Fast Machine Learning Inference

Elham E Khoda
University of California, San Diego

Accelerated AI Algorithms for Data-Driven Discovery

FCC Workshop 2024
June 13, 2024

Thanks!

Many thanks to

Javier Duarte (UCSD)

for helping me prepare the slides

Lindsey Gray, Jennet Dickinson, Nhan Tran (Fermilab), Shih-Chieh Hsu (UW), Dylan Rankin (Penn)

for helping with inputs for the presentation

Introduction

Introduction

- Machine learning has already changed the way we do particle physics *from trigger/data acquisition to event reconstruction, simulation, data analysis, and interpretation*

Introduction

Snowmass CompF03 Report, arXiv:2209.07559

- Machine learning has already changed the way we do particle physics *from trigger/data acquisition to event reconstruction, simulation, data analysis, and interpretation*
  - It is an essential and versatile tool that we use to improve existing approaches

Introduction

Snowmass CompF03 Report, [arXiv:2209.07559]

- Machine learning has already changed the way we do particle physics *from trigger/data acquisition to event reconstruction, simulation, data analysis, and interpretation*
    - It is an essential and versatile tool that we use to improve existing approaches
    - It enables fundamentally new approaches

Introduction

Snowmass CompF03 Report, arXiv:2209.07559

- Machine learning has already changed the way we do particle physics *from trigger/data acquisition to event reconstruction, simulation, data analysis, and interpretation*
    - It is an essential and versatile tool that we use to improve existing approaches
    - It enables fundamentally new approaches

Introduction

- Machine learning has already changed the way we do particle physics *from trigger/data acquisition to event reconstruction, simulation, data analysis, and interpretation*
  - It is an essential and versatile tool that we use to improve existing approaches
  - It enables fundamentally new approaches

- In this talk, I'll focus on fast inference of ML and how they can shift the paradigm

Computing Hardware

Image: Microsoft

Computing Hardware

Image: [Microsoft]

Computing Hardware

Image: [Microsoft]

