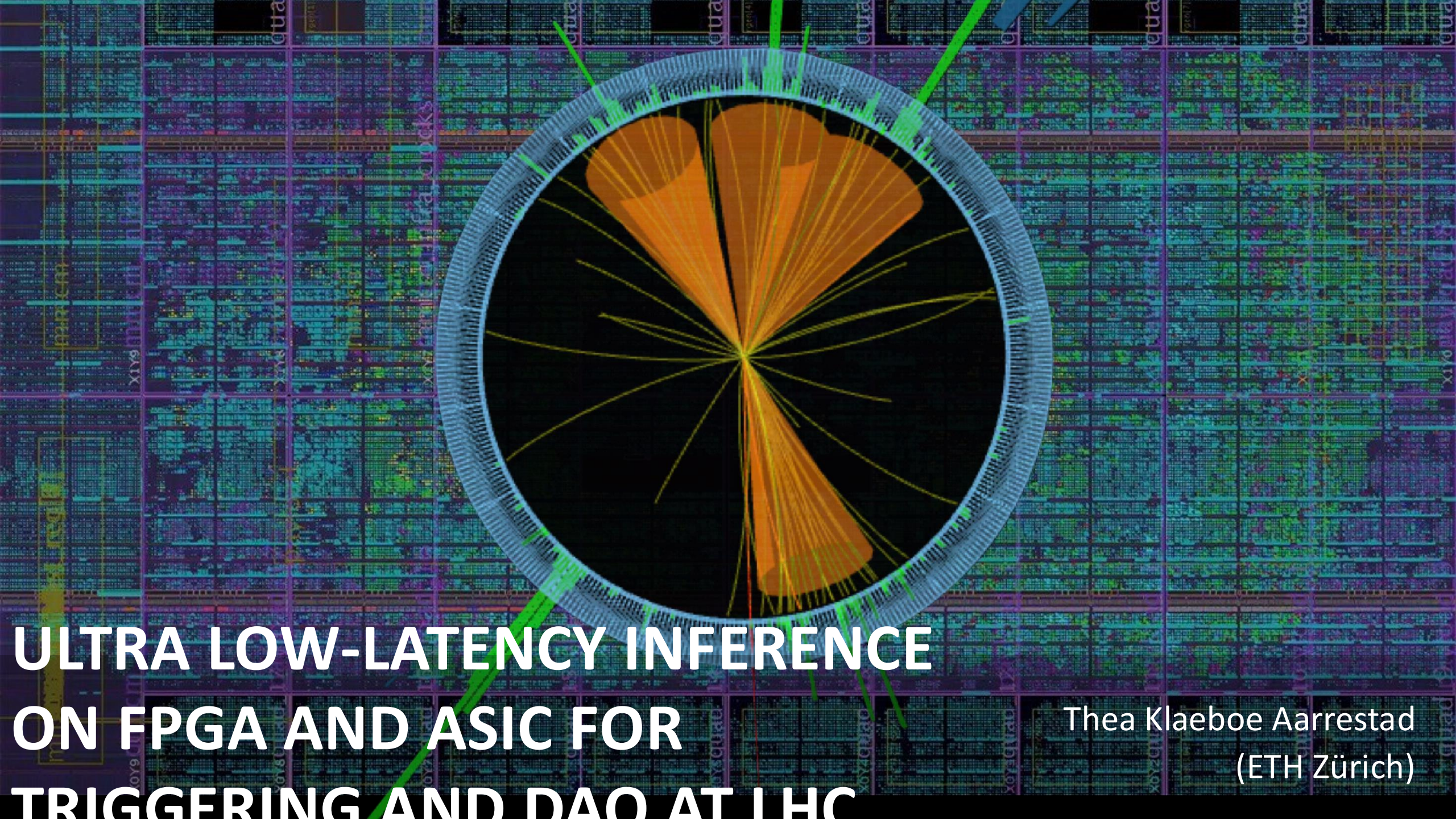
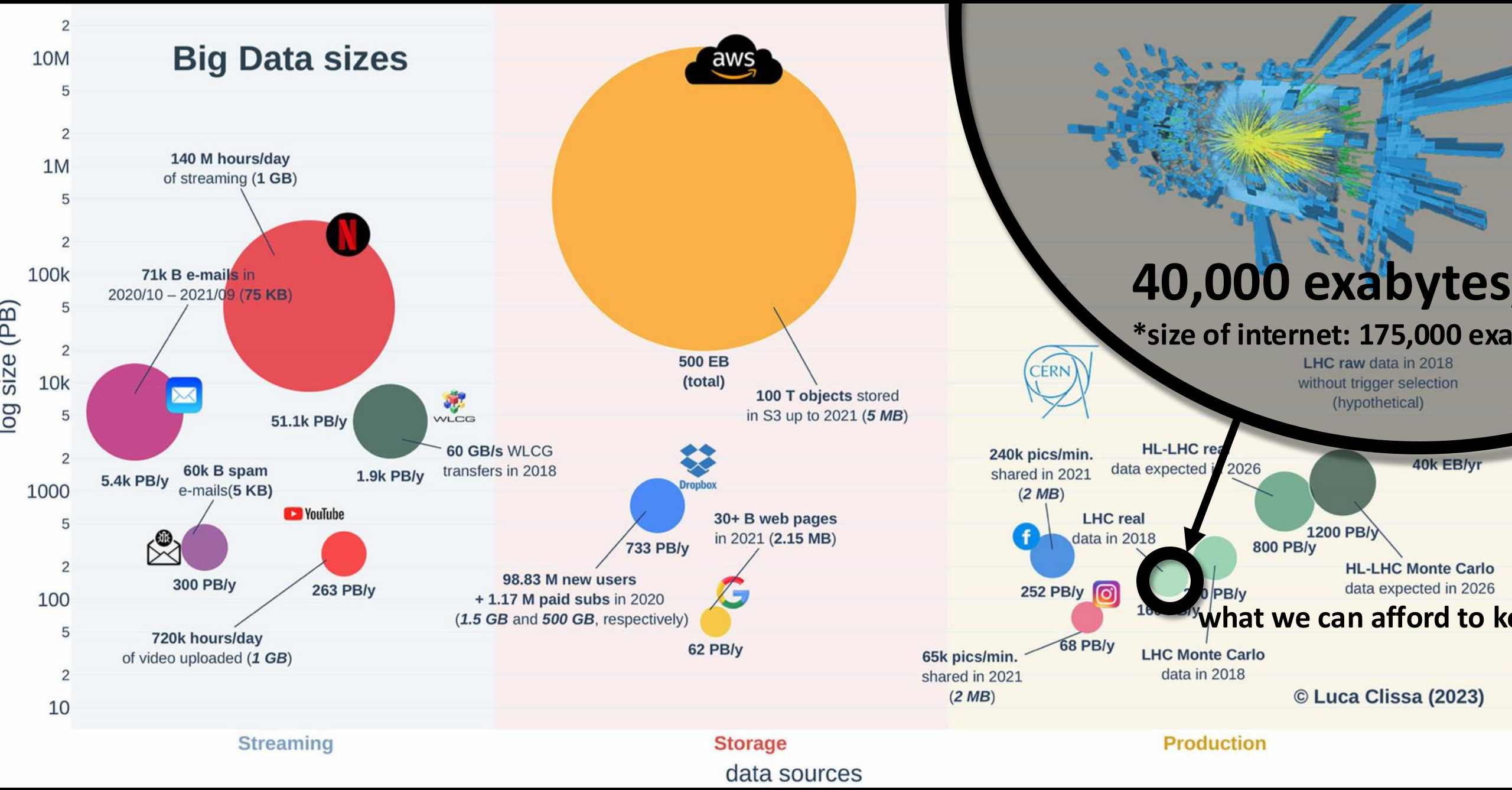
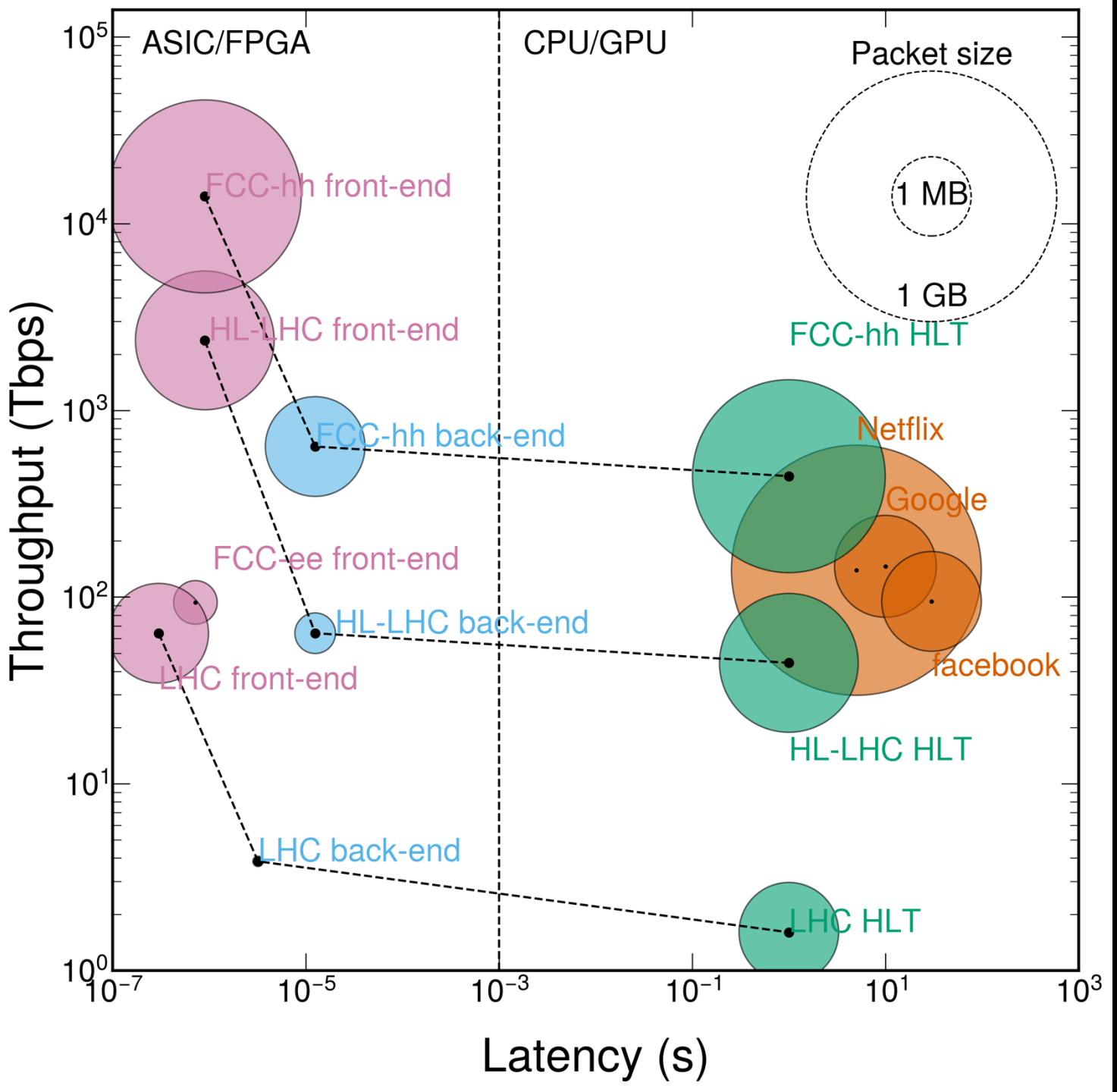


ULTRA LOW-LATENCY INFERENCE ON FPGA AND ASIC FOR TRIGGERING AND DAQ AT LHC

Thea Klaeboe Aarrestad
(ETH Zürich)


Big Data sizes
log size (PB)
2
10M
5
2
1M
5
2
100k
5
2
10k
5
2
1000
5
2
100
5
2
10
140 M hours/day
of streaming (1 GB)
71k B e-mails in
2020/10 – 2021/09 (75 KB)
5.4k PB/y
60k B spam
e-mails(5 KB)
300 PB/y
720k hours/day
of video uploaded (1 GB)
263 PB/y
51.1k PB/y
1.9k PB/y
60 GB/s WLCG
transfers in 2018
500 EB
(total)
100 T objects stored
in S3 up to 2021 (5 MB)
733 PB/y
98.83 M new users
+ 1.17 M paid subs in 2020
(1.5 GB and 500 GB, respectively)
30+ B web pages
in 2021 (2.15 MB)
62 PB/y
40,000 exabytes
*size of internet: 175,000 exa
LHC raw data in 2018
without trigger selection
(hypothetical)
240k pics/min.
shared in 2021
(2 MB)
252 PB/y
65k pics/min.
shared in 2021
(2 MB)
68 PB/y
LHC real
data in 2018
LHC Monte Carlo
data in 2018
800 PB/y
1200 PB/y
40k EB/yr
HL-LHC Monte Carlo
data expected in 2026
what we can afford to k
© Luca Clissa (2023)
Streaming
Storage
Production
data sources

How much data goes through our system



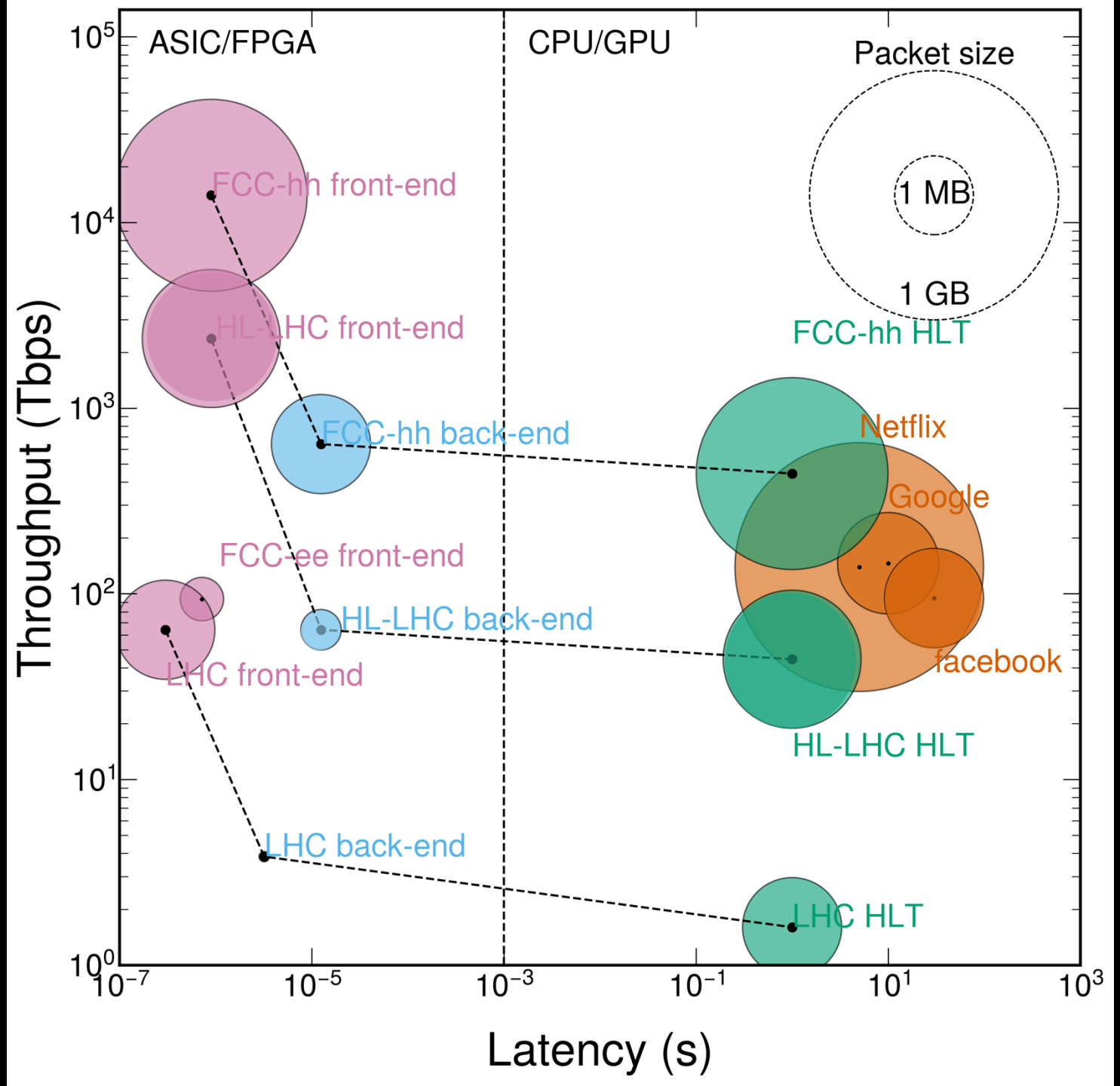
unit size of transmitted data (e.g., event size, webpage size)

