QDIPS: Deep Sets Network for FPGA investigated for high speed inference on ATLAS

CHEP 2024, Kraków, Poland

Swiss National Science Foundation

Deep Sets neural networks are useful applications for variable sized, unordered inputs (e.g. tracks associated to jets): Made permutation invariant to input order.

QDIPS is a *quantised* version of "**fastDIPS**" (fast Deep Impact Parameter Sets) [1], an ATLAS low-level *Deep Sets* based jet flavour tagging algorithm [2] used in the Run 3 High Level Trigger. We aim to run QDIPS on FPGA as demonstrator for a high throughput heterogeneous compute TDAQ system.

QDIPS Project Goals

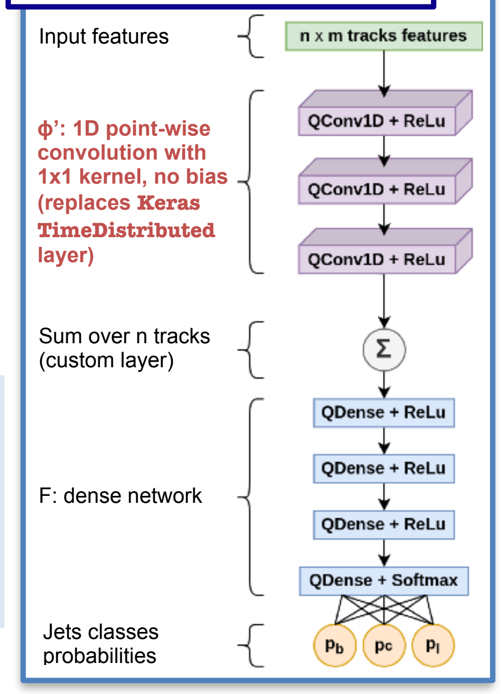
- ✦ Adapt architecture to FPGA: Does performance match up to full-precision CPU performance?
- ✦ Fit it on FPGA: Does it fit and use fraction of FPGA resources?

We use QKeras (quantised machine learning) & HLS4ML (high level synthesis translation for FPGA) [3,4].

DIPS architecture for HLS4ML

Not directly translatable from original: Needed replacements for Keras TimeDistributed layer (see Φ') & Masking of empty tracks: We avoid bias in Φ' so empty tracks don't contribute to sum.

QDIPS Neural Network Architecture
(changes w.r.t. fastDIPS highlighted)



Sample Inputs

We use ATLAS Monte Carlo Upgrade Reconstruction samples with 200 proton interactions per bunch-crossing. Event Filter "fast online reco quality" tracks are emulated using a *fast emulation* tool [5] so that: p_T^{trk} > 2 GeV, $\epsilon_{trk} = 98\,\%$, σ_{trk} smeared by factor 5 w.r.t. offline reco tracks.

