Artificial Intelligence Accelerated Discoveries: are we there yet?

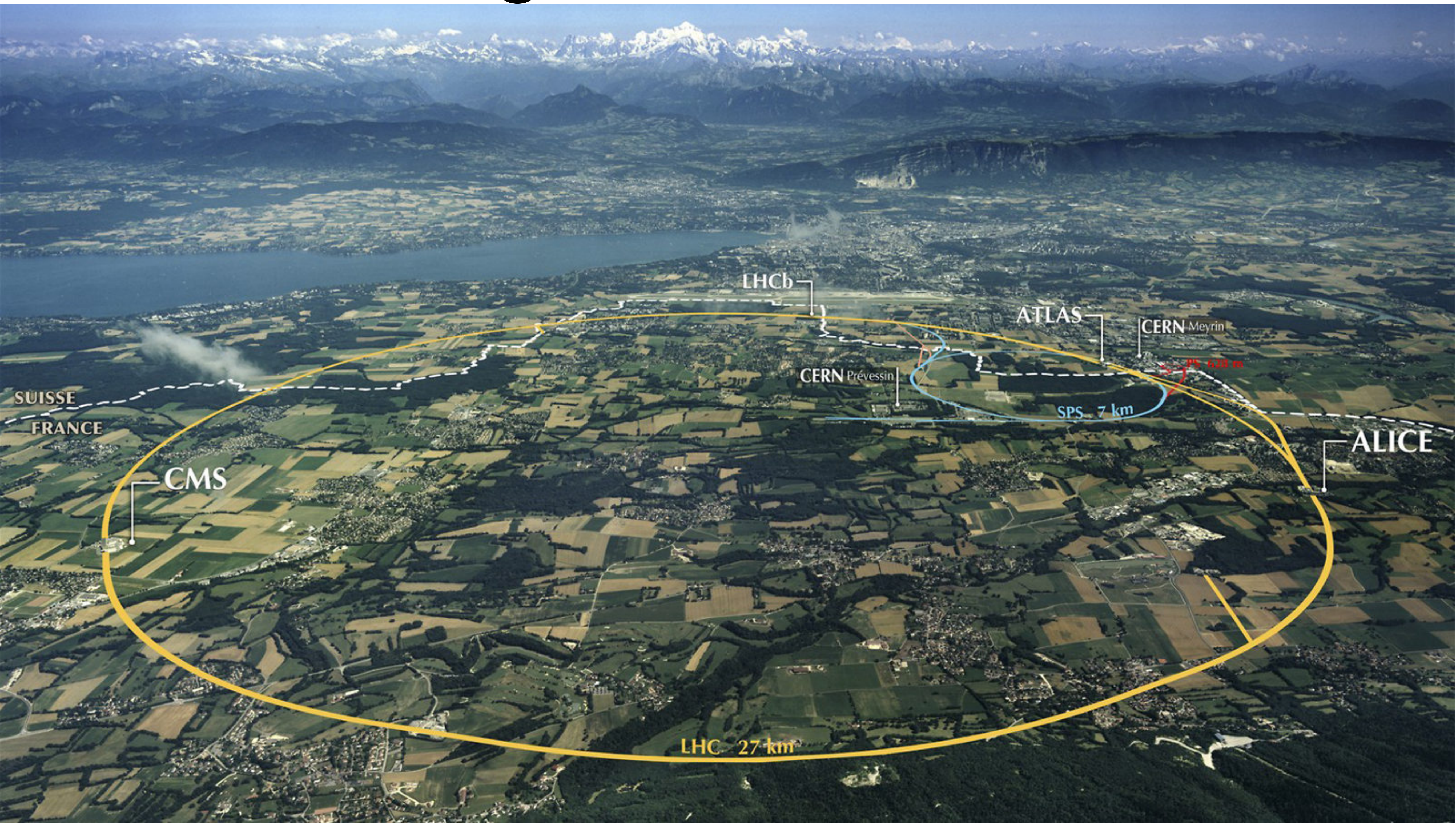
Mia Liu
Purdue University

Seminar@BNL, Nov 16. 2023

CMS

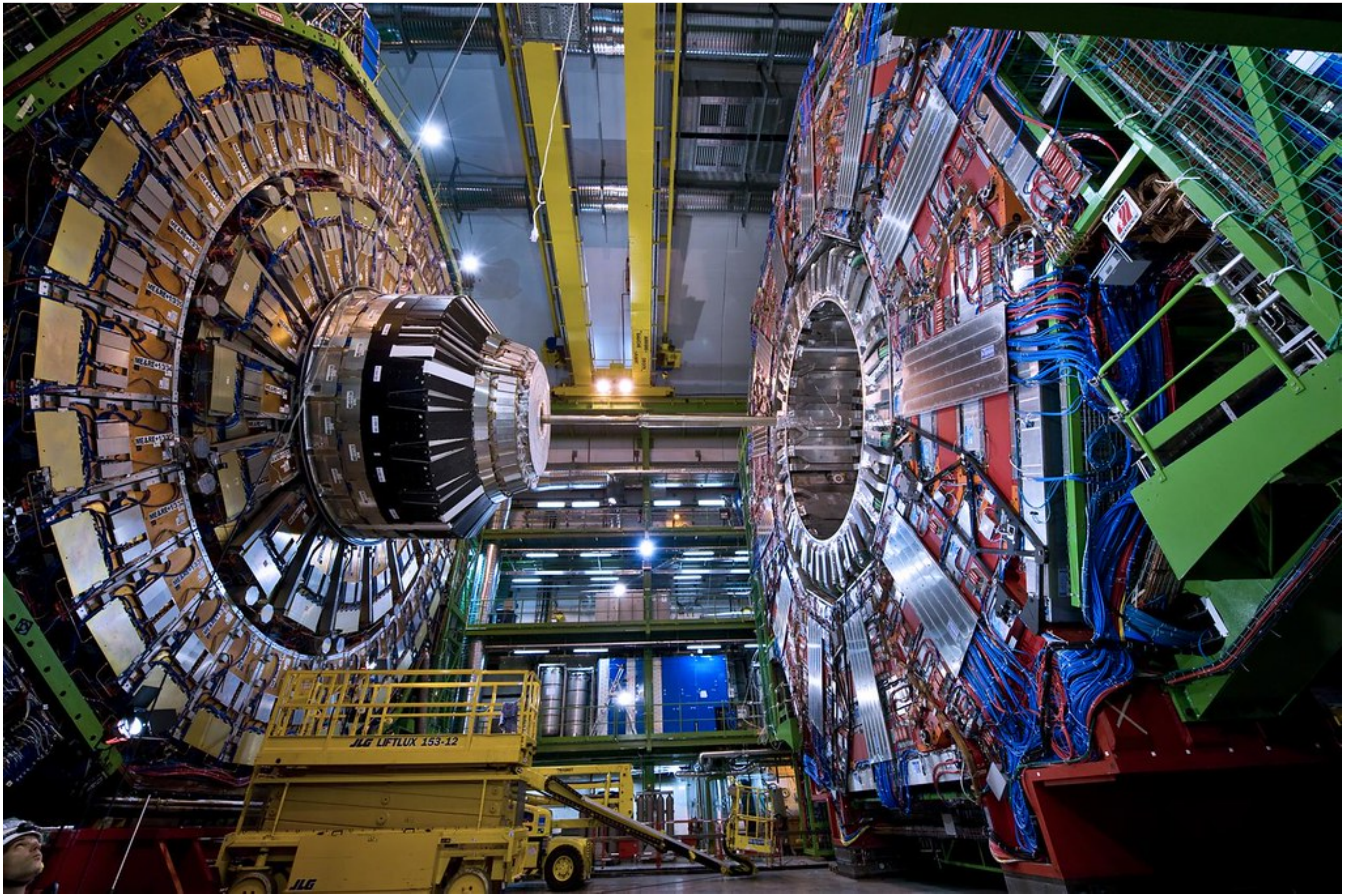
a3d3 Accelerated AI Algorithms for Data-Driven Discovery

The Large Hadron Collider



https://x.com/JannaLevin/status/1512067673311506432?s=20

The CMS detector



Artificial Intelligence *Accelerated* Discoveries

- Advanced data analysis: discoveries not possible with existing datasets
- Fast: big data processing
- Fast: real-time accelerations

- I will not talk about large language models
- Will focus on examples from the LHC
- Briefly mention connection to other science domains

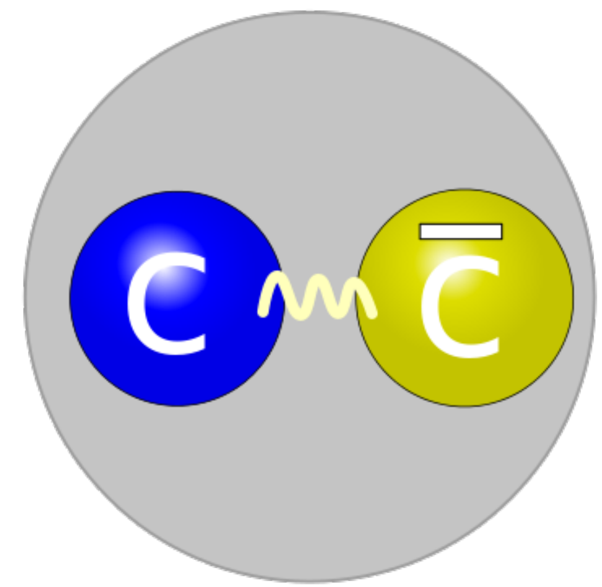
Search For Toponium

What is quarkonium?

Meson formed by heavy quarks.

C-quarks: Charmonium. aka. J/ψ mesons

B-quarks: Bottomonium. Upsilon mesons



Toponium [edit]

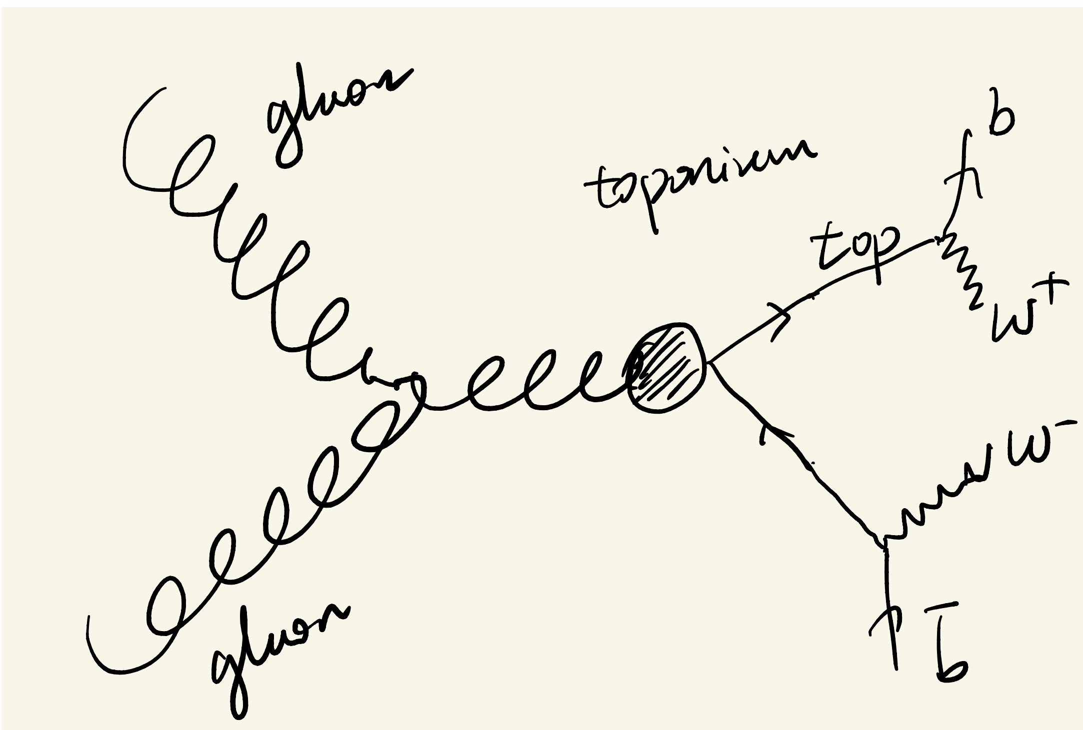
Main article: Theta meson

The theta meson hasn't been and isn't expected to be observed in nature, as top quarks decay too fast to form mesons in nature (and be detected).

This section **needs expansion**. You can help by adding to it. *(April 2017)*

Wikipedia

Toponium: Bound states of top quark pair



Top quark: heaviest fundamental particle observed

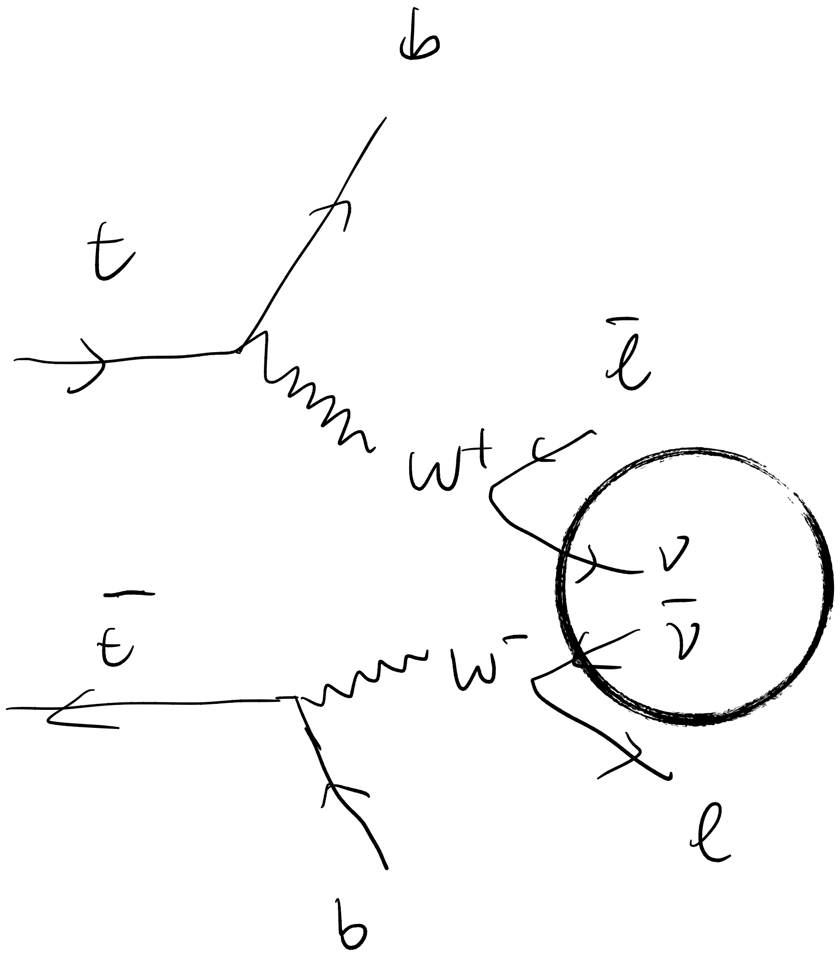
$$\underbrace{\frac{1}{m_t}}_{\substack{\text{production} \\ 10^{-27}\ \text{s}}} < \underbrace{\frac{1}{\Gamma_t}}_{\substack{\text{lifetime} \\ 10^{-25}\ \text{s}}} < \underbrace{\frac{1}{\Lambda_{QCD}}}_{\substack{\text{hadronization} \\ 10^{-24}\ \text{s}}} < \underbrace{\frac{m_t}{\Lambda^2}}_{\substack{\text{spin-flip} \\ 10^{-21}\ \text{s}}}$$

Toponium: bound state formed by top quarks that are relatively long lived

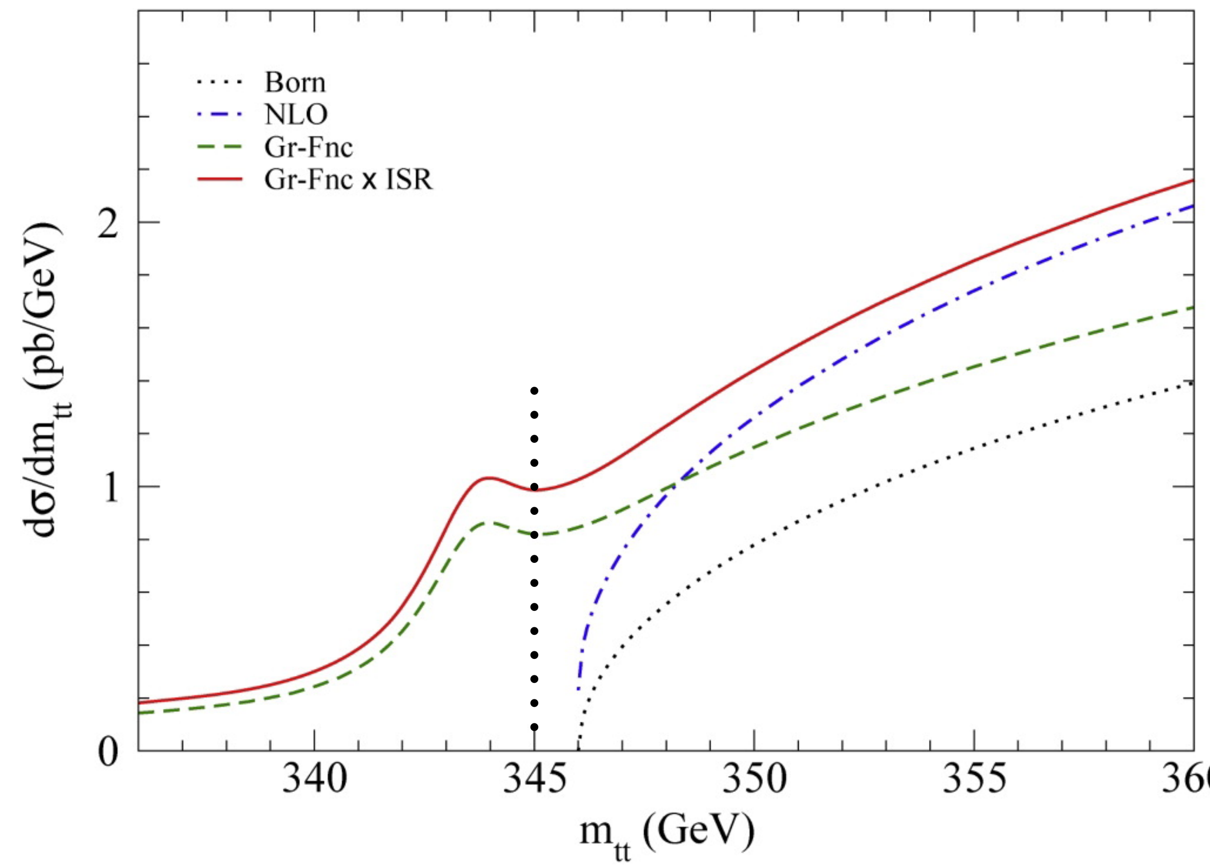
0.7% of the top pairs would form spin singlet toponium η_t

More than 100 millions of top pairs produced in LHC Run 2

Search for Toponium with Dilepton Events



Missing
transverse
energy

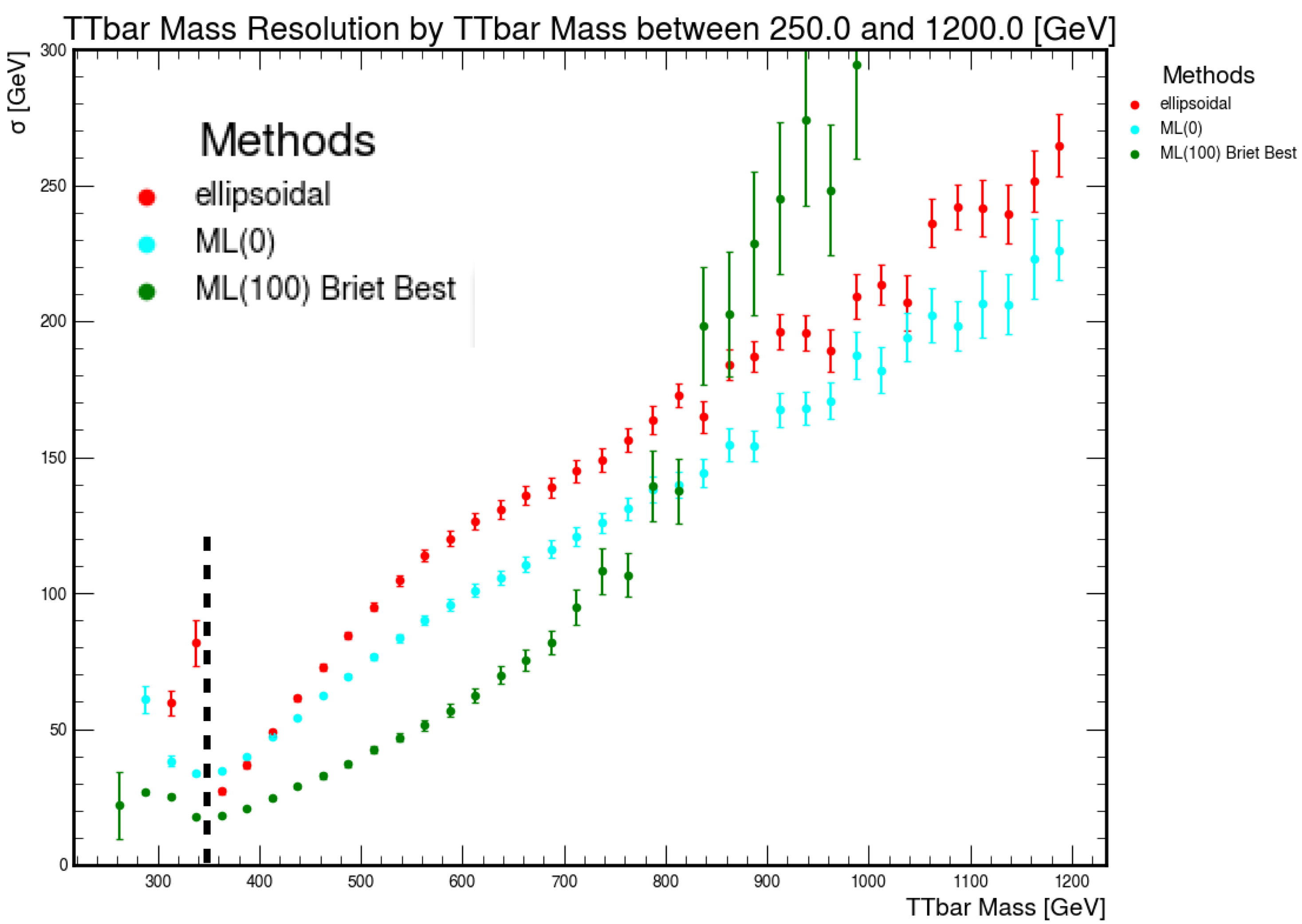


$t\bar{t} \to 2l$: two leptons carry spin correlation information of the two top quarks, can be used for toponium and ttbar separation

Many methods have been studied in order to analyze Tevatron/LHC 2l ttbar events. [0603011.pdf, PhysRevLett.80.2063]

Toponium reconstruction: need good resolution near $M_{t\bar{t}}$ threshold.

Top quark reconstruction: non-ML Method



Analytical solutions with "ellipse" method

Neutrinos

$$
\begin{aligned}
\not{E}_x &= p_{\nu_x} + p_{\bar{\nu}_x}, \\
\not{E}_y &= p_{\nu_y} + p_{\bar{\nu}_y}, \\
E_\nu^2 &= m_\nu^2 + p_{\nu_x}^2 + p_{\nu_y}^2 + p_{\nu_z}^2, \\
E_{\bar{\nu}}^2 &= m_{\bar{\nu}}^2 + p_{\bar{\nu}_x}^2 + p_{\bar{\nu}_y}^2 + p_{\bar{\nu}_z}^2,
\end{aligned}
$$

W

$$
\begin{aligned}
m_{W^+}^2 &= (E_{\ell^+} + E_\nu)^2 - (p_{\ell_x^+} + p_{\nu_x})^2, \\
&\quad -(p_{\ell_y^+} + p_{\nu_y})^2 - (p_{\ell_z^+} + p_{\nu_z})^2, \\
m_{W^-}^2 &= (E_{\ell^-} + E_{\bar{\nu}})^2 - (p_{\ell_x^-} + p_{\bar{\nu}_x})^2, \\
&\quad -(p_{\ell_y^-} + p_{\bar{\nu}_y})^2 - (p_{\ell_z^-} + p_{\bar{\nu}_z})^2,
\end{aligned} \tag{1}
$$

top

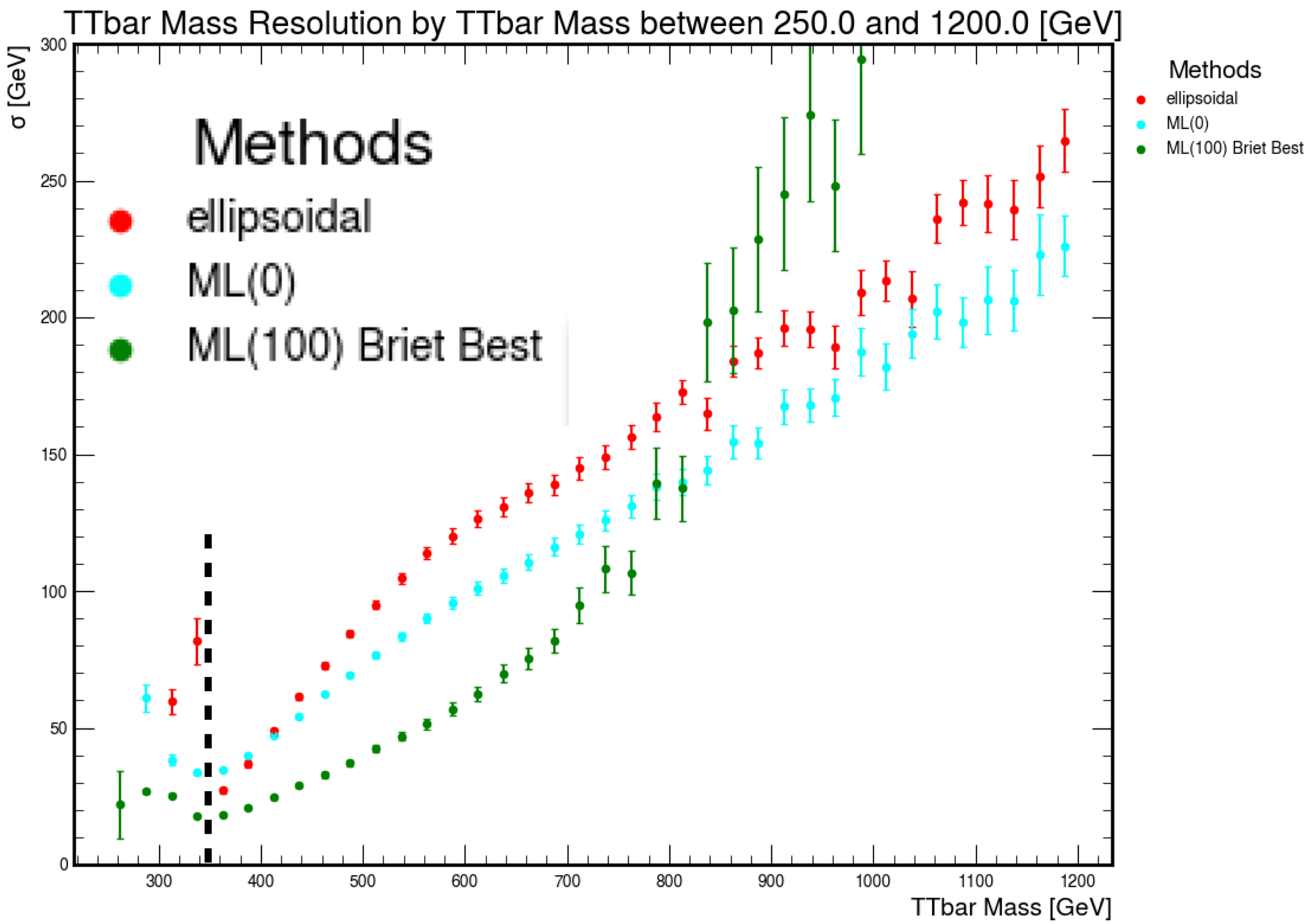
$$
\begin{aligned}
m_t^2 &= (E_b + E_{\ell^+} + E_\nu)^2 - (p_{b_x} + p_{\ell_x^+} + p_{\nu_x})^2, \\
&\quad -(p_{b_y} + p_{\ell_y^+} + p_{\nu_y})^2 - (p_{b_z} + p_{\ell_z^+} + p_{\nu_z})^2, \\
m_{\bar{t}}^2 &= (E_{\bar{b}} + E_{\ell^-} + E_{\bar{\nu}})^2 - (p_{\bar{b}_x} + p_{\ell_x^-} + p_{\bar{\nu}_x})^2, \\
&\quad -(p_{\bar{b}_y} + p_{\ell_y^-} + p_{\bar{\nu}_y})^2 - (p_{\bar{b}_z} + p_{\ell_z^-} + p_{\bar{\nu}_z})^2.
\end{aligned}
$$

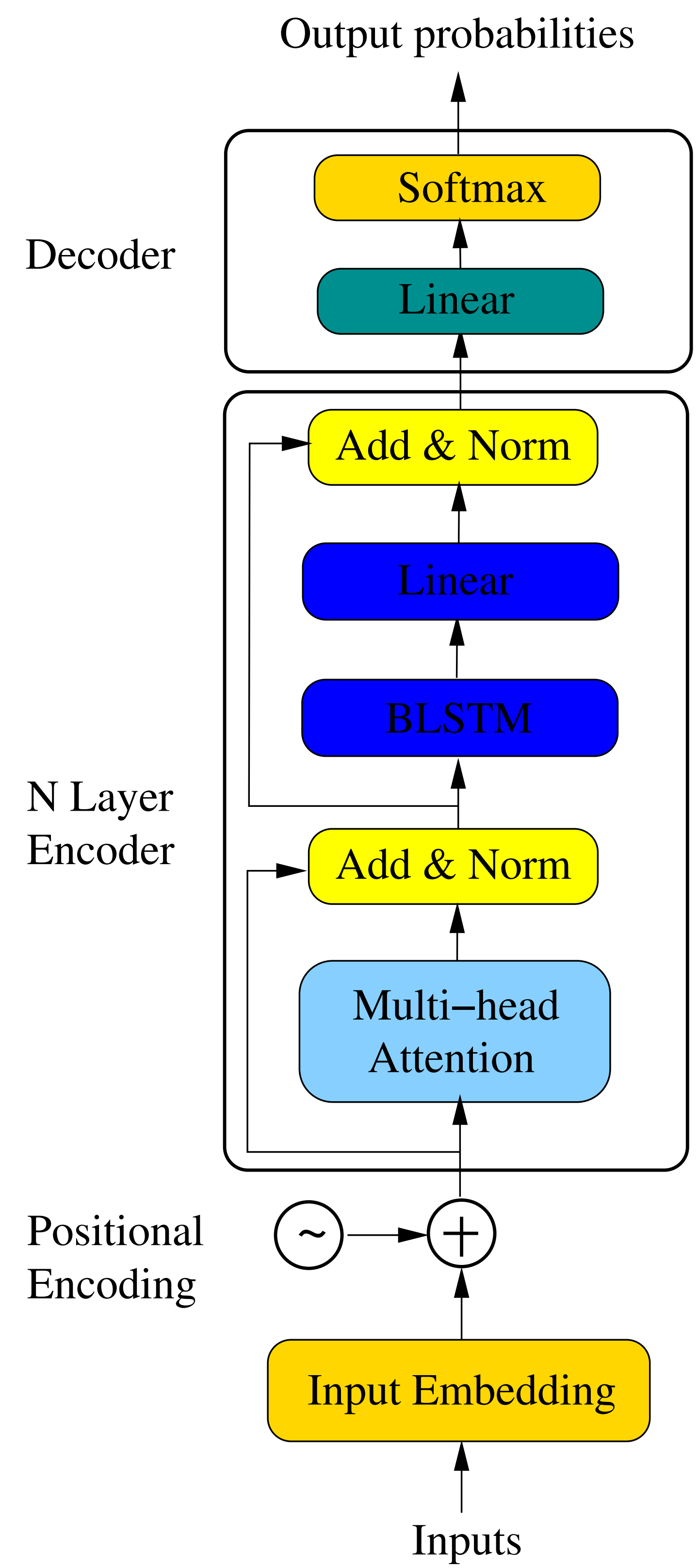
Hard constraints on particle masses, detector resolution and extra jet radiation effects ignored

