



*Boosting discoveries through hardware-based
machine learning, FPGAs, and embedded systems*
Chicagoland Workshop on LHC Analysis with Machine Learning

David W. Miller

Enrico Fermi Institute



THE UNIVERSITY OF
CHICAGO

January 15, 2019



Outline

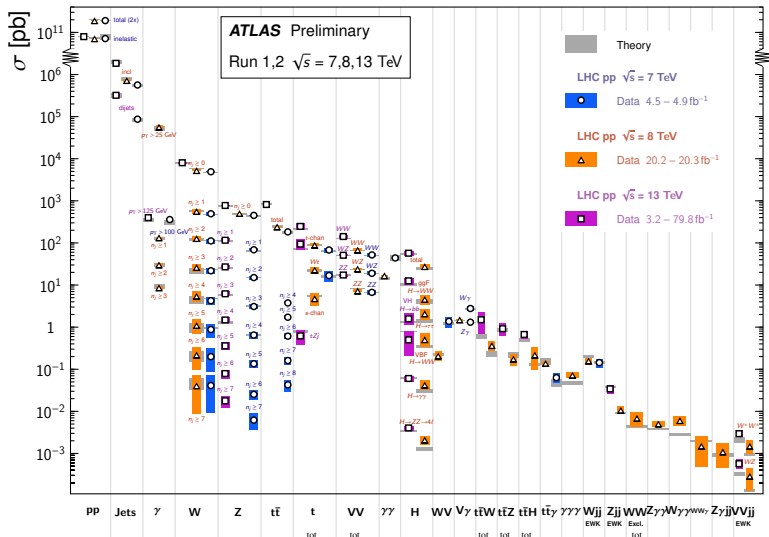
- 1 *Challenges of the Energy and Luminosity Frontier*
- 2 *ATLAS Phase I & II Hadronic Trigger Systems*
- 3 *Machine learning using FPGAs and MPSoCs*
- 4 *Summary and conclusions*

The overwhelming hadronic environment of the LHC

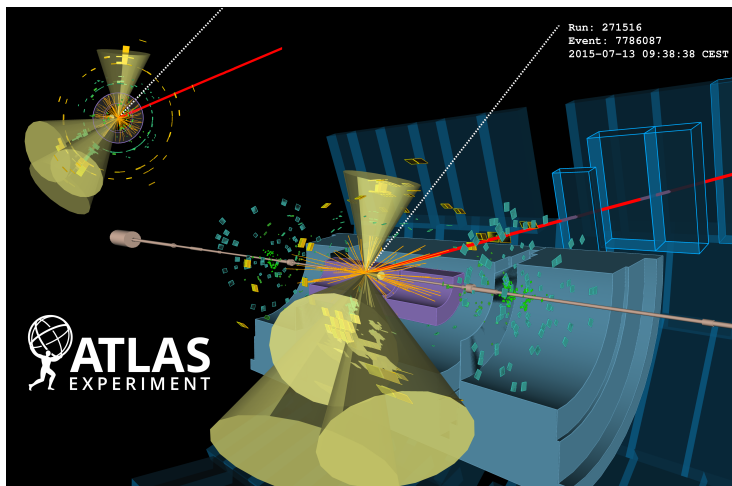
HL-LHC: $\mathcal{L}_{inst} = 10^{35} \text{ cm}^{-2} \text{ s}^{-1} = 0.1 \text{ pb}^{-1} \text{ s}^{-1} = 30 \text{ kHz}$ of dijet events

Standard Model Production Cross Section Measurements

Status: July 2018

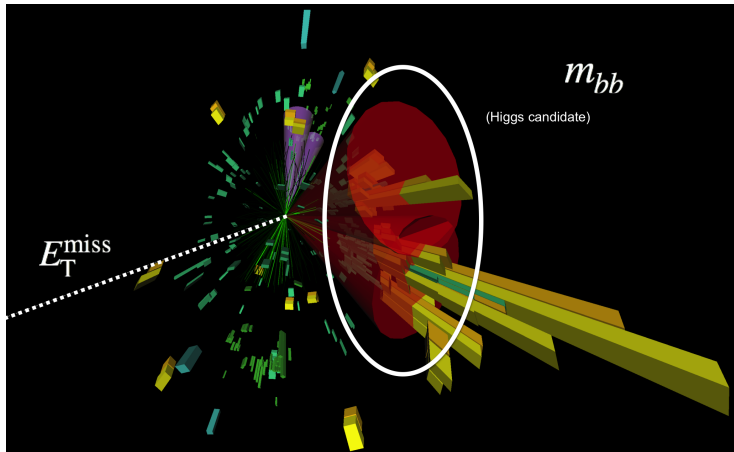


Hadronic final states: major part of LHC physics program



- Physics may be **compromised due to trigger & data proc. limitations**
- Even if we *can* trigger, **offline data management may be a bottle-neck**

Hadronic final states: major part of LHC physics program



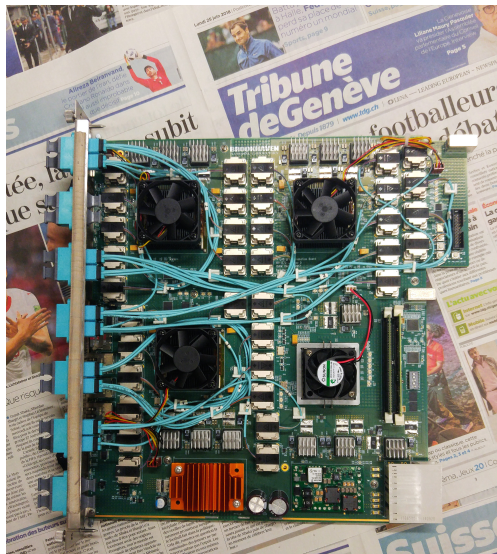
- Physics may be **compromised due to trigger & data proc. limitations**
- Even if we *can* trigger, **offline data management may be a bottle-neck**