Computing Hardware

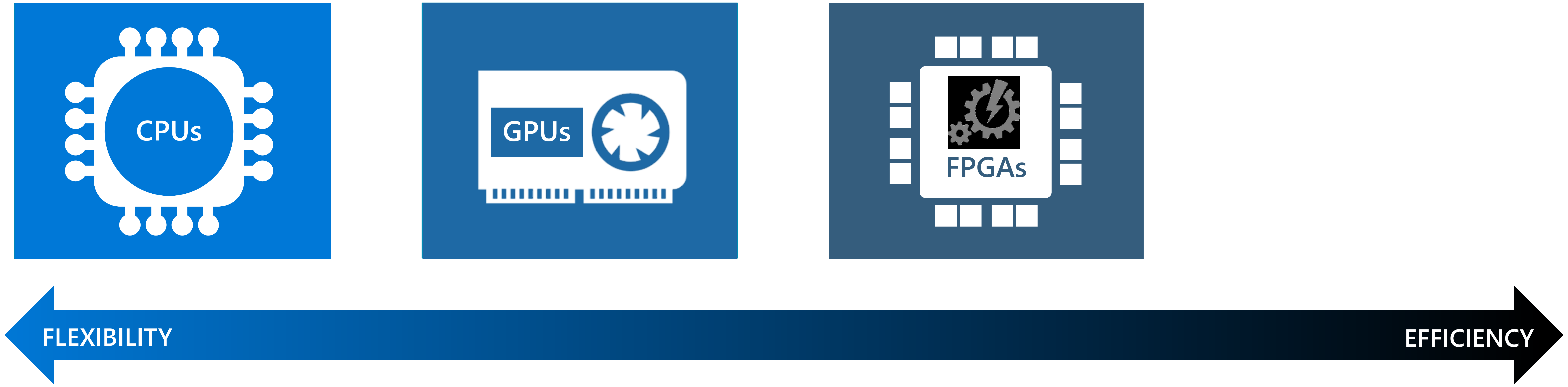


Image: Microsoft

Computing Hardware

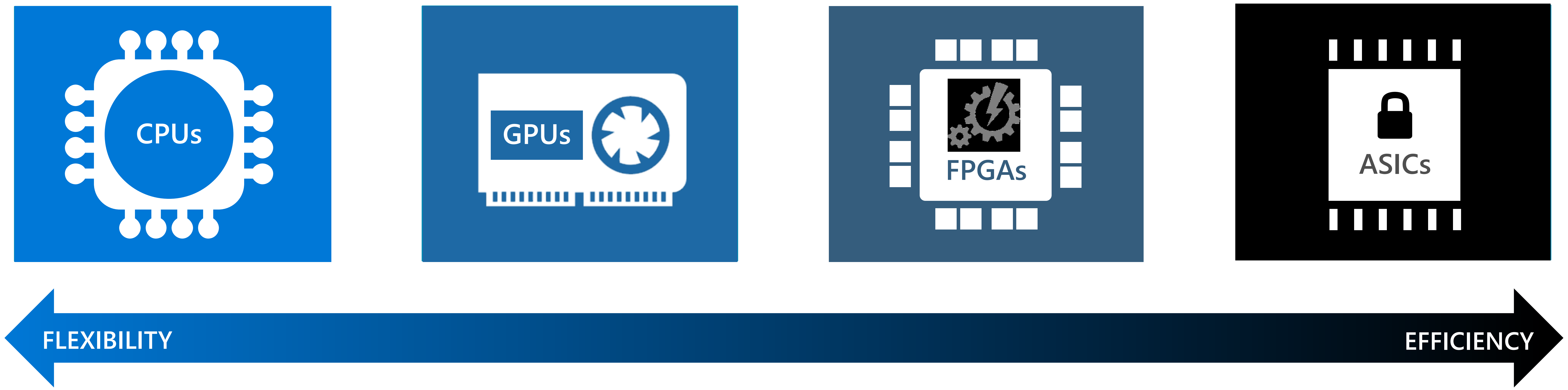


Image: Microsoft

Computing Hardware

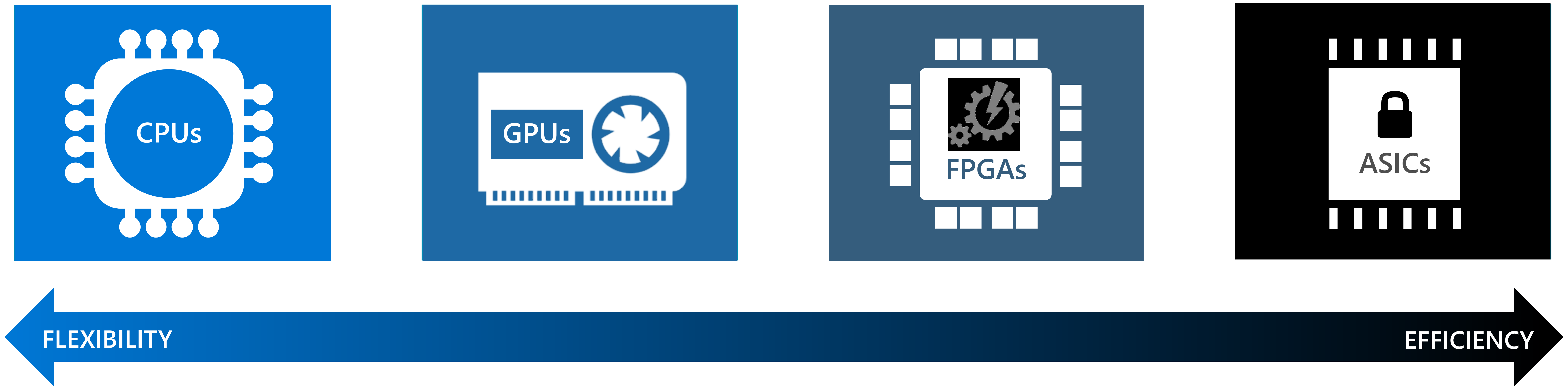


Image: [Microsoft]