Poor reconstruction below threshold and for boosted region

Fail to find solutions/Numerical methods, e.g. Neutrino reweighting: computationally expensive

With Transformers





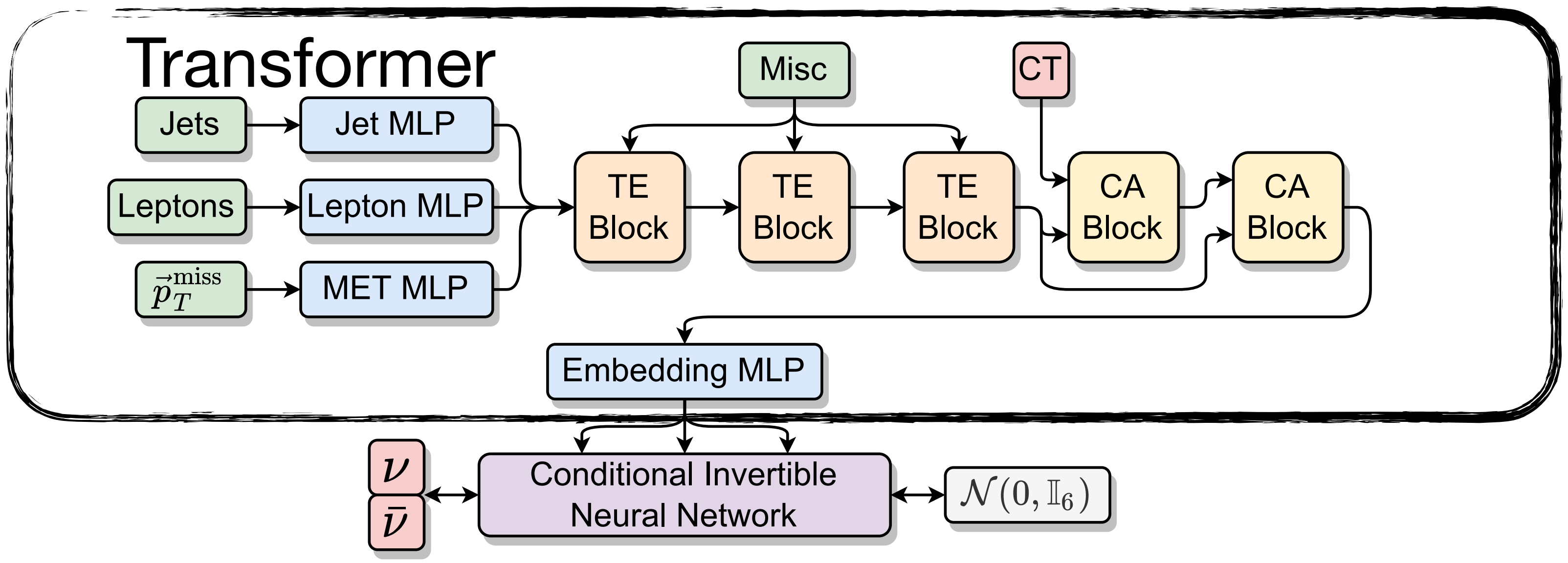
Method based on transformer-LSTM model

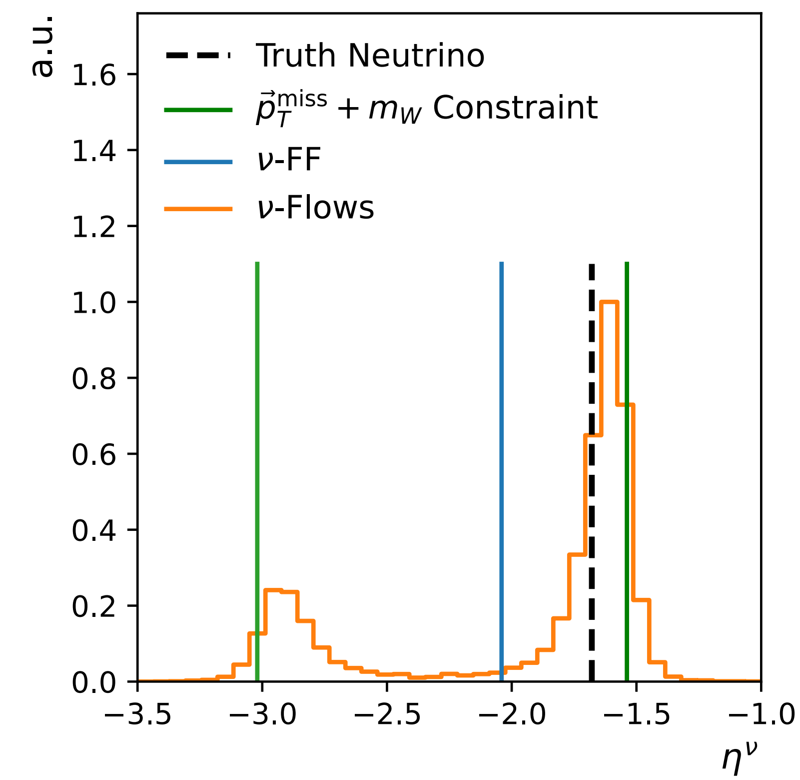
Four momentum of leptons, jets, missing transverse momentum as sequence input

Regress to neutrino true momentum:

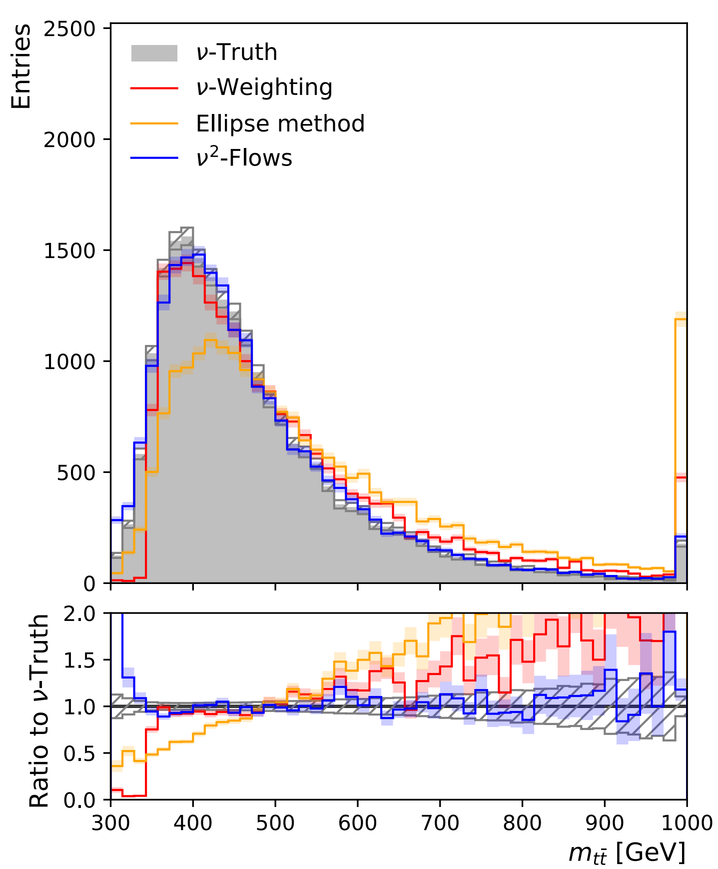
Particle mass, kinematic constraints included in the Loss function

Make it Probabilistic: Conditional Normalizing Flow





https://arxiv.org/pdf/2307.02405.pdf



Instead of regressing to true neutrino momentum:

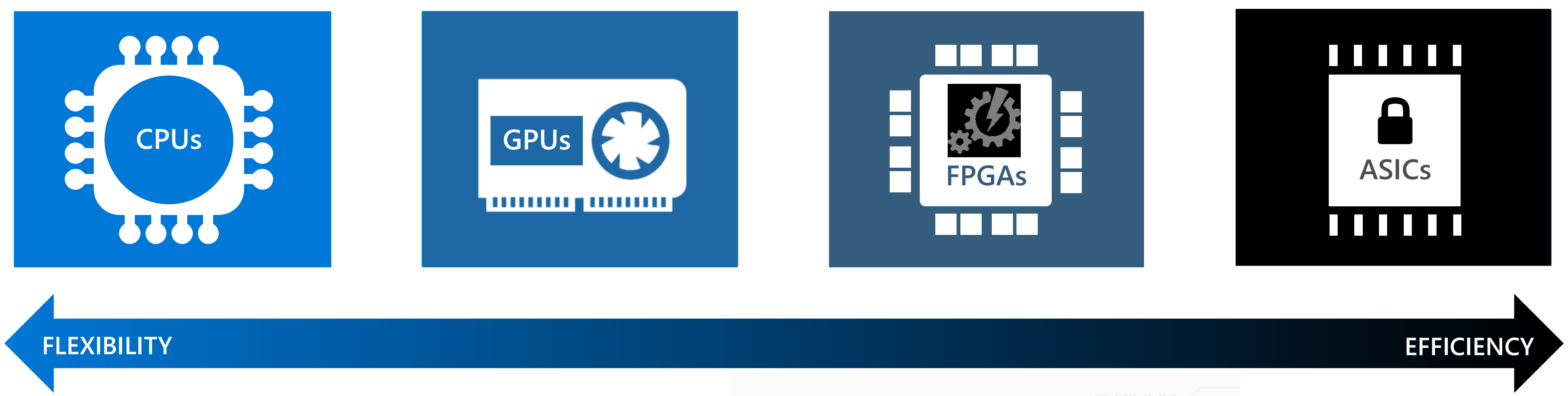
Conditional probability distribution of neutrino momentum given event content; likelihood for final statistical interpretation;

Improvement in threshold and boosted regions, also faster (computing cost shifted to training)

Promising for toponium search and entanglement, bells inequality studies & BSM search

Everyone is doing some ML now, how to process data efficiently?

Heterogeneous Computing for ML/AI



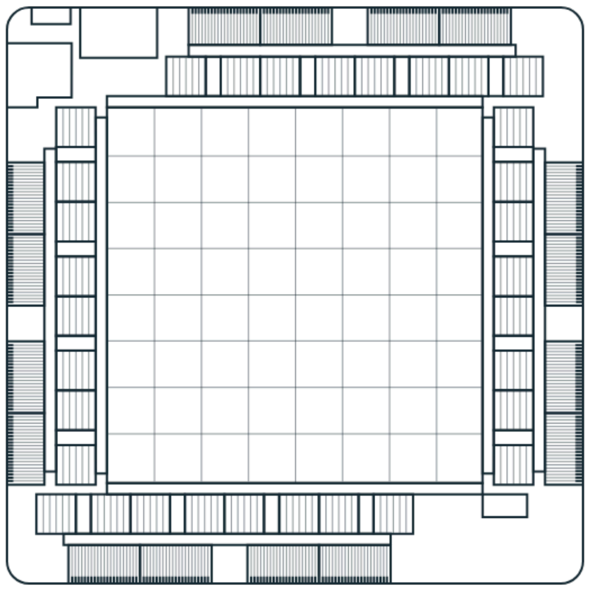
Advances driven by
big data explosion
& machine learning

A 5 year old slide, message remains…

Discontinued: October 18, 2022

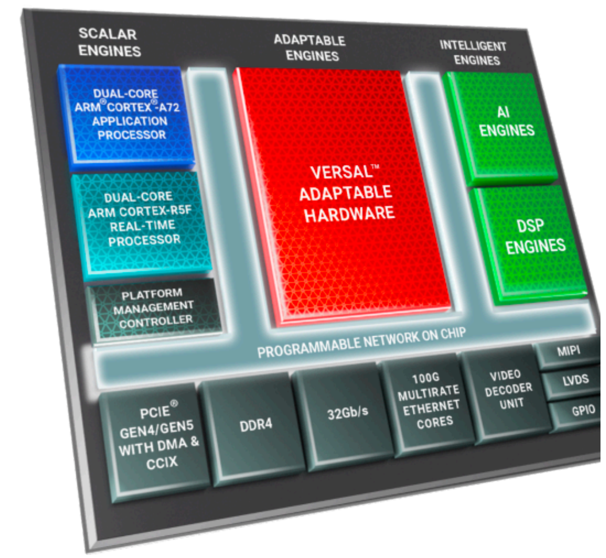
‘AI chips’ in 2023

Meta



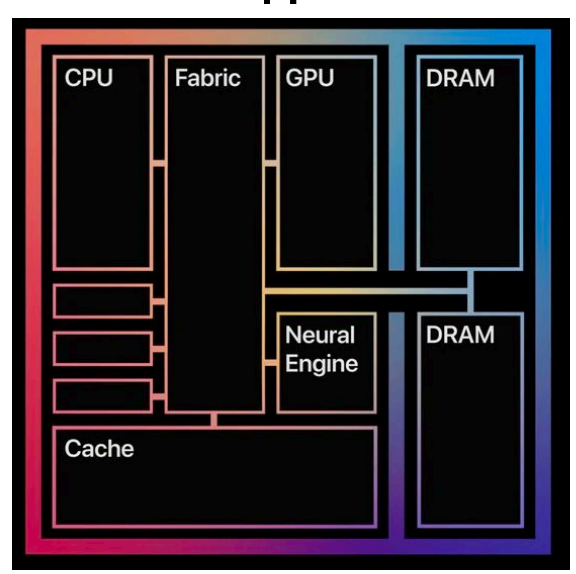
Groq

AMD / Xilinx

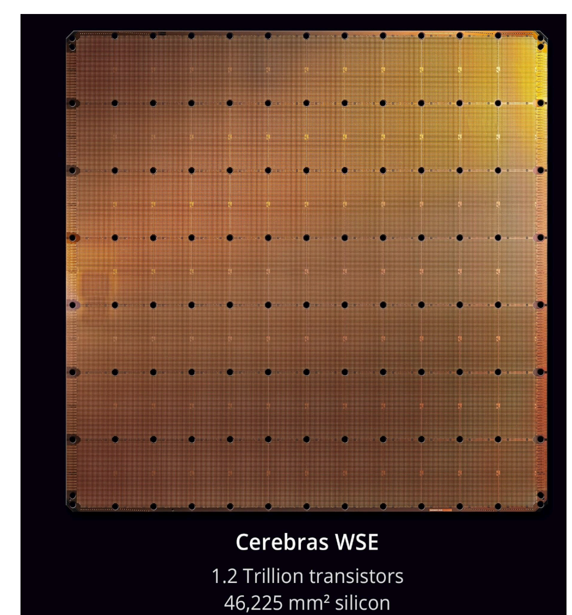


Graphcore

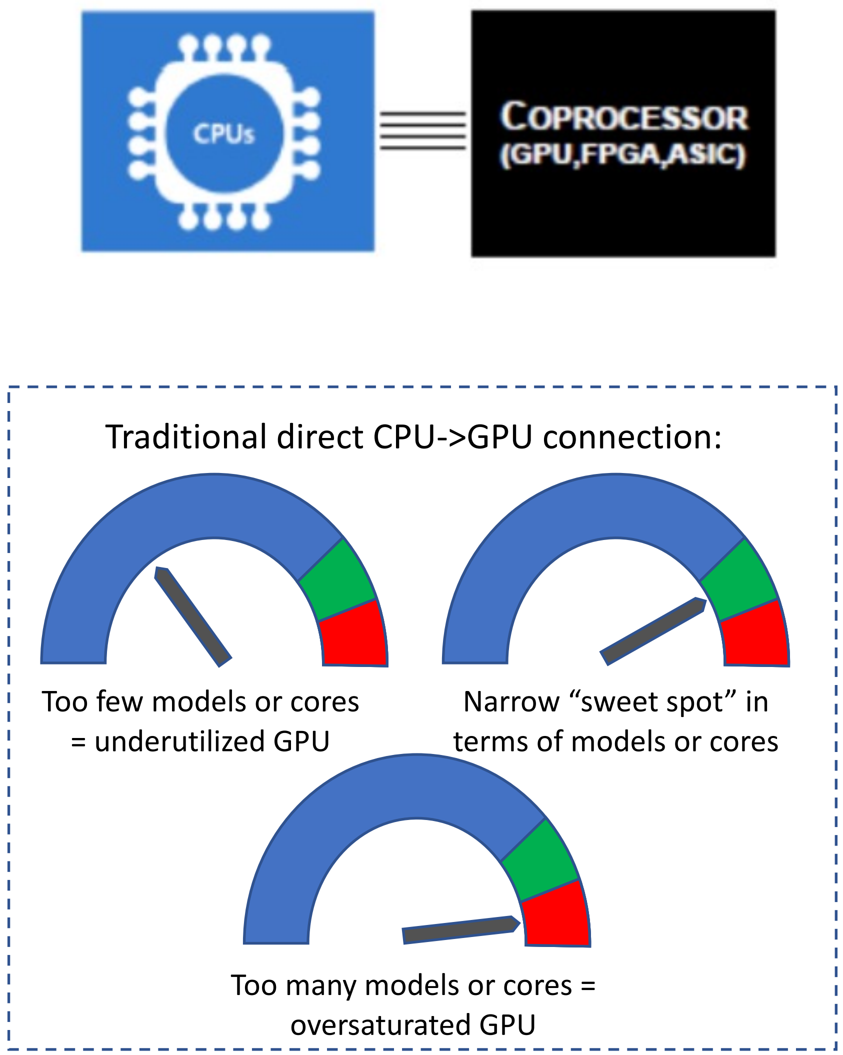
Apple



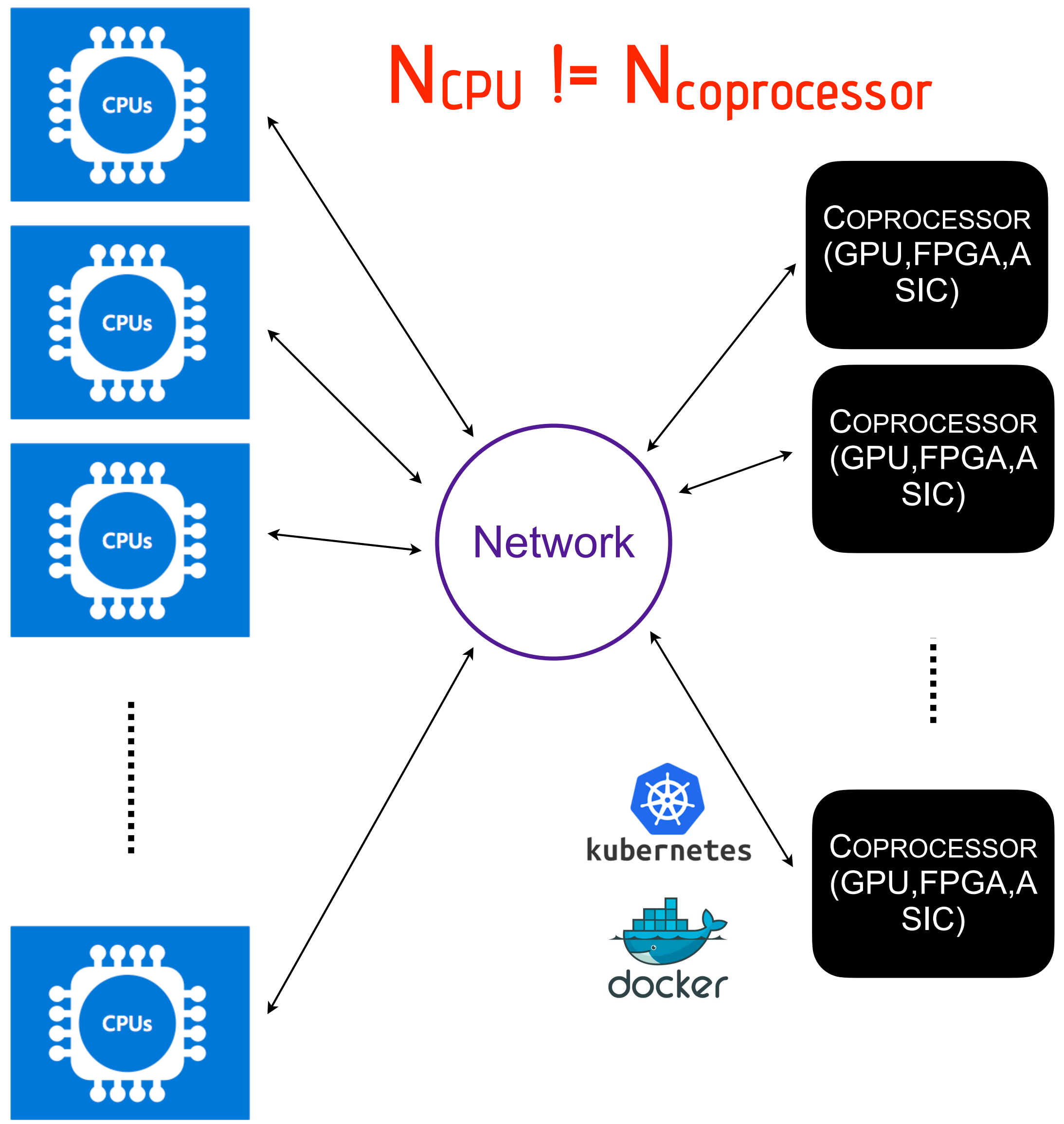
Cerebras



Optimal Acceleration hardware usage

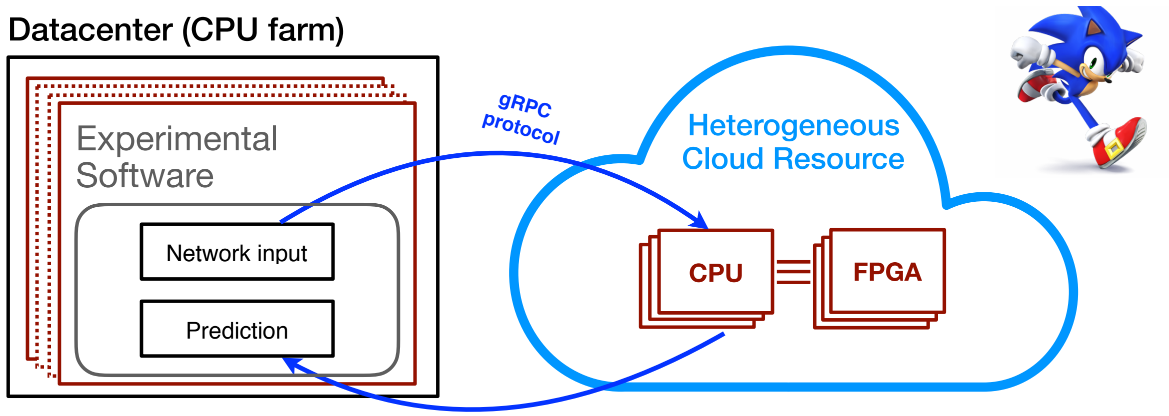


Inflexible & Expensive



Complex, Requires R&D

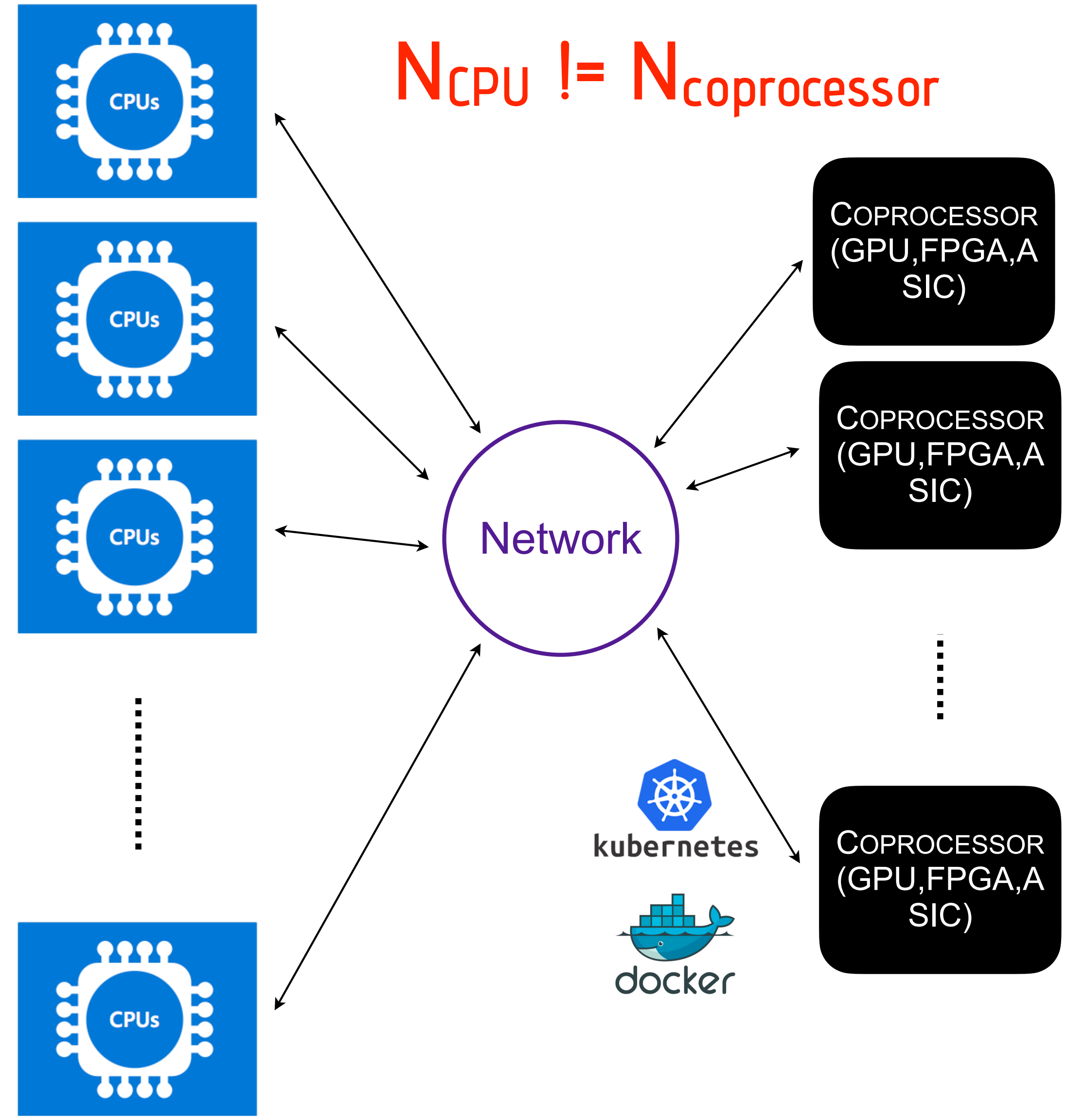
Alternative solution : as-a-service



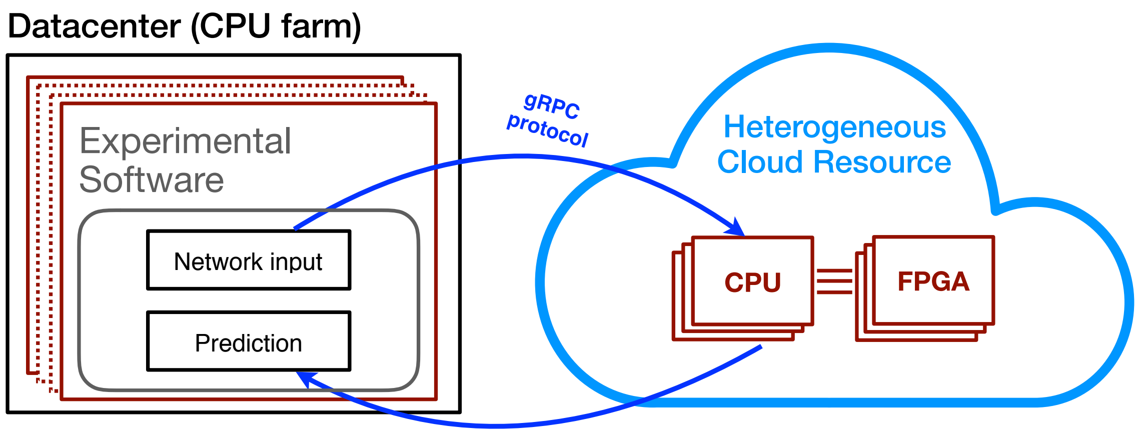
Services for **O**ptimized **N**etwork **I**nference on **C**o-processors. https://arxiv.org/pdf/1904.08986.pdf

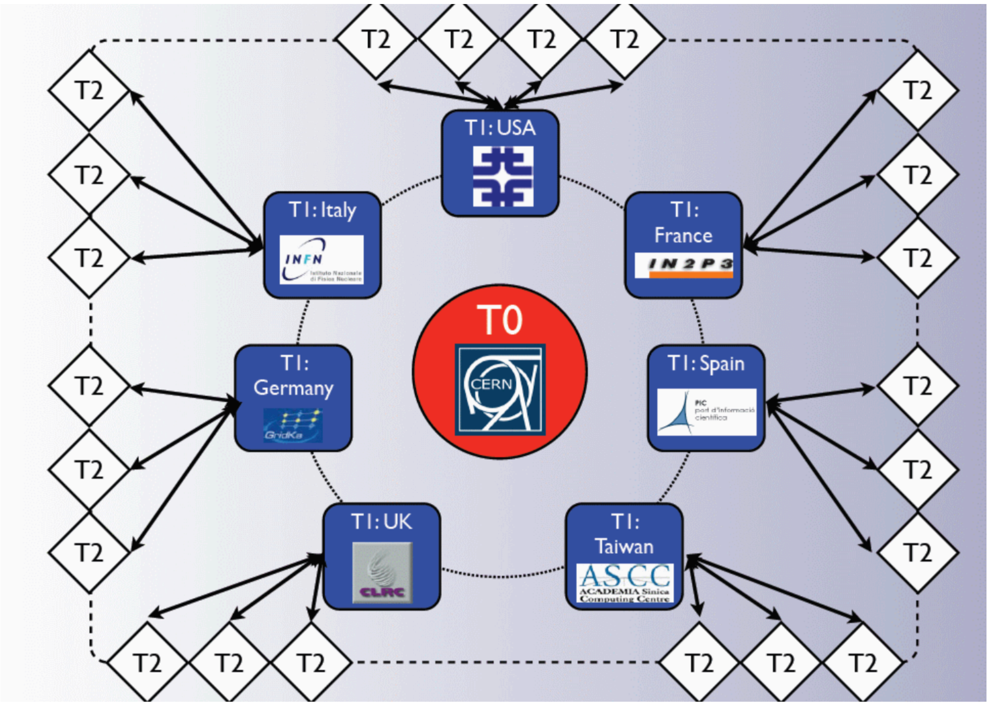
Flexible & Adaptable - right-size the system based on compute needs, maximize e.g. GPU acceleration; task-based optimization;

Scalable & Portable - co-processor disassociated from existing CPU infrastructure; reduce client software dependency on server hardware

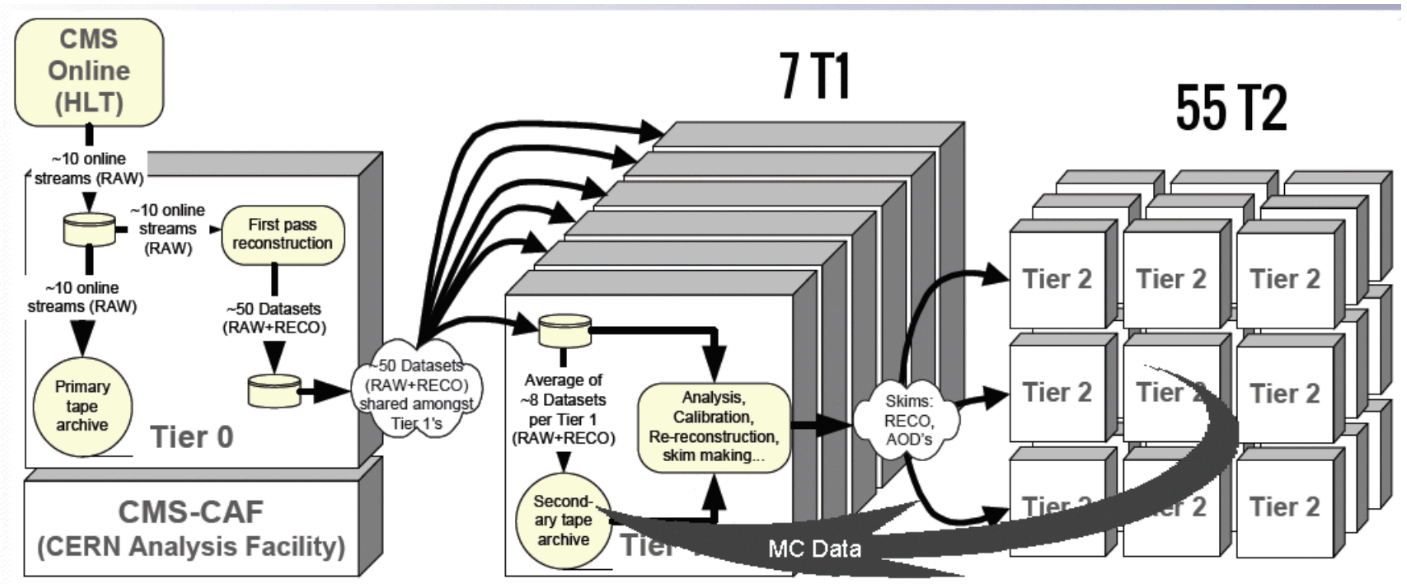


How to deploy SONIC in CMS





NVIDIA Triton Inference Server

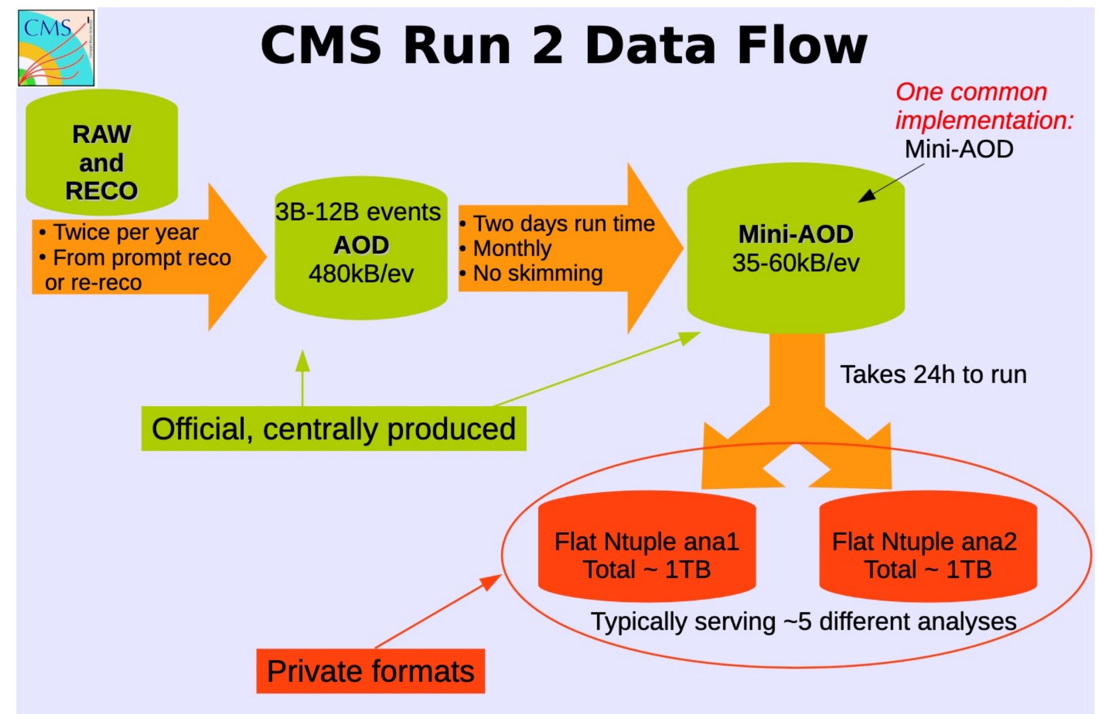


First demonstration in large experiment data processing

MiniAOD step of the CMS data processing: ML inferences consume 10% of the total processing time.

