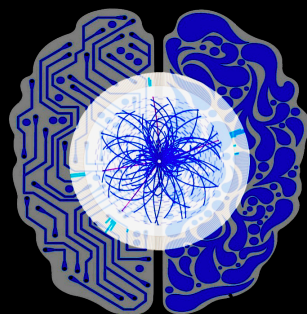
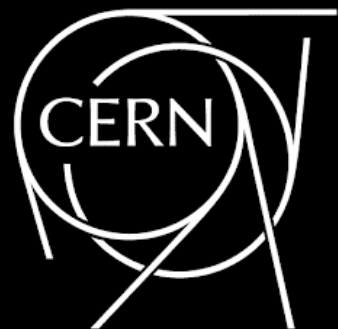


Experimental highlights: Edge AI for real-time systems in HEP

Jennifer Ngadiuba (Fermilab)

ML4Jets 2024
LPNHE, Paris
November 4-8, 2024



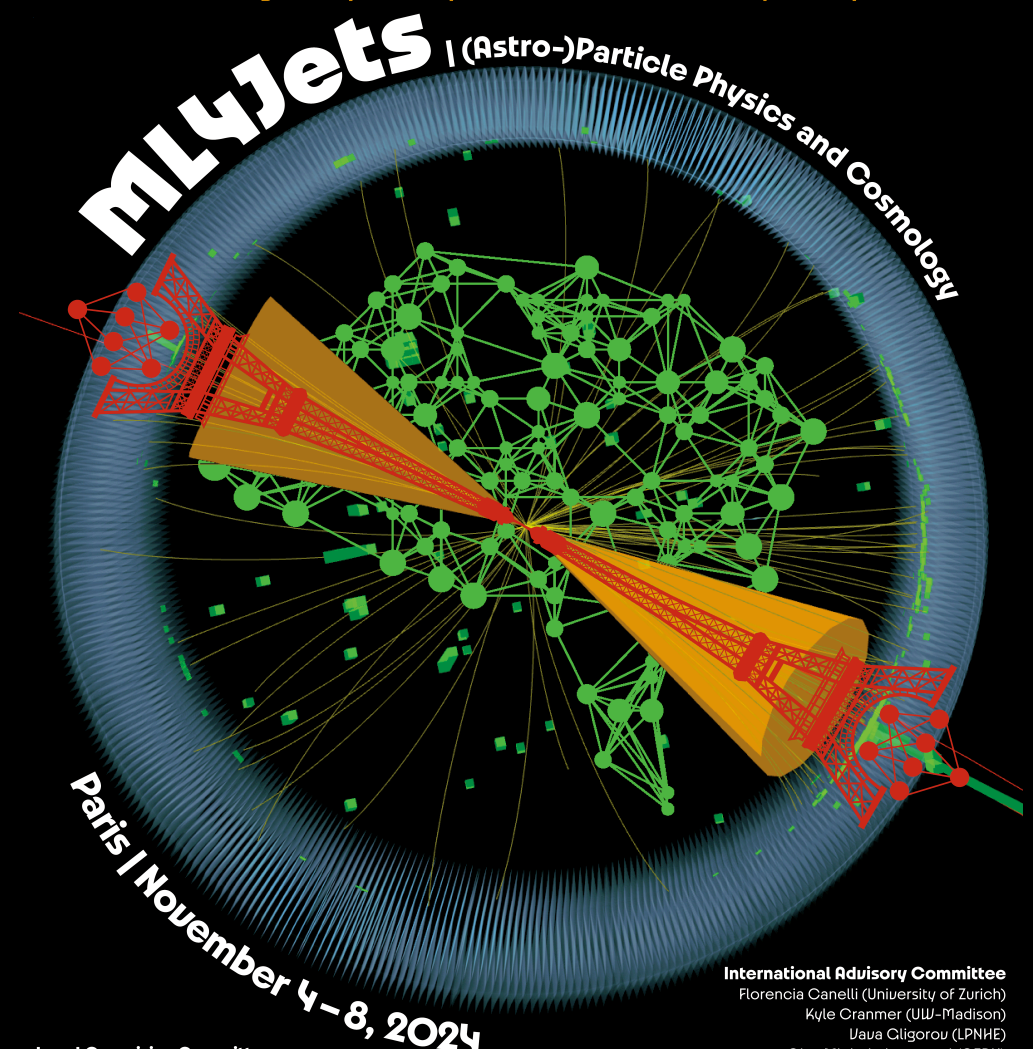
FastML Lab



NextGen
Next Generation Triggers

Fermilab

Tagging ■ Reconstruction ■ Detector Simulation ■ Event Generation ■ Astrophysics
Unfolding ■ Theory ■ Anomaly Detection ■ Uncertainties ■ Interpretability



Local Organizing Committee

Anja Butter (LPNHE)
Reina Camacho (LPNHE)
Benjamin Fuks (LPTHE)
Nabil Carroum (LPNHE)
Mark Goodsell (LPTHE)
Bertrand Laforge (LPNHE)
Bogdan Malaescu (LPNHE)
David Rousseau (IJCLab)



<https://indico.cern.ch/event/1386125/>

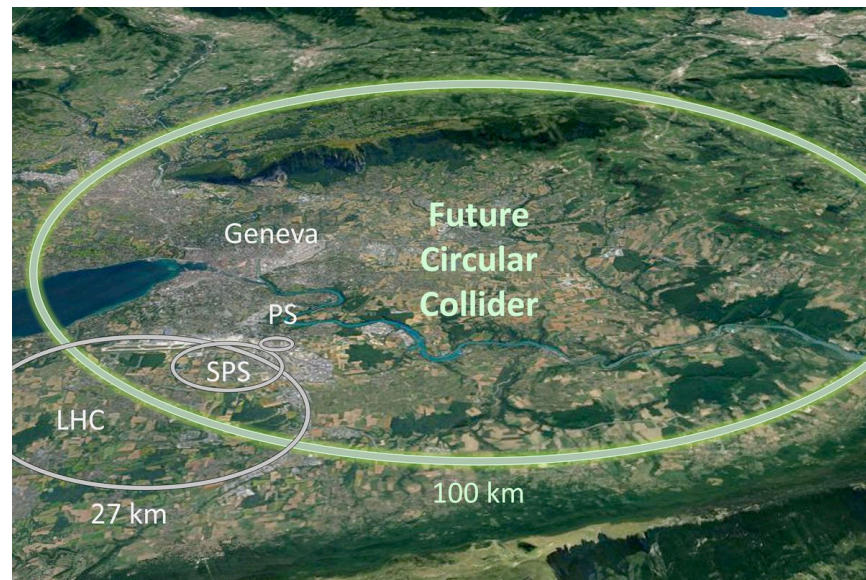
International Advisory Committee

Florencia Canelli (University of Zurich)
Kyle Cranmer (UW-Madison)
Uva Olgorou (LPNHE)
Gian Michele Innocenti (CERN)
Gregor Kasieczka (Hamburg)
Ben Nachman (LBNL)
Mihoko Nojiri (KEK)
Maurizio Pierini (CERN)
Tilman Plehn (Heidelberg)
David Shih (Rutgers)
Jesse Thaler (MIT)
Sofia Vallecorsa (CERN)



Big Science in 21st century

Probing the **fundamental structure of nature** requires complex experimental devices, large infrastructures and big collaborations.



The Large Hadron Collider

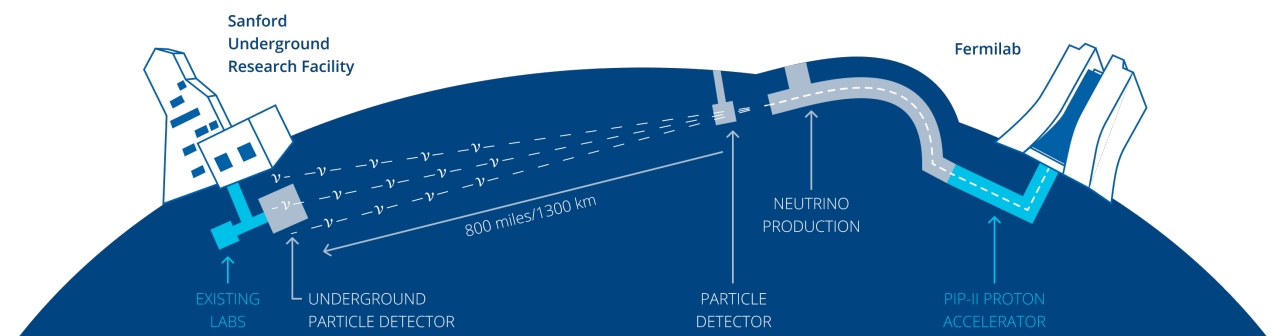


LIGO/VIRGO interferometers



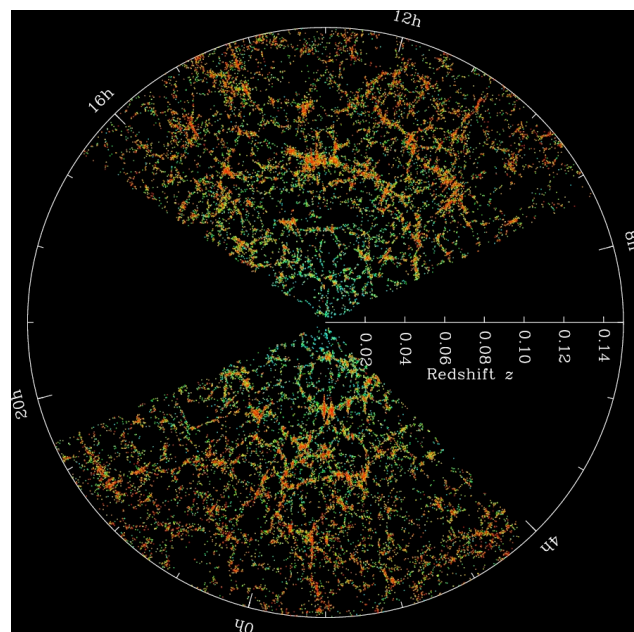
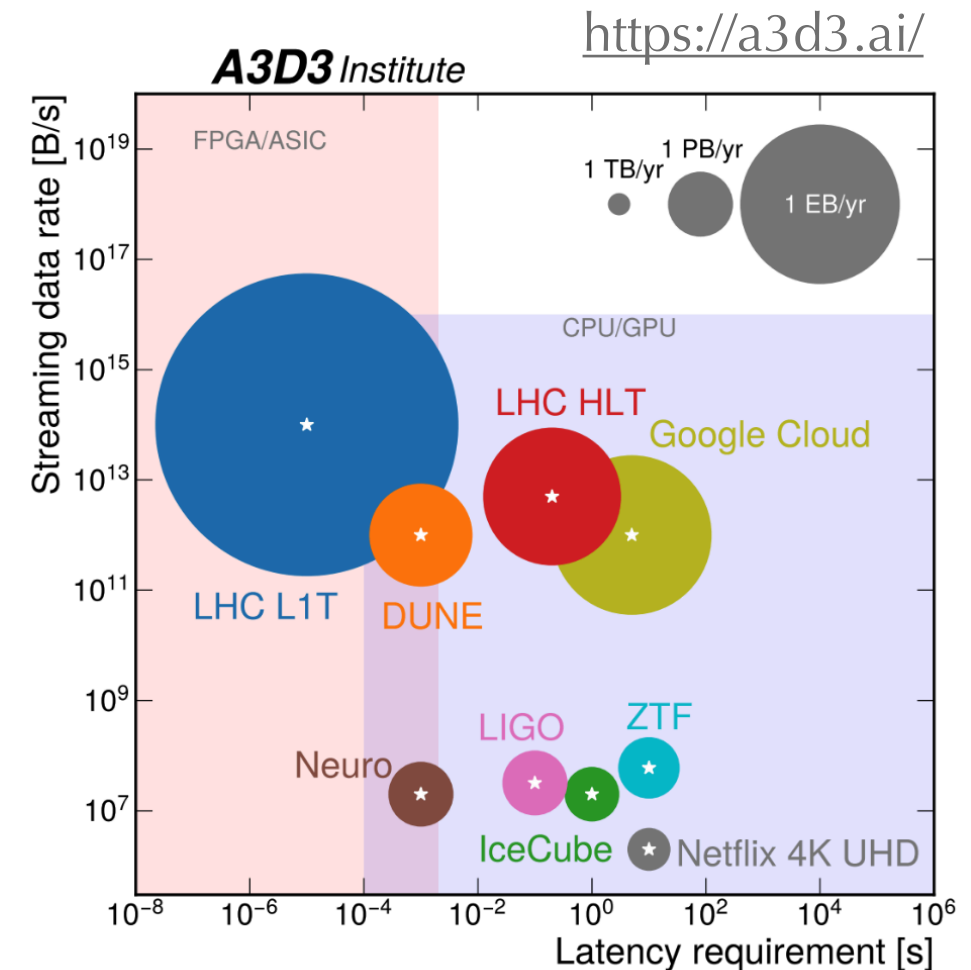
Vera C. Rubin Observatory

The DUNE neutrino experiment

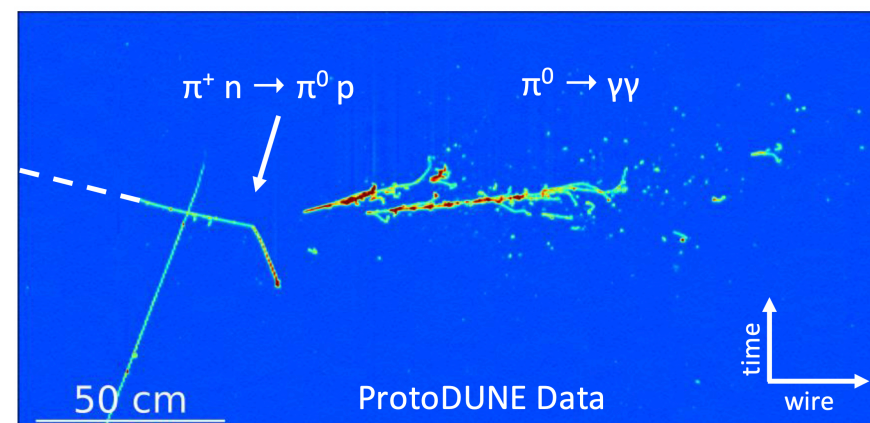


Big Science = Big Data

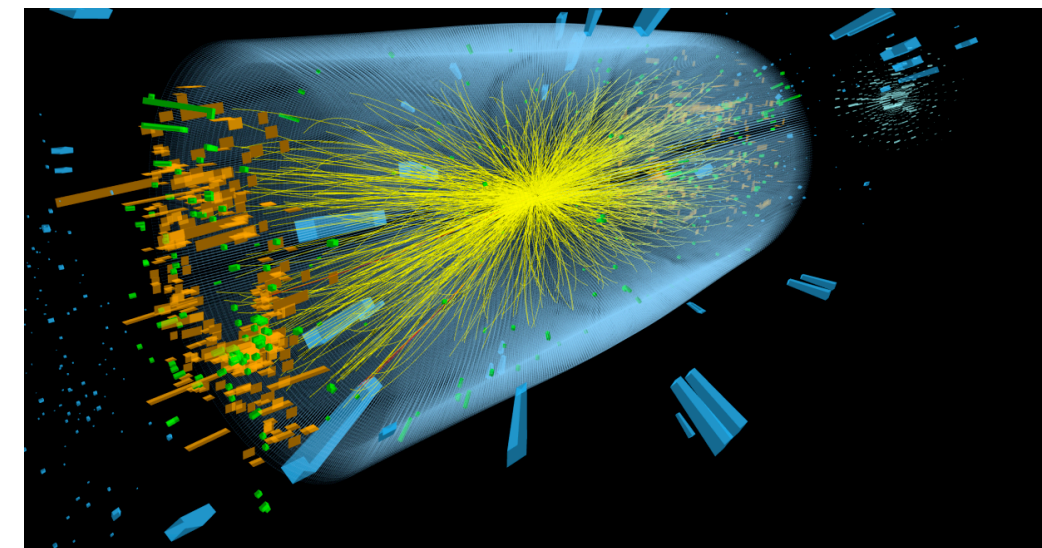
- Increasingly complex data both in **volume** and **dimensionality**
- Increasing need for **efficient and accurate data processing pipelines**
- Challenge in **simulating expectations** for what experiments may observe
- But also need for innovative **data & discovery driven** physics analyses approaches



Sloan Digital Sky Survey



Interactions in LArTPC

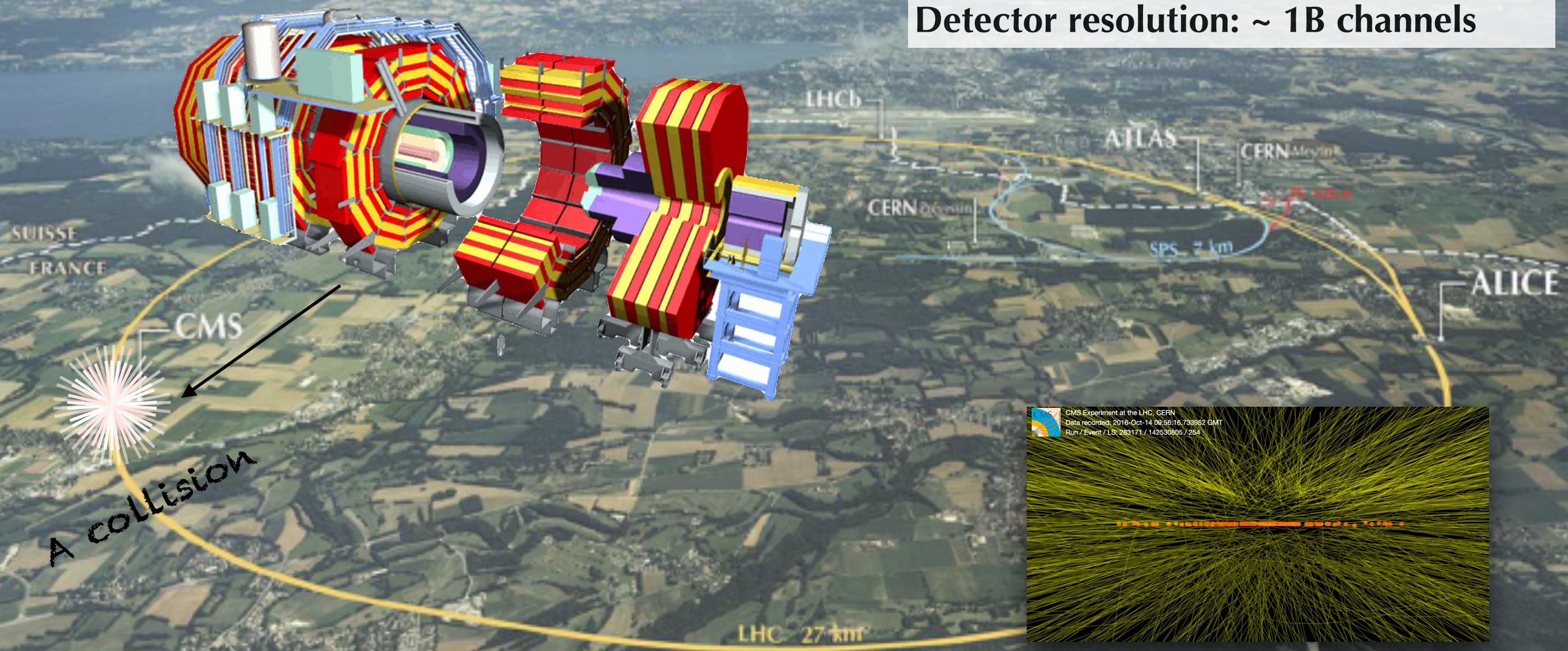


A LHC collision

Big Data @ the Energy Frontier

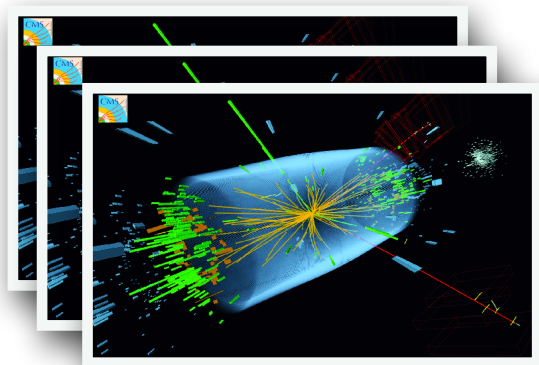
The Large Hadron Collider (LHC)

Collision frequency: 40 MHz
Particles per collision: $O(10^3)$
Detector resolution: $\sim 1\text{B}$ channels

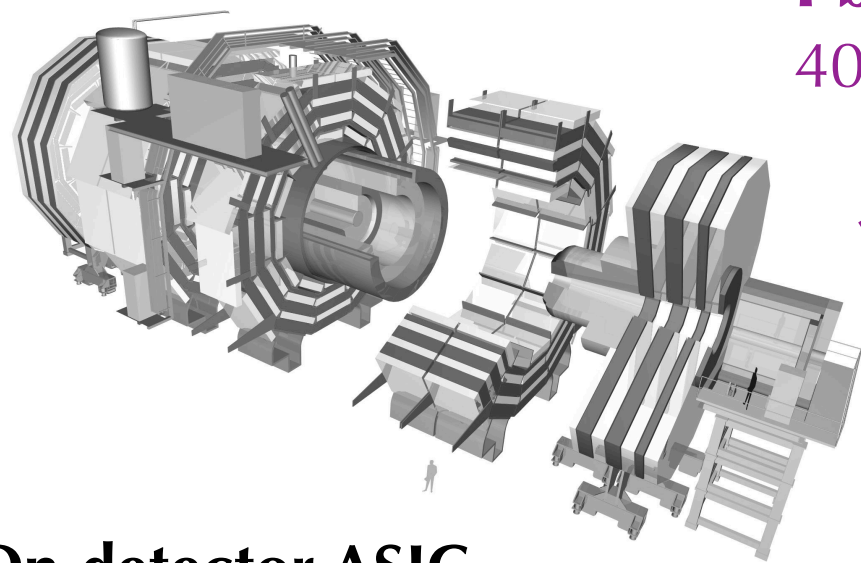


Extreme data rates of $\sim\text{Pb/s}$!

Data reduction workflow @ LHC



CMS Experiment
40 MHz collision rate
~1B detector channels



**On-detector ASIC
compression**
~100 ns latency

Pb/s
40 MHz

FPGA filter stack
~ μ s latency

**Level-1
Trigger**

10s Tb/s
100s kHz

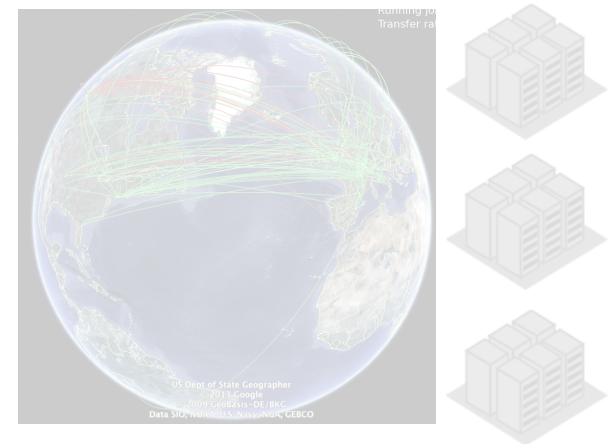
10s Gb/s
~5 kHz

**Offline
analysis**

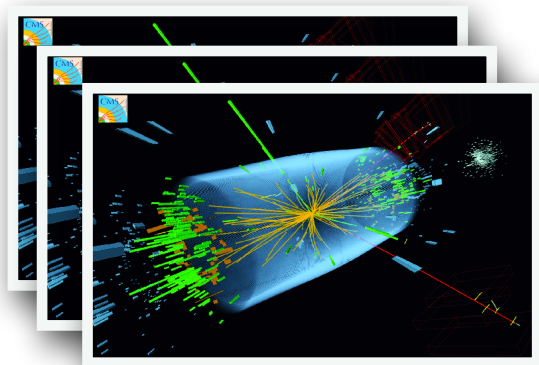
**High-Level
Trigger**

On-prem CPU/GPU filter farm
~100 ms latency

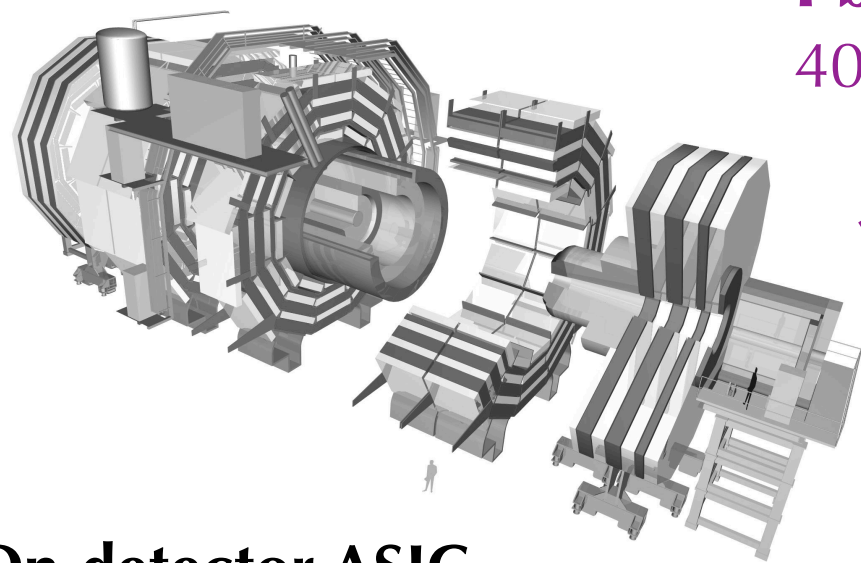
Worldwide
computing grid
Exabyte-scale
datasets



Data reduction workflow @ LHC

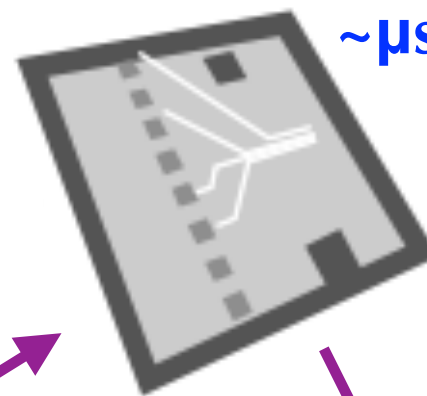


CMS Experiment
40 MHz collision rate
~1B detector channels



**On-detector ASIC
compression**
~100 ns latency

Pb/s
40 MHz



FPGA filter stack
~ μ s latency

**Level-1
Trigger**

10s Tb/s
100s kHz

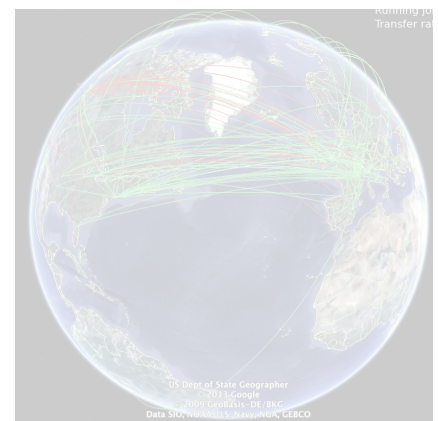


**High-Level
Trigger**

On-prem CPU/GPU filter farm
~100 ms latency

10s Gb/s
~5 kHz

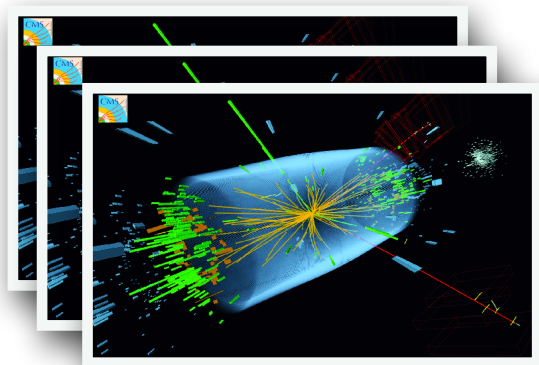
**Offline
analysis**



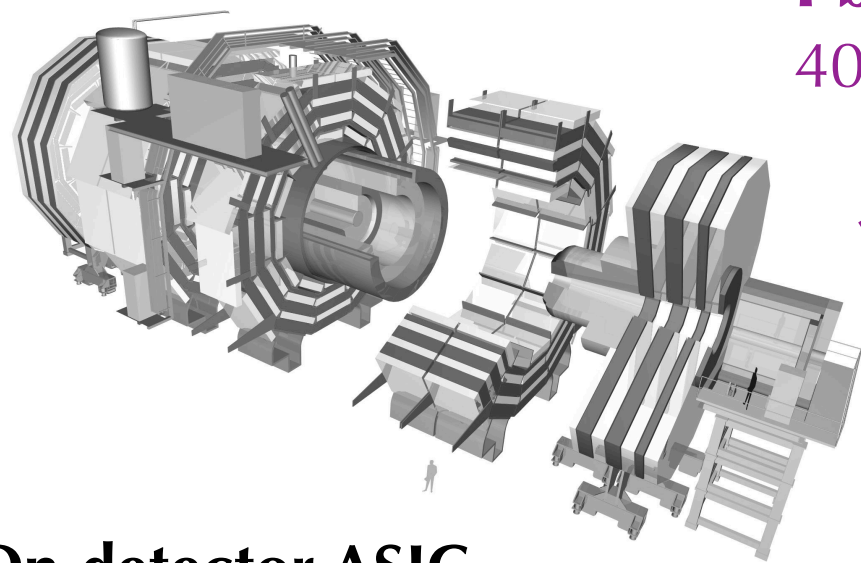
Worldwide
computing grid
Exabyte-scale
datasets



Data reduction workflow @ LHC



CMS Experiment
40 MHz collision rate
~1B detector channels



**On-detector ASIC
compression**
~100 ns latency

Pb/s
40 MHz

FPGA filter stack
~ μ s latency

**Level-1
Trigger**

10s Tb/s
100s kHz

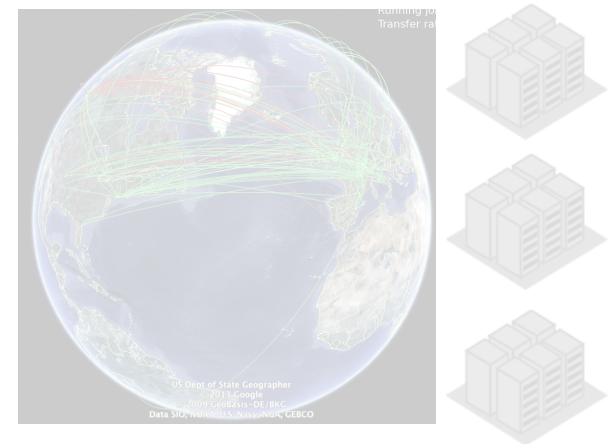
10s Gb/s
~5 kHz

**Offline
analysis**

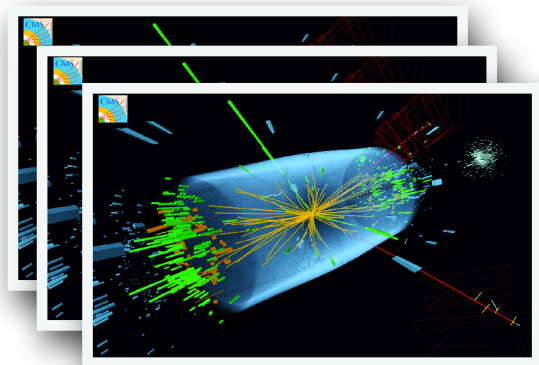
**High-Level
Trigger**

On-prem CPU/GPU filter farm
~100 ms latency

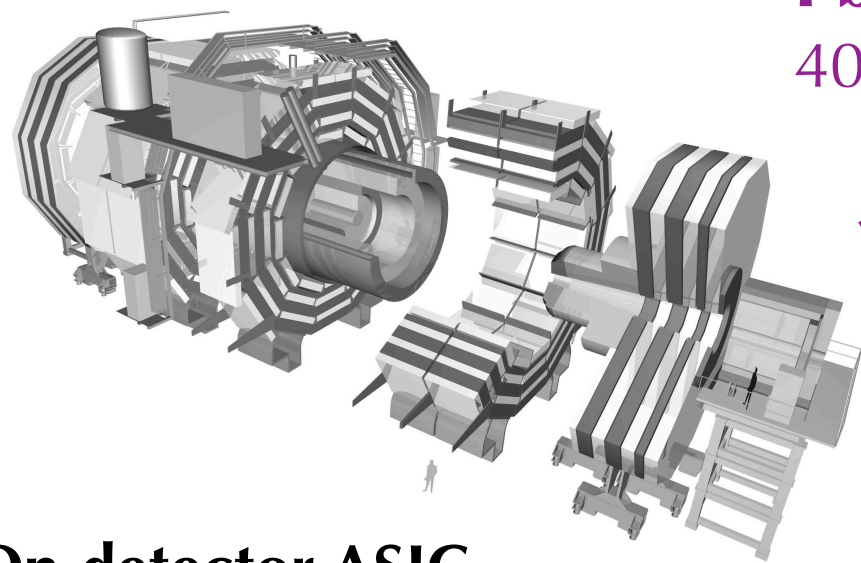
Worldwide
computing grid
Exabyte-scale
datasets



Data reduction workflow @ LHC



CMS Experiment
40 MHz collision rate
~1 B detector channels



**On-detector ASIC
compression**
~100 ns latency

Pb/s
40 MHz

FPGA filter stack
~ μ s latency

**Level-1
Trigger**

10s Tb/s
100s kHz

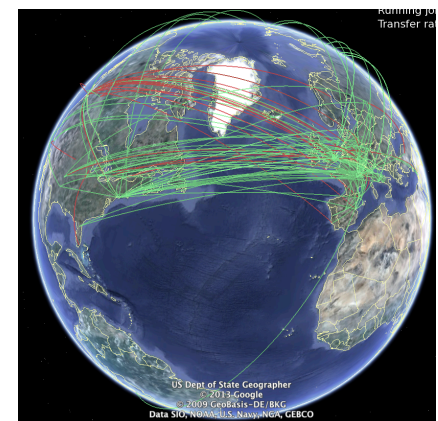
10s Gb/s
~5 kHz

**Offline
analysis**

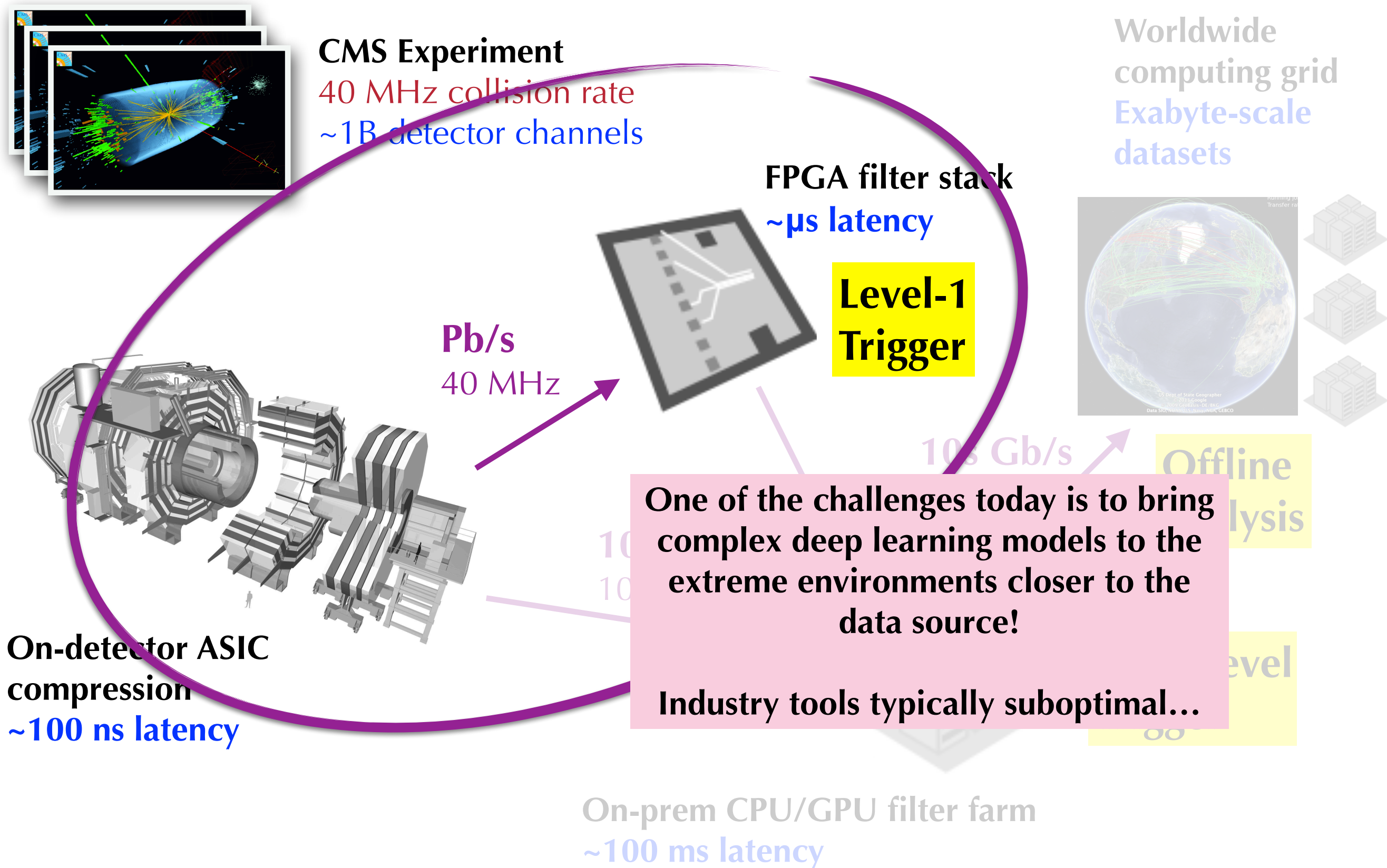
**High-Level
Trigger**

On-prem CPU/GPU filter farm
~100 ms latency

**Worldwide
computing grid**
Exabyte-scale
datasets

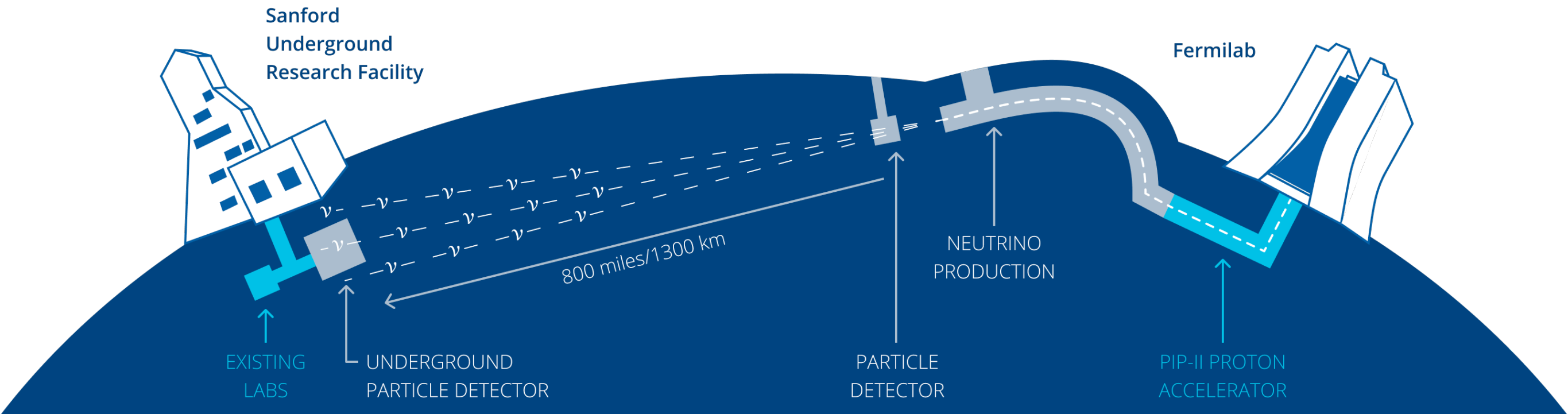


Data reduction workflow @ LHC



Big data @ the Intensity Frontier

The Deep Underground Neutrino Experiment (DUNE)