Time to process

arXiv:2511.03883

ON-DETECTOR

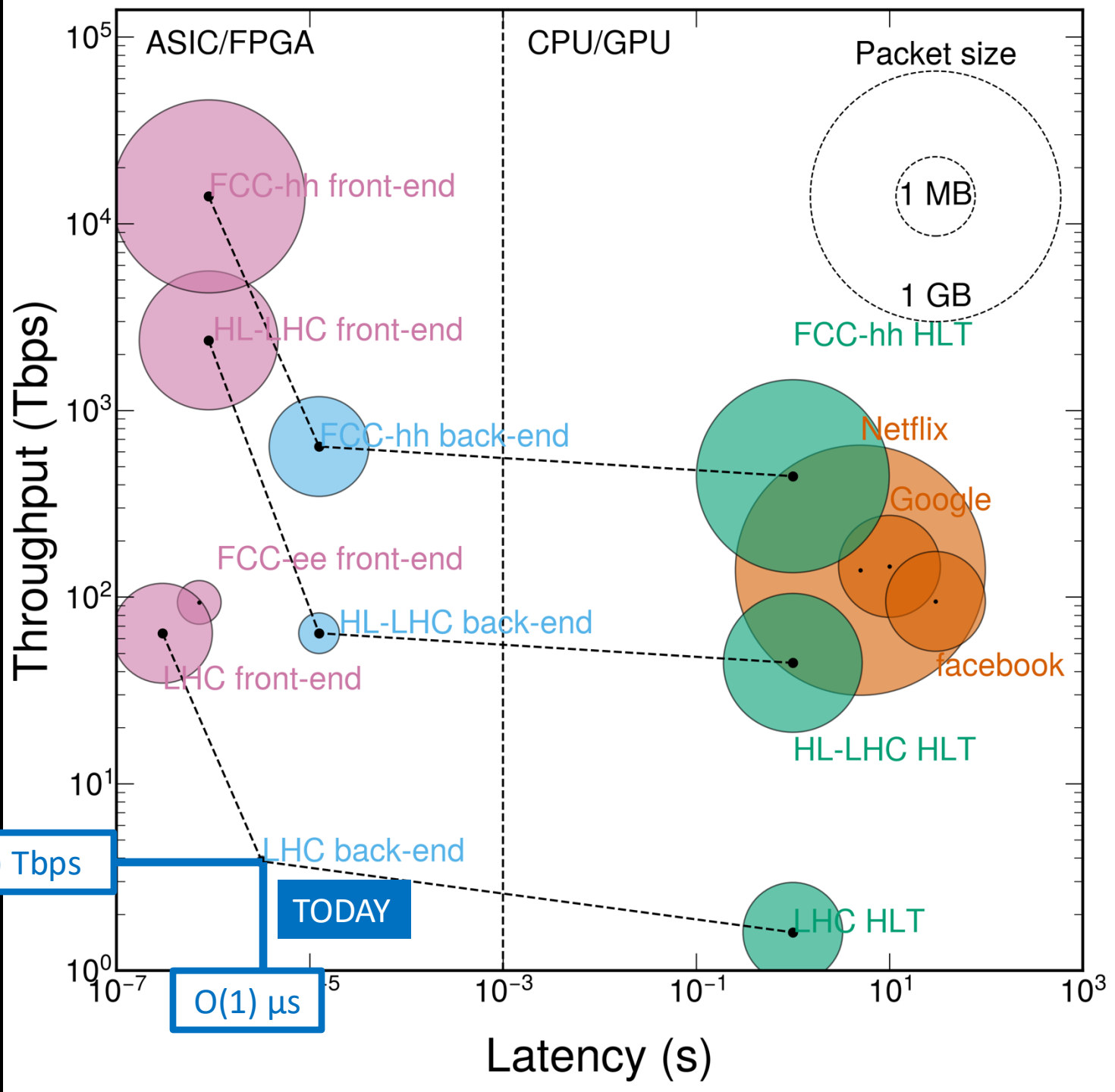
HARDWARE TRIGGERS



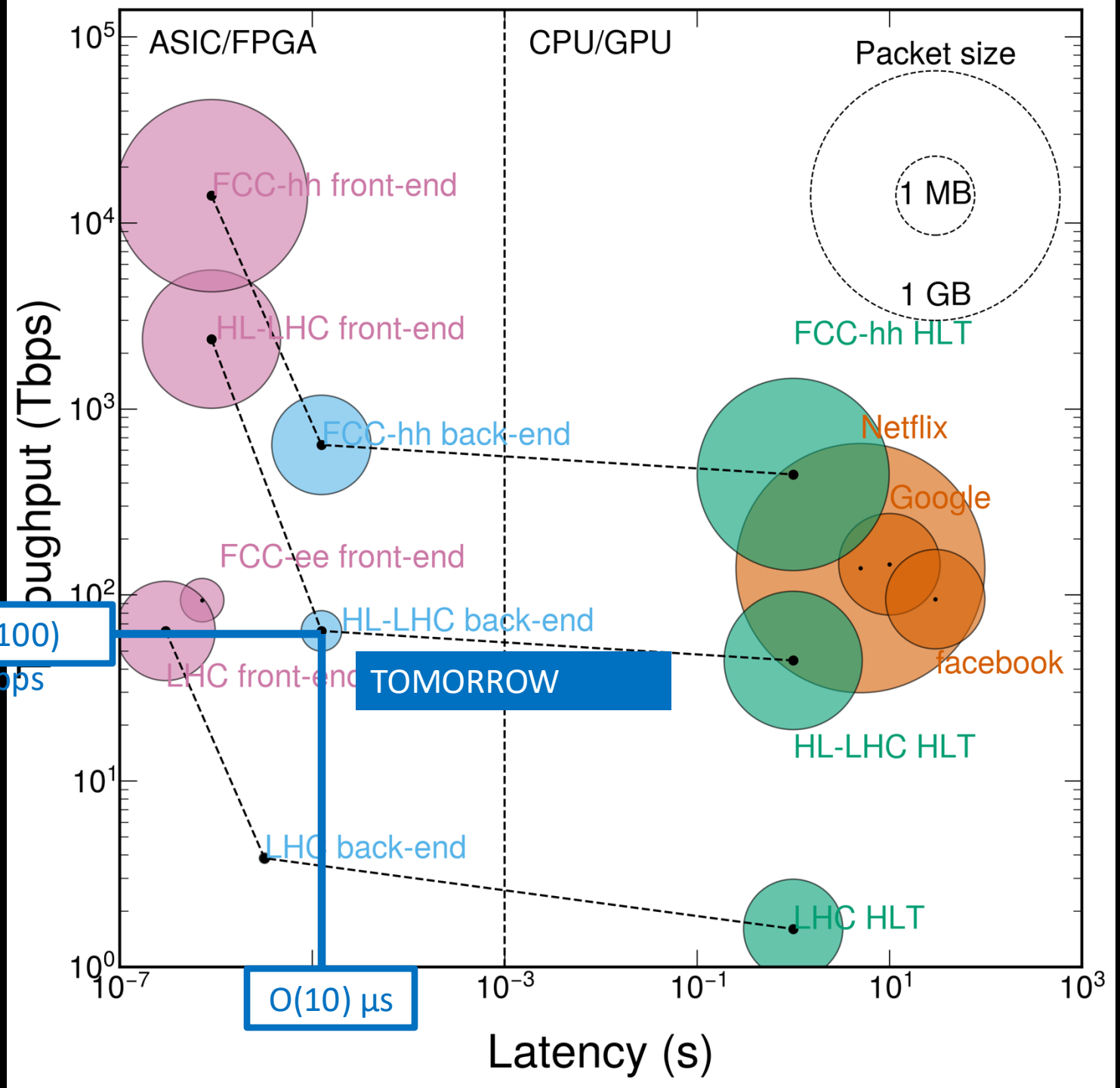
INDUSTRY

SOFTWARE TRIGGERS

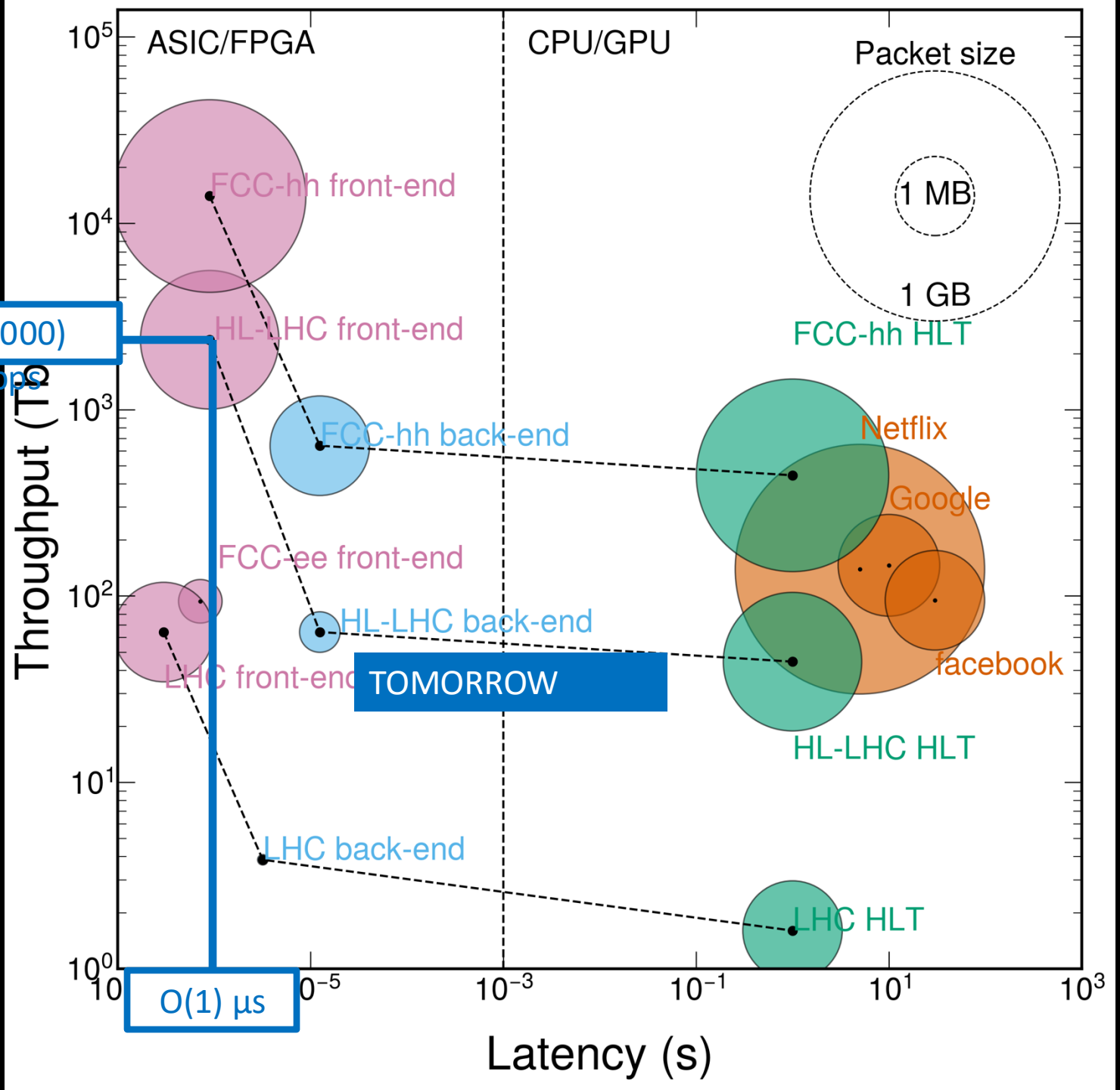
How did we get to operating 50 nanosecond ML triggers?



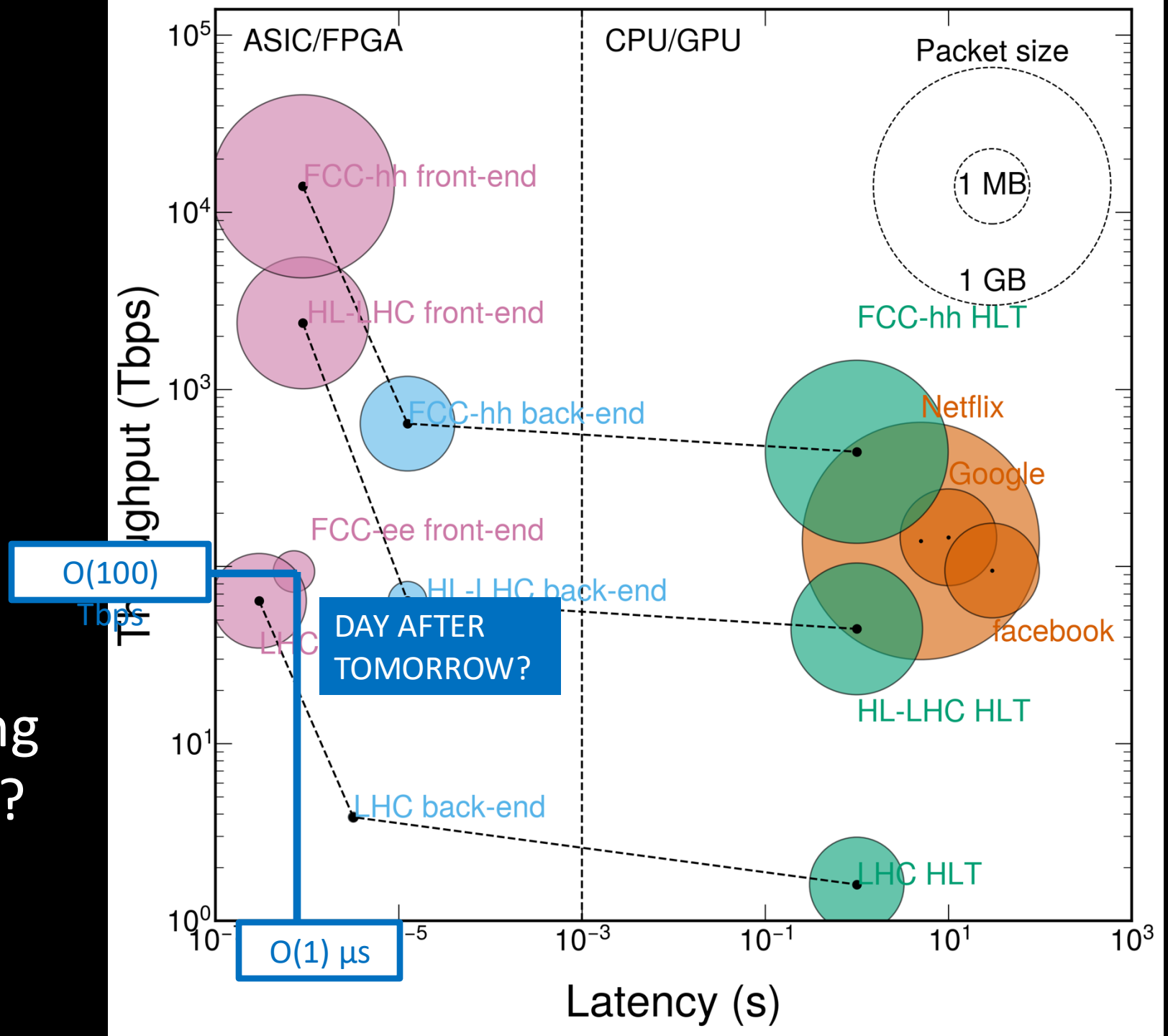
How will we maintain physics performance with x10 higher throughput?

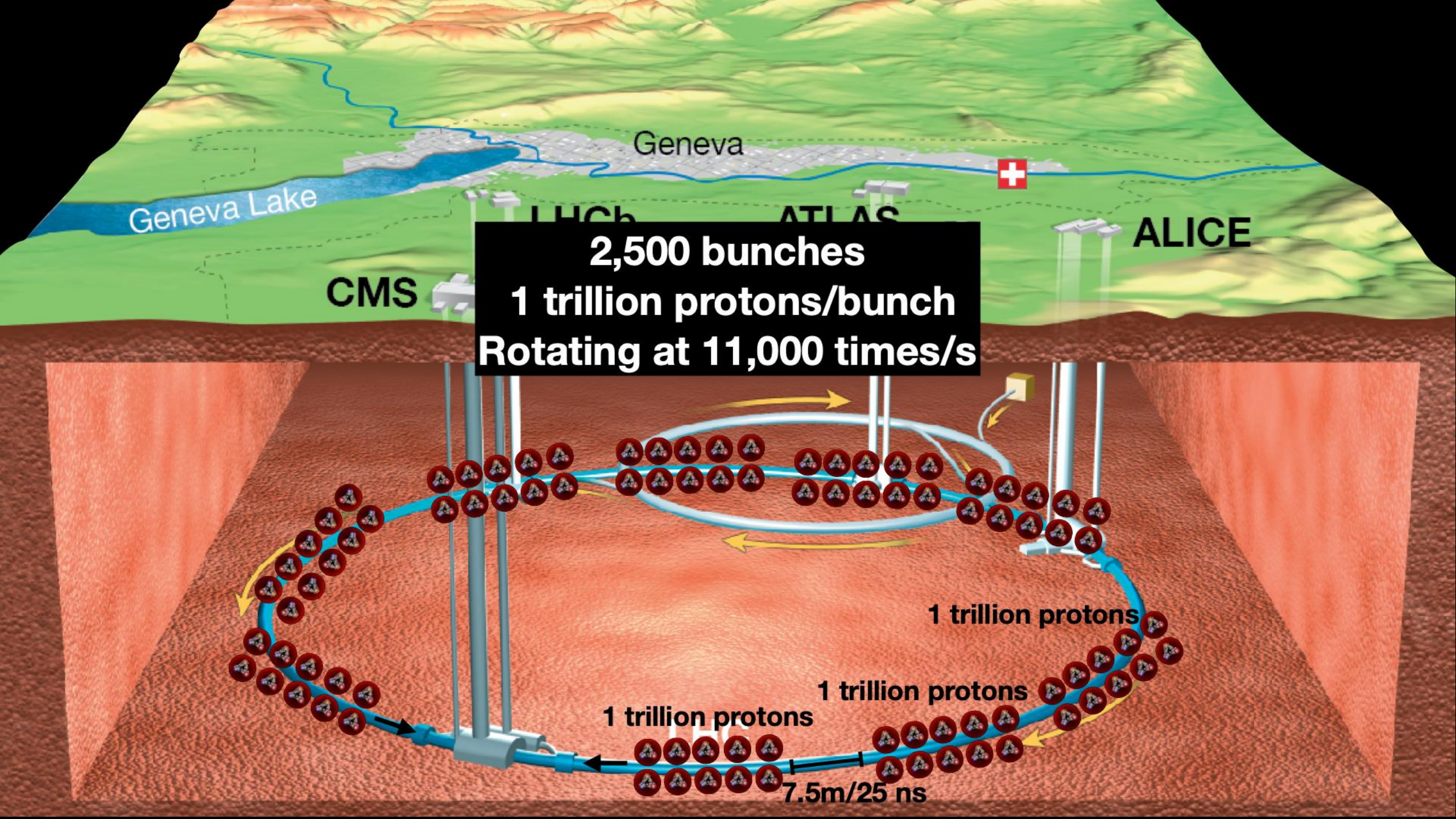


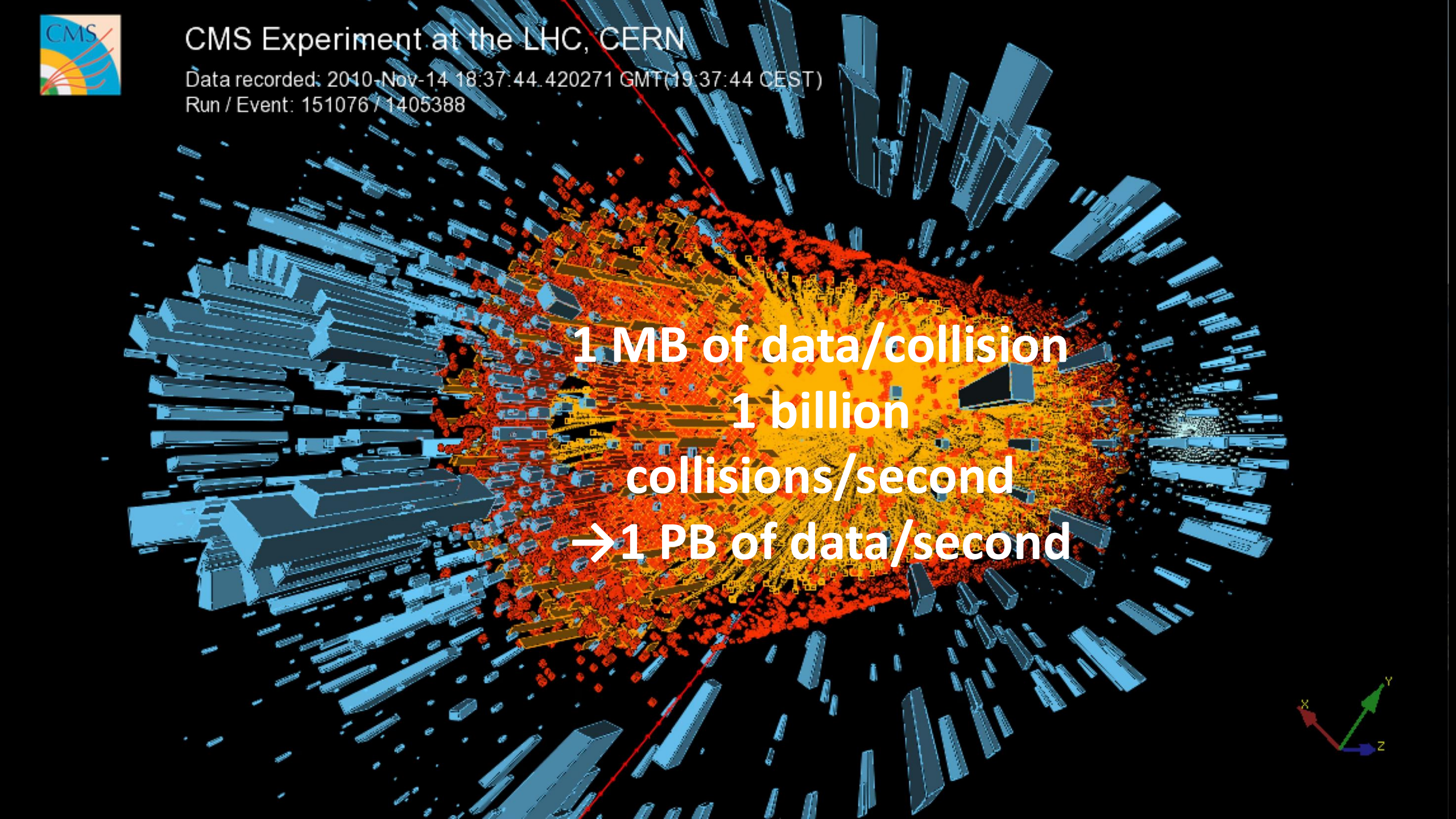
How are we getting ML in our front-ends to handle increasing granularity?

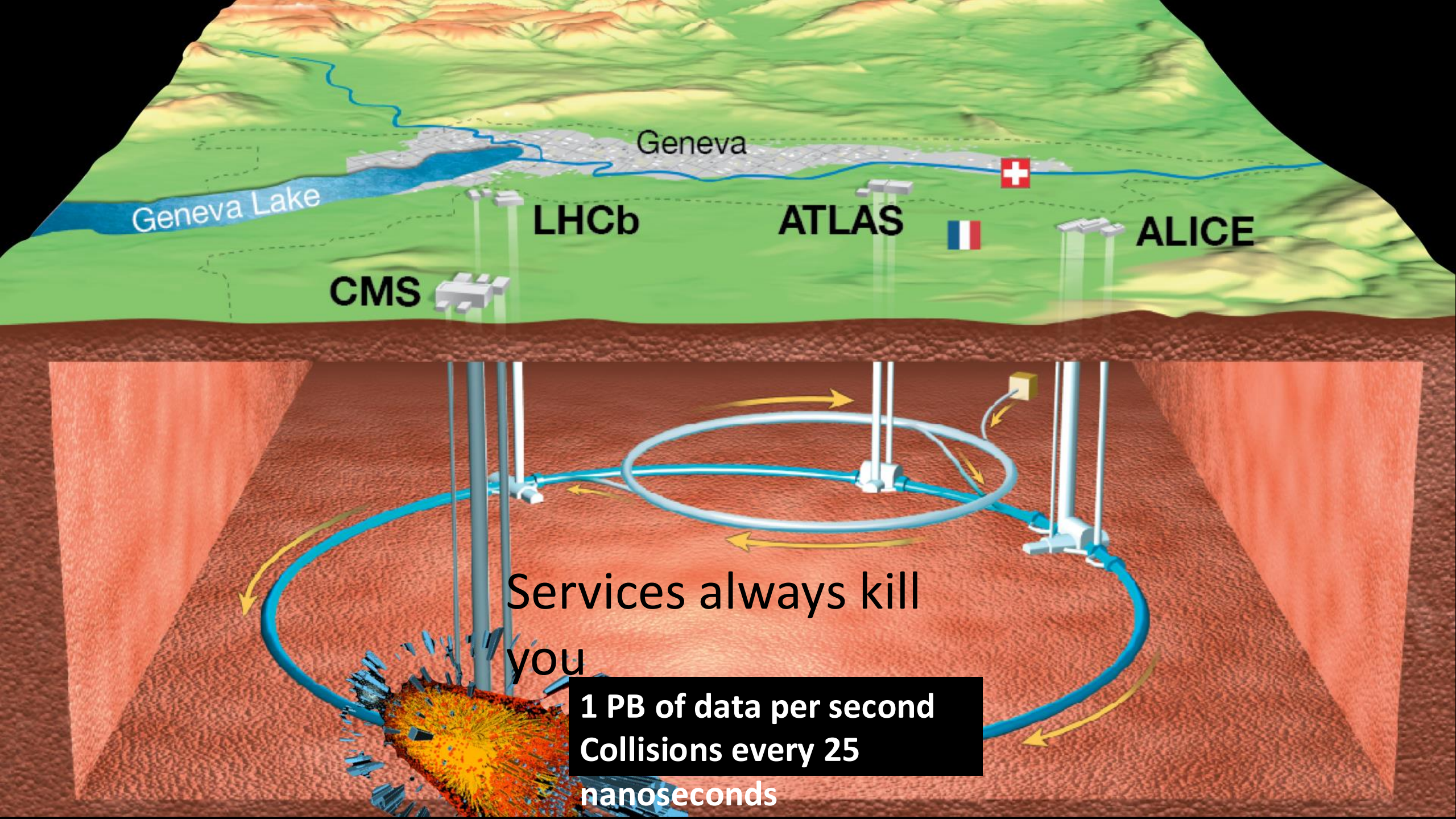


Streaming readout?