Outline

- 1 *Challenges of the Energy and Luminosity Frontier*
- 2 *ATLAS Phase I & II Hadronic Trigger Systems*
- 3 *Machine learning using FPGAs and MPSoCs*
- 4 *Summary and conclusions*

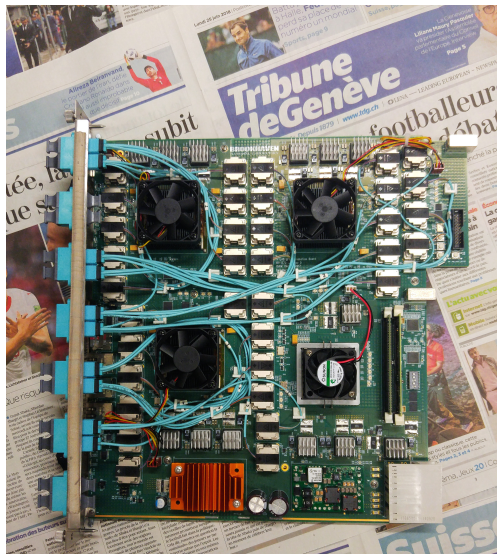
Triggering on complex hadronic final states

Go GLOBAL: global feature extraction trigger (gFEX) for ATLAS Run 3



Triggering on complex hadronic final states

Go GLOBAL: global feature extraction trigger (gFEX) for ATLAS Run 3

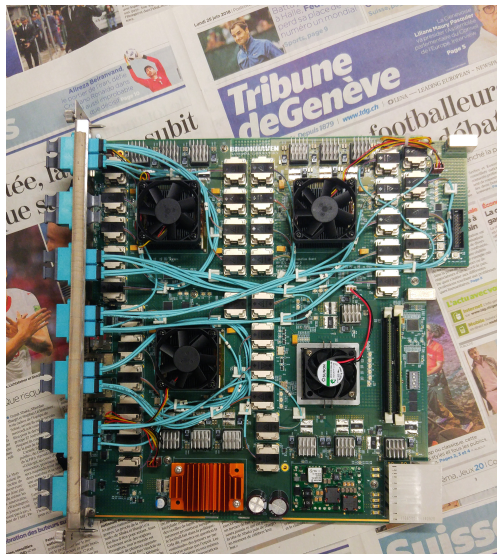


Goal

analyze **event-level features** for characteristics of moderate p_T (~ 100 's of GeV) signatures of **new and key physics processes**

Triggering on complex hadronic final states

Go GLOBAL: global feature extraction trigger (gFEX) for ATLAS Run 3



Goal

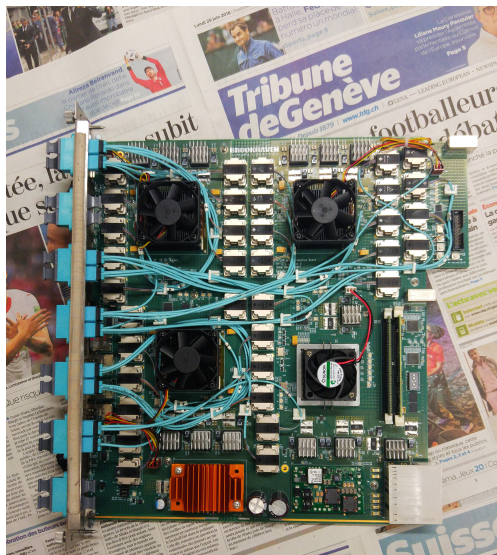
analyze **event-level features** for characteristics of moderate p_T (~ 100 's of GeV) signatures of **new and key physics processes**

Strategy

input entire calorimeter **onto a single trigger board**

Triggering on complex hadronic final states

Go GLOBAL: global feature extraction trigger (gFEX) for ATLAS Run 3



D.W. Miller (EFI, Chicago)

Goal

analyze **event-level features** for characteristics of moderate p_T (~ 100 's of GeV) signatures of **new and key physics processes**

Strategy

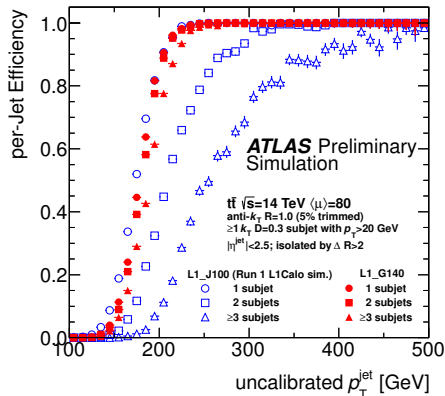
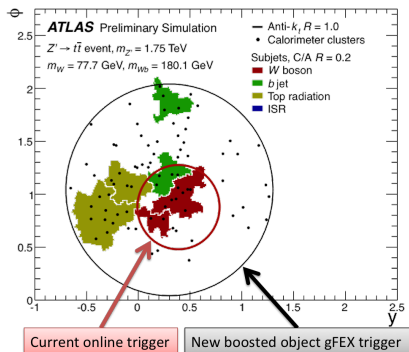
input entire calorimeter **onto a single trigger board**

Tactics

- coarse towers (0.2×0.2)
- state-of-the-art FPGAs
- MPSoC for control, *additional processing*