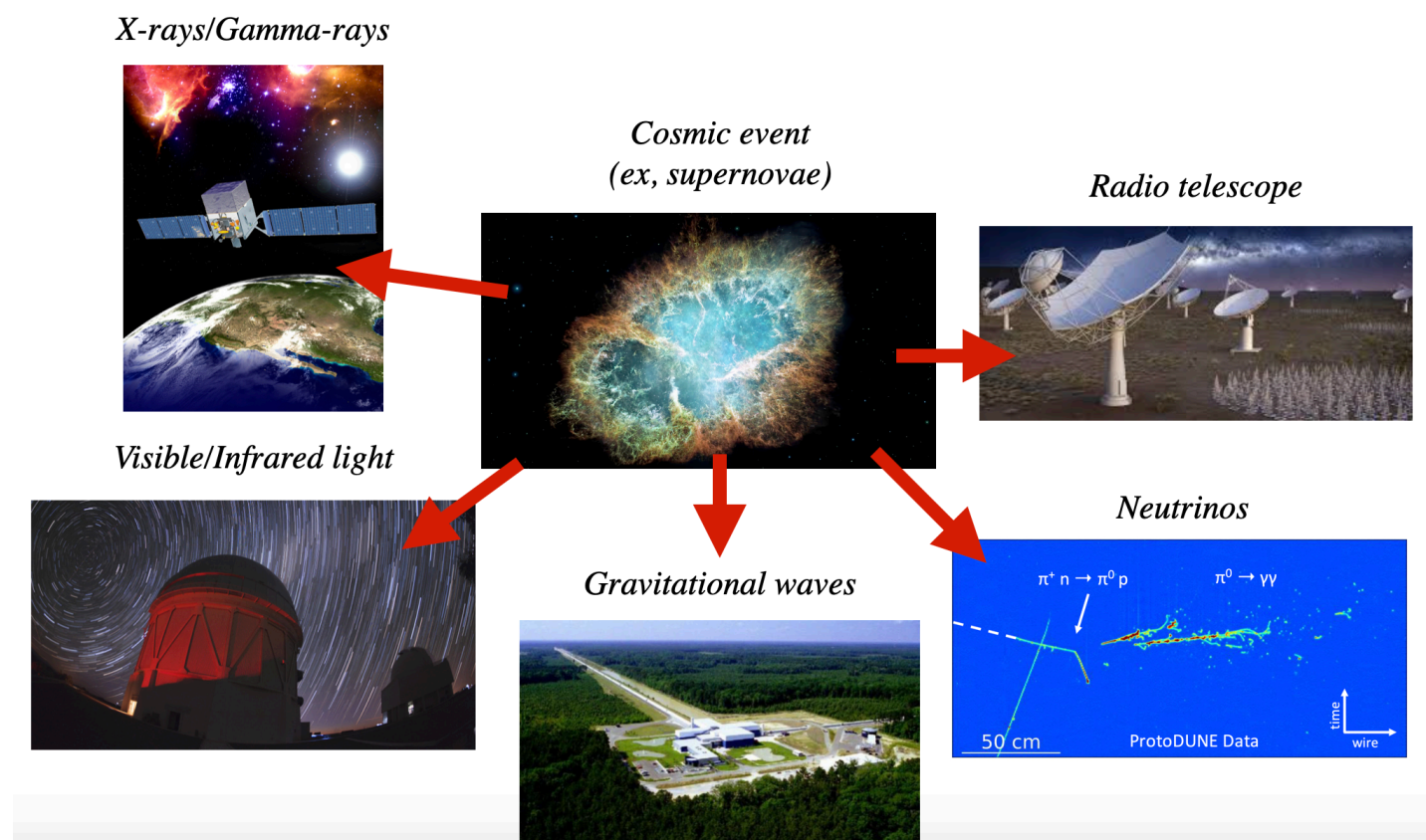
- Next generation neutrinos oscillation experiment now under construction and R&D to start operations in late 2020s
- Massive far detector 1 mile underground comprising **70k tons of Liquid Argon** and advanced technology to record neutrino interactions with extraordinary precision
- Uncompressed continuous readout of modules will yield **O(100) Tb/s** → unprecedented for this type of experiment!

Multi-messenger astronomy

Multi-messenger astronomy probes the Universe using different cosmic messengers

- Two notable examples:

- **neutron star merger (GW170817):**
gravitational wave (Ligo/Virgo) + electromagnetic signal (Fermi and INTEGRAL telescopes)
- **blazar (TXS 0506+056):**
high-energy neutrino (IceCube) + electromagnetic signals (Fermi, MAGIC and others)

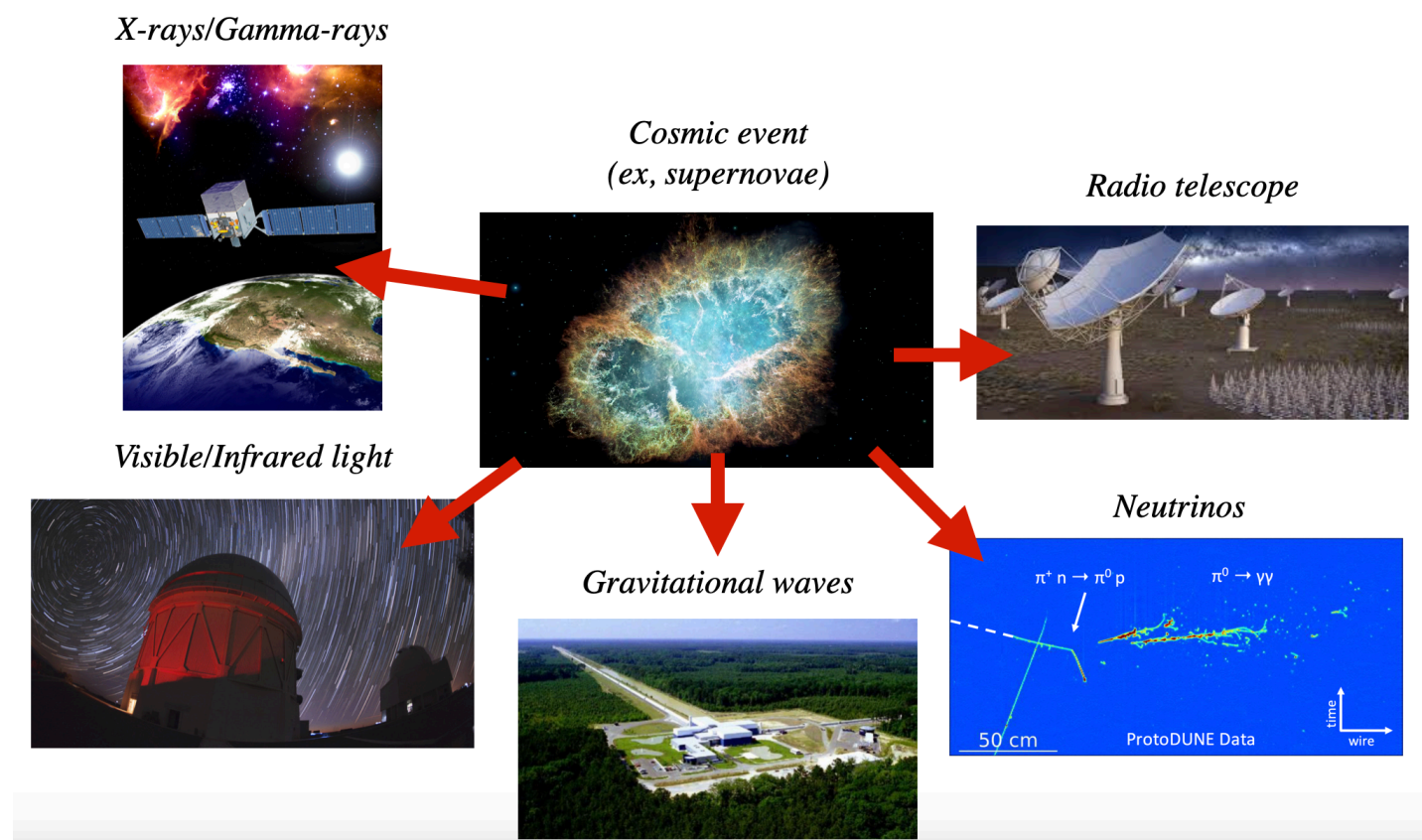


Multi-messenger astronomy

Multi-messenger astronomy probes the Universe using different cosmic messengers

- Two notable examples:

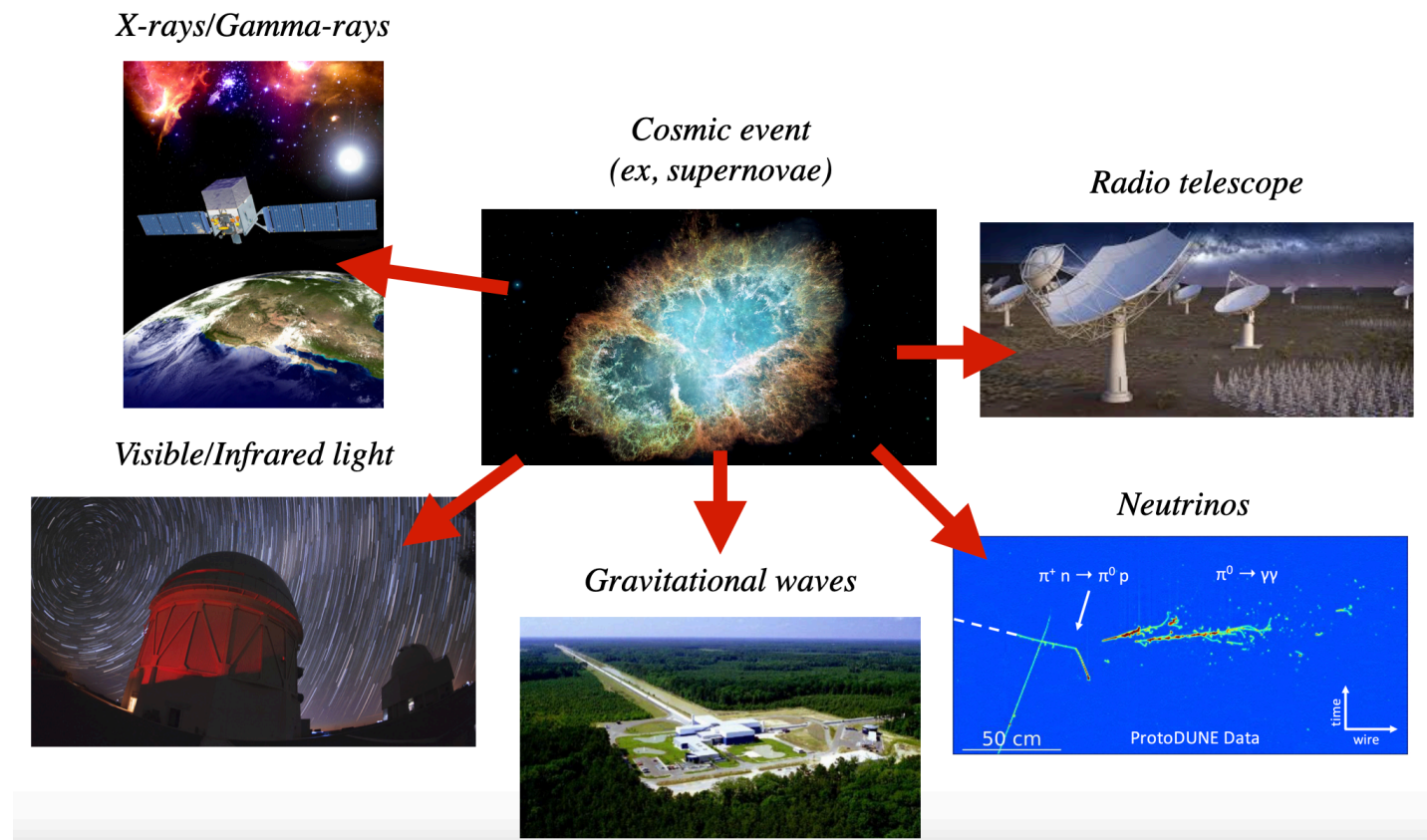
- **neutron star merger (GW170817):**
gravitational wave (Ligo/Virgo) + electromagnetic signal (Fermi and INTEGRAL telescopes)
- **blazar (TXS 0506+056):**
high-energy neutrino (IceCube) + electromagnetic signals (Fermi, MAGIC and others)



- **Timing and pointing accuracy** crucial to deliver alerts for a potential cosmological event to many different instruments around the globe
- This to become challenging with increase in size and sensitivity to a larger volume of space

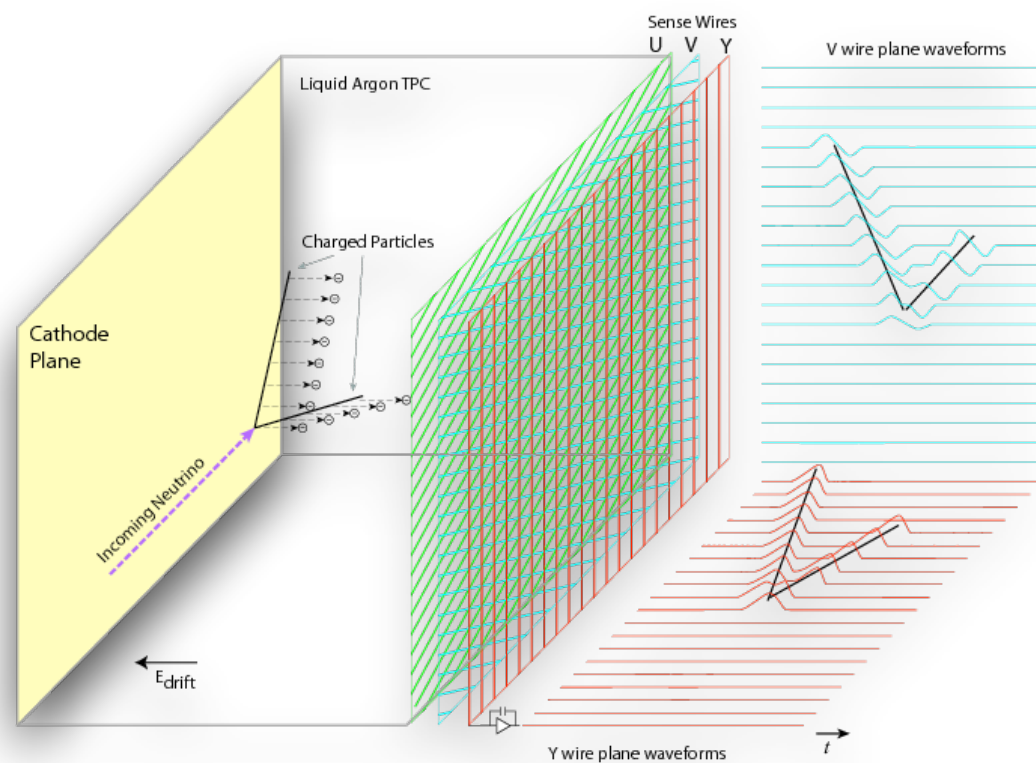
Multi-messenger astronomy w/ Neutrinos

- **Core-collapse supernovae are a huge source of neutrinos of all flavours**
 - 99% of energy released is carried away by neutrinos
- Rich information embedded in neutrino signal plus associated gravitational and electromagnetic signals
 - **supernova physics:** core-collapse mechanism, black hole formation, nucleosynthesis, ...
 - **particle physics:** flavor transformation in SN core, mass ordering, BSM...
- **Detection and pointing in real-time in large scale neutrino experiments is an active field of research!**

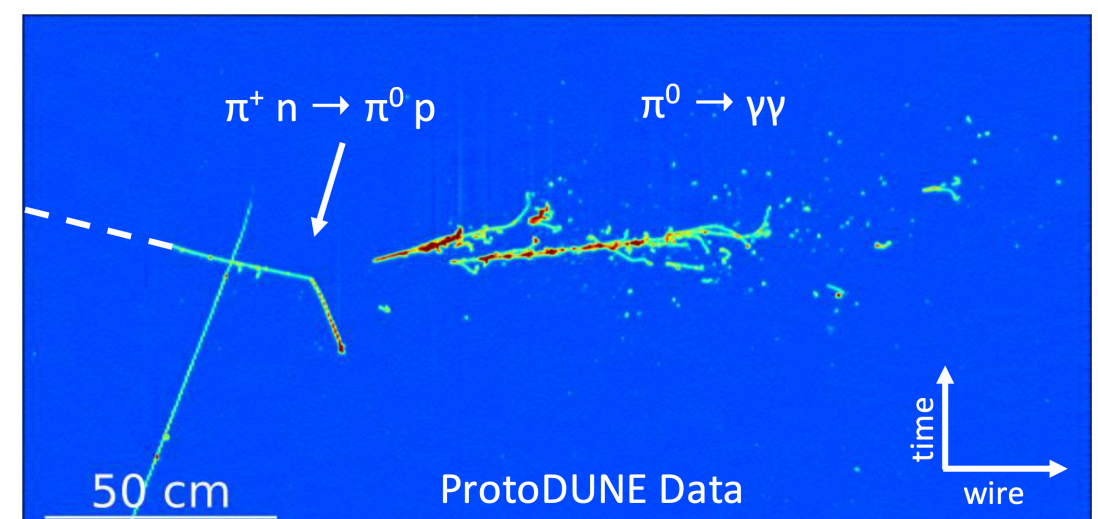
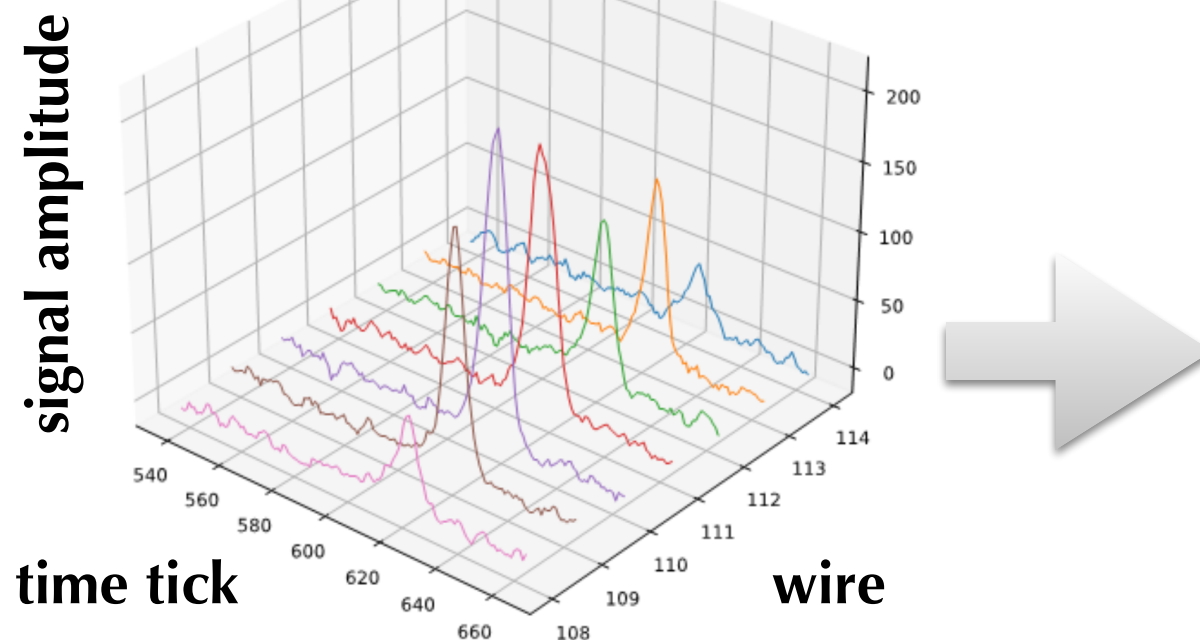


Big data @ the Intensity Frontier

Operating principle of a LArTPC

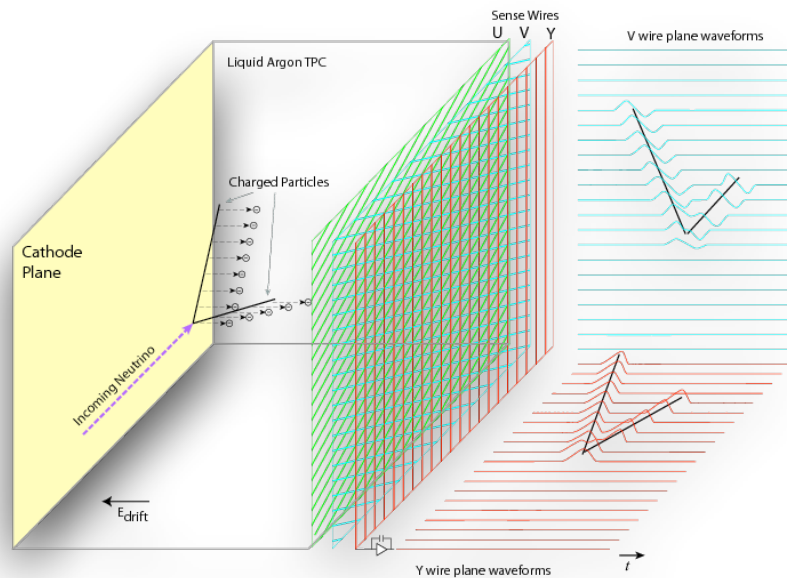


- Neutrinos interacting with the LAr produce charged particles, which in turn produce electrons
- Electrons are collected by anode wires
- The signal from each wire channel is a wave form
- There are 3 planes of wires for a full 3D reconstruction of the interaction
- The result is a continuous stream of 3D images of detector volume yielding a **high-resolution “video”**



Big data @ the Intensity Frontier

Detector South Dakota



4.8 TB/s
100 seconds: 480 TB
100 Gbps

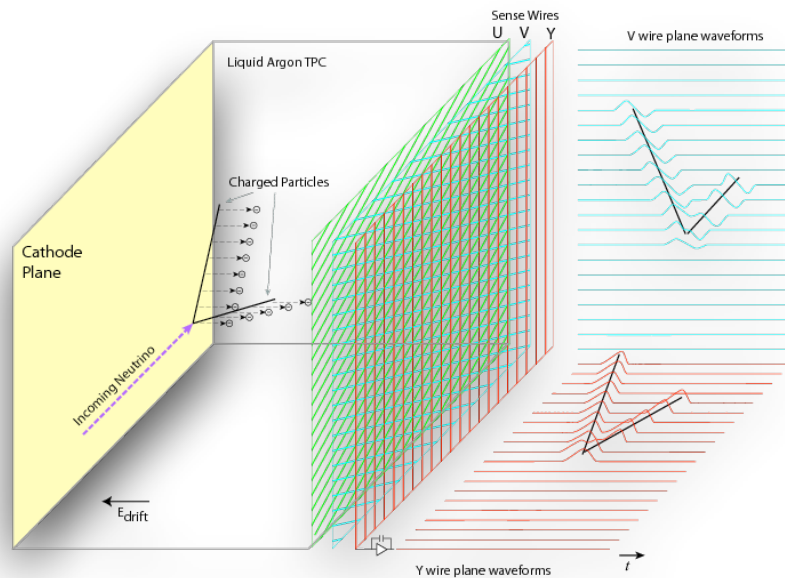


HPC Fermilab, Illinois



Big data @ the Intensity Frontier

Detector South Dakota



4.8 TB/s
100 seconds: 480 TB
100 Gbps

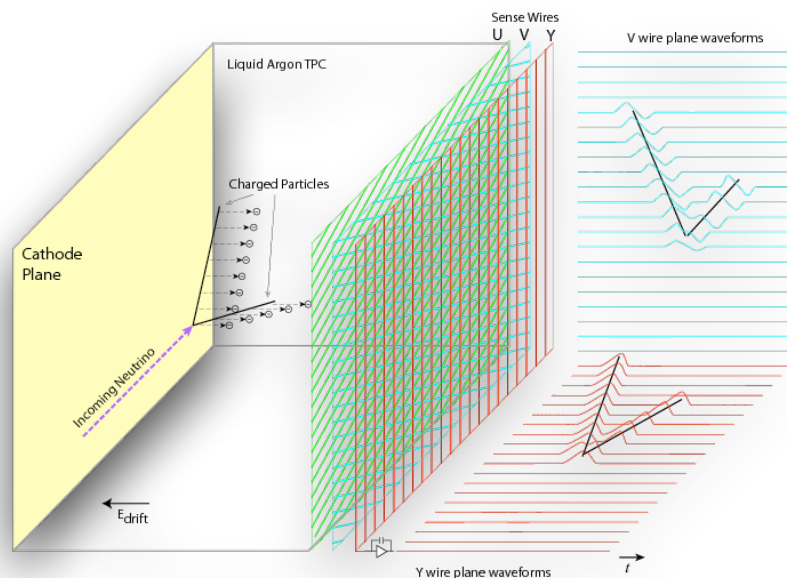
At least 12 hours before we
can detect a supernova and
reconstruct point of origin!

HPC Fermilab, Illinois



Big data @ the Intensity Frontier

Detector South Dakota



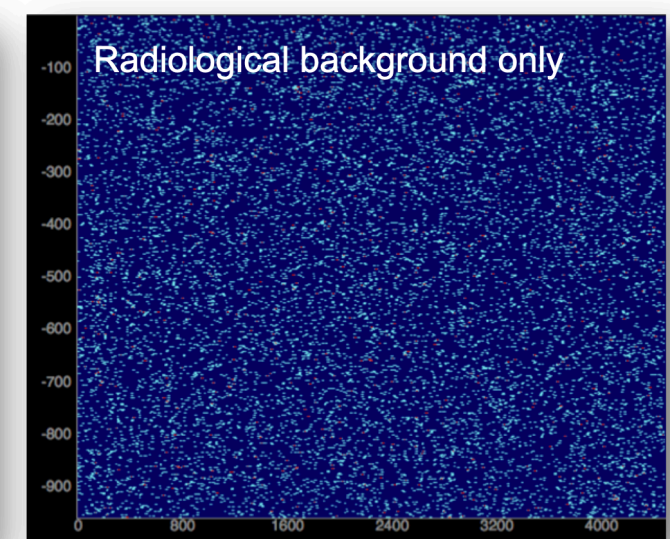
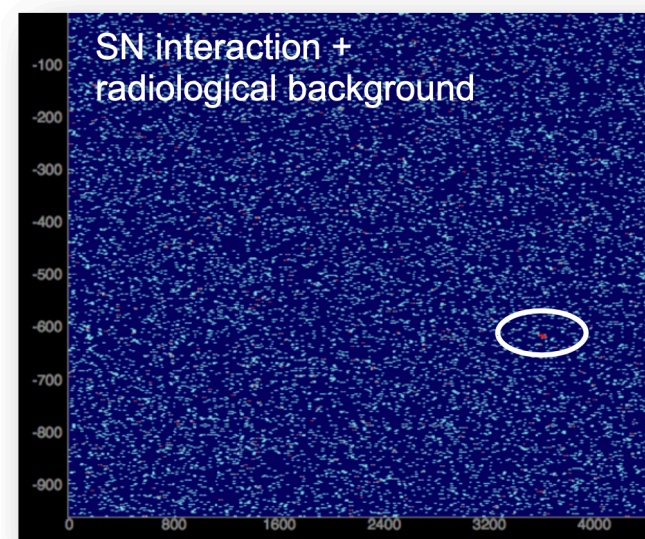
4.8 TB/s
100 seconds: 480 TB
100 Gbps

At least 12 hours before we
can detect a supernova and
reconstruct point of origin!

HPC Fermilab, Illinois



- Aggressive data reduction must happen underground **close to the data source**
- Must be smart as neutrinos from supernova are challenging → **Machine Learning**
- Very limited power underground requires dedicated hardware → **FPGAs**

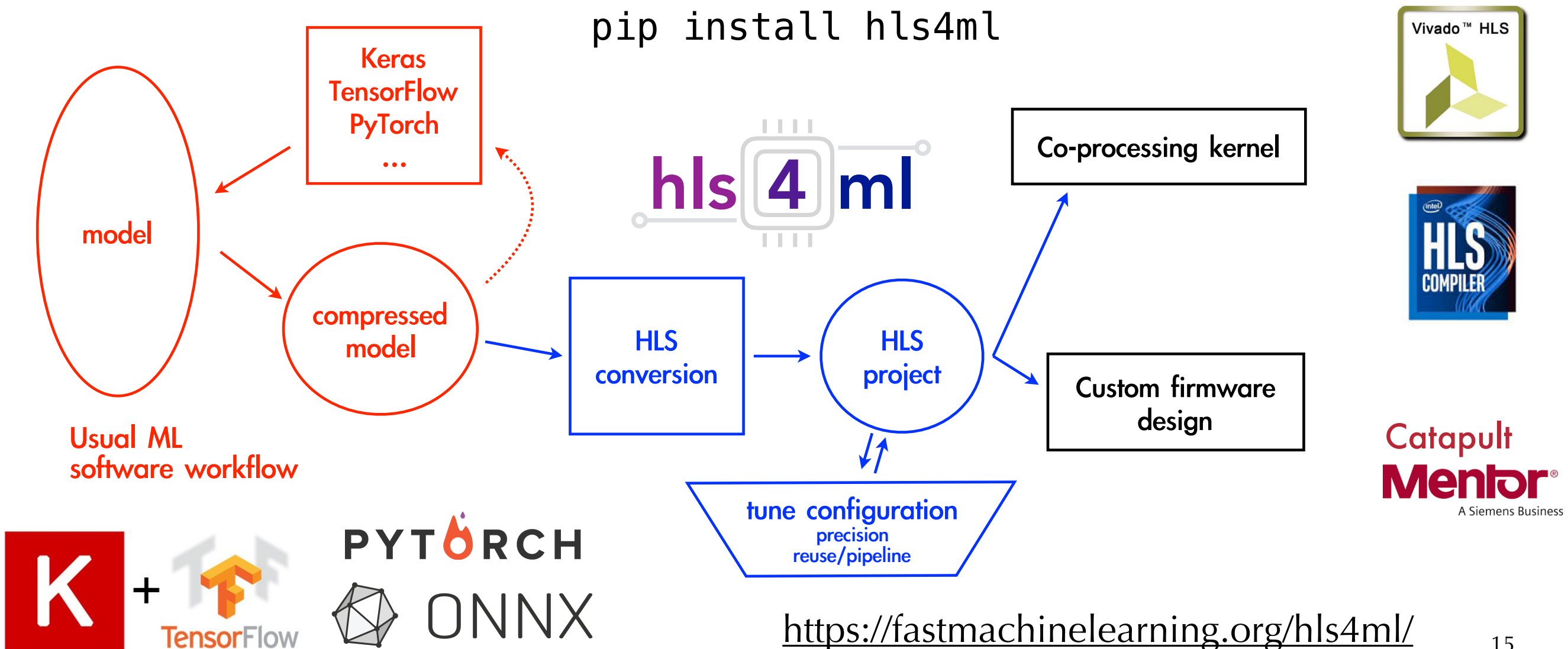


Bring ML models to hardware for real-time AI

high level synthesis for machine learning

**A tool to efficiently program the FPGA hardware for Neural Networks
with experimental constraints in mind!**

Many use cases in HEP and beyond... and still growing!
(see [Fast Machine Learning For Science Workshop Oct '24](#))



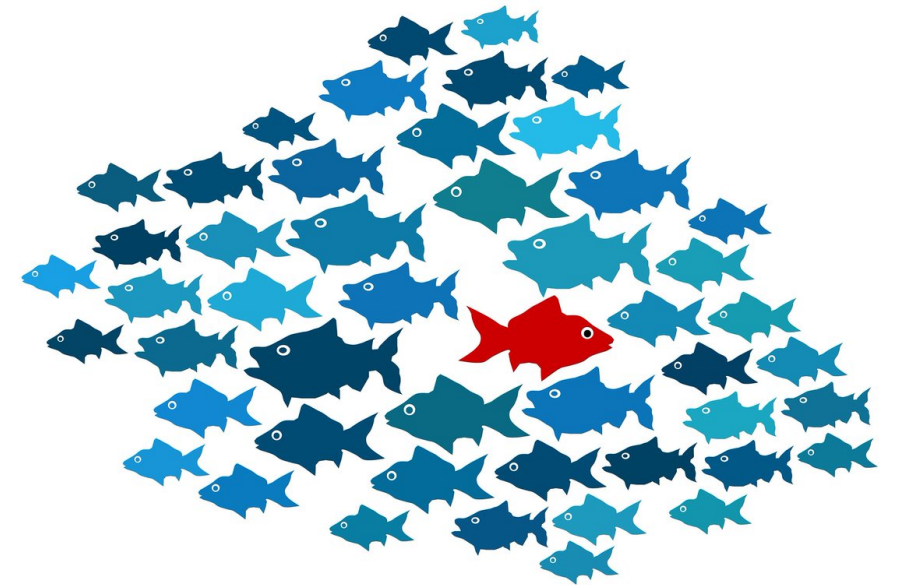
A few applications at the LHC

Anomaly detection

Machine learning based anomaly detection algorithms can be used to look at our data without model assumptions

Main idea: **learn directly from data** **how the standard model looks like**

⇒ **eliminate signal priors** and search for **anything anomalous** wrt standard model