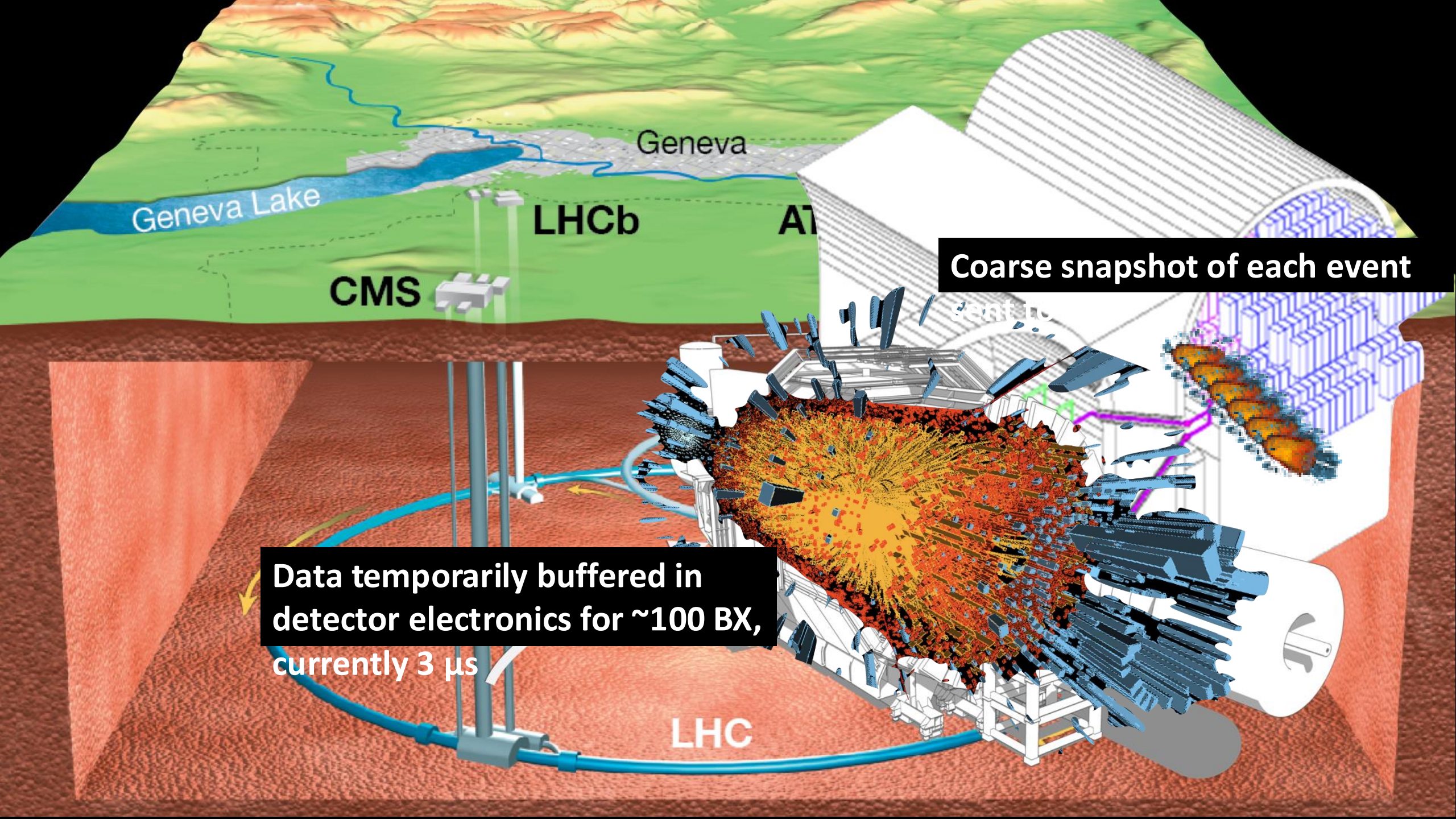

Geneva
Geneva Lake
ATLAS
ALICE
CMS
2,500 bunches
1 trillion protons/bunch
Rotating at 11,000 times/s
1 trillion protons
1 trillion protons
1 trillion protons
7.5m/25 ns

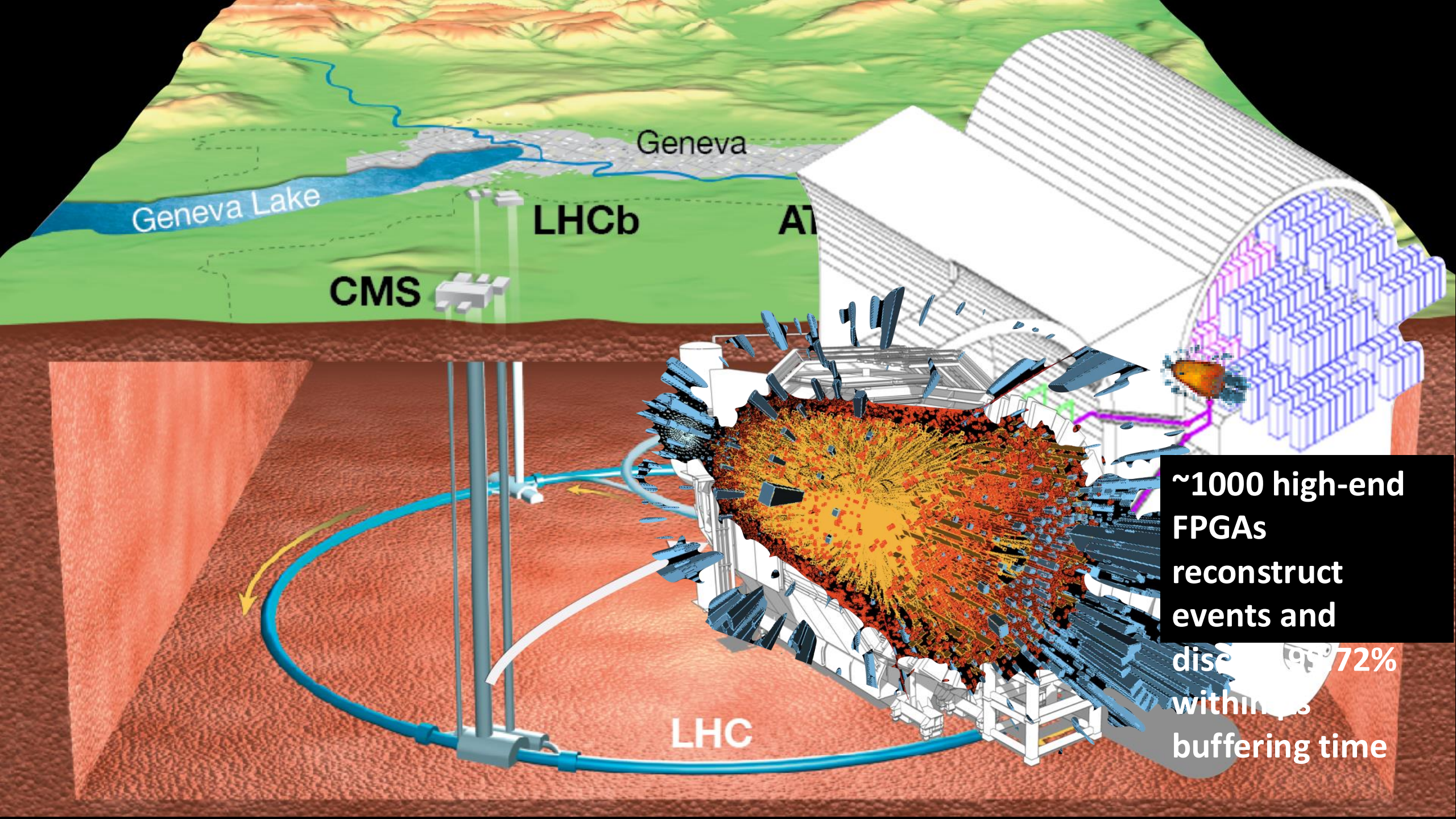

CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)
Run / Event: 151076 / 1405388
1 MB of data/collision
1 billion
collisions/second
→1 PB of data/second

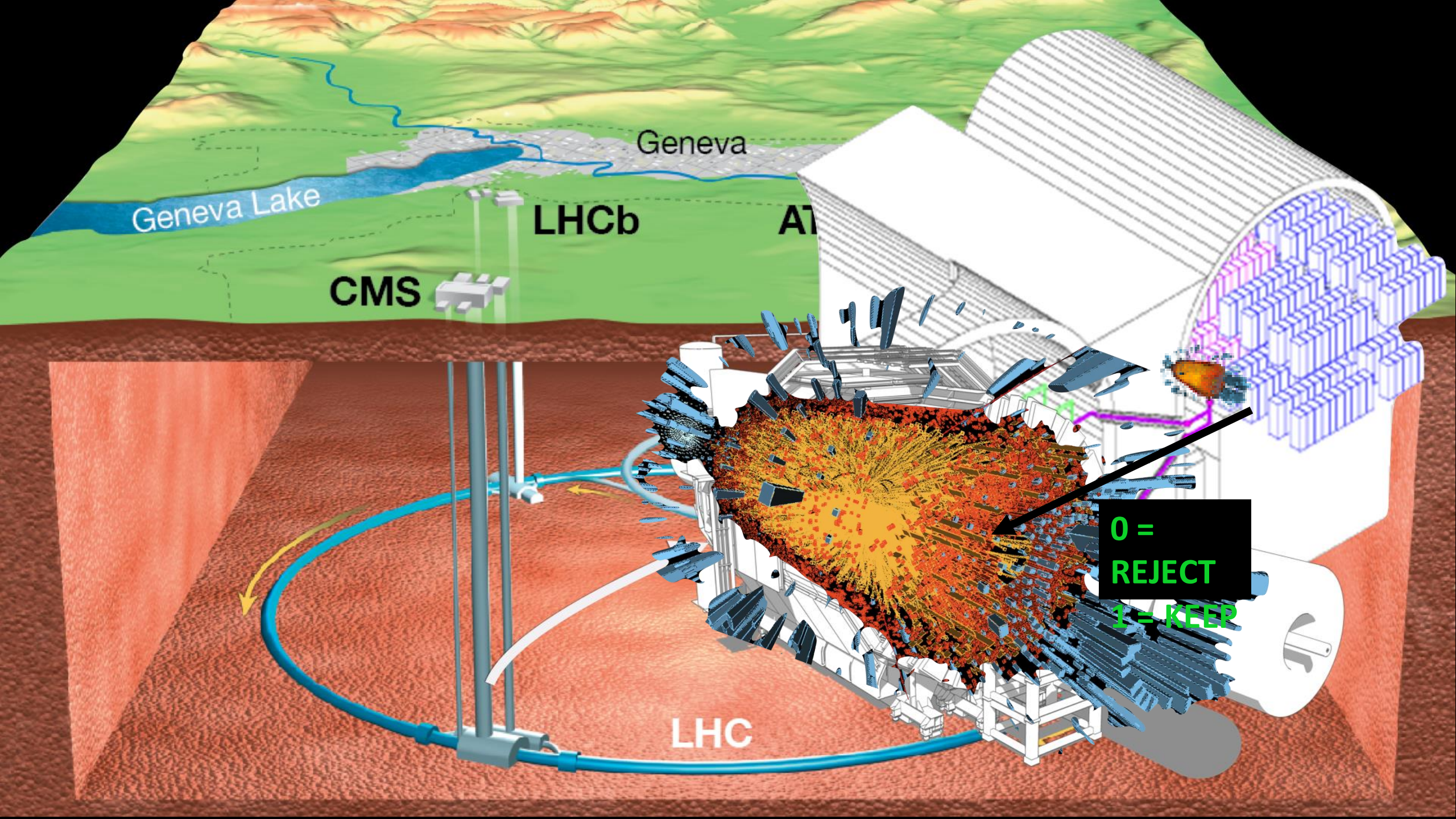


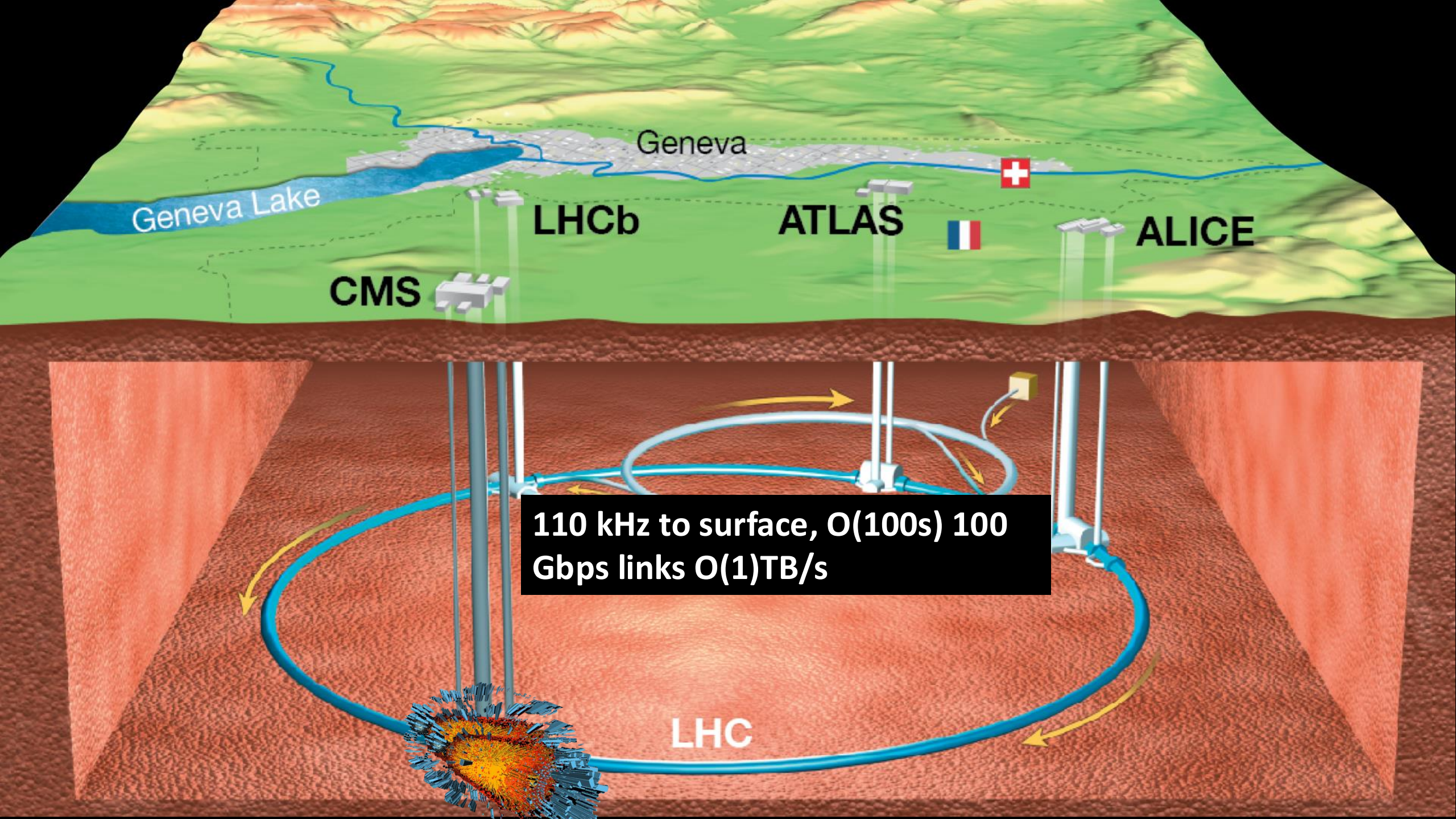
Services always kill you

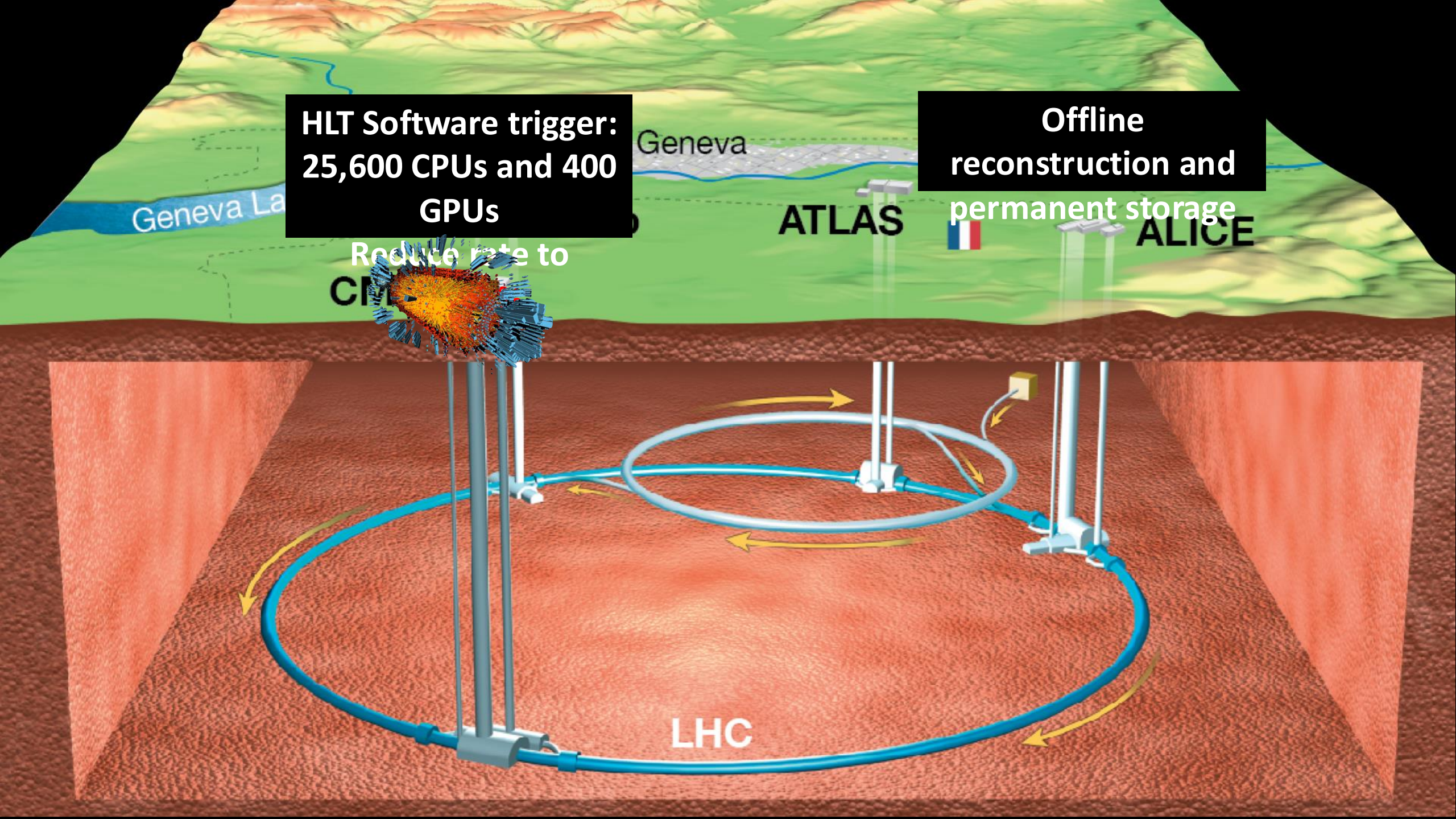
1 PB of data per second
Collisions every 25 nanoseconds


Geneva
Geneva Lake
LHCb
CMS
LHC
Coarse snapshot of each event
Data temporarily buffered in detector electronics for ~100 BX, currently 3 µs


Geneva
Geneva Lake
LHCb
CMS
LHC
~1000 high-end FPGAs reconstruct events and
buffering time

