FPGA, the software, the knowledge and a plan.
  - We only lacked a subject for experimentation:
    - We have seen an important need of DNN computing power in LHCb: MC sim for CALO, Sprucing, etc…
    - As a first example, we are working with Firenze University to implement their GAN model for parametrizing the LHCb tracking resolution.
  - Everything was ready!
  - Let’s do it!

UNIVERSITÀ
DEGLI STUDI
FIRENZE

RECLAIM-DAQ

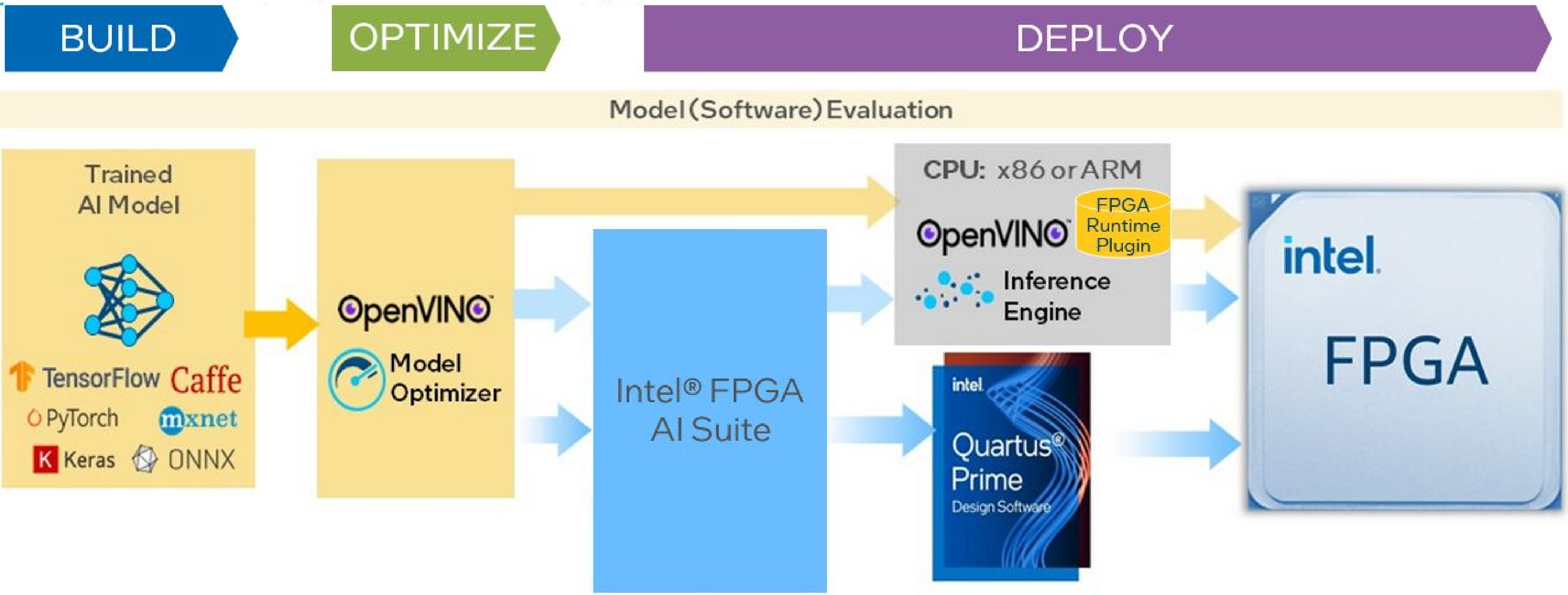
- How does it work?



- The FPGA is filled with a system that can multiply matrices and do AI related functions like Activation Functions, Pooling, etc.
- The vendor provides example designs with different capabilities and sizes.