Computing Hardware

First stage of
LHC *trigger*

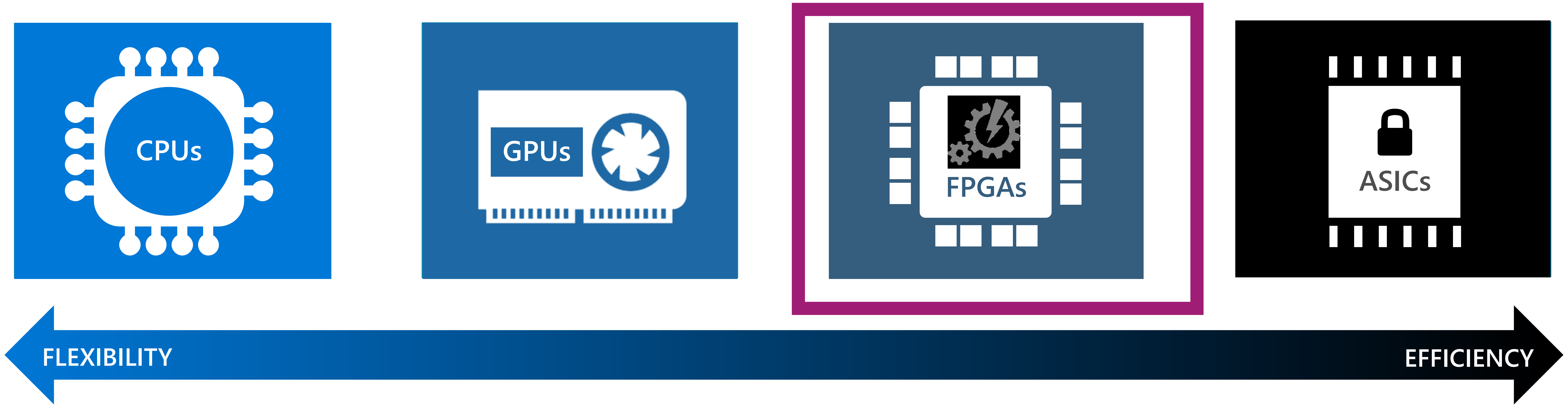


Image: [Microsoft]

Computing Hardware

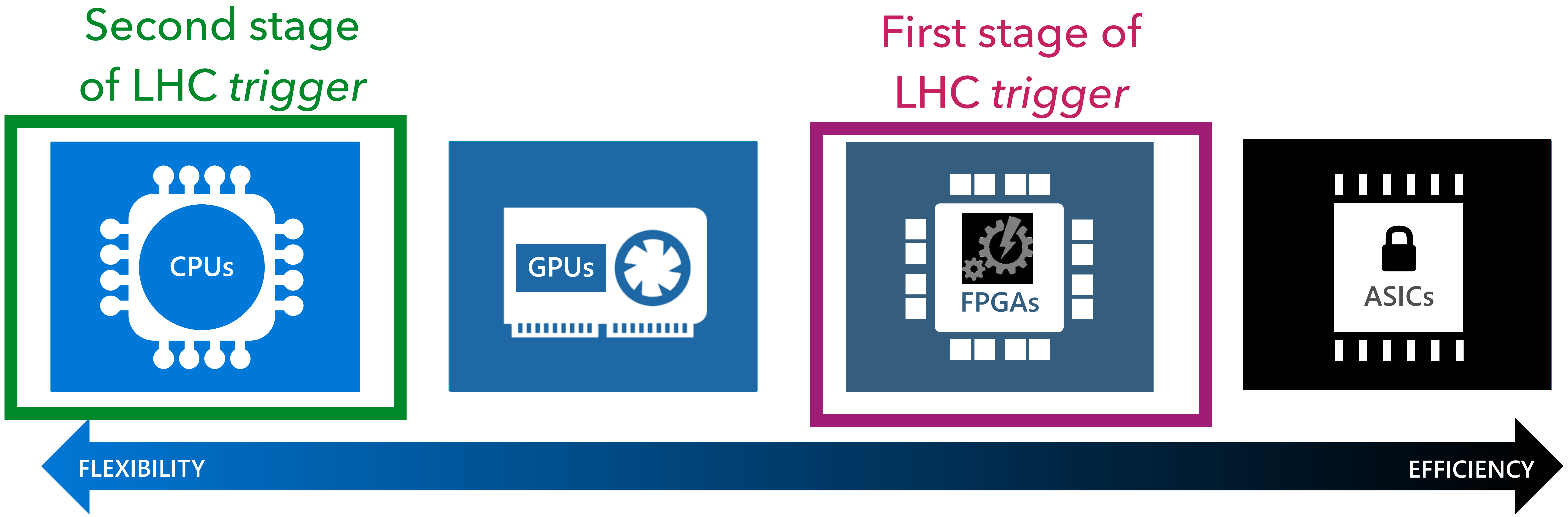
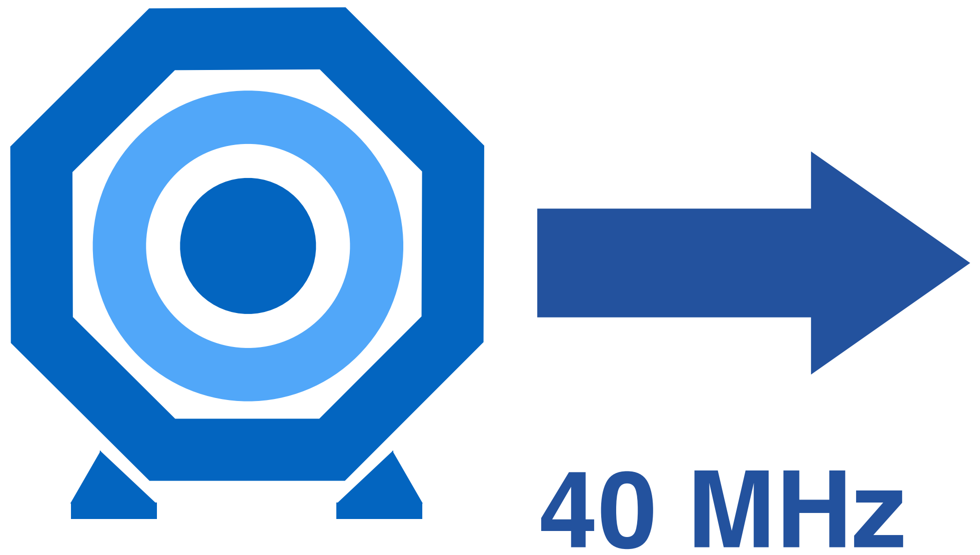