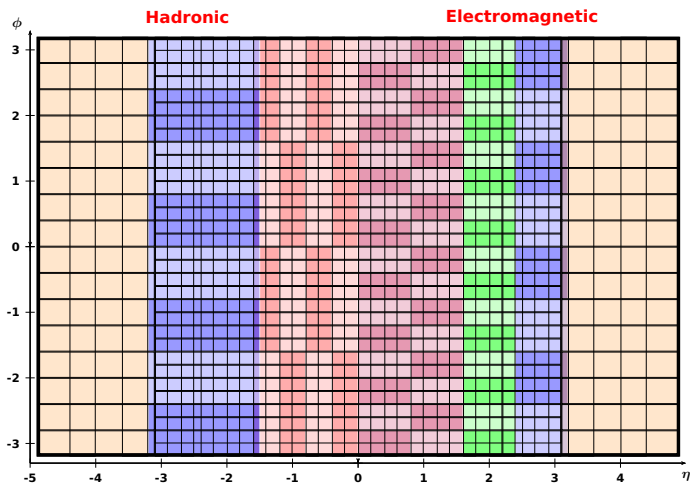
gFEX Performance for Run 3

- **Signal: e.g. boosted tops**
- **Compare to Run 2 triggers**



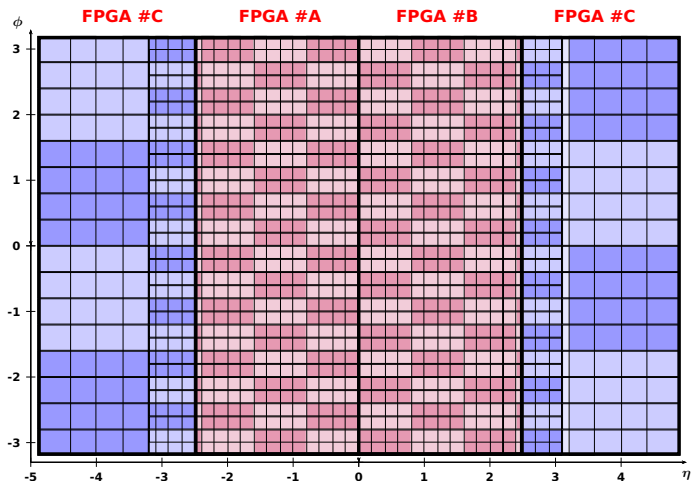
gFEX can efficiently identify jet structure at 300 GeV!

The gFEX trigger design



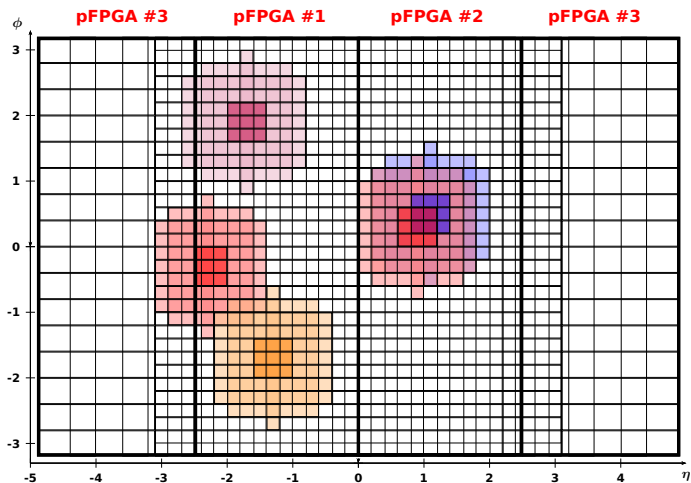
- Implement **new algorithms** using **state-of-the-art FPGAs + SoCs**
- **Image-like** event format is **well-suited for computer vision & ML**

The gFEX trigger design



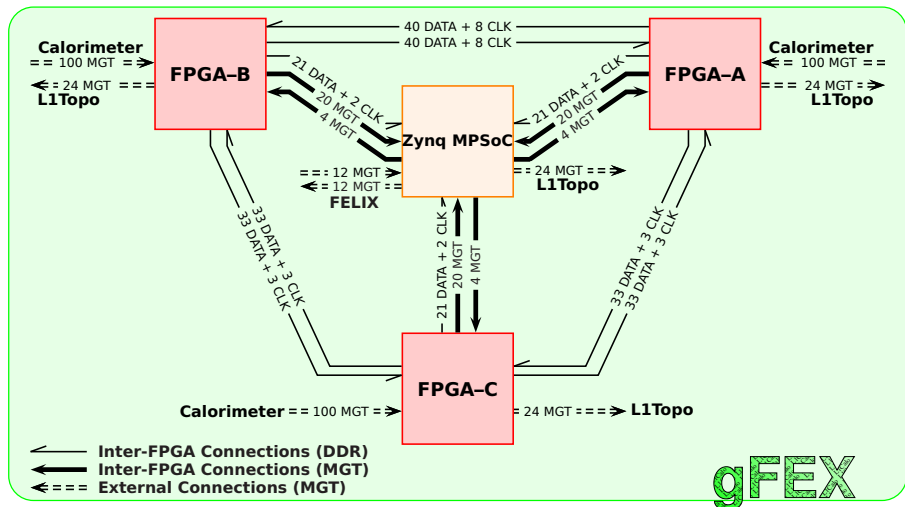
- Implement **new algorithms** using **state-of-the-art FPGAs + SoCs**
- **Image-like** event format is **well-suited for computer vision & ML**