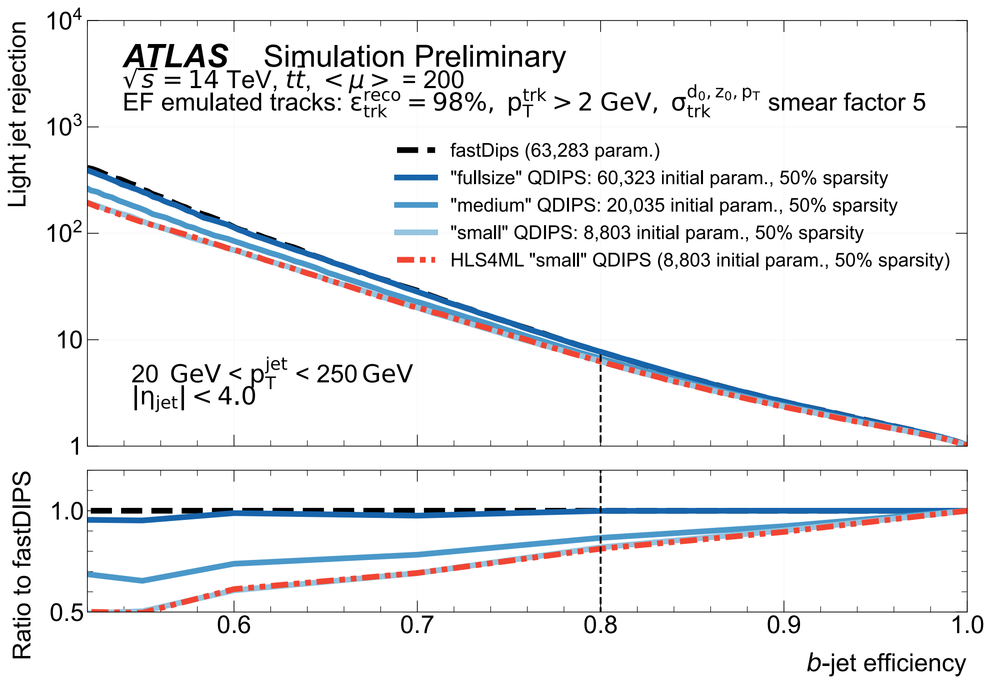


Figure 1: b-jet tagger performance roc curve for different QDIPS model sizes. The performance of the original for-CPU fastDIPS (dashed black line) as well as the HLS4ML-translated QDIPS tagger (red dot-dashed line) are shown for comparison.

Performance versus model size

Aggressively **scaling down model size** has small performance impact @ 80% b-jet identification ϵ (Fig.1):

"full size" → "medium" QDIPS: light jet rej. 7.7 → 6.7 (-13%)

"full size" → "small" QDIPS: light jet rej. 7.7 → 6.3 (-19%)

Optimising bit precision

A bit precision scan shows that **homogeneously decreasing precision** from 16-bit to 8-bit has small impact on performance (Fig.2) e.g.:

16-bit → 8-bit "small" QDIPS: light jet rej. 6.3 → 5.7 (-10%)

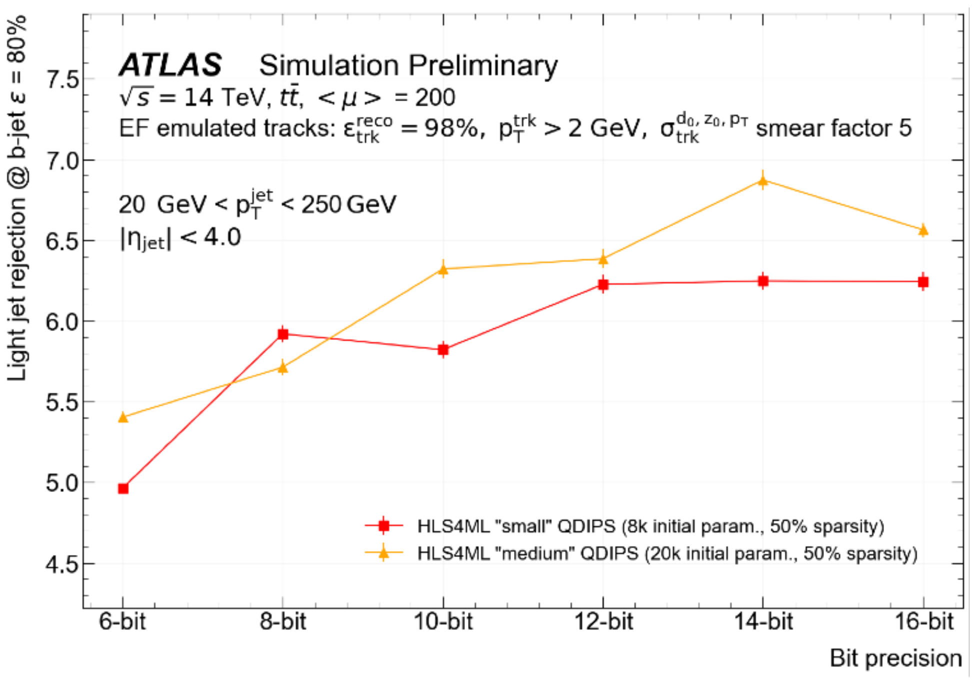


Figure 2: The light jet rejection at 80% b-jet tagging efficiency is shown for different uniformly applied bit precisions for "small" QDIPS (red squares) and "medium" QDIPS (yellow triangles).

Resource usage

Synthesising 8-bit QDIPS for an AMD Alveo U250 FPGA shows we can balance latency vs resource usage with the reuse factor (no. of times a computing unit is reused) (Fig. 3). Reuse factors of > 64 limits usage to 5% LUT resources, allowing us to instantiate 4 cores on one board (Fig. 4).