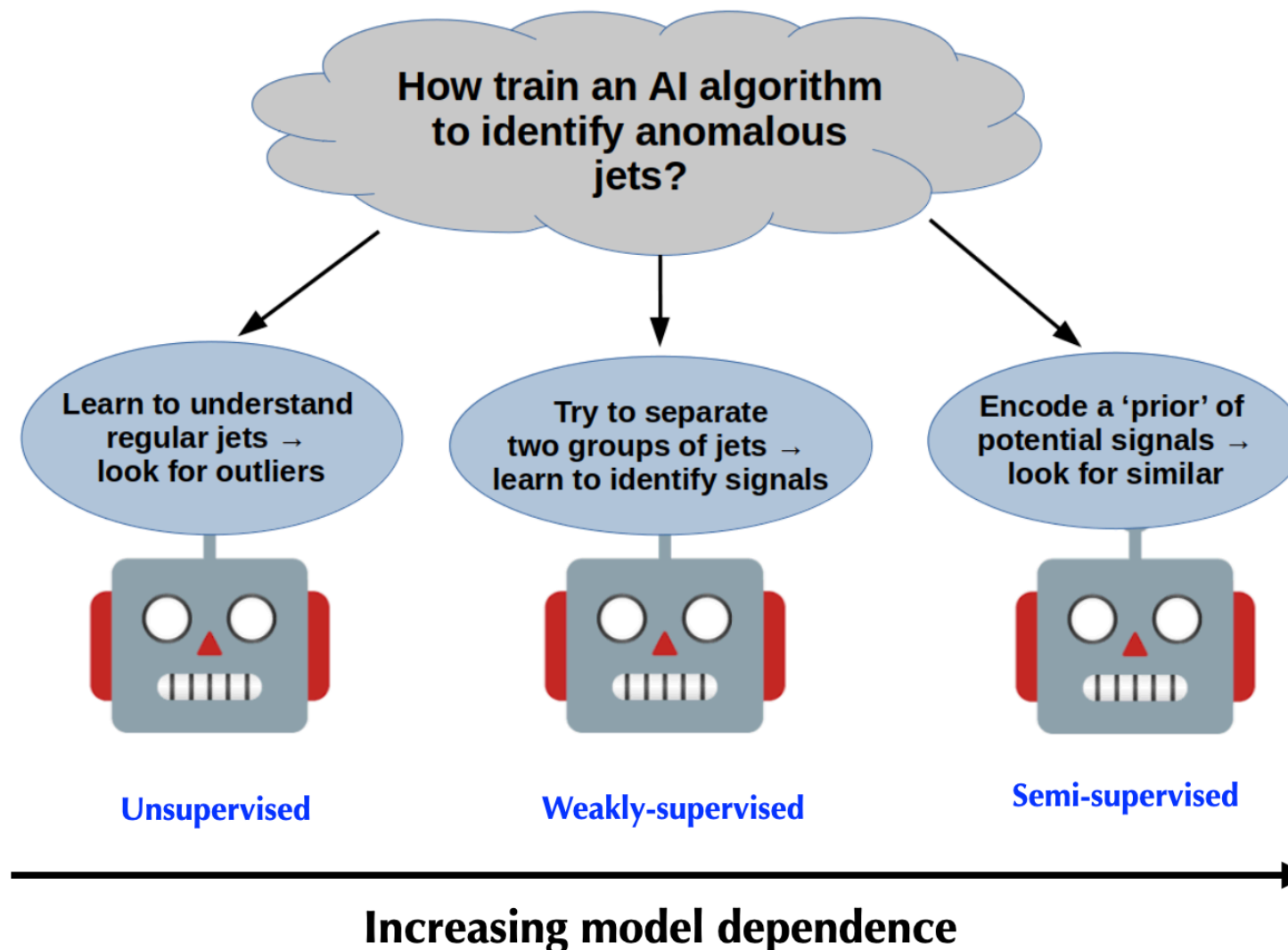
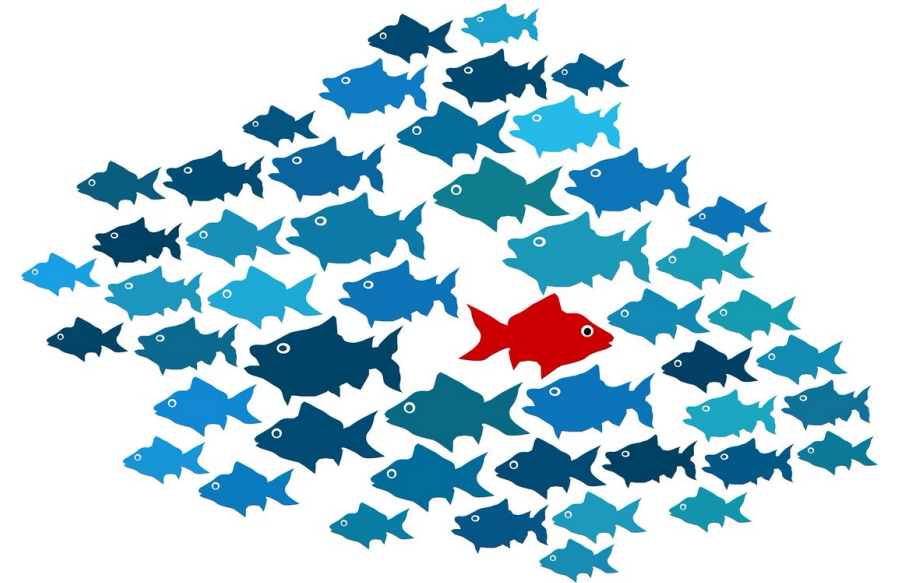


Anomaly detection

Machine learning based anomaly detection algorithms can be used to look at our data without model assumptions

Main idea: learn directly from data **how the standard model looks like**

⇒ eliminate signal priors and search for **anything anomalous** wrt standard model



**How train an AI algorithm
to identify anomalous
jets?**

**Learn to understand
regular jets →
look for outliers**



Unsupervised

**Two ATLAS searches using autoencoders
in final states with:**

- two boosted jets [\[PRD 108 \(2023\) 052009\]](#)
- lepton + jet(s) and photon + jet(s) [\[PRL 132 \(2024\) 081801\]](#)

Increasing model dependence

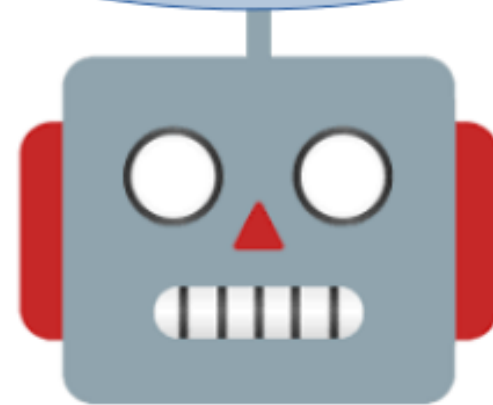
**How train an AI algorithm
to identify anomalous
jets?**

**Try to separate
two groups of jets →
learn to identify signals**

**One ATLAS search using
the CWoLA method:**

boosted dijet final state

[\[PRL 125, 131801 \(2020\)\]](#)



Weakly-supervised

Increasing model dependence

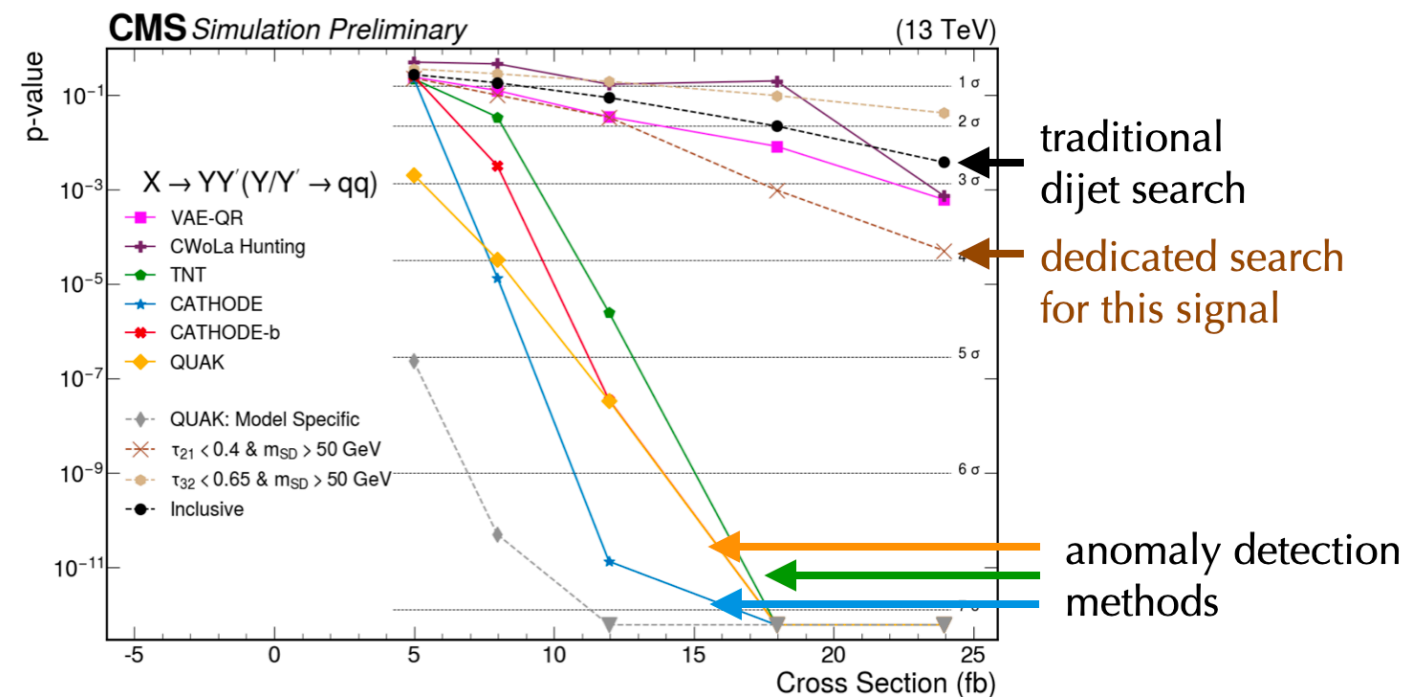
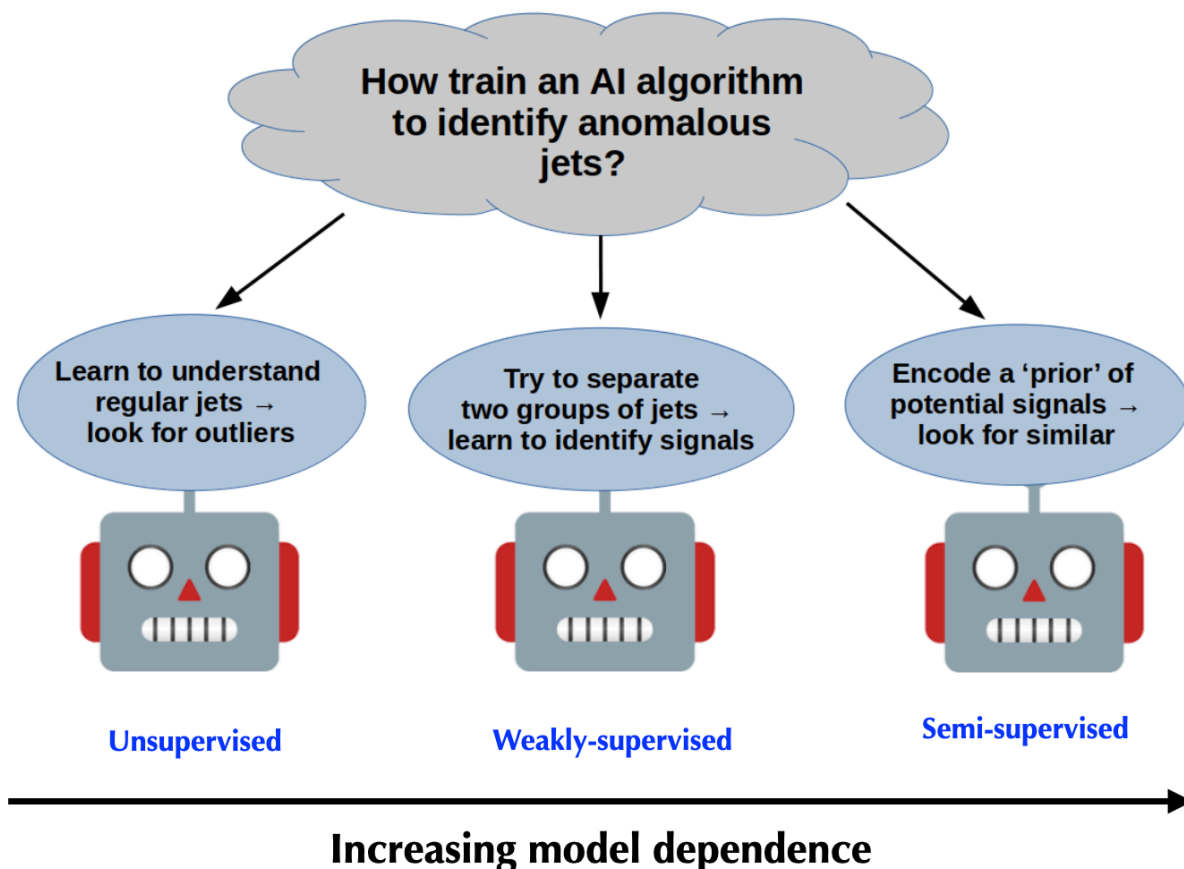
Anomaly detection

Model-agnostic search for dijet resonances with anomalous jet substructure in proton-proton collisions at $\sqrt{s} = 13$ TeV

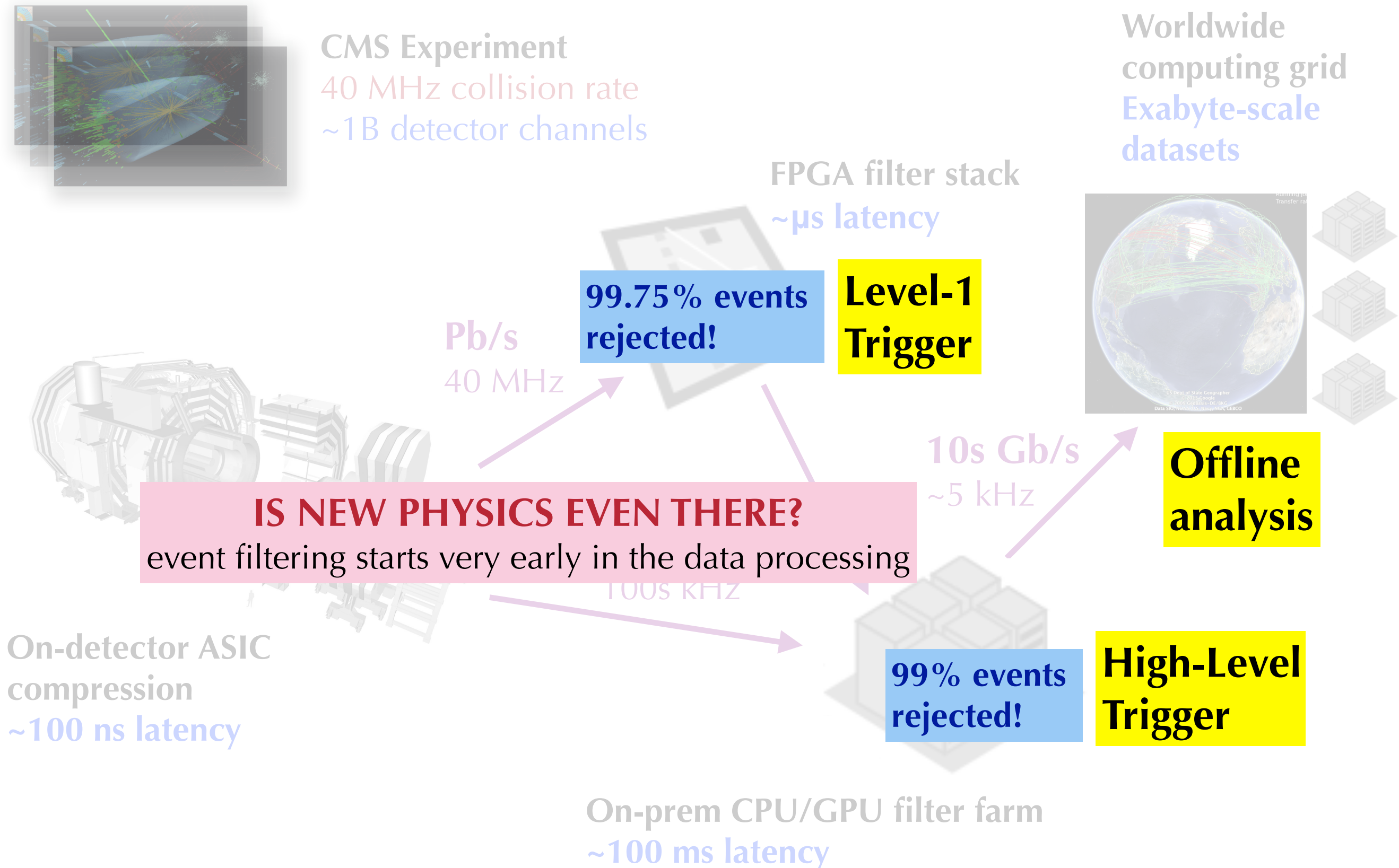
The CMS Collaboration

[CMS-PAS-EXO-22-026](#)

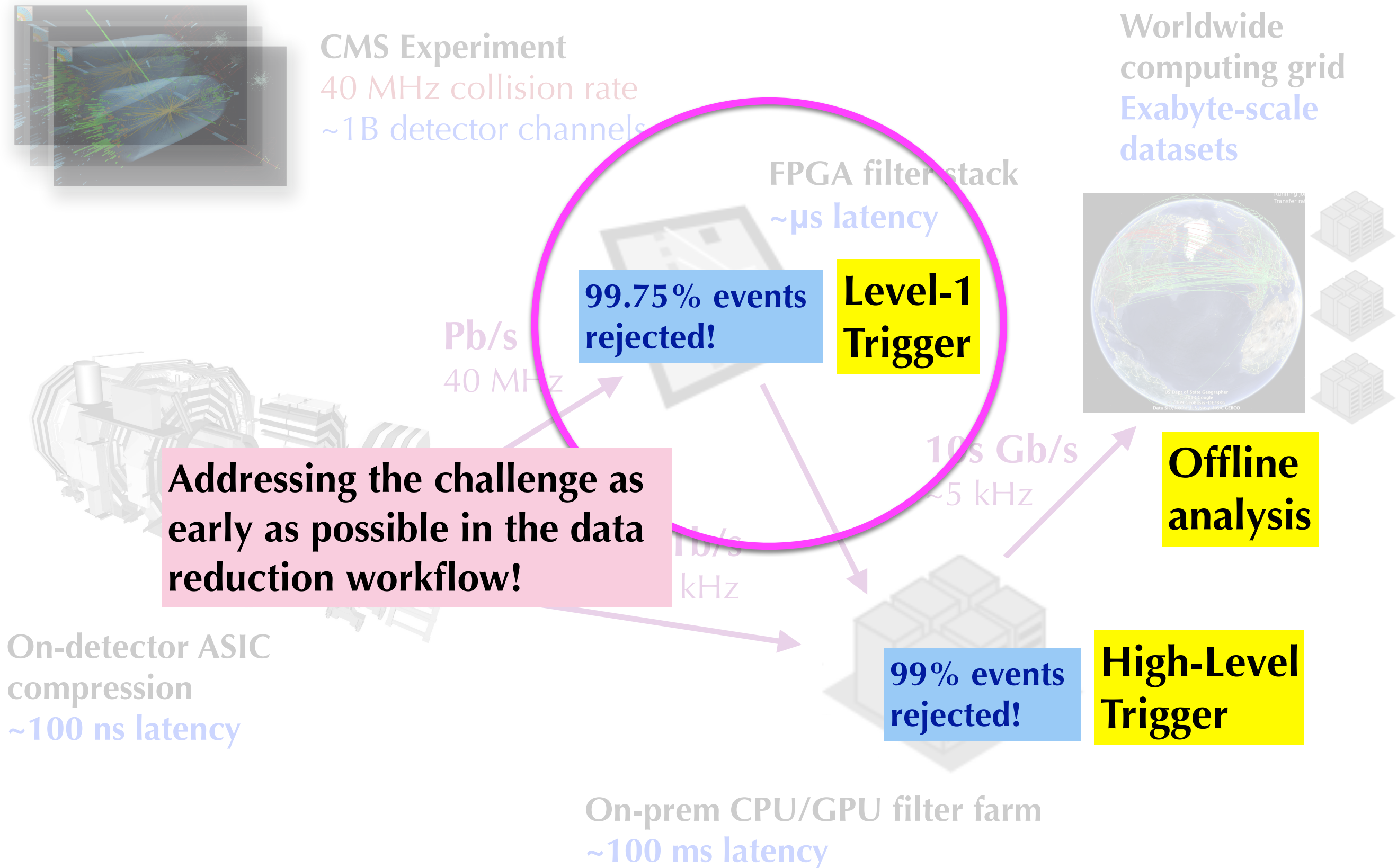
Use and compare all methods



Data reduction workflow @ LHC



Data reduction workflow @ LHC



Ultra-fast anomaly detection @ CMS

Learn typicality: by training on Zero Bias dataset

CMS establishing a new trigger paradigm with sub- μ s autoencoders for anomaly detection!



	p_T	η	ϕ
MET		N/A	
4 e/γ			
4 μ			
10 jets			

From calorimeter and muon trigger system:

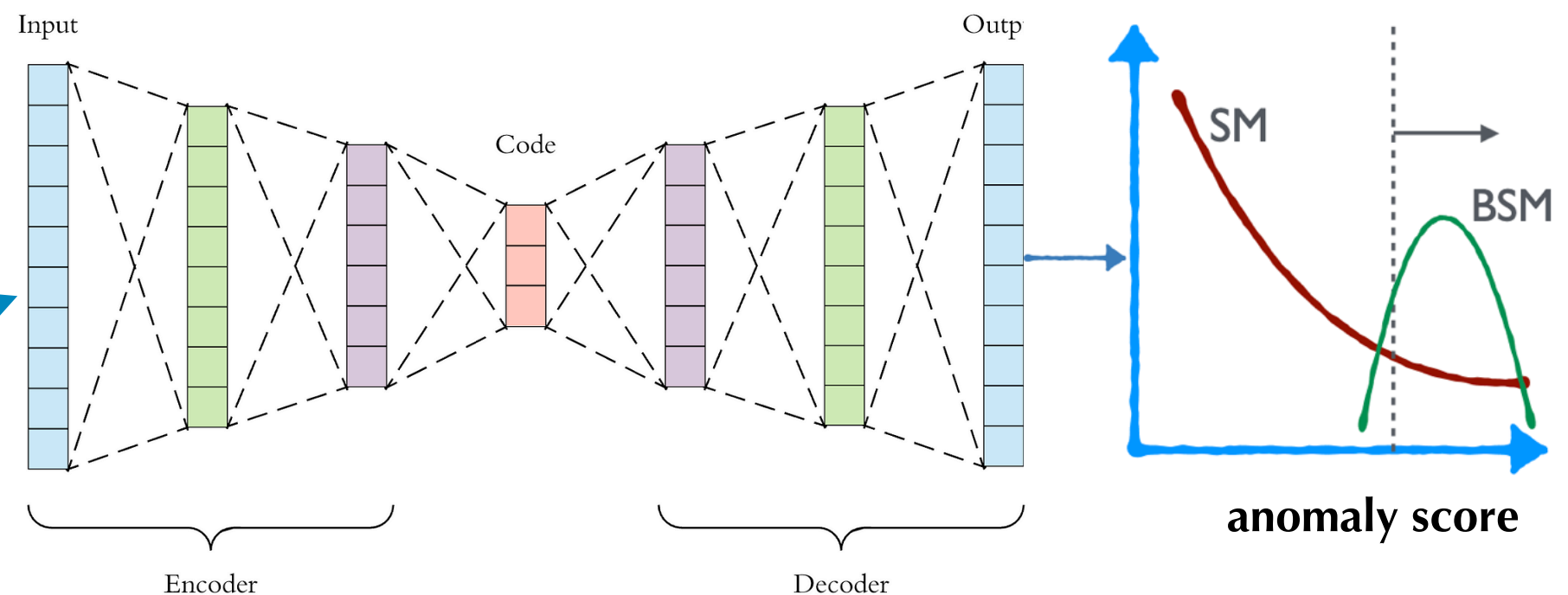
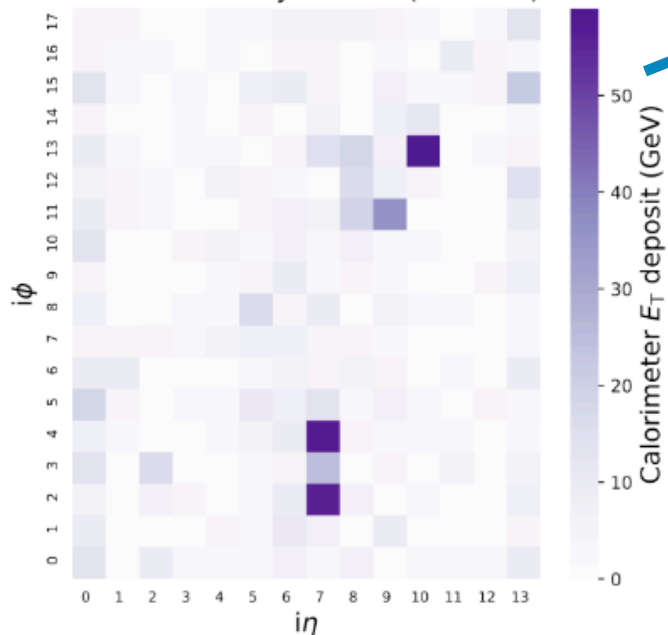
Objects: 10 jets, 4 muons, e/γ , MET

Features: p_T , η , ϕ (in raw integer values)

Architecture: MLP



CMS Preliminary 2023 (13.6 TeV)

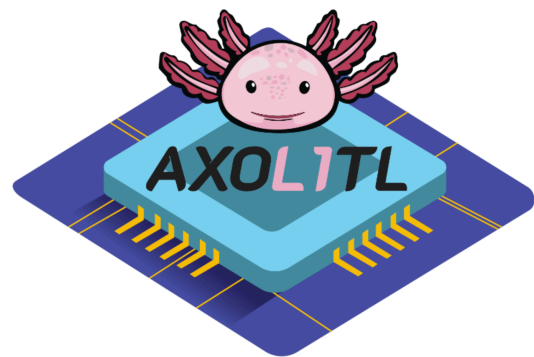


Low-level inputs: aggregated calorimeter towers
Architecture: 2D CNN w/ knowledge distillation

[\[CMS-DP-2023-086\]](#)

Ultra-fast anomaly detection @ CMS

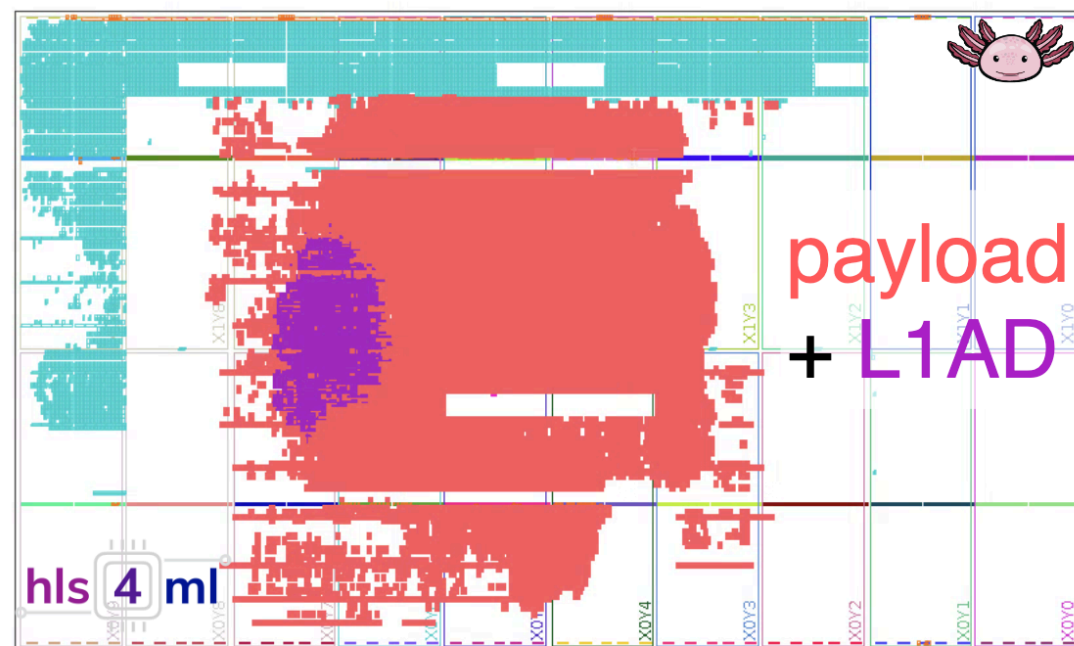
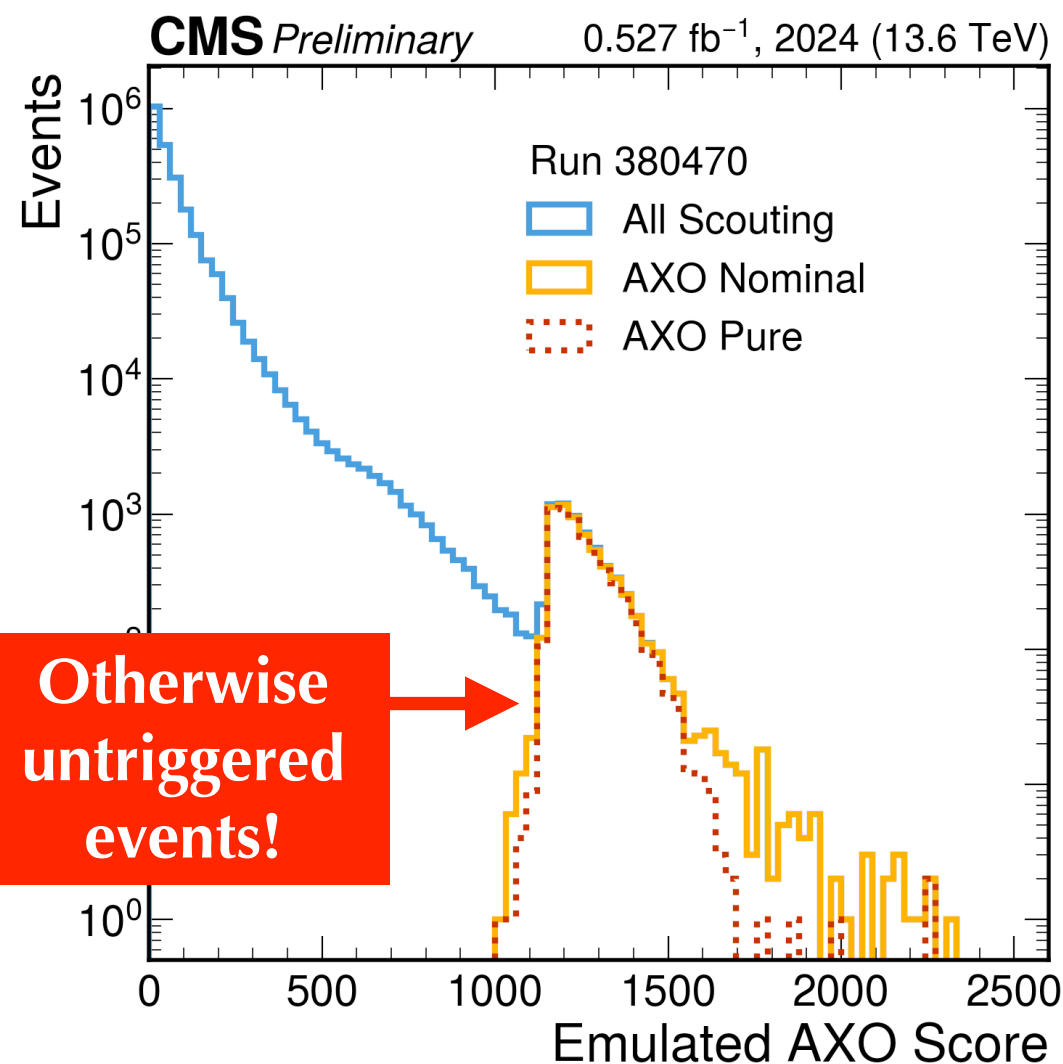
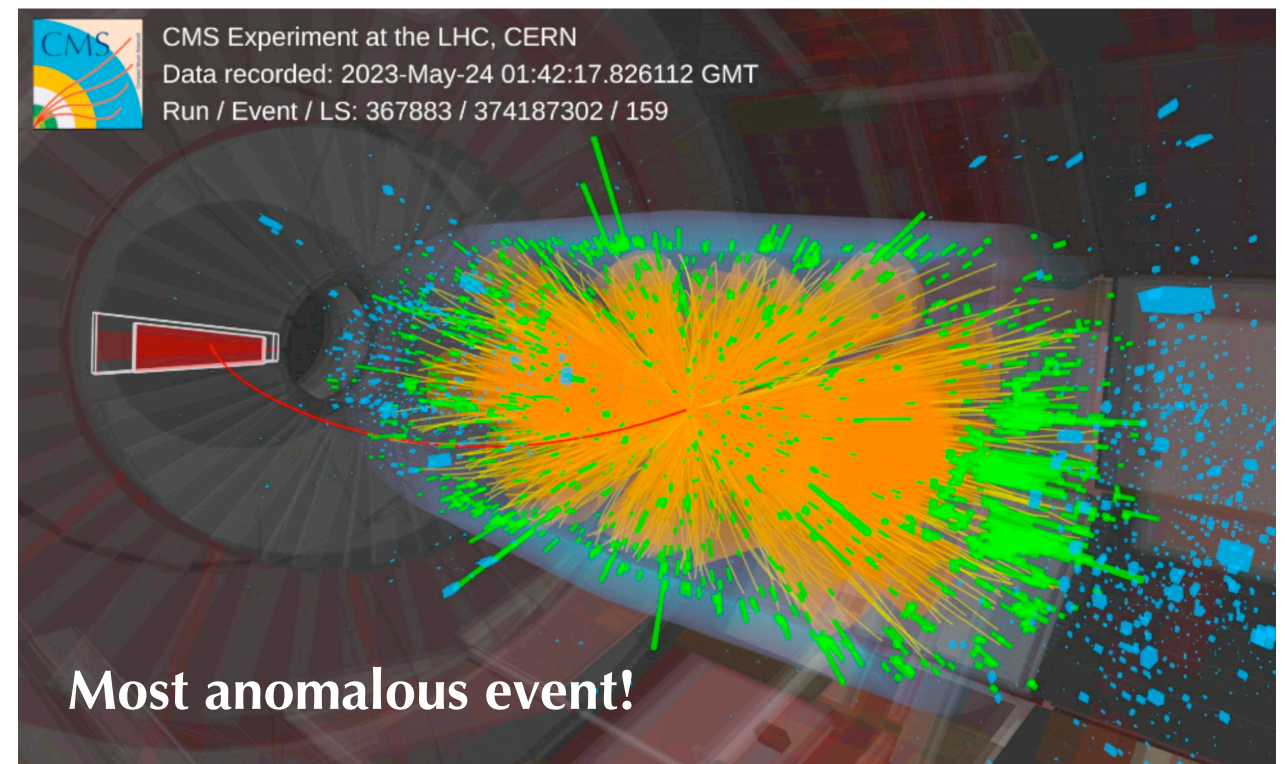
Anomaly eXtraction Online Level-1 Trigger aLgorithm



**Online since
Spring this year!**

**Full analysis and
interpretation of dataset
ongoing... stay tuned!**

[CMS-DP-2023-079](#)
[CMS-DP-2024-059](#)



hls 4 ml

	Latency	LUTs	FFs	DSPs	BRAMs
AXOL1TL	2 ticks 50 ns	2.1%	~0	0	0

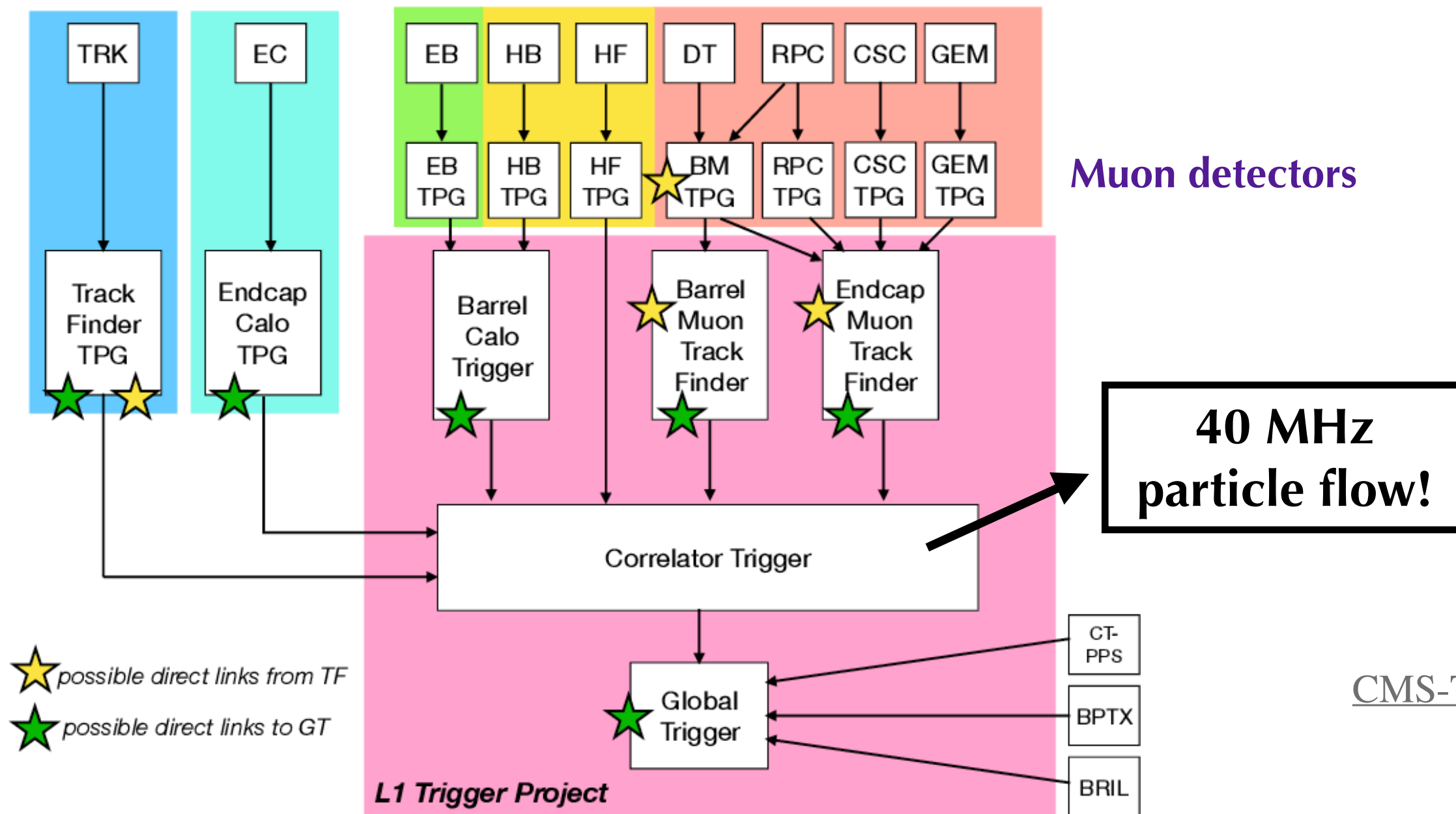
The HL-LHC challenge: CMS Phase 2

At HL-LHC, up to 200 pile-up interactions: CMS is upgrading the L1T and HLT to enable the same physics program we are doing now (at @60 PU)