The gFEX trigger design



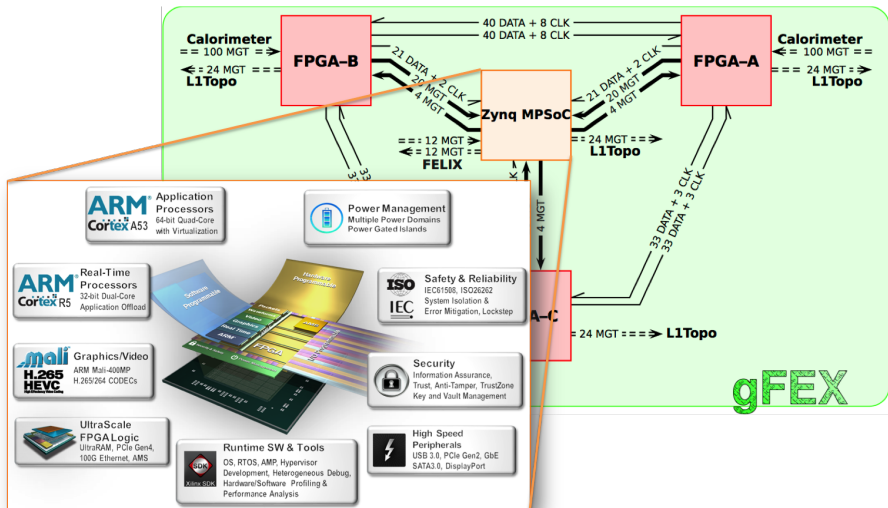
- Implement **new algorithms** using **state-of-the-art FPGAs + SoCs**
- **Image-like** event format is **well-suited for computer vision & ML**

gFEX Design: Virtex 7 & Zynq UltraScale+



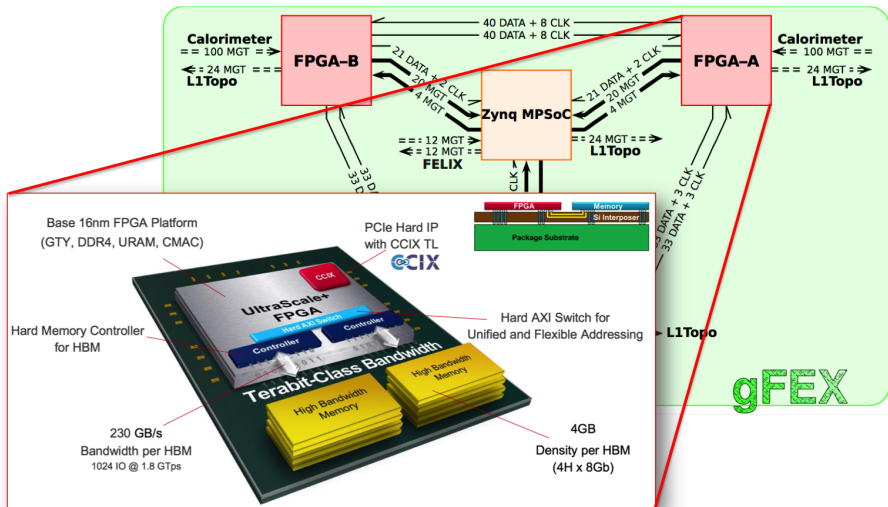
2.3 Tb/s of calorimeter data received by gFEX

gFEX Design: Virtex 7 & Zynq UltraScale+



2.3 Tb/s of calorimeter data received by gFEX

gFEX Design: Virtex 7 & Zynq UltraScale+

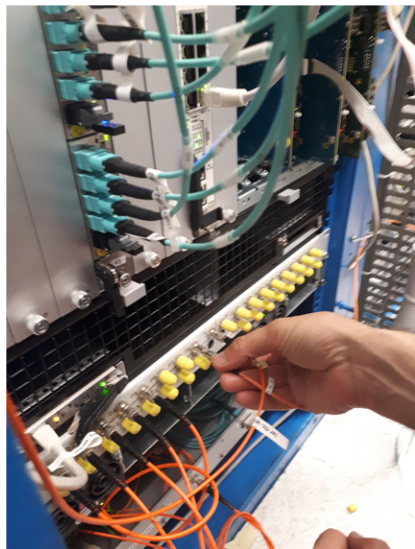


gFEX

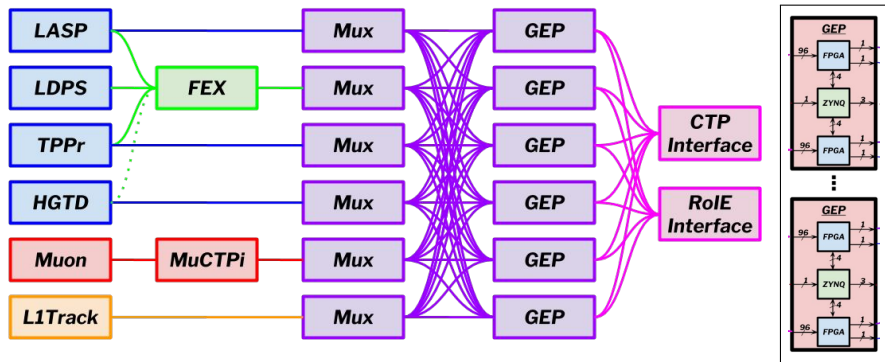
2.3 Tb/s of calorimeter data received by gFEX

gFEX already recorded Stable Beams data in Run 2!