Image: [Microsoft]

HL-LHC Data Processing

Compute Latency

1 ns — 1 μs — 1 ms — 1 s



40 MHz

ASICs

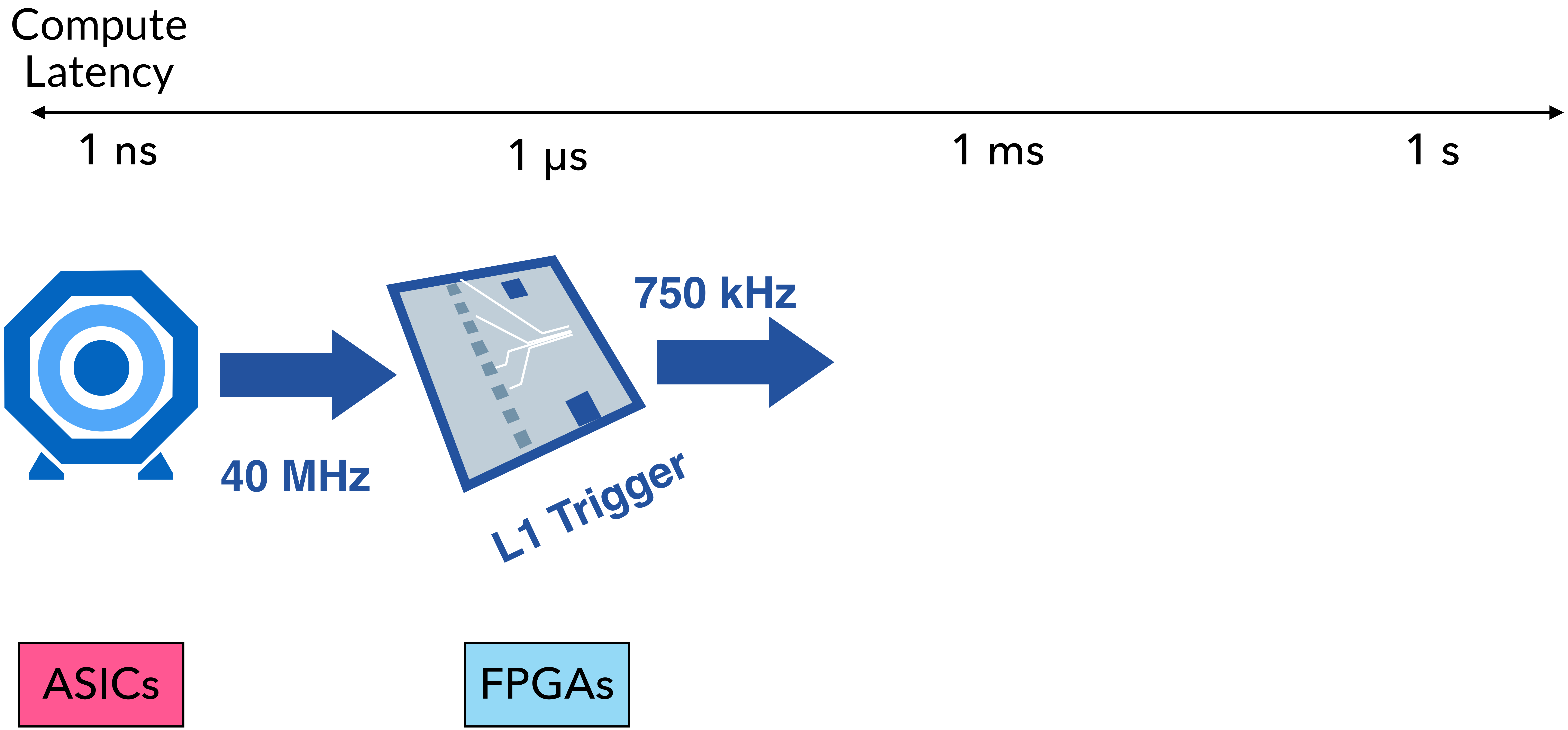
Challenges:

Each collision produces O(10^3) particles

The detectors have O(10^8) sensors

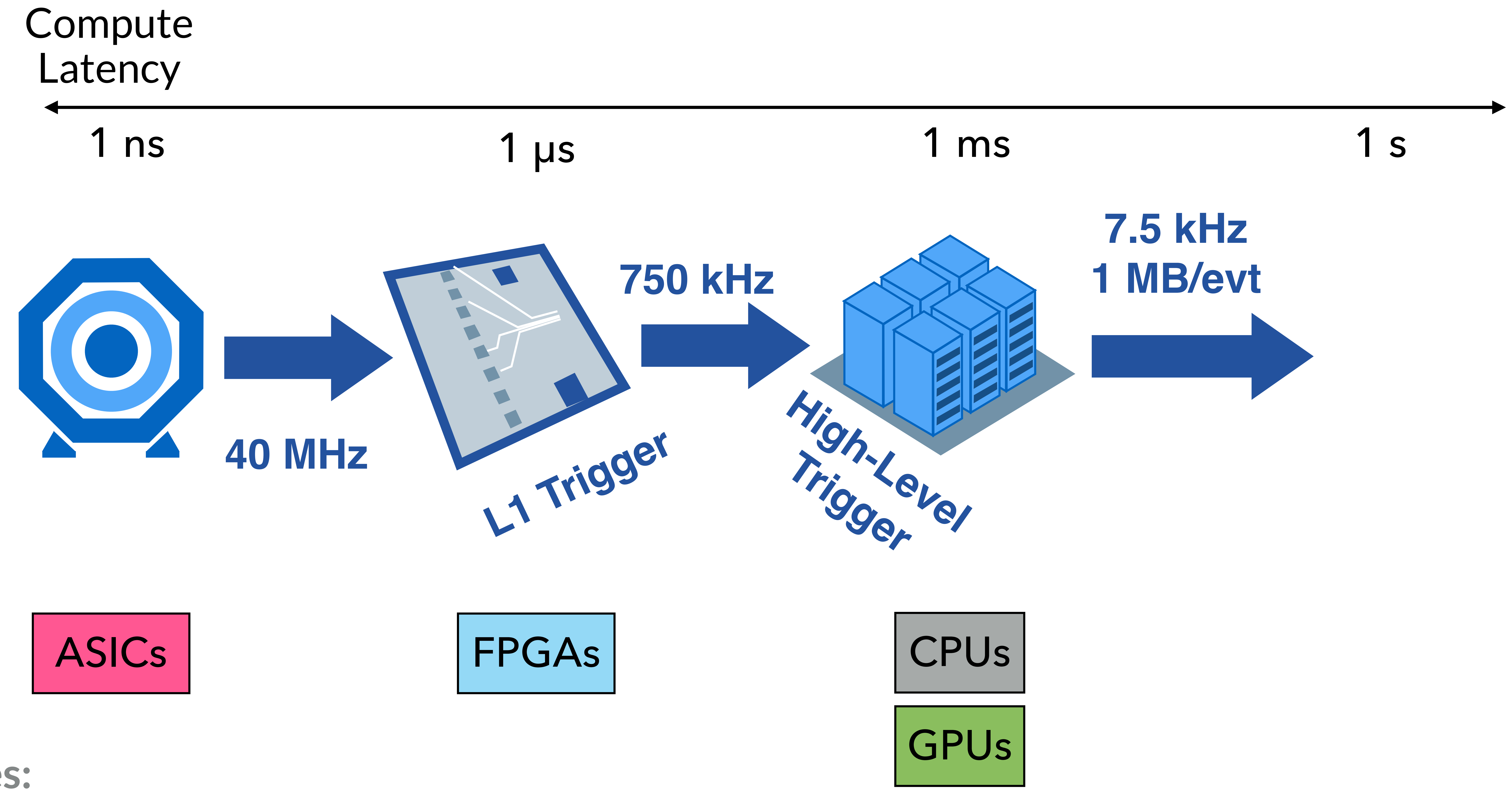
Extreme data rates of O(100 TB/s)