Tests at Purdue CMS Tier-2 data center, GCP and Fermilab.

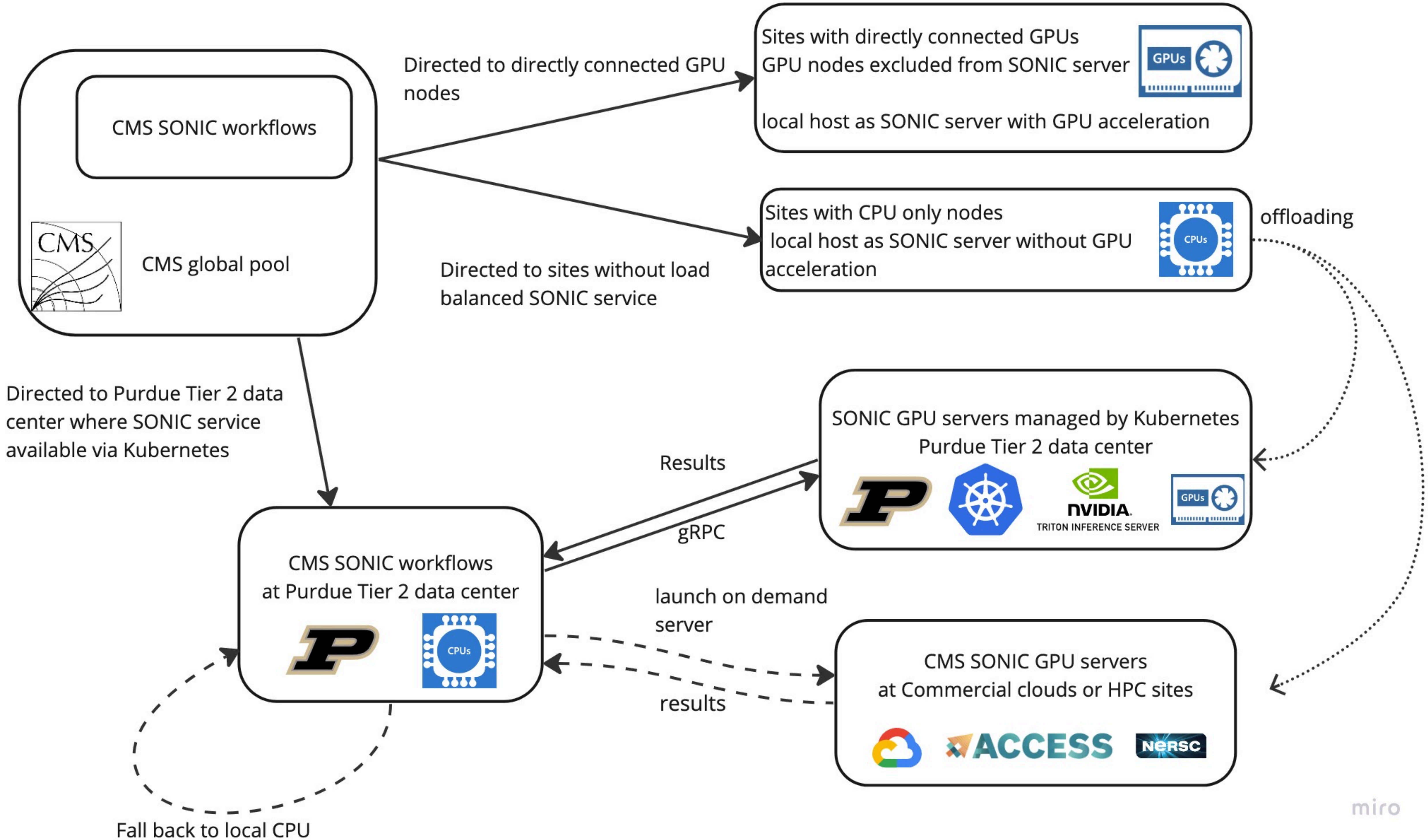
Public results



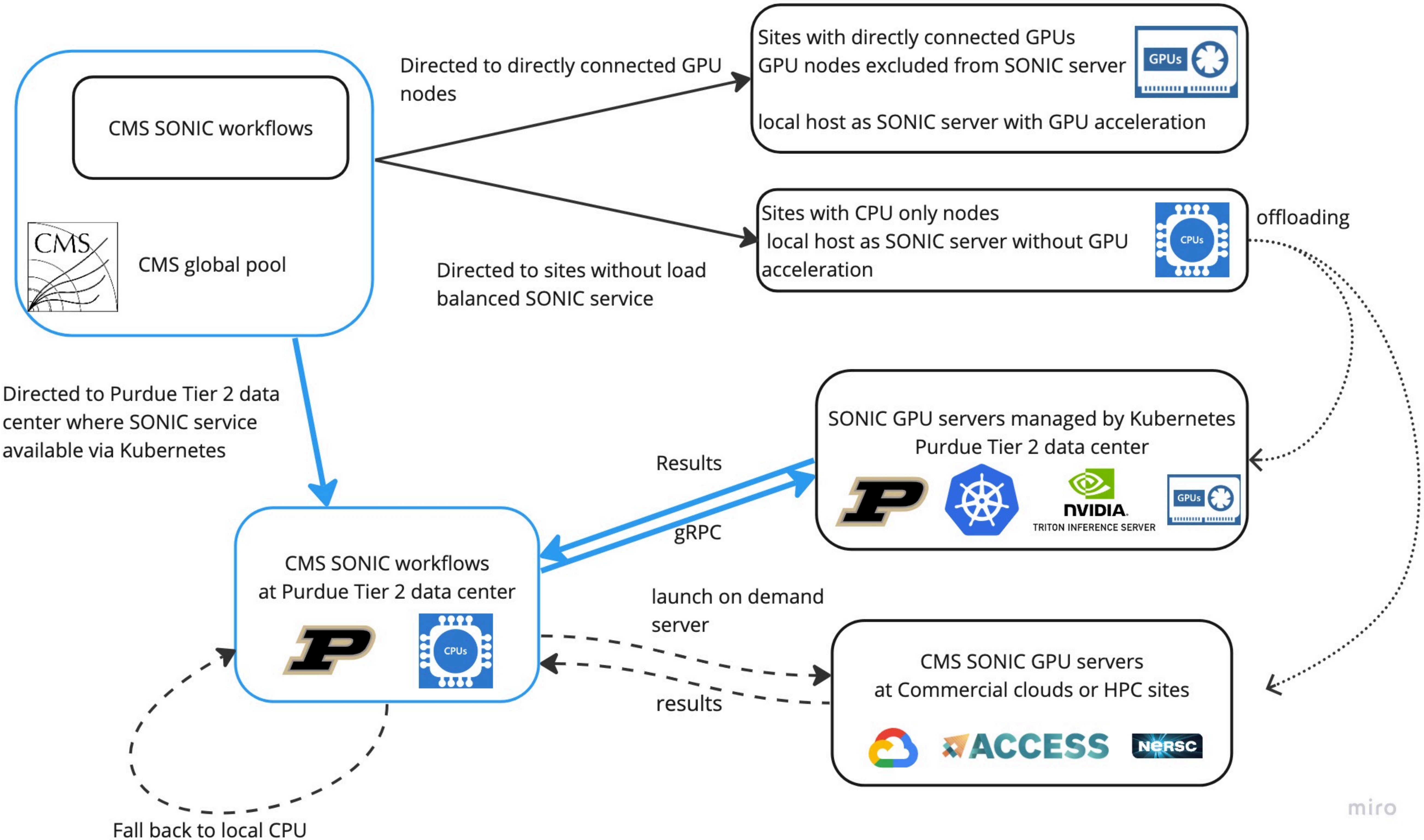
Mini-AOD production typically takes about 0.5 seconds per event on production grid nodes

Algorithm	Time [ms]	Fraction [%]	Input [MB]
PN-AK4	42.4	4.3	0.04
PN-AK8	11.4	1.1	0.003
DeepMET	13.2	1.3	0.33
DeepTau	21.1	2.1	1.18
ParticleNet+DeepMET+DeepTau	88.1	8.8	1.55
Total	993.3	100.0	—

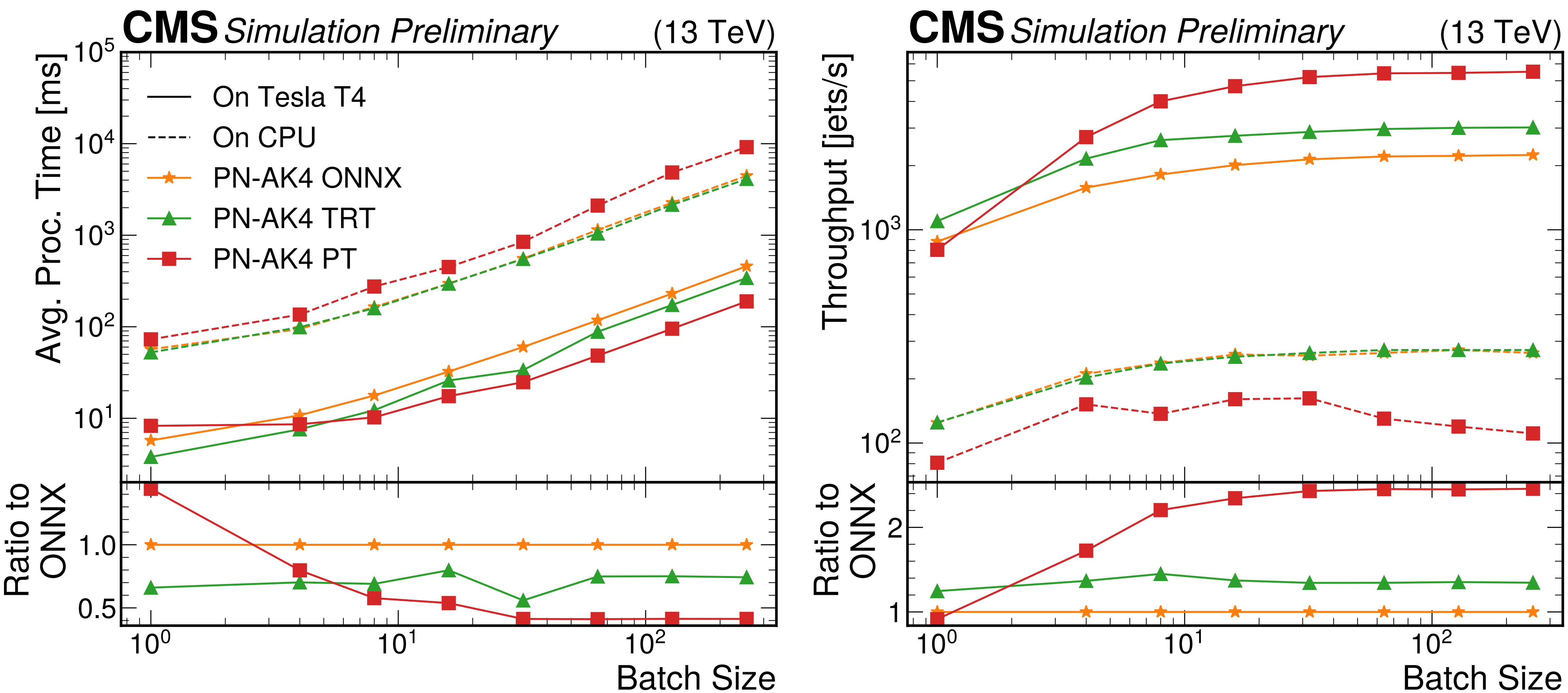
GPU as-a-service for CMS



Load-balanced Triton service available on site

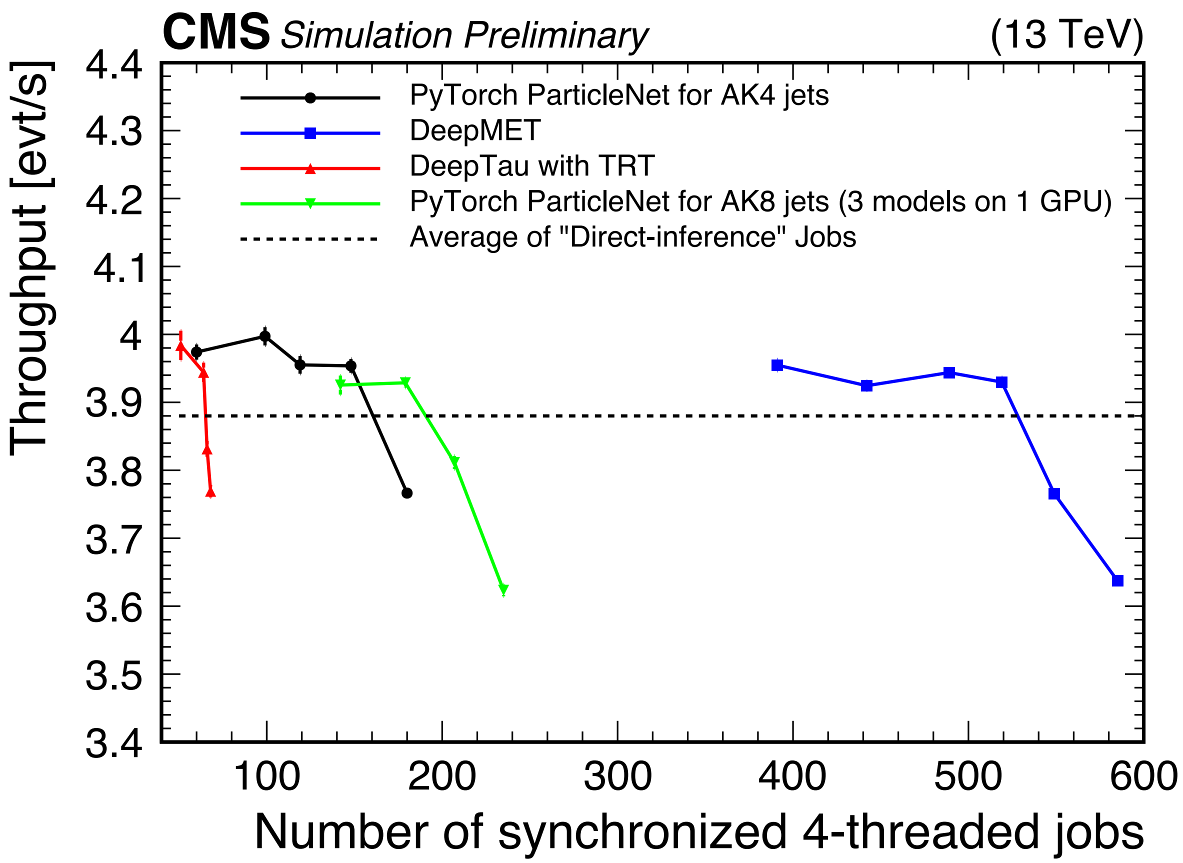


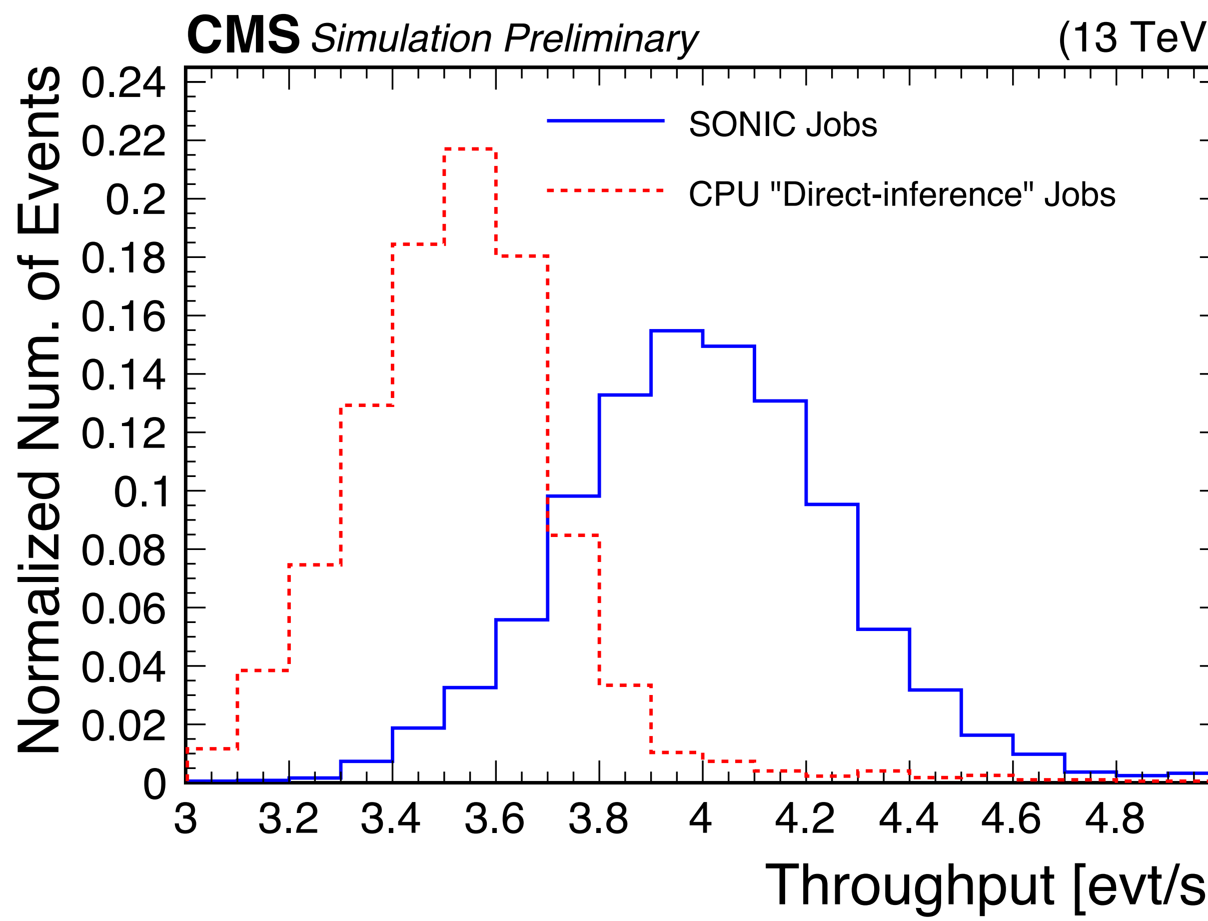
Single Model Optimization



- Asynchronous inference requests
- Triton supports ONNX, TensorFlow, PyTorch, Scikit-Learn, etc. Triton model analyzer tool optimizes server settings. e.g. batch size, dynamic batching window.

CPU/GPU ratio optimization





One ML model offloaded for each test

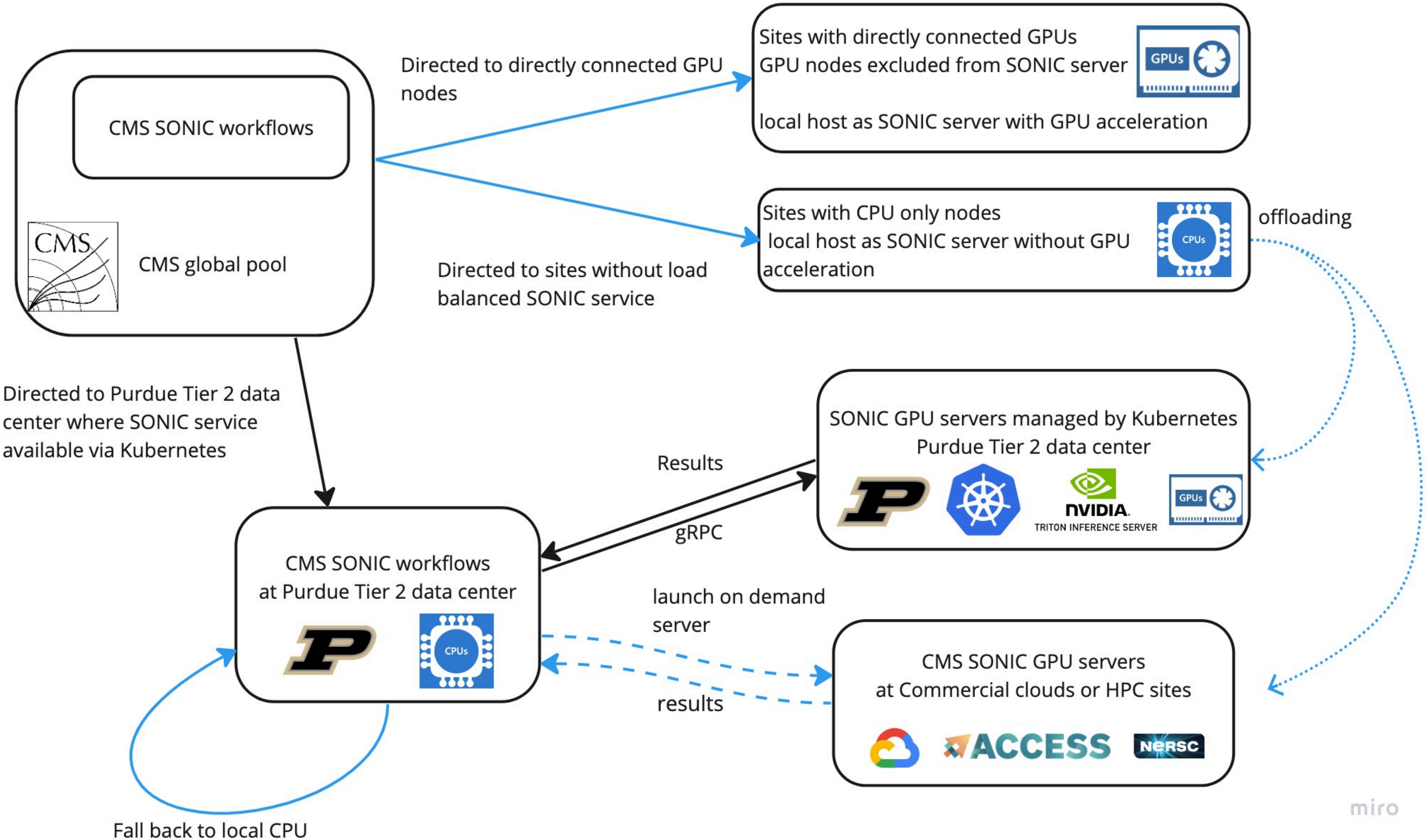
Optimal CPU client jobs / GPU ratio decided by acceleration factor/saturation point

Scale-out test at GCP:

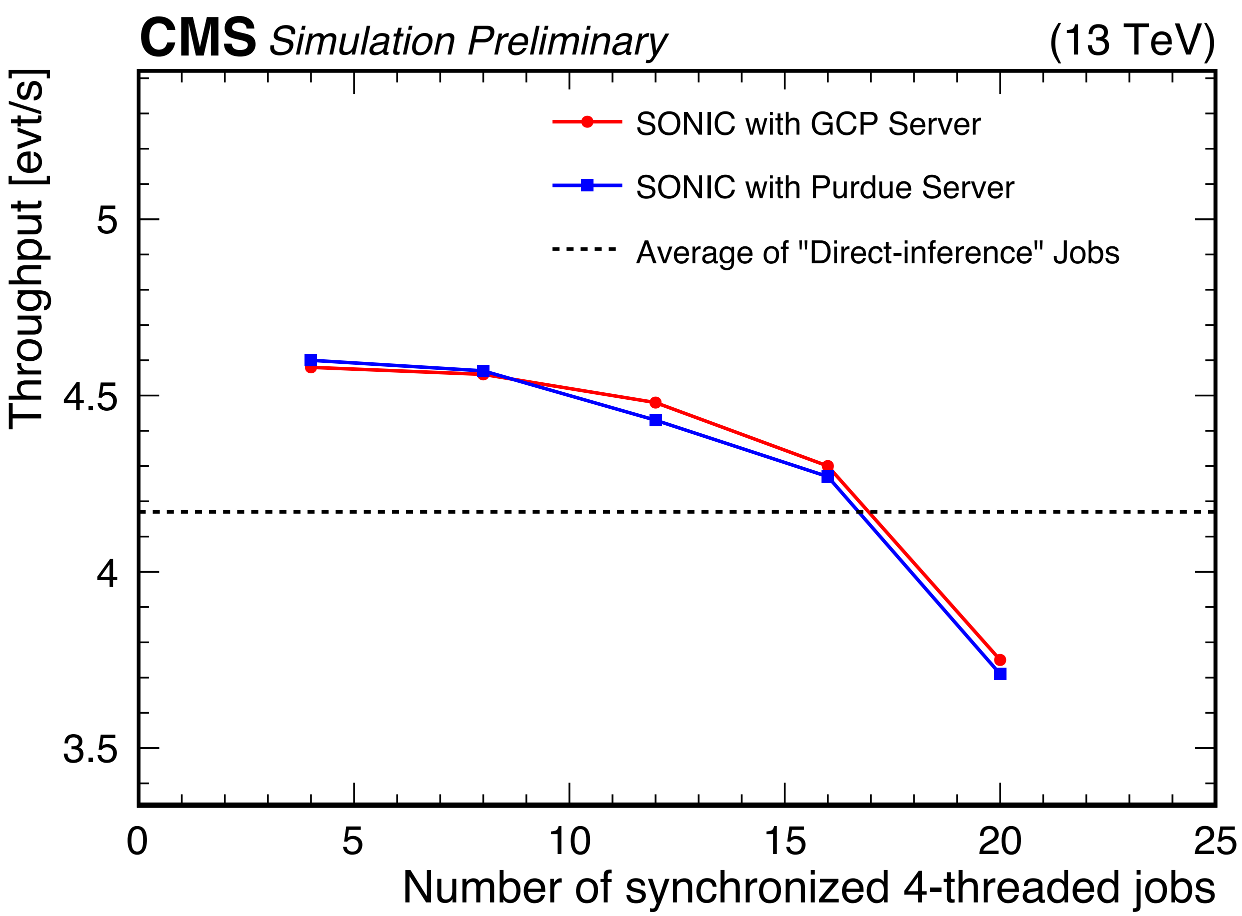
10,000 CPU cores (2,500 4-threaded) client jobs, 100 Tesla T4 GPUs with Kubernetes as load balancer

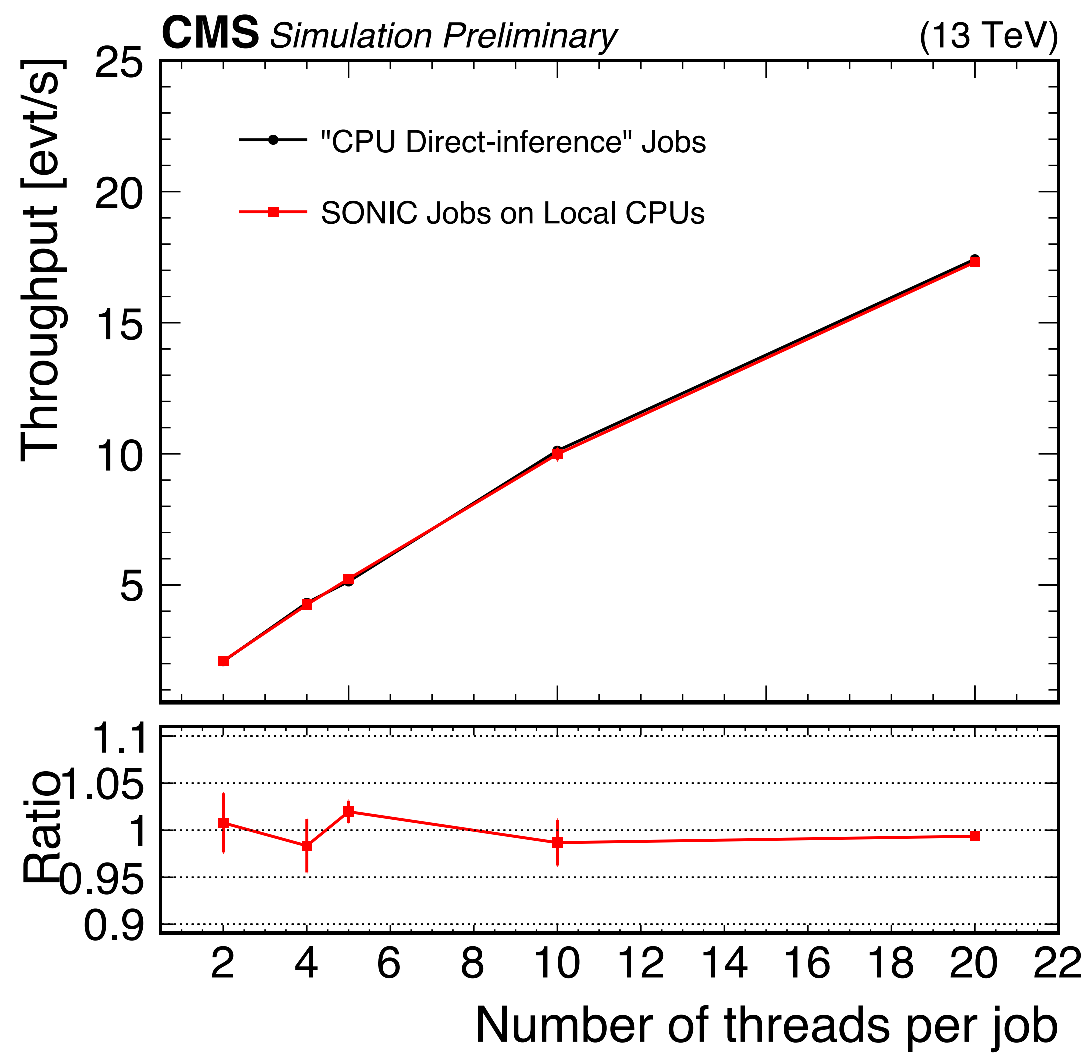
Peak network usage was ~15 GB/s (total bandwidth coming into GPU cluster)

Plan B?