RECLAIM-DAQ

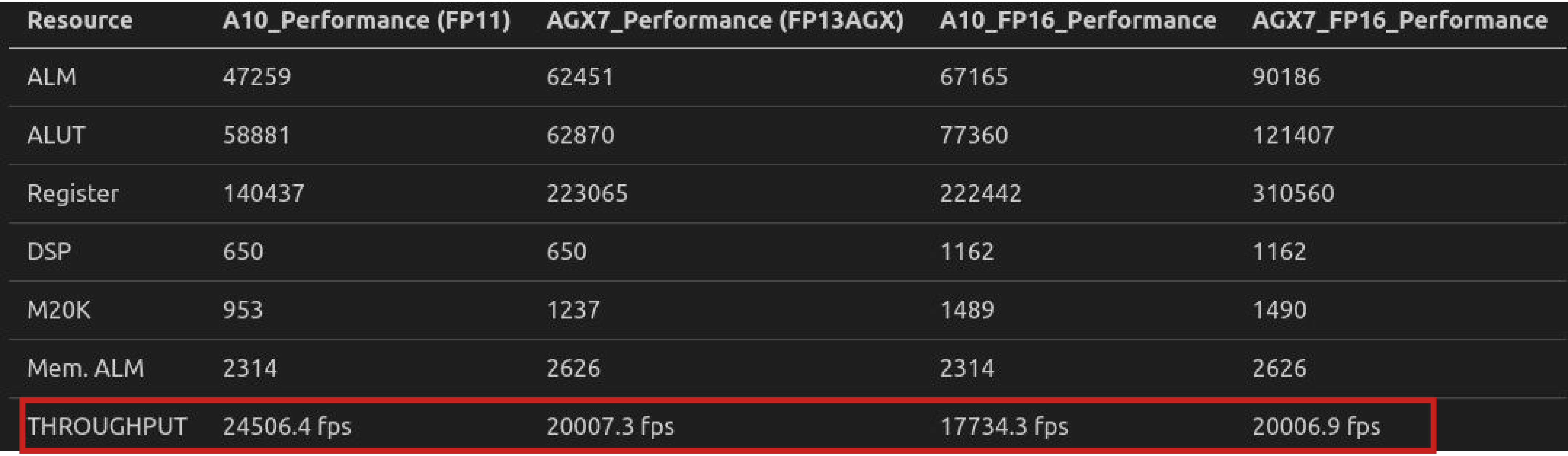
- How does it work?



- It is also possible to tailor a design for a given model. The size of the Processing Engine, buffer sizes, etc… are tuned for best performance.
- We can also choose the precision: FP11, FP16, etc...
- The tool helps with the software stack too; an API is provided with examples.

RECLAIM-DAQ

- Did it work?
- With the example architectures on our AI model:

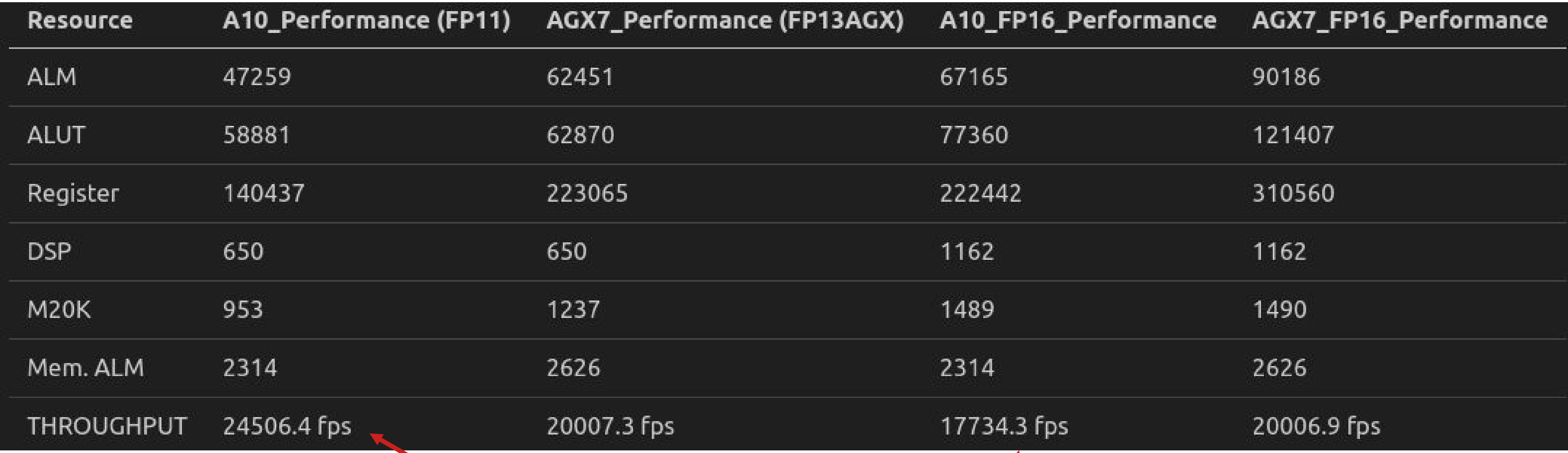


Resource	A10_Performance (FP11)	AGX7_Performance (FP13AGX)	A10_FP16_Performance	AGX7_FP16_Performance
ALM	47259	62451	67165	90186
ALUT	58881	62870	77360	121407
Register	140437	223065	222442	310560
DSP	650	650	1162	1162
M20K	953	1237	1489	1490
Mem. ALM	2314	2626	2314	2626
THROUGHPUT	24506.4 fps	20007.3 fps	17734.3 fps	20006.9 fps

  - Estimations by compiler tool
  - Made with different precisions (low and high) for the two cards: the TELL40 currently installed and the TELL400 of the upgrade.

