UNIVERSITÉ DE GENÈVE

ATLAS EXPERIMENT

Vitis Accelerator Backend for HSL4ML

CHIPP 2024 Annual Meeting

Quentin Berthet, Kostas Axiotis
Université de Genève
Prof: Anna Sfyrla

19/06/2024

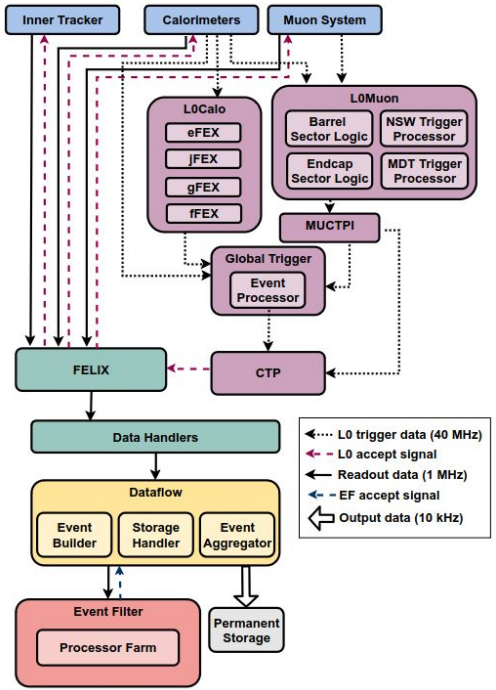
High Luminosity LHC (challenges)

High Luminosity consequences

- High pile-up up to 200 events per bunch crossing (previously ~ <40>)
- High granularity detectors that need to be read out
  - New, all-silicon Inner Tracker (ITk)
  - front-end/back-end electronics updates
- Larger event size ~5.2 MB (~2 MB previously)

Operating points for ATLAS TDAQ

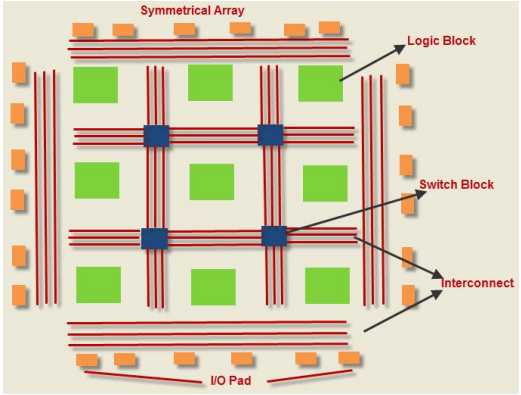
- L1 latency increase to ~ 10 μs (~2.5 μs previously)
- Readout rate increase to 1-4 MHz (100 kHz previously)
- Rate to permanent storage ~ 10 kHz (~ 1 kHz previously)



Field Programmable Gate Array characteristics and architecture

FPGA advantages

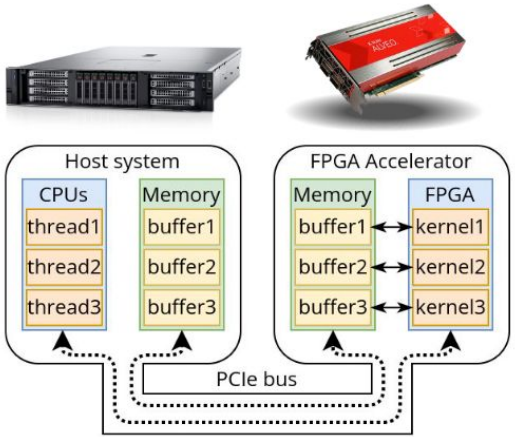
- Flexible processing
- Flexible input/output
- Parallel processing
- Multiclock systems
- Better with time-critical operations
- Predictable project cycles