40 MHz tracking!

Calorimeters

- * input data from 2 Tb/s to 63 Tb/s
- * latency of 12.5 μ s to take decision

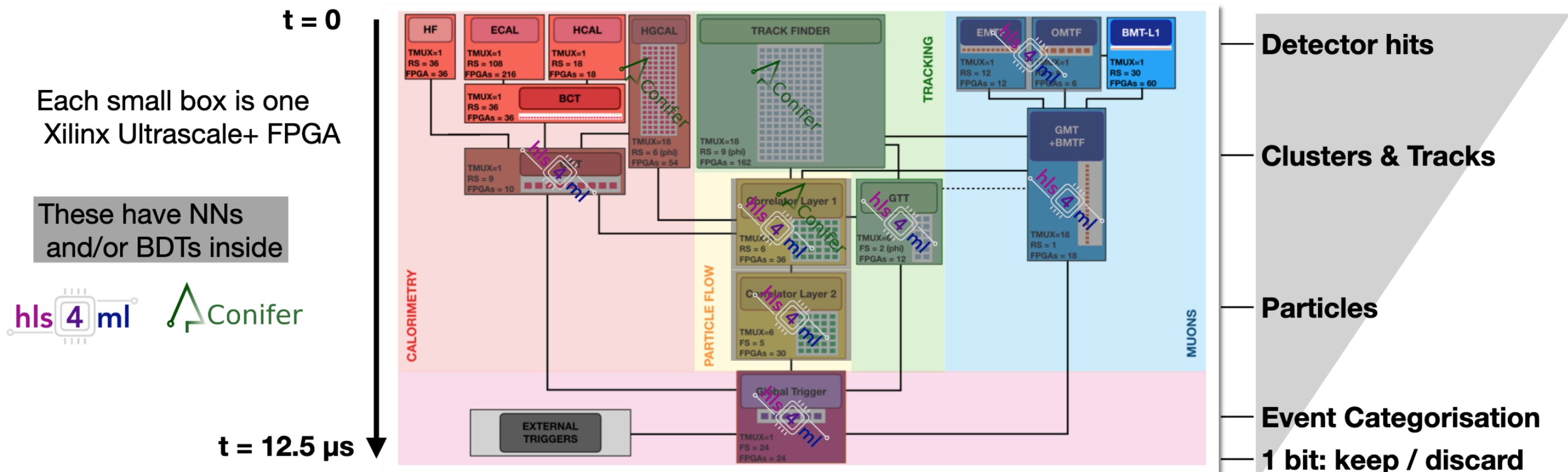


The HL-LHC challenge: CMS Phase 2

At HL-LHC, up to 200 pile-up interactions: CMS is upgrading the L1T and HLT to enable the same physics program we are doing now (at @60 PU)

With significantly more powerful compute we expect ML to be well embedded into L1T to exploit higher information granularity:

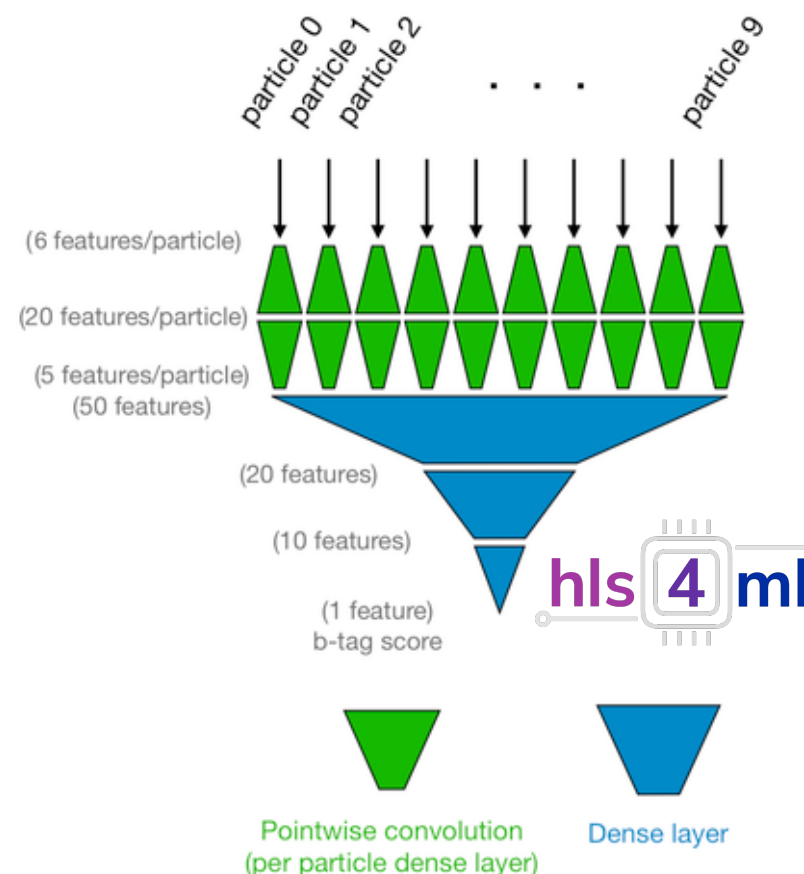
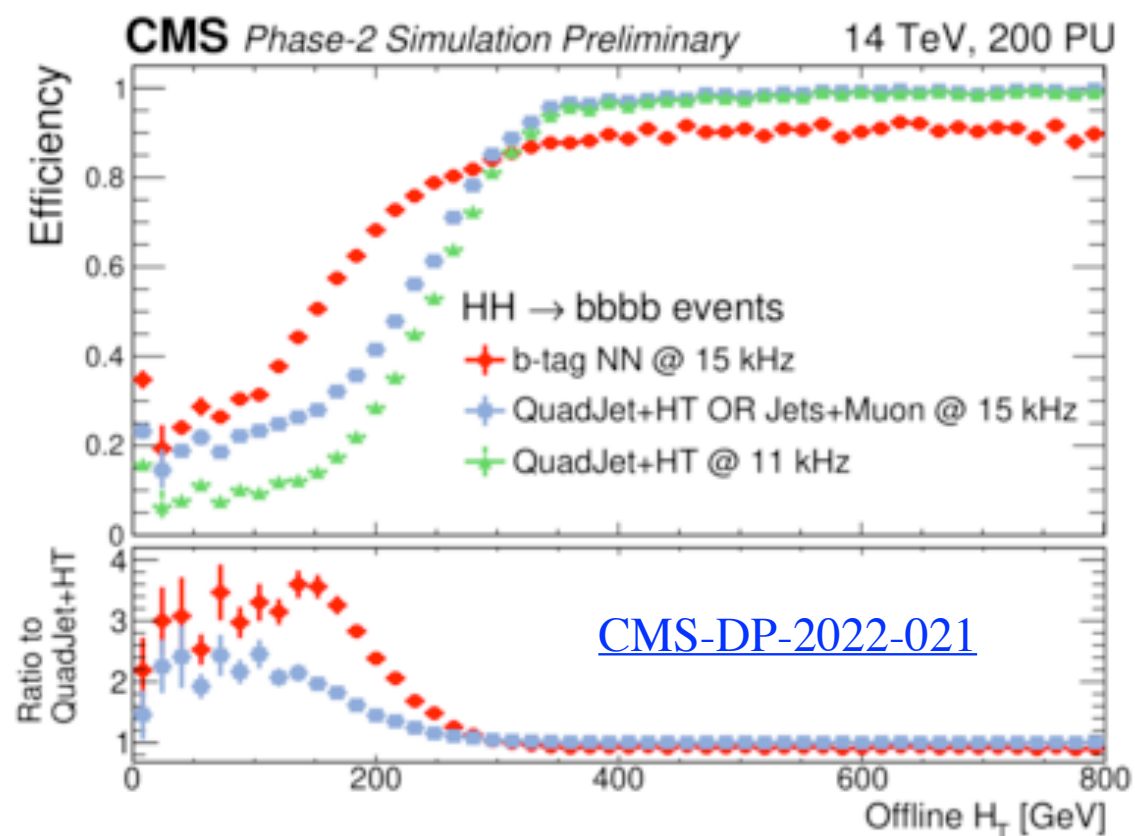
Around 20 projects (NNs, BDTs) in development accounting for 25 billion ML inferences per second



Example: b-tagging

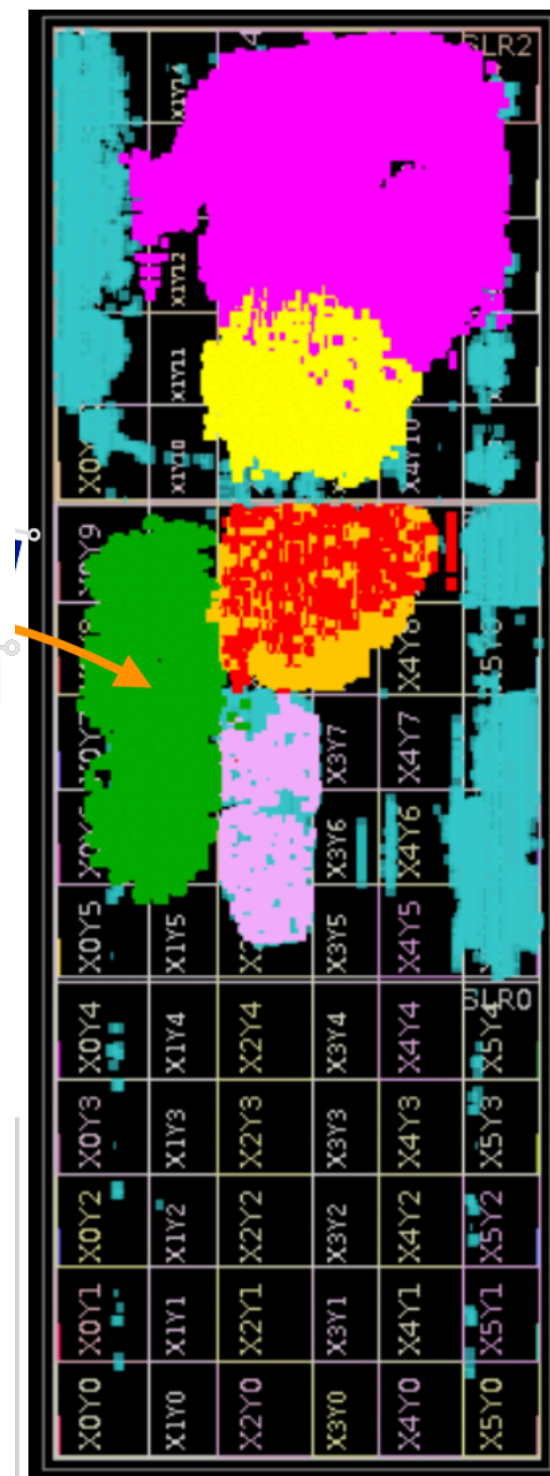
- Two 1D CNN layers acting on first 10 PUPPI candidates:
 - particle kinematics and type
 - vertex information

- Demonstrated **improved acceptance at low m_{HH} for the $HH \rightarrow 4b$ process** (compared to traditional cut-based trigger algorithms)



Particles Receiving
Jet Constituents Finding
Jet Axis computation
Sorting and Buffering
B-tagging NN

AMD VU9P



Finding the best NN architecture

- At offline level: chose the architecture with highest accuracy even if not efficient...
- For edge applications this is not an option: crucial to **co-design the architecture with the application and its constraints**

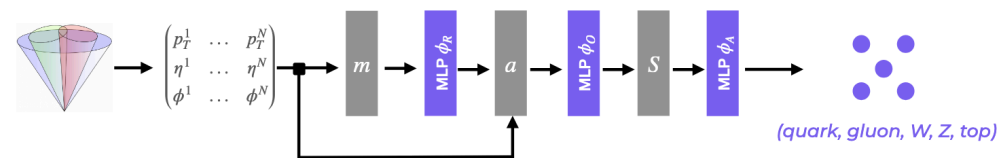
a) Multilayer Perceptron MLP



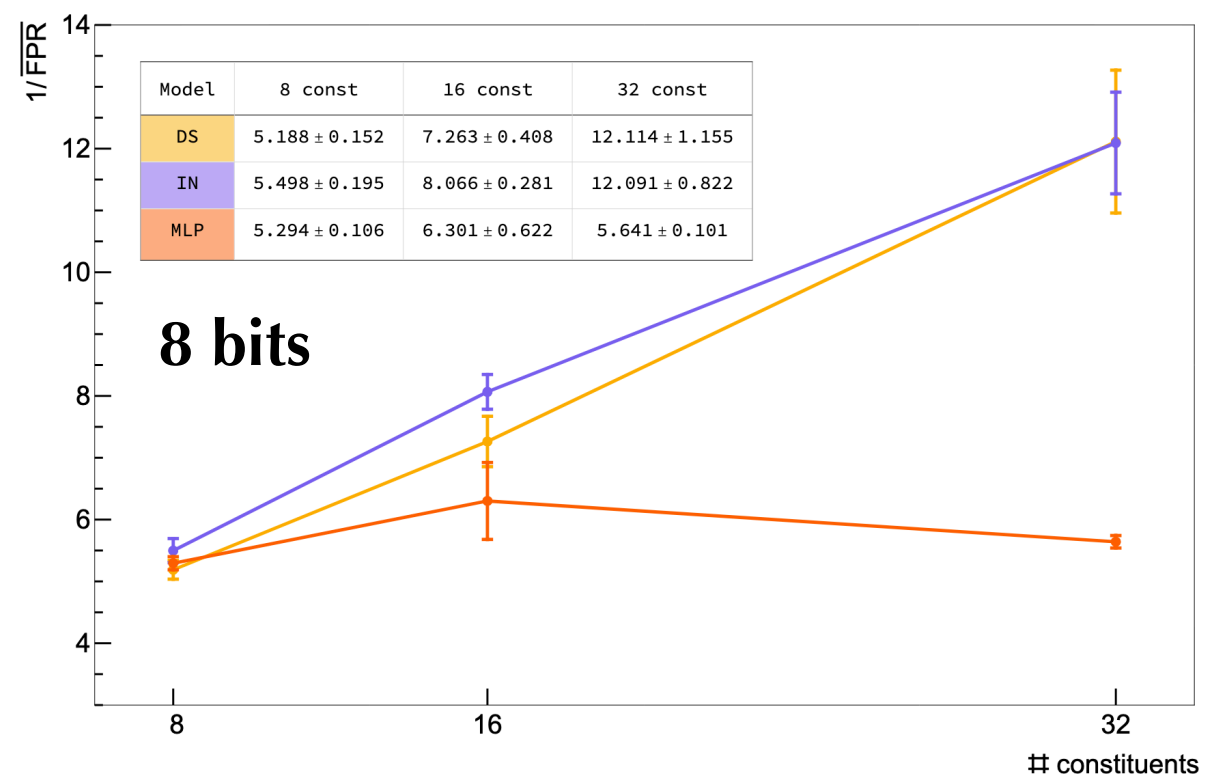
b) Deep Sets DS



c) Interaction Network IN



[arXiv.2402.01876](https://arxiv.org/abs/2402.01876)



**Graph NNs at
O(100) ns latency!**



FPGA: Xilinx Virtex UltraScale+ VU112D								
Architecture	Constituents	RF	Latency [ns] (cc)	II [ns] (cc)	DSP	LUT	FF	BRAM18
MLP	8	1	105 (21)	5 (1)	262 (2.1%)	155,080 (9.0%)	25,714 (0.7%)	4 (0.1%)
	16	1	100 (20)	5 (1)	226 (1.8%)	146,515 (8.5%)	31,426 (0.9%)	4 (0.1%)
	32 ^a	1	105 (21)	5 (1)	262 (2.1%)	155,080 (7.2%)	25,714 (0.7%)	4 (0.1%)
DS	8	2	95 (19)	15 (3)	626 (5.1%)	386,294 (22.3%)	121,424 (3.5%)	4 (0.1%)
	16	4	115 (23)	15 (3)	555 (4.5%)	747,374 (43.2%)	238,798 (6.9%)	4 (0.1%)
	32 ^a	8	130 (26)	10 (2)	434 (3.5%)	903,284 (52.3%)	358,754 (10.4%)	4 (0.1%)
IN	8	2	160 (32)	15 (3)	2,191 (17.8%)	472,140 (27.3%)	191,802 (5.5%)	12 (0.2%)
	16	4	180 (36)	15 (3)	5,362 (43.6%)	1,387,923 (80.3%)	594,039 (17.2%)	52 (1.9%)
	32 ^a	8	205 (41)	15 (3)	2,120 (17.3%)	1,162,104 (67.3%)	761,061 (22.0%)	132 (2.5%)

Finding the best NN architecture

- Many offline applications moving to SOA **transformer architectures** → not trivial mapping of MHA to FPGA circuit
 - attention map requires N^2 computations
 - softmax in those computation is slow and expensive
 - large weights matrices easily saturate memory
- First vanilla solutions for HEP being explored recently
- Expect more R&D in this direction in the near future
 - e.g., aggressive quantization and pruning of weights and/or attention scores, low rank matrices, ...

[Wayne Luk, et al.](#)

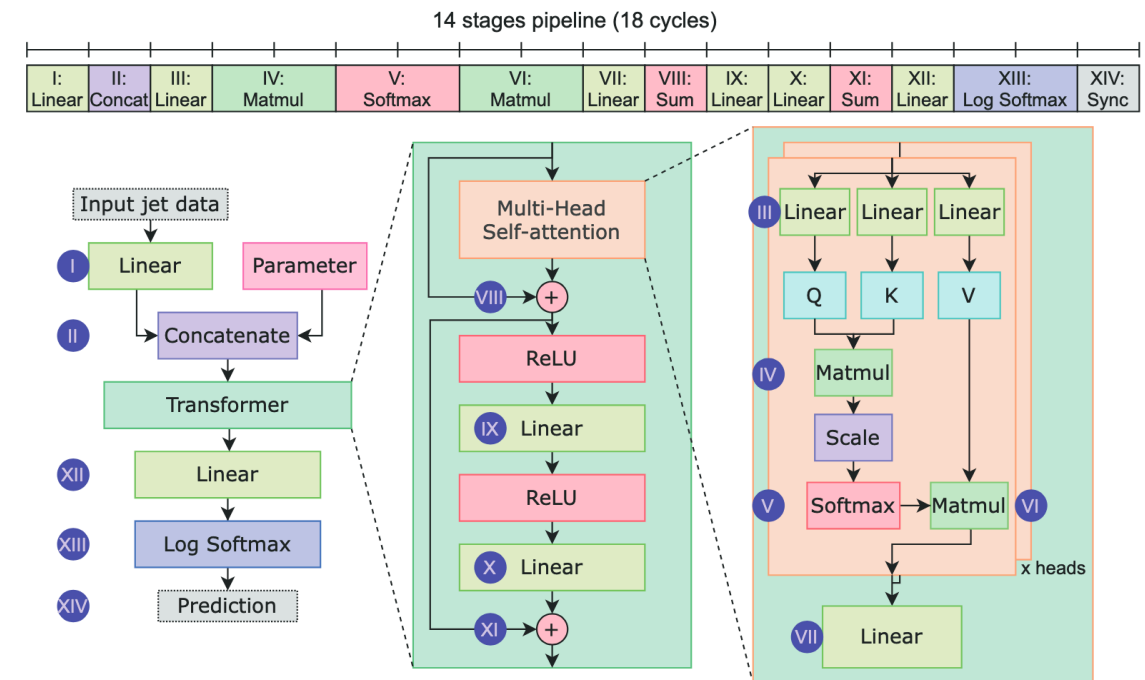


TABLE III
FPGA RESOURCES UTILIZATION

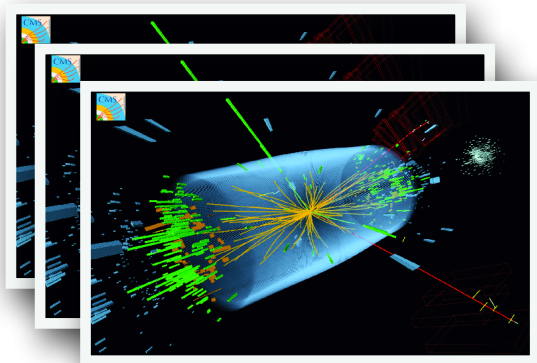
	BRAM 18K	DSP48E	FF	LUT
Total used	12	4,351	58,942	298,881
Available	5,376	12,288	3,456,000	1,728,000
Utilization	0.22%	35.41%	1.71%	17.30%

90 ns inference time

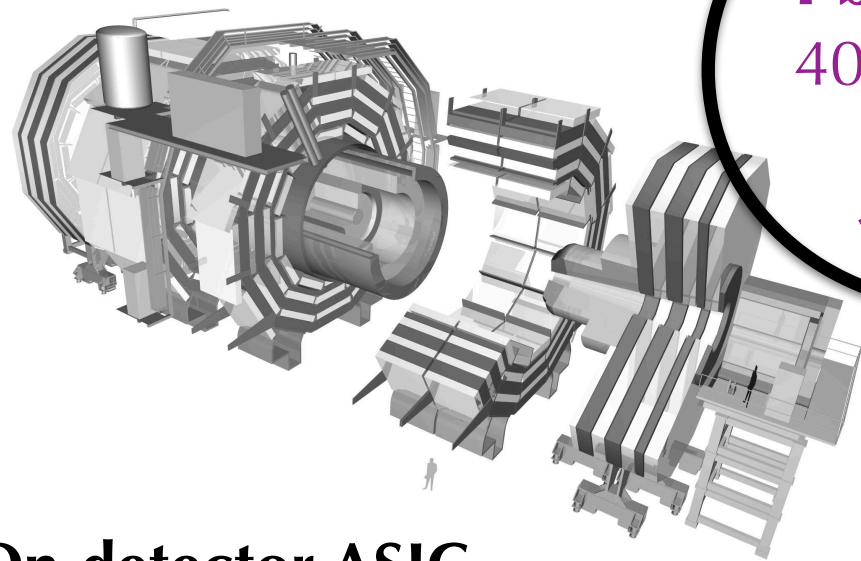
73% accuracy on jet tagging

16 jet level features

AI @ Extreme Edge



CMS Experiment
40 MHz collision rate
~1B detector channels



**On-detector ASIC
compression**
~100 ns latency

Pb/s
40 MHz

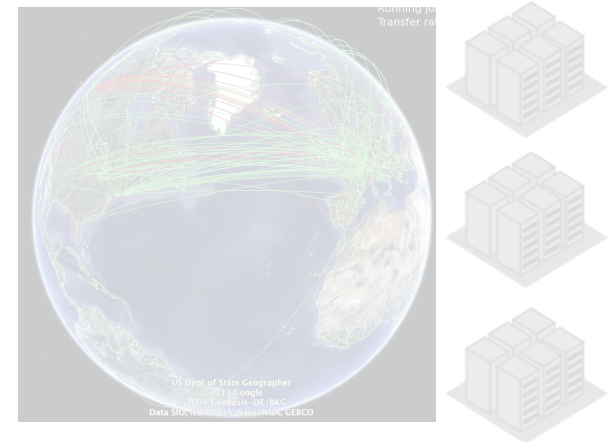
FPGA filter stack
~ μ s latency

**Level-1
Trigger**

ASICs typically used at the front end for sensors read out: directly embed ML in here to allow intelligent data compression at the very edge

100s KHz

Worldwide
computing grid
Exabyte-scale
datasets



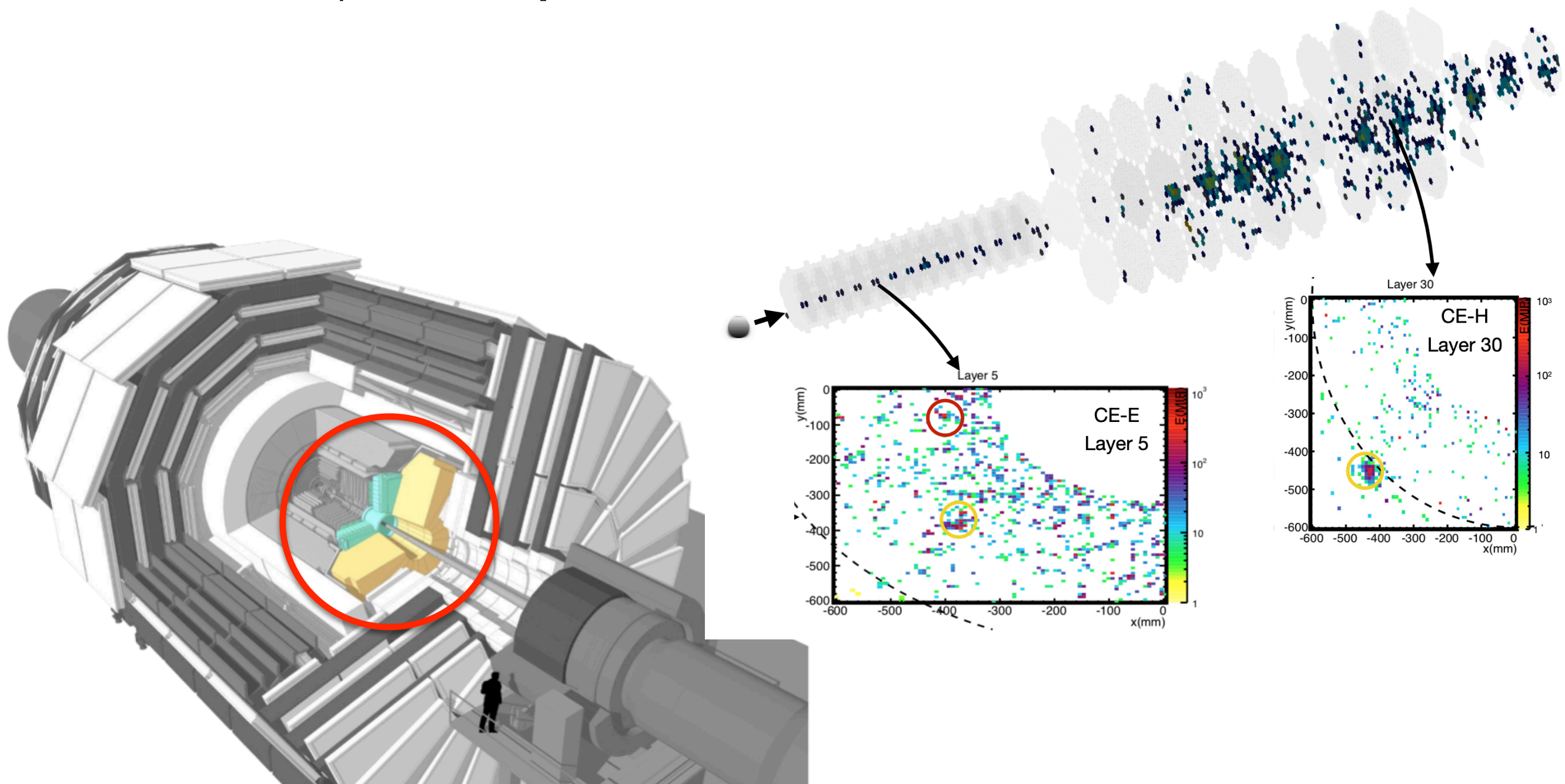
**High-Level
Trigger**

On-prem CPU/GPU filter farm
~100 ms latency

Example:

High-granularity calorimeter @ HL-LHC

Novel technology for future CMS endcap calorimeter:
50 layers with unprecedented number of readout channels (6M)!

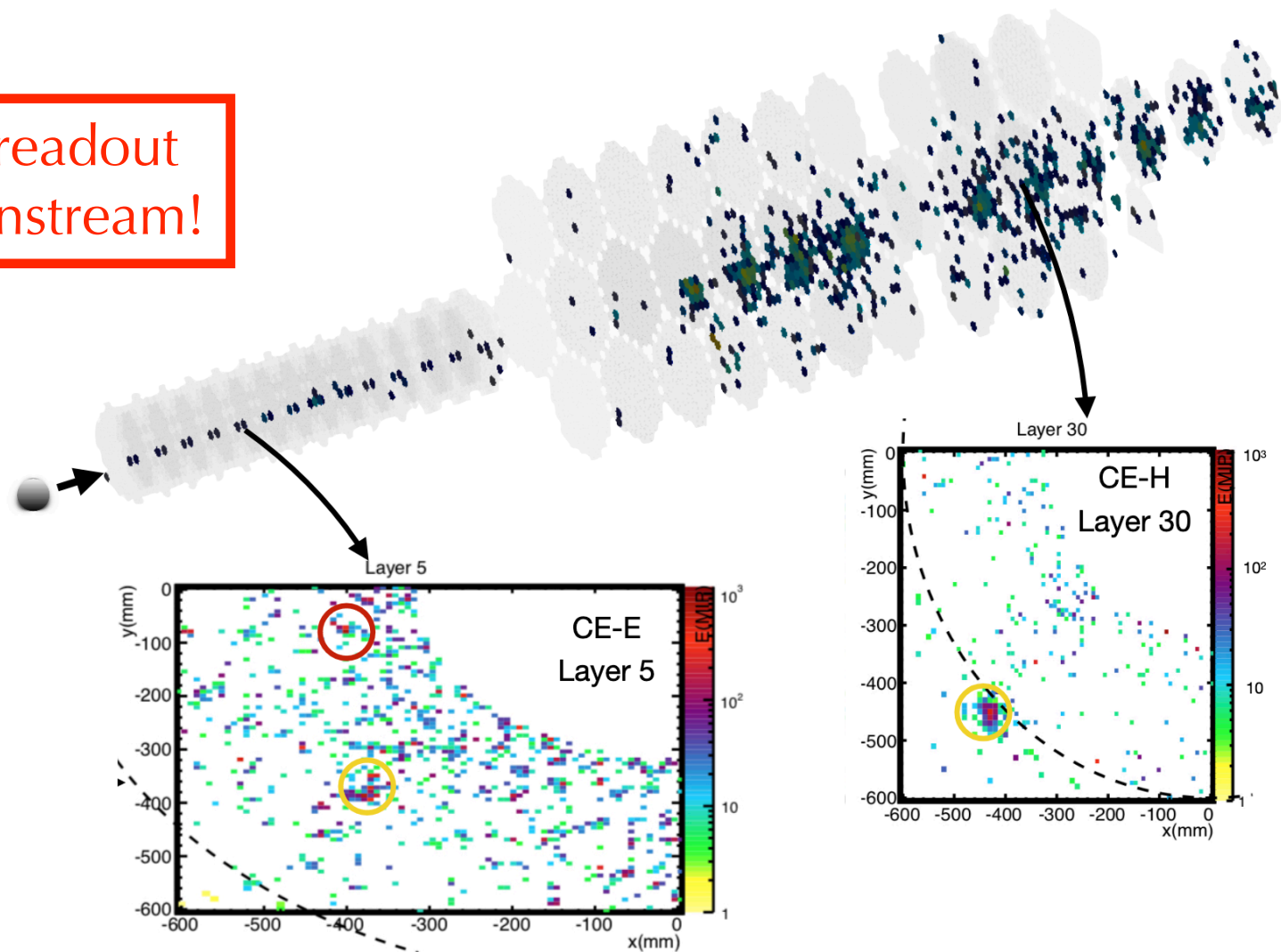
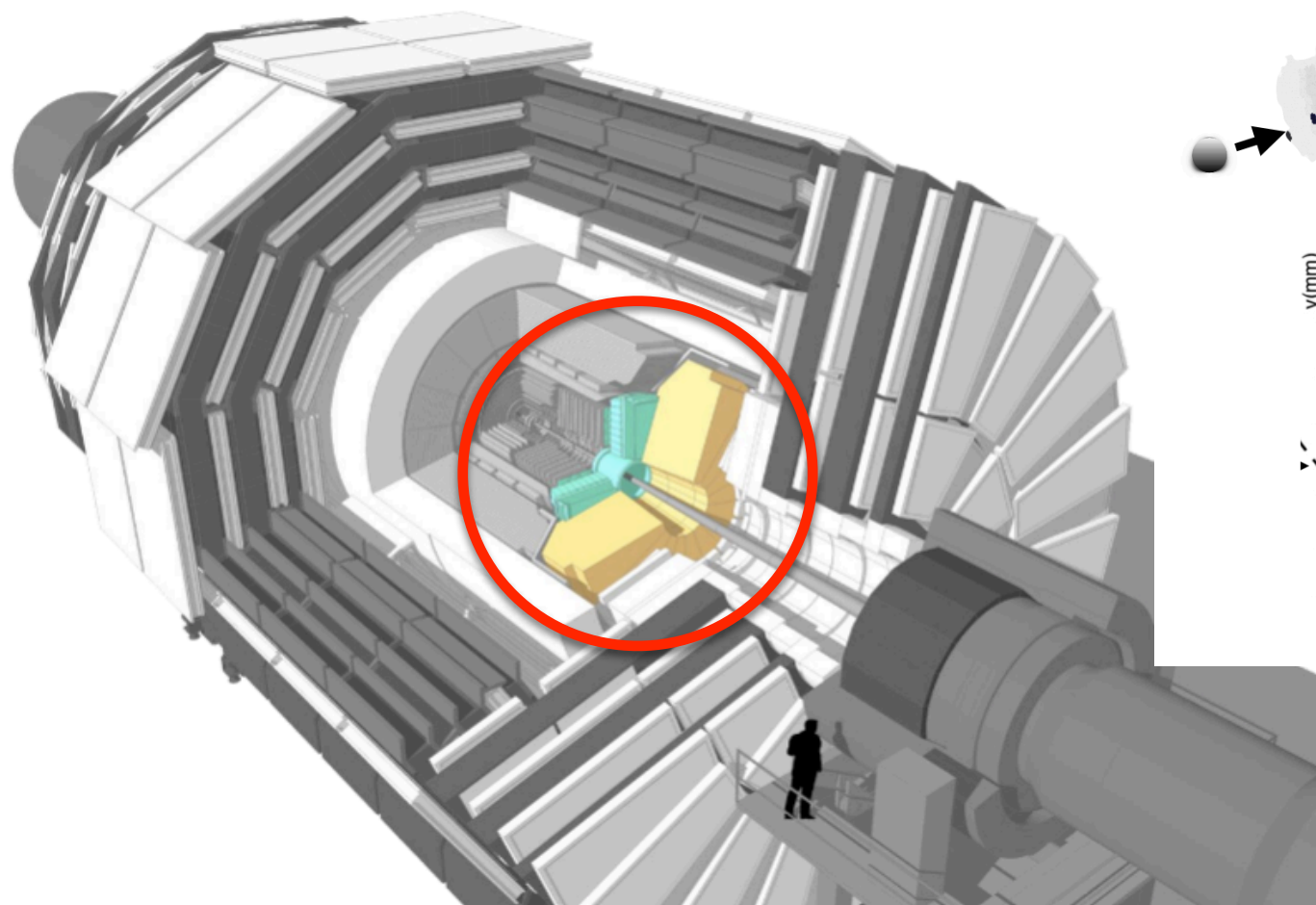


Example:

High-granularity calorimeter @ HL-LHC

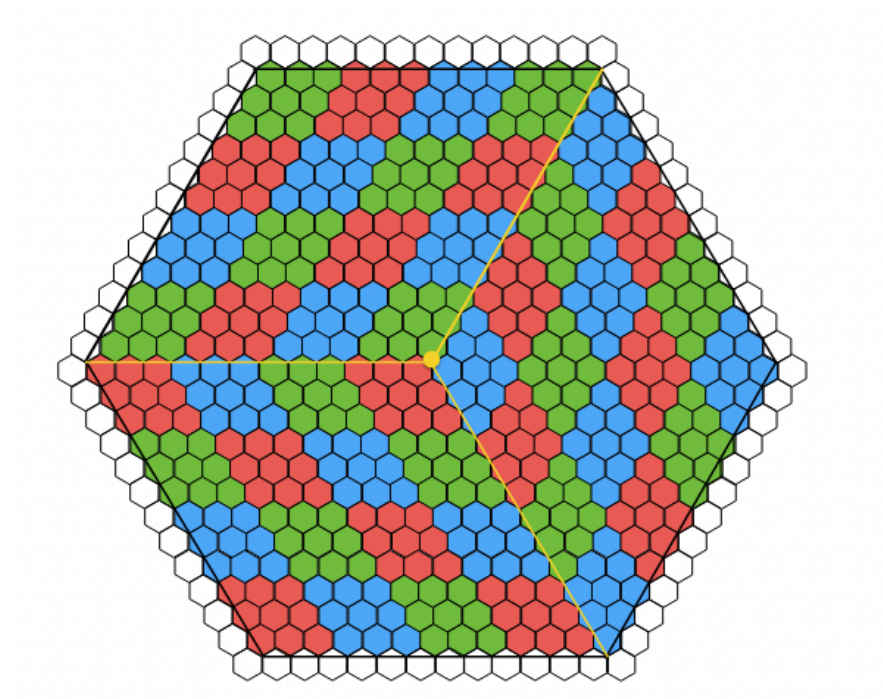
Novel technology for future CMS endcap calorimeter:
50 layers with unprecedented number of readout channels (6M)!