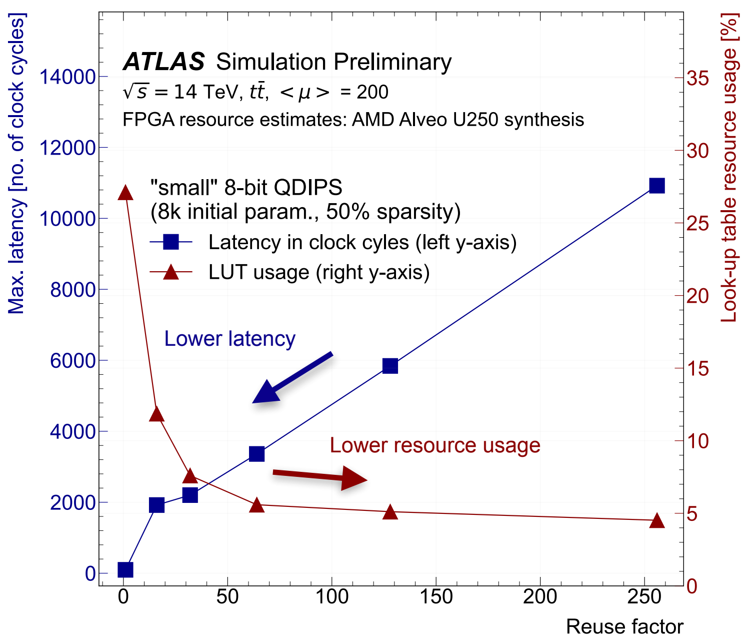


Figure 3: The LUT usage (red squares) & latency (blue squares) versus reuse factor shows the trade-off between resource reduction and latency.

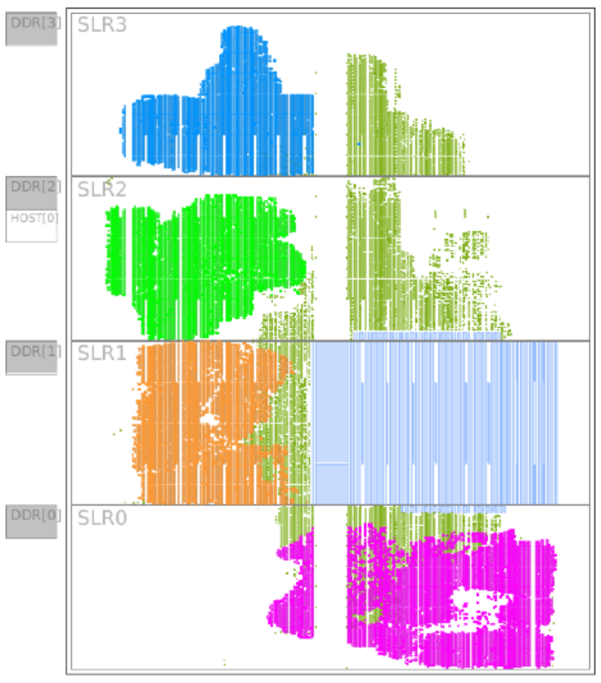


Figure 4: Floor plan showing an AMD Alveo U250 implementation with 4 model instantiations.

QDIPS model	Reuse factor	Max. latency	LUT used	Light jet rejection (% rel. to fastDIPS)
8-bit "small"	16 256	9.6μs 54.6 μs	11.86% 4.52%	5.9±0.05 (−23%)
8-bit "medium"	32 256	13.2μs 55.0μs	18.04% 9.16%	5.7±0.05 (−25%)

Table 1: Resource estimates on AMD Alveo U250 synthesis at 200 MHz, and flavour tag performance at 80% b-jet tagging ϵ (statistical uncertainty given).

Summary

FPGAs are power efficient, low latency alternatives to CPUs/GPUs for accelerated computing at the ATLAS Event Filter at the HL-LHC. We demonstrated that we can fit a Deep Sets network on a AMD Alveo U250 accelerator with down to 5% LUT usage and 25% decrease in performance compared to CPU, enabling multiple cores for increased throughput. The low latencies achieved may also make it applicable to hardware triggering in future following further optimisations.

References
[1] "Fast b-tagging at the high-level trigger of the ATLAS experiment in LHC Run 3" by the ATLAS coll., arXiv:2306.09738 (JINST 2023)
[2] "Deep Sets based Neural Networks for Impact Parameter Flavour Tagging in ATLAS", ATL-PHYS-PUB-2020-014
[3] "Fast inference of deep neural networks in FPGAs for particle physics", Duarte, Javier and others, arXiv:1804.06913 (JINST 2018)
[4] https://github.com/fastmachinelearning/hls4ml
[5] "Performance studies of tracking-based triggering using a fast emulation", ATL-DAQ-PUB-2023-001

Link to plots: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/PhysicsAndPerformancePhaseIIUpgradePublicResults

Claire Antel, Quentin Berthet, Stefano Franchelucci, Anna Sfyrla (University of Geneva), on behalf of the ATLAS collaboration

UNIVERSITÉ DE GENÈVE
Département de physique nucléaire et corpusculaire
ATLAS EXPERIMENT