CMS SONIC workflows
CMS global pool
Directed to directly connected GPU nodes
Sites with directly connected GPUs
GPU nodes excluded from SONIC server
local host as SONIC server with GPU acceleration
Sites with CPU only nodes
local host as SONIC server without GPU acceleration
Directed to sites without load balanced SONIC service
offloading
Directed to Purdue Tier 2 data center where SONIC service available via Kubernetes
SONIC GPU servers managed by Kubernetes
Purdue Tier 2 data center
Results
gRPC
CMS SONIC workflows
at Purdue Tier 2 data center
launch on demand server
results
CMS SONIC GPU servers
at Commercial clouds or HPC sites
Fall back to local CPU
miro

Remote server/Fall-back CPU





No added latency observed in remote offloading.

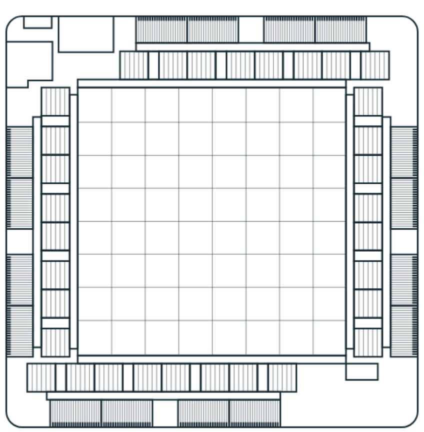
More studies on memory overhead etc, see paper.

More and Next

- SONIC with FPGA co-processors
- SONIC for CMS data analysis
- SONIC for GraphCore IPU
- SONIC to offload customized algorithms in GPU
- Future: SONIC for AMD GPUs

……

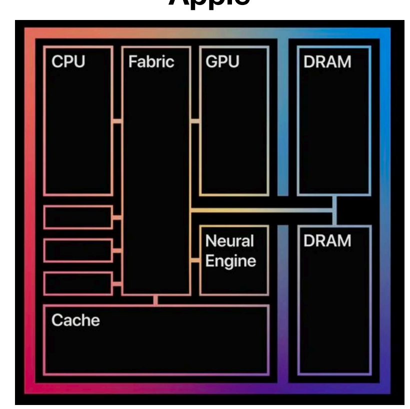
Meta



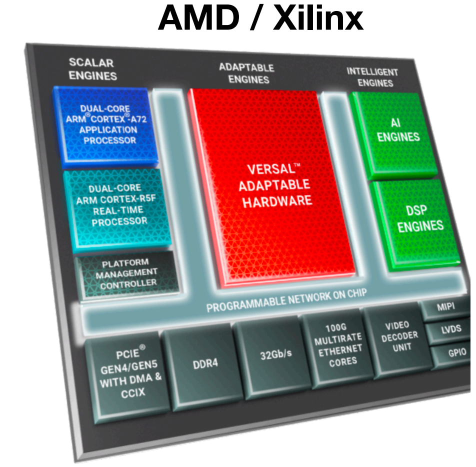
Graphcore

Groq

Apple

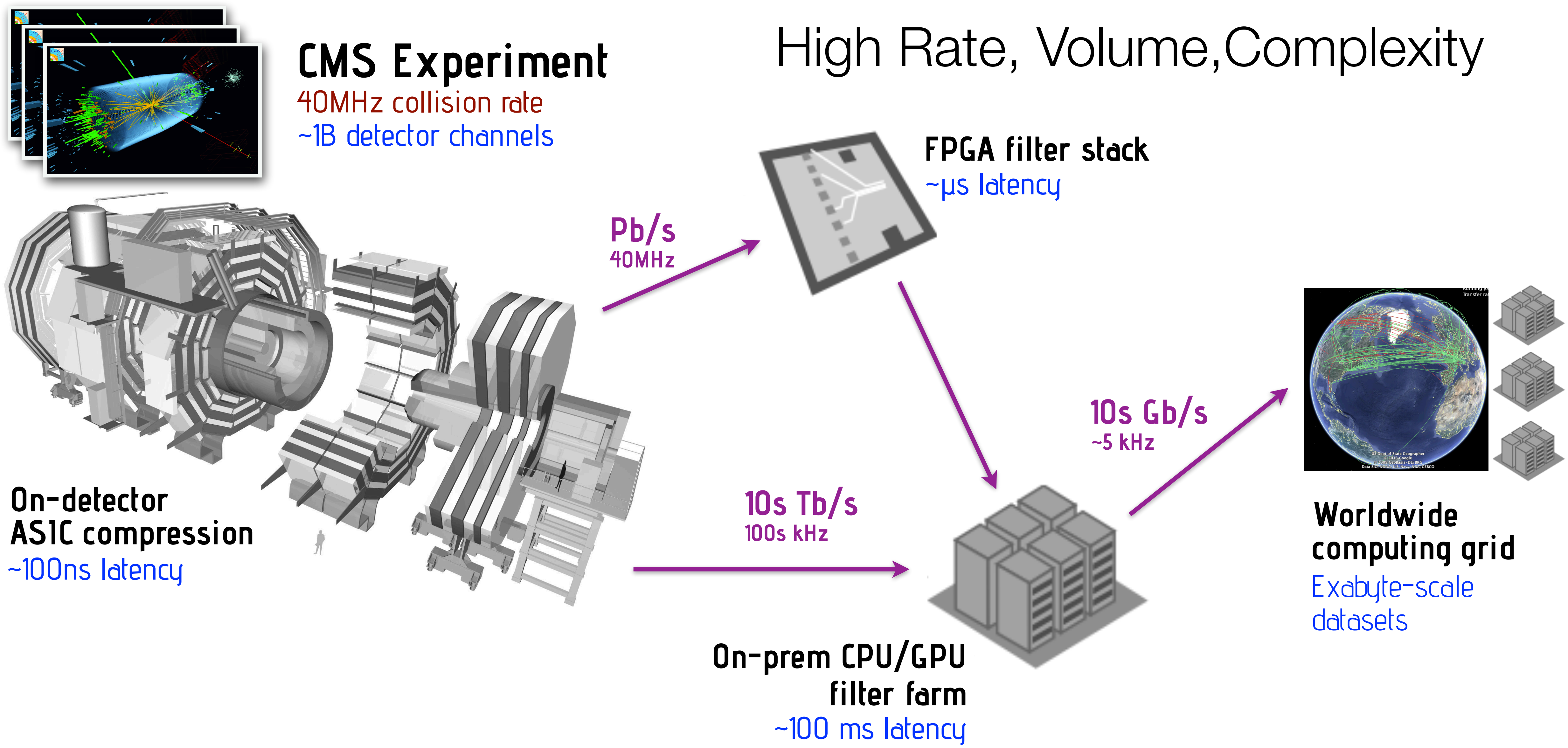


AMD / Xilinx



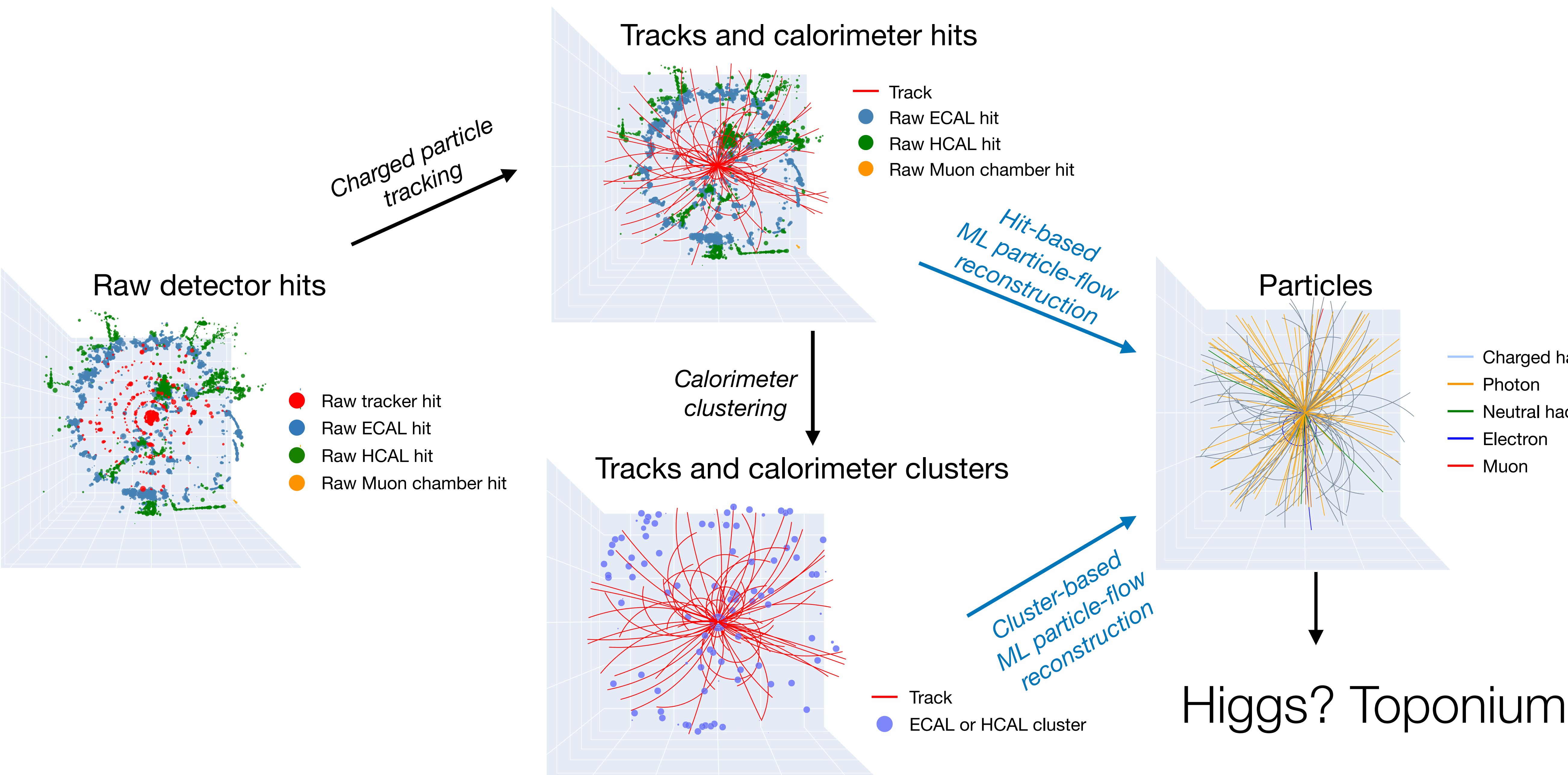
Bringing AI to Data Creation

From Collisions to Discoveries



Science with Big data: Multi-tier Data Processing

From Collisions to Discoveries



Higgs? Toponium?

Dark Matter?

Learning grouping of detector elements

End to End Detection of Exotic patterns?

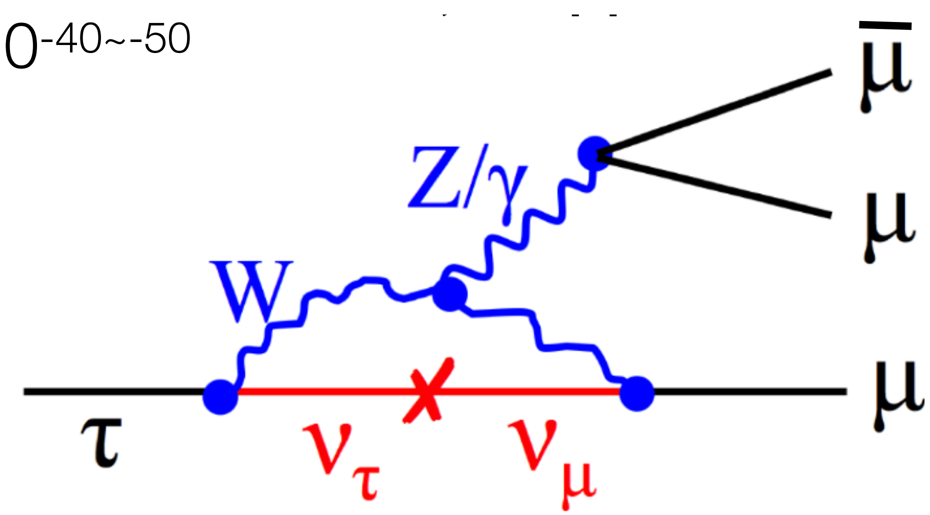
Lepton flavor violating τ->3μ decay

Flavor of particles: a mystery yet to be solved.

Quarks and neutrinos mix, do charged leptons also mix and why?

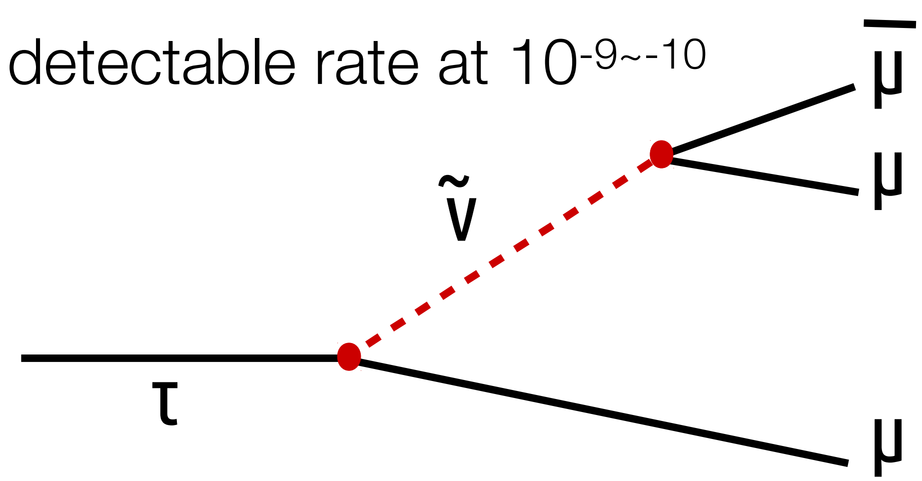
search for lepton flavor violating decays

Neutrino oscillations : $10^{-40\sim-50}$

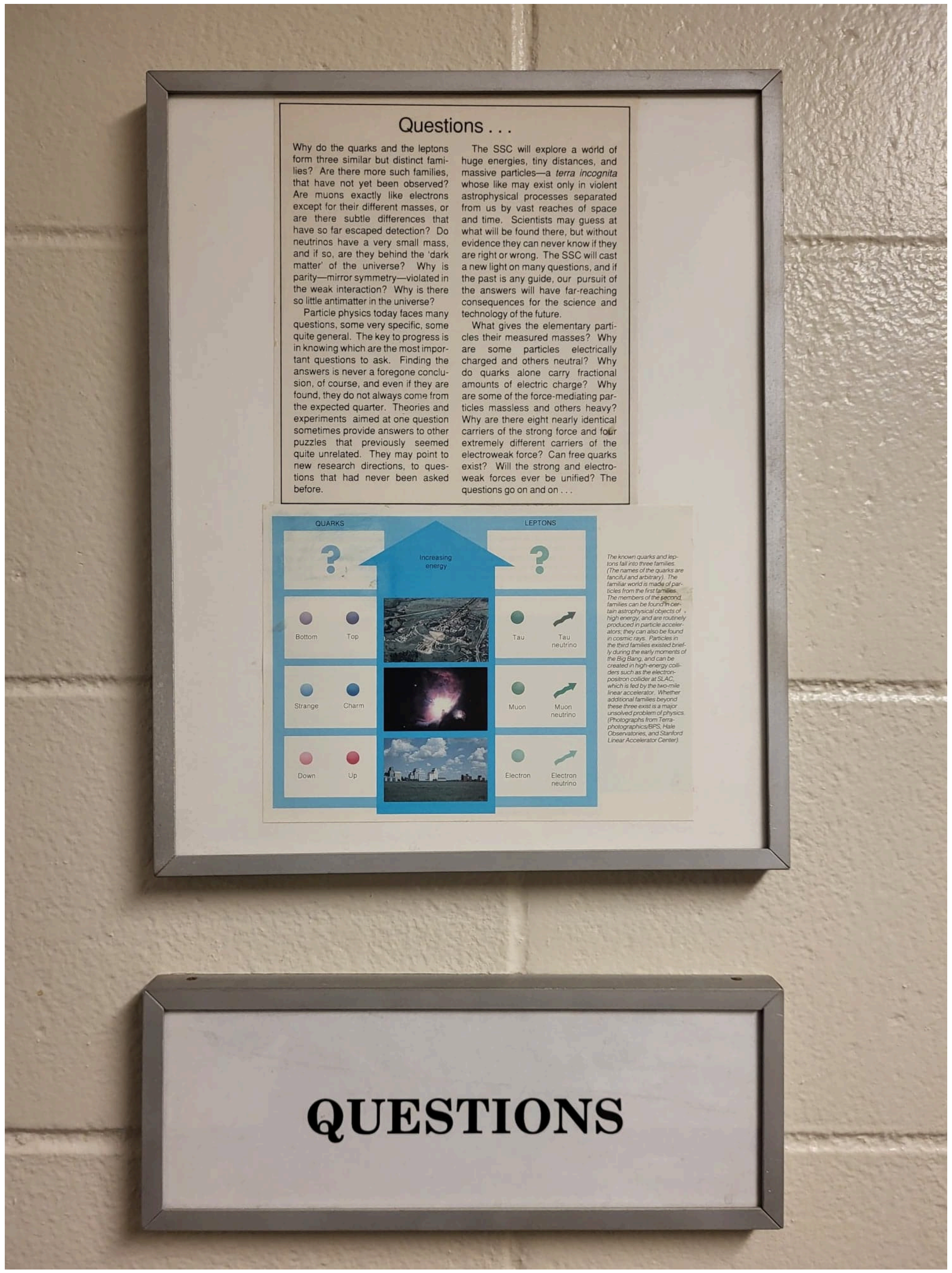


New physics results in

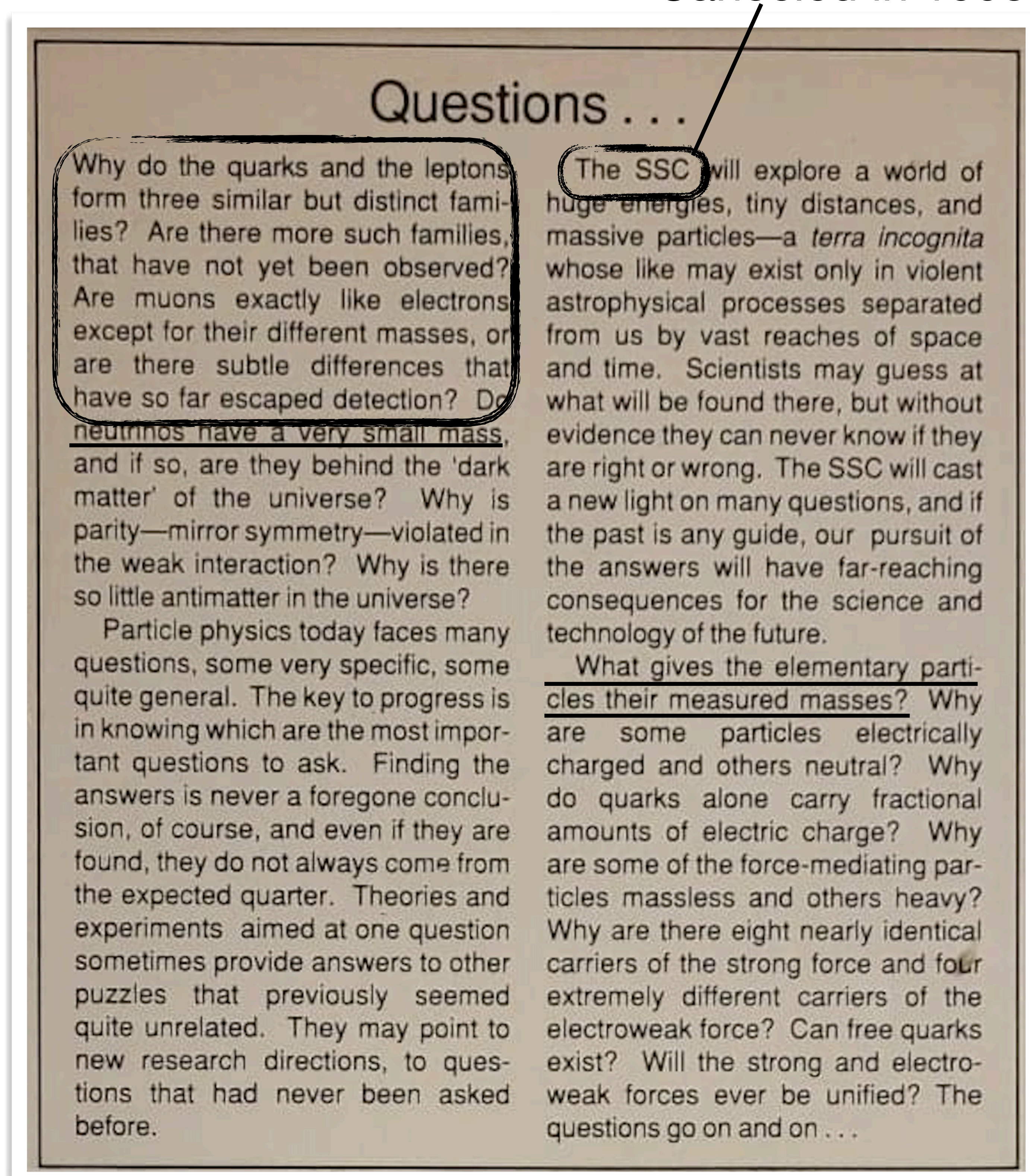
detectable rate at $10^{-9\sim-10}$



Hallway in Purdue Physics Building



Canceled in 1993



Questions . . .

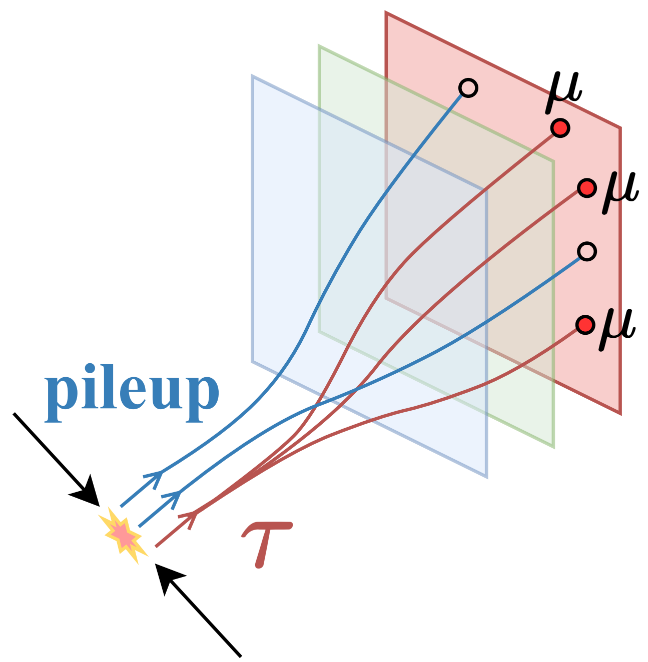
Why do the quarks and the leptons form three similar but distinct families? Are there more such families, that have not yet been observed? Are muons exactly like electrons except for their different masses, or are there subtle differences that have so far escaped detection? Do neutrinos have a very small mass, and if so, are they behind the 'dark matter' of the universe? Why is parity—mirror symmetry—violated in the weak interaction? Why is there so little antimatter in the universe?

Particle physics today faces many questions, some very specific, some quite general. The key to progress is in knowing which are the most important questions to ask. Finding the answers is never a foregone conclusion, of course, and even if they are found, they do not always come from the expected quarter. Theories and experiments aimed at one question sometimes provide answers to other puzzles that previously seemed quite unrelated. They may point to new research directions, to questions that had never been asked before.

The SSC will explore a world of huge energies, tiny distances, and massive particles—a *terra incognita* whose like may exist only in violent astrophysical processes separated from us by vast reaches of space and time. Scientists may guess at what will be found there, but without evidence they can never know if they are right or wrong. The SSC will cast a new light on many questions, and if the past is any guide, our pursuit of the answers will have far-reaching consequences for the science and technology of the future.

What gives the elementary particles their measured masses? Why are some particles electrically charged and others neutral? Why do quarks alone carry fractional amounts of electric charge? Why are some of the force-mediating particles massless and others heavy? Why are there eight nearly identical carriers of the strong force and four extremely different carriers of the electroweak force? Can free quarks exist? Will the strong and electroweak forces ever be unified? The questions go on and on . . .

Charged lepton flavor violating decay

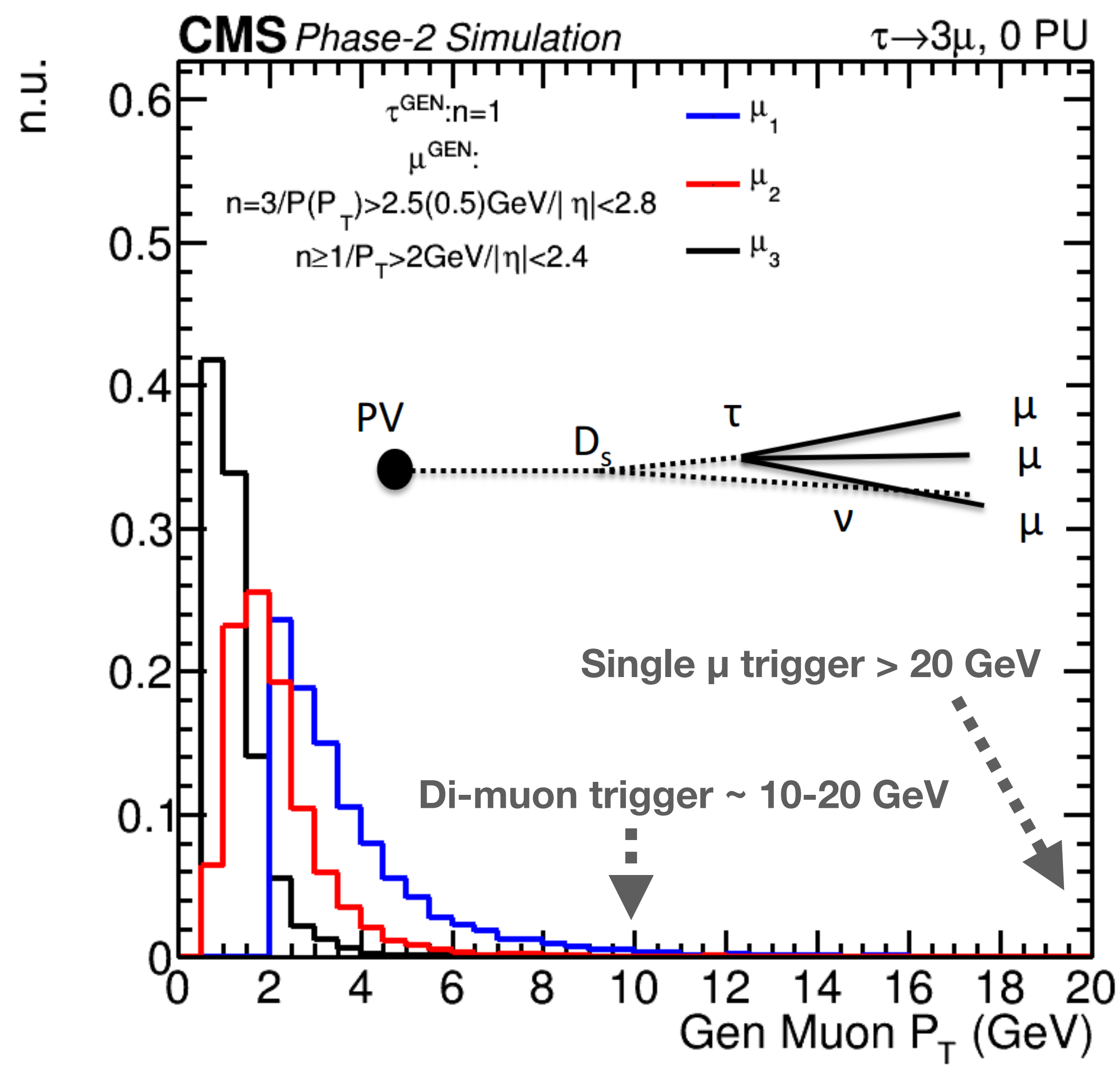


Experimentally challenging

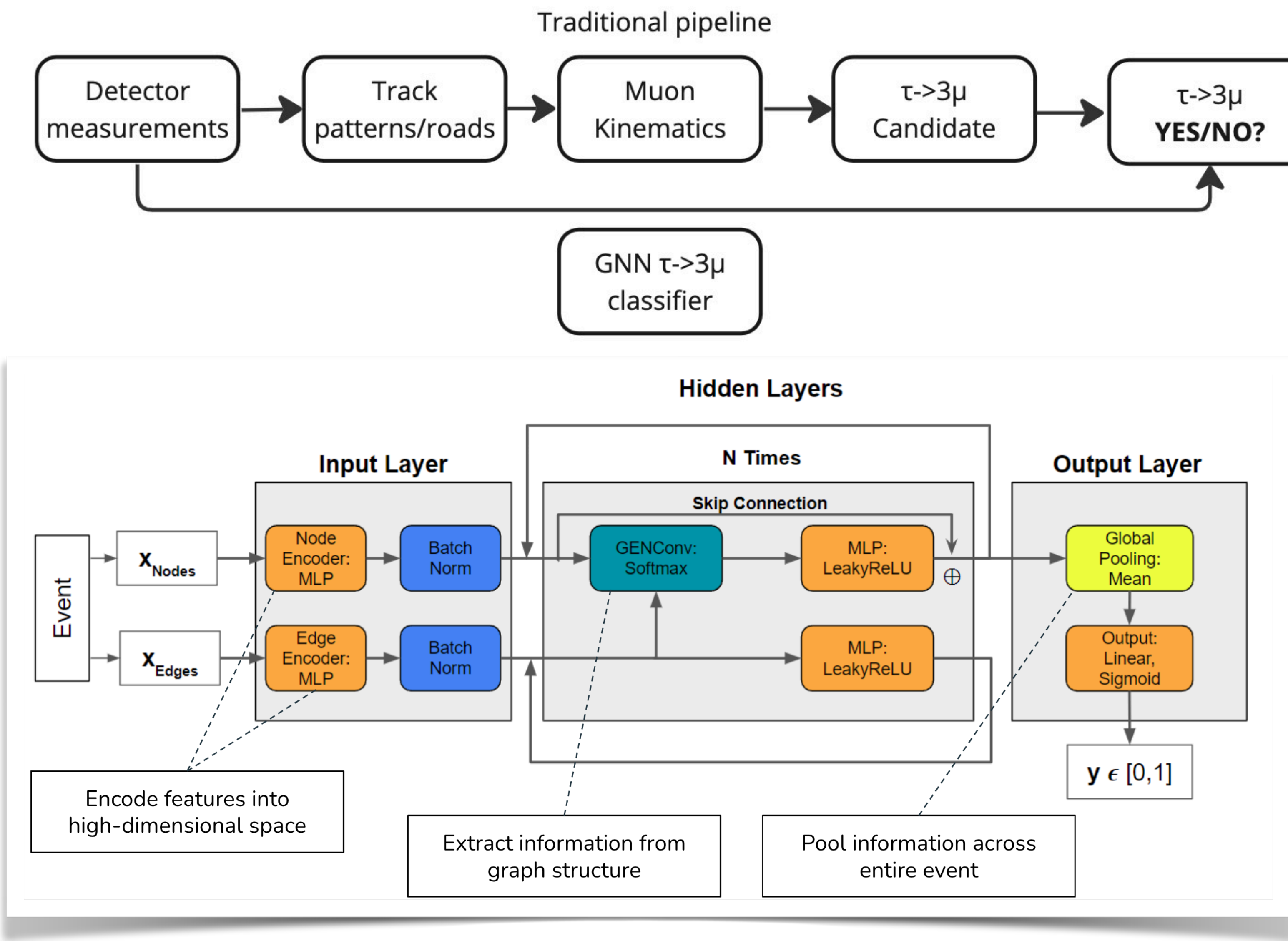
Mτ ~ 1.7 GeV : Mμ ~ 100 MeV

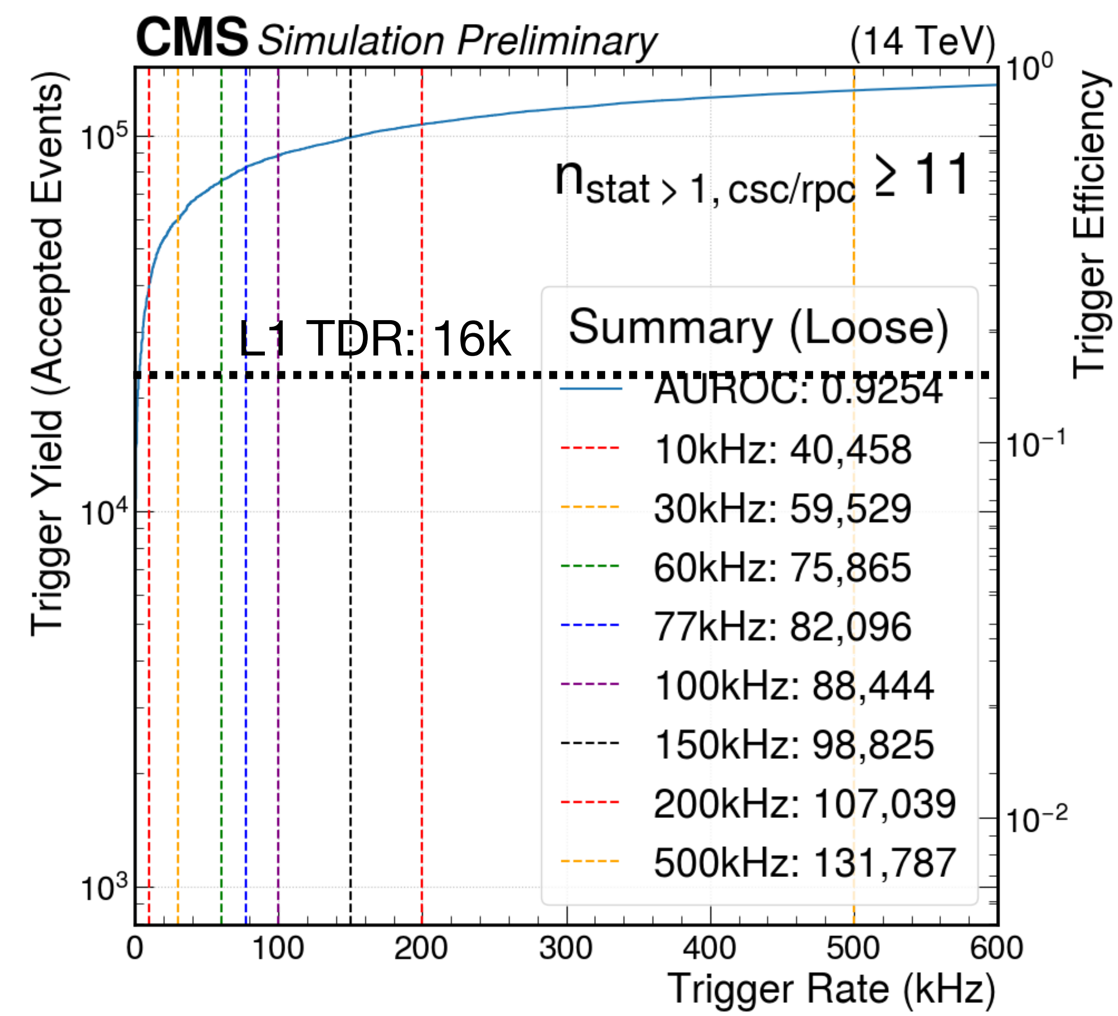
Collimated, low pT, forward muons: flagged "uninteresting" with nominal triggers