Not enough bandwidth and latency to readout
and put together all these channels downstream!



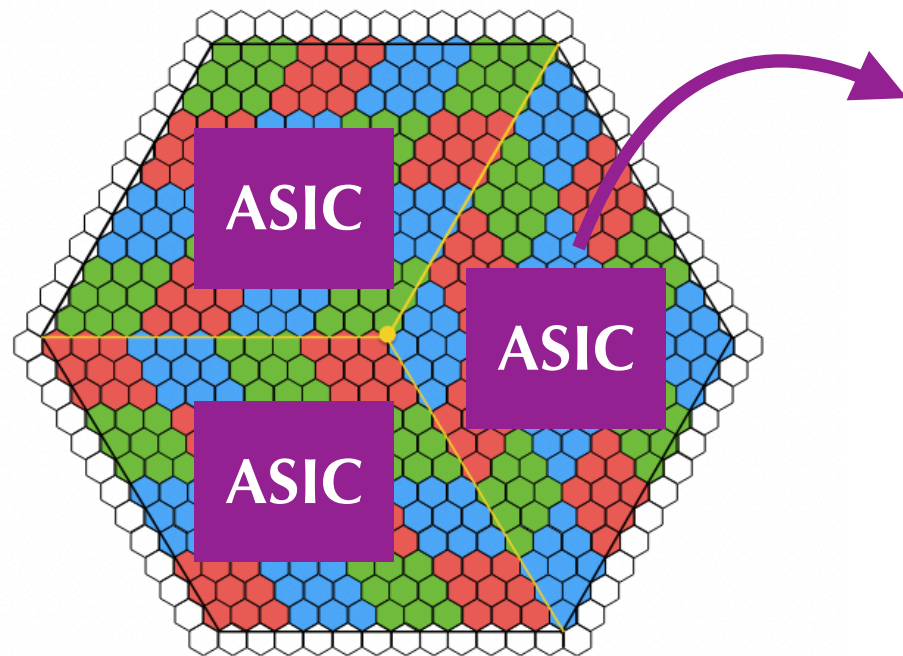
Example: CMS HG calorimeter

One module = 432 sensors

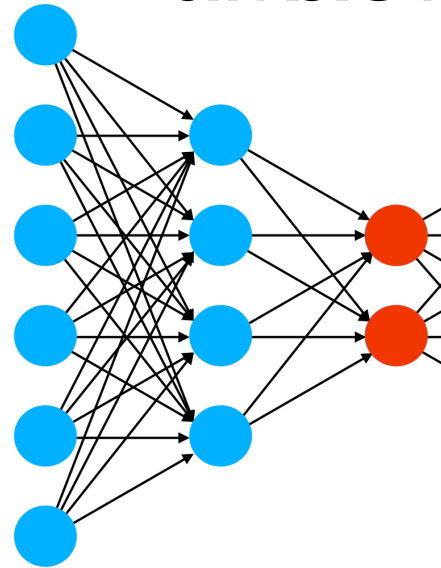


Example: CMS HG calorimeter

One module = 432 sensors



**Encode to N bits
on ASIC with NN**



Compress data on sensor in ASIC:

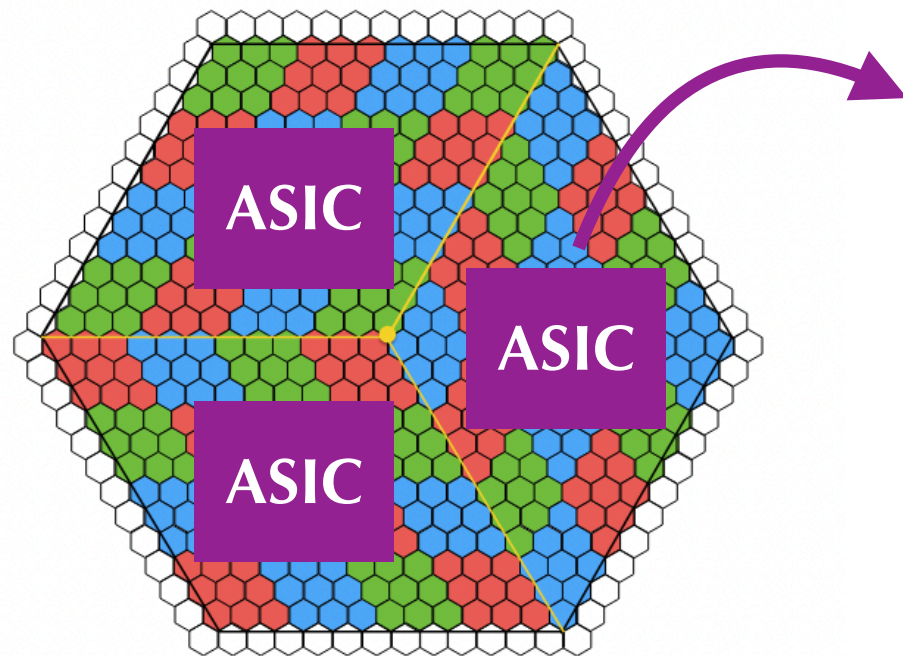
High radiation

Cooled to -30 C → low power

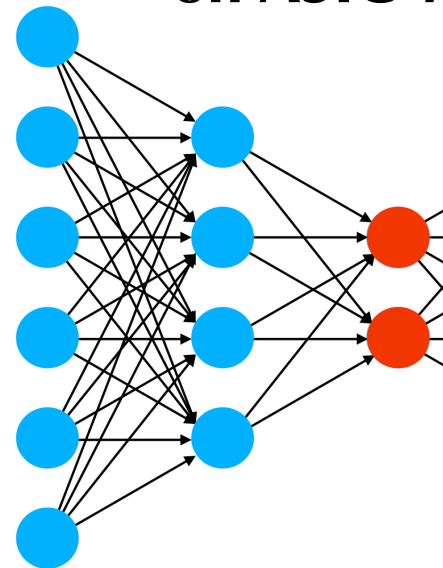
O(100) ns latency

Example: CMS HG calorimeter

One module = 432 sensors



**Encode to N bits
on ASIC with NN**

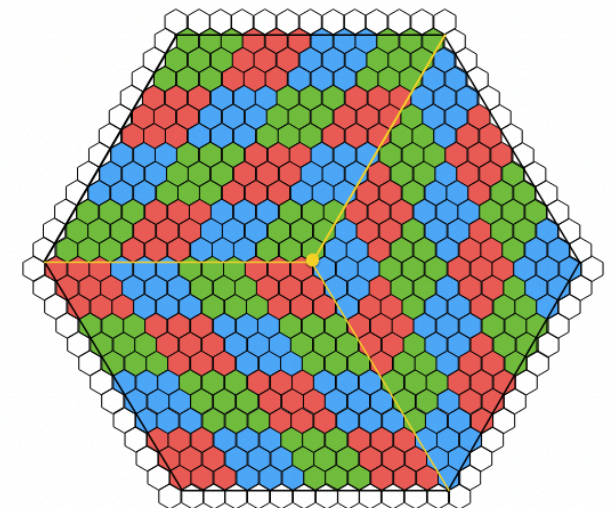
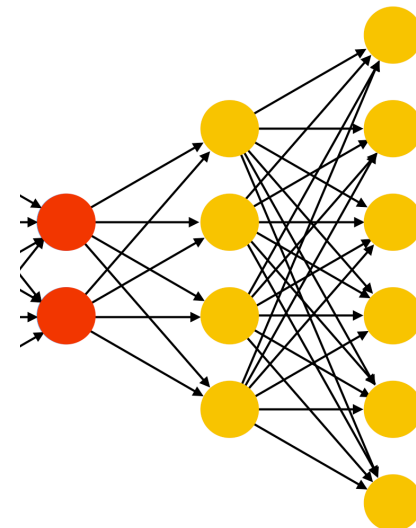


Transmit encoded data

**Reconstruct or do latent space
analysis on downstream
processors (FPGAs)**

Compress data on sensor in ASIC:

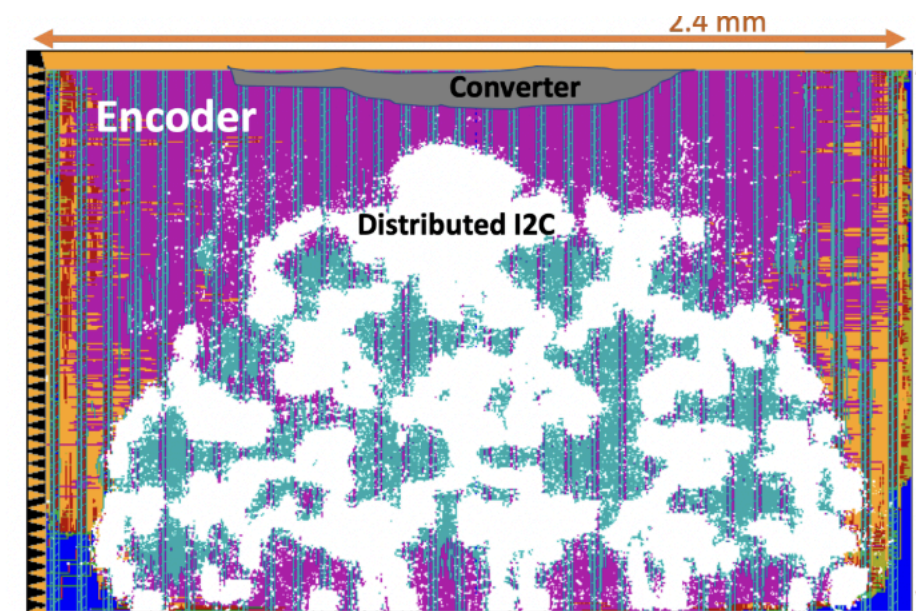
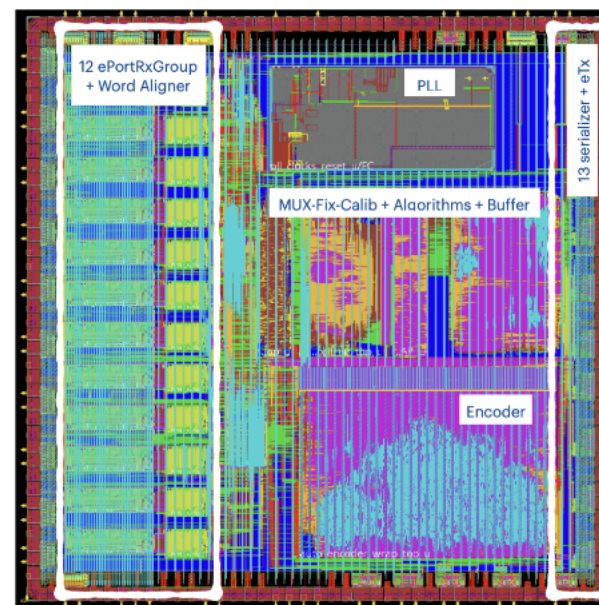
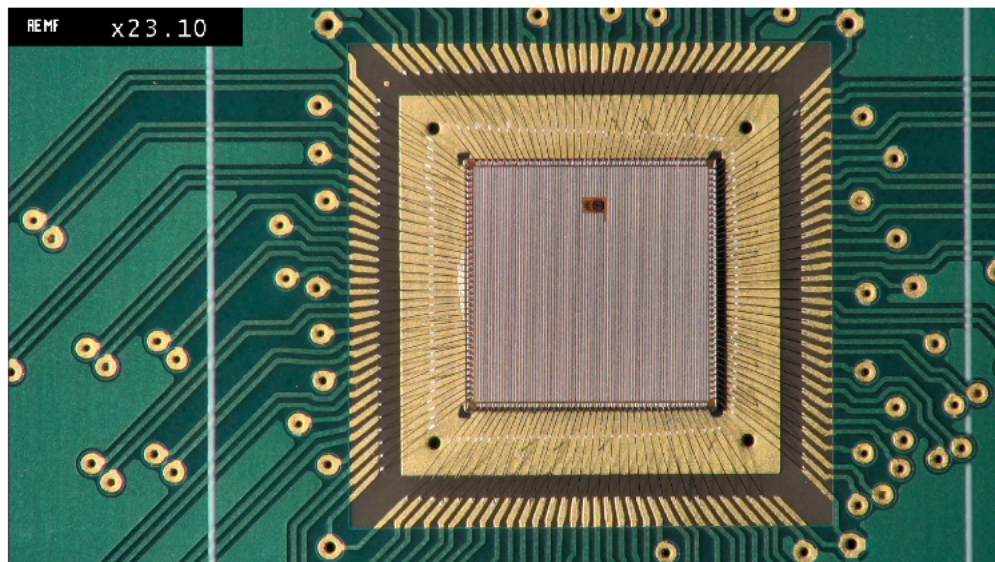
High radiation
Cooled to -30 C → low power
O(100) ns latency



AI @ Extreme Edge

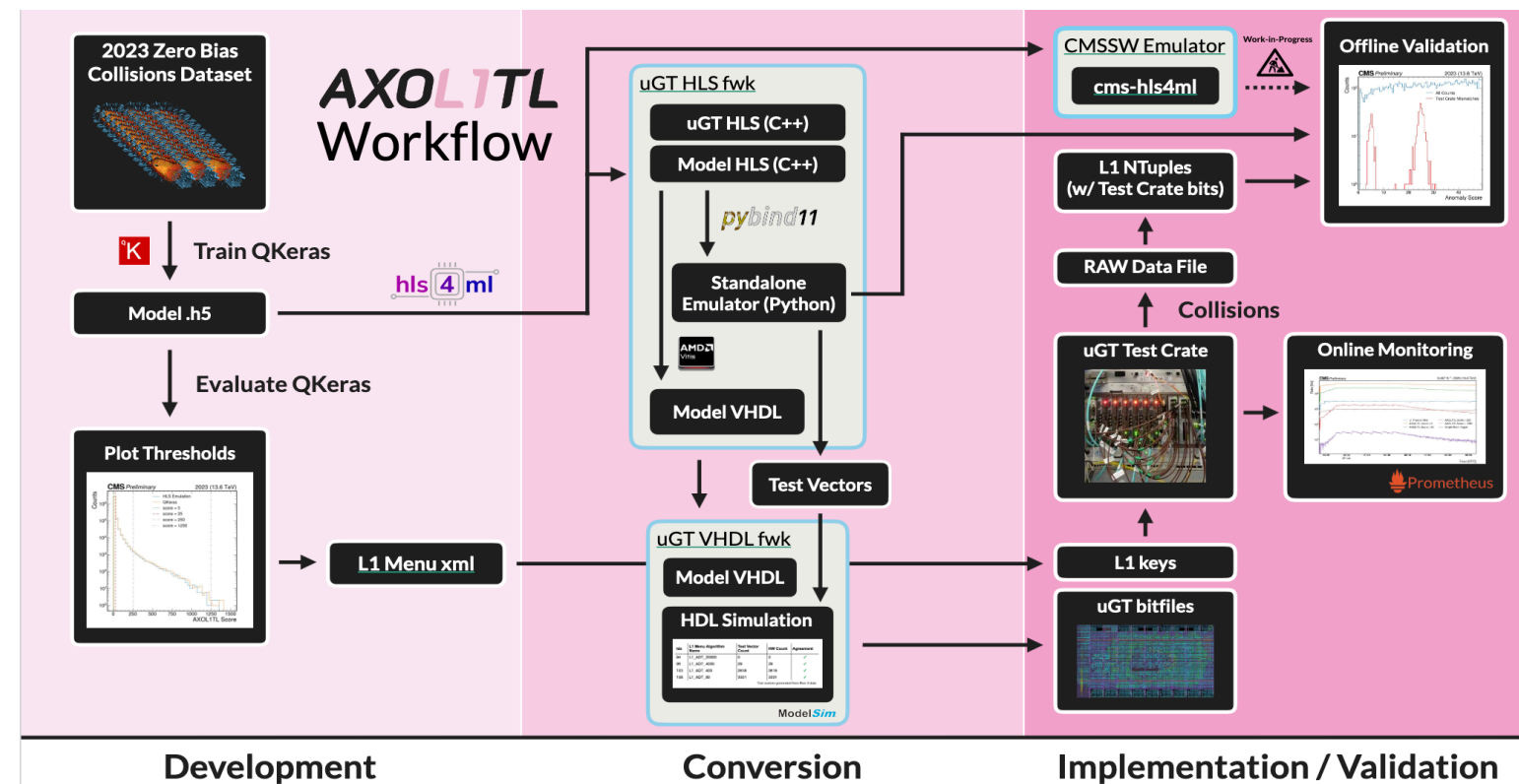


- **Tiny and heavily quantized NN**
- NN IP block created for the ASIC with **Catapult HLS (Mentor/Siemens)** and **hls4ml**
 - NN architecture is fixed, weights can be reprogrammed over I2C
 - NN parameters (weights and biases) triplicated for radiation tolerance → 200% overhead
- Developed in parallel a tool — [FKeras](#) — that performs **bit-level sensitivity study of each weight in the NN**
 - allows to prioritize which bits need protection and which may be safely disregarded, reducing resource overhead

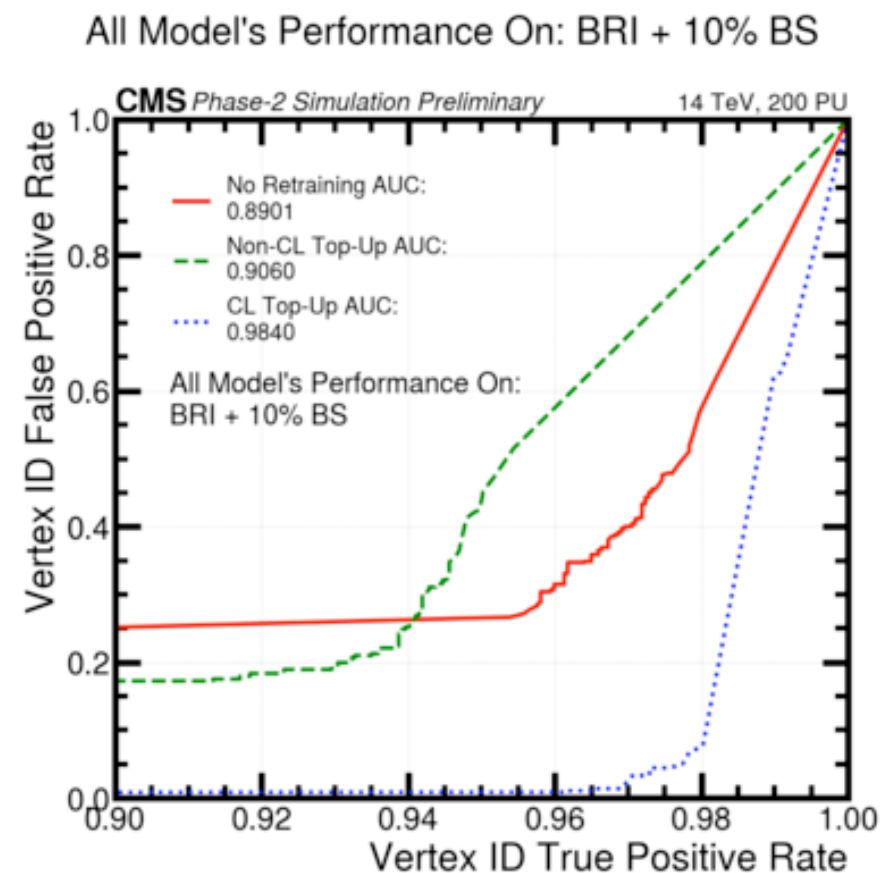


Future challenges

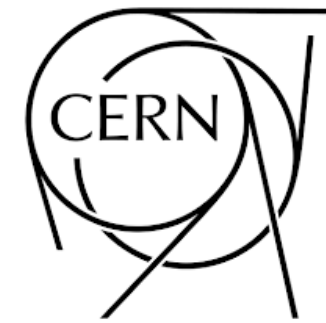
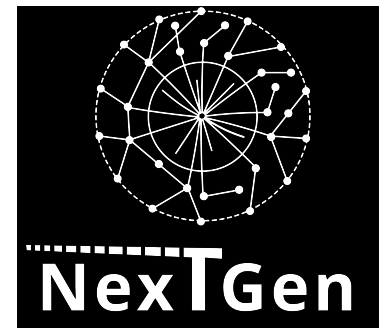
- Put everything together (fast multi-models simultaneous training)
- Robustness vs conditions (e.g., continual learning)
- Automatization (MLOps and system-control strategies)
- Hyperpars optimization beyond offline applications (e.g., through surrogate models for neural architecture search)
- Explore transformers and expensive self-supervised learning strategies on new hardware (e.g. AI engines or distributed computing)



Example: continual learning for
CMS Phase 2 tracker degradation
[\[CMS-DP-2023-022\]](#)



The Next Generation Triggers

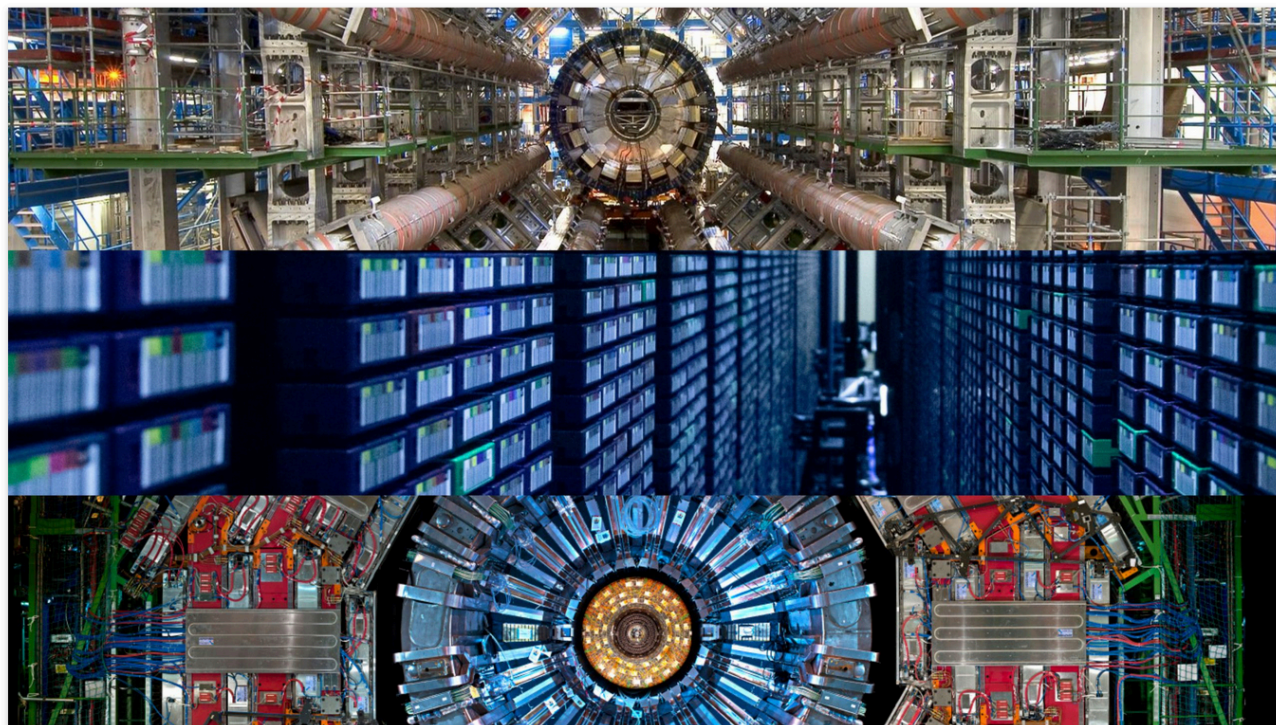


[CERN news](#)

The next-generation triggers for CERN detectors

The recently launched Next-Generation Triggers project is set to remarkably increase the efficiency, sensitivity and modelling of CERN experiments

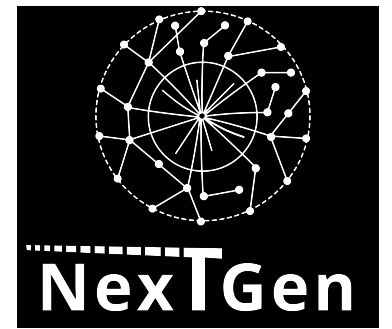
11 APRIL, 2024 | By [Antonella Del Rosso](#)



From top to bottom: ATLAS, CERN Data Centre and CMS (Image: CERN)

- Eric & Wendy Schmidt foundation fund a CERN project that will *enhance the physics reach of the ATLAS and CMS experiments at HL-LHC and beyond* using novel technologies:
 - neural network optimization
 - quantum-inspired algorithms
 - high-performance computing and FPGAs
 - theoretical modelling

The Next Generation Triggers

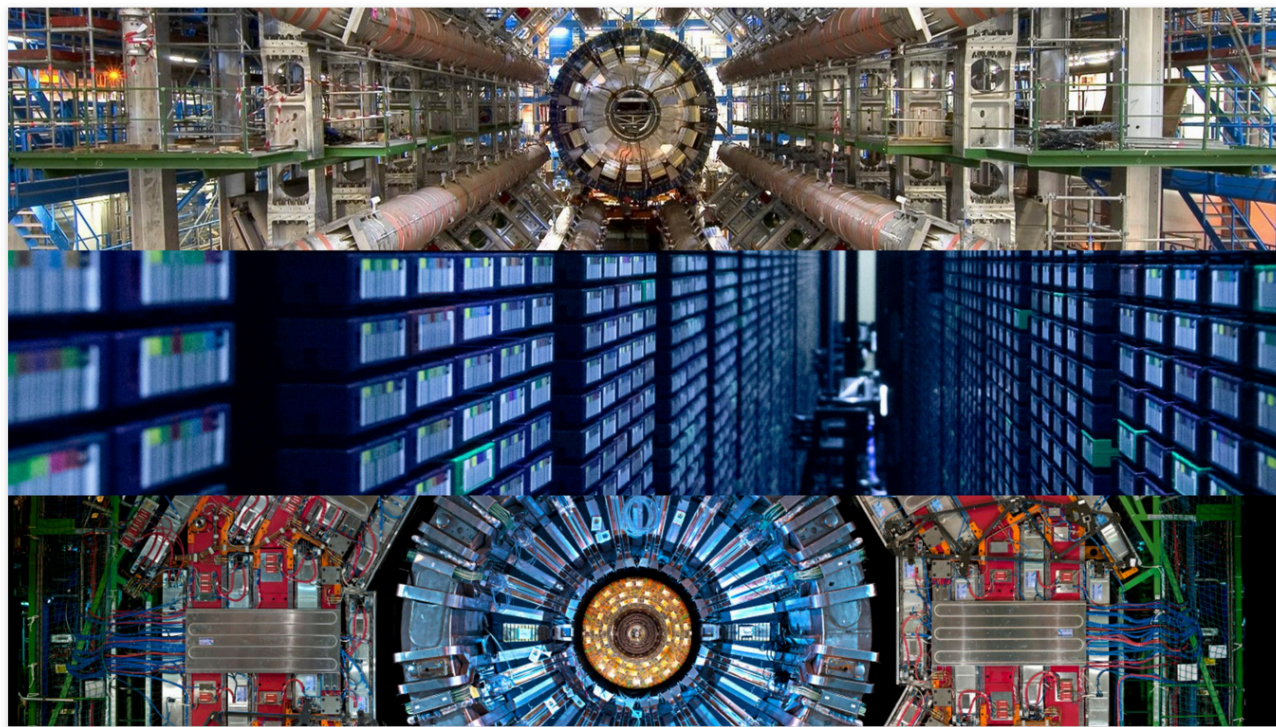


[CERN news](#)

The next-generation triggers for CERN detectors

The recently launched Next-Generation Triggers project is set to remarkably increase the efficiency, sensitivity and modelling of CERN experiments

11 APRIL, 2024 | By [Antonella Del Rosso](#)



From top to bottom: ATLAS, CERN Data Centre and CMS (Image: CERN)

- Eric & Wendy Schmidt foundation fund a CERN project that will *enhance the physics reach of the ATLAS and CMS experiments at HL-LHC and beyond* using novel technologies:
 - neural network optimization
 - quantum-inspired algorithms
 - high-performance computing and FPGAs
 - theoretical modelling

The [NextGen Triggers project](#) will mark a new chapter in high-energy physics, leveraging upgraded event-selection systems and data-processing techniques to unlock a realm of discoveries.

Summary

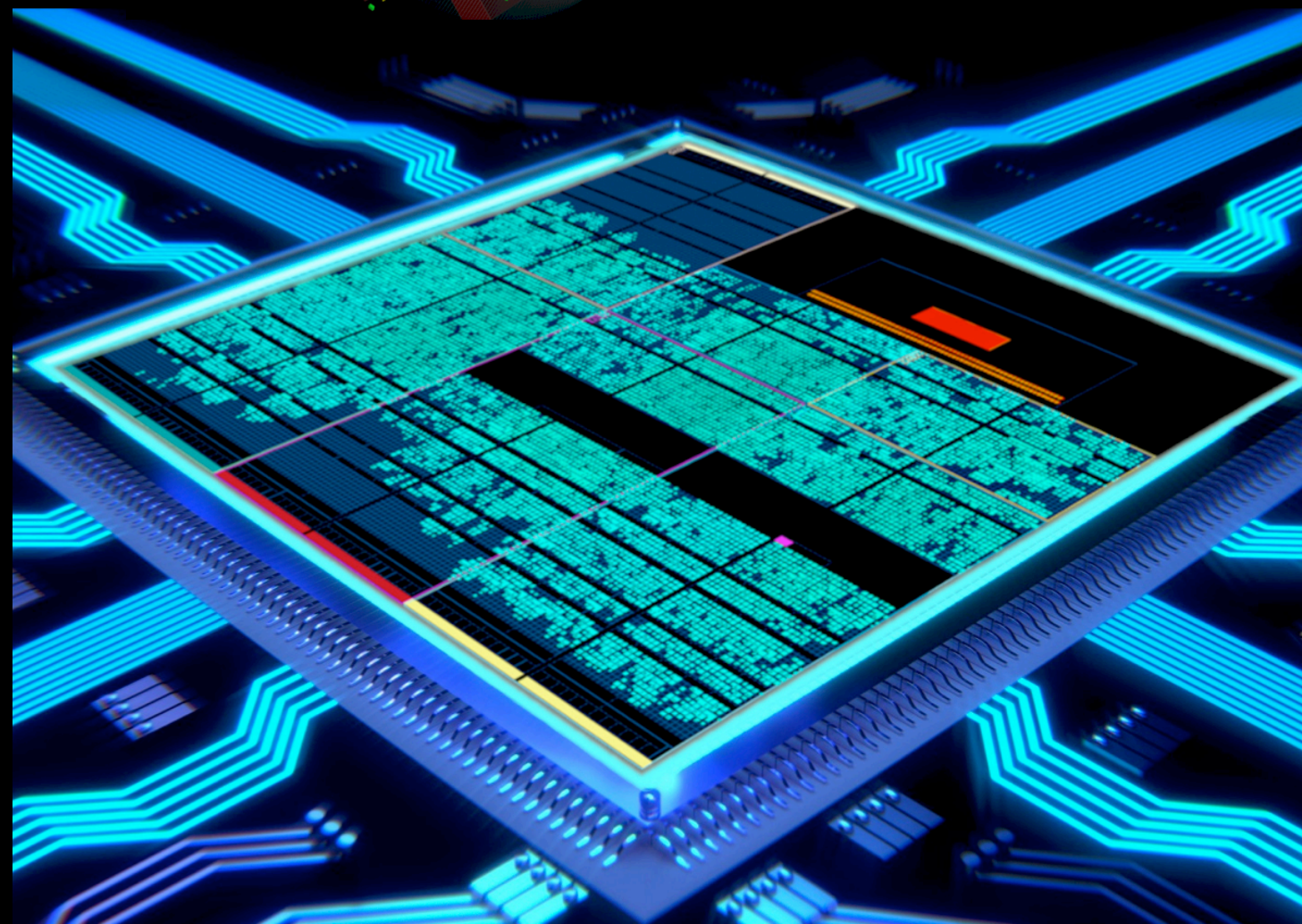
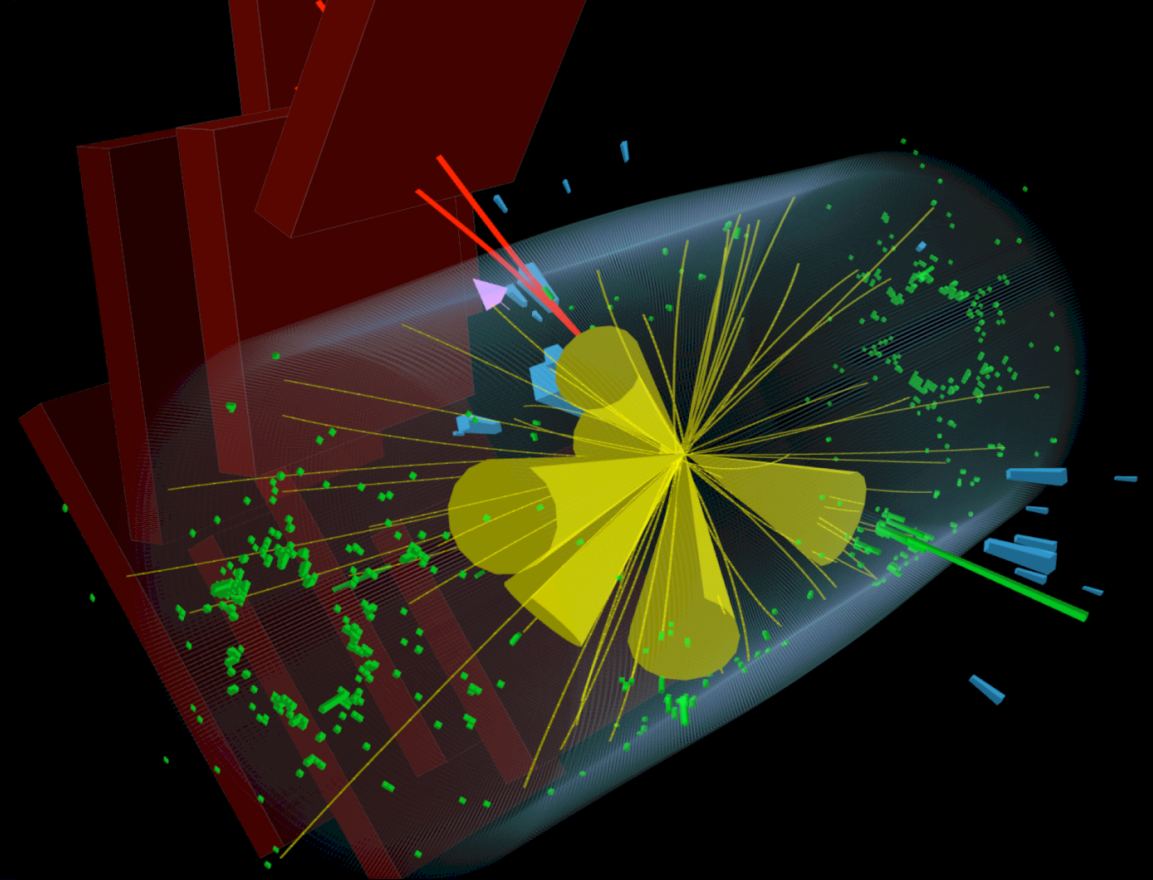
- **We hope to understand the fundamental structure of nature**
 - we expect new phenomena to answer those questions
 - but these are rare so we build large scale experimental setups
- **The challenge ahead is big**
 - more data, more complex data, not enough resources
- **This is why we need to push ML to the edge**
 - to do more with less (faster & better)
- **And hopefully discover new phenomena!**



CMS Experiment at the LHC, CERN

Data recorded: 2016-Oct-11 10:44:24.059904 GMT

Run / Event / LS: 282842 / 47118579 / 25



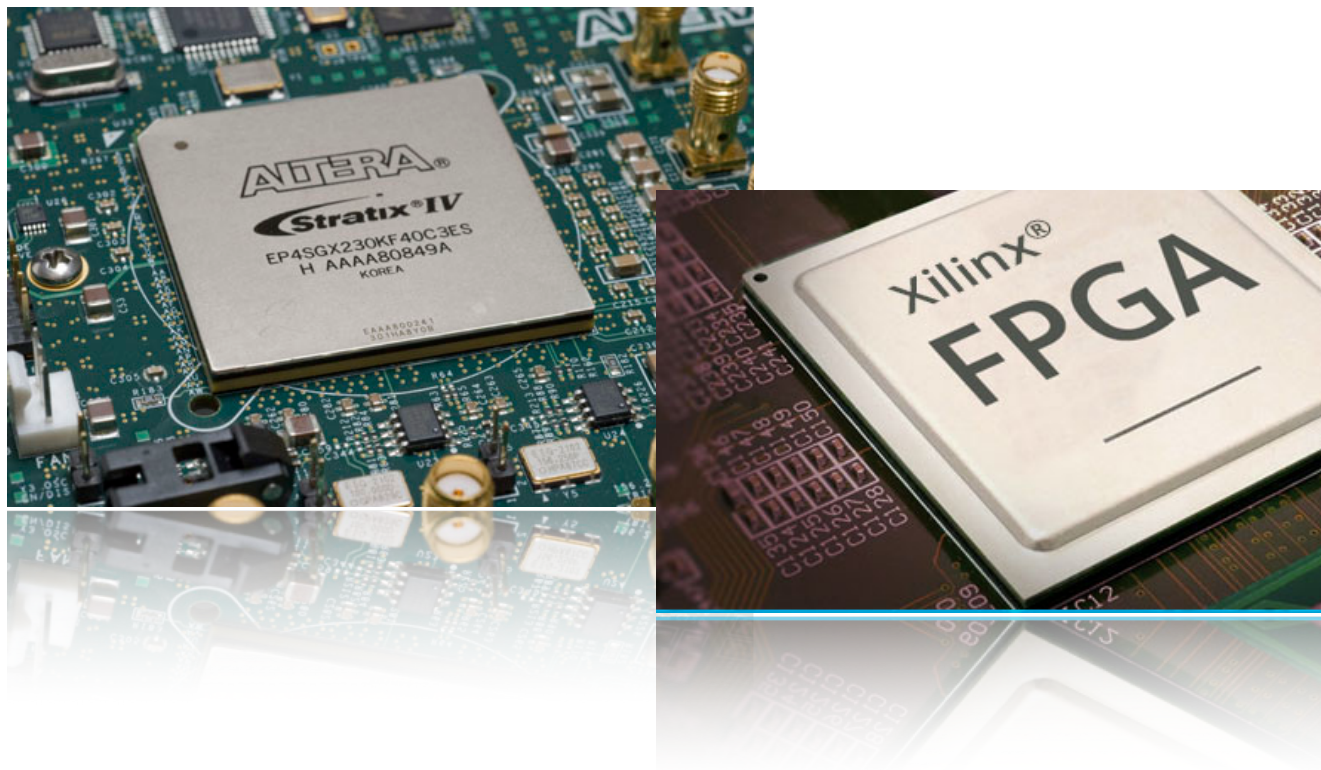
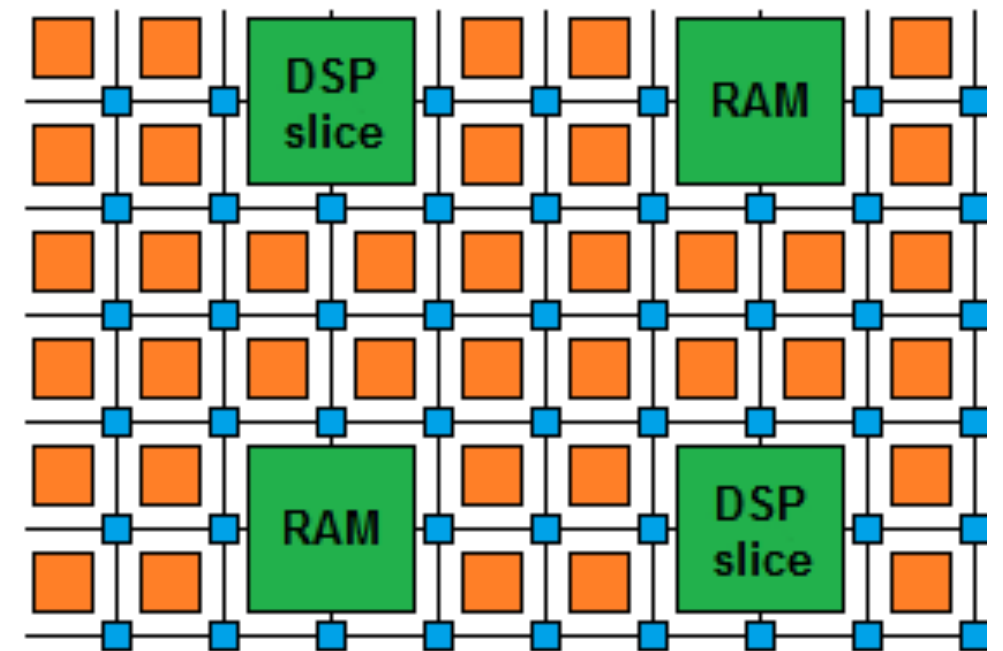
Backup

What are FPGAs?