- **gFEX recorded data Stable Beams data on Oct 16, 2018!**
 - Calorimeter back-end system is a prototype for the Phase I/II upgrade (Run 3 & 4)
 - This is a major milestone, but there is certainly more to come



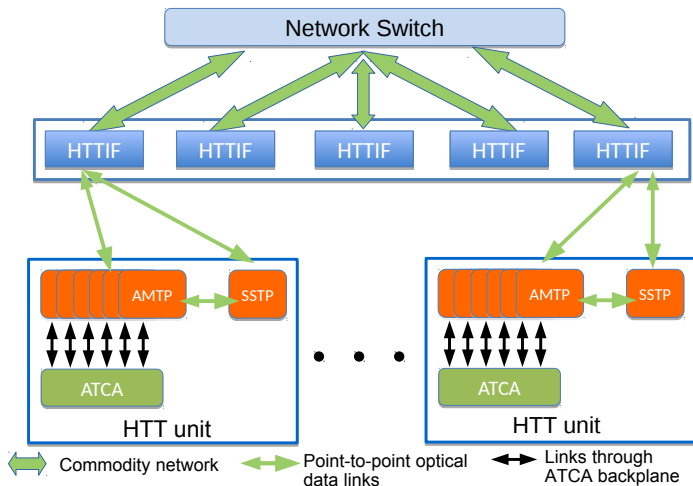
Run 3 ideas for Run 4 reality: Global Event Processor



Receives trigger object information from **all systems (jets, electrons, muons, timing, and possibly tracks)**. Makes **global trigger decision** about the event.

Built from a *common module* with both a **Zynq and two processor FPGAs**.

Run 3 ideas for Run 4 reality: Hardware Track Triggers

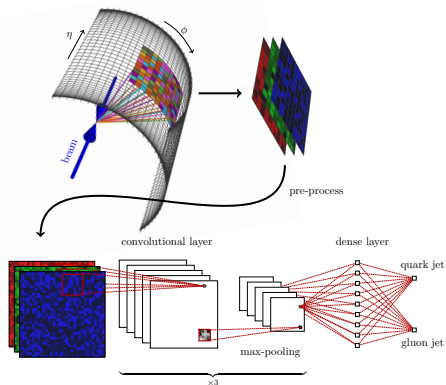


Regional tracking at 1 MHz and global tracking at 100 kHz accomplished with associative memory **ASICs** (AMTP) with tracking in **FPGAs** (SSTP)

Outline

- 1 *Challenges of the Energy and Luminosity Frontier*
- 2 *ATLAS Phase I & II Hadronic Trigger Systems*
- 3 *Machine learning using FPGAs and MPSoCs*
- 4 *Summary and conclusions*

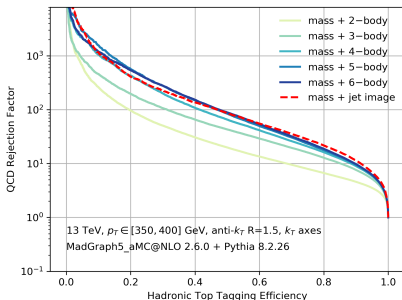
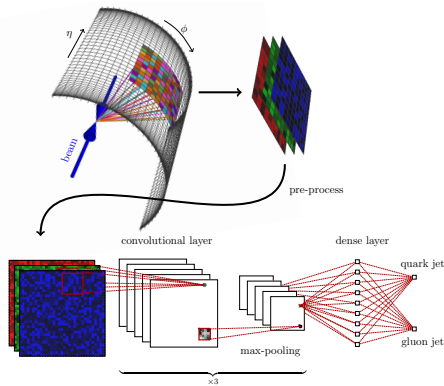
Convolutional neural networks (CNN) for jet identification



Komiske, Metodiev, Schwartz
(arXiv:1612.01551)

Convolutional neural networks (CNN) for jet identification

Effectively the same CNN from Komiske, *et al.* can be used for top-tagging, using either high-level observables or jet images.



Komiske, Metodiev, Schwartz
(arXiv:1612.01551)

Moore, Nordström, Varma, Fairbairn
(arXiv:1807.04769)

(CNNs here use 4 layers, 64 filters in the conv. layers, and 128 node dense layer.)

Teaching the machines to learn! ML on MPSoCs & FPGAs

There is a significant benefit to modern MPSoC devices:

- Execute **high-level applications** on **CPU/RPU**
- Perform **low/fixed latency operations** on **FPGA**
- Offload **simple vector/matrix operations** to **GPU**

Teaching the machines to learn! ML on MPSoCs & FPGAs

There is a significant benefit to modern MPSoC devices:

- Execute **high-level applications** on **CPU/RPU**
- Perform **low/fixed latency operations** on **FPGA**
- Offload **simple vector/matrix operations** to **GPU**

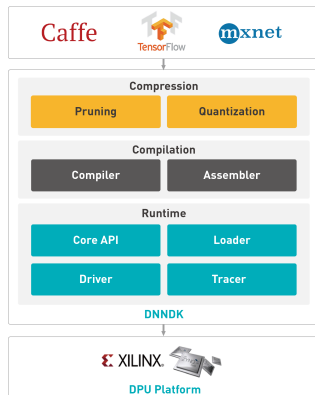
And we can execute complex ML applications using CNN directly on these devices!

Teaching the machines to learn! ML on MPSoCs & FPGAs

There is a significant benefit to modern MPSoC devices:

- Execute **high-level applications** on **CPU/RPU**
- Perform **low/fixed latency operations** on **FPGA**
- Offload **simple vector/matrix operations** to **GPU**

And we can execute complex ML applications using CNN directly on these devices!