Current signal acceptance: < 10^{-11}



End-to-end τ->3μ detection





- 5 times more events triggered compared to the L1T TDR
- Can also detect long-lived particles resulting in muon “showers”

To make it a reality…

- An actual end-to-end reconstruction

Irregular computation patterns in graph generation

There are some workaround (Local Sensitivity Hashing)

Measurements association with particles

Some ‘auxiliary studies’ in robustness and interpretability:

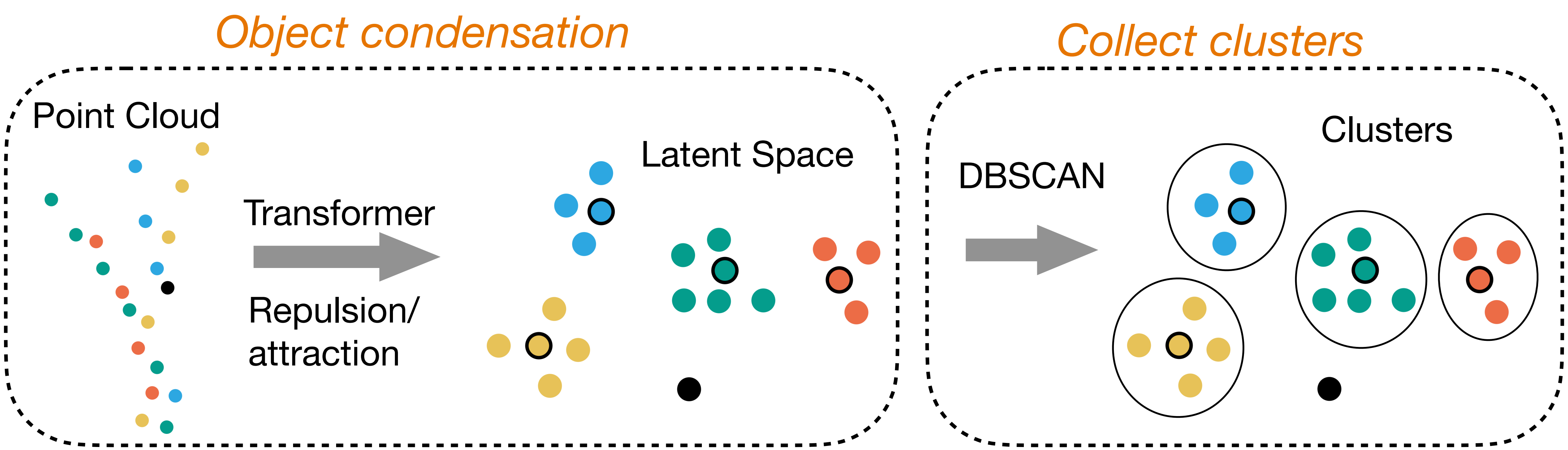
https://arxiv.org/abs/2210.16966, https://arxiv.org/abs/2201.12987

- Will share some attempts on tracking datasets
- Put it in Level-1 Trigger of CMS

HLS4ML for GNNs/Transformers

Co-design tools essential for scientific applications

End to End Reconstruction



Tracking dataset generated with ACTS software

Tracking for a full event has 50k+ points

kNN Graph Construction can be $O(n^2)$

GNNs have lots of irregular computations

Separating a full event into multiple sections. Extra overheads, hard to recover tracks across sections

Efficient Sparse Transformer

Contrastive learning with hard negative mining

With comparable accuracy

Can be trained **end-to-end**

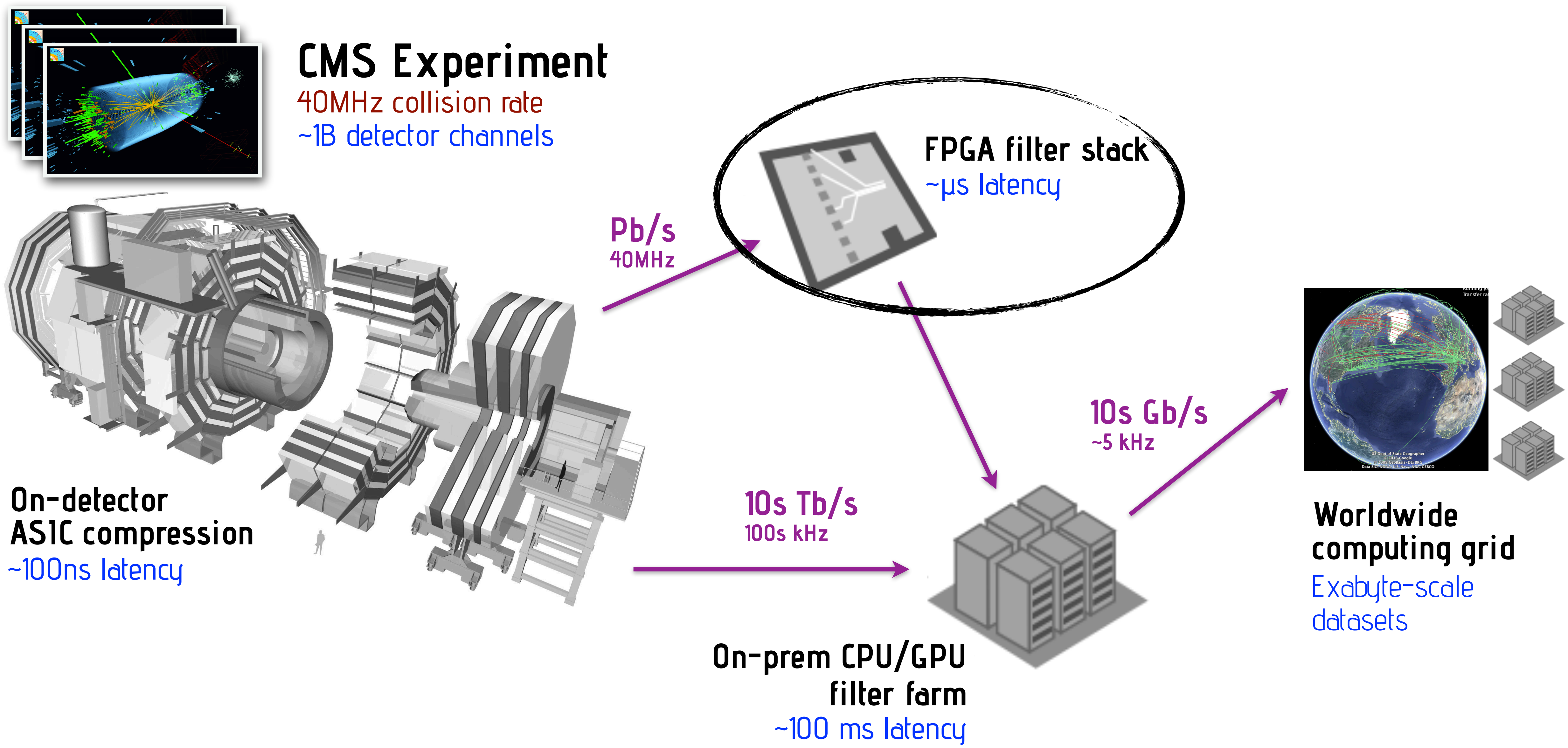
Computations are **parallelizable** and **regular,** $O(n \log n)$ complexity

Speedup on a GPU (Quadro RTX 6000)

100x faster on a full event

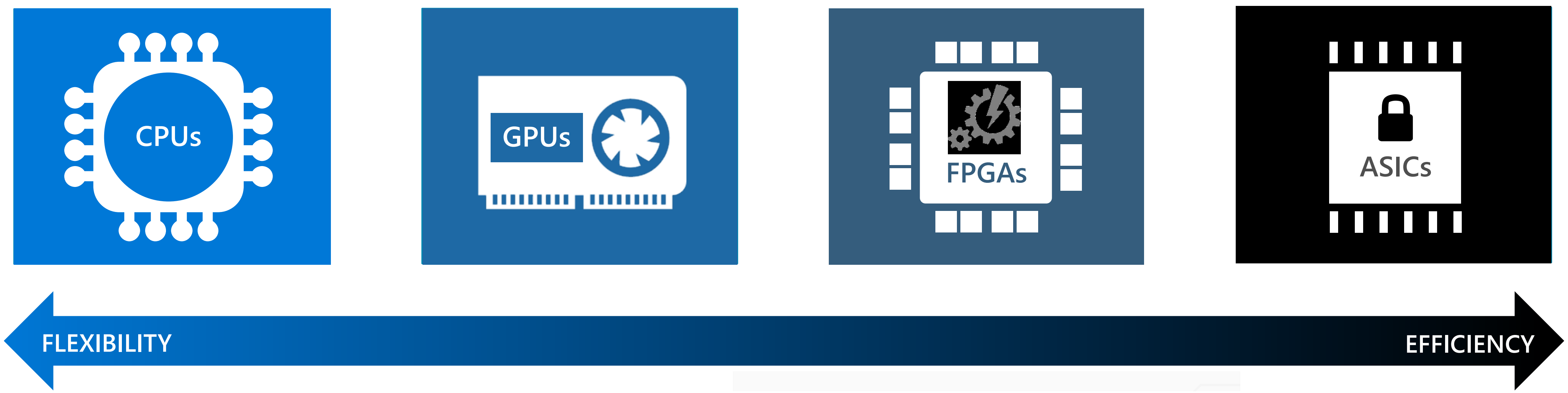
20x faster on 1/10 event

We want to put it here



Science with Big data: Multi-tier Data Processing

Specialized Hardware



L1 trigger

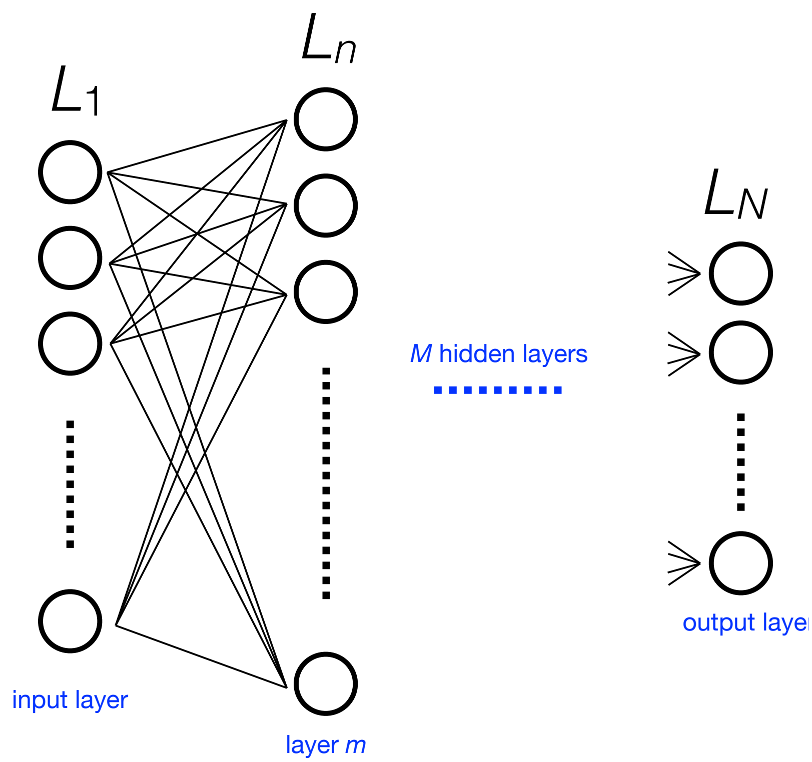
On-detector electronics

Co-design tool: crucial for prototyping AI at edge solutions

Algorithm hardware co-design for limited computing

Prototype with manageable programming barrier for domain scientists

Mapping NN onto FPGAs



$$\mathbf{x}_n = g_n(\mathbf{W}_{n,n-1}\mathbf{x}_{n-1} + \mathbf{b}_n)$$

Activation functions

Precomputed, and stored in BRAMs

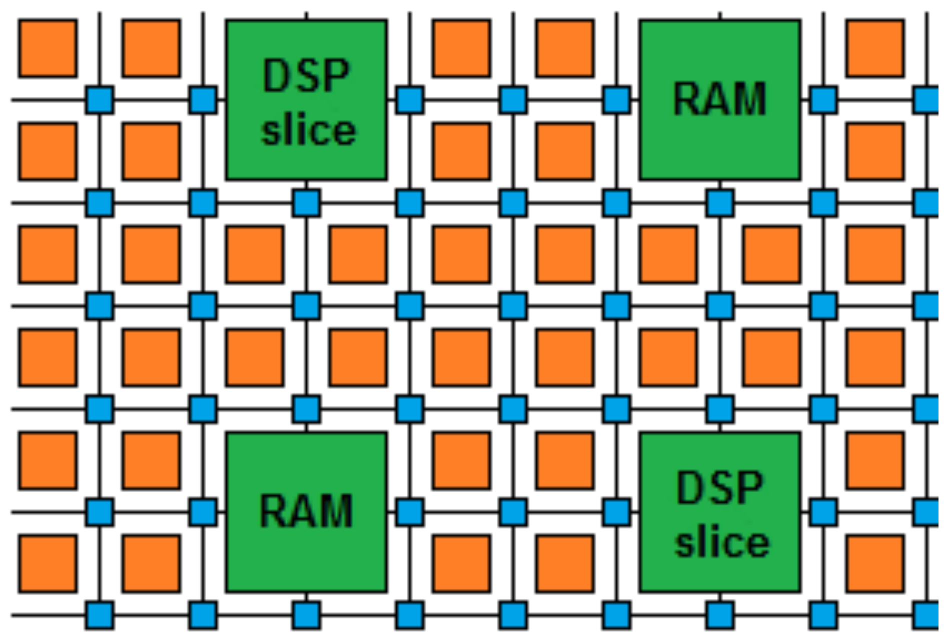
Multiplications

DSPs

Addition

Logic cells

FPGA diagram



Logic cell:

Flip-flops (FF) and

look up tables (LUTs)

Digital Signal Processors

(DSPs)

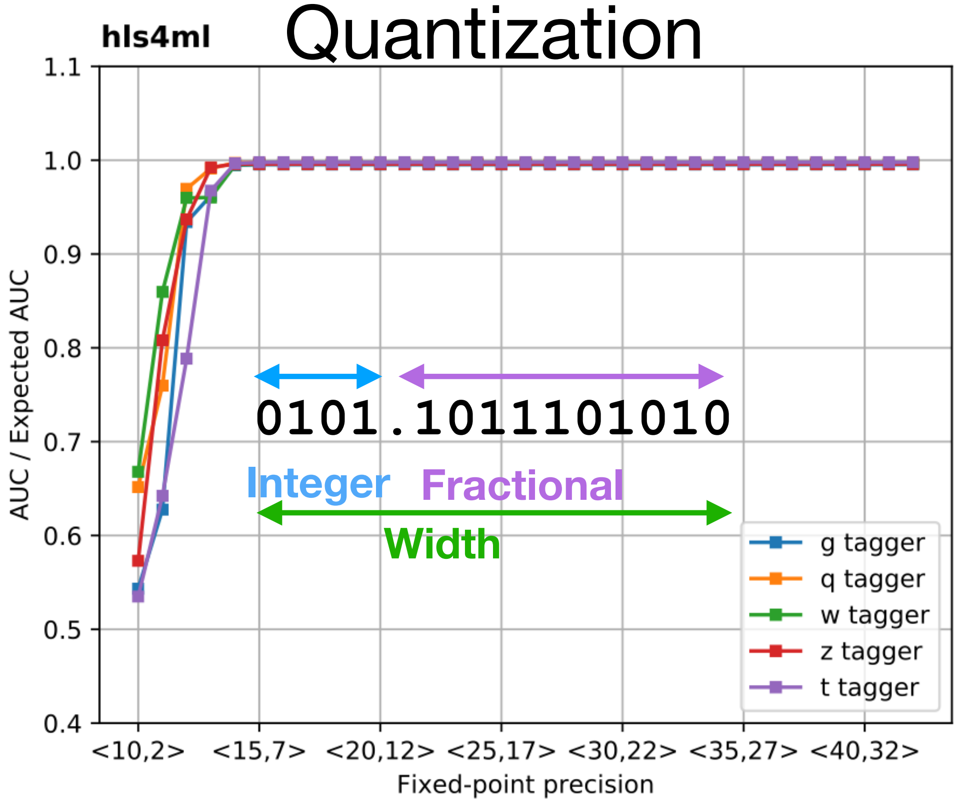
$$N_{\text{multiplications}} = \sum_{n=2}^{N} L_{n-1} \times L_n$$

Virtex Ultrascale+ VU9P

→ 6800 DSPs
1M LUTs
2M FFs
75 Mb BRAM

Efficient Algorithms

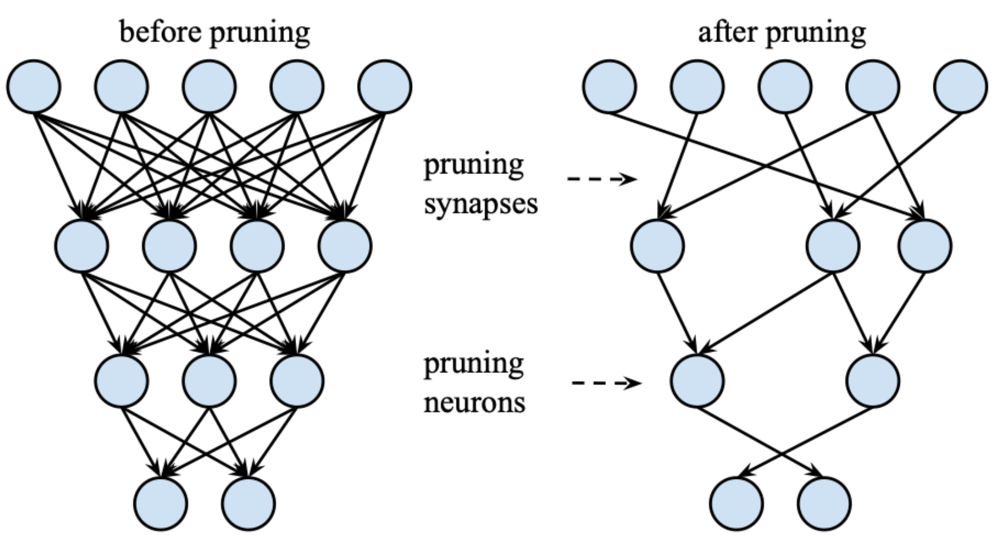
Quantization



Inspired by
Phil's Fast ML for Science workshop

Efficient Algorithms

Compression



Efficient Algorithms

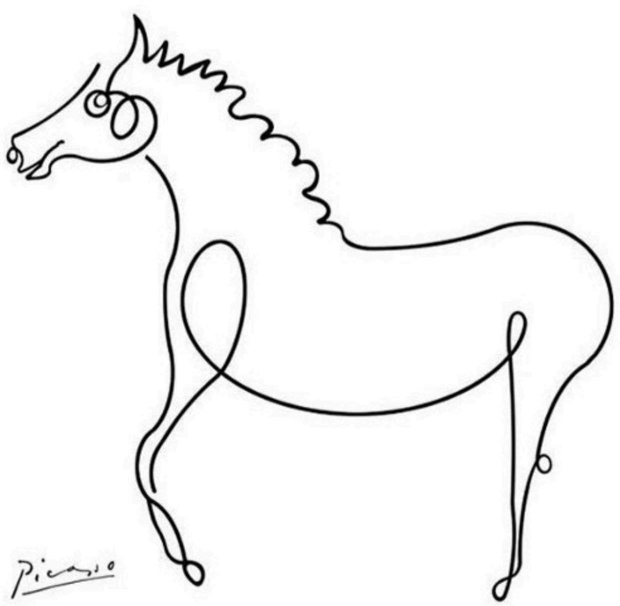
Quantization

Compression/Pruning

‘Ultimate optimization’ of ‘bits of information’

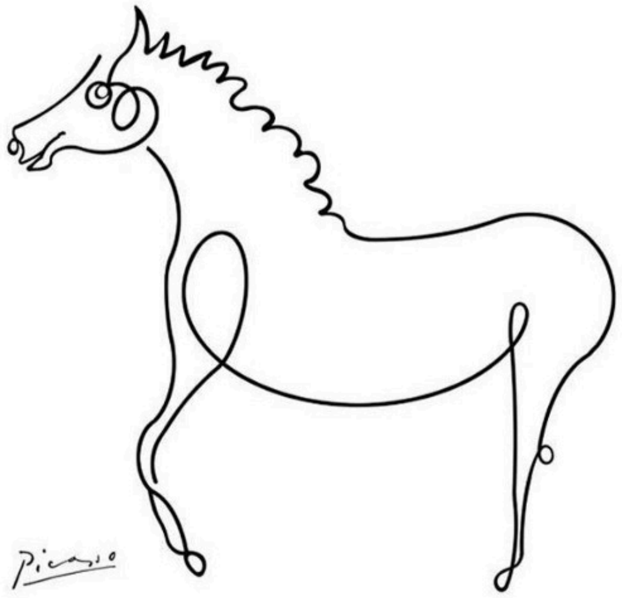
https://arxiv.org/abs/2102.11289
https://arxiv.org/abs/2304.06745

Efficient Algorithms



Compress it creatively:
knowledge distillation. e.g

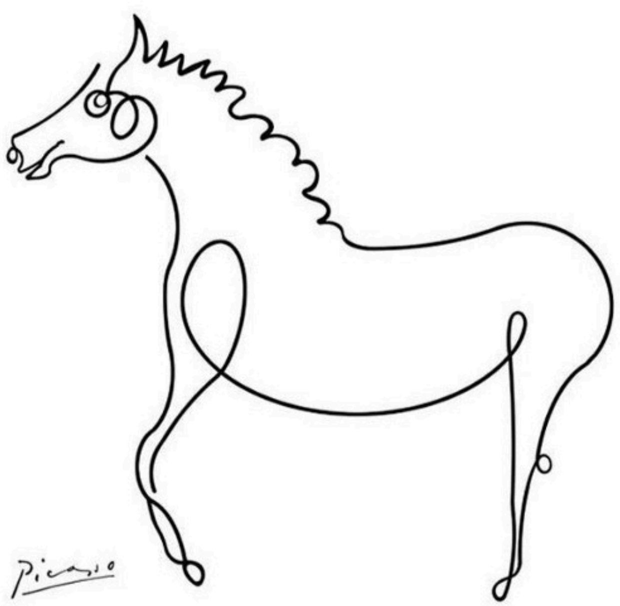
Efficient Algorithms



馬

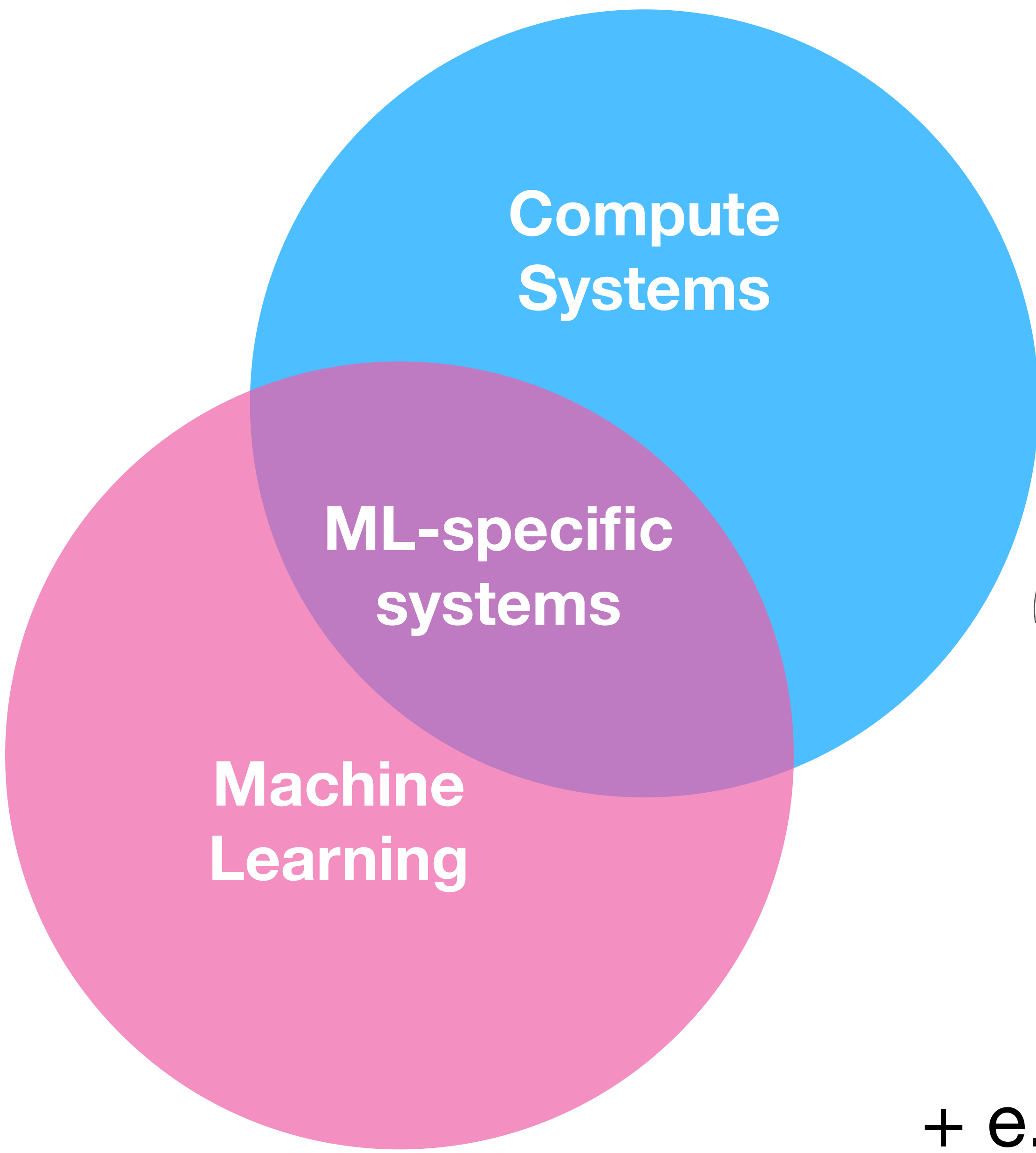
Neural Architecture search e.g. EfficientNet for image detection