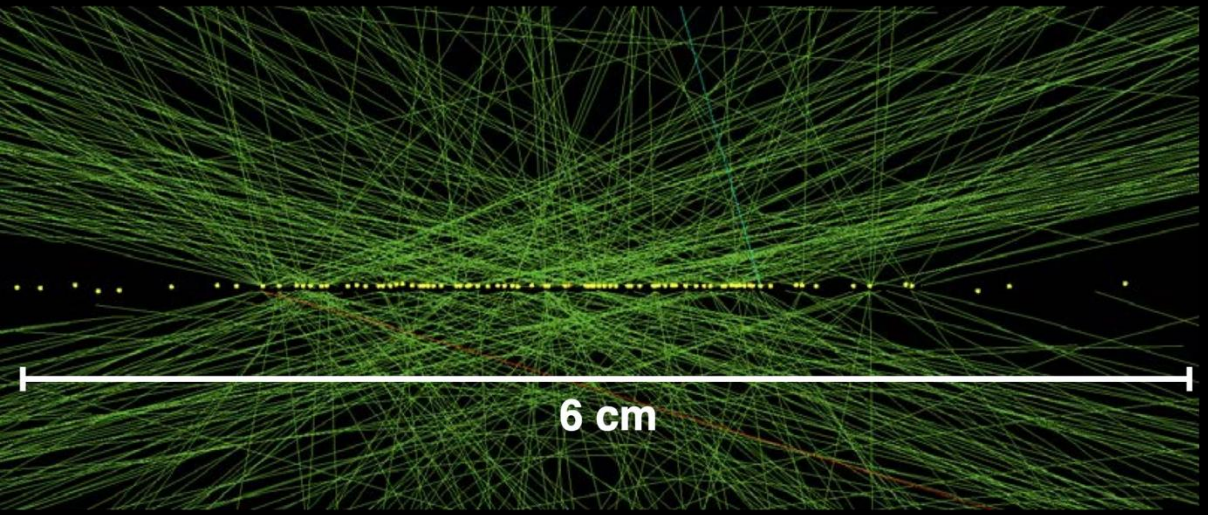

Geneva
Geneva Lake
LHCb
CMS
LHC
0 =
REJECT
1 = KEEP


Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
110 kHz to surface, O(100s) 100 Gbps links O(1)TB/s
LHC


HLT Software trigger: 25,600 CPUs and 400 GPUs
Offline reconstruction and permanent storage
Reduce rate to
Geneva
ATLAS
ALICE
LHC

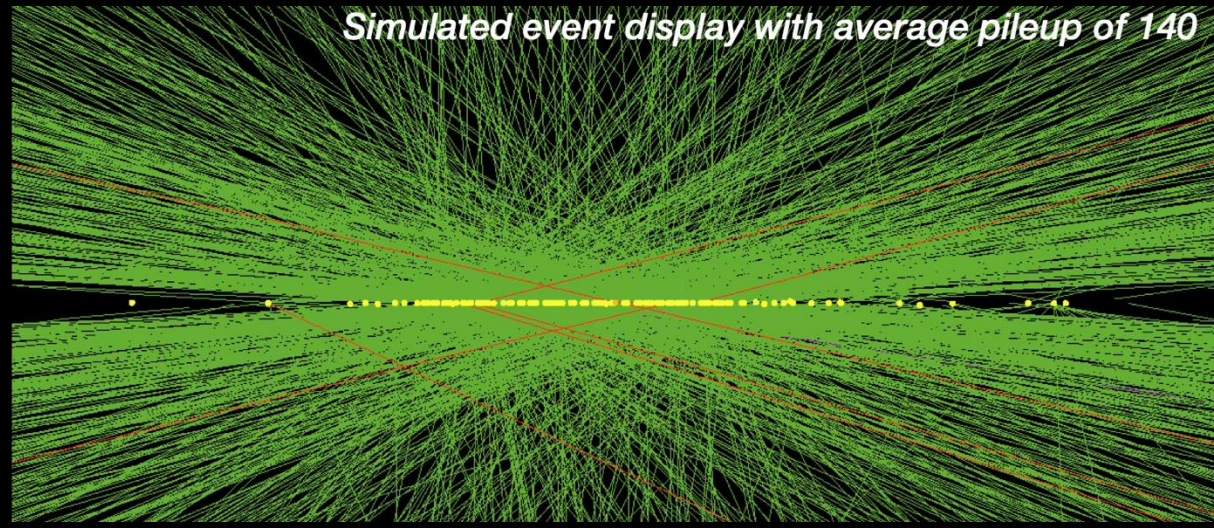
LHC

78 vertices (average 60)

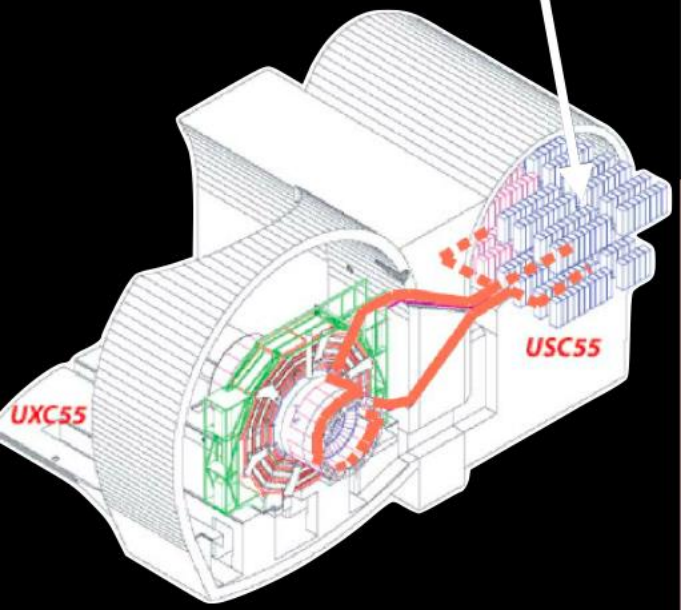


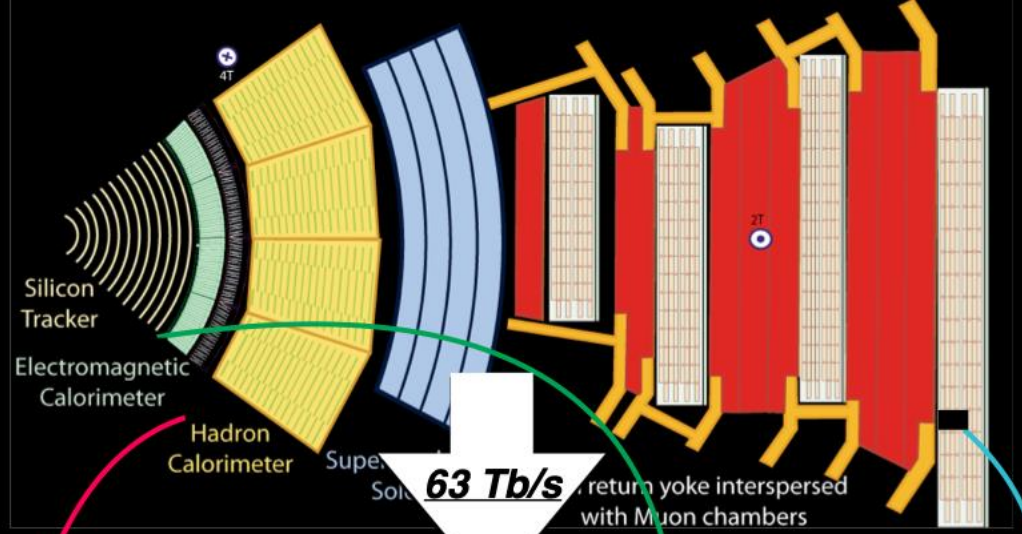
HL-LHC

200 vertices (average 140)



Need to upgrade detectors and algorithms to maintain physics performance in challenging environment





Input: 2 Tb/s → 63 Tb/s
Latency: 4 μs → 12 μs
Output: 100kHz→750 kHz

4 Tb/s

CALORIMETRY:
370 FPGAs
*54 for HGCAL only!

Xilinx Ultrascale+ FPGAs

TRACKING
174 FPGAs

6.4 Tb/s

MUONS:
96 FPGAs 5 μs

PARTICLE FLOW:
66 FPGAs

CALORIMETRY

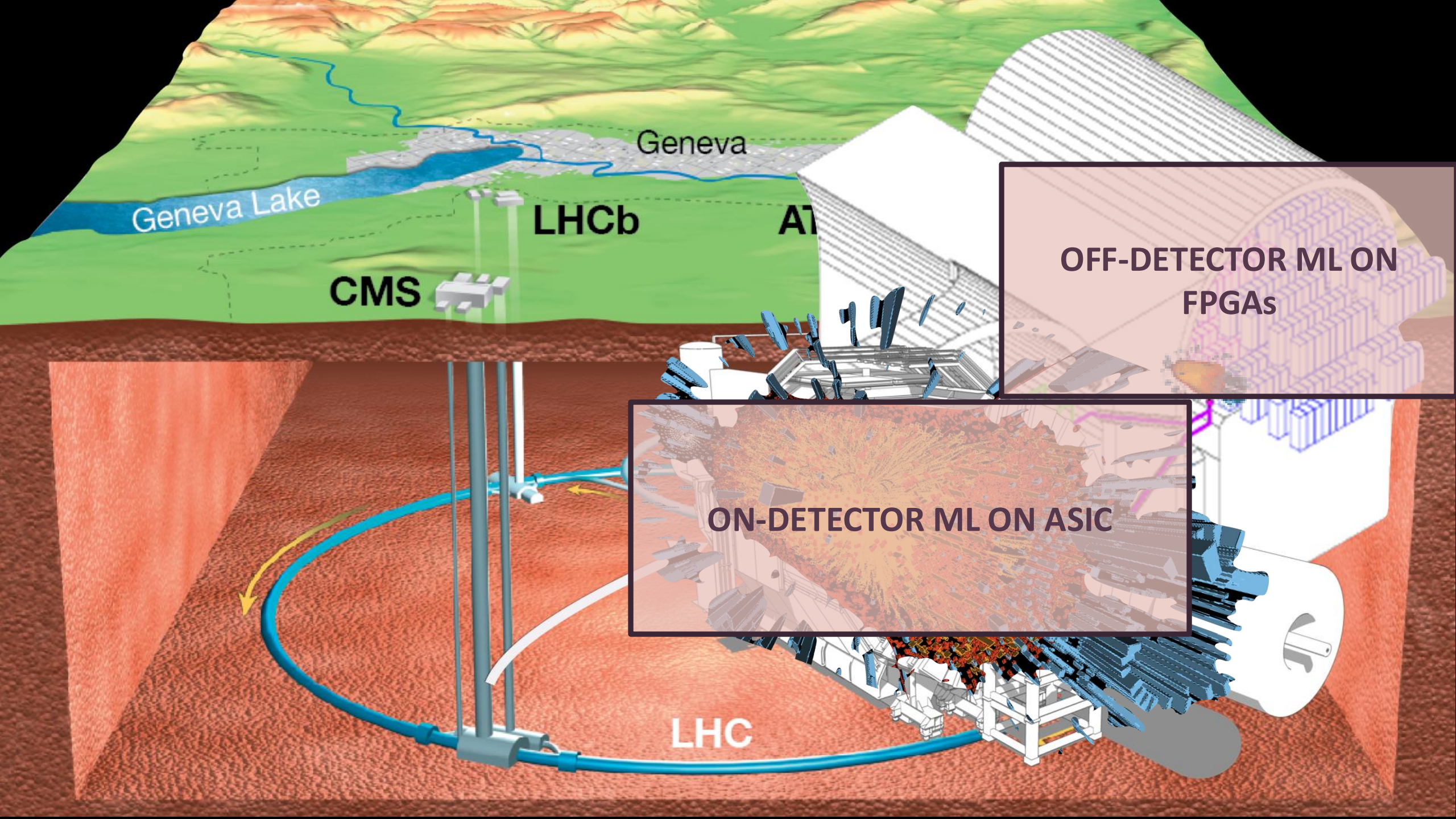
PARTICLE FLOW

Will read this out parasitically at 8Tb/s!

GLOBAL TRIGGER:
12 FPGAs

EXTERNAL TRIGGERS

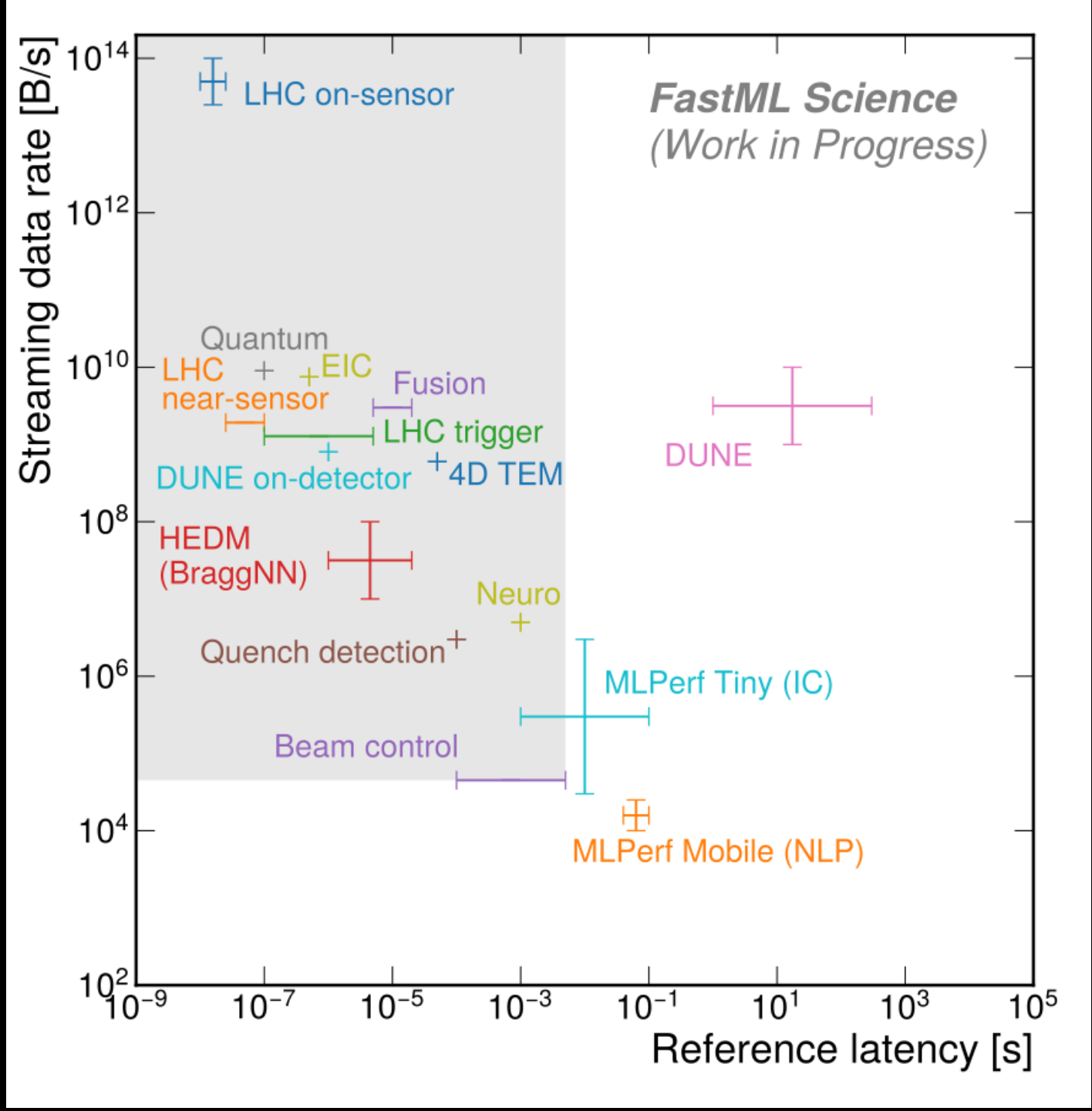
Trigger accept/reject
12.5 μs


Geneva
Geneva Lake
LHCb
CMS
LHC
OFF-DETECTOR ML ON FPGAs
ON-DETECTOR ML ON ASIC

Fast Machine Learning for HEP

Our tasks not well-represented by industry-driven benchmarks

Must develop our own tools, datasets and data challenges



ML inference at low latency

HEP Tools and Communities

Co-processing kernel
(Xilinx accelerators/SoCs)

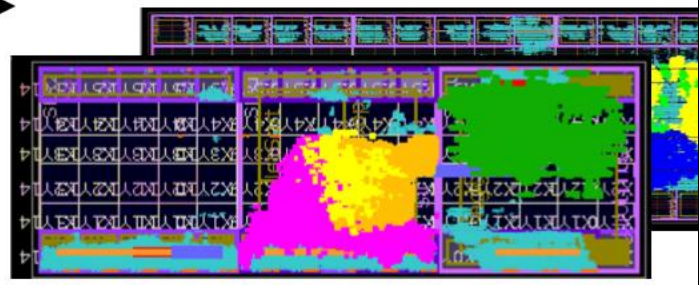
Model
(quantized/pruned)

HEP hardware ML libraries:

hls4ml

Conifer

FPGA custom designs
(eg trigger algorithms)



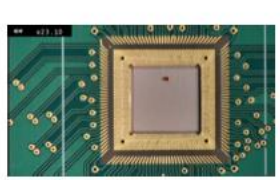
HEP quantization libraries:

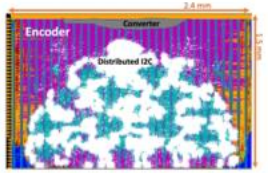
QONNX

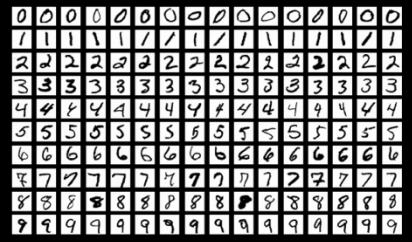
HGQ

QKERAS

ASICs







Conv2D

ReLU

MaxPool2D

Conv2D

ReLU

MaxPool2D

Flatten

Dense

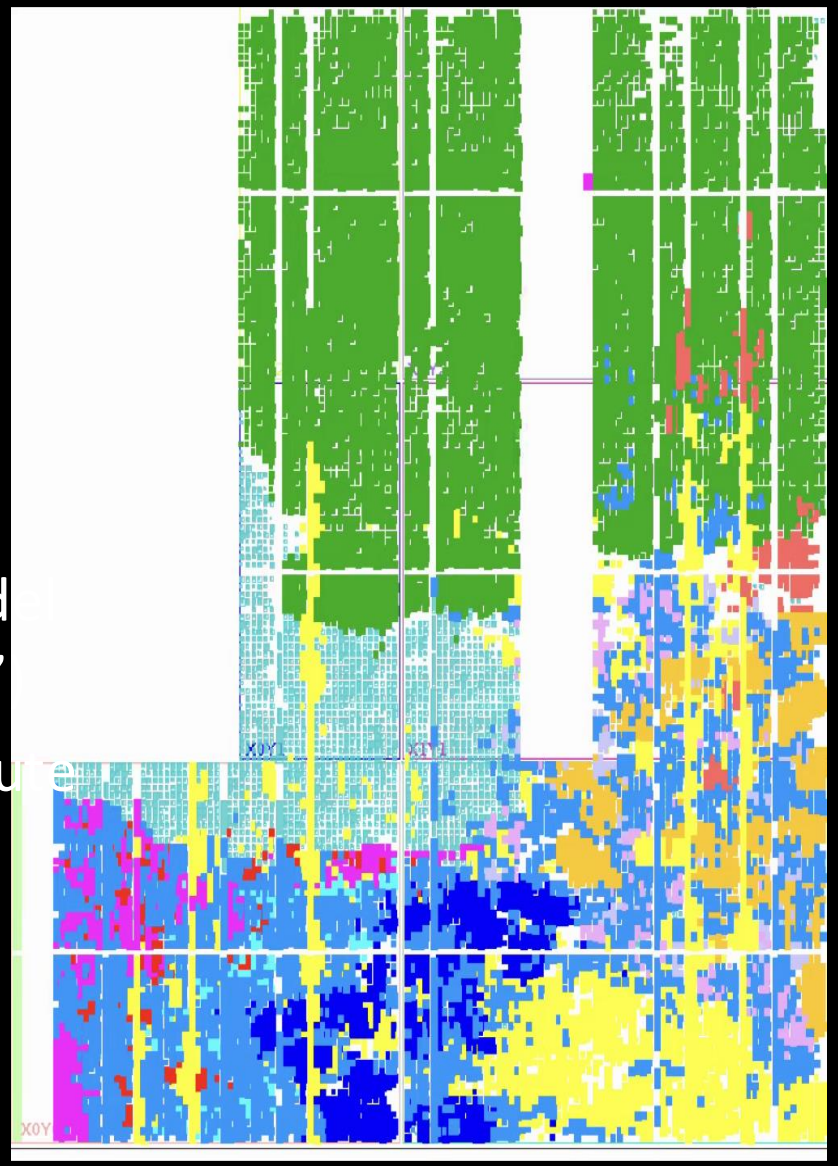
Softmax

Prediction

hls4ml

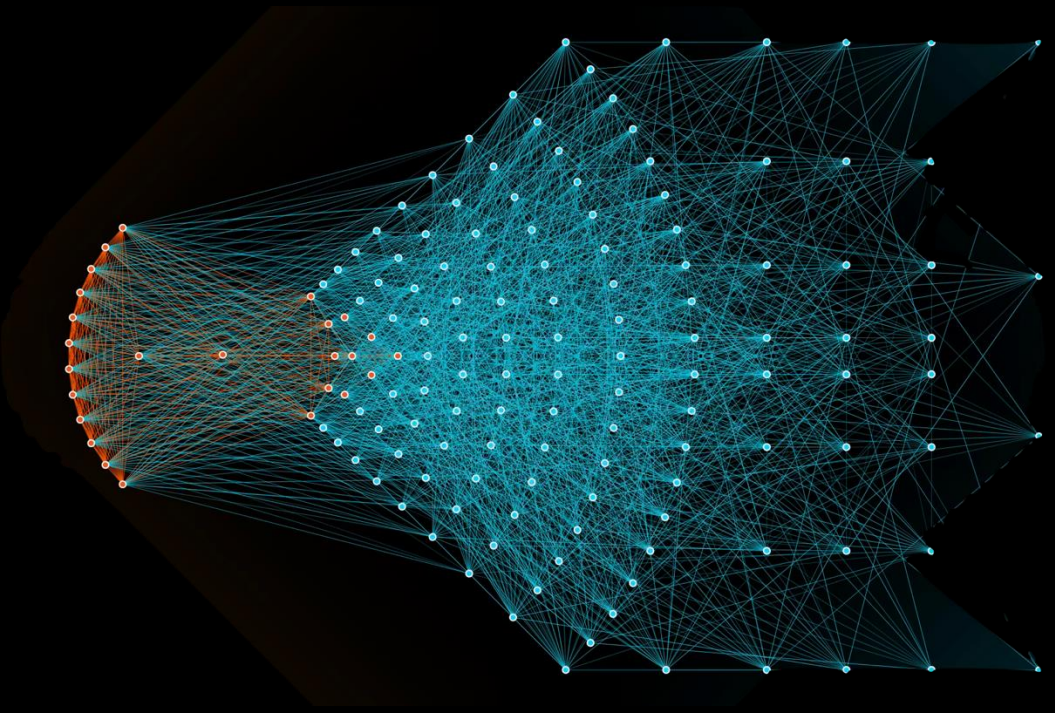
Data flow architecture

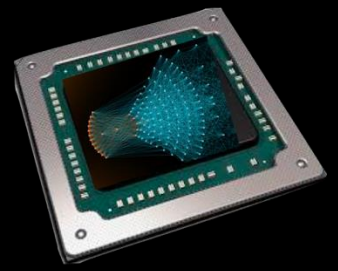
- Tailored hardware for a mod
  (“Decisions are design time”
- Each layer is separate compu
  unit
- **Stay on-chip**