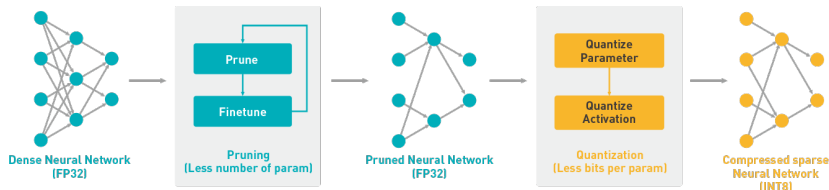
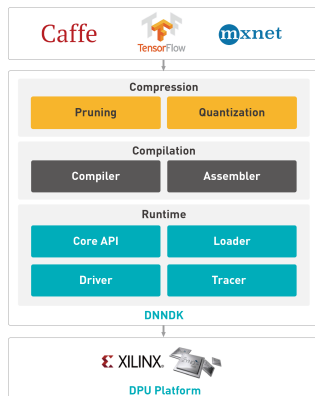


Teaching the machines to learn! ML on MPSoCs & FPGAs

There is a significant benefit to modern MPSoC devices:

- Execute **high-level applications** on **CPU/RPU**
- Perform **low/fixed latency operations** on **FPGA**
- Offload **simple vector/matrix operations** to **GPU**

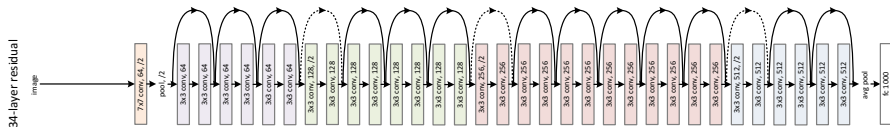
And we can execute complex ML applications using CNN directly on these devices!



Proof-of-principle with the gFEX Zynq: ResNet-50

Implement **ResNet-50** neural network for image classification on **our Zynq UltraScale+ MPSoC for gFEX**

- Dramatically larger network!
- Thousands of filters
- ~10 billion operations!!
- **merely a Proof-of-principle**



Work conducted by Emily Smith (grad student),
in collaboration with Giordon Stark (UC Santa Cruz)
and two UChicago undergraduates Jack Huang, Ben Warren.

Proof-of-principle with the gFEX Zynq: ResNet-50

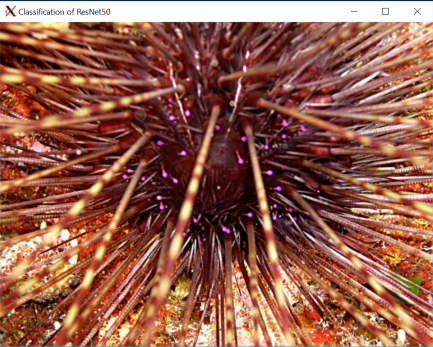
```
ksn@5@em91-pc --
Run ResNet50 CONV layers ...
DPU CONV Execution time: 13607us
DPU CONV Performance: 566.62GOPS
Run ResNet50 FC layers ...
DPU FC Execution time: 236us
DPU FC Performance: 16.9492GOPS
top[0] prob = 0.993850 name = English setter
top[1] prob = 0.001493 name = clumber, clumber spaniel
top[2] prob = 0.001493 name = Brittany spaniel
top[3] prob = 0.001163 name = English springer, English springer spaniel
top[4] prob = 0.000705 name = Great Pyrenees

Load image : 2ILSVRC2012_test_00068213.JPEG

Run ResNet50 CONV layers ...
DPU CONV Execution time: 13595us
DPU CONV Performance: 567.12GOPS
Run ResNet50 FC layers ...
DPU FC Execution time: 236us
DPU FC Performance: 16.9492GOPS
top[0] prob = 0.915599 name = rock beauty, Holocanthus tricolor
top[1] prob = 0.075157 name = king penguin, Aptenodytes patagonica
top[2] prob = 0.001768 name = anemone fish
top[3] prob = 0.000835 name = fiddler crab
top[4] prob = 0.000650 name = toucan

Load image : 2ILSVRC2012_test_00042675.JPEG

Run ResNet50 CONV layers ...
DPU CONV Execution time: 13586us
DPU CONV Performance: 567.496GOPS
Run ResNet50 FC layers ...
DPU FC Execution time: 235us
DPU FC Performance: 17.0213GOPS
top[0] prob = 0.977076 name = jaguar, panther, Panthera onca, Felis onca
top[1] prob = 0.017896 name = leopard, Panthera pardus
top[2] prob = 0.000891 name = tiger, Panthera tigris
top[3] prob = 0.000540 name = cheetah, chetah, Acinonyx jubatus
top[4] prob = 0.000540 name = tiger cat
```



The image shows a window titled "Classification of ResNet50" displaying a close-up photograph of a sea urchin. The urchin's spines are dark purple and brown, with some spines having bright yellow and orange tips. The background is a mix of brown and orange, suggesting a rocky or sandy seabed. The window is overlaid on a terminal window showing the execution of ResNet-50 on three different images, with performance metrics and classification results for each.

Image processing at the level of $\mathcal{O}(\text{ms})$, expected to decrease to $\mathcal{O}(\mu\text{s})$ for jet network and 30×30 “images” (i.e. gFEX events).

Outline

- 1 *Challenges of the Energy and Luminosity Frontier*
- 2 *ATLAS Phase I & II Hadronic Trigger Systems*
- 3 *Machine learning using FPGAs and MPSoCs*
- 4 *Summary and conclusions*

Summary

- **Major challenge** to measurements and searches in hadronic final states at the future LHC will be **triggering and data management**
- Run 3 trigger systems (gFEX) have **unique and novel capabilities** as part of both baseline design and **ML on MPSoC & FPGAs**
 - Co-processor applications using FPGA+CPU+GPU would be very interesting!
- Clear **opportunities for the Phase II trigger system** in terms of hadronic final state physics, tracking, and more for the trigger system currently planned
 - *There is much to be explored in Hardware-based Track Triggers for HLT!*
- **Strong involvement with scalable systems, hardware accelerators, and even data management plans for the “offline” world may be essential to realize gains further in physics potential**