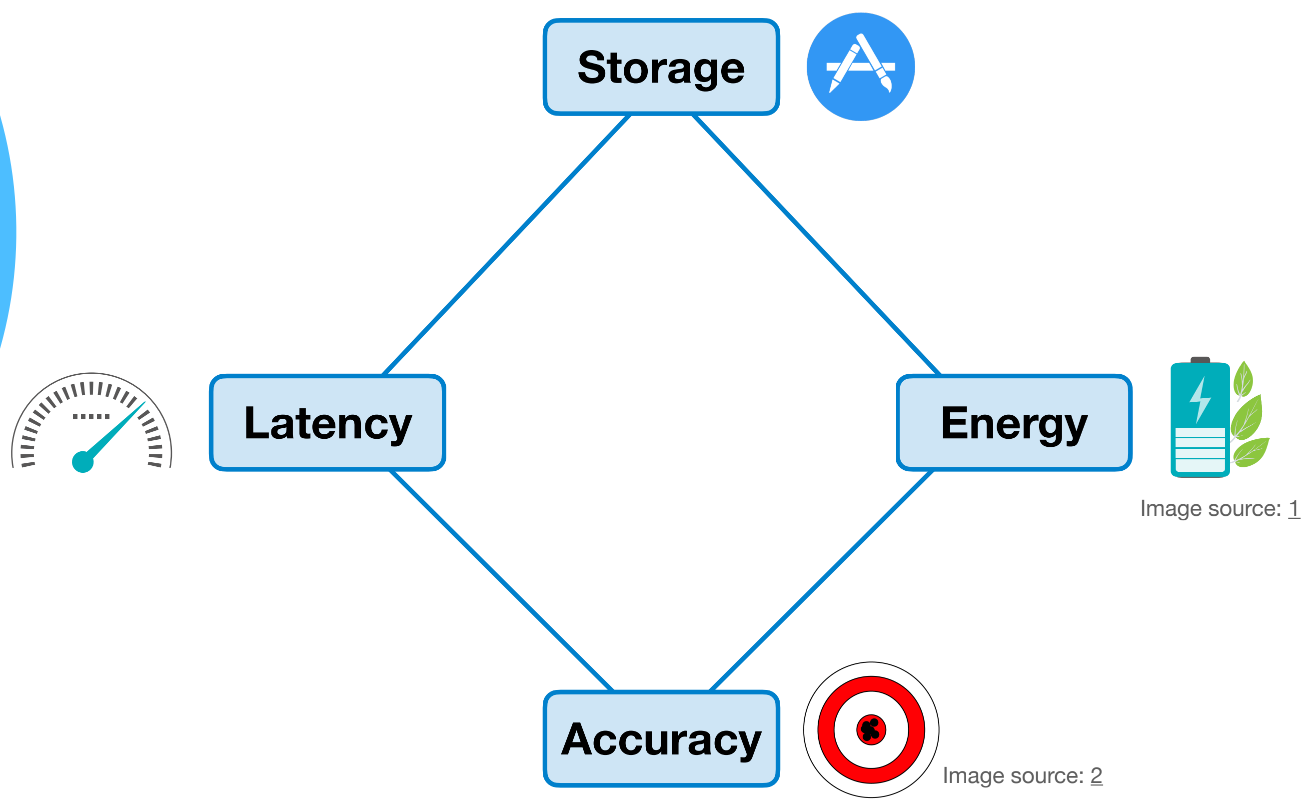
Efficient Algorithms



馬 马

Algorithm System Co-design for Your Metrics

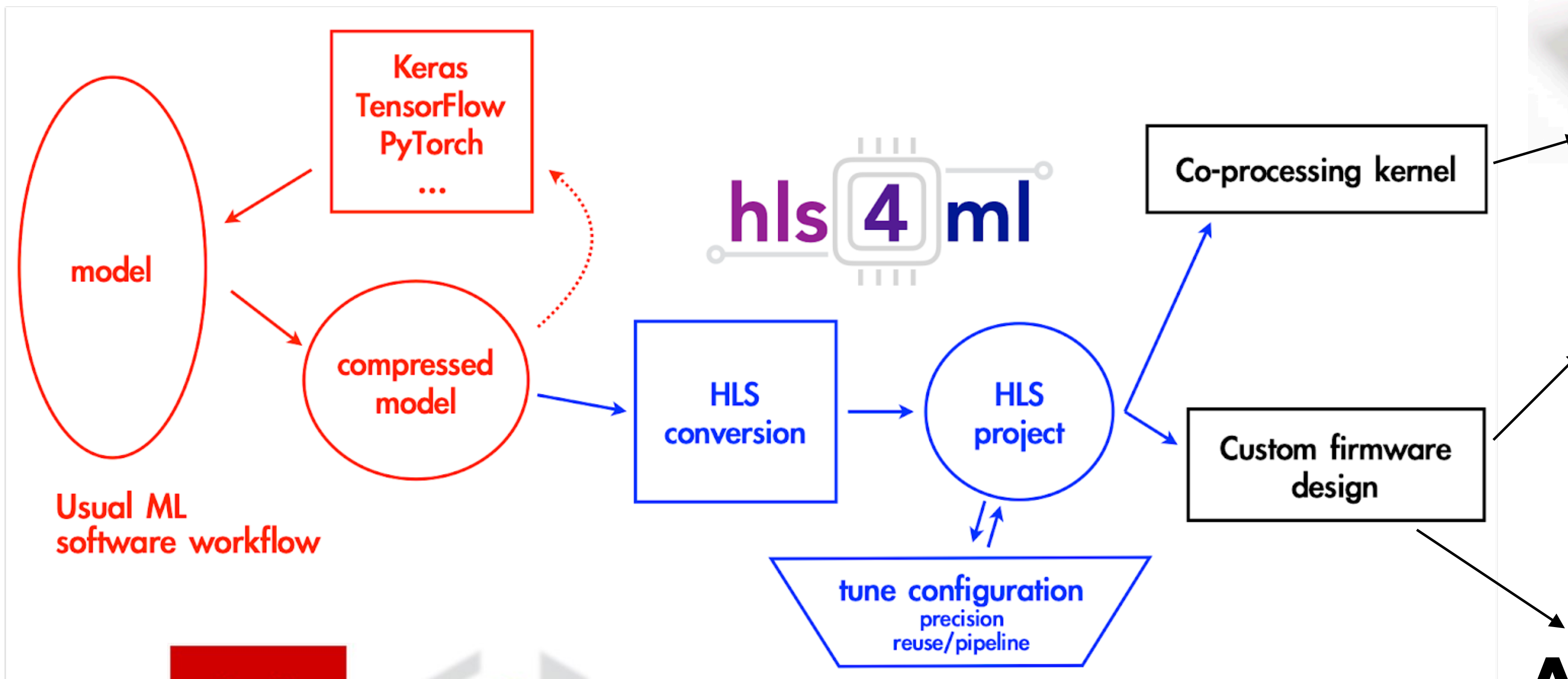


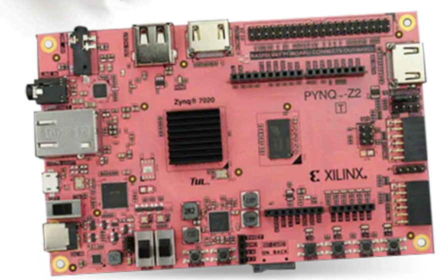


+ e.g. Radiation Environment: Fault Tolerant

Another dimension of Co-design

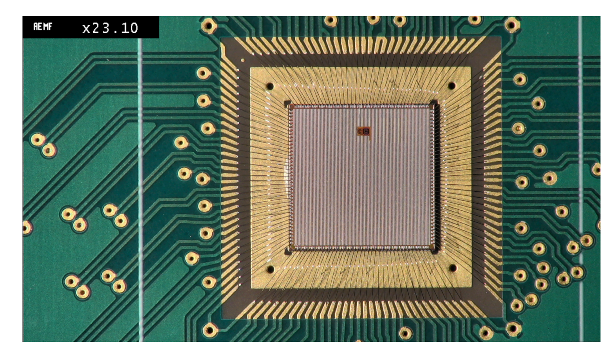
Connecting domain scientists with prototype solutions





VivadoAccelerator for Xilinx accelerators / SoCs

ASIC



(Q) K + TensorFlow

(Q) ONNX

PYTORCH

Vivado™ HLS

HLS COMPILER

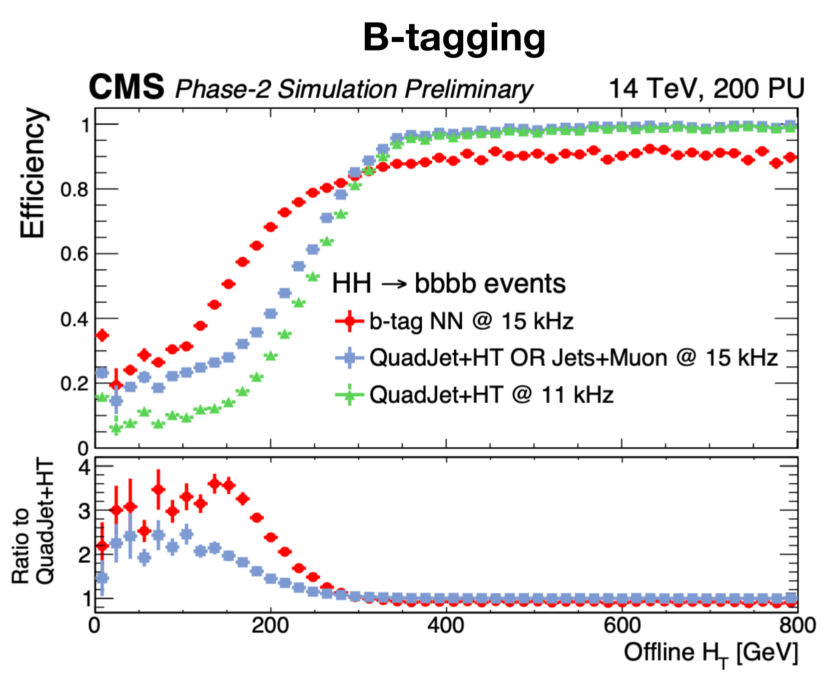
Mentor® A Siemens Business

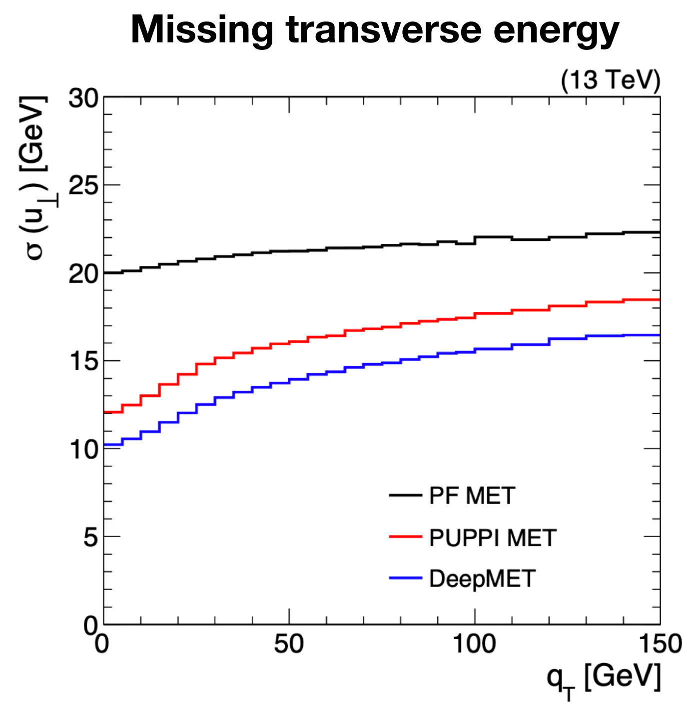
Catapult

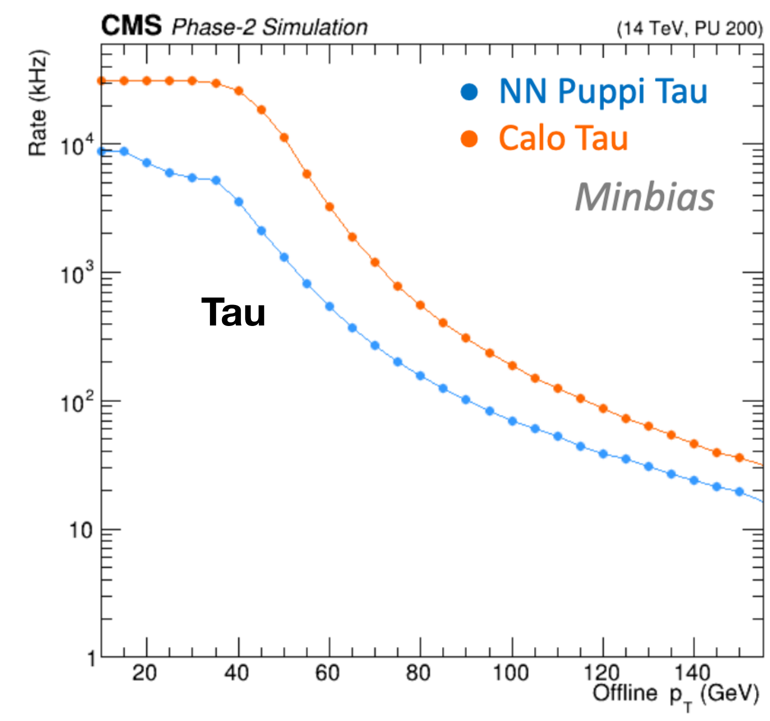
Coming soon

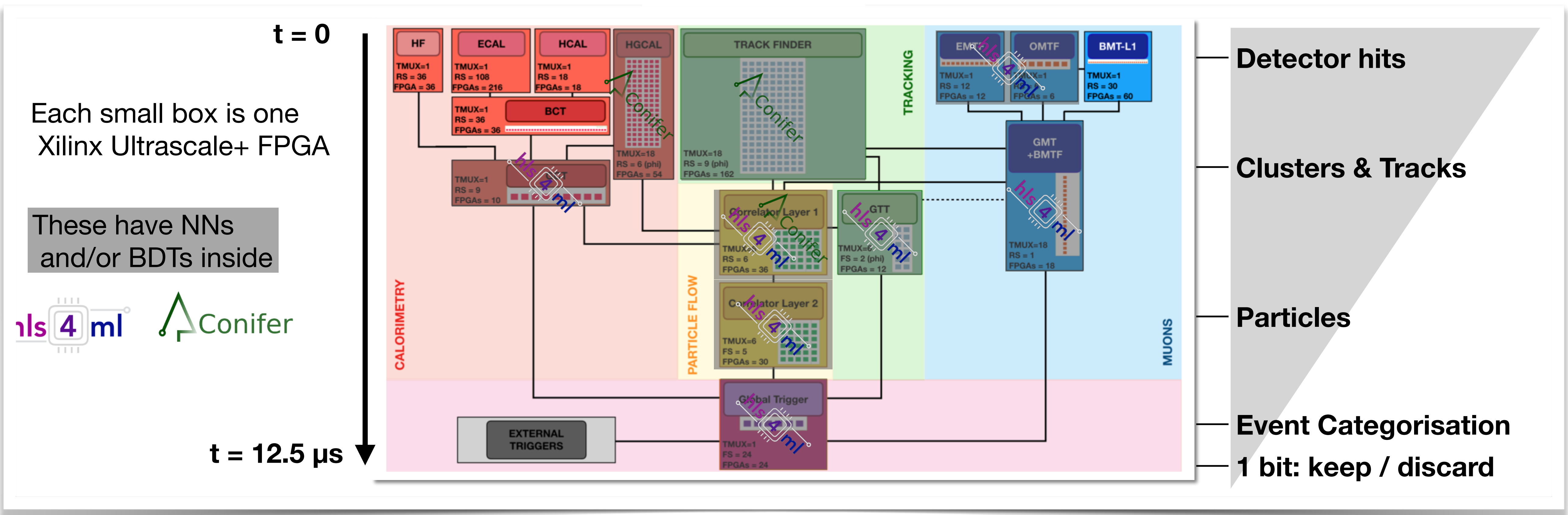
HLS4ML: to aid prototype science application solutions

ML everywhere in CMS Phase 2 L1 Trigger









Towards Scalable, Flexible, Adaptable GNN/ transformer with HLS4ML

- **hls4ml: great support for MLP and CNN Keras models.**
- Support of parsing PyTorch models: this has been improved!
- **Some (non-trivial) engineering work to support GNN/transformers:**
- Tau3mu Detection: MessagePassing layers, and meet 100 ns latency!
- **Long term: need to improve hls4ml code generation**
  - Current code generation in hls4ml is based on naive string generation - i.e., it becomes a mess very fast for anything complex.

sPHENIX tracking GNN hls4ml synthesis results

Extremely preliminary - DO NOT TRUST NUMBERS

- Network inputs: nodes=80, edges=100
- Input network
  - Can be parallelized to be “nodes” times faster (i.e., 15ns)

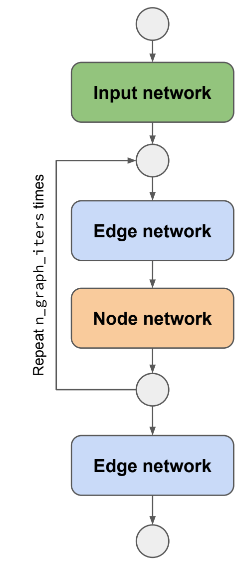
Latency	BRAMs	DSPs	FFs	LUTs
1.2 us	6.5%	0.3%	5%	7.5%

- Edge network

Latency	BRAMs	DSPs	FFs	LUTs
3 us	15%	2%	20%	65%

- Node network (results from HLS synthesis, vivado synthesis OOM'd)
  - Neet to optimize the scatter_add function (expecting ~2us for the net)

Latency	BRAMs	DSPs	FFs	LUTs
12 us	42%	7%	-	-



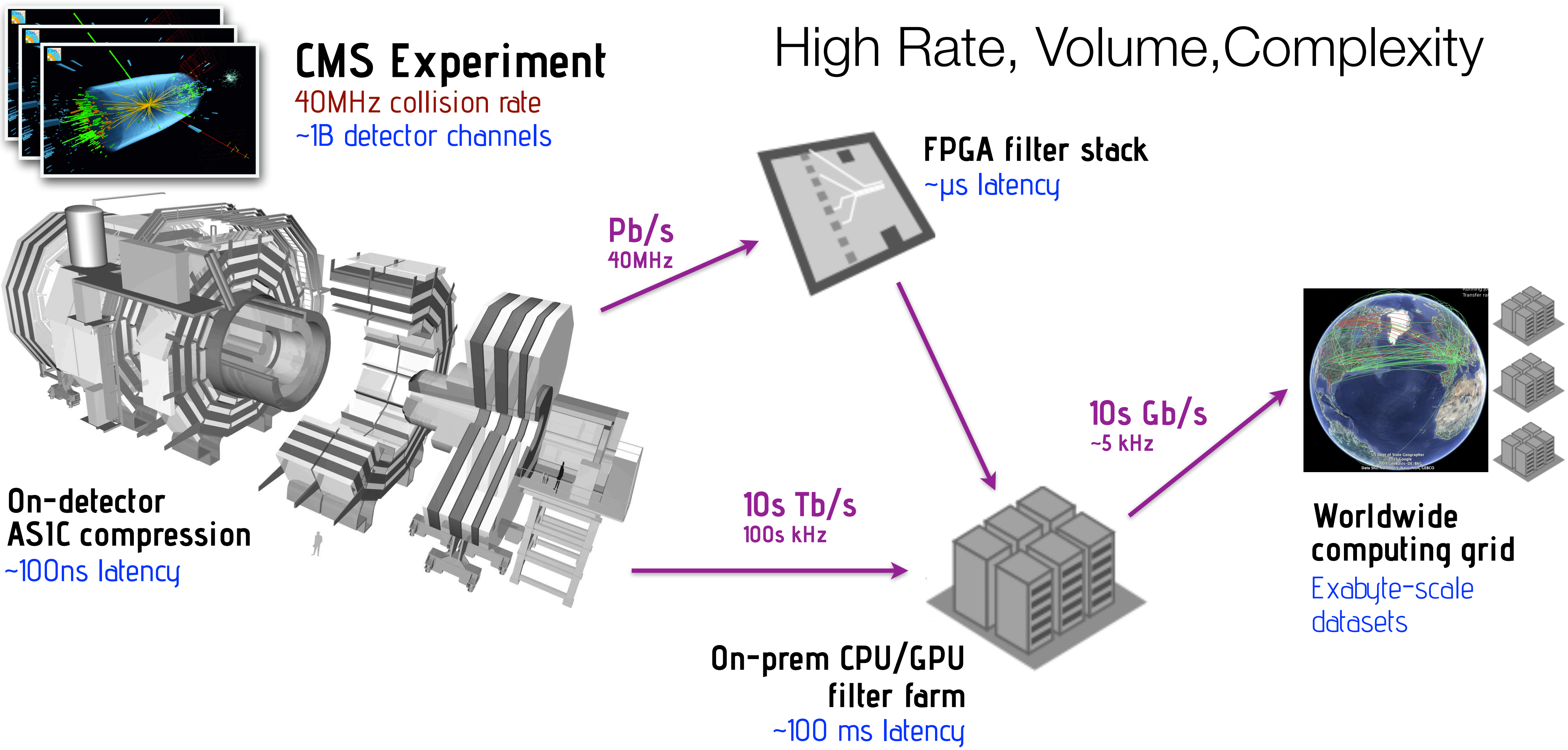
Example: <u>Extended operations supported in hls4ml to implement a GNN developed for track reconstruction in the sPhenix trigger</u>

- Added missing operations for GNN: Scatter_* “getitem”, “gather”, “ones()” and “zeros()” etc

What else?

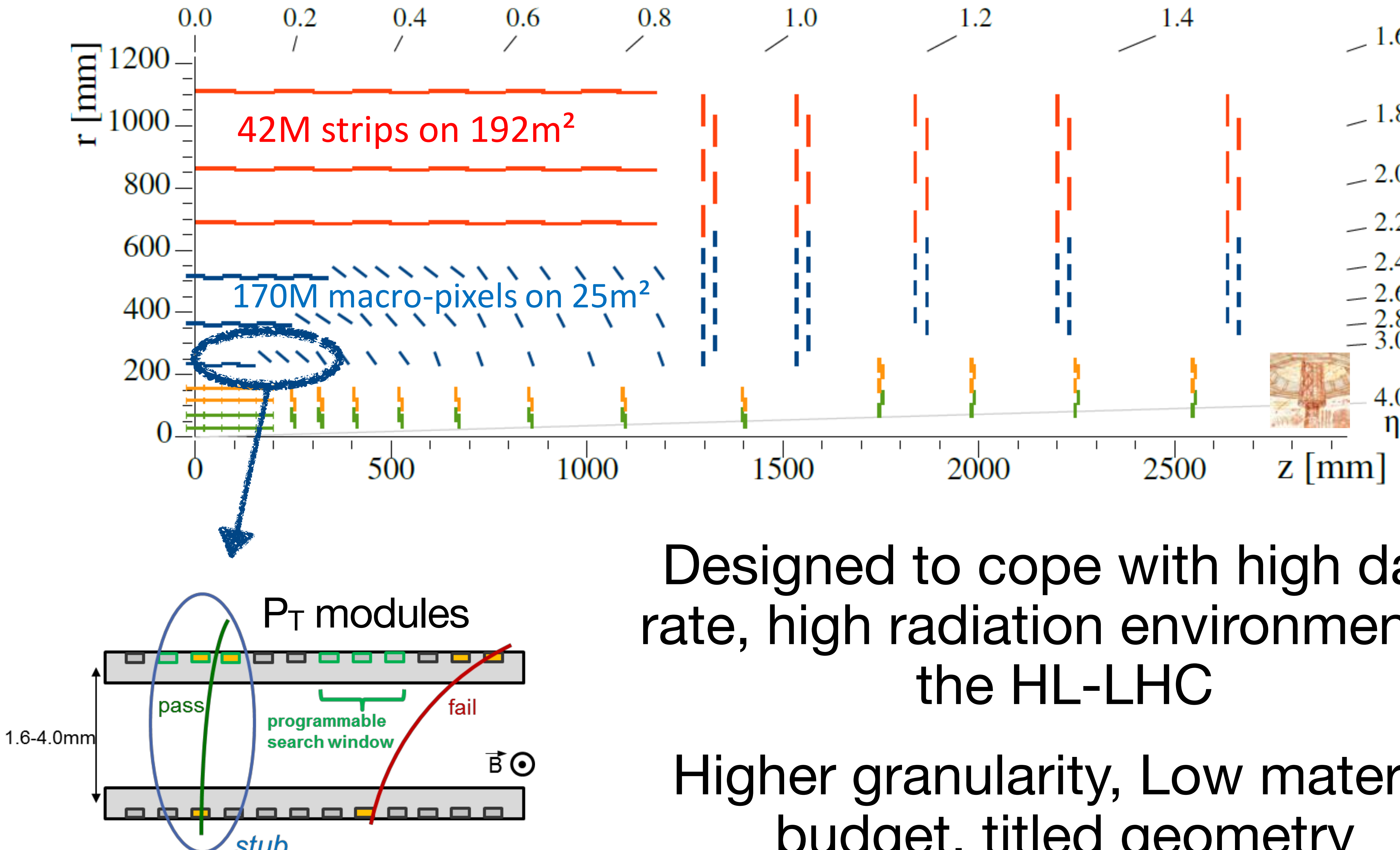
On-Chip?

High Rate, Volume,Complexity



Science with Big data: Multi-tier Data Processing

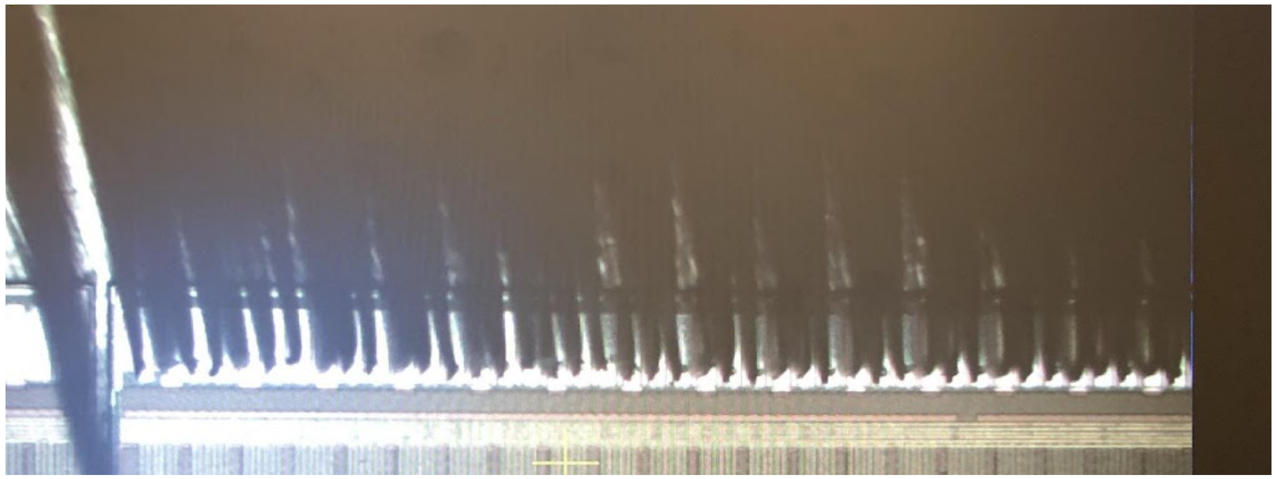
P_T modules to provide hardware trigger capabilities

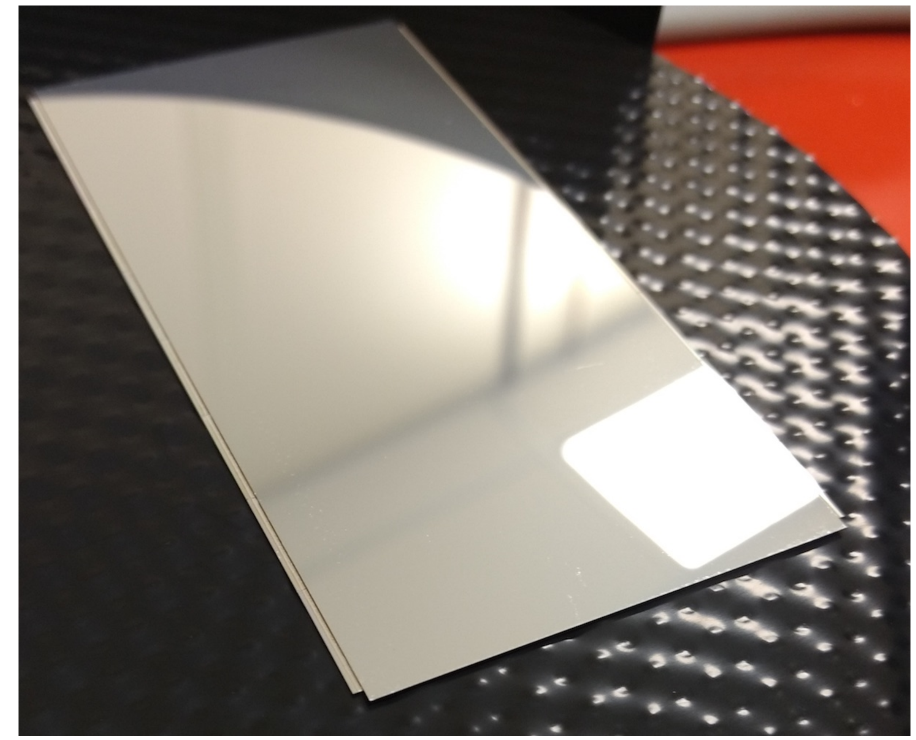


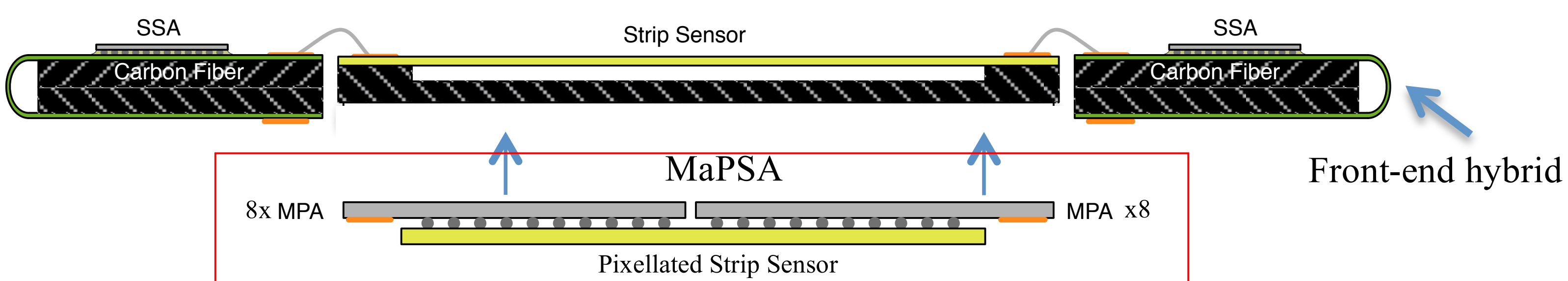
Designed to cope with high data rate, high radiation environment at the HL-LHC

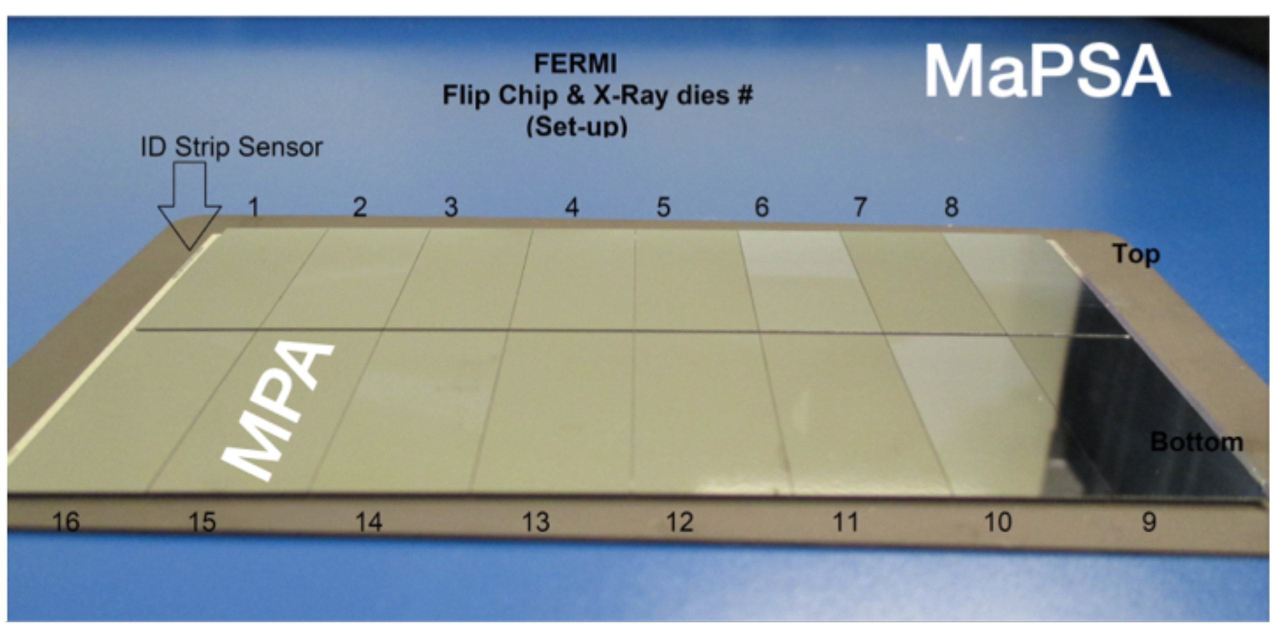
Higher granularity, Low material budget, titled geometry