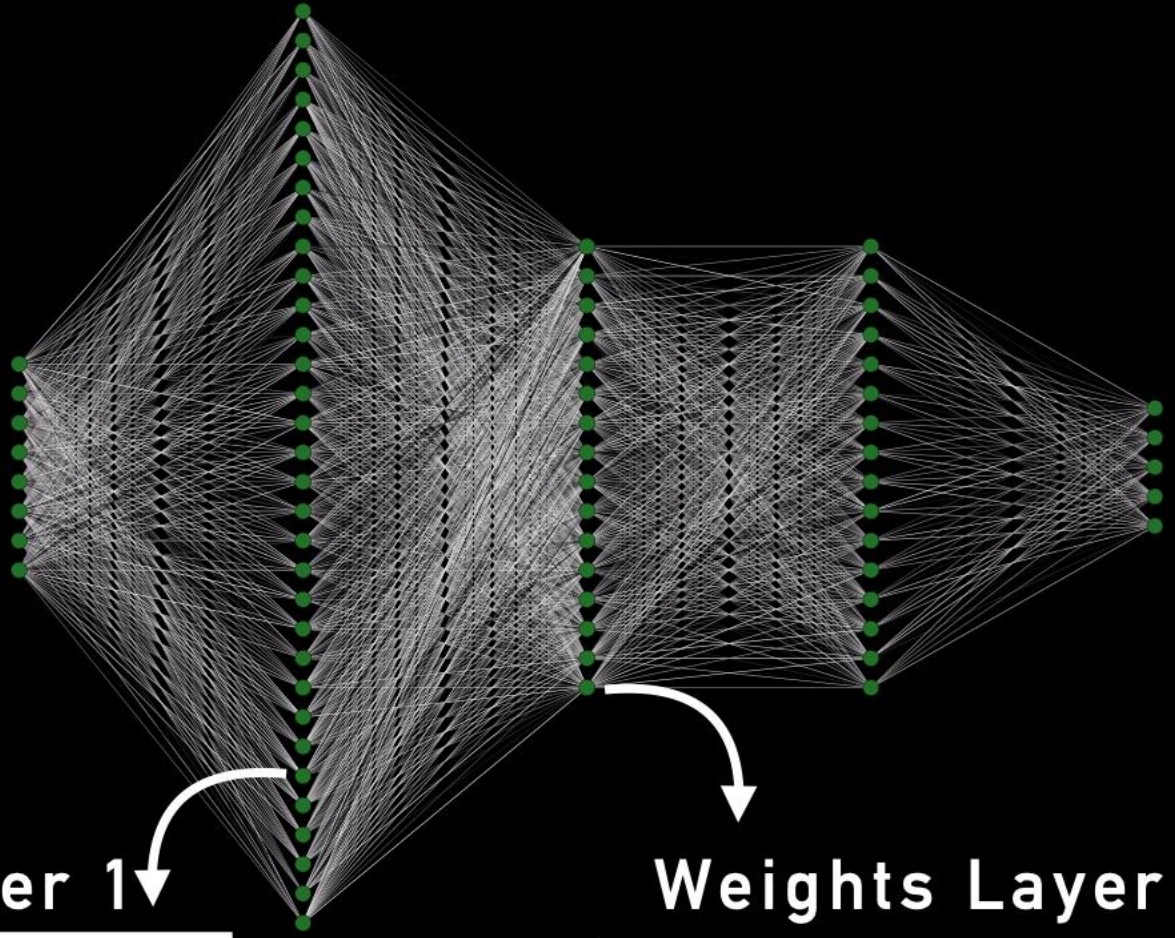
IDEALLY

REALITY

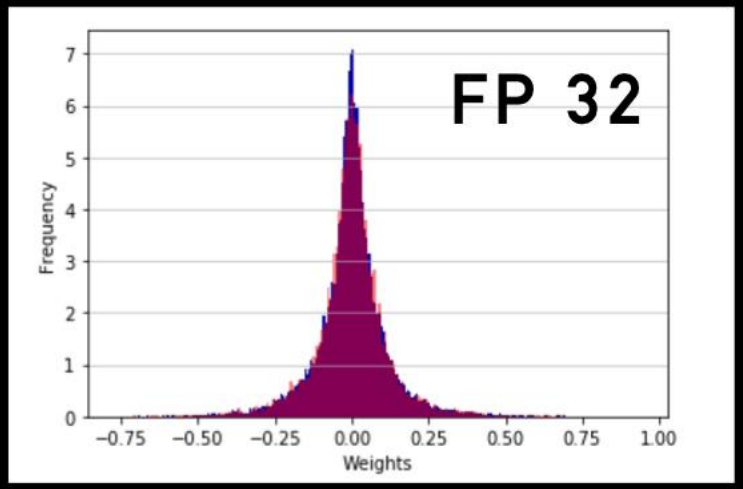




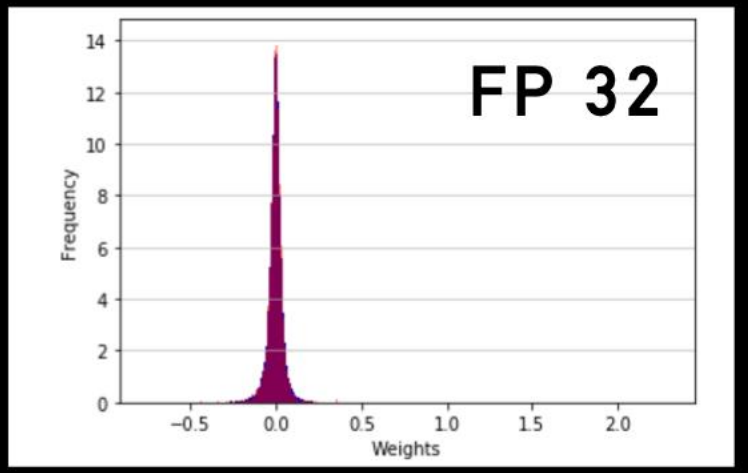
Quantization
Pruning
Parallelisation
Knowledge

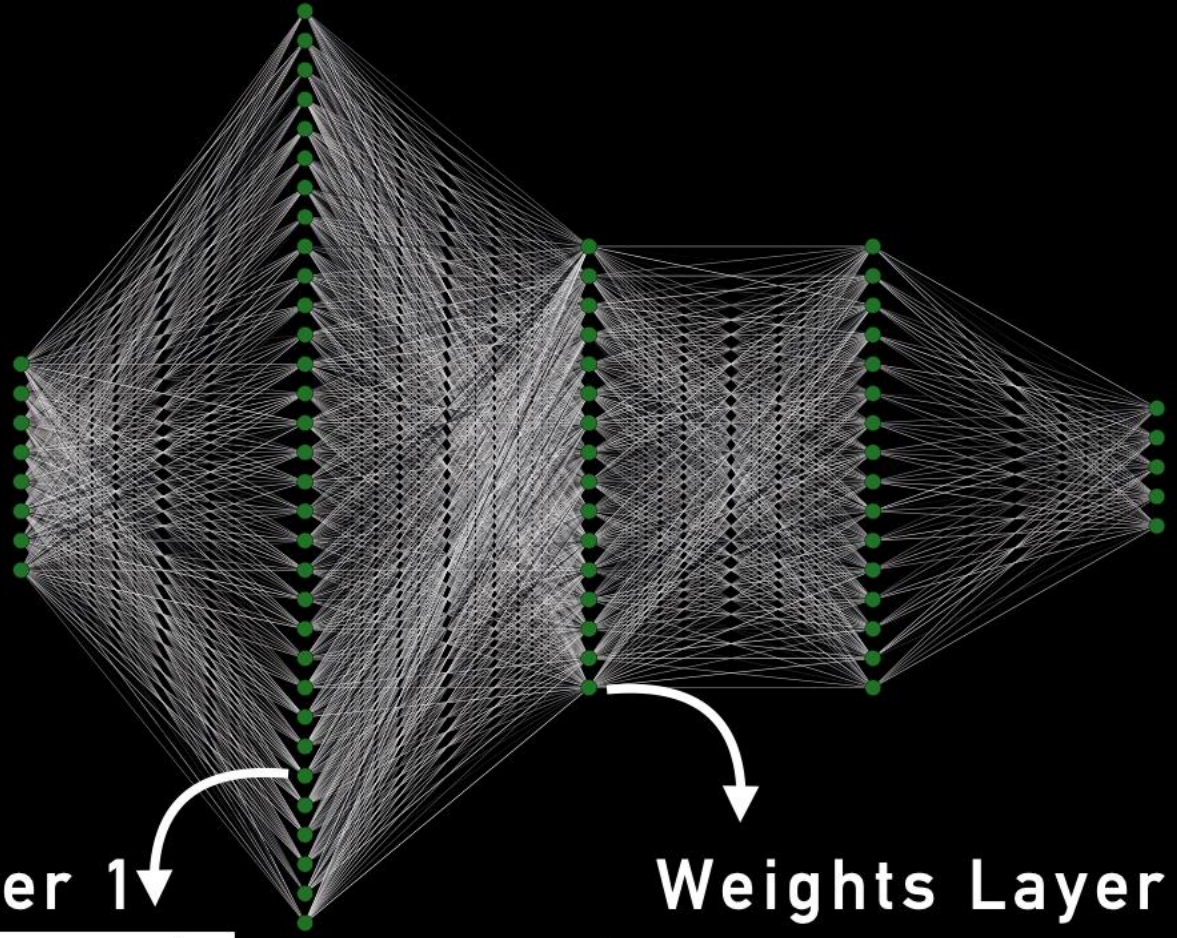


Weights Layer 1



Weights Layer 2





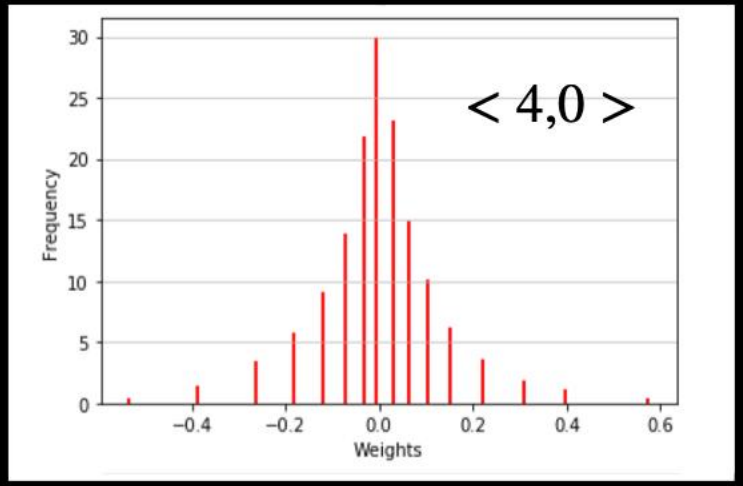
Fixed point

0101.1011101010

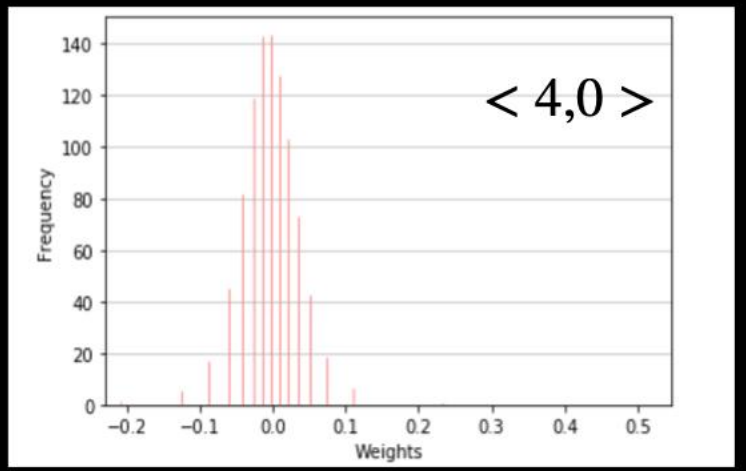
integer — fractional

width

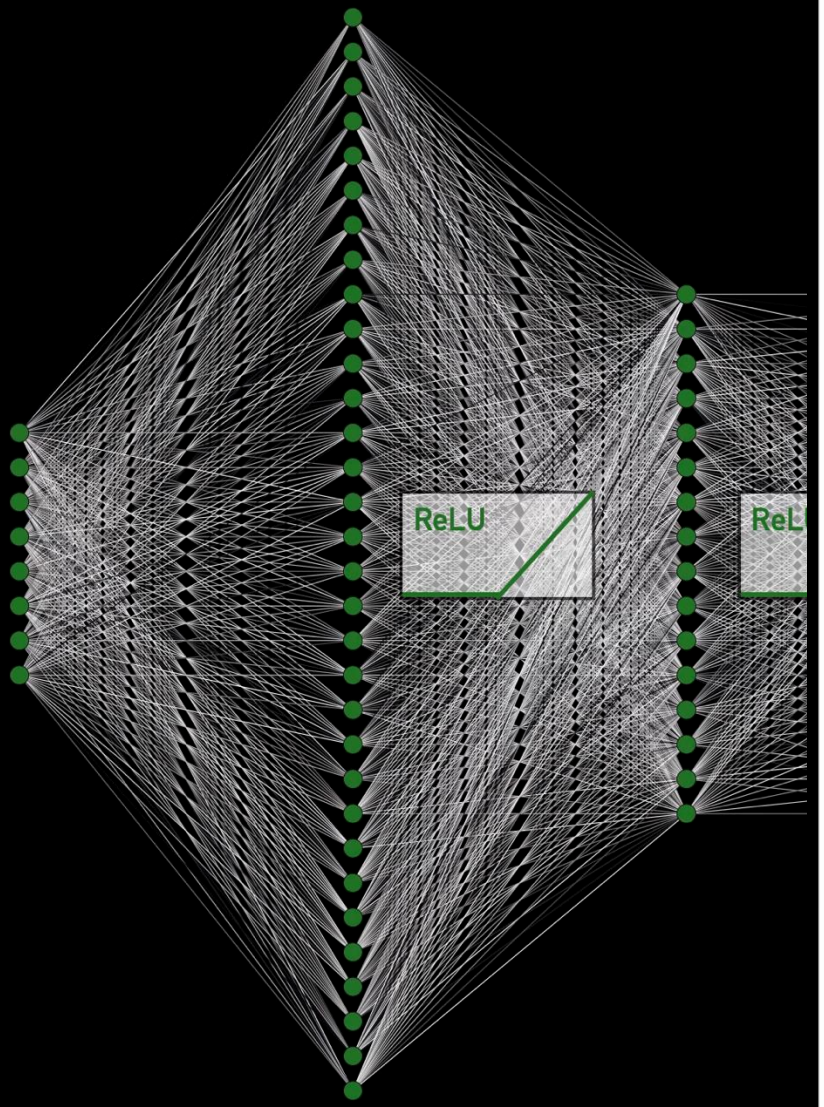
Weights Layer 1



Weights Layer 2



QKERAS



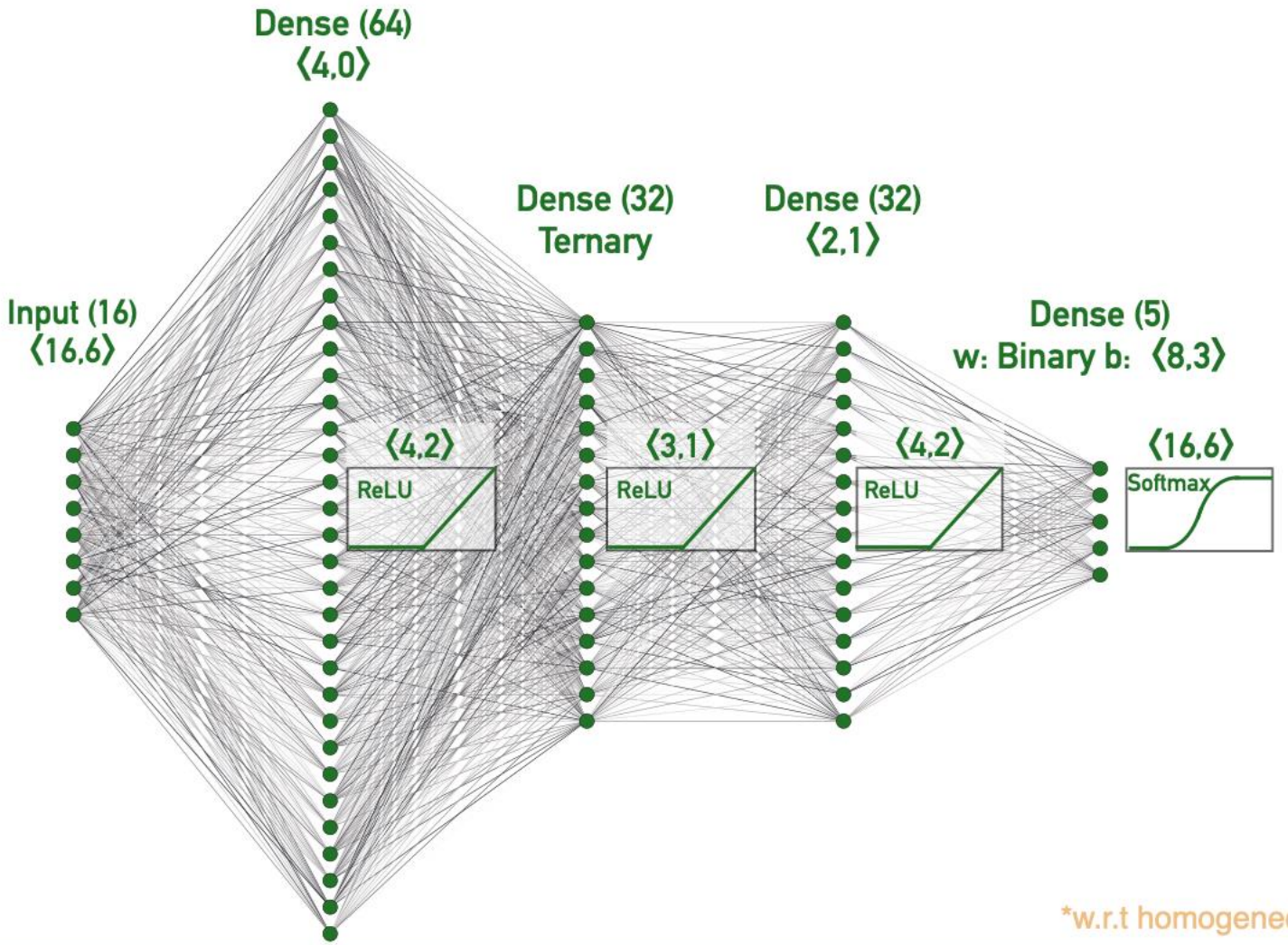

```
from tensorflow.keras.layers import Input, Activation
from qkeras import quantized_bits
from qkeras import QDense, QActivation
from qkeras import QBatchNormalization

x = Input((16))
x = QDense(64,
    kernel_quantizer = quantized_bits(6,0,alpha=1),
    bias_quantizer   = quantized_bits(6,0,alpha=1))(x)
x = QBatchNormalization()(x)
x = QActivation('quantized_relu(6,0)')(x)
x = QDense(32,
    kernel_quantizer = quantized_bits(6,0,alpha=1),
    bias_quantizer   = quantized_bits(6,0,alpha=1))(x)
x = QBatchNormalization()(x)
x = QActivation('quantized_relu(6,0)')(x)
x = QDense(32,
    kernel_quantizer = quantized_bits(6,0,alpha=1),
    bias_quantizer   = quantized_bits(6,0,alpha=1))(x)
x = QBatchNormalization()(x)
x = QActivation('quantized_relu(6,0)')(x)
x = QDense(5,
    kernel_quantizer = quantized_bits(6,0,alpha=1),
    bias_quantizer   = quantized_bits(6,0,alpha=1))(x)
x = Activation('softmax')(x)
```