Field Programmable Gate Arrays
are reprogrammable integrated circuits

Contain many different building blocks
(‘resources’) which are connected together as you
desire

FPGA diagram



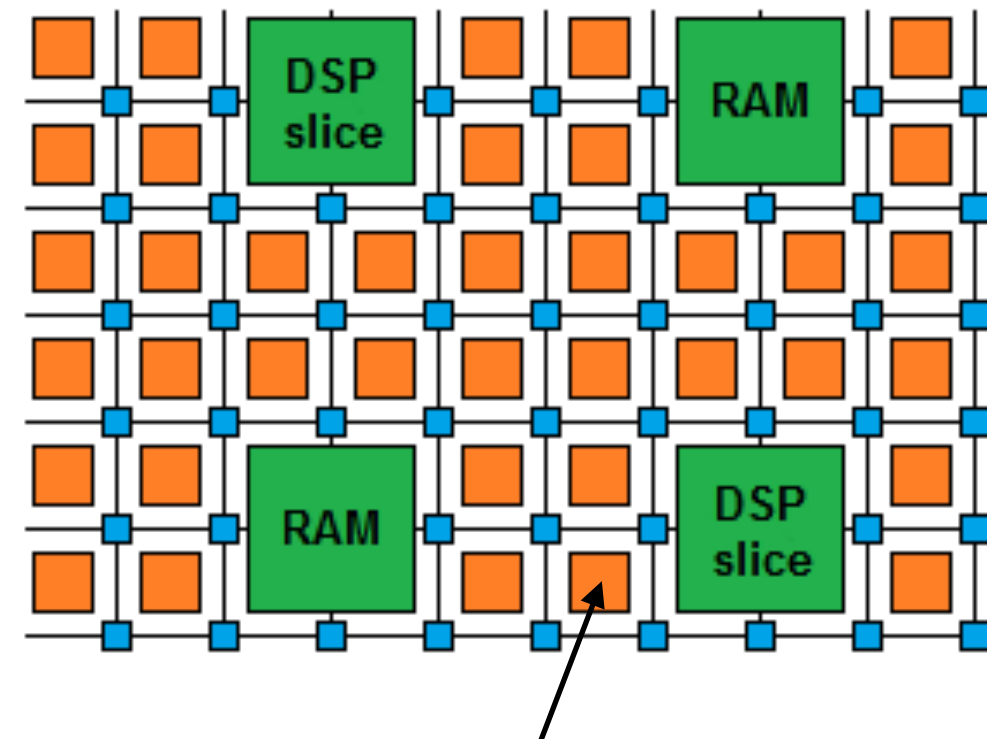
What are FPGAs?

Field Programmable Gate Arrays
are reprogrammable integrated circuits

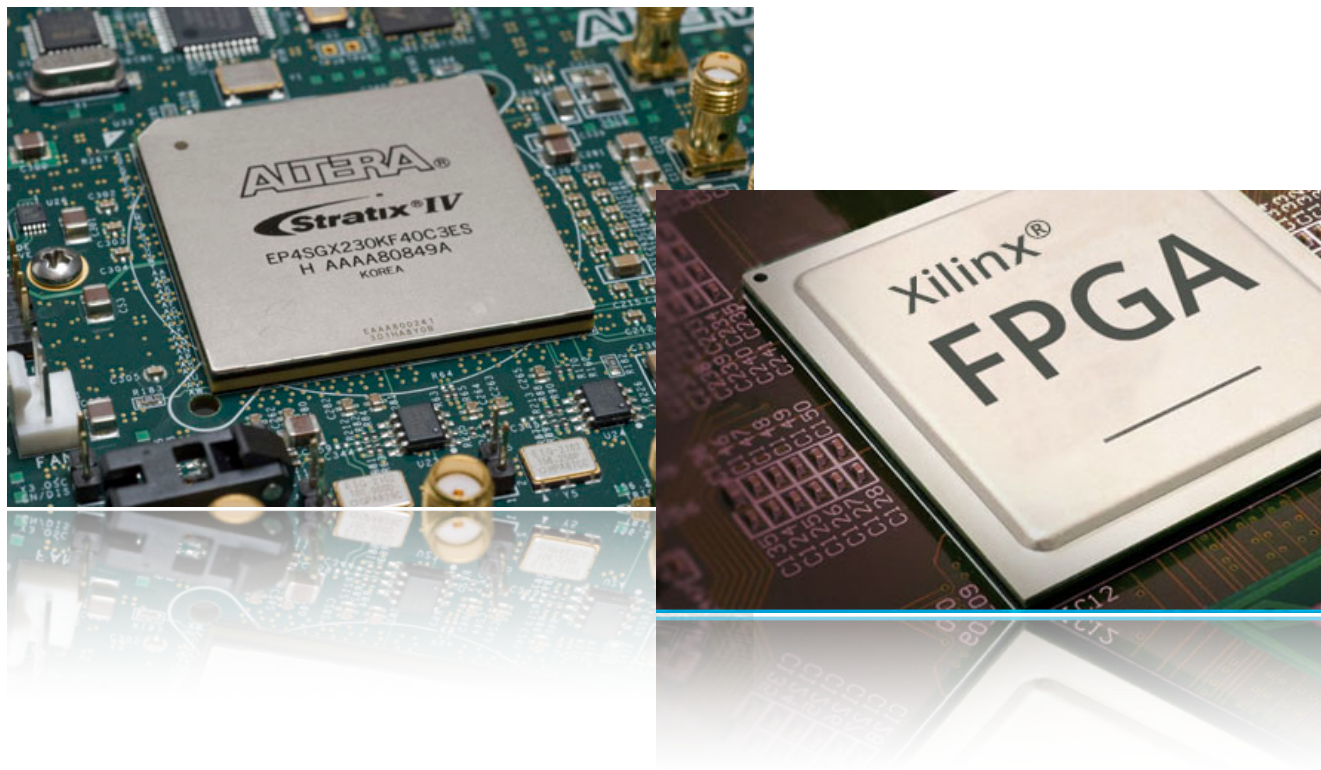
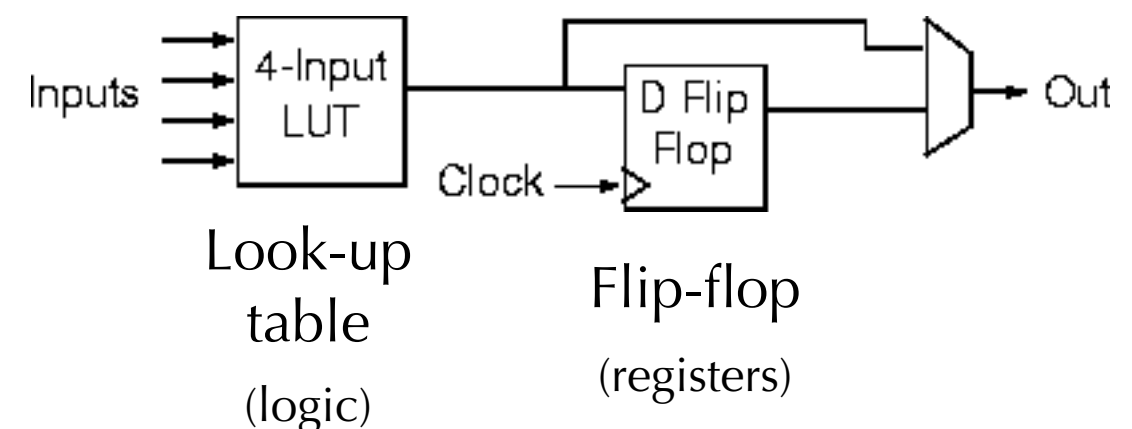
Look Up Tables (LUTs) perform arbitrary functions on small bitwidth inputs (2-6 bits)
→ used for boolean operations, arithmetics, memory

Flip-flops register data in time with the clock pulse

FPGA diagram



Logic cell



What are FPGAs?

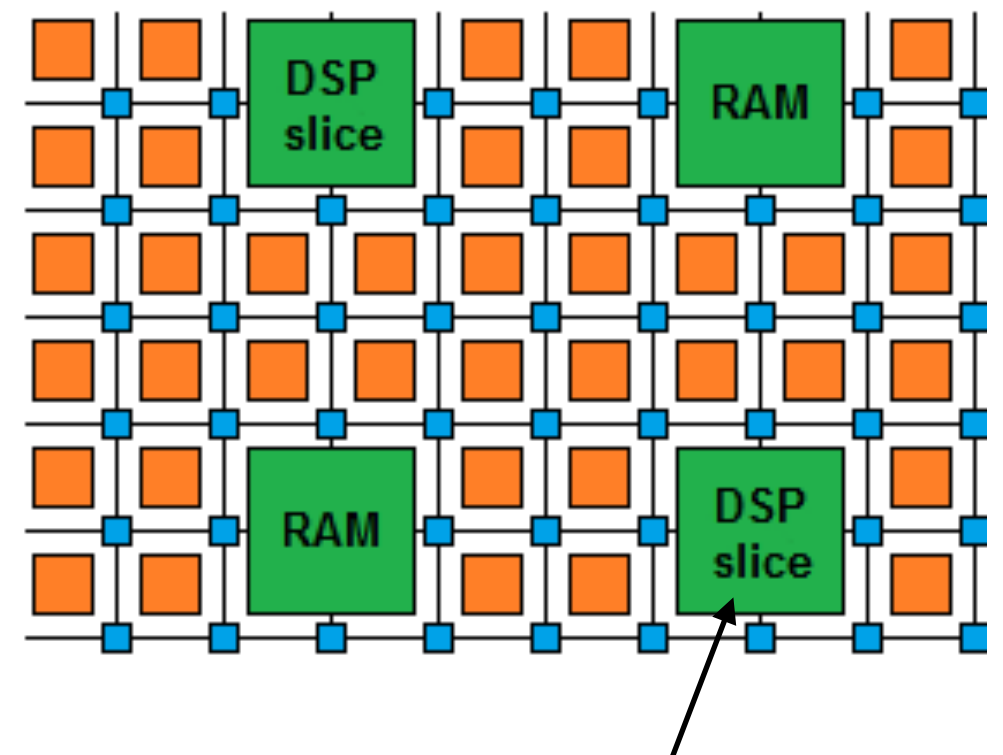
Field Programmable Gate Arrays
are reprogrammable integrated circuits

DSPs are specialized units for multiplication and arithmetic

→ faster and more efficient than LUTs for these type of operations

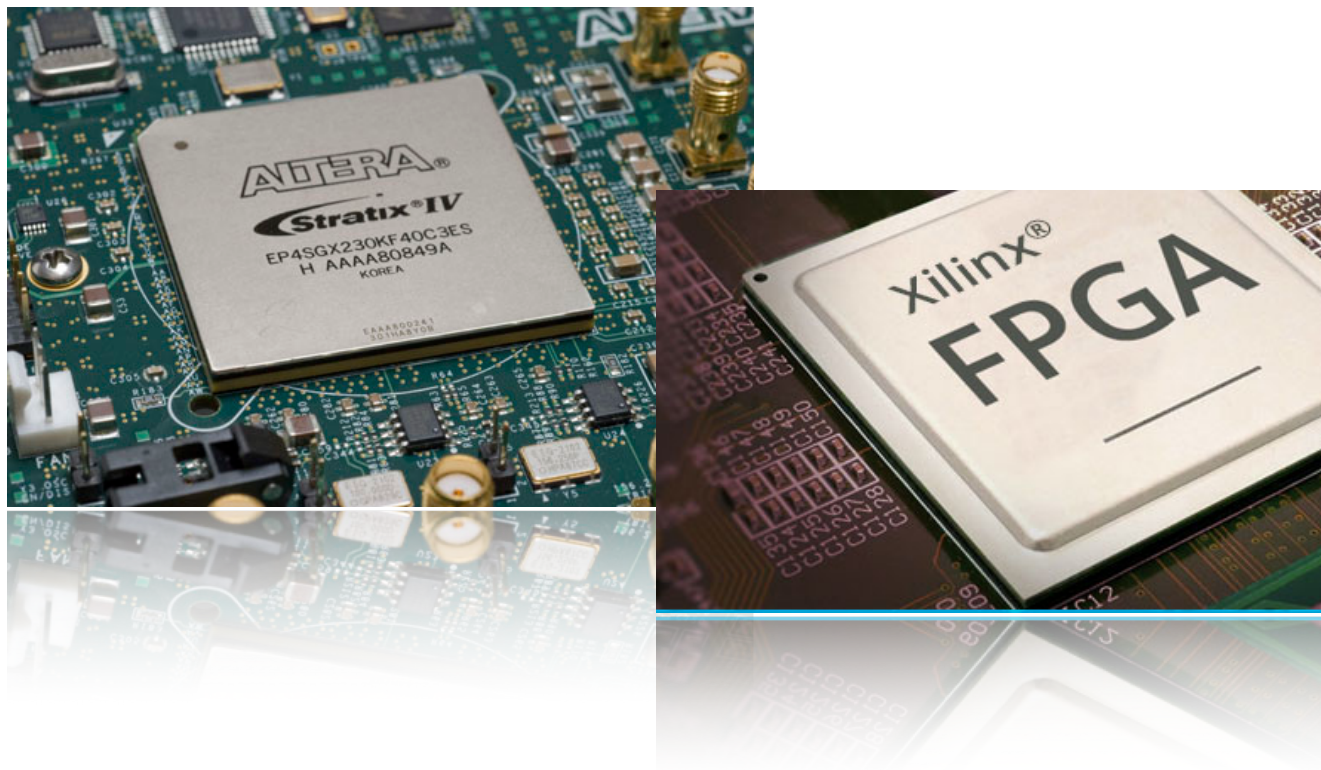
→ for deep learning, they are often the most precious resource

FPGA diagram



Also contain embedded components:

Digital Signal Processors (DSPs): logic units used for multiplications



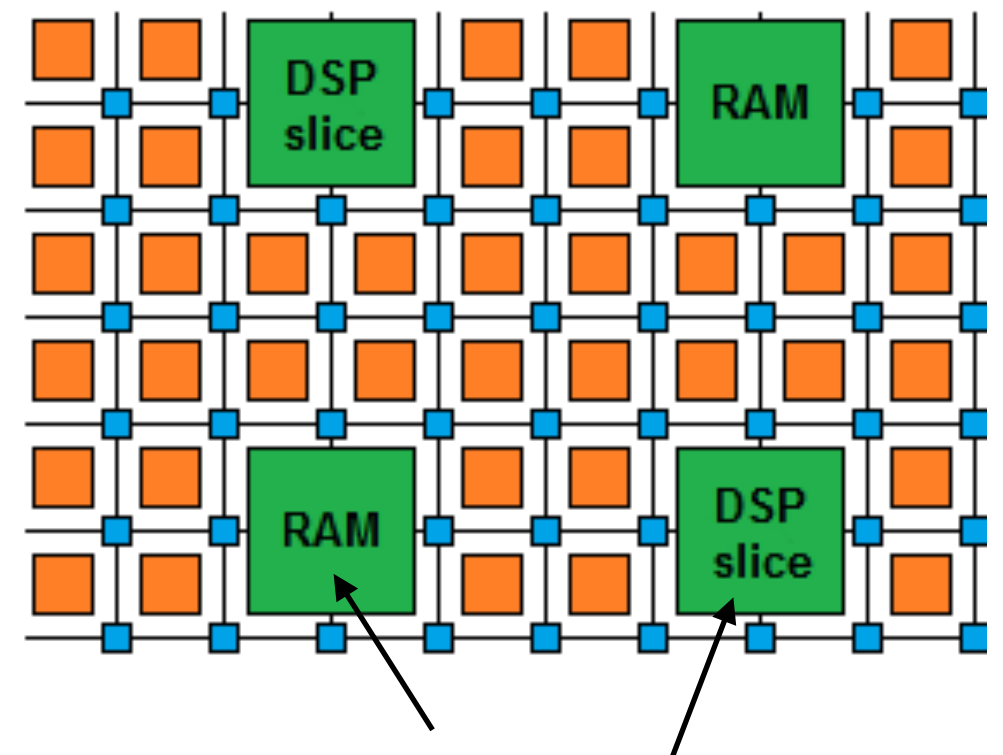
What are FPGAs?

Field Programmable Gate Arrays
are reprogrammable integrated circuits

BRAMs are small, fast memories (ex, 18 Kb each)
→ more efficient than LUTs when large memory is required

Modern FPGAs have ~100 Mb of BRAMs,
chained together as needed

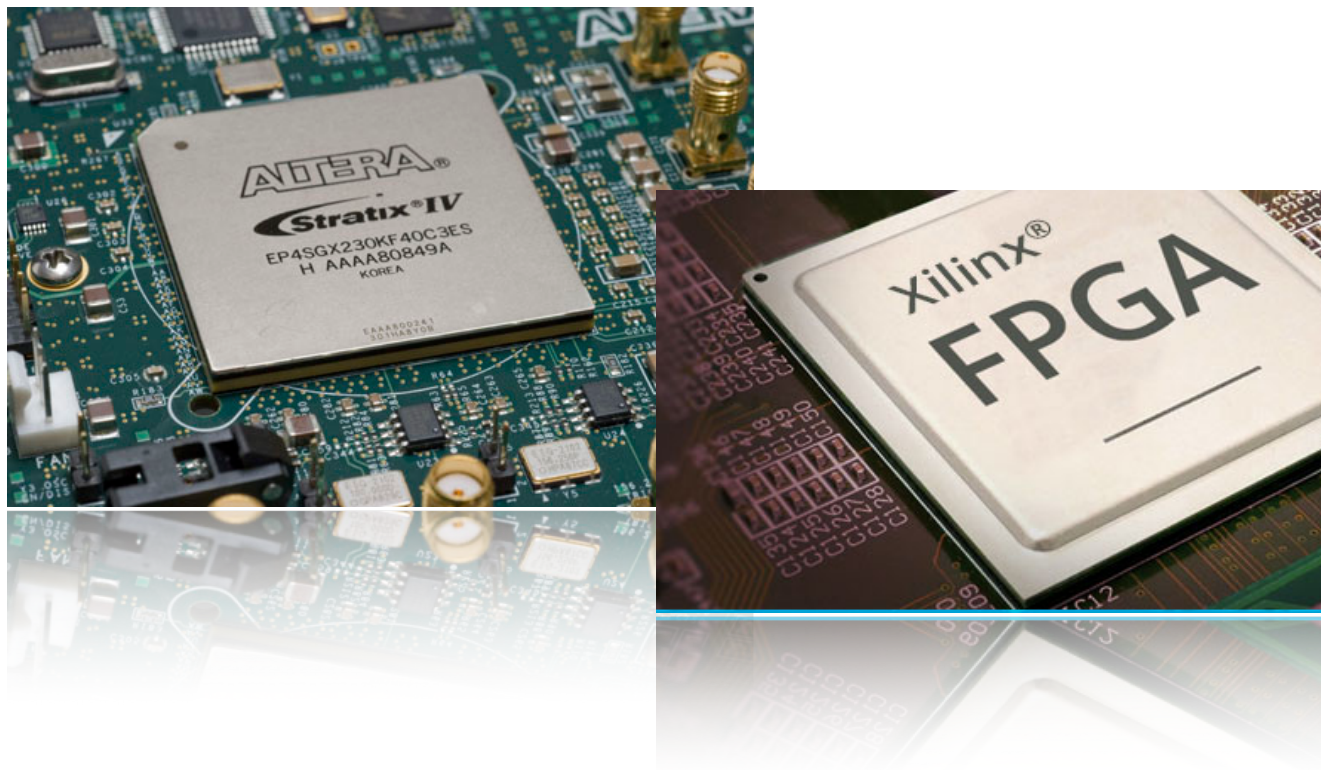
FPGA diagram



Also contain embedded components:

Digital Signal Processors (DSPs): logic units used for multiplications

Random-access memories (RAMs): embedded memory elements



What are FPGAs?

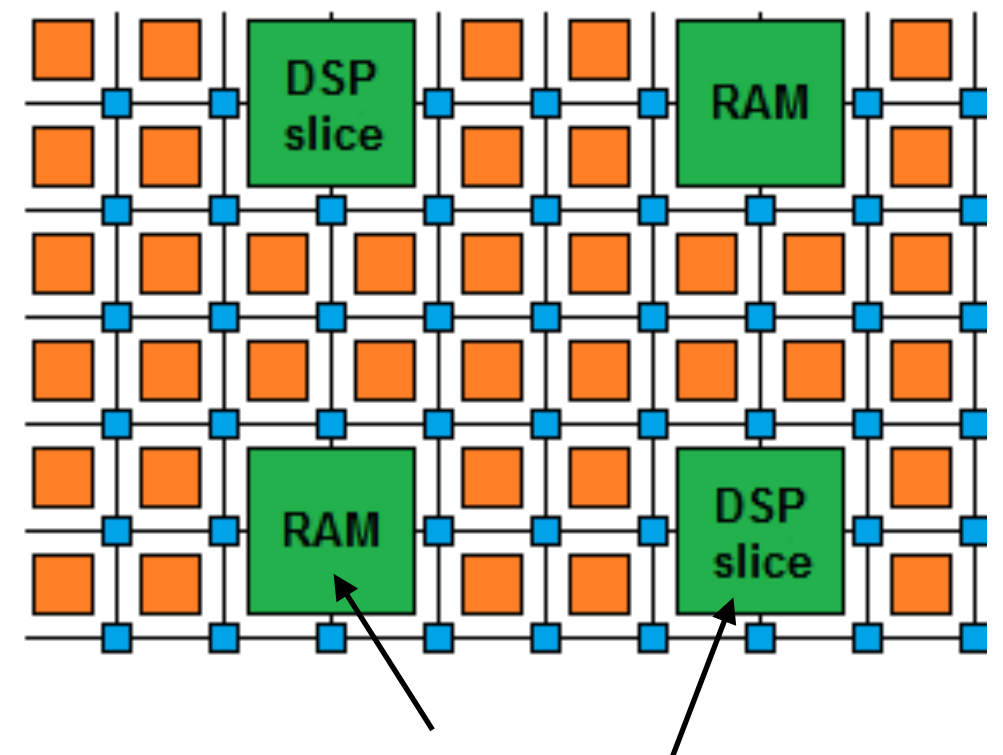
Field Programmable Gate Arrays
are reprogrammable integrated circuits

Contain array of **logic cells** embedded with **DSPs**, **BRAMs**, etc.

Support **highly parallel** algorithm implementation

Low power per Op (relative to CPU/GPU)

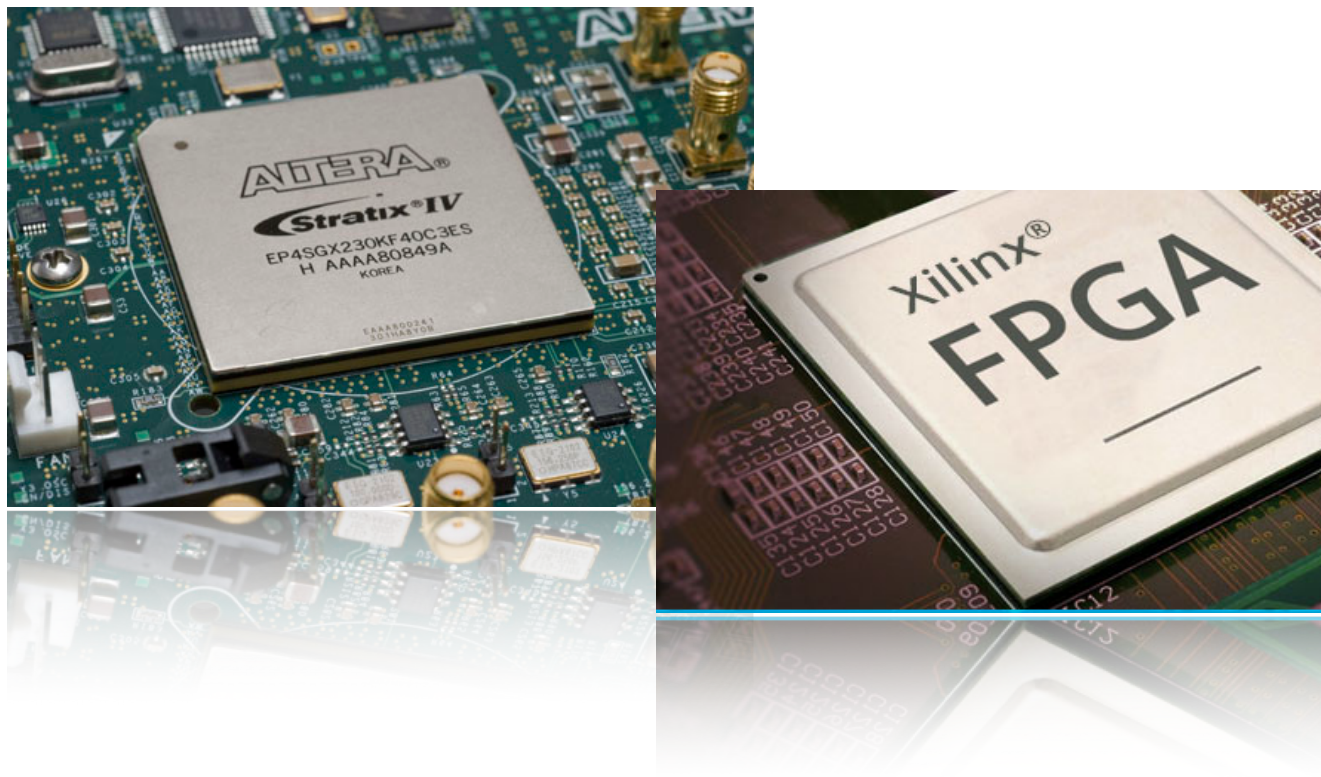
FPGA diagram



Also contain embedded components:

Digital Signal Processors (DSPs): logic units used for multiplications

Random-access memories (RAMs): embedded memory elements



Why are FPGAs fast?

- Fine-grained / resource parallelism
 - use the many resources to work on different parts of the problem simultaneously
 - allows us to achieve **low latency**
- Most problems have at least some sequential aspect, limiting how low latency we can go
 - but we can still take advantage of it with...
- Pipeline parallelism
 - instruct the FPGA to work on different data simultaneously
 - allows us to achieve **high throughput**



Like a production line for data...

How are FPGAs programmed?

Hardware Description Languages

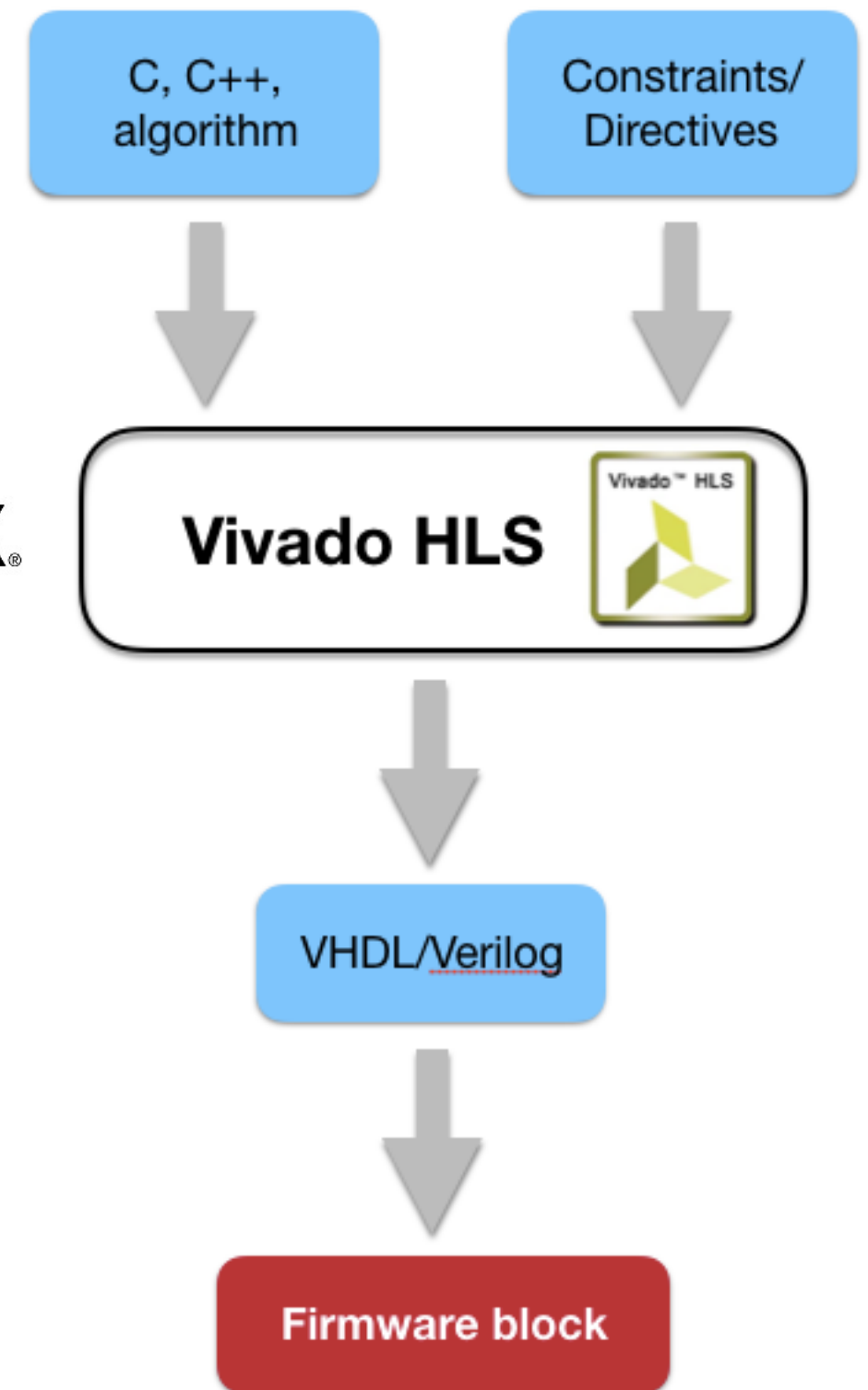
HDLs are programming languages which describe electronic circuits

High Level Synthesis

generate HDL from more common C/C++ code
pre-processor directives and constraints used to optimize the timing

drastic decrease in firmware development time!

See [Xilinx Vivado HLS](#), [Intel HLS](#), [Catapult HLS](#)



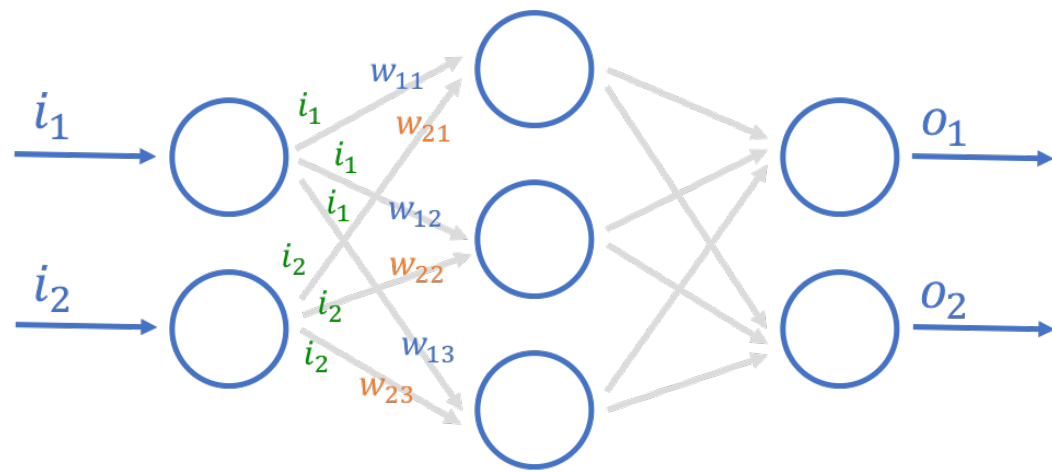
Neural Network inference on FPGA

Neural network inference
=
matrix multiplication



Efficient implementation on FPGA uses
DIGITAL SIGNAL PROCESSORS

There are about 5–10k DSPs in modern
FPGAs!