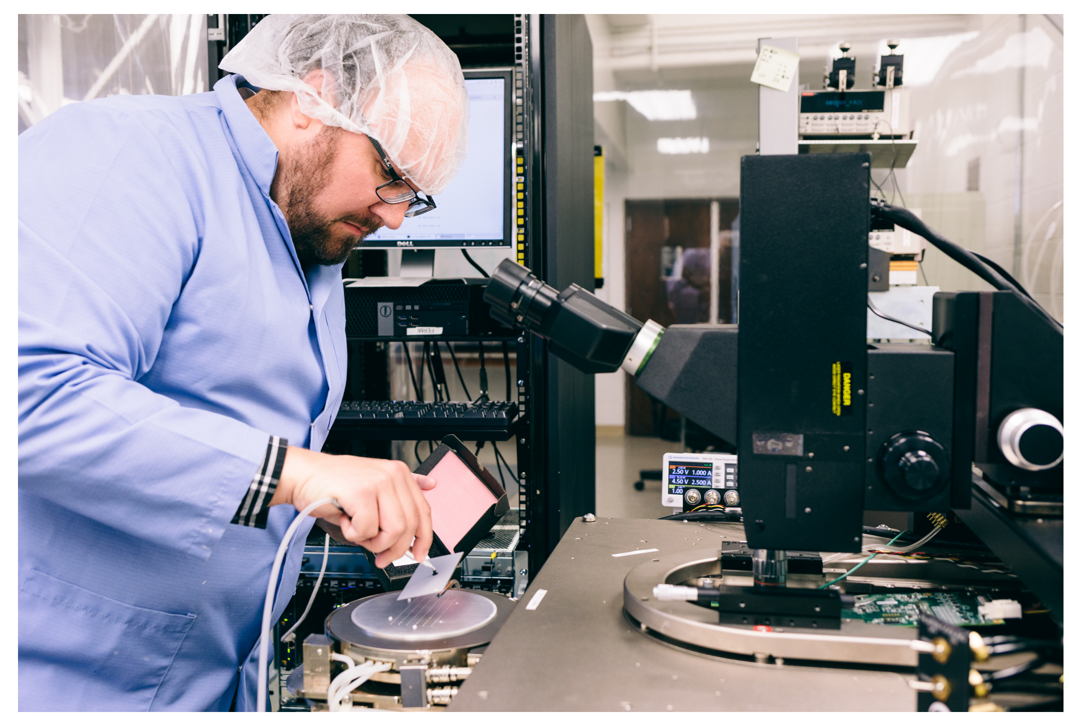
Testing at Purdue



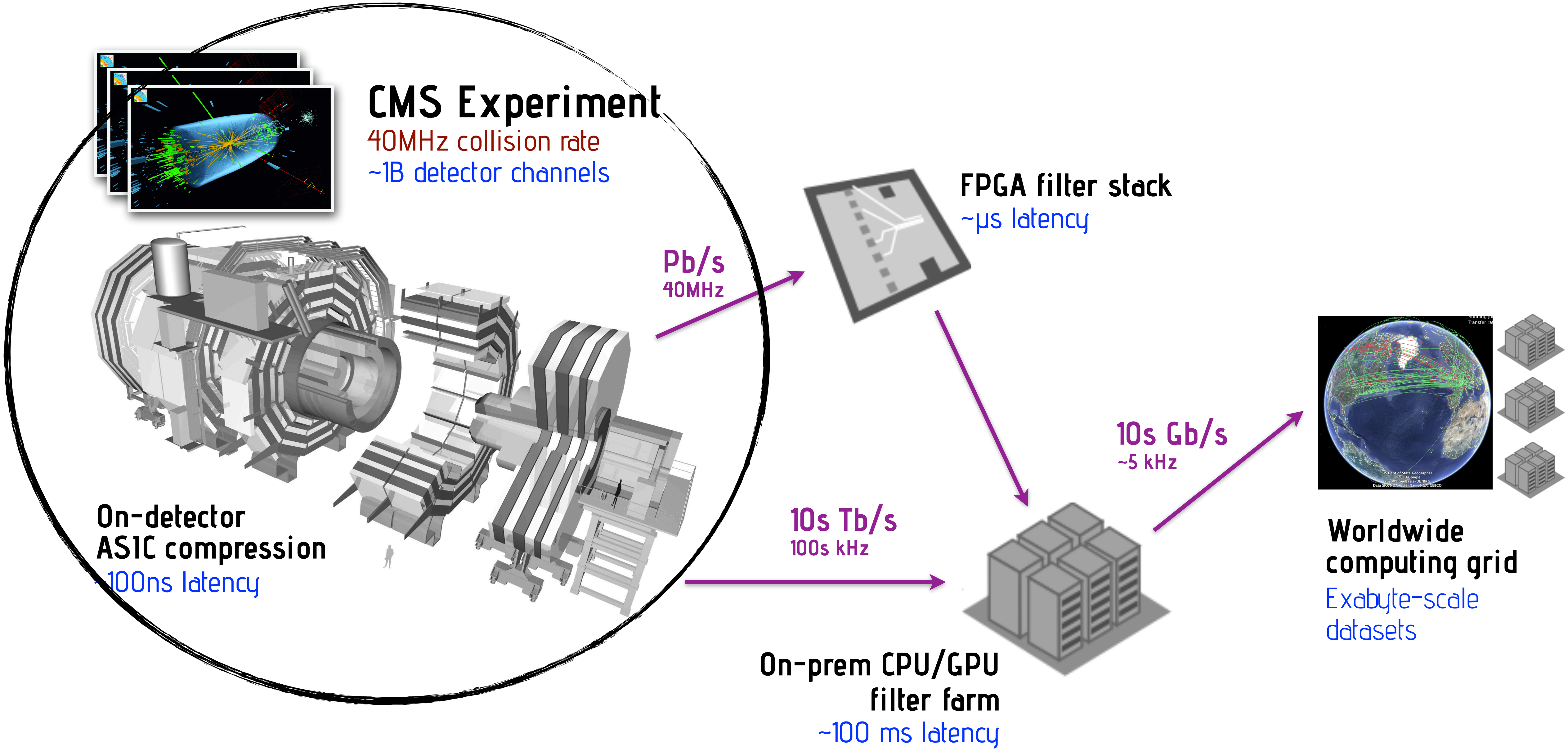








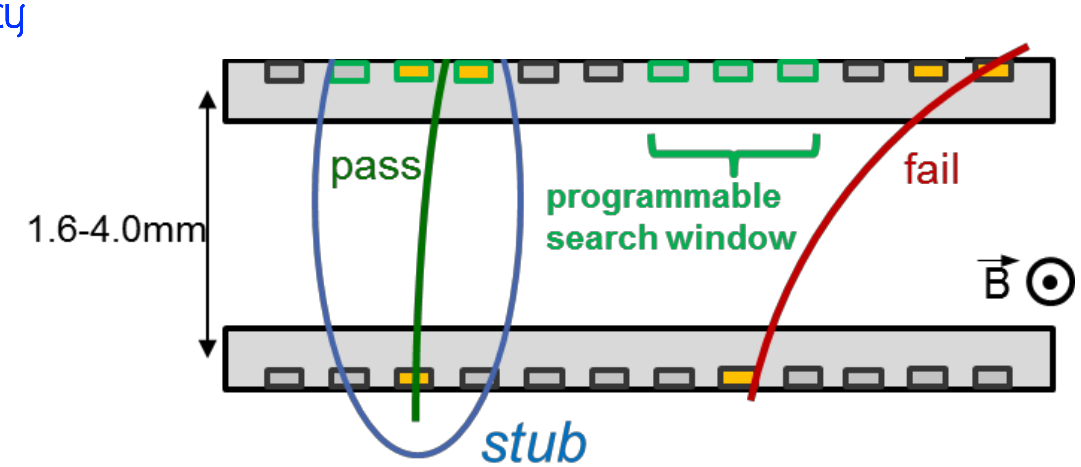
‘Pt modules’ for pixels?



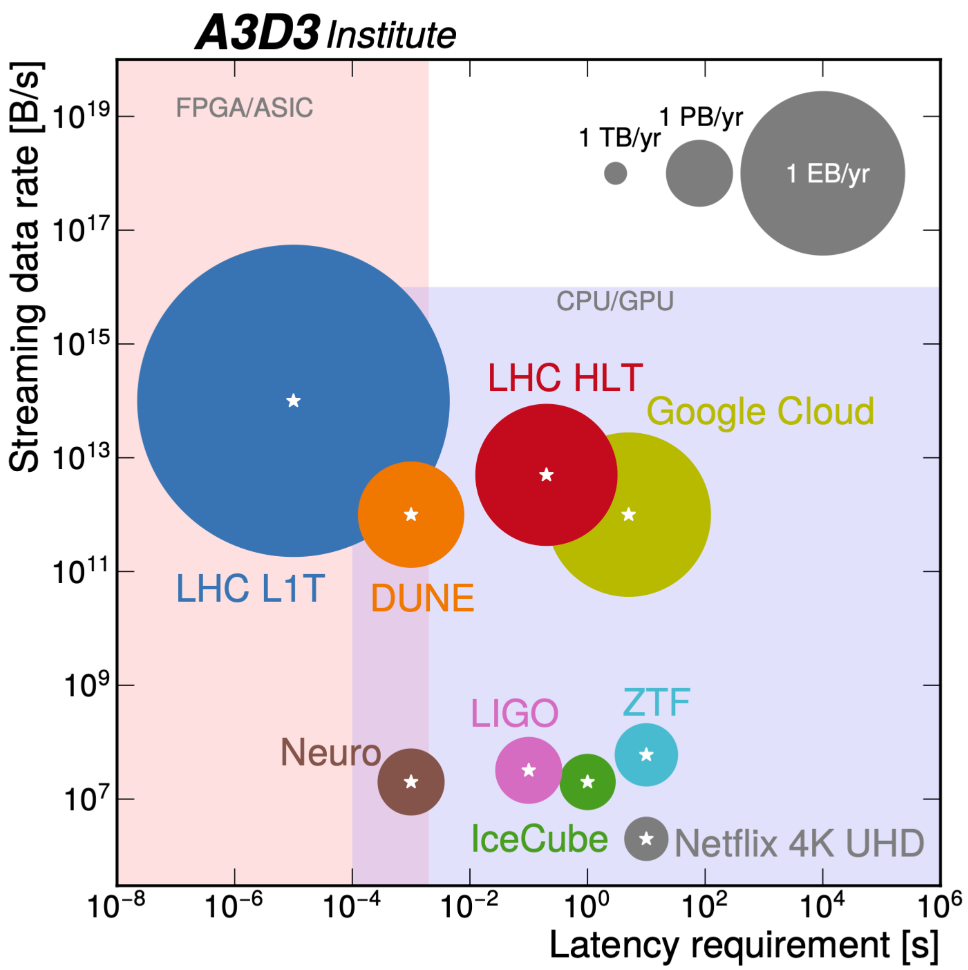
Enabled by HLS4ML catapult

See talk by Jieun, and Mathieu

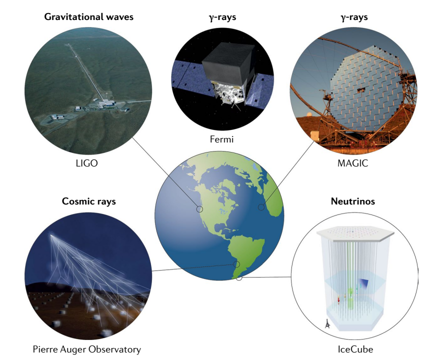
Symbolic form, plus symbolic regression?



Accelerated AI Opportunities

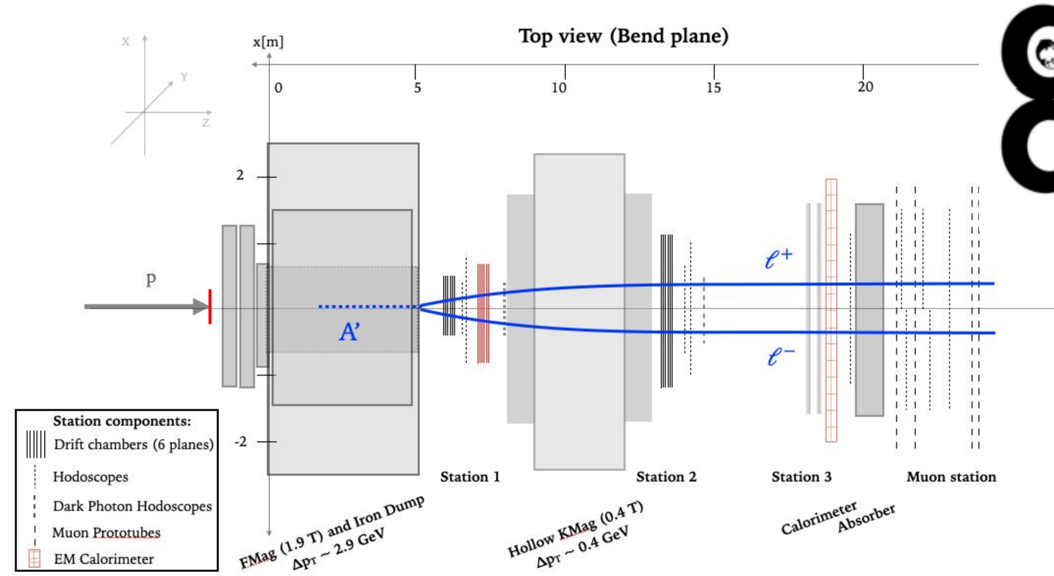


Accelerated AI Algorithms for Data-Driven Discovery



MMA

Neuro Science



dark quest

Dark sector signature
SpinQuest: muon final states
DarkQuest: e,γ,π,...

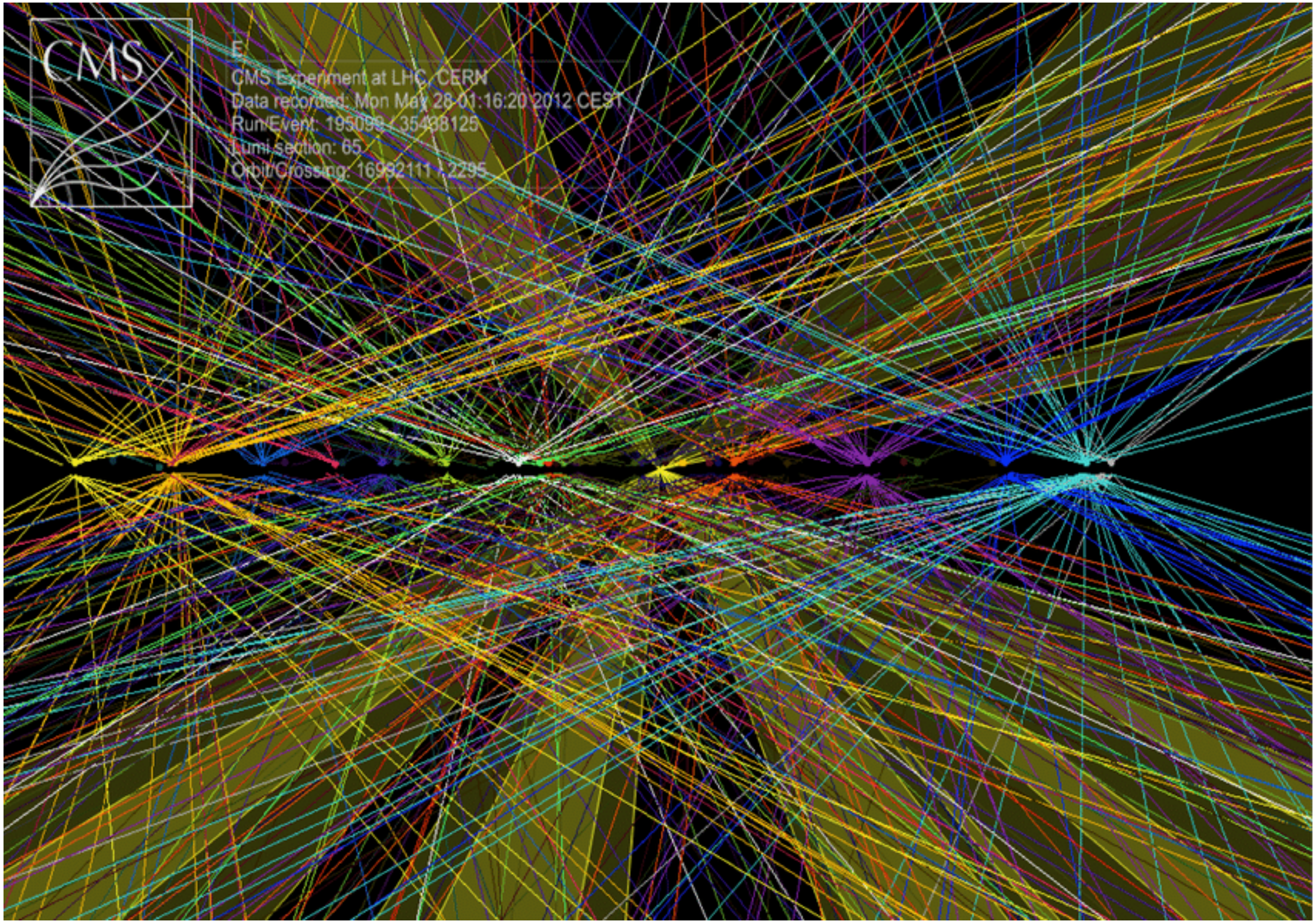
System upgrades
Existing EMCal from PHENIX
Tracking MWPC available
Tensor polarized deuteron target

NSF A3D3 institute: Domain Scientists, Computer Scientists and System Experts
Impact broader science domains Fast ML for Science Workshop

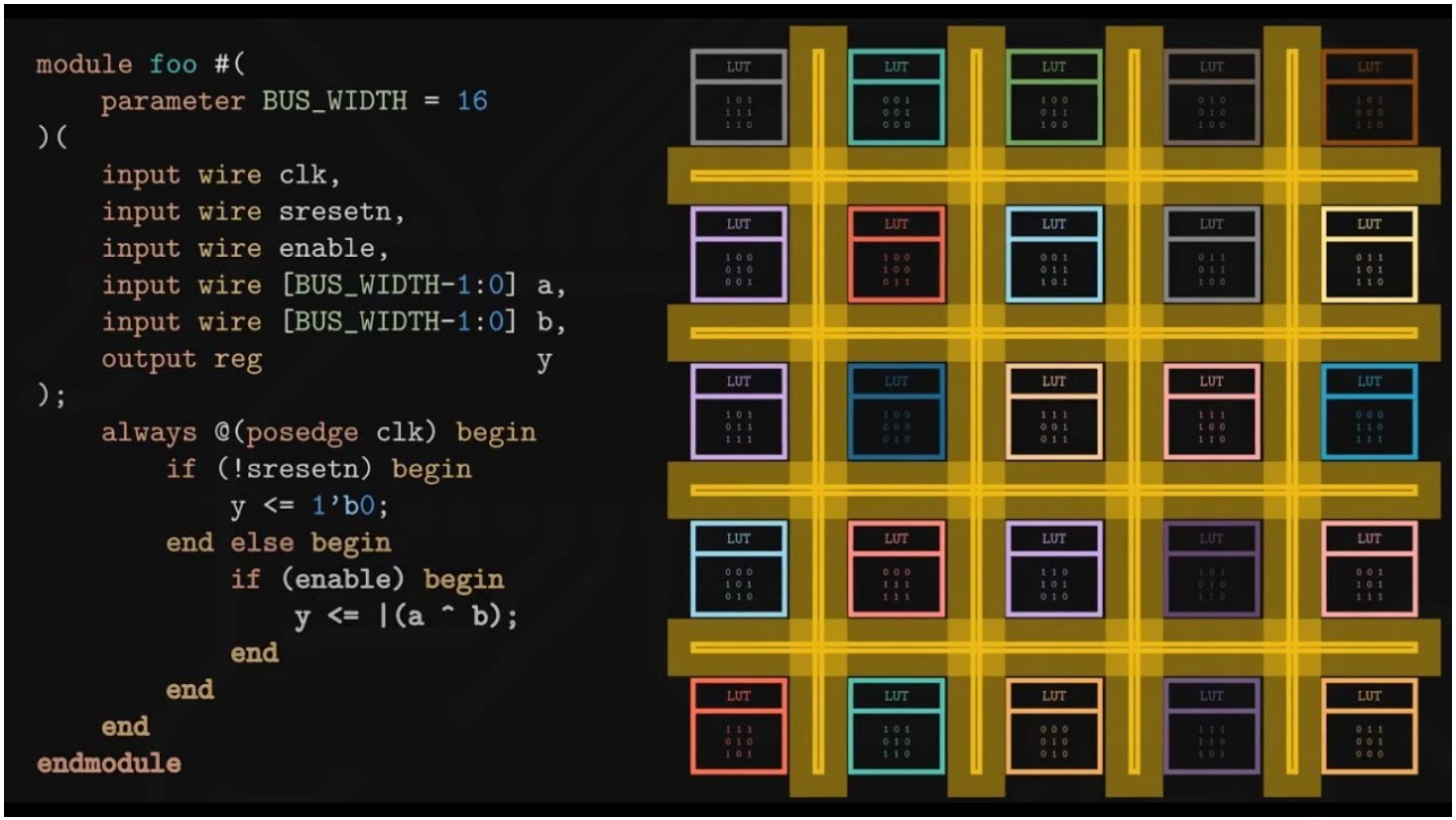
Final Remark

- Many existing and emerging opportunities in advancing our science results with ML/AI
  - Was only able to highlight a few today
- Advanced data analysis techniques, real-time system improvement.
- Elegant solutions for individual cases —> towards foundational models
- Lots of Fun

Overlapping proton collisions



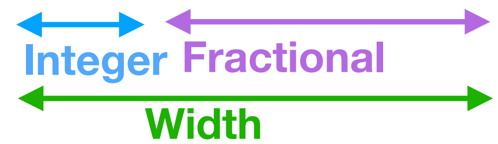
Programming FPGAs



Quantization

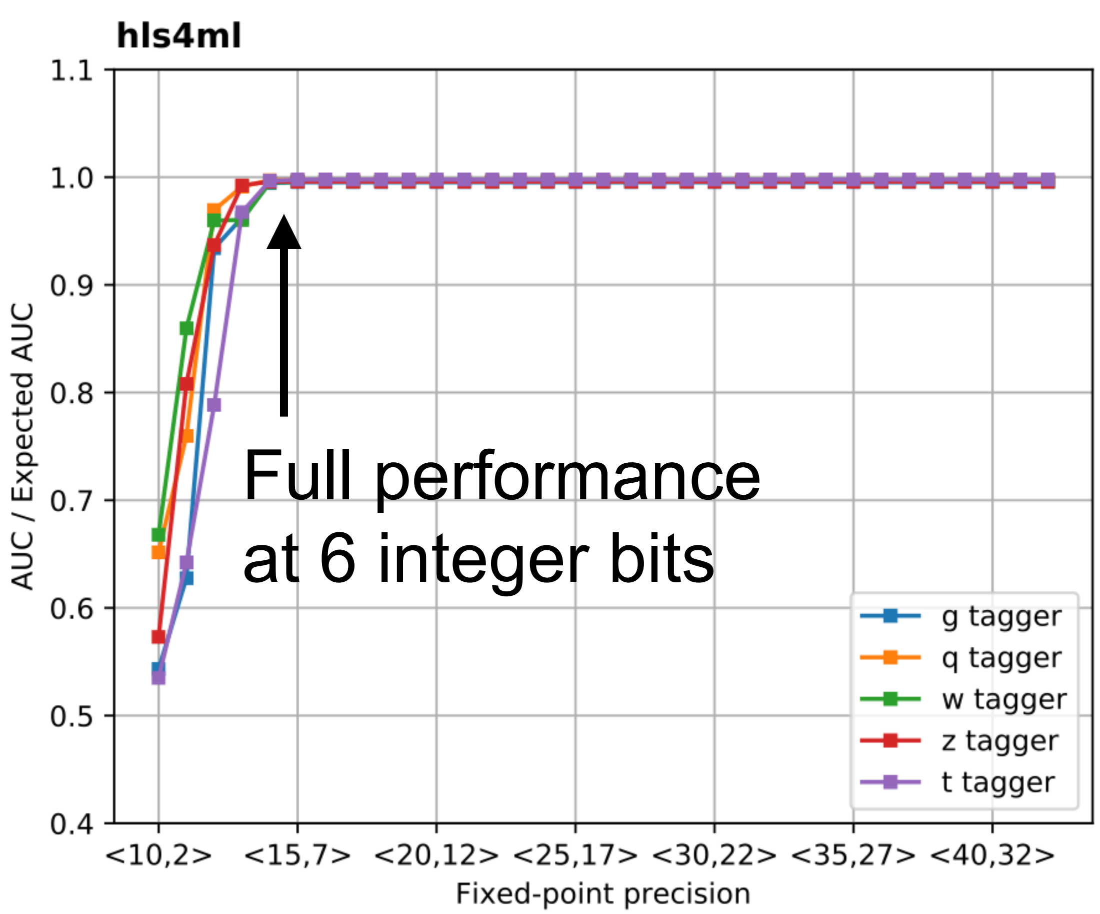
ap_fixed<width bits, integer bits>

`0101.1011101010`



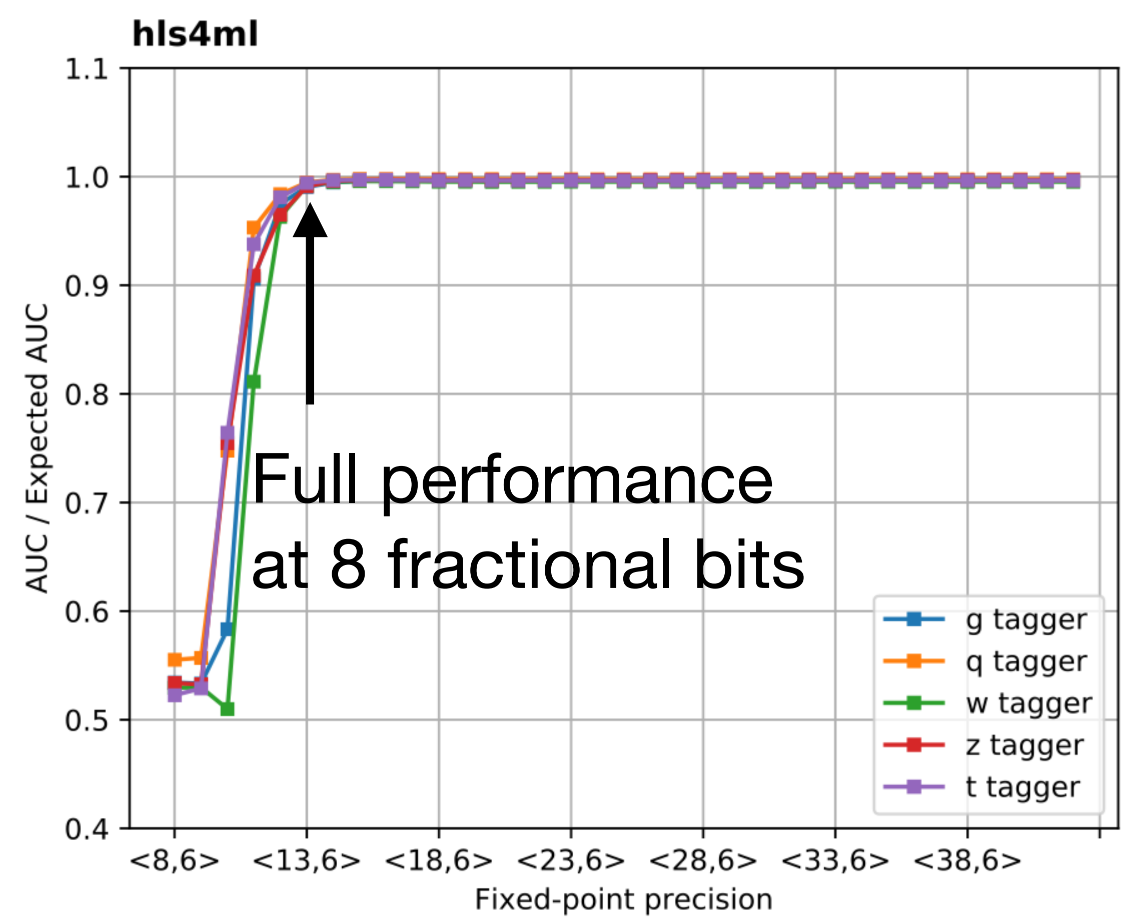
Scan integer bits

Fractional bits fixed to 8

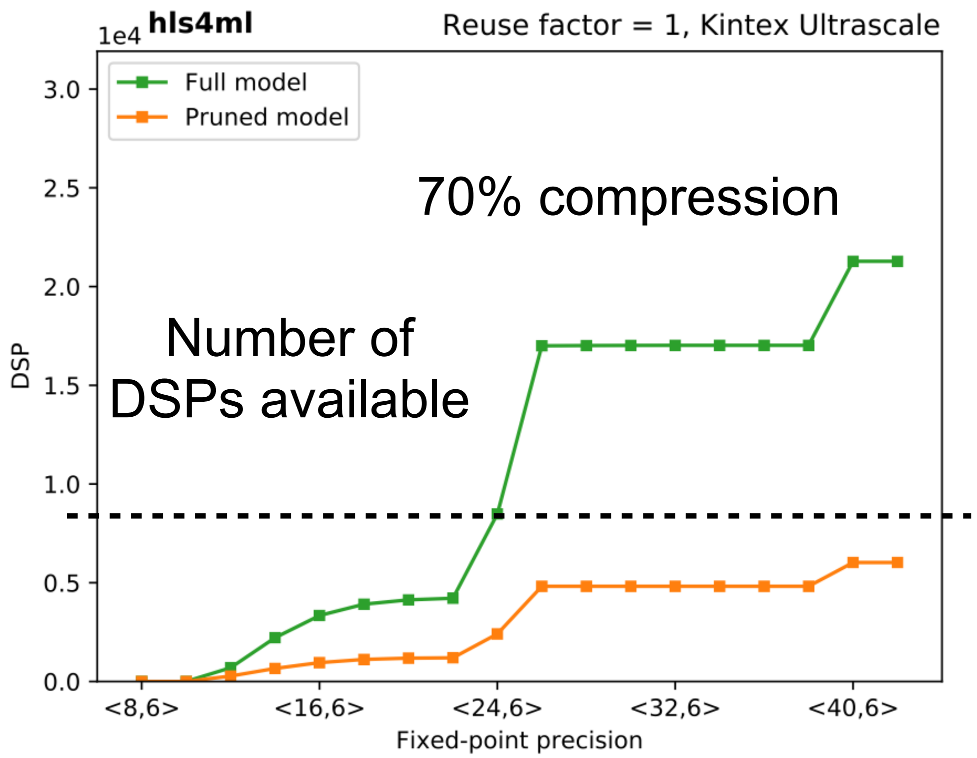


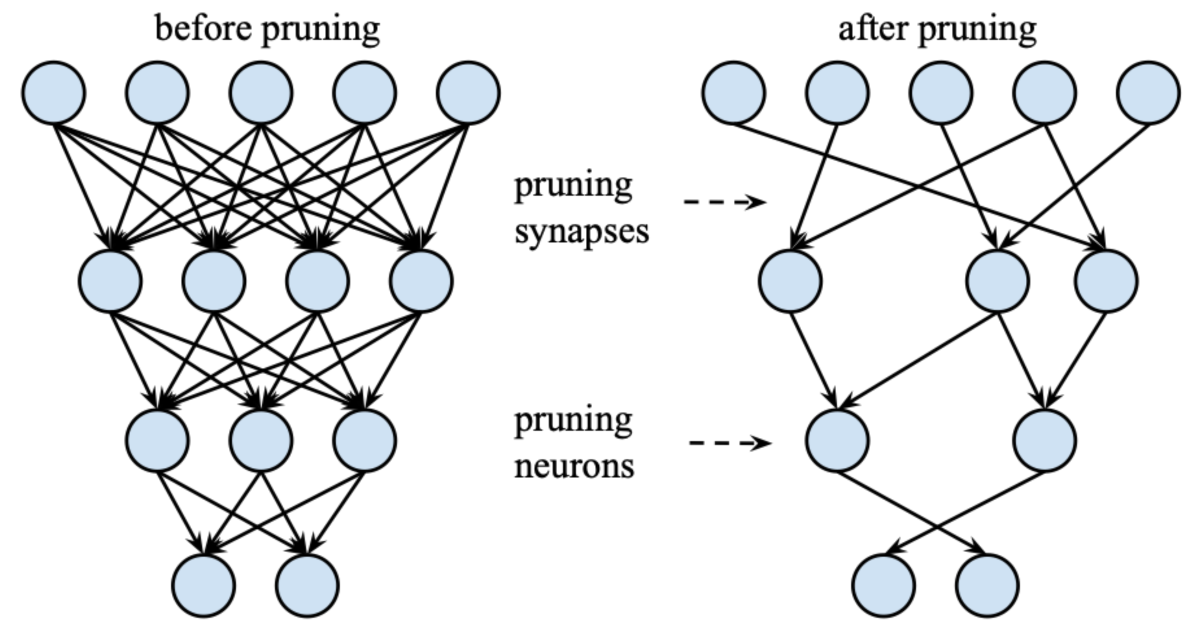
Scan fractional bits

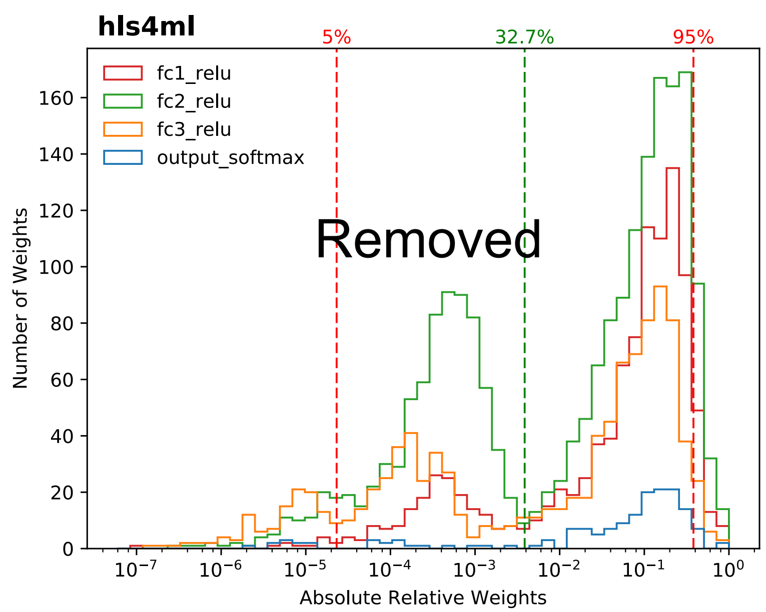
Integer bits fixed to 6



Pruning (Network compression)