RECLAIM-DAQ

- Did it work?
- But these are estimations, what about the real numbers?

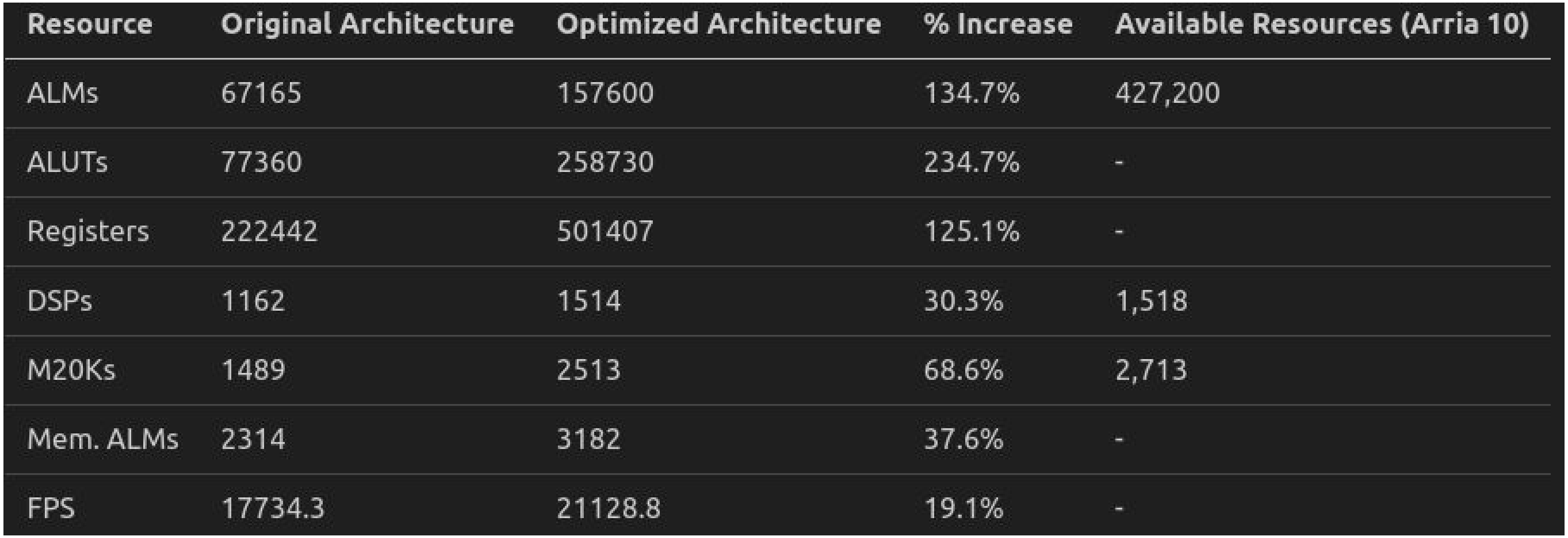


Resource	A10_Performance (FP11)	AGX7_Performance (FP13AGX)	A10_FP16_Performance	AGX7_FP16_Performance
ALM	47259	62451	67165	90186
ALUT	58881	62870	77360	121407
Register	140437	223065	222442	310560
DSP	650	650	1162	1162
M20K	953	1237	1489	1490
Mem. ALM	2314	2626	2314	2626
THROUGHPUT	24506.4 fps	20007.3 fps	17734.3 fps	20006.9 fps

- We don’t have measurements for AGX7, as these cards are not available yet.
- But our measurements of Arria 10 cards show:
  - A10_Performance(FP11): 25024 fps
  - A10_FP16_Performance: 18596 fps
- If any, estimations are typically pessimistic...