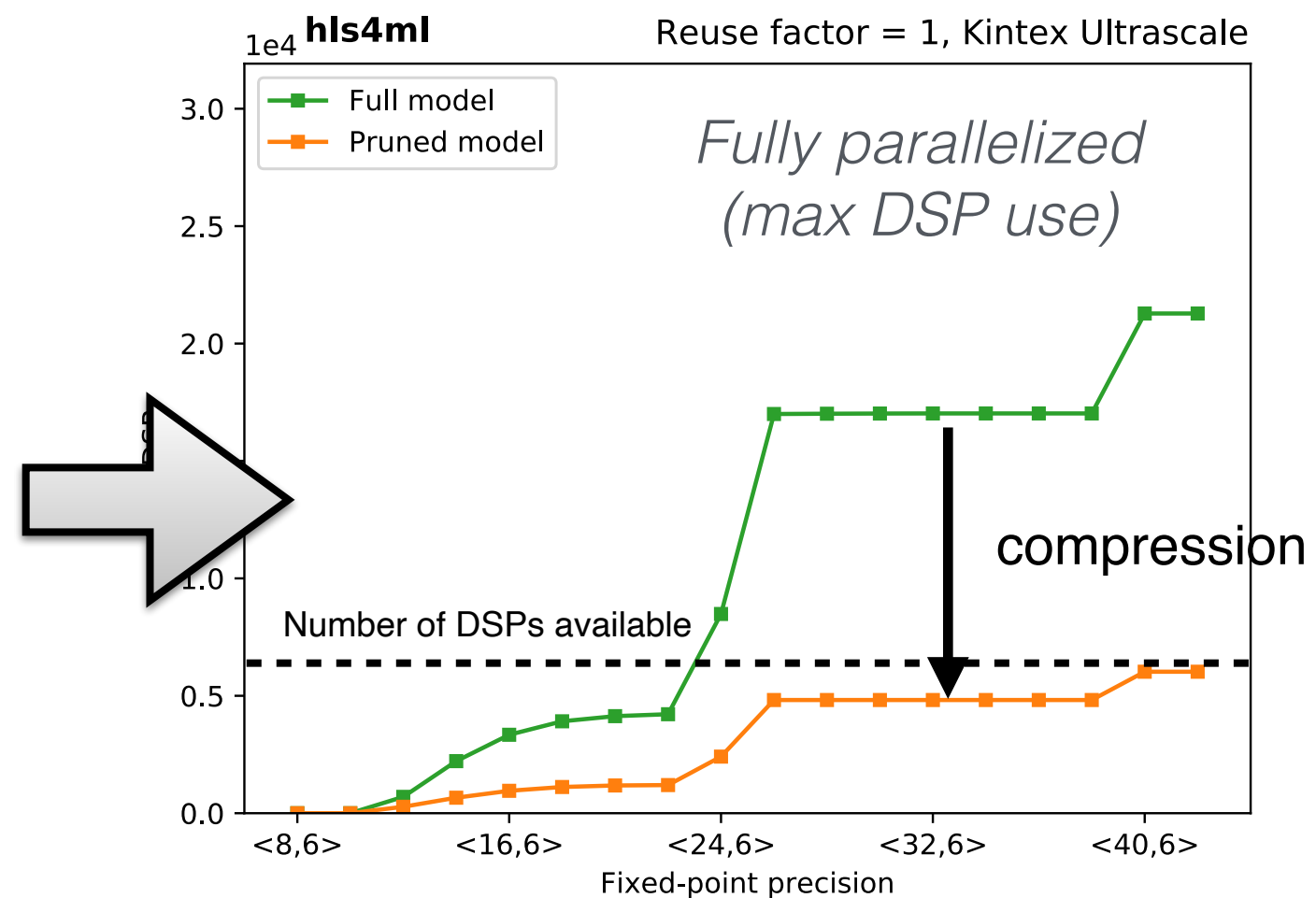
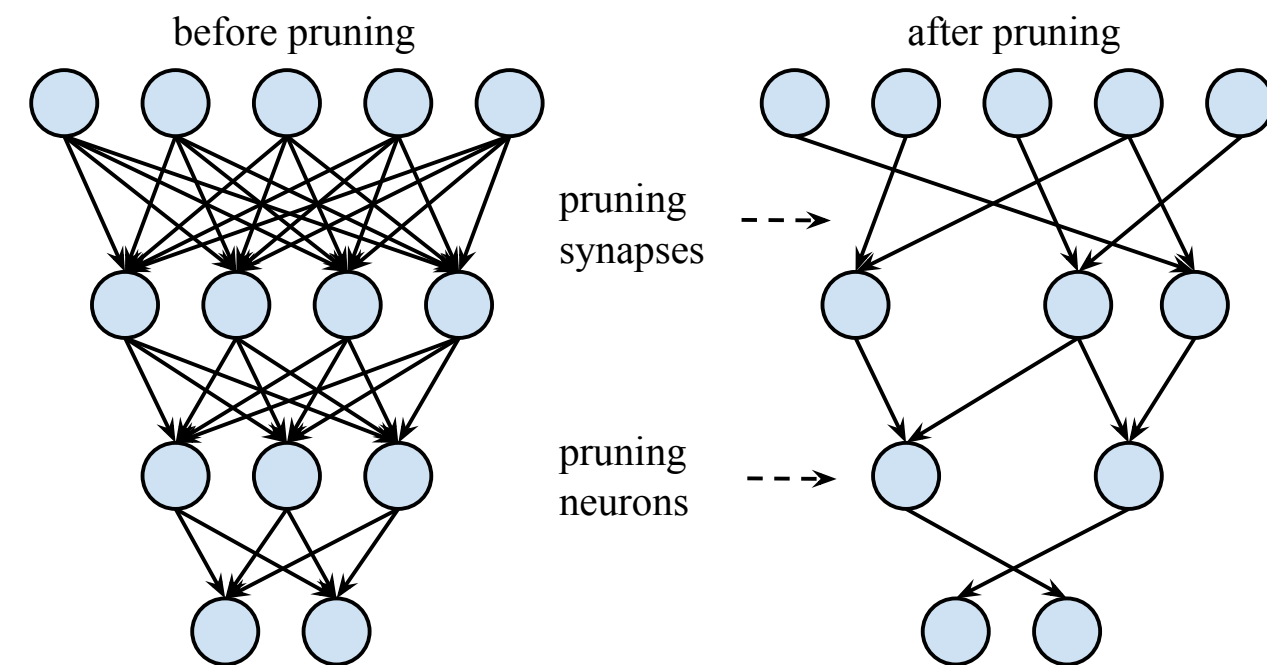
$$\begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix} \cdot \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} (w_{11} \times i_1) + (w_{21} \times i_2) \\ (w_{12} \times i_1) + (w_{22} \times i_2) \\ (w_{13} \times i_1) + (w_{23} \times i_2) \end{bmatrix}$$



ex: Xilinx Virtex Ultrascale +

Make the model fit on one chip

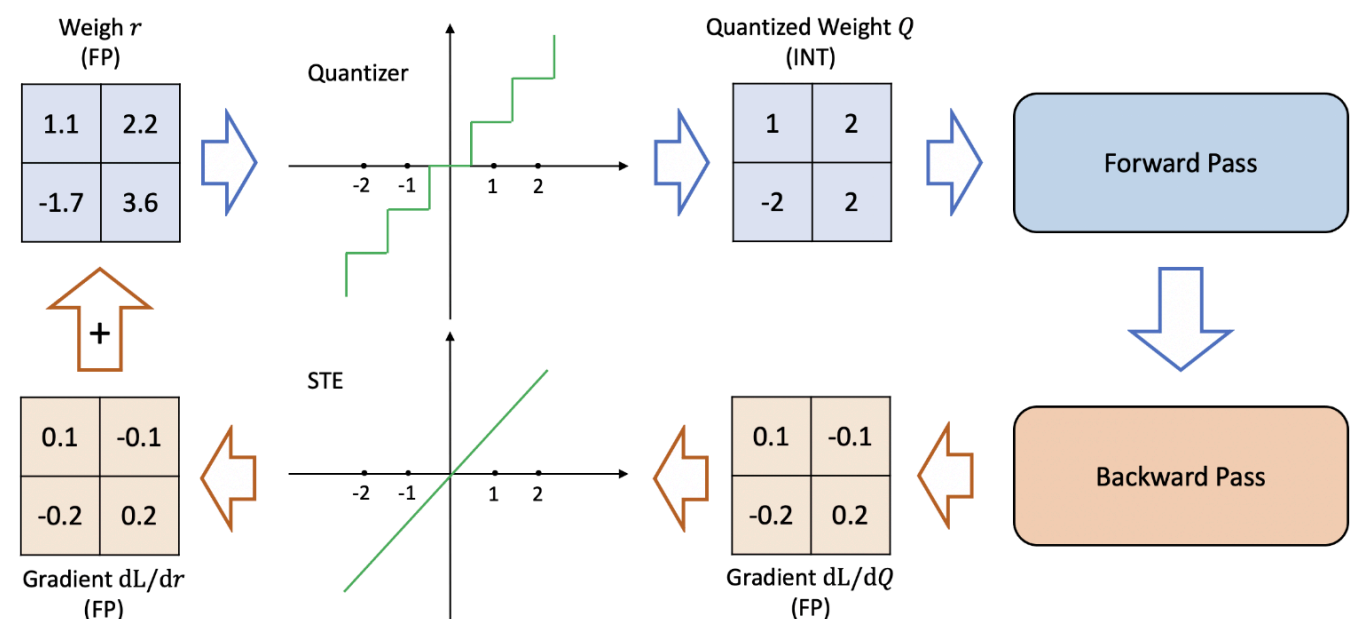
- Some tricks are needed here:
 - **Compression/pruning:** remove the connections that play little role for final decision



70% compression ~ 70% fewer DSPs

Quantization-aware training

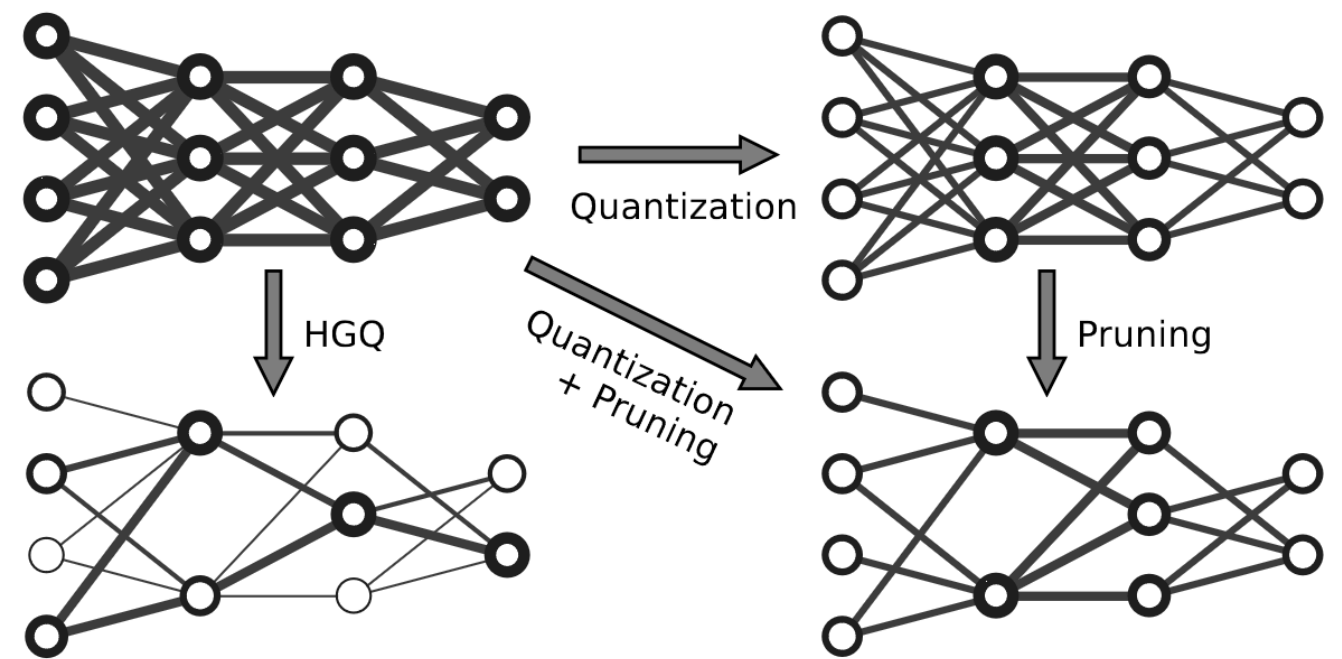
- **Post-training quantization can affect accuracy**
 - for a given bit allocation, the loss minimum at floating-point precision might not be the minimum anymore
- One could **specify quantization while look for the minimum during training**
 - quantization functions applied to weights and activations only in the forward pass
 - use Straight Through Estimator for back propagation step
- **Our workflow:** quantization-aware training with [Google QKeras](#) and firmware design with [hls4ml](#) for most efficient NN inference on chip!



A. Gholami et al, arxiv.2103.13630

High-Granularity Quantization

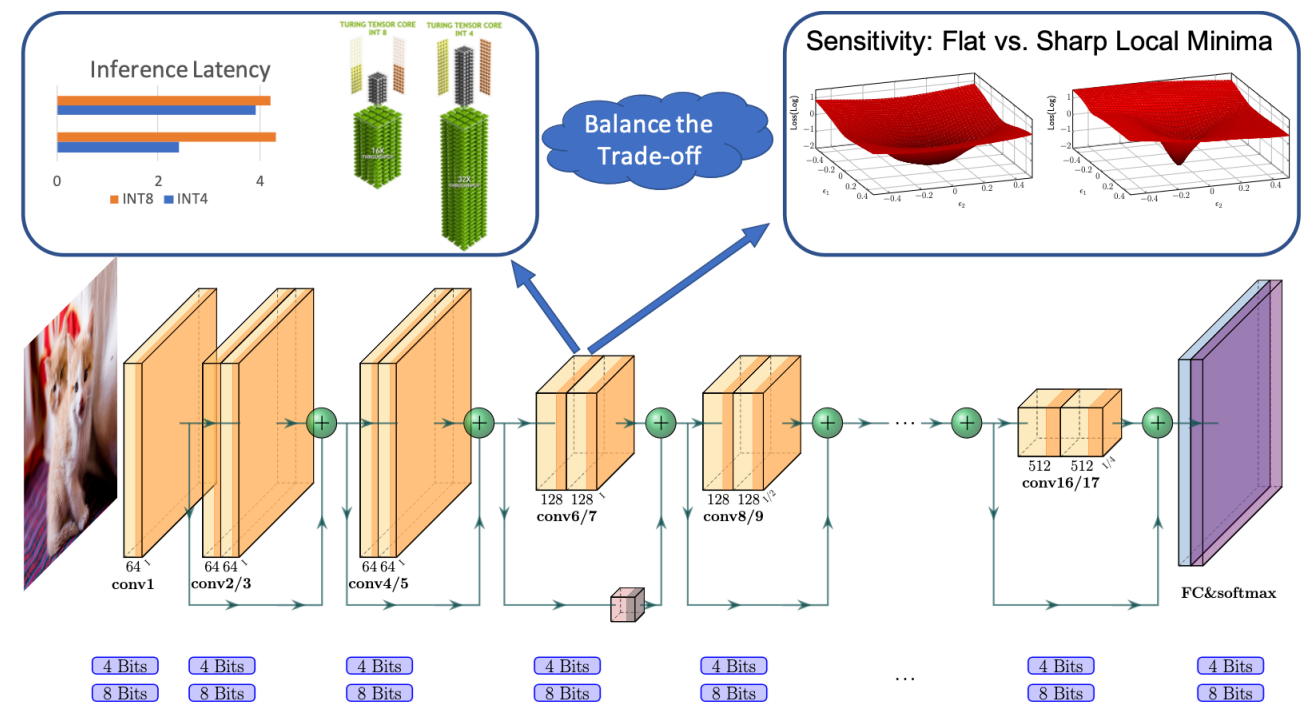
- **The wish:** squeeze even more NN inference performance when each parameter in the network may have its unique bitwidth
- **Limitations of QKeras:**
 - bitwidths for NN parameters are optimized in predefined, structured block (e.g., per layer)
 - bitwidth is not part of optimization
→ need to run your own hyperparameter scan
- **Solution: optimize the individual bitwidths alongside the NN accuracy using gradient descent**



Other quantization methods

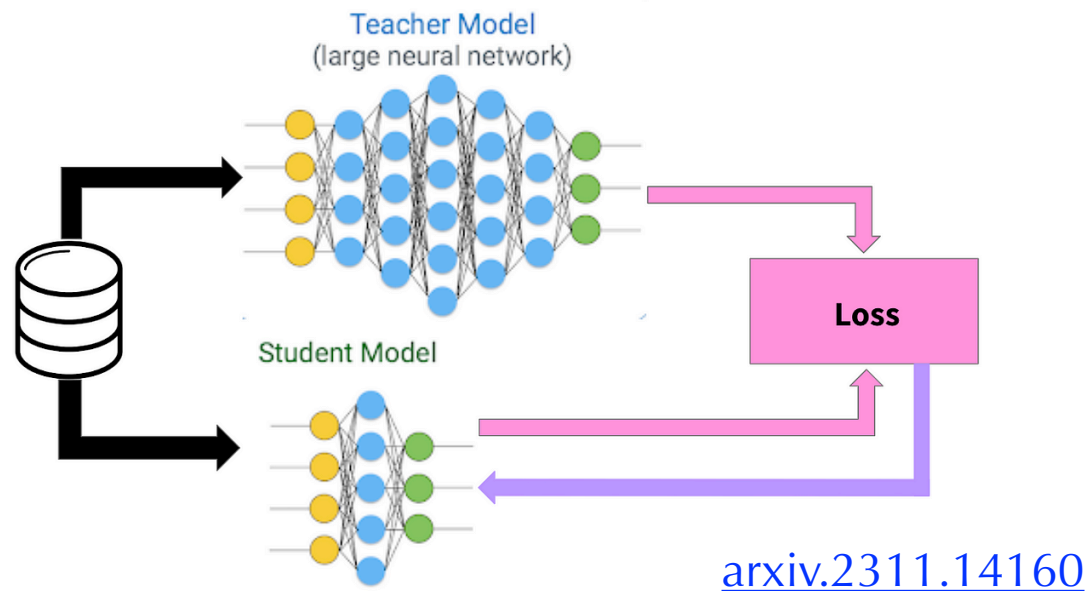
HAWQ — Hessian AWare Quantization

- mixed-precision quantization tool written for PyTorch
- main idea: **sensitive** layers are kept at higher precision than less **sensitive** layers
- problem: search space is **exponential** to the number of layers in models
- solution: use ILP to find the optimal trade-off between model perturbation (through Hessian trace) and application-specific constraints (latency, BOPs, size limit,...)
- Scales linearly w.r.t to the number of layers and bitwidth options



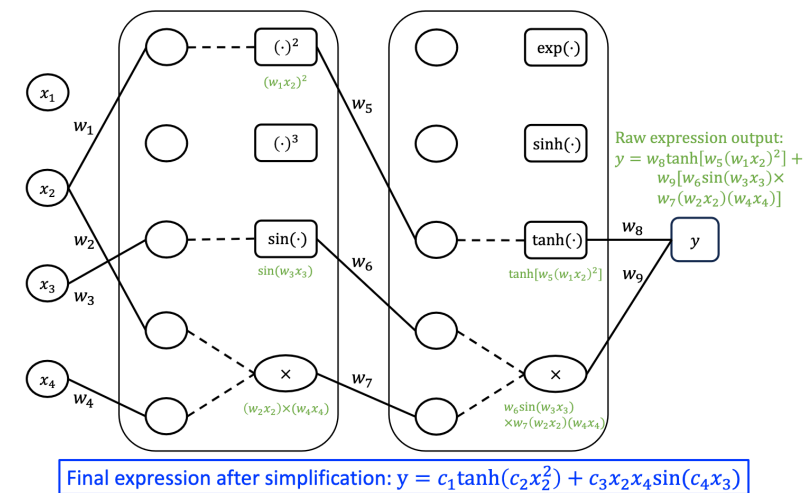
[arxiv.2011.10680](https://arxiv.org/abs/2011.10680)

Efficiency beyond quantization



Knowledge distillation

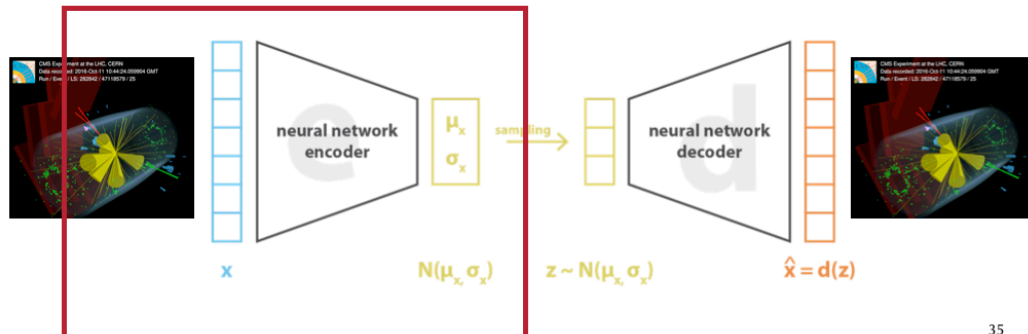
- Allows to deploy smaller student NN at similar accuracy of more complex teacher NN
- And to transfer powerful inductive bias to the student NN



Symbolic regression

- Trained with gradient-based approach can achieve high sparsity and compact representation
- Mathematical operations can be implemented efficiently in HLS with LUTs

Fast autoencoders for anomaly detection



[arxiv.2311.17162](https://arxiv.org/abs/2311.17162), [2401.08777](https://arxiv.org/abs/2401.08777), [2108.03986](https://arxiv.org/abs/2108.03986)

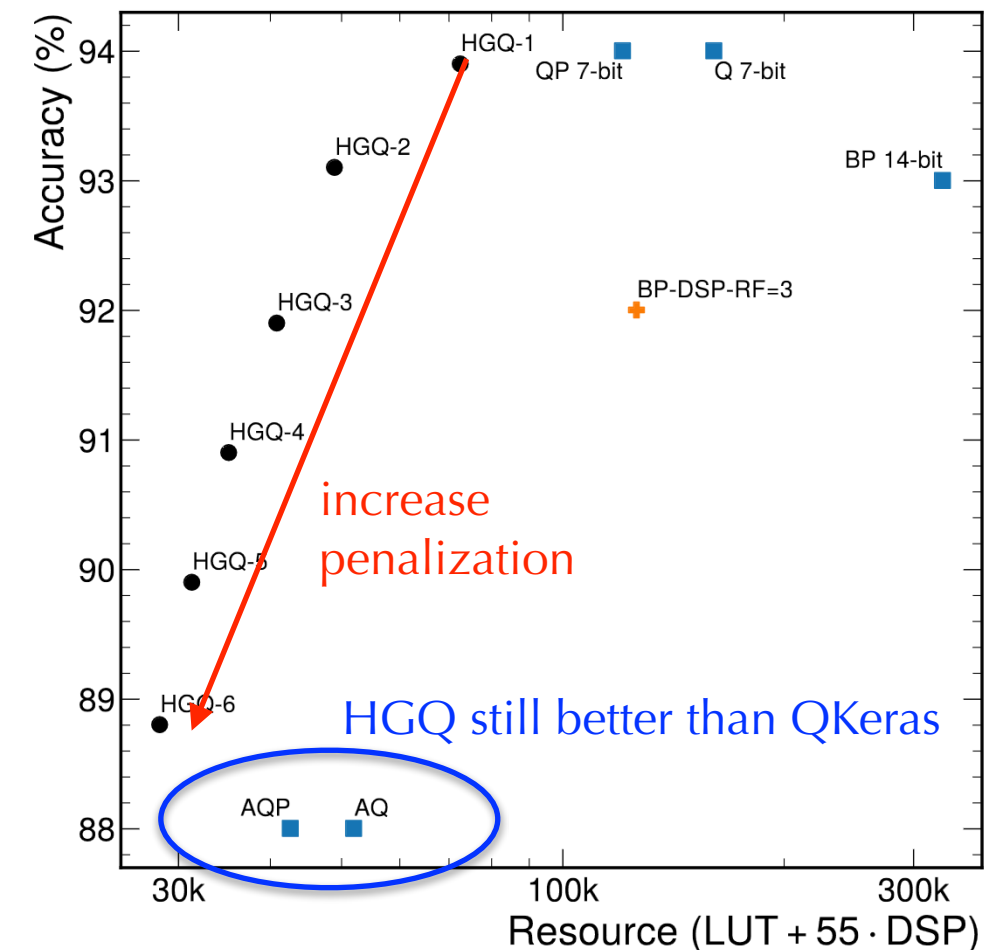
- If variational, define anomaly metric in latent space \rightarrow deploy only encoder in inference \rightarrow half latency and model size
- Informed latent representations can lead to more efficient model \rightarrow SSL and compact foundation models/transformers

High-Granularity Quantization

- **Solution: optimize the individual bitwidths alongside the NN accuracy using gradient descent**

- **How:**

- treat the bitwidths as continuous variables
- introduce surrogate gradients for discrete variables such as bitwidths
- introduce a novel on-chip resource consumption metric that when incorporated into the loss function penalizes larger bitwidths efficiently
- pruning integrated naturally in the optimization step (gradient descent reduces certain bitwidths to zero)



Gradient-based Automatic Mixed Precision Quantization for Neural Networks On-Chip

Chang Sun,^{1,2,*} Thea K. Årrestad,¹ Vladimir Loncar,^{3,4} Jennifer Ngadiuba,⁵ and Maria Spiropulu²

¹ETH Zurich (Zurich, Switzerland)

²California Institute of Technology (CA, USA)

³Massachusetts Institute of Technology (MA, USA)

⁴Institute of Physics Belgrade (Belgrade, Serbia)

⁵Fermi National Accelerator Laboratory (IL, USA)

**Fully supported
in hls4ml!**

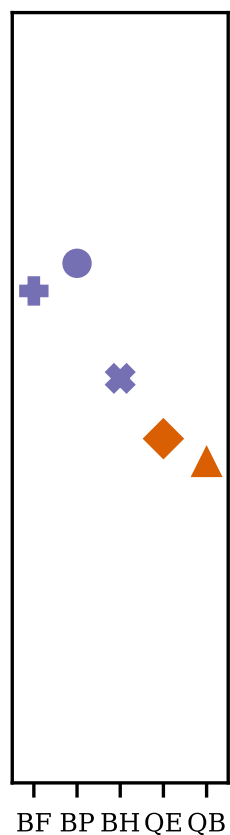
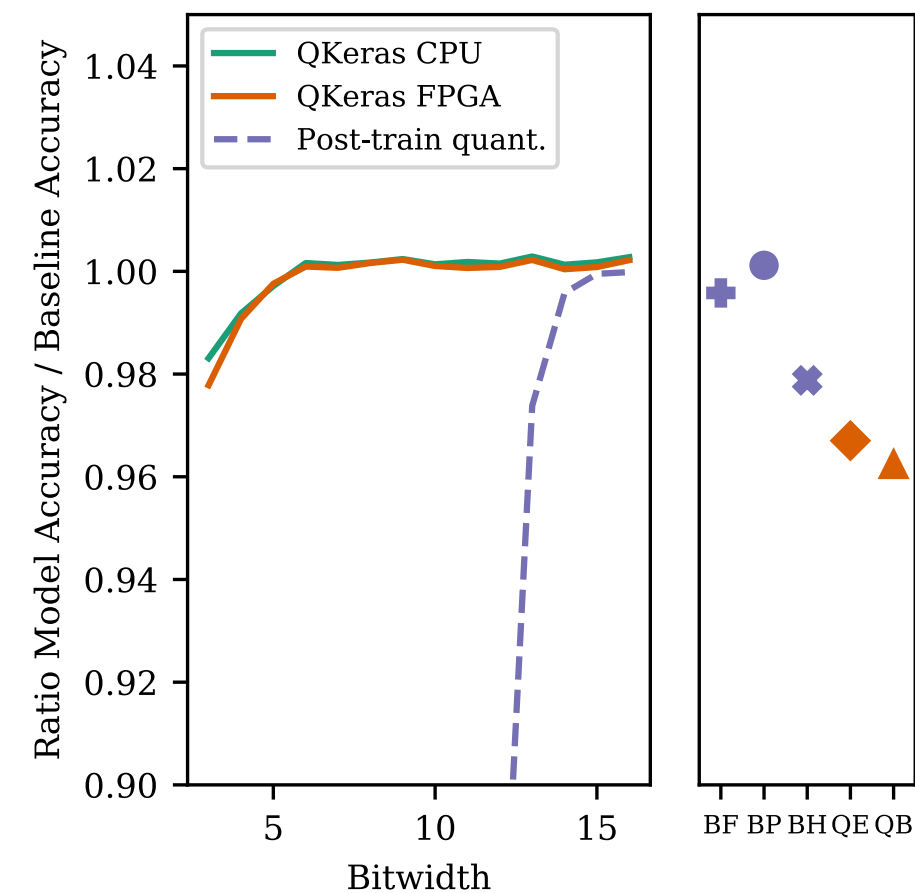


Fermilab
ETH zürich



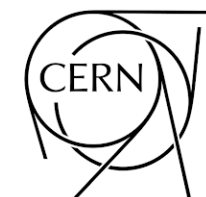
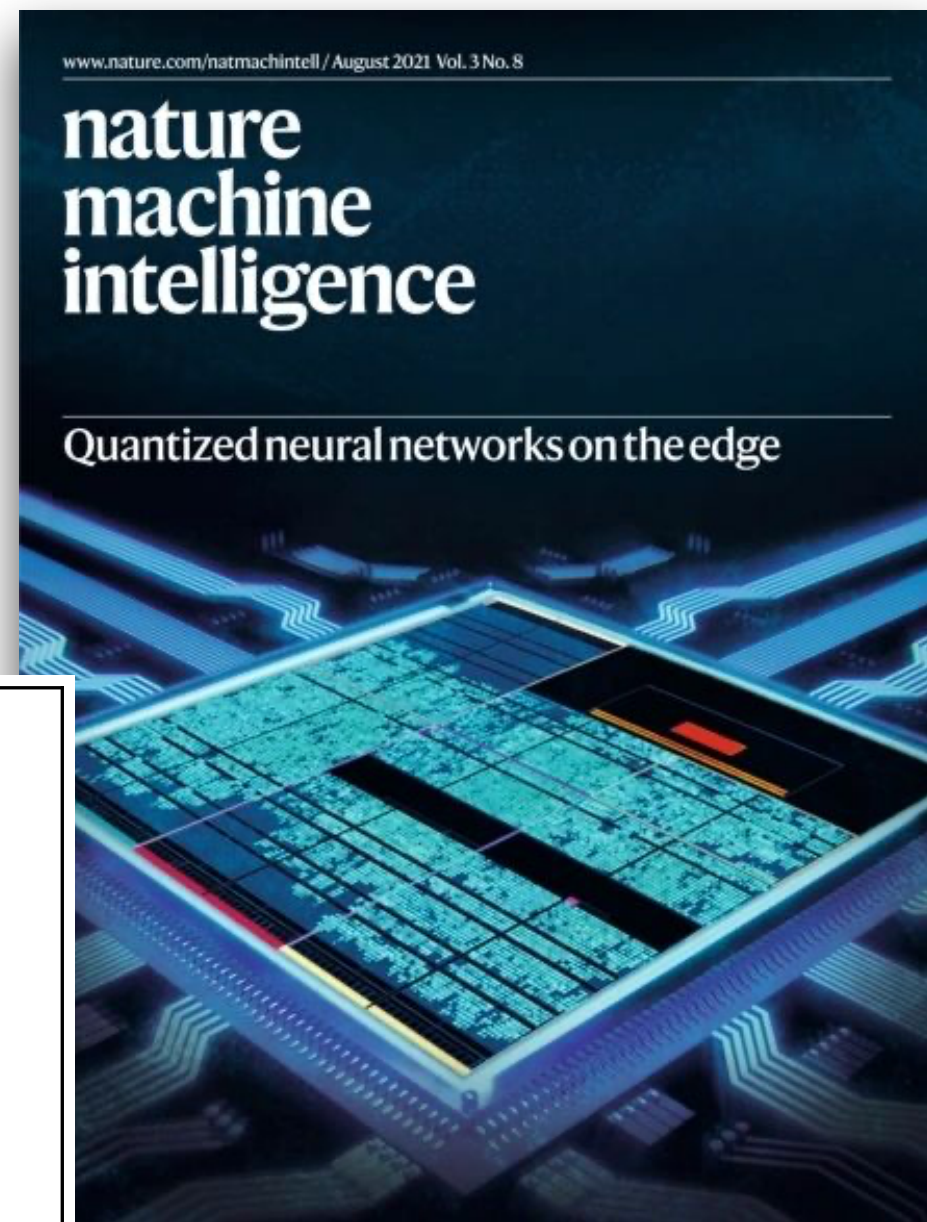
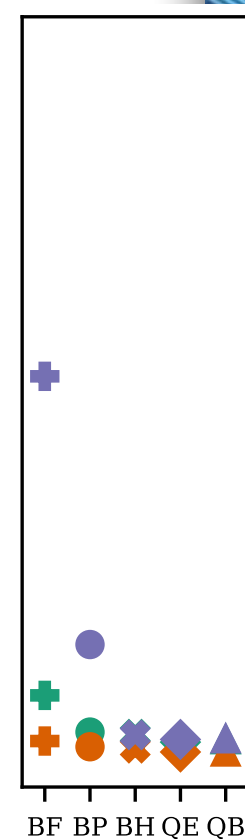
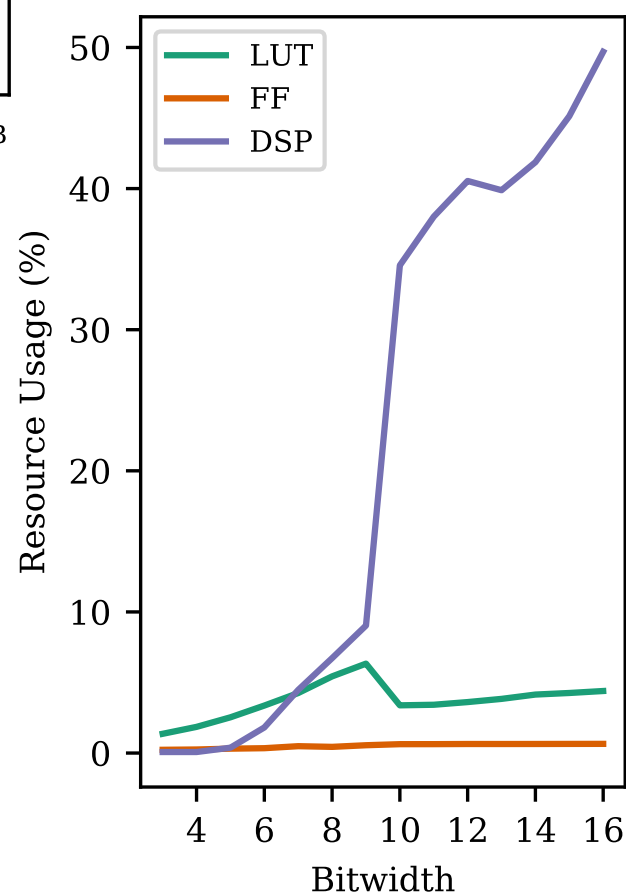
[Paper on arxiv ready for
submission!](#)

QKeras & hls4ml



Matches ap_fixed exactly!