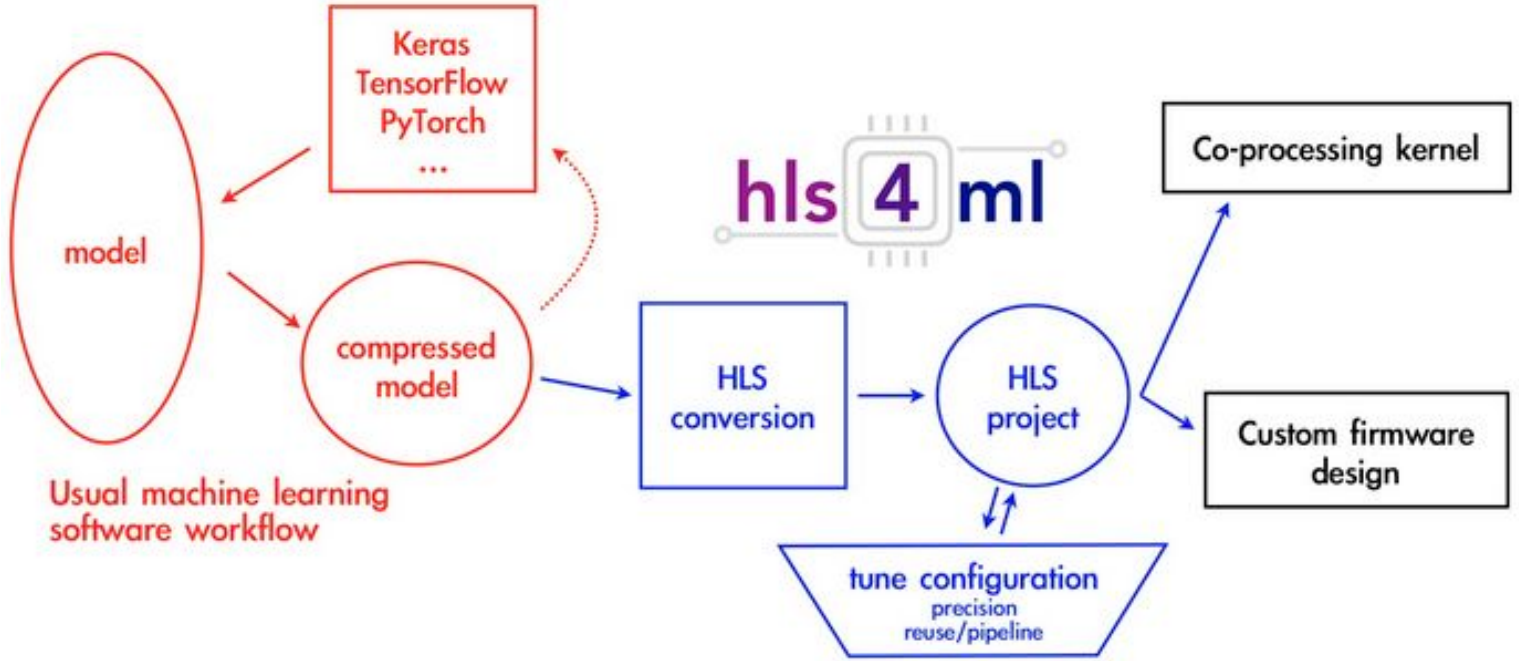
FPGA as an accelerator

Specialized and efficient hardware used to off-load computing from general-purpose CPUs.

- Alveos: Off-the-shelf AMD accelerators:
  - FPGA paired with memory (DDR or HBM)
  - PCIe bus
- Software developer oriented



High Level Synthesis For Machine Learning



HLS4ML - Vitis Accelerator Backend Motivation

Vivado Flow:

- Supports standard Vivado IP creation.
- Primarily targets embedded system applications.

Vitis Flow:

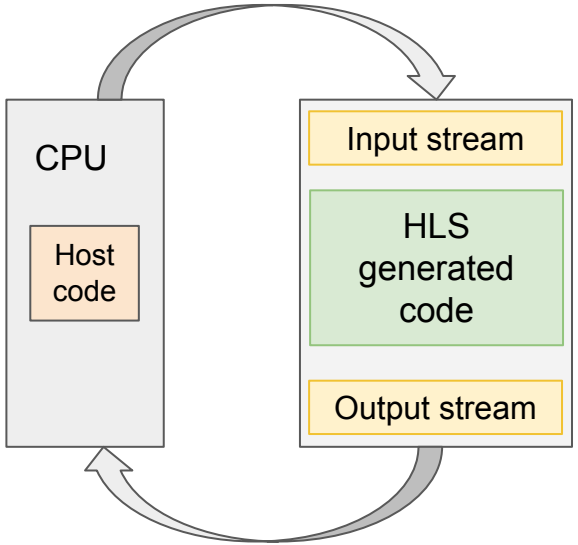
- Designed for Vivado IP only.
- Lacks direct support for kernel expression.

Vivado Accelerator Flow:

- Not compatible with Alveo accelerator cards.

In Development: Vitis Accelerator Backend

- **Objective:** Enable direct kernel expression.
- **Potential Application:** Event filtering in high-performance environments.
- **Prototyping:** Leveraging FPGA technology for testing and development.