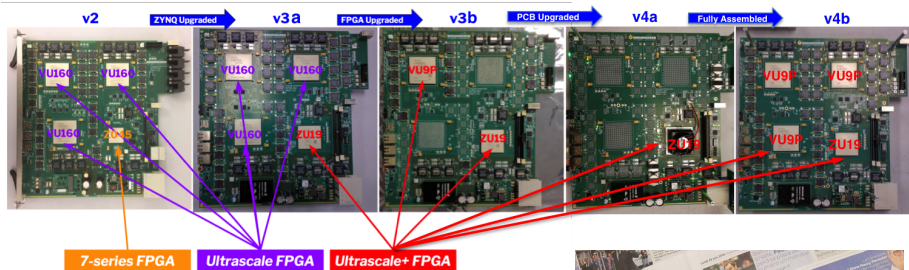
Outline

5 *Bonus material*

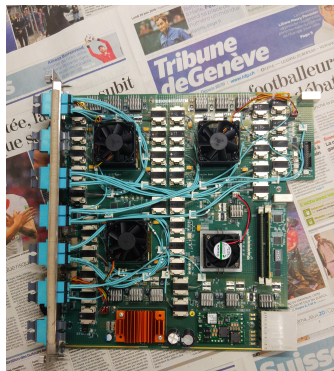
Appendix

5 *Bonus material*

gFEX prototypes and production boards



- **Prototype** (1×VU9P + 1×ZU19) used for integration and commissioning at CERN since Q1 2018.
- **Final board** (3×VU9P + 1×ZU19) delivered to CERN on 25 June, 2018, 5 years from proposal to delivery!
- **Installation** ~now, ready for Run 3



gFEX Multi-Processor System-on-Chip: Zynq Ultrascale+

ARM[®] Application Processors
Cortex[™] A53
64-bit Quad-Core with Virtualization



Power Management
Multiple Power Domains
Power Gated Islands

ARM[®] Real-Time Processors
Cortex[™] R5
32-bit Dual-Core
Application Offload



Safety & Reliability
IEC61508, ISO26262
System Isolation &
Error Mitigation, Lockstep

mali[™] Graphics/Video
H.265 HEVC
ARM Mali-400MP
H.265/264 CODECS



Security
Information Assurance,
Trust, Anti-Tamper, TrustZone
Key and Vault Management



UltraScale
FPGA Logic
UltraRAM, PCIe Gen4,
100G Ethernet, AMS

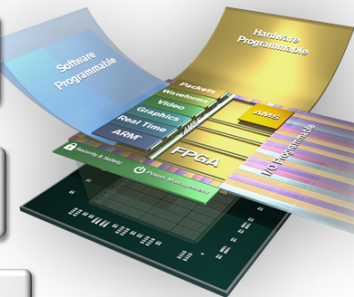


Runtime SW & Tools

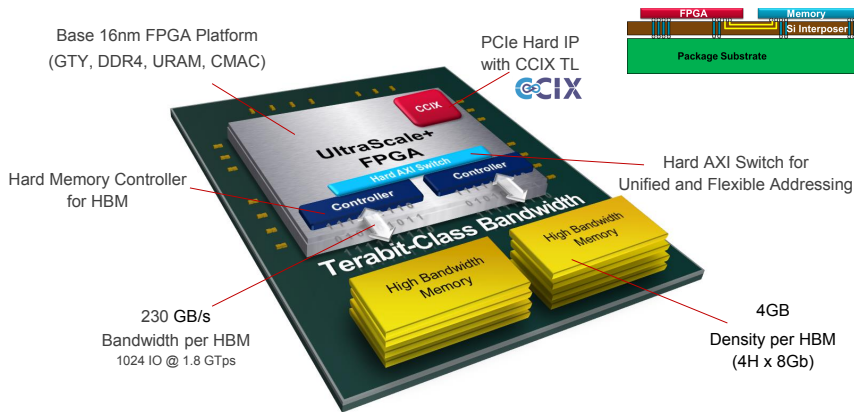
OS, RTOS, AMP, Hypervisor
Development, Heterogeneous Debug,
Hardware/Software Profiling &
Performance Analysis



High Speed
Peripherals
USB 3.0, PCIe Gen2, GbE
SATA3.0, DisplayPort



gFEX Virtex 7 Ultrascale+ Processor FPGAs



gFEX Multi-Processor System-on-Chip: Zynq Ultrascale+

ARM[®] Application Processors
Cortex[™] A53
64-bit Quad-Core with Virtualization



Power Management
Multiple Power Domains
Power Gated Islands

ARM[®] Real-Time Processors
Cortex[™] R5
32-bit Dual-Core
Application Offload



Safety & Reliability
IEC61508, ISO26262
System Isolation &
Error Mitigation, Lockstep

mali
H.265 HEVC
High Efficiency Video Coding
Graphics/Video
ARM Mali-400MP
H.265/264 CODECS



Security
Information Assurance,
Trust, Anti-Tamper, TrustZone
Key and Vault Management



UltraScale
FPGA Logic
UltraRAM, PCIe Gen4,
100G Ethernet, AMS

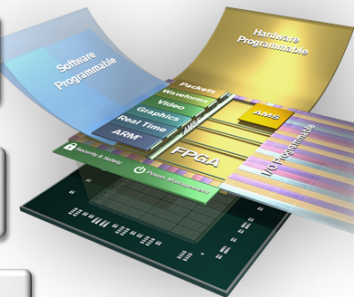


Runtime SW & Tools

OS, RTOS, AMP, Hypervisor
Development, Heterogeneous Debug,
Hardware/Software Profiling &
Performance Analysis