RECLAIM-DAQ

- Can you optimize it for this given model?
  - Yes, for instance, this is with FP16, the worst case scenario:

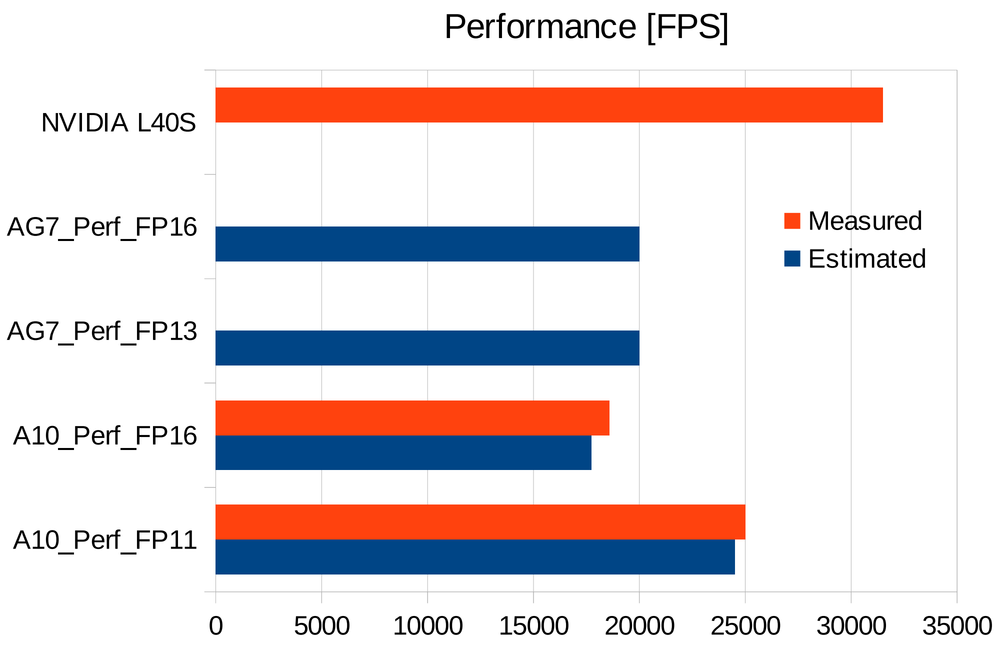


Resource	Original Architecture	Optimized Architecture	% Increase	Available Resources (Arria 10)
ALMs	67165	157600	134.7%	427,200
ALUTs	77360	258730	234.7%	-
Registers	222442	501407	125.1%	-
DSPs	1162	1514	30.3%	1,518
M20Ks	1489	2513	68.6%	2,713
Mem. ALMs	2314	3182	37.6%	-
FPS	17734.3	21128.8	19.1%	-

  - There is some gain, but not groundbreaking.
  - We were told in some cases it is actually better to use 2 smaller inference engines than 1 large one.
  - We are working on that...

RECLAIM-DAQ

- How does this compare with a GPU?



- We think there are still things to be improved, but as it is today, we have measured ~ 2/3 of a very performant GPU.
- If we can reclaim 555 FPGAs, (**<u>for this model</u>**) we could have the equivalent of 370 high-end GPUs without a major effort and at ZERO cost.
  - *<u>This is very model dependent! (we have models with ~x2 performance in CTA!)</u>*

Conclusions

- We propose using the DAQ clusters for computing, including the FPGAs.
  - Our **case of study** in LHCb shows that we could reclaim:
    - 11.776 CPU cores
    - 555 GPUs
    - 555 FPGAs
      - **In our example**, equivalent to 370 high-end GPUs
- Still many things to do:
  - This is only a proof of principle that this can be done:
    - We gave an Intel example because LHCb uses it, but other experiments will use other brands.
  - We want **a *transparent* framework** such that a user doesn't have to fight the system: just give us the model, the data, and we compute it for you. We are starting to work in this direction.
  - … but promising results already

Conclusions

- Other experiments have complete triggers made with FPGAs, plus the DAQ, plus others:
  - Large HEP experiments spend millions in FPGAs and the community is not taking full profit of this investment.
  - We want to get in contact with DAQ designers so new cards make this easy. In many cases the effort would be minimal.
  - Other experiments have Xilinx FPGAs. We also plan on developing for those too.
- Studies done in LHCb show how running algorithms in FPGAs can improve energy efficiency significantly [1]. We want this for our computing.
  - We can save electricity and CO_2 emissions.
- We aim to help having a responsible and efficient use of resources.
  - We can contribute to have less e-waste

[1] G. Bassi et al., "A FPGA-Based Architecture for Real-Time Cluster Finding in the LHCb Silicon Pixel Detector," in IEEE Transactions on Nuclear Science, vol. 70, no. 6, pp. 1189-1201, June 2023, doi: 10.1109/TNS.2023.3273600.