QDips on FPGA accelerator

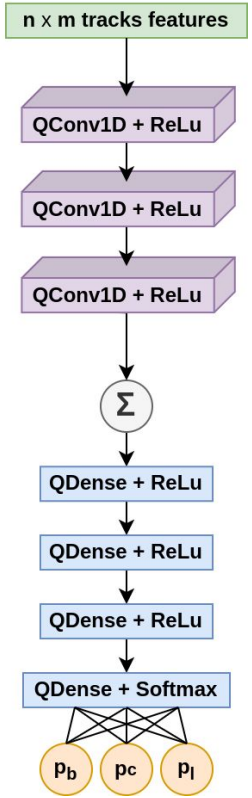
Introduced on
Stefano and Luca's talk

Deep Sets Model for Jet b-Tagging in ATLAS:

- Offline Use
- In Trigger

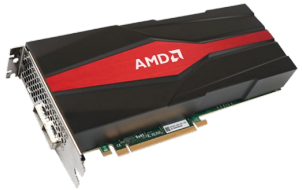
Quantization Aware Training with QKeras:

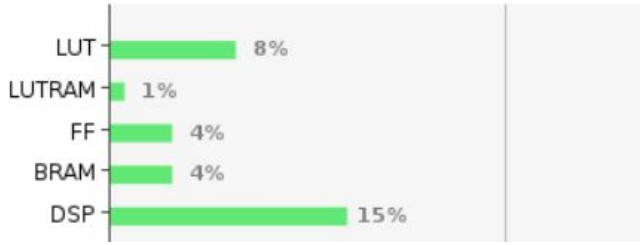
- Employs QKeras for training the Deep Sets model with quantization awareness, ensuring that the model remains effective even when constrained to lower precision, which is crucial for deployment on hardware with limited resources.

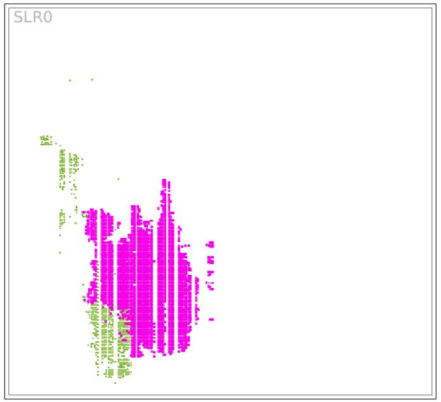


QDips on FPGA accelerator - 1 Kernel Implementation

Board	Versal VCK5000
Frequency	300 MHz
Batch size	256
Throughput	~100000 Predictions/s



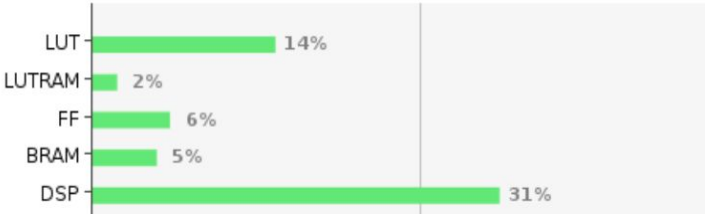


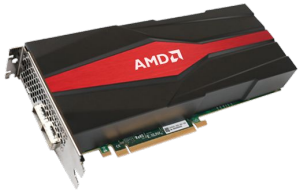


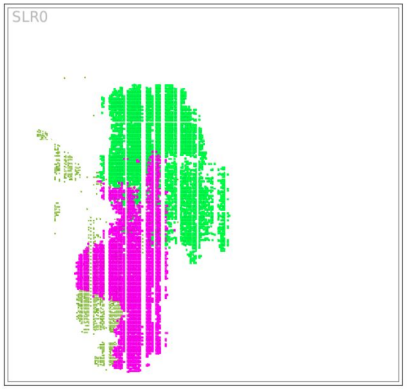
10us per inference **including** the **transfer** to and from PCIe and DDR loading from kernel

QDips on FPGA accelerator - 2 Kernel Implementation

Board	Versal VCK5000
Frequency	300 MHz
Batch size	256
Throughput	~210000 Predictions/s







5us per inference **including** the **transfer** to and from PCIe and DDR loading from kernel

Summary

Challenges of the High Luminosity LHC:

- Increased data rates and complexity
- Necessity for efficient Trigger and Data Acquisition (TDAQ) systems

Custom FPGA Boards and FPGA Accelerators:

- FPGA solutions for high-performance data processing
- Benefits of custom FPGA designs in addressing LHC challenges

HLS4ML:

- Streamlining the implementation of ML algorithms on FPGAs

The Vitis Accelerator Backend:

- Integration of the Vitis Accelerator
- Targeting Alveo boards for enhanced performance and flexibility

Validation with QDips Algorithm:

- Demonstration of Vitis Accelerator’s effectiveness
- Successful validation using the QDips algorithm

Thank you!