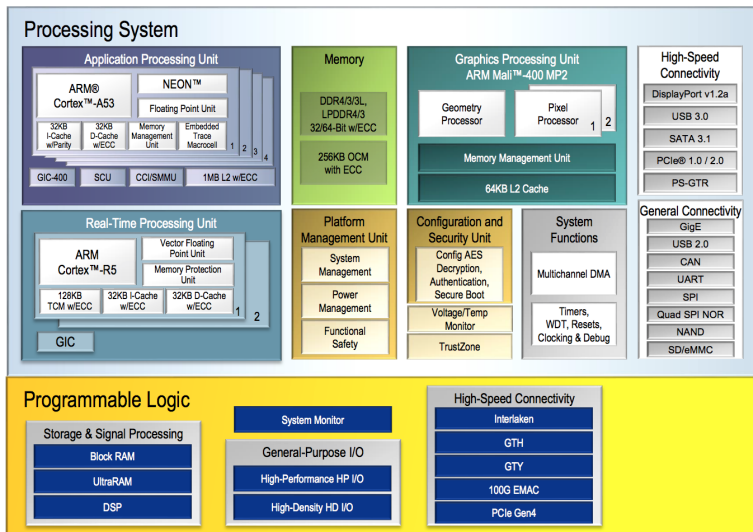


High Speed
Peripherals
USB 3.0, PCIe Gen2, GbE
SATA3.0, DisplayPort

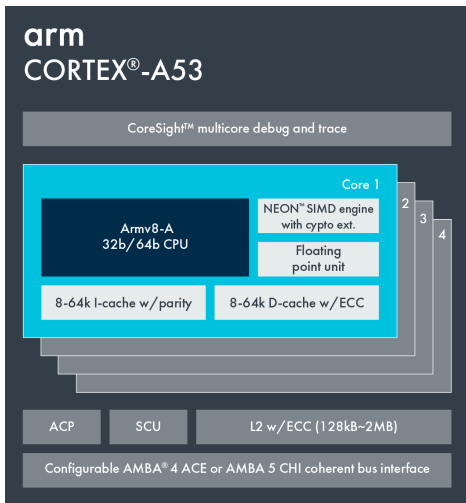


gFEX Multi-Processor System-on-Chip: Zynq UltraScale+

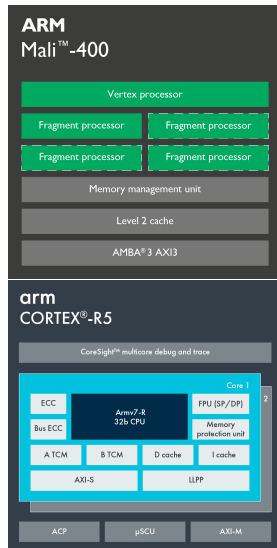
Zynq® UltraScale+™ MPSoCs: EG Block Diagram



Zynq Ultrascale+ Processors



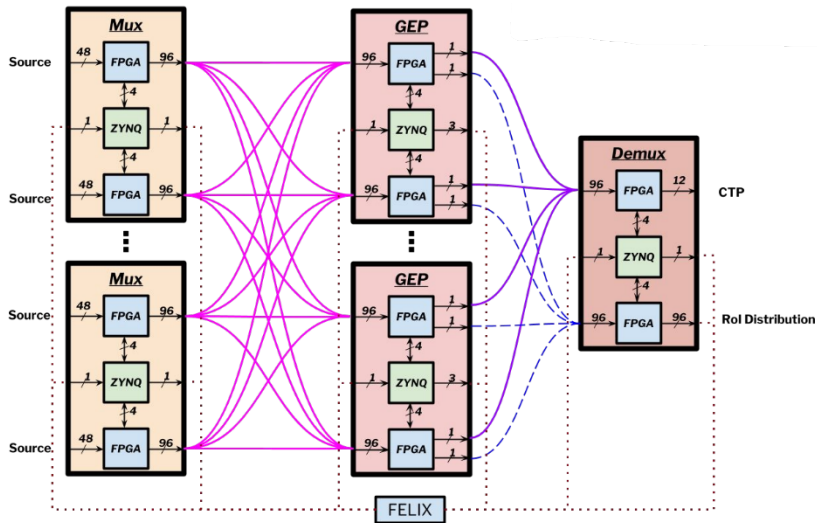
64 bit ARM quad-core processor



Processor and Zynq FPGA comparison for gFEX boards

gFEX version	Processor FPGA			Zynq	
	v1	v2/v3	v3/v4	v1/v2	v3/v4
FPGA type	VX690T	VU160	VU9P	Z7045	ZU19
Logic Cells (M)	0.7	2.0	2.6	0.4	1.1
CLB (M)	0.9	1.9	2.4	0.3	1.0
Total RAM (Mb)	52.9	115.2	345.9	17.6	70.6
DSP slices (K)	3.6	1.6	6.8	0.4	2.0

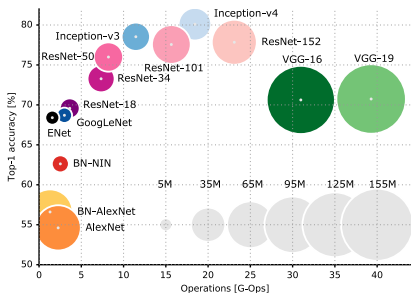
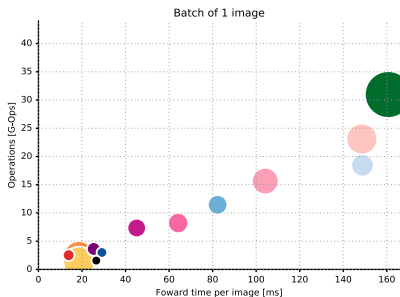
Global Event Processor Information



Zynq MPSoC also available on future Phase II trigger system.

Industrial neural networks and ResNet-50

From Canziani, Culurciello, Paszke “An Analysis of Deep Neural Network Models for Practical Applications” (arXiv:1605.07678)



“operations count represent a good estimation of inference time.”