Nature Machine Intelligence 3 (2021)

Nature Machine Intelligence 3 (2021)

AutoQKeras



*w.r.t homogeneously quantized 6 bit model

Model	Accuracy (%)	Precision								$\frac{E}{E_{Q6}}$	$\frac{\text{Bits}}{\text{Bits}_{Q6}}$
		Dense	ReLU	Dense	ReLU	Dense	ReLU	Dense	Softmax		
QE	72.3	⟨4, 0⟩	⟨4, 2⟩	Ternary	⟨3, 1⟩	⟨2, 1⟩	⟨4, 2⟩	w: Stoc. bin. b: ⟨8, 3⟩	⟨16, 6⟩	0.27	0.18*

High Granularity Quantization

```
from tensorflow.keras.layers import Input
from HGQ import HQuantize, HDense

inp = Input((16,))
out = HQuantize(name='inp_q', beta=beta)(out)
out = HDense(64, activation='relu', beta=beta)(out)
out = HDense(32, activation='relu', beta=beta)(out)
out = HDense(32, activation='relu', beta=beta)(out)
out = HDense(5, activation='linear', beta=beta)(out)

hgq_model = Model(inp, out)
```

arxiv:2405.00645

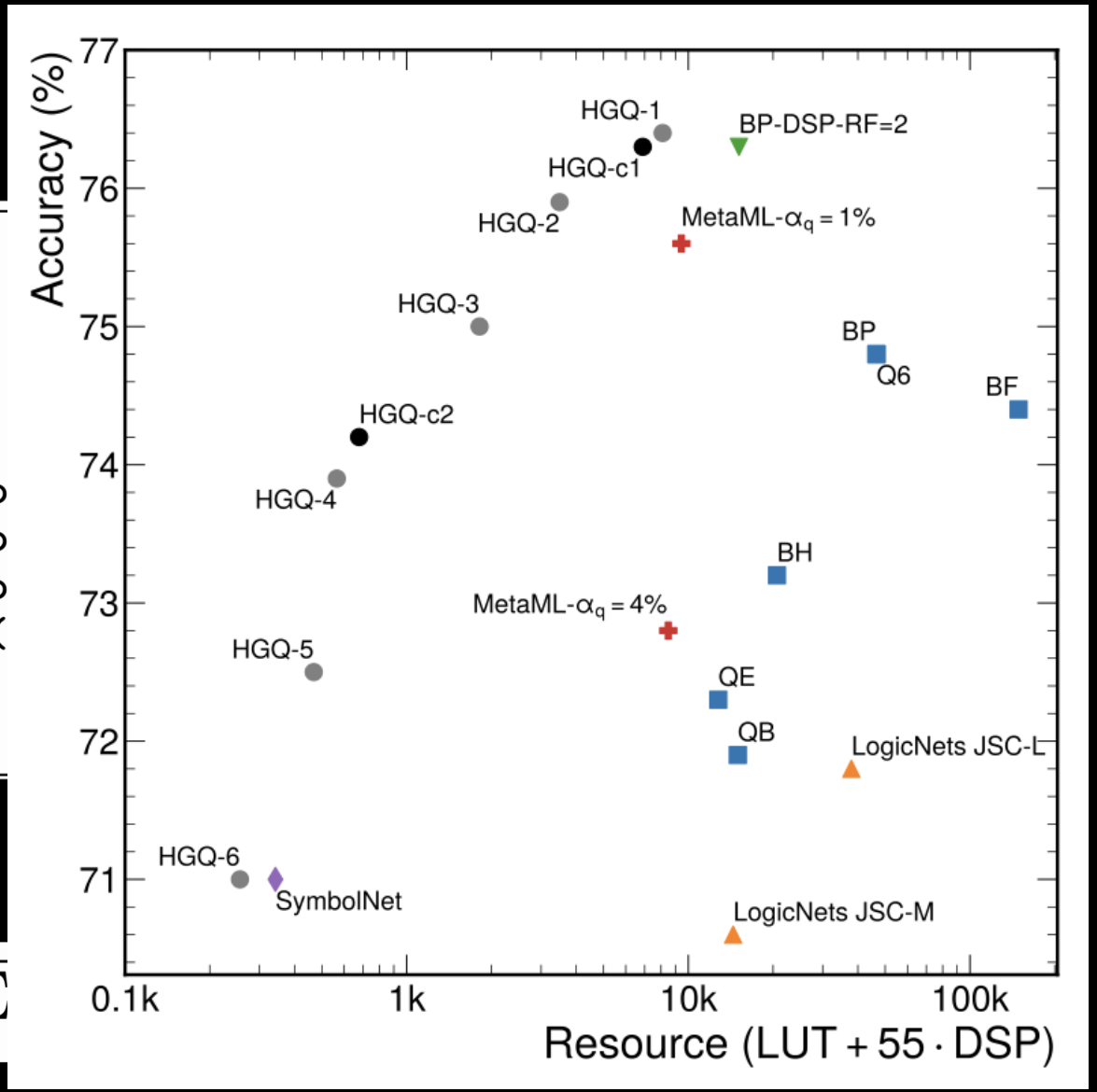
High Granularity Quantization

```
from tensorflow.keras.layers import Input
from HGQ import HQuantize, HDense

inp = Input((16,))
out = HQuantize(name='inp_q', beta=beta)(out)
out = HDense(64, activation='relu', beta=beta)(o
out = HDense(32, activation='relu', beta=beta)(o
out = HDense(32, activation='relu', beta=beta)(o
out = HDense(5, activation='linear', beta=beta)(

hgq_model = Model(inp, out)
```

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \beta \cdot \overline{\text{E}}$$



arxiv:2405.00645

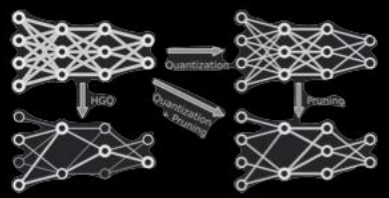
hls4ml + AMD Research Labs: QONNX

Frontend (QAT)

Backend (compiler)

QKeras

PyTorch Brevitas



HQG

QONNX

FINN AMD

hls4ml

Marius Köppel
A. Pappalardo et al. (2022)

Standard industry tools couldn't reach the specs that we require

Better: ↑ ↓ ↑

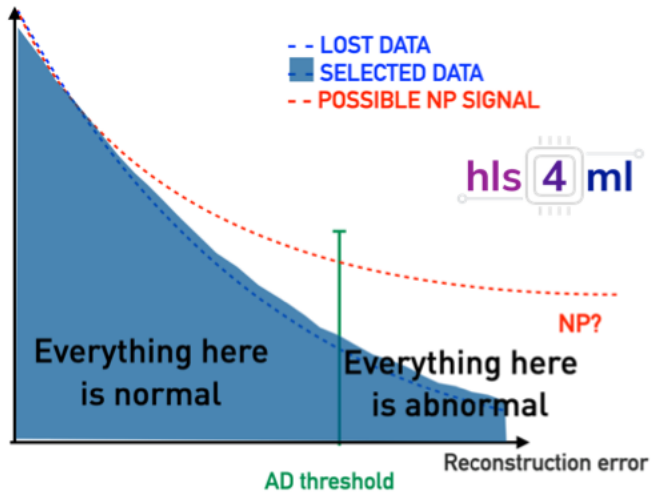
Platform	Framerate (FPS)	Power (W)	pixels / s / W ($\times 10^6$)
hls4ml	**1500**	**3.7**	26.5
Vitis AI	40	9.4	1.1
Ratio hls4ml / Vitis AI	38	0.4	**24.0**

ZCU102
(Zynq Ultrascale+ 9)

2024: Neural hardware triggers

CMS:

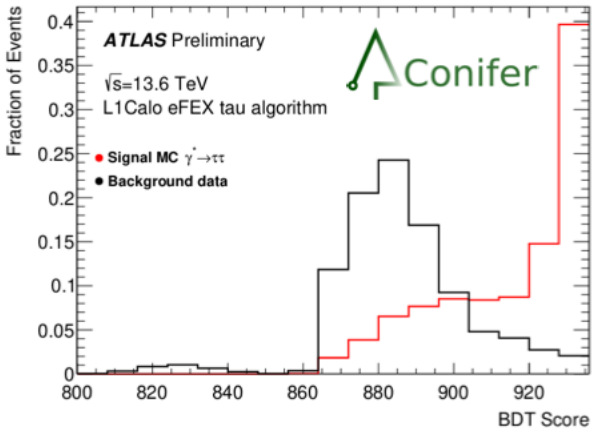
- Anomaly detection in 50 ns



CMS DP2023_079

ATLAS:

- BDT selecting candidate τ events in <100ns

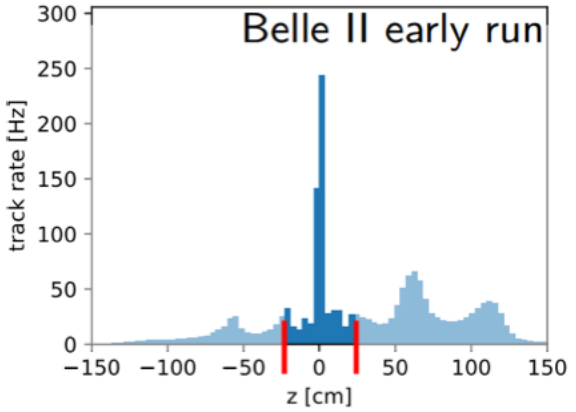


L1CaloTriggerPublicResults

Belle-2:

- Neural track trigger since April 2021

Reject $|z| \neq 0$ cm tracks

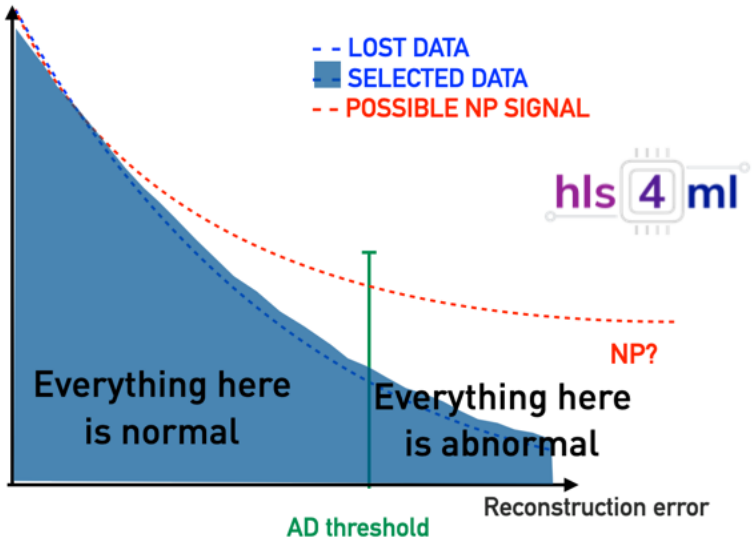


arxiv:2402.14962

2024: Neural hardware triggers making decisions in LHC experiments!

CMS Experiment:

- Anomaly detection in 50 ns
- 300 events/second

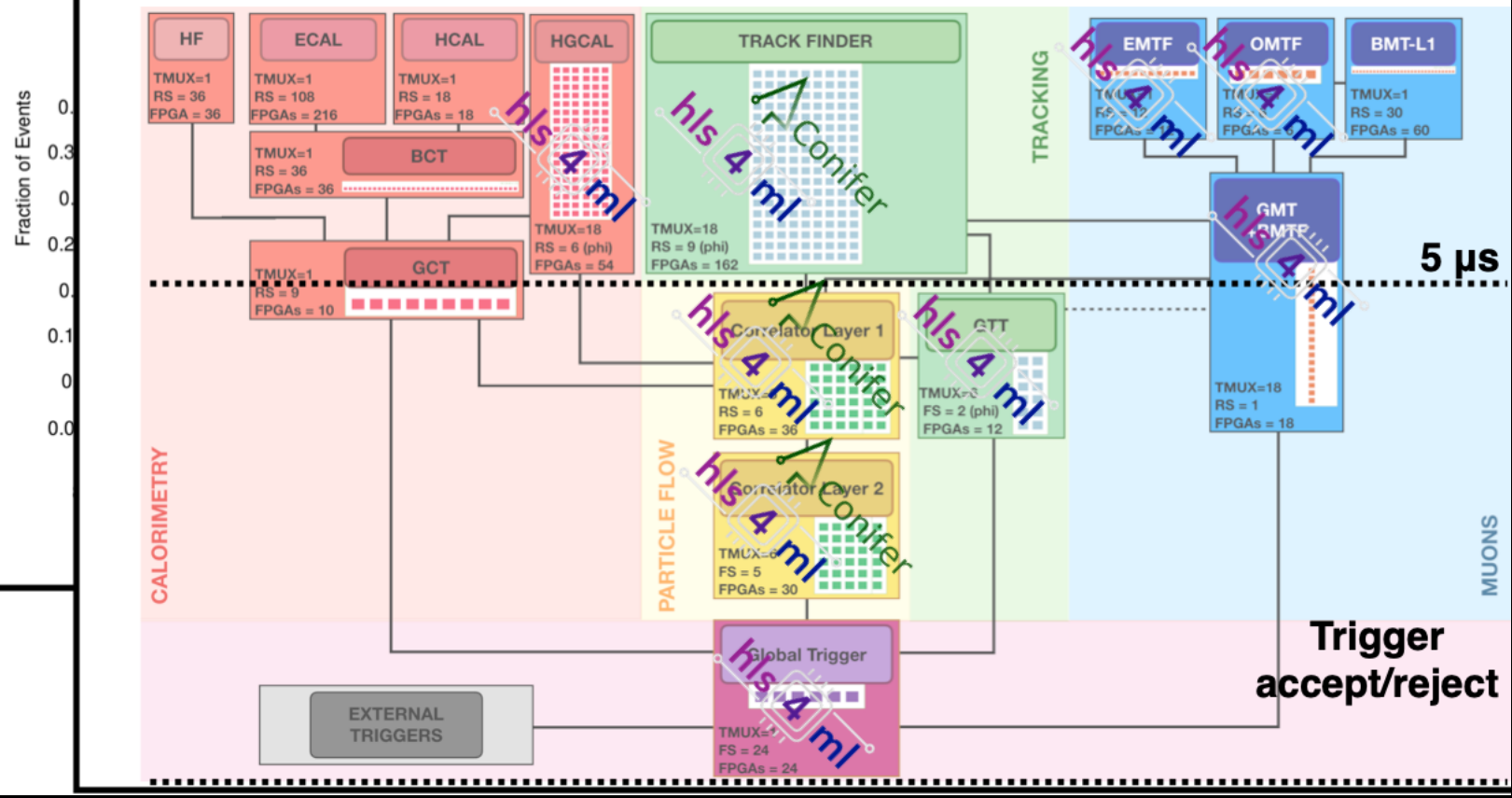


CMS DP2023_079

ATLAS Experiment:

- BDT selecting candidate τ lepton events in <100 ns

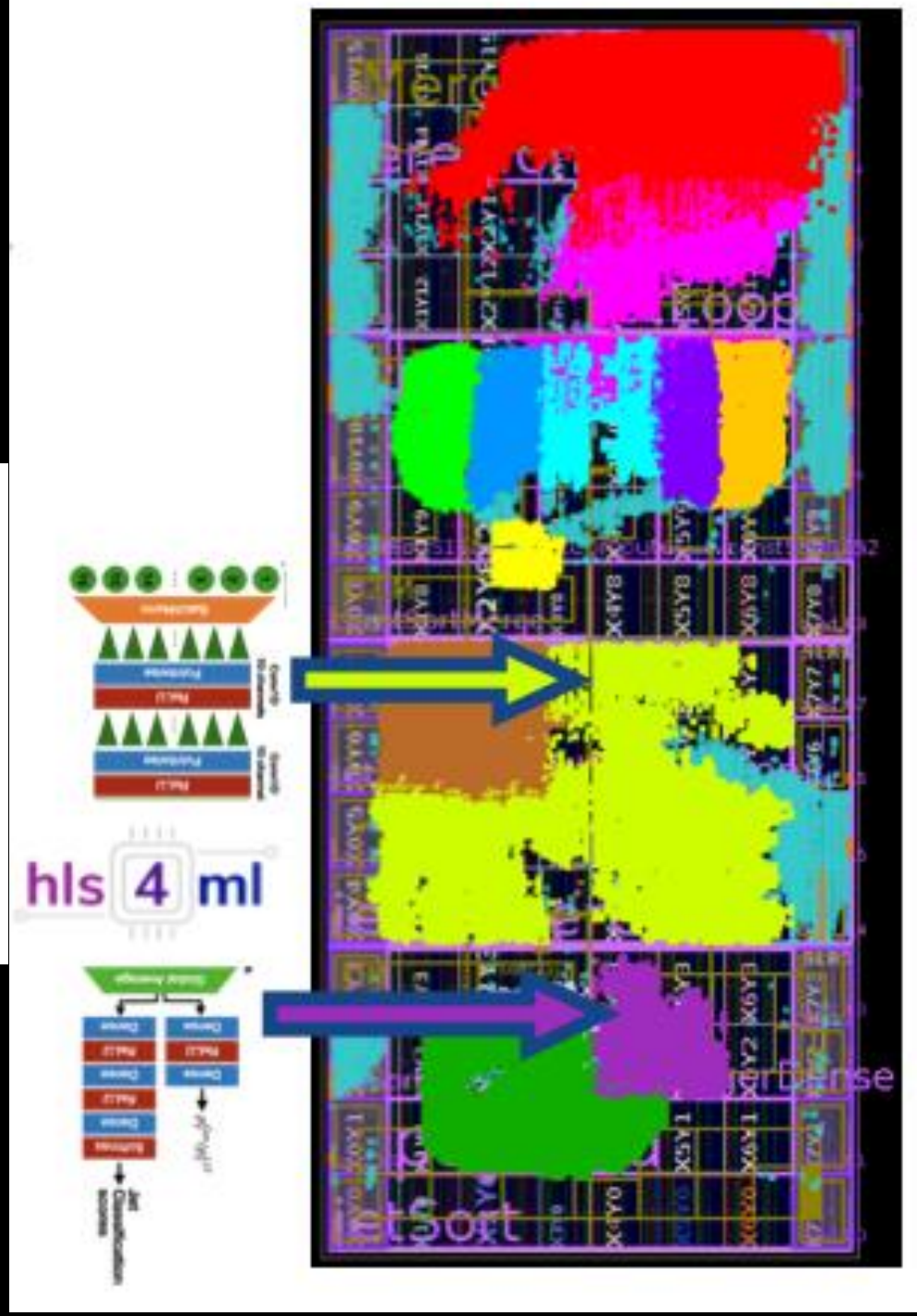
2031: ~20 billion inferences/s during HL-LHC