- DSPs (used for multiplication) are often limiting resource
  - maximum use when fully parallelized
  - DSPs have a max size for input (e.g. 27x18 bits), so number of DSPs per multiplication changes with precision
- Compression with L1 norm penalty term: penalizes small weights

Trade-off Between Efficiency and Accuracy

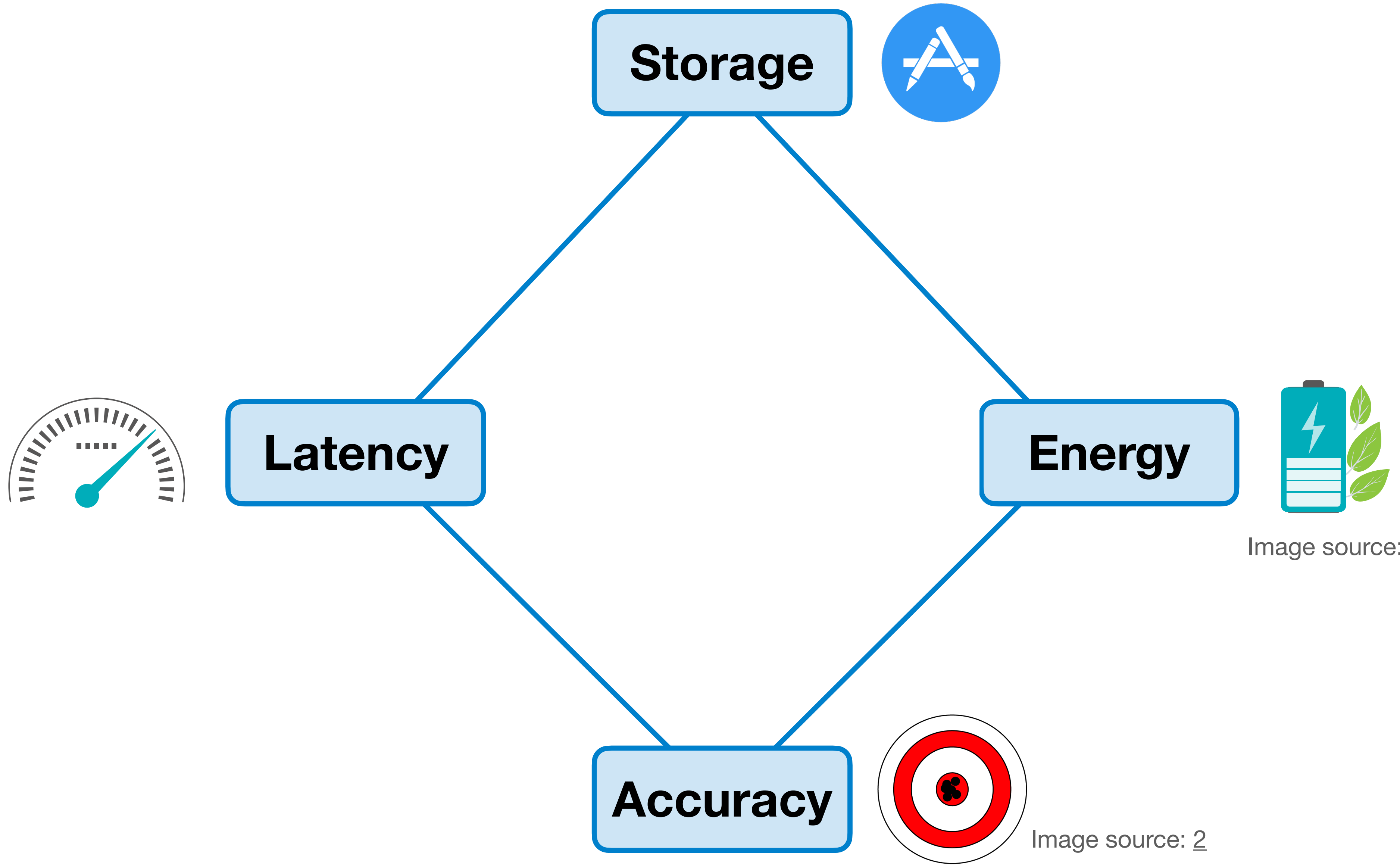
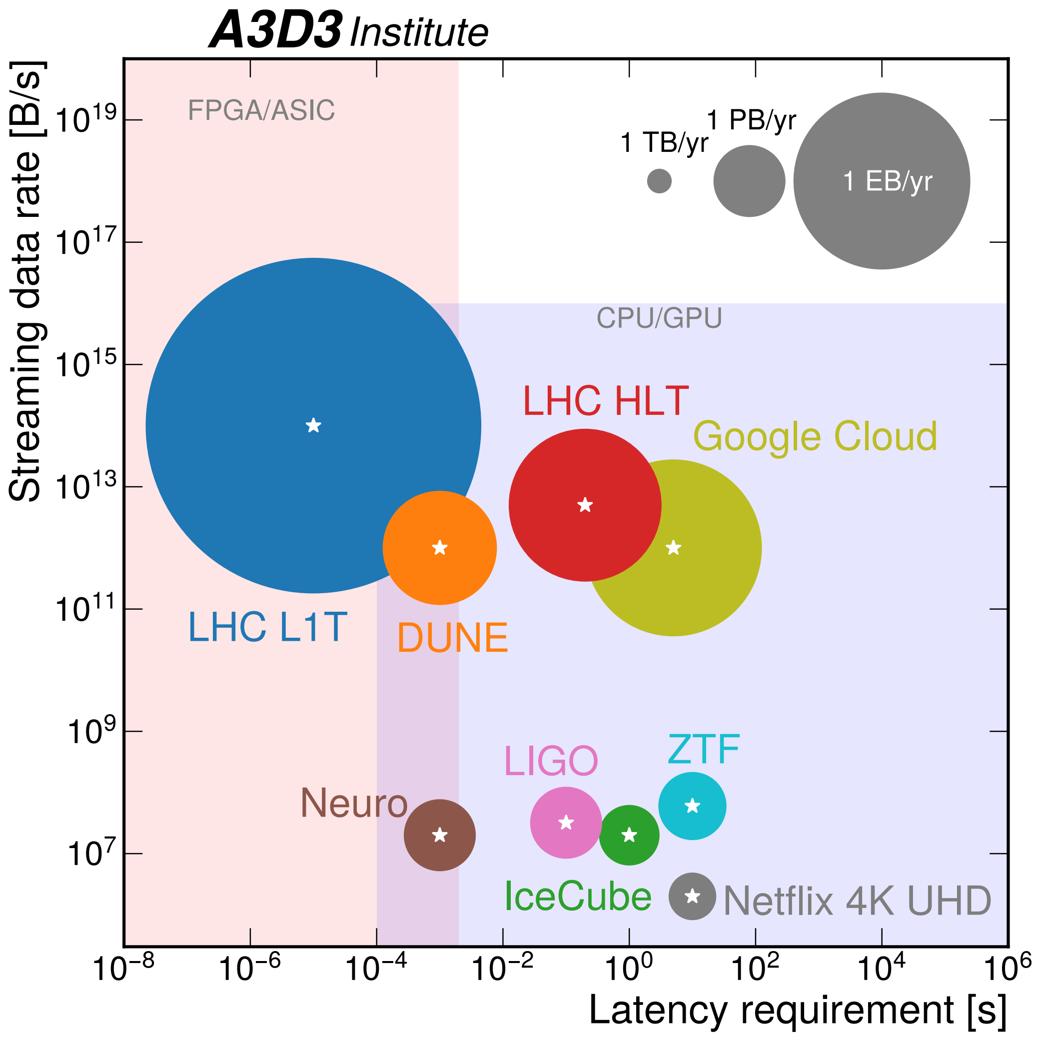


Image source: 1

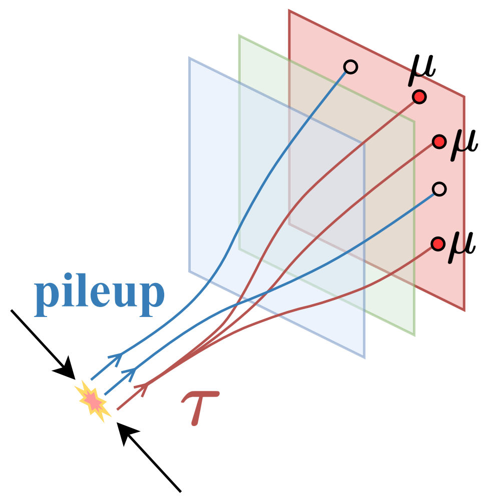
Image source: 2

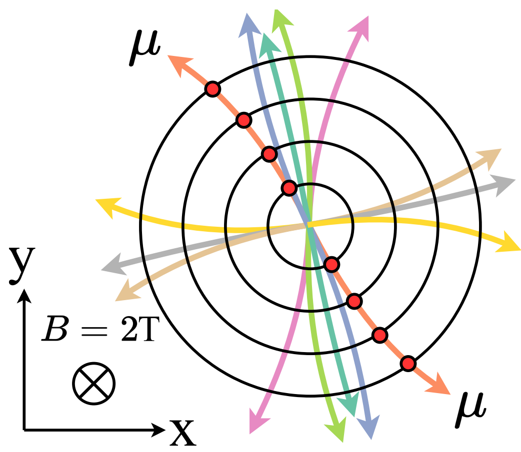
Summary

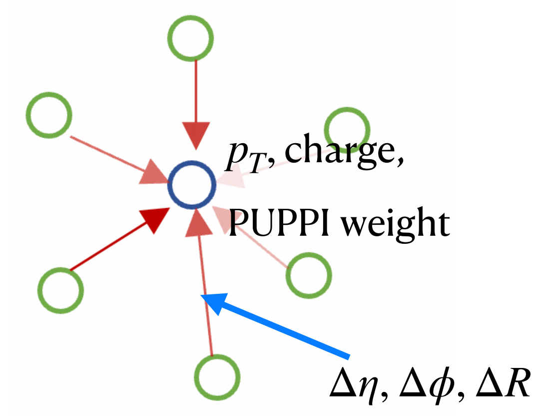


A3D3 Institute
Streaming data rate [B/s]
Latency requirement [s]
FPGA/ASIC
CPU/GPU
1 TB/yr
1 PB/yr
1 EB/yr
LHC L1T
DUNE
LHC HLT
Google Cloud
Neuro
LIGO
IceCube
ZTF
Netflix 4K UHD

Geometric Deep Learning for Particle physics







Elusive New Physics Footprint

Flavor: observed pattern, no underlying symmetry

Collimated, low momentum, impossible to detect in 100 ns

Graph Neural Network: explores the correlations between signal hits on top of background hits

Extensions: Dark photon decaying to sub-GeV Electrons

Contrastive learning with efficient transformers for particle tracking

Learn from data: semisupervised learning

Publications

https://arxiv.org/abs/2203.15823
(JHEP and NeurIPS 2020AI4Science)

https://arxiv.org/abs/2201.12987
(ICML 2022)

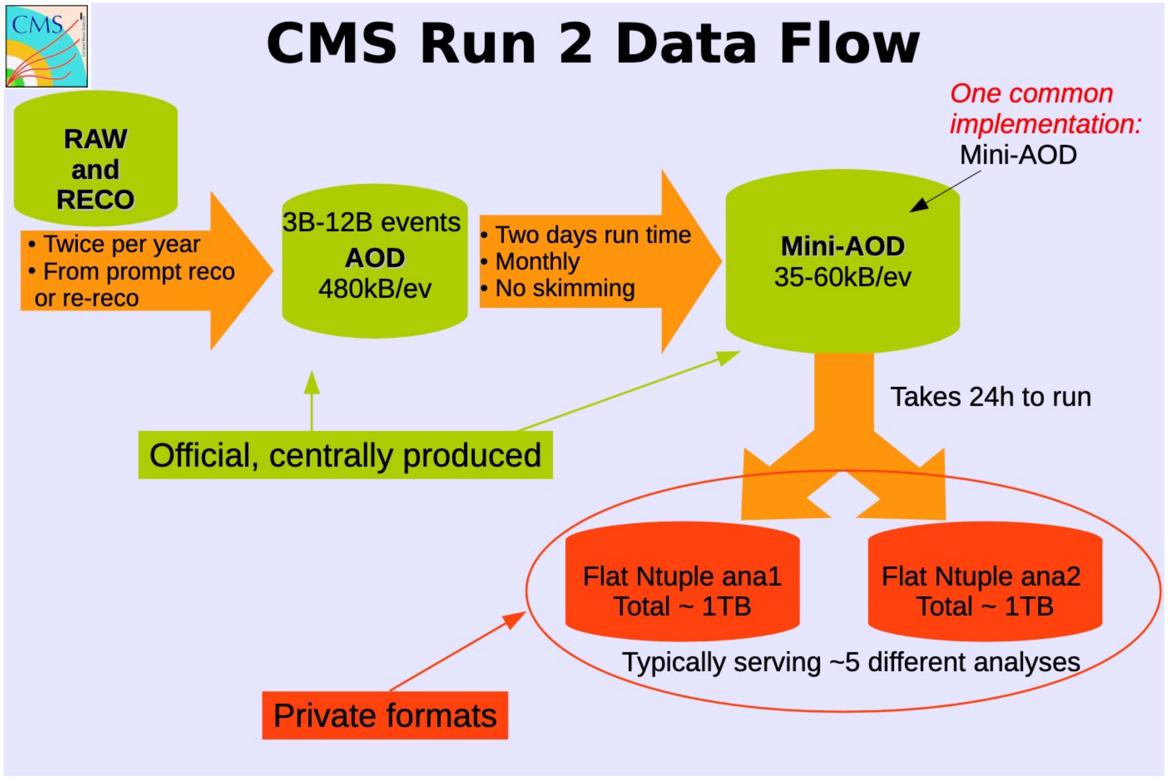
https://arxiv.org/abs/2210.16966
(ICLR 2023 and NeurIPS 2022)

First demonstration in large experiment data processing

[First demonstration of integrating SONIC, with tests at Purdue CMS Tier-2 data center, paper just came out.]

MiniAOD step of the CMS data processing: ML inferences consume 10% of the total processing time

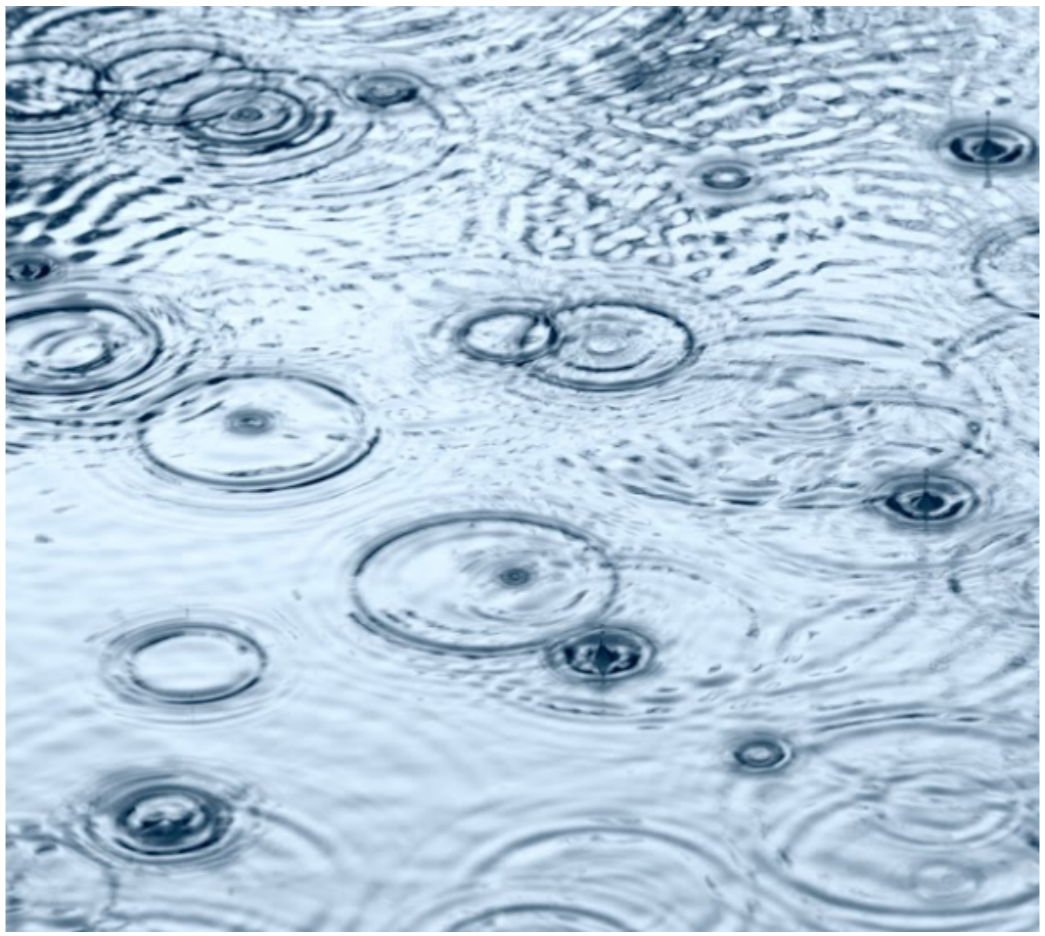
Algorithm	Time [ms]	Fraction [%]	Input [MB]
PN-AK4	42.4	4.3	0.04
PN-AK8	11.4	1.1	0.003
DeepMET	13.2	1.3	0.33
DeepTau	21.1	2.1	1.18
ParticleNet+DeepMET+DeepTau	88.1	8.8	1.55
Total	993.3	100.0	—



Mini-AOD production typically takes about 0.5 seconds per event on production grid nodes

The Fast & Furious

LHC

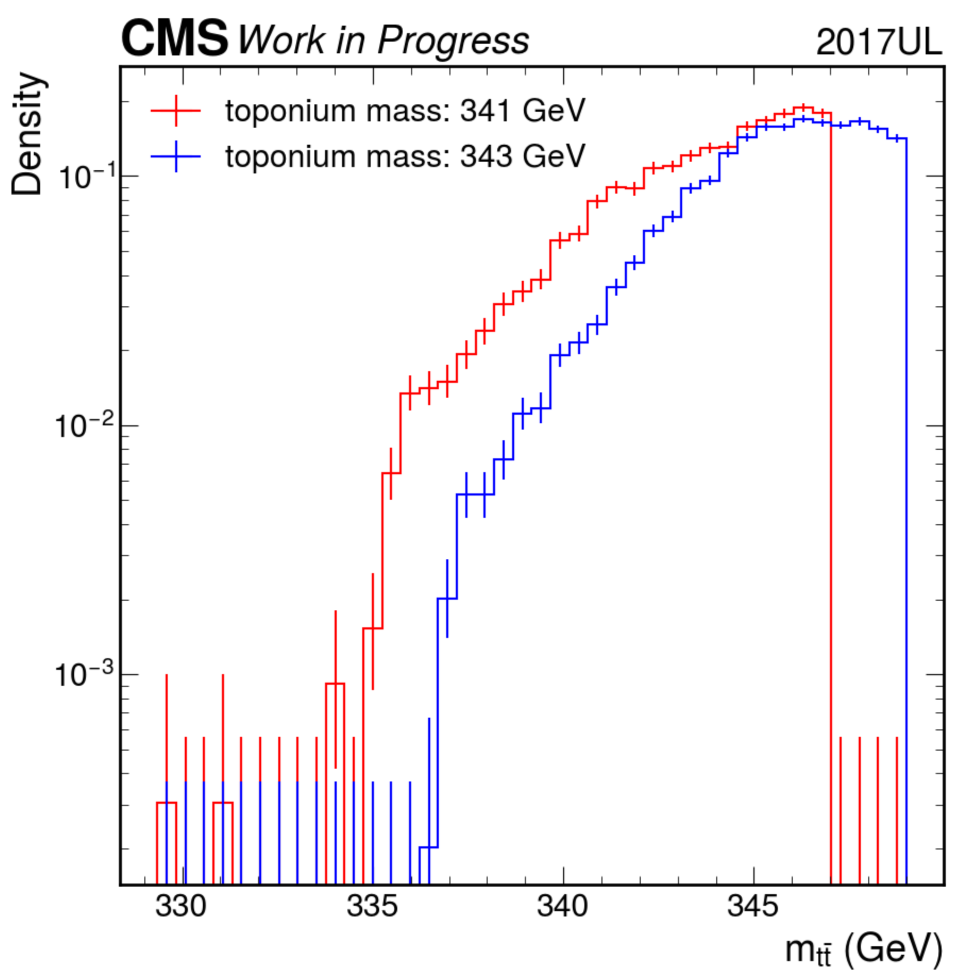


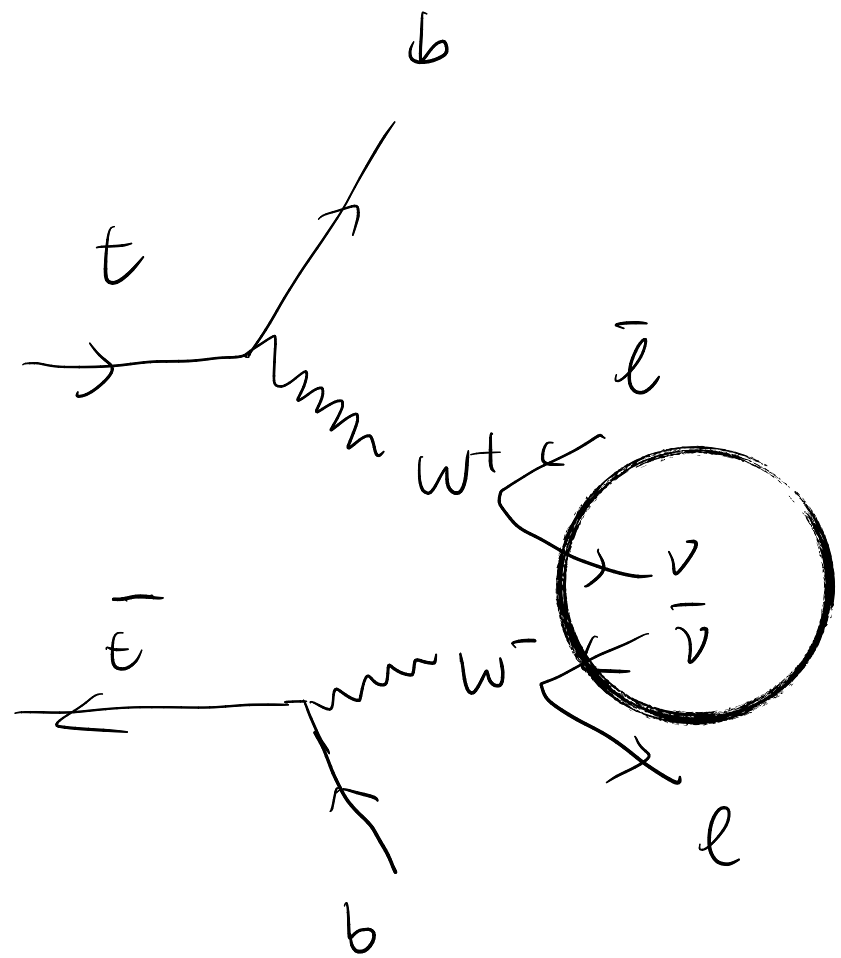
The beams in the LHC are made up of bunches of protons, spaced seven meters (25 ns) apart, with each one containing more than 100 billion protons.

HL-LHC

Higher pileup, fine granularity detectors, advanced algorithms to capture rare new physics

Search for Toponium with Dilepton Events





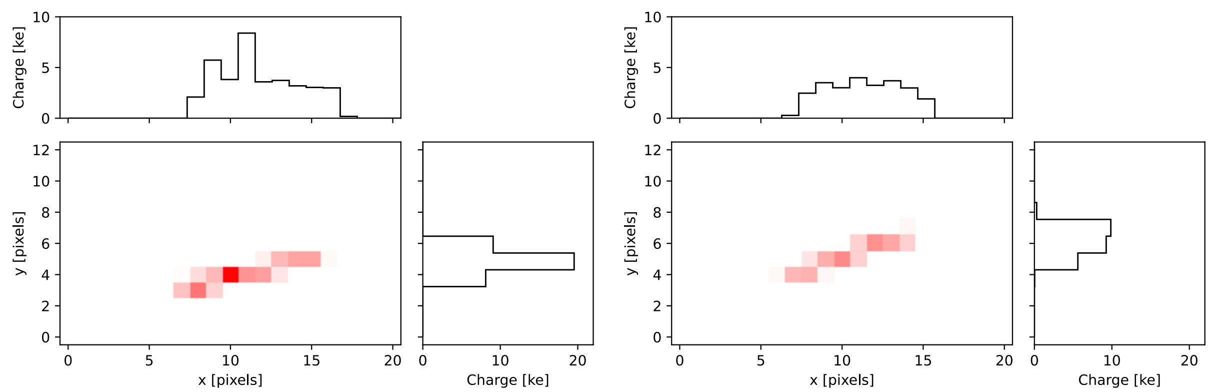
Missing transverse energy

Dilepton channel: two leptons carry spin correlation information of the two top quarks, can be used for toponium and ttbar separation

Many methods have been studied in order to analyze Tevatron/LHC 2l ttbar events. [0603011.pdf,PhysRevLett.80.2063]

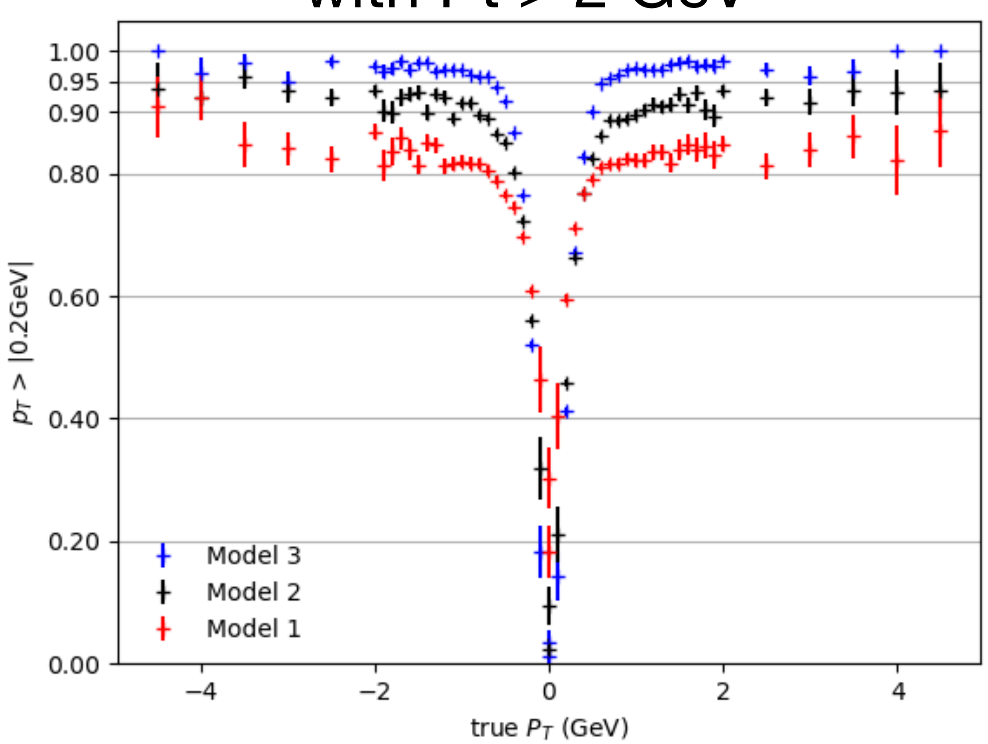
Toponium mass reconstruction: need good resolution near Mttbar threshold.

SmartPixel

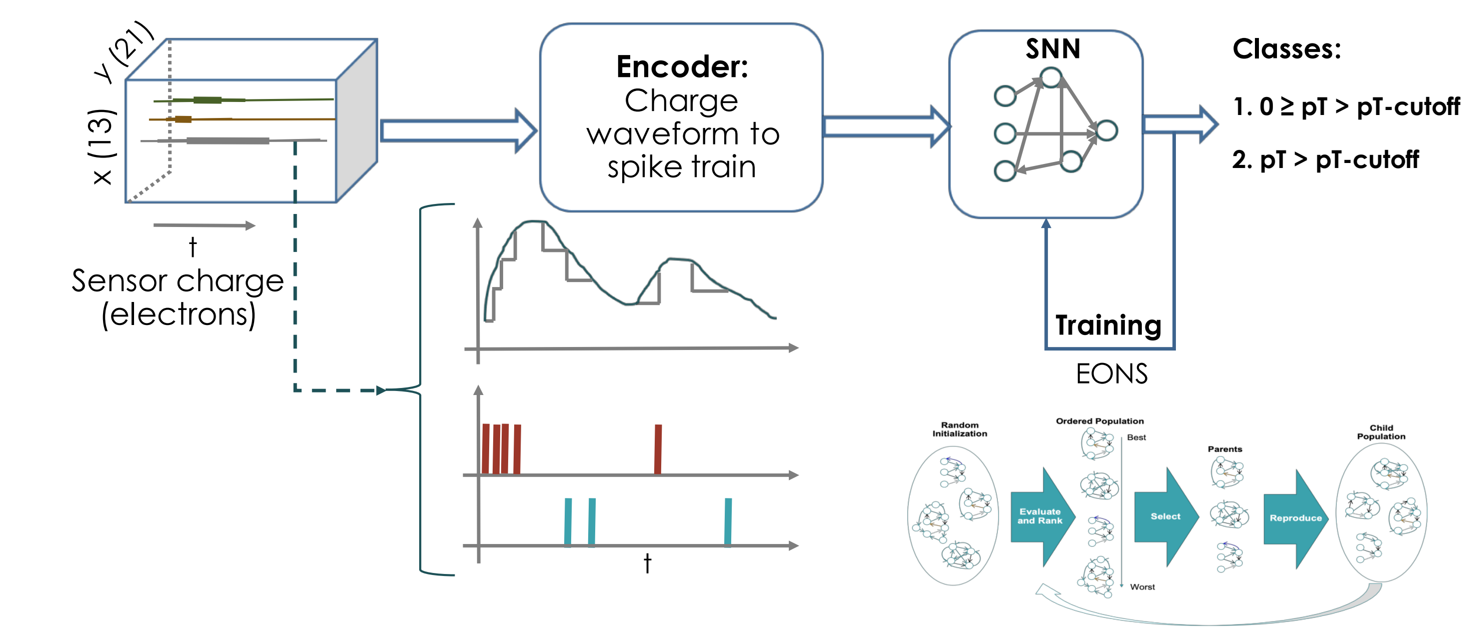


Model	Sig. efficiency	Bkg. rejection
Model 1	84.8 %	26.6 %
Model 2	93.3 %	25.1 %
Model 3	97.6 %	21.7 %

Efficiency of detecting particles with Pt > 2 GeV



Neuromorphic Approach: Processing of Pixel Arrays as Spikes with SNNs



https://physics.paperswithcode.com/paper/smart-pixel-sensors-towards-on-sensor