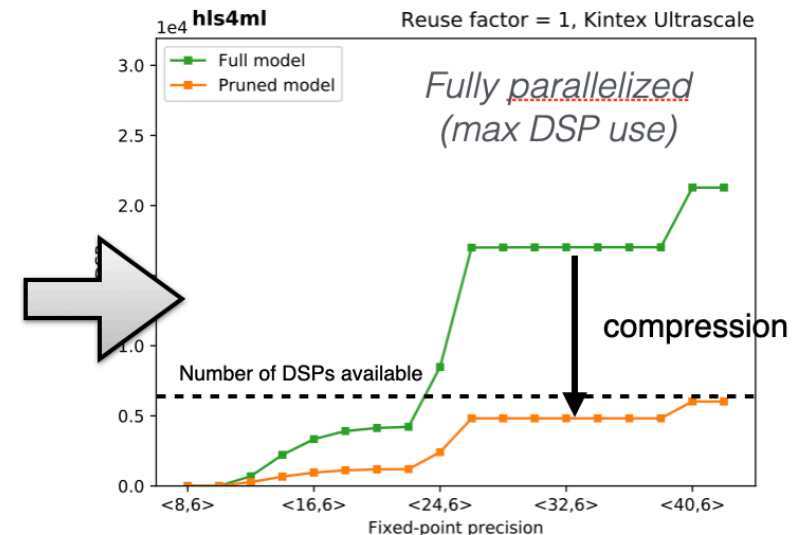
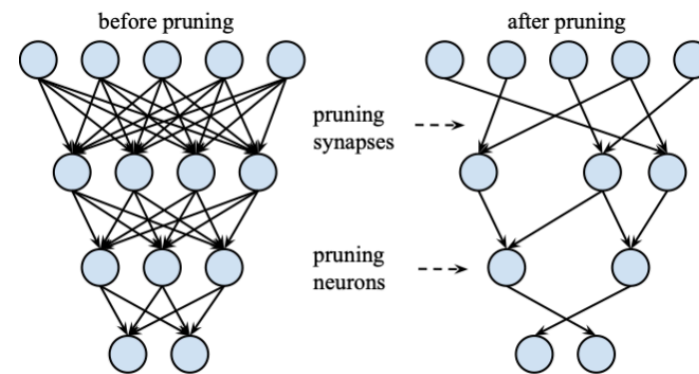
- same granularity as hls4ml
- same precision at training and inference



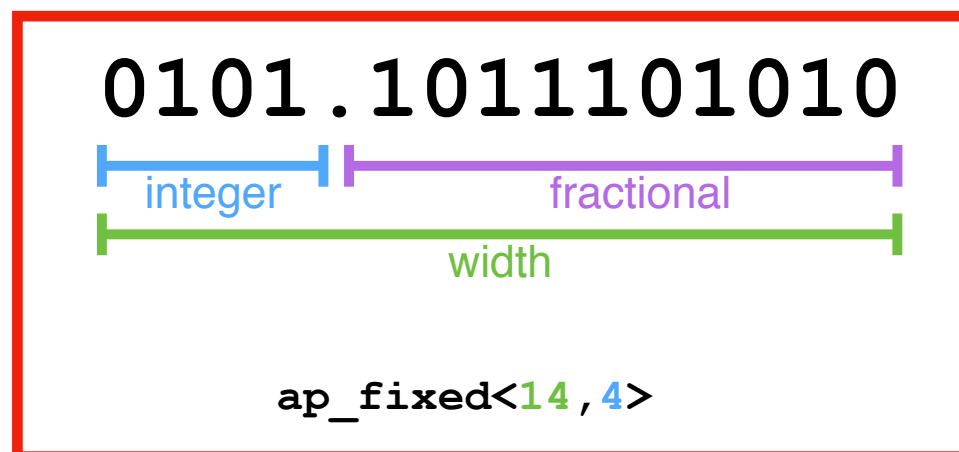
Make the model fit on one chip

- Some tricks are needed here:

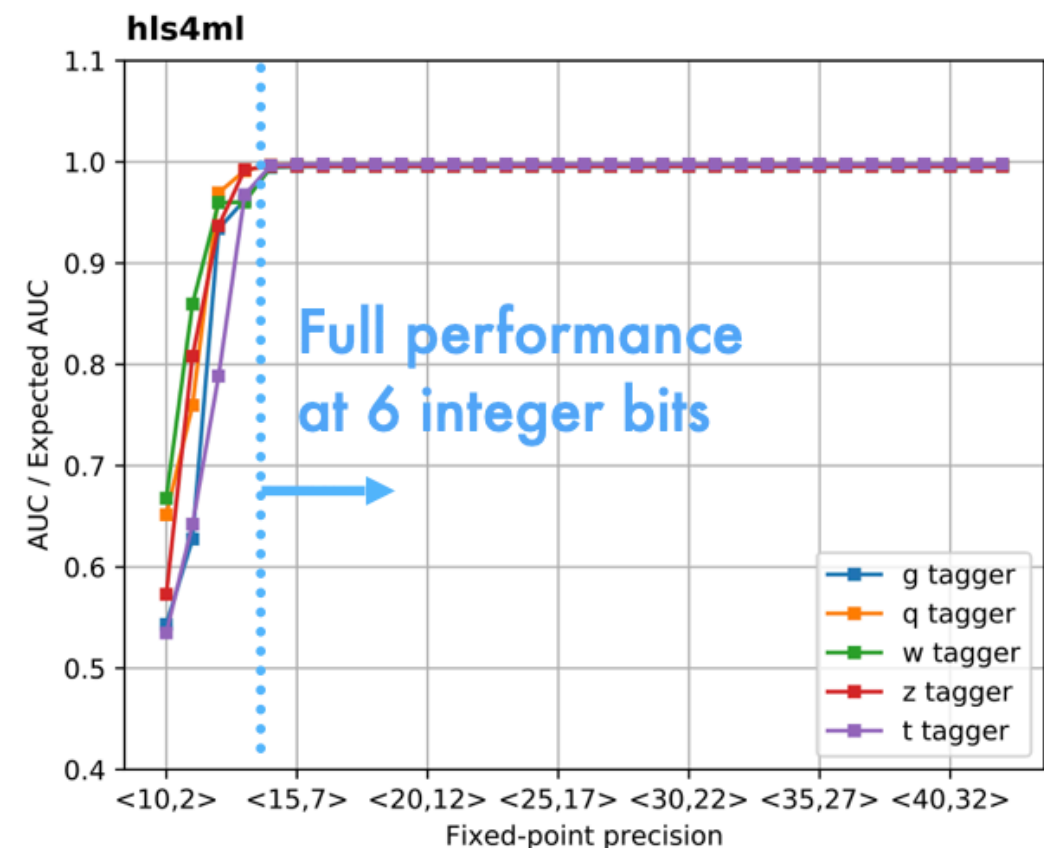
- **Compression/pruning:** remove the connections that play little role for final decision



- **Quantisation:** represents numbers with few bits reduce resources



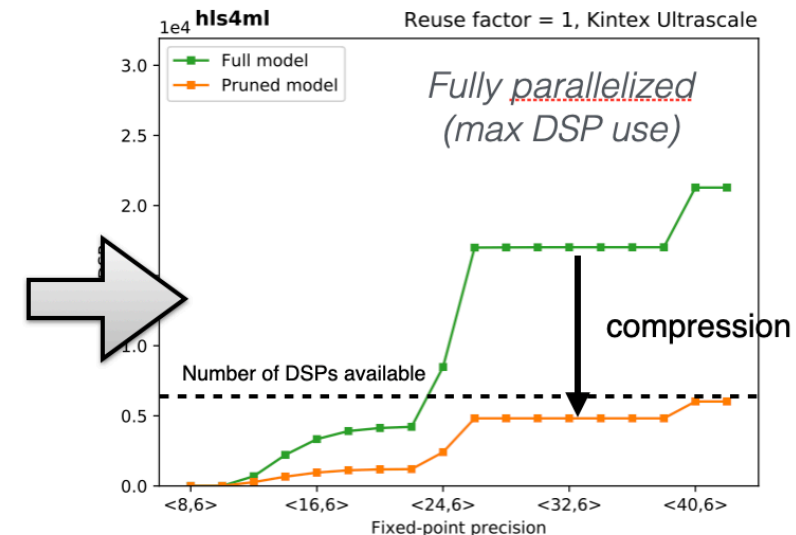
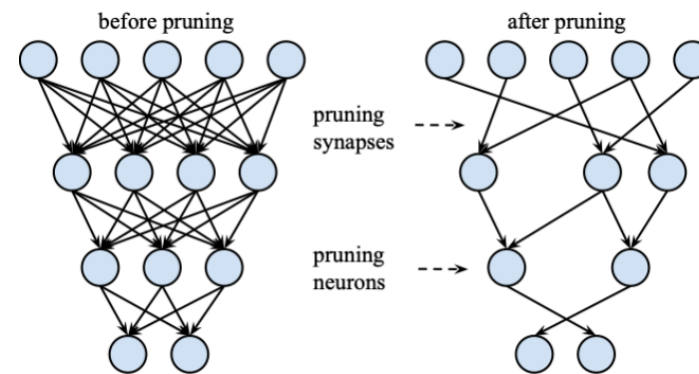
Scan integer bits
Fractional bits fixed to 8



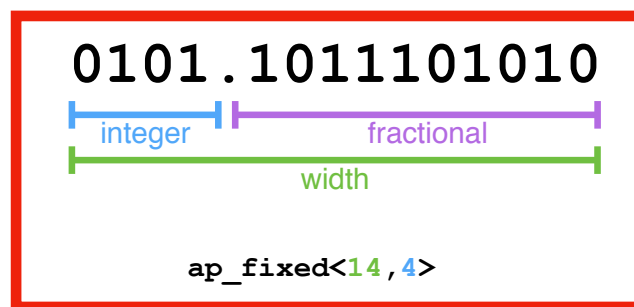
Make the model fit on one chip

- Some tricks are needed here:

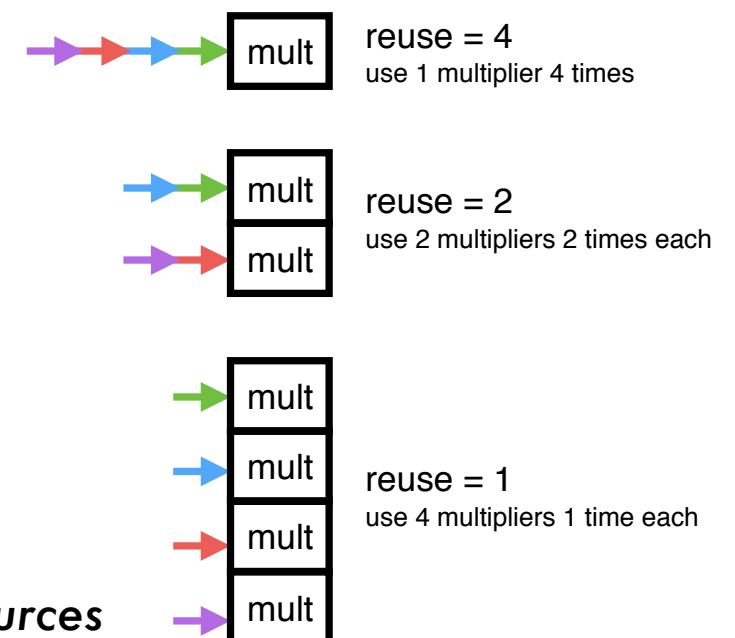
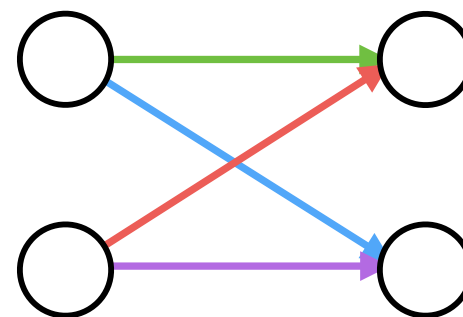
- **Compression/pruning:** remove the connections that play little role for final decision



- **Quantisation:** represents numbers with few bits reduce resources



- **Parallelization:** allocate resources for each operation (run all network in one clock) vs spread calculation across several clock cycles

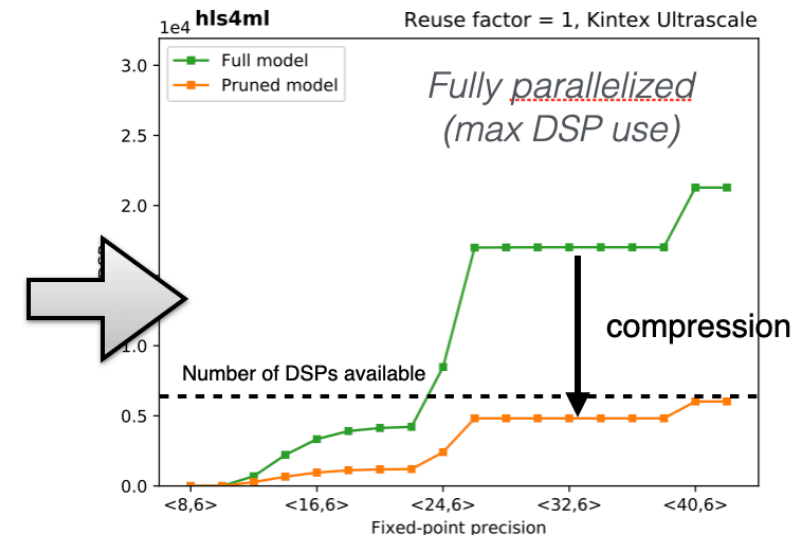
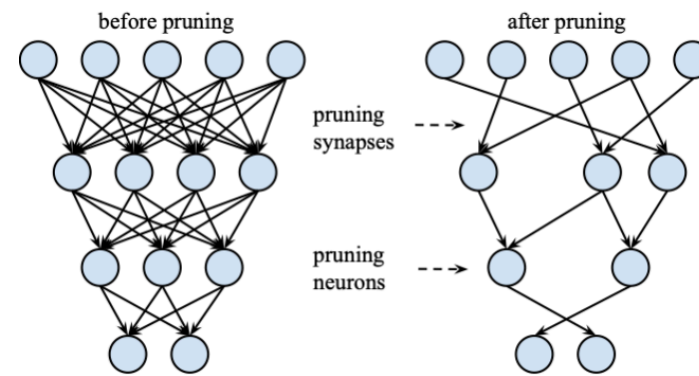


more parallelization → more resources

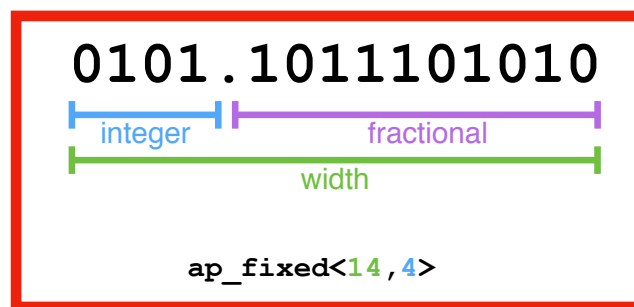
Make the model fit on one chip

- Some tricks are needed here:

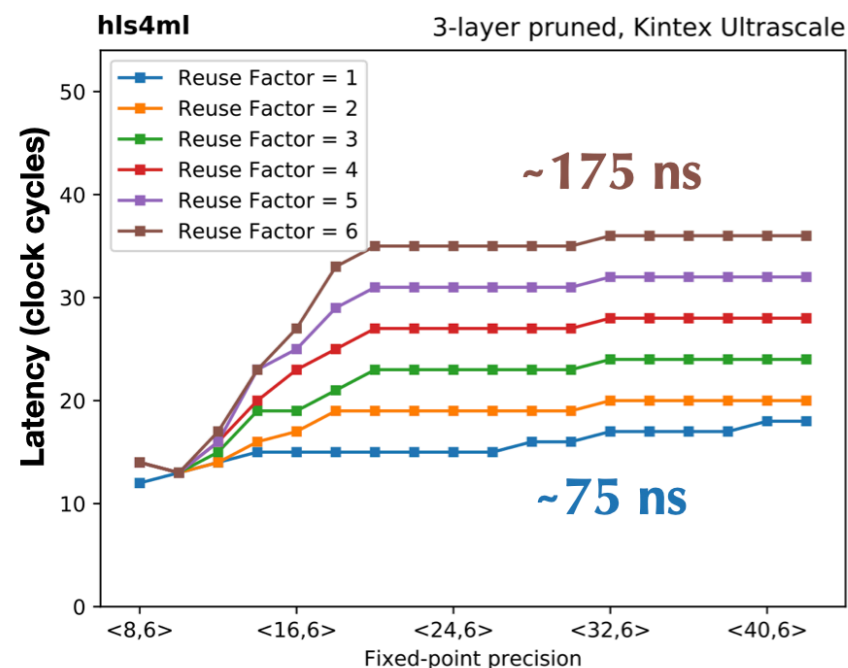
- **Compression/pruning:** remove the connections that play little role for final decision



- **Quantisation:** represents numbers with few bits reduce resources



- **Parallelization:** allocate resources for each operation (run all network in one clock) vs spread calculation across several clock cycles



Longer latency

Each mult. used 6x

Each mult. used 3x

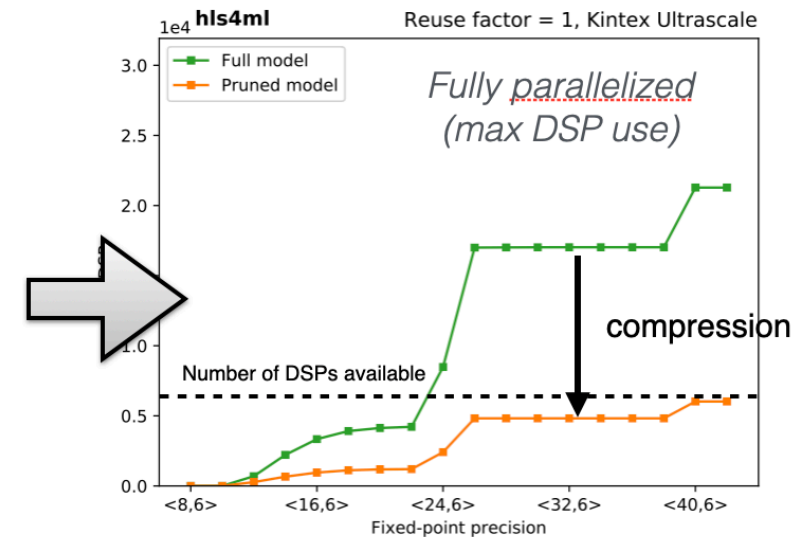
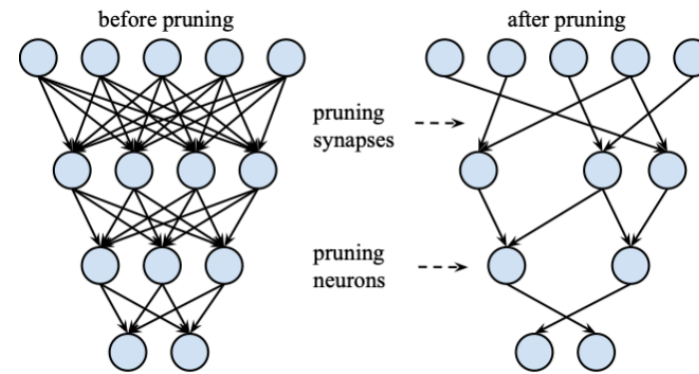
Fully parallel
Each mult. used 1x

More resources

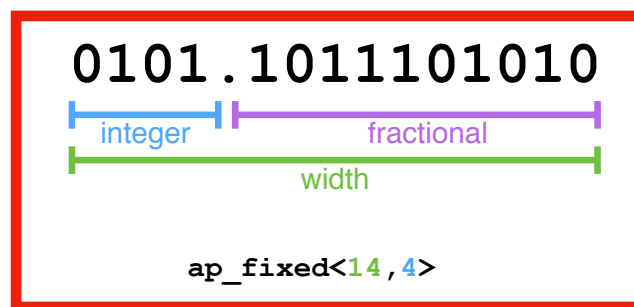
Make the model fit on one chip

- Some tricks are needed here:

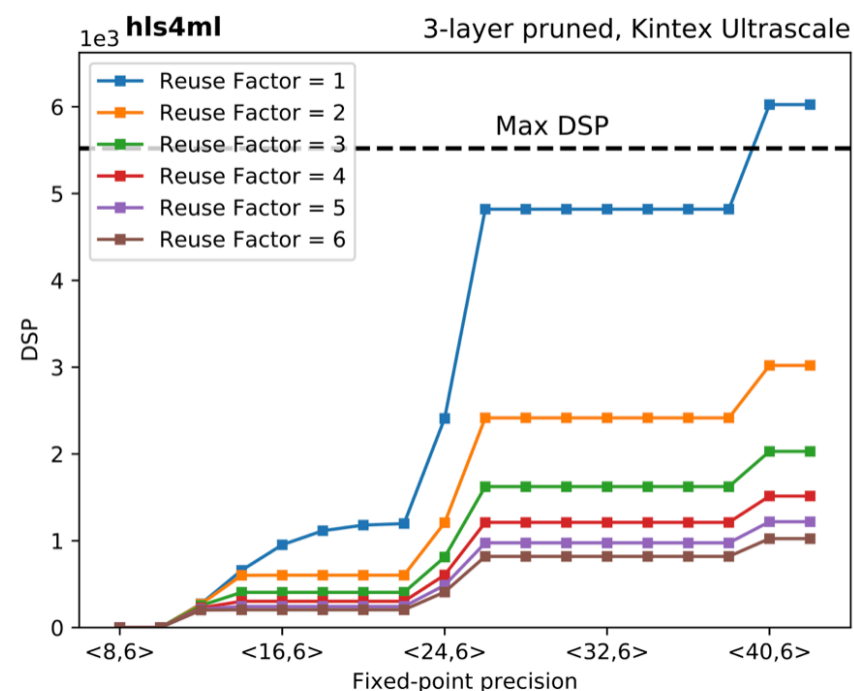
- **Compression/pruning:** remove the connections that play little role for final decision



- **Quantisation:** represents numbers with few bits reduce resources



- **Parallelization:** allocate resources for each operation (run all network in one clock) vs spread calculation across several clock cycles



More resources

Fully parallel
Each mult. used 1x

Each mult. used 2x

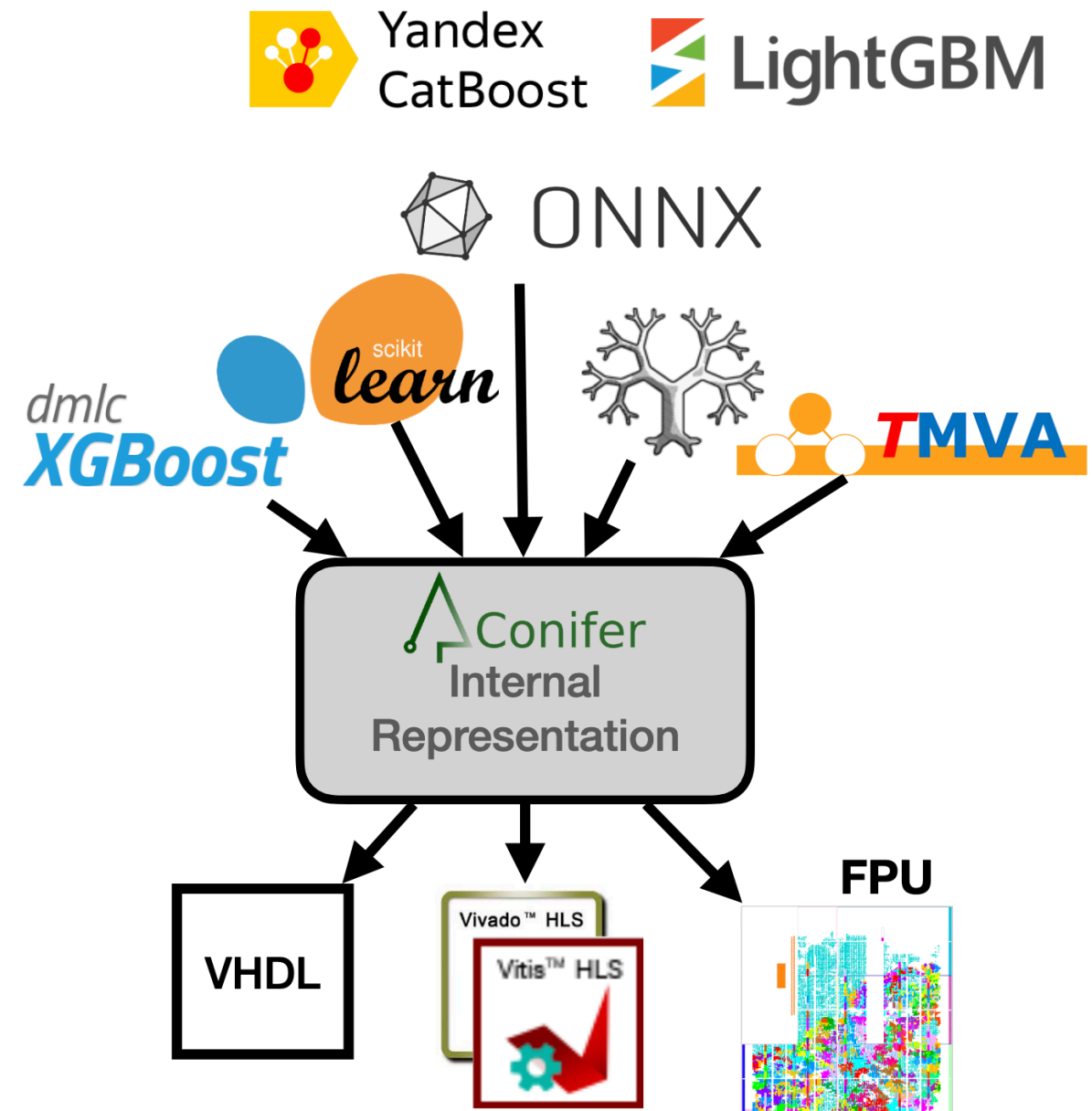
Each mult. used 3x

⋮

Longer latency

The Conifer tool for BDTs

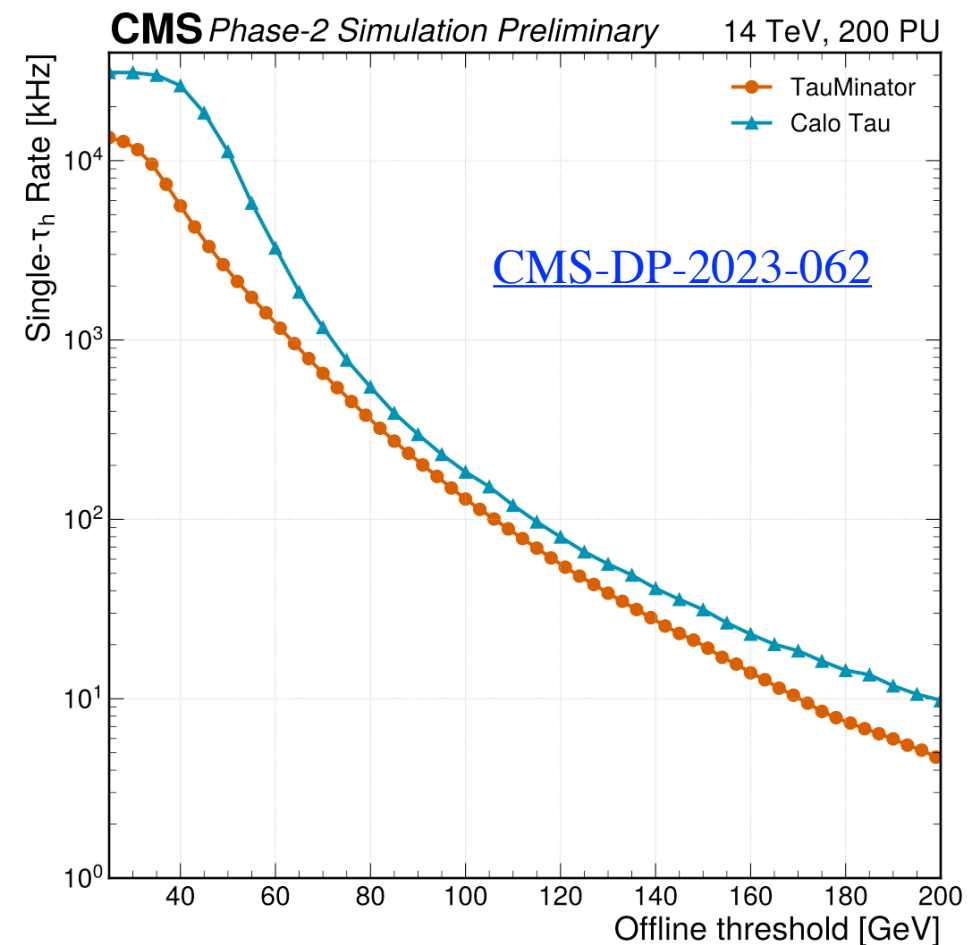
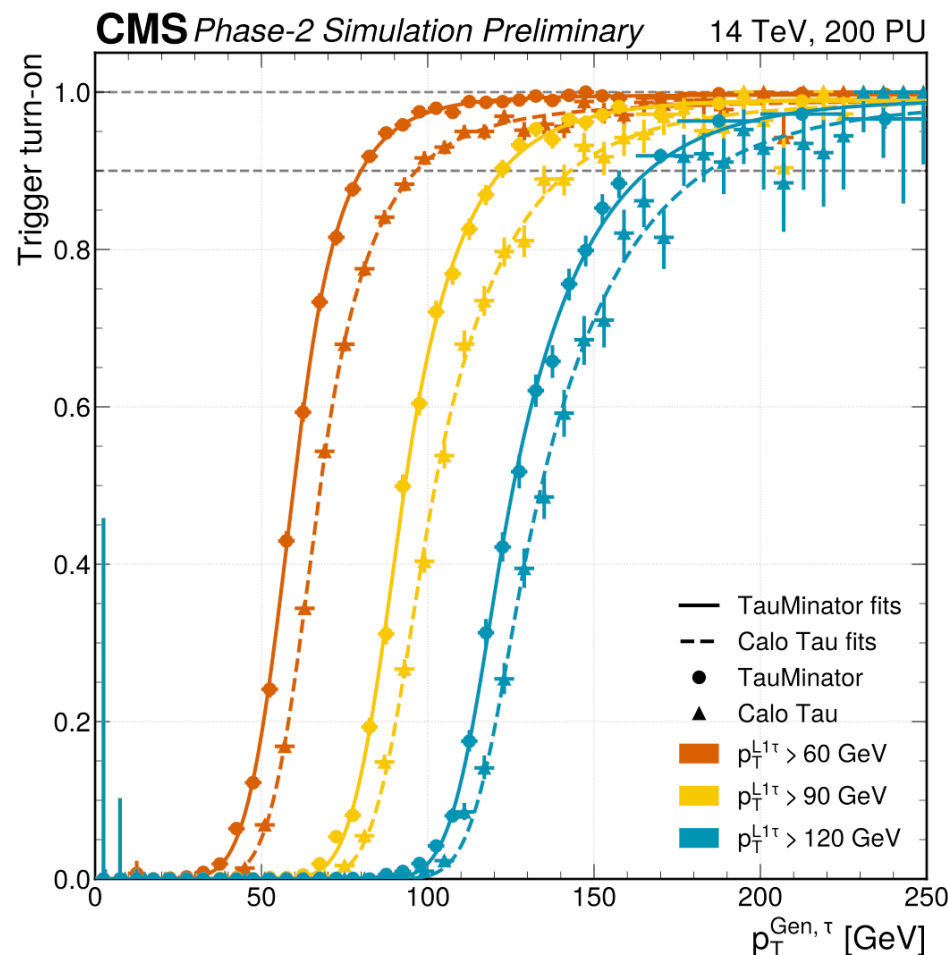
- Conifer is to DFs as hls4ml is to NNs
- Very much like hls4ml, conifer has frontends, an Internal Representation, and backends
- Frontend support for popular BDT training libraries
- Backends: HLS, (hand-written) VHDL, Forest Processing Unit (FPU)
- Conifer maps DFs onto FPGA logic: Implemented with high parallelism for low latency and high throughput



A few applications at the LHC

Hadronic τ reconstruction

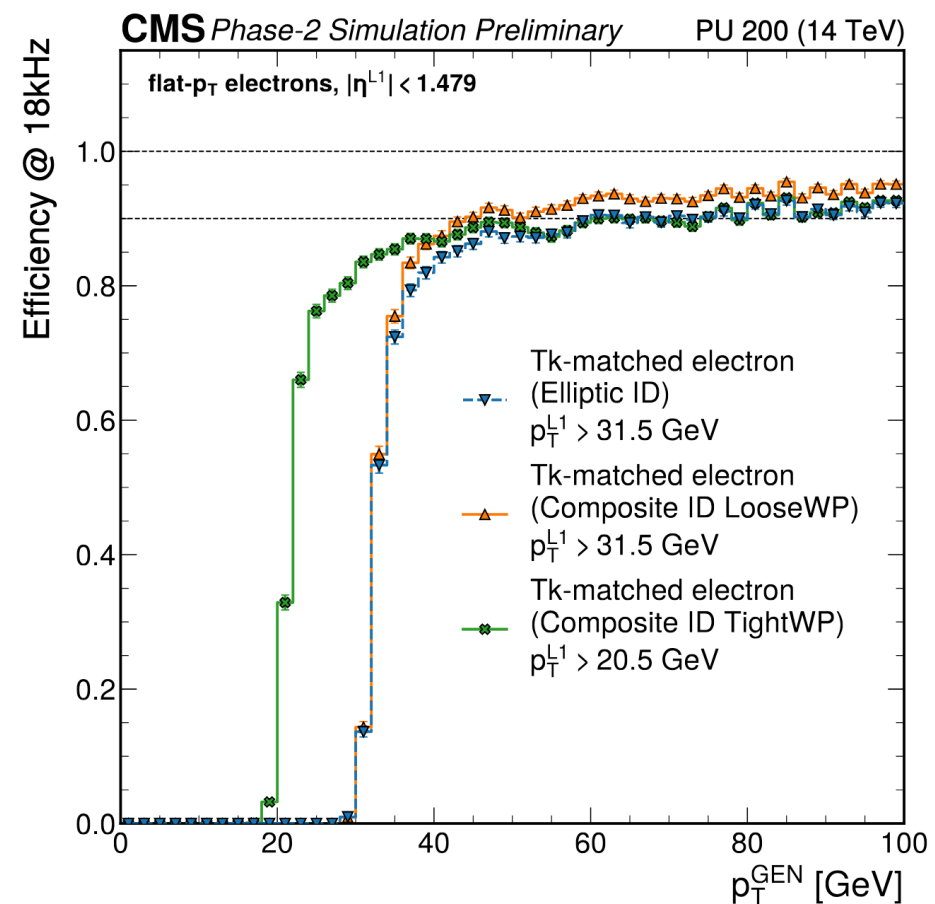
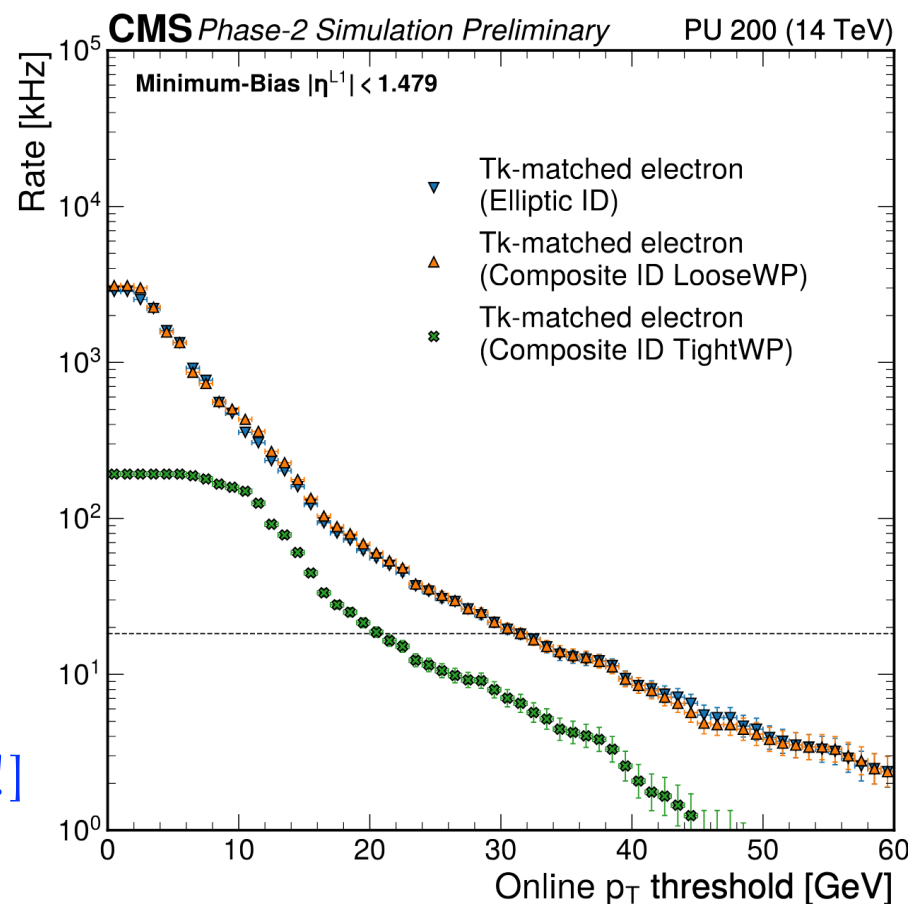
- **2-layers 2D CNN for ID and calibration** with 2D images of seeded calorimeter clusters
 - for the HGCal endcap additional inputs of 3D cluster shape included
- Quantization and pruning applied to achieve **55.6 ns latency @ 360 MHz and $< 1\%$ DSPs** on VU13P AMD chip for a single instance of the NN



A few applications at the LHC

Electron identification

- PF electrons will be reconstructed by linking a track with a calorimeter cluster
- Baseline kinematic approach used distance and p_T compatibility to make a link
- **New BDT approach** combines calorimeter cluster shape variables, track qualities, and track-matching features
- Improved electron reconstruction efficiency at **27.8 ns latency @ 180 MHz and $< 1\%$ DSPs** on VU13P AMD chip



[DP Note soon!]

The CMS L1 Scouting system

- **L1T Data Scouting:** acquire and analyse the L1 Trigger information for all events
- Look for physics signatures identifiable with **just coarse L1 information** but that would evade the L1T → HLT → Offline chain, e.g.:
 - too large “irreducible” backgrounds, e.g. narrow resonances of low mass
 - complex signatures exceeding the computing capabilities of the L1 system
 - signal identification requires time-correlation across several BXs, e.g. slow or long-lived BSM
- FPGA-equipped boards that receive L1 data via optical links and transfer it to PCs and the software world via TCP/IP or PCI express
- At HL-LHC: can profit from much improved L1T object reconstruction quality
- **However, prohibitive downstream bandwidth and storage** → to store all L1 info at 40 MHz a factor $O(10)$ compression/reduction needed
 - opportunity to explore AI methods for data reduction or compression, e.g. through SSL