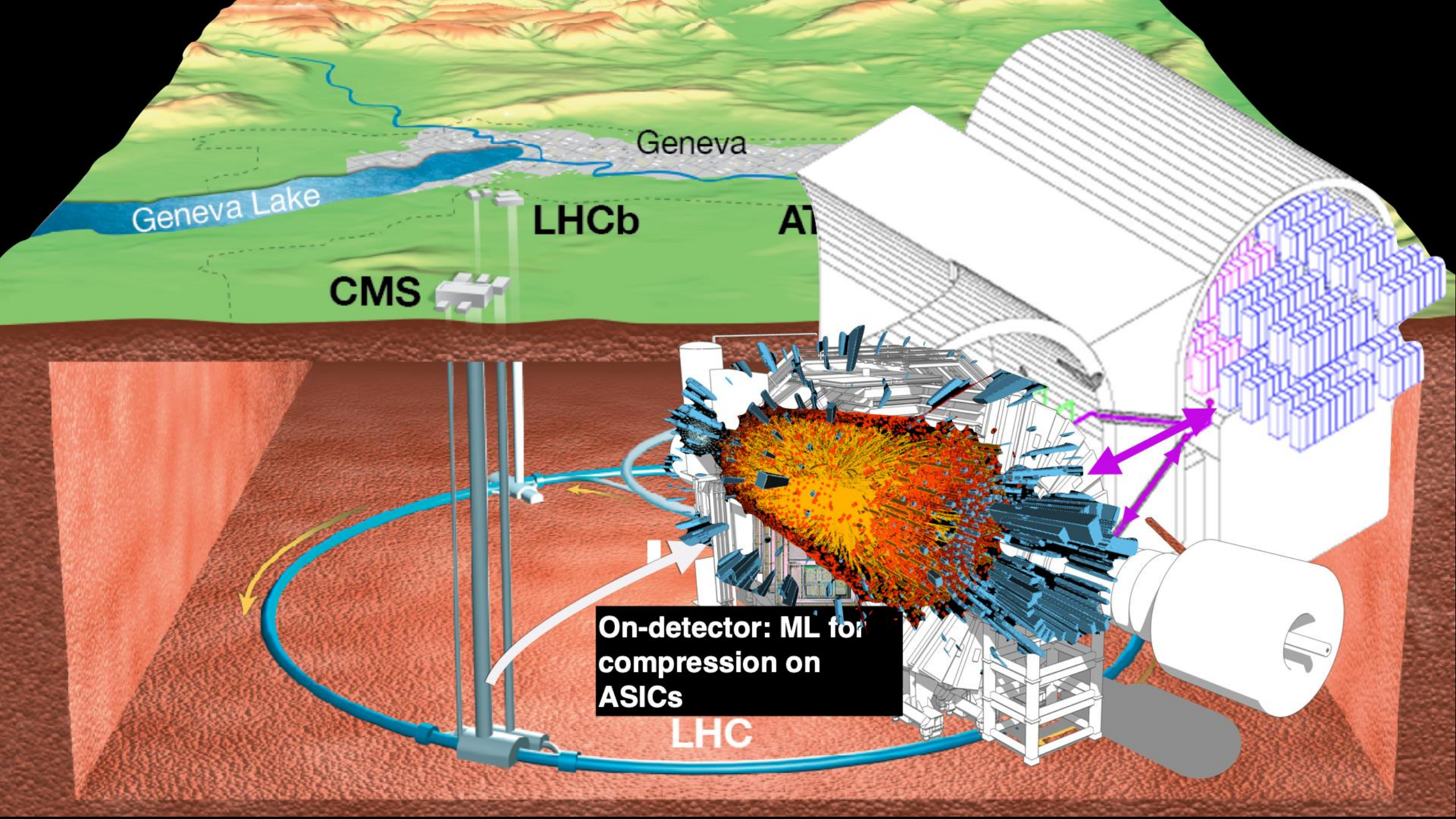


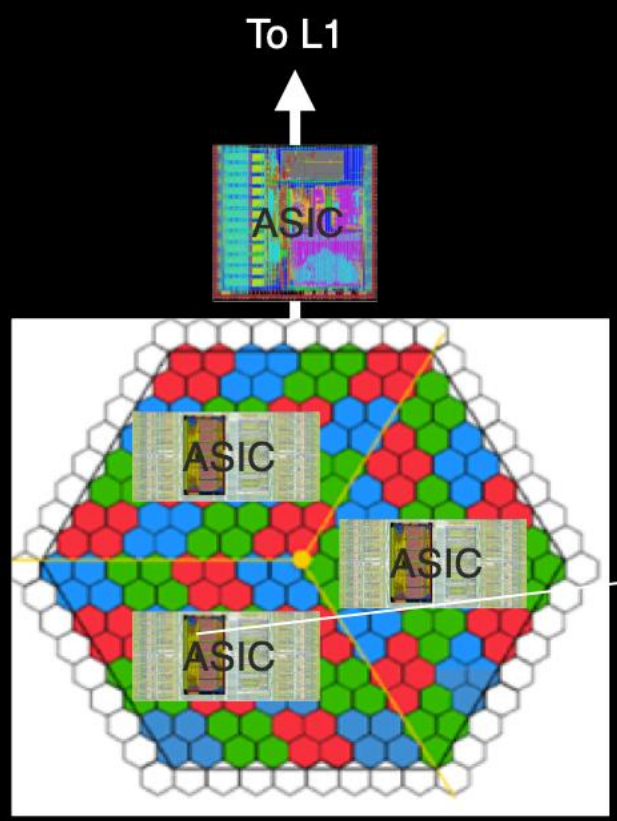
HL-LHC CMS will run with **FPGA track finding, particle flow, jet clustering and flavour tagging in < 12 µs**

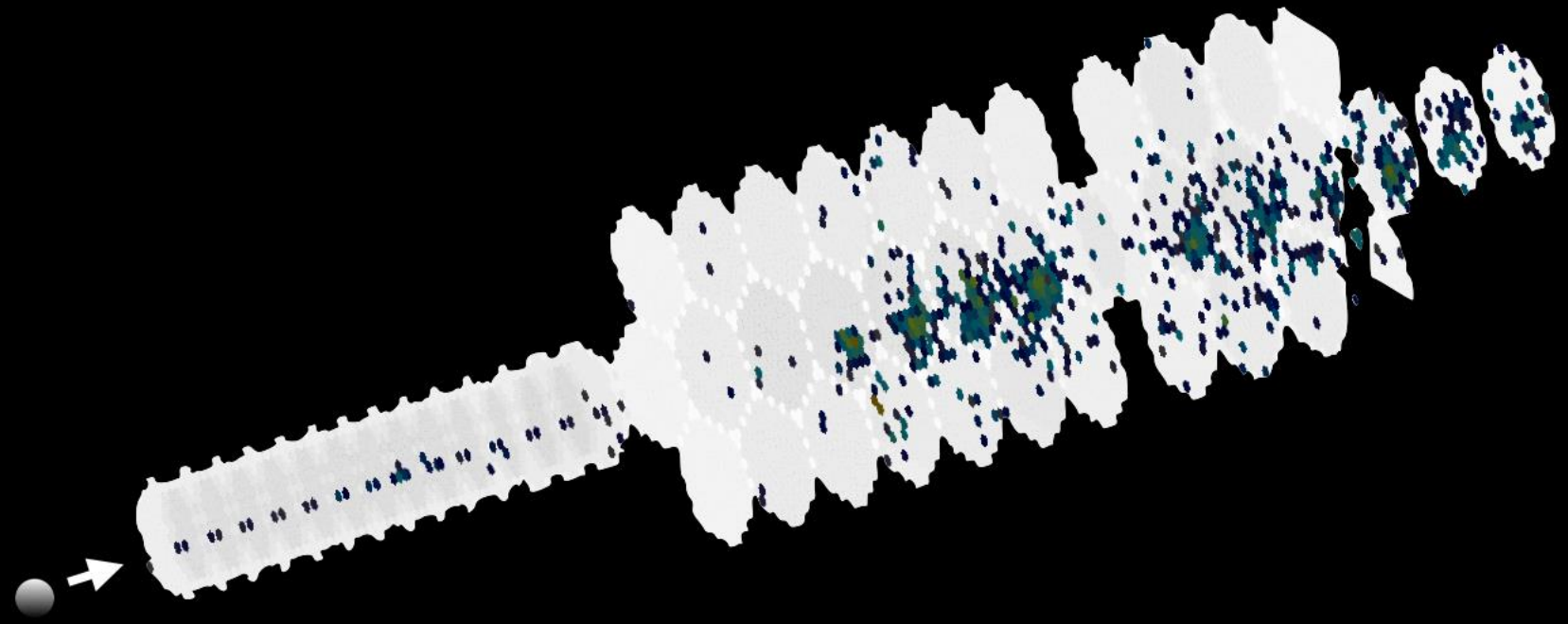
- Reco. jets; sort the constituents by p_T; compute input variables; per-constituent NN part; per-jet NN part; sort jets by p_T

Component	Latency (µs)	LUT	FF	DSP	BRAM
Preprocessing (incl. sort)	0.10	6%	4%	1%	0.6%
Jet Tag NN	0.19	7%	5%	14%	1%
Jet Reco	0.74	9%	7%	5%	0%
Total System	1.01	25%	15%	19%	1%



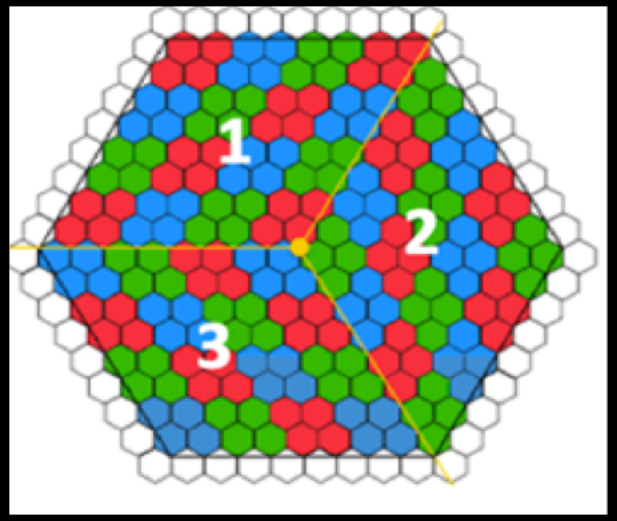

Geneva
Geneva Lake
LHCb
CMS
On-detector: ML for compression on ASICs
LHC

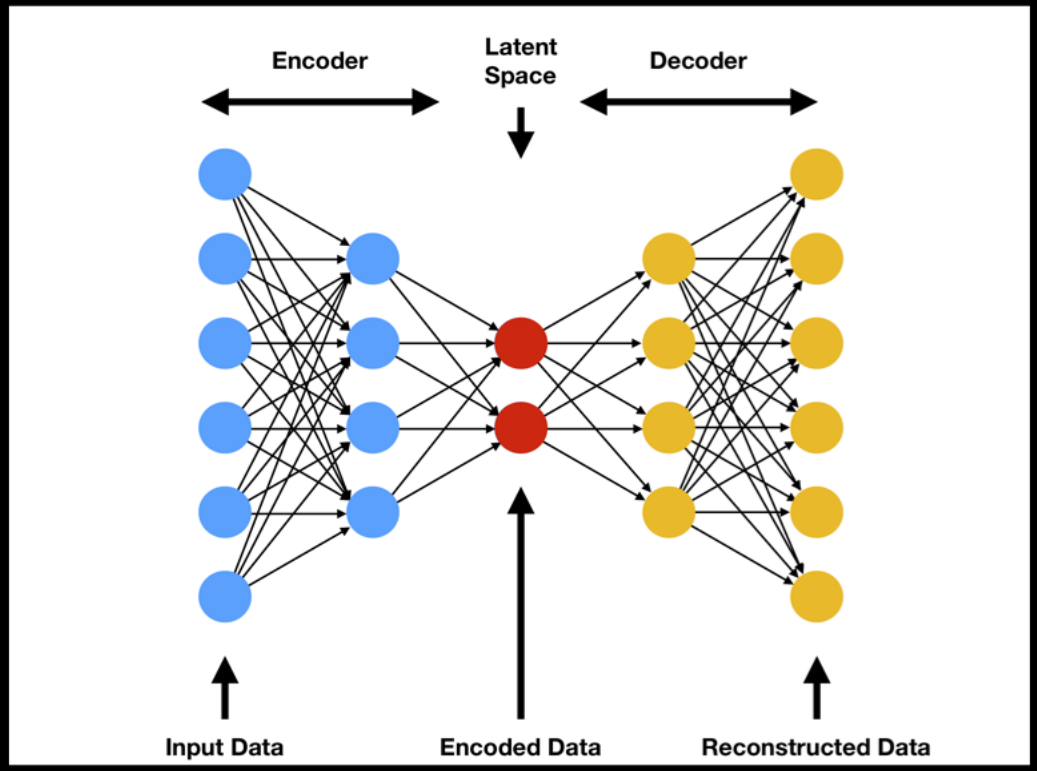


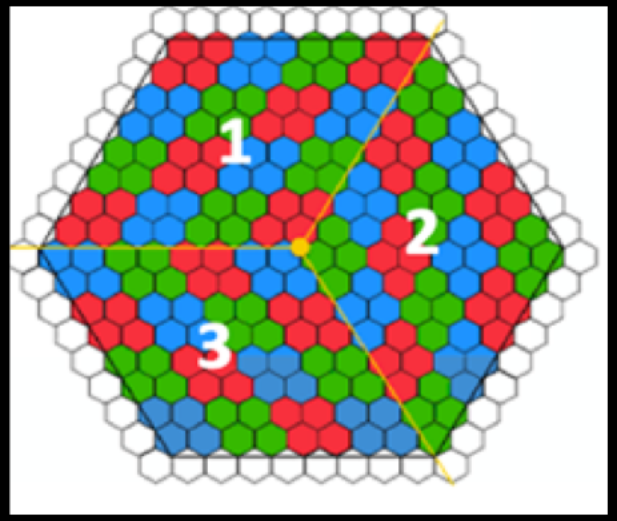


Must compress ON DETECTOR

- High radiation
- Cooled to -30 → low power
- 1.5 μs latency

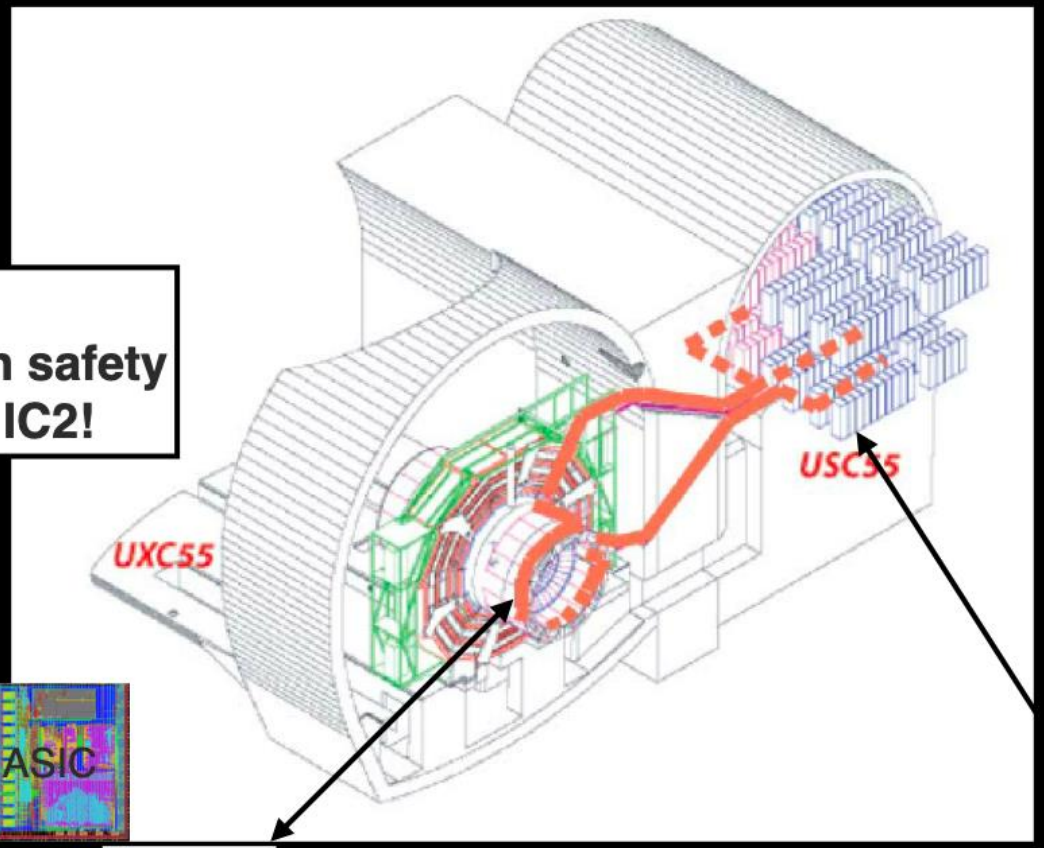


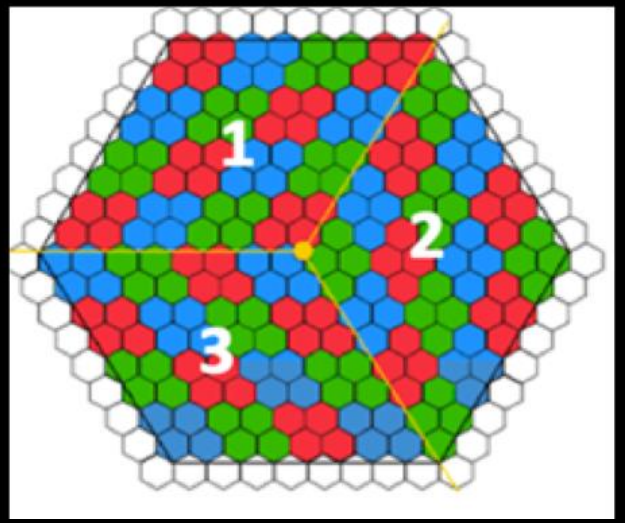


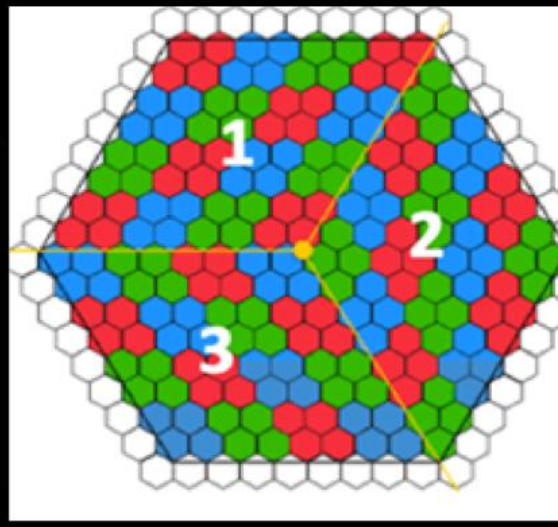


Variational Autoencoder

- 75-100 mW
- Triplicated w/b for radiation safety
  Reprogrammable w/b over IC2!







https://newsroom.sw.siemens.com/en-US/siemens-catapult-ai-nn/

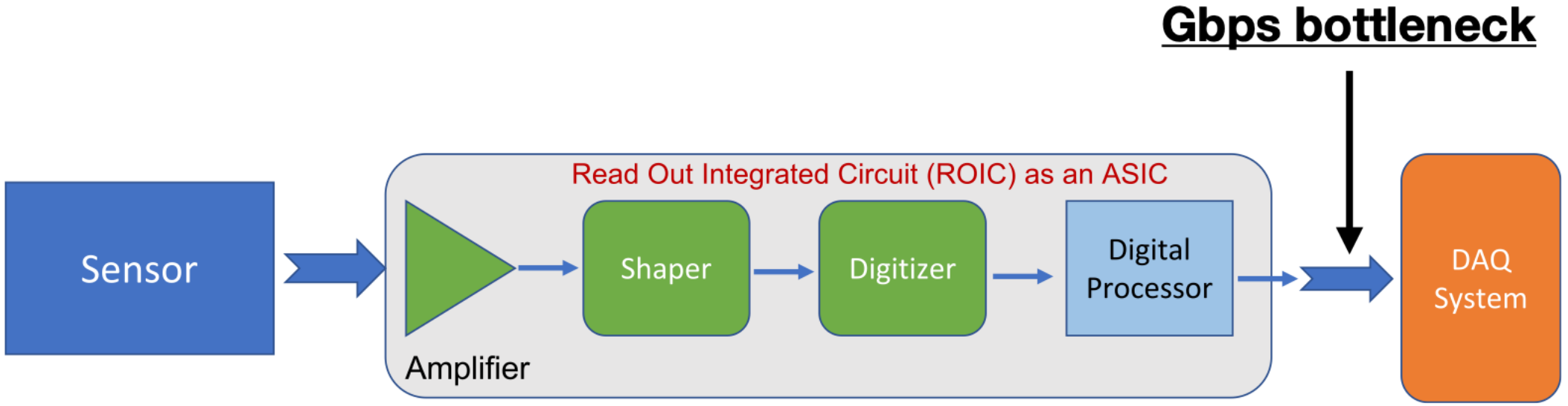
PRESS RELEASE

Siemens simplifies development of AI accelerators for advanced system-on-chip designs with Catapult AI NN

May 21, 2024
Plano, Texas

Catapult AI NN brings together hls4ml, an open-source package for machine learning hardware acceleration, and Siemens' Catapult™ HLS software for High-Level Synthesis. Developed in close collaboration with Fermilab, a U.S. Department of Energy Laboratory, and other leading contributors to hls4ml, Catapult AI NN addresses the unique requirements of machine learning accelerator design for power, performance, and area on custom silicon.