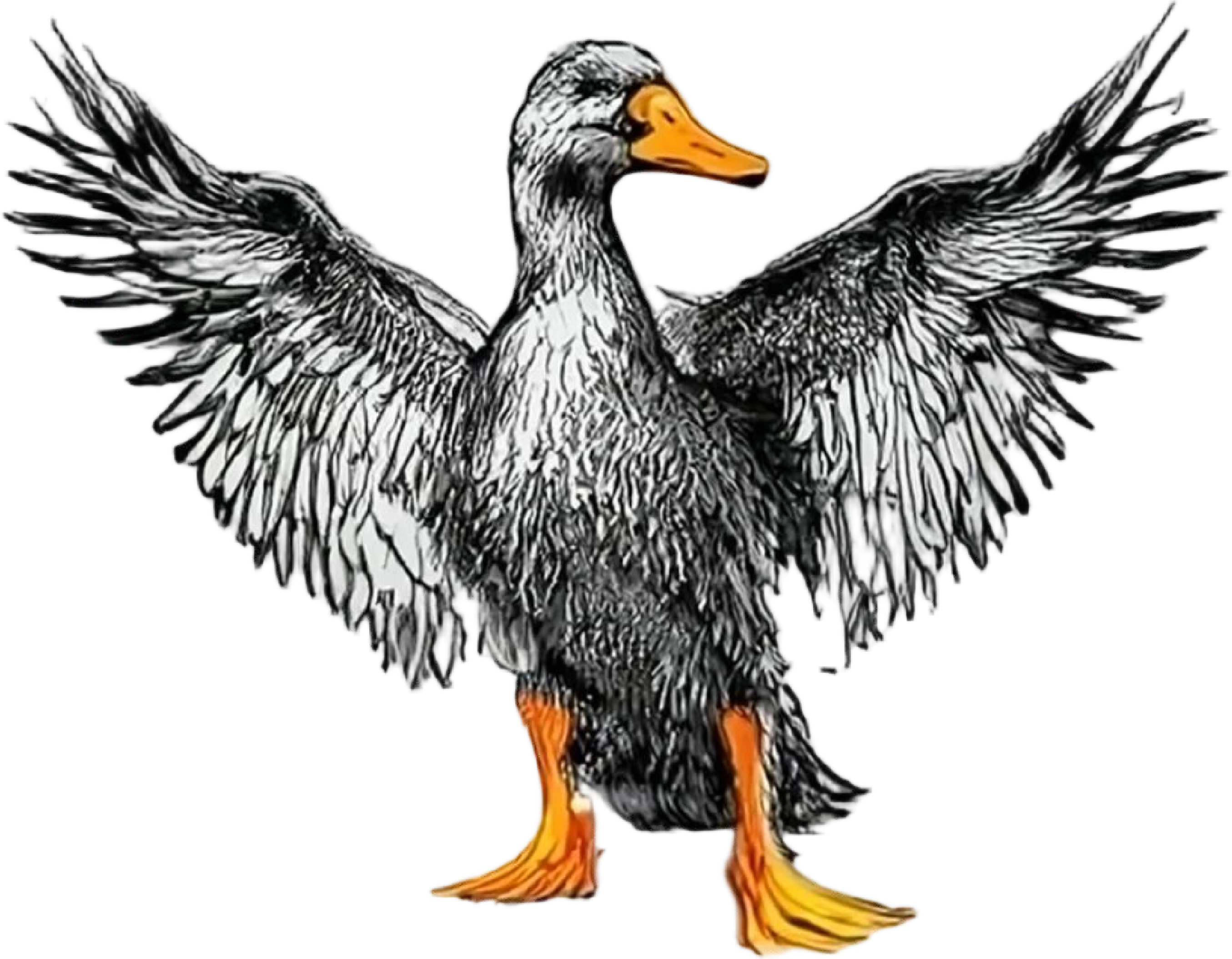
Conclusions

- Do you have a DNN model that you would want to try in an FPGA?
- Do you want to get into FPGAs and AI ? Edge computing? Heterogeneous computing?
- Do you like the idea but you work in an experiment with a different DAQ ?

<u>We are looking for collaborations</u>

<u>Please come talk to us!</u>

Questions?



RECLAIM-DAQ

A framework for reclaiming the DAQ for computing