Data readout bottleneck



Example vertex detector inner layer from before (#241 FCC-ee):

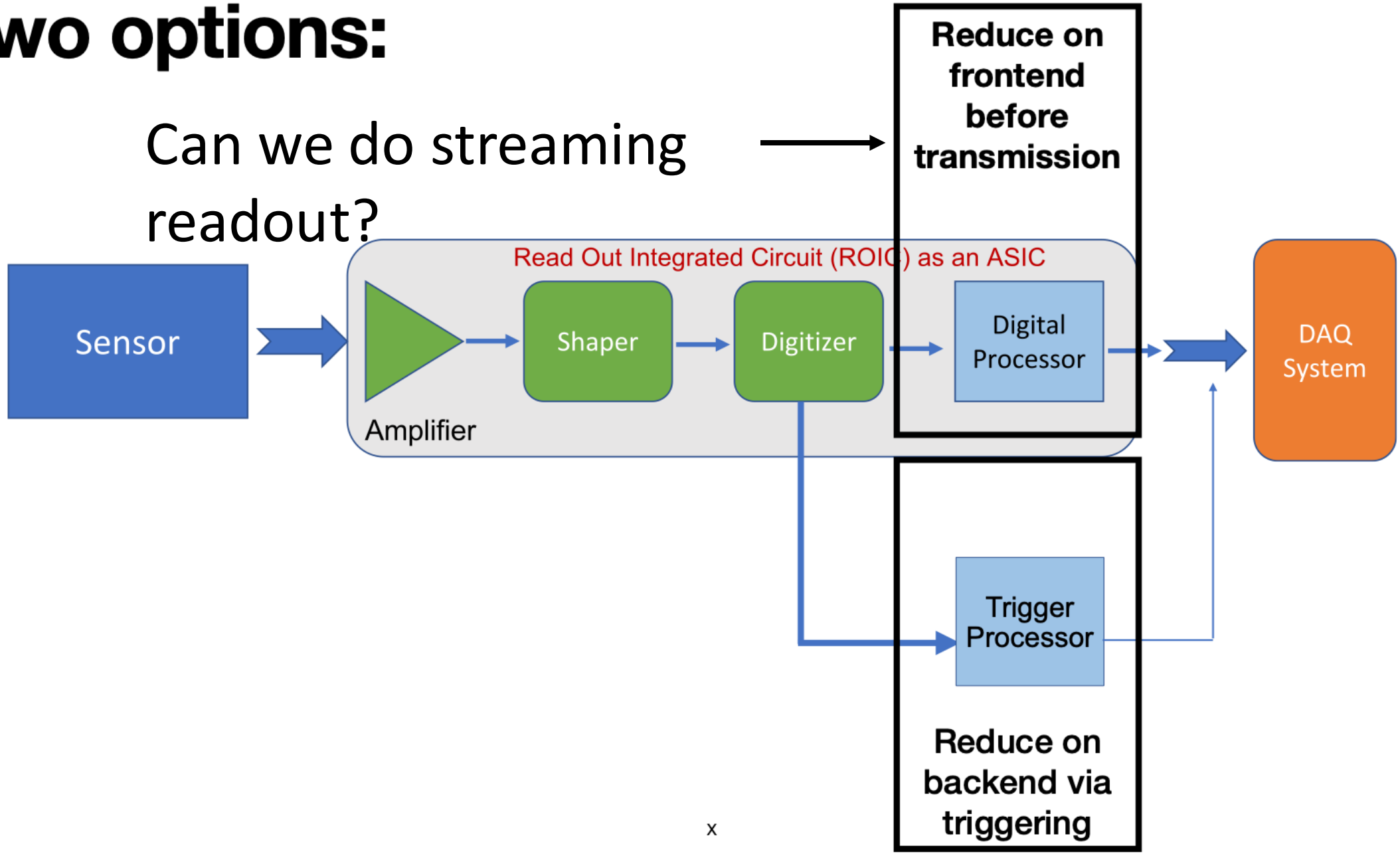
32 bits pixel data in inner layer @ 200 MHz/cm^2

→ **24.4 Gbit/s**

2.2 Tb/s for Layer 1 (x10 less in Layer 2)

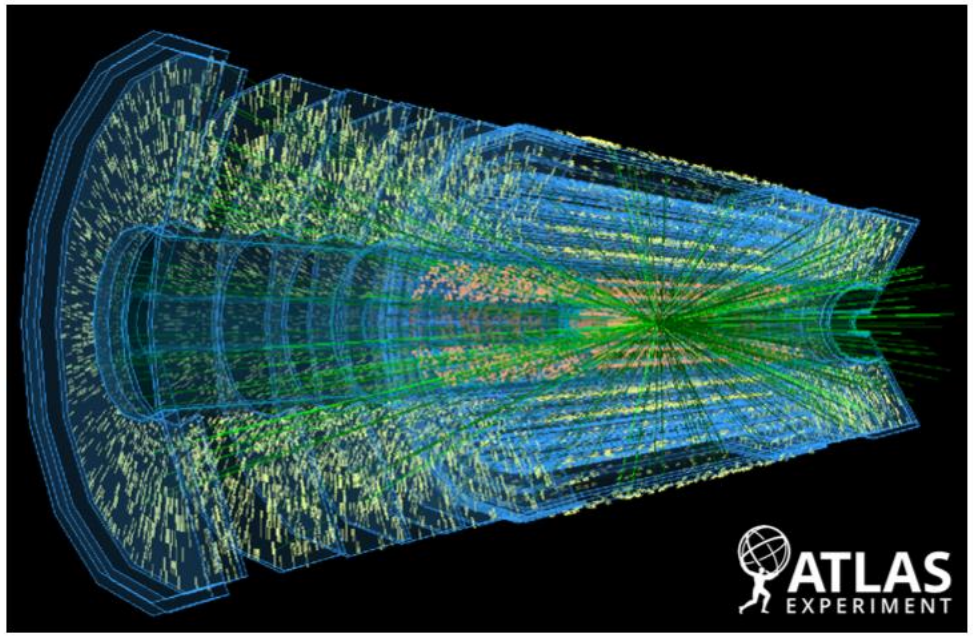
Two options:

Can we do streaming readout?

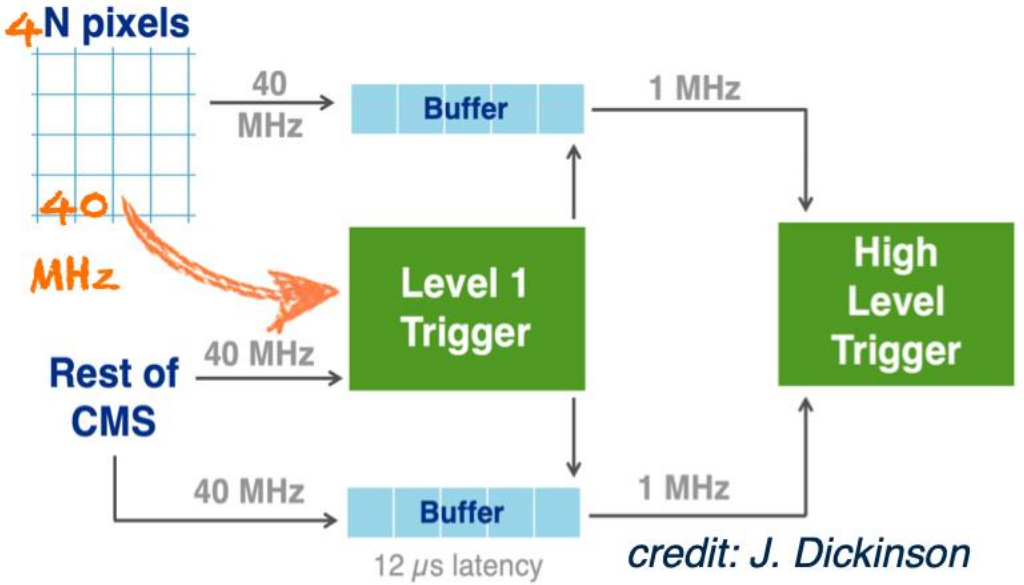


Smart Pixels

https://fastmachinelearning.org/smart-pixels/



- Reduce silicon data via **in-pixel intelligence**
  - **frontend filtering**: discard low-p_T track data (< 2 GeV)
  - **feature extraction**: Extract particle position and angle in pixel front-end ASICs from charge in single pixel layer
- Bandwidth savings of 57-75%!

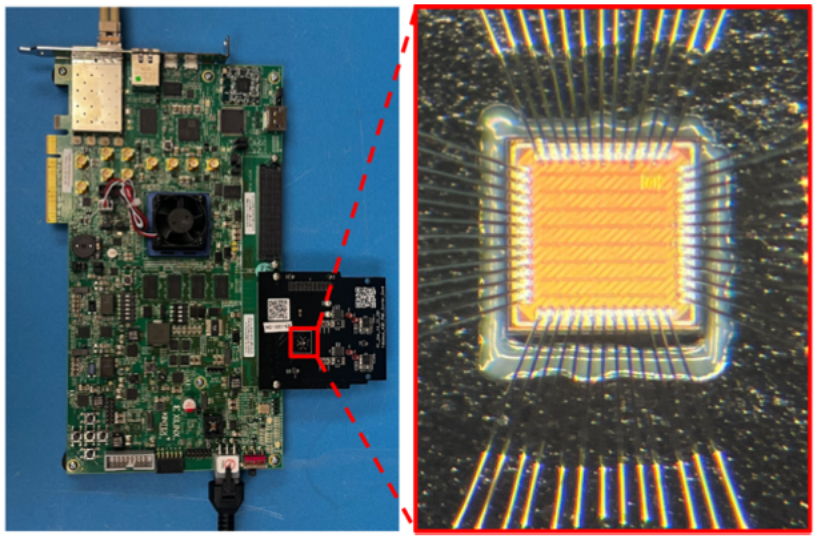


Reconfigurable logic in ASIC design

The Embedded FPGA framework

- Pathway to implementing ML "at source”
  - Fully reconfigurable logic on detector frontend
- Open source (FABulous, OpenFPGA)
  - potential to apply to variety of subsystems/ fields (SuperKEKB, FCC-ee, DUNE, free-electron lasers)

28nm CMOS ASIC (1x1mm)



Gonski et. Al

aie4ml

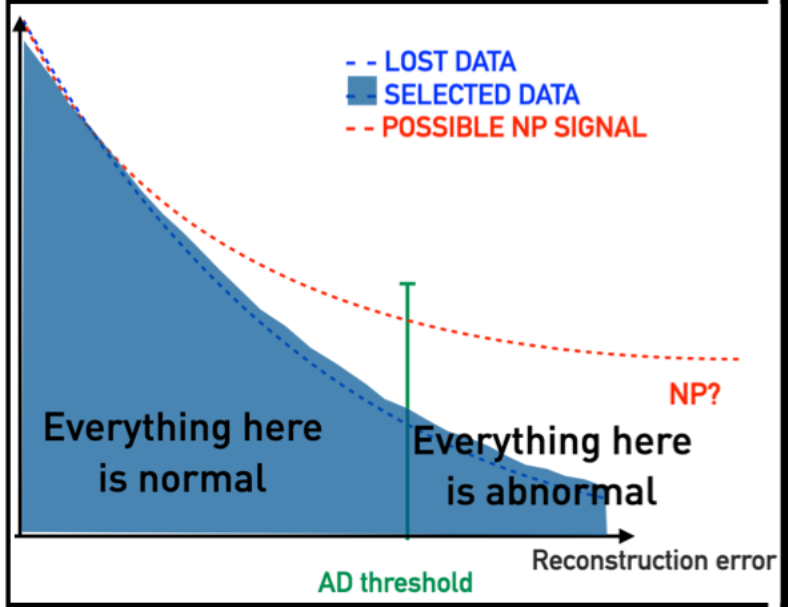
AI engines can offer higher compute density, provide necessary high speed I/O interfaces. Make them easier to use for trigger experiments!

hls4ml for AI engines

Status:
linear layers and ReLu implemented
scales workloads across AIE array
support for Gen2/Gen3 devices

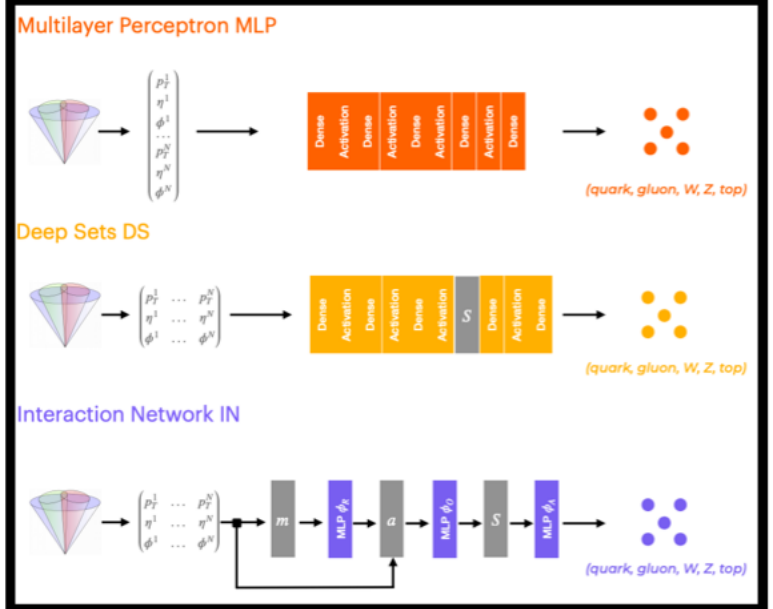
NGT WP 1.2

Anomaly detection with VAEs in 50 ns



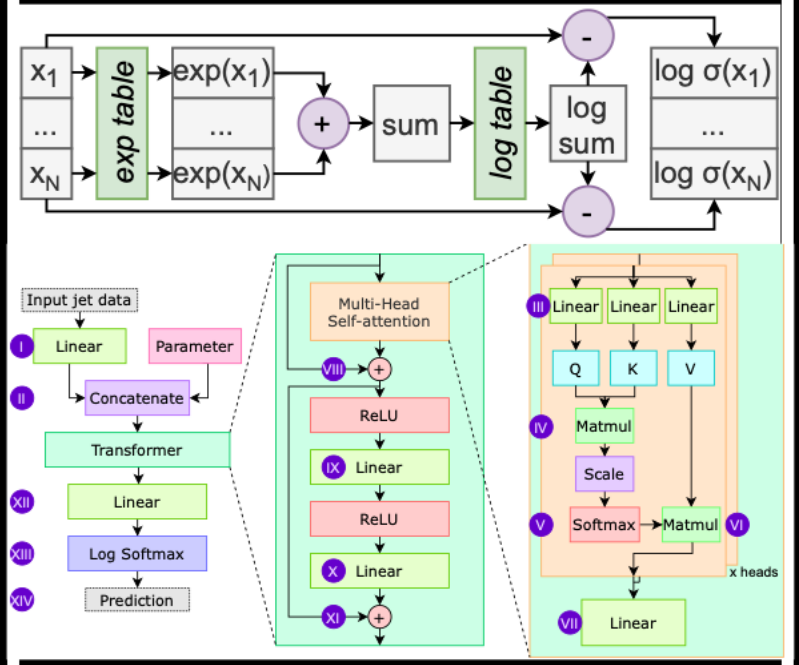
CMS DP2023_079
E. Govorkova et al (2022)

Quantised Interaction Networks and Deep Sets in <160 ns



P. Odagiu et al. 2024

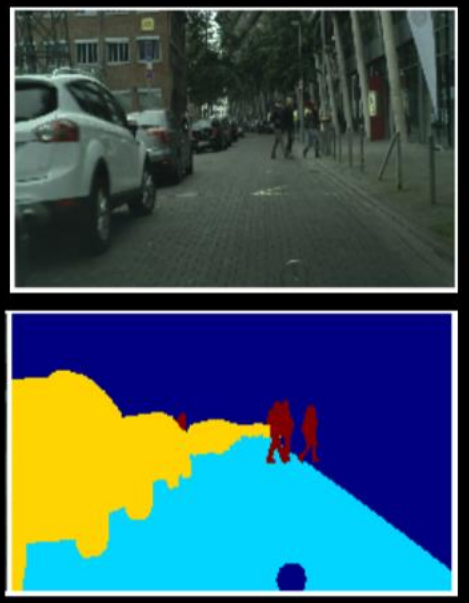
Fully on-chip transformers in 90 ns (500k flops, 9k param)



IEEE ICFTP 2022
arxiv:409.05207

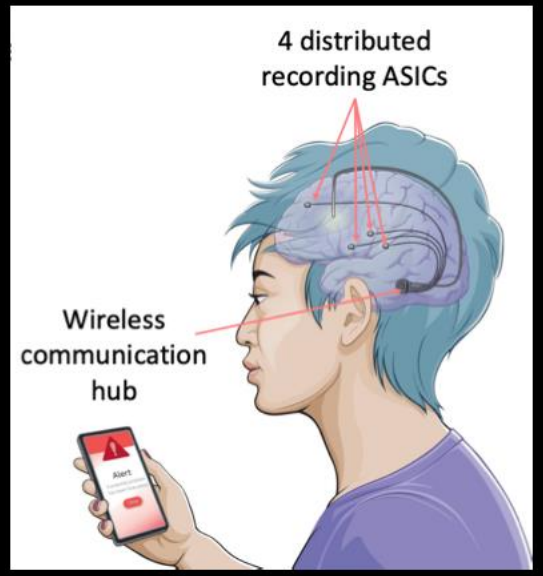
...and outside

Semantic segmentation for autonomous vehicles



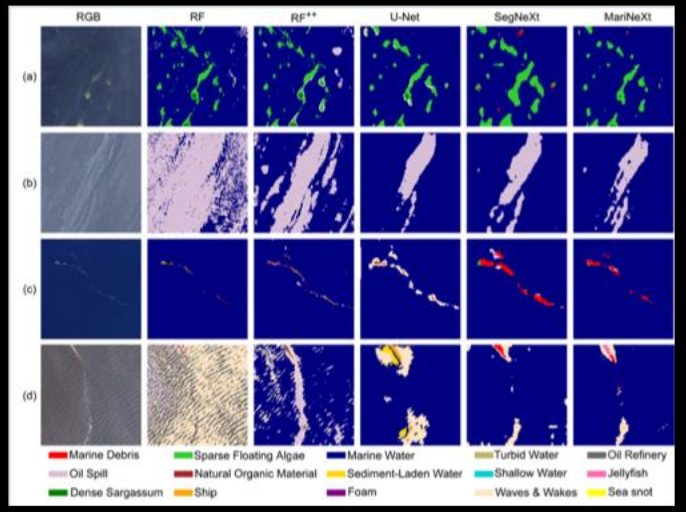
N. Ghielmetti et al. 2022

Seizure Predicting Brain Implant



W. Lemaire et al. 2022

Earth monitoring in satellites

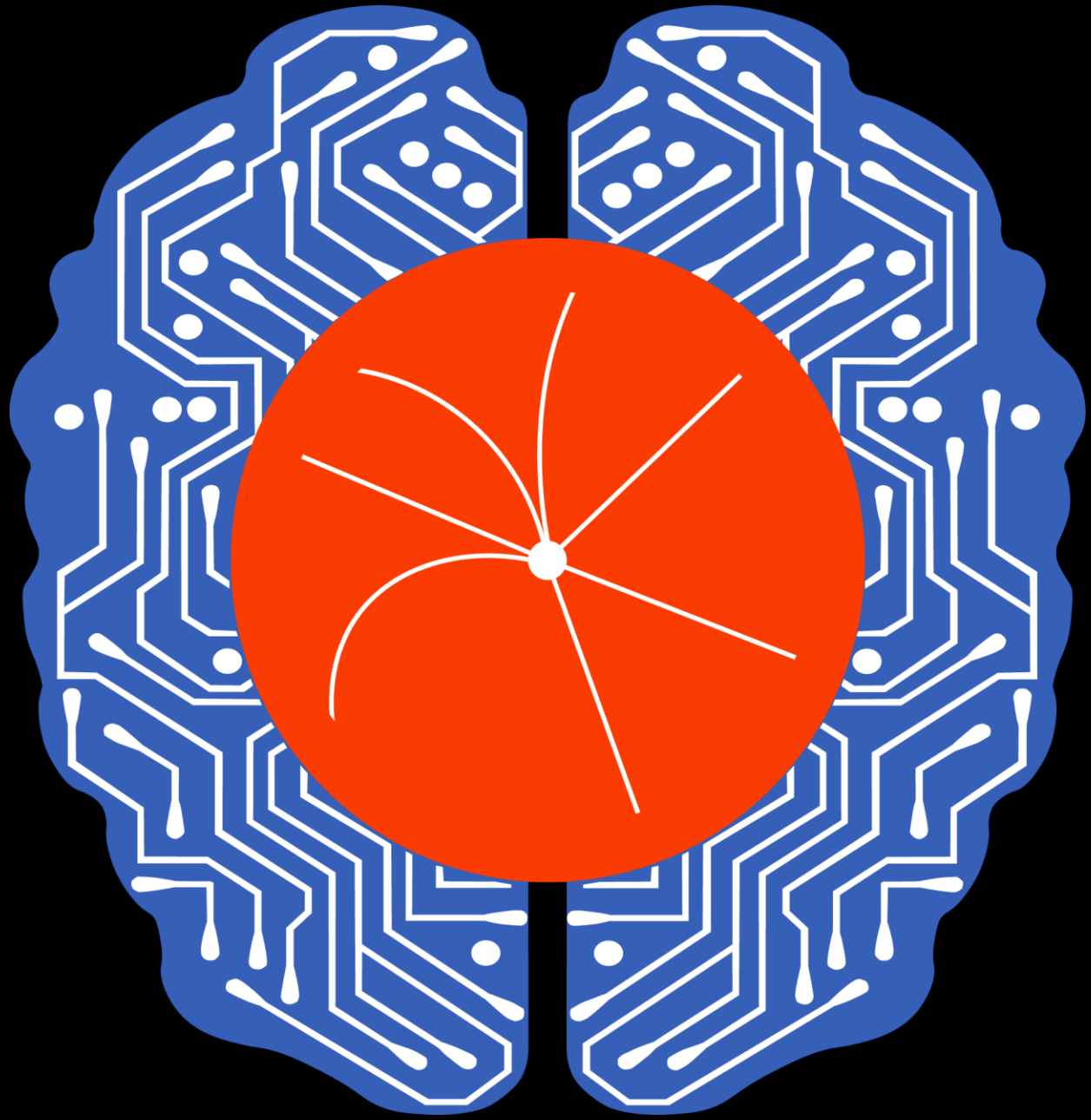


Edge SpAIce, S. Summers

- MLPerf tinyML benchmarking
- For fusion science phase/mode monitoring
- Crystal structure detection
- Triggering in DUNE
- Accelerator control
- Magnet Quench Detection
- Food contamination detection
- Quantum control etc....

FastML Lab

Keep in touch with community by signing up to the e-group hls-fml@cern.ch (sign up here!) and attending meetings (listed here)



BACKUP

Smart Pixels

Currently being built

Technology	65nm CMOS	28nm CMOS
Pixel ROIC size	50x50 μm^2	25x25 μm^2
Pixel Sensor size	100x25	50x12.5 μm^2
Pixels	394x400 = 157.6k	788x800 = 0.63M
Detection threshold	~1000e-	~500e-
Hit rate	< 3GHz/cm^2	< 3GHz/cm^2
Readout rate	1MHz	40MHz (?)
Digital buffer	12.5 µs	(?)
Readout Bandwidth	1-4 links @ 1.28Gbps	Photonic link @ 30-100 Gbps
Radiation tolerance	500Mrad at -15°C	1Grad at -15°C
Power	1 W /cm^2	1 W /cm^2

J. Dickinson

Jieun Yoo et al 2024 Mach. Learn.: Sci. Technol. 5 035047