HL-LHC Data Processing



Challenges:
Each collision produces $O(10^3)$ particles
The detectors have $O(10^8)$ sensors
Extreme data rates of O(100 TB/s)

HL-LHC Data Processing

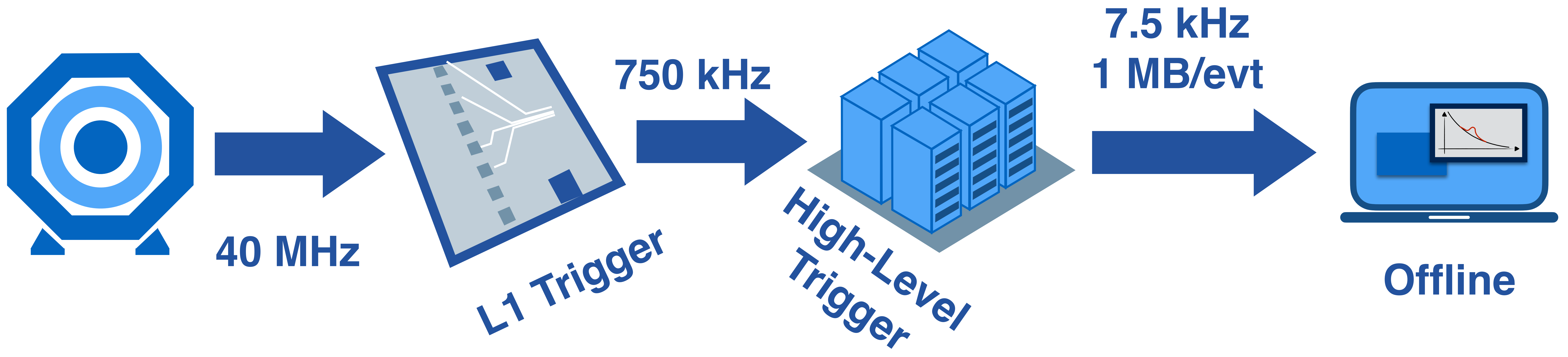


Challenges:
Each collision produces O(10^3) particles
The detectors have O(10^8) sensors
Extreme data rates of O(100 TB/s)

HL-LHC Data Processing

Compute Latency

1 ns — 1 μs — 1 ms — 1 s



1 ns	1 μs	1 ms	1 s
ASICs	FPGAs	CPUs	CPUs
		GPUs	GPUs
		FPGAs	FPGAs
		Other processors: IPU, TPU ..	Other processors: IPU, TPU ..

Exabyte-scale datasets

Challenges:

Each collision produces O(10^3) particles

The detectors have O(10^8) sensors

Extreme data rates of O(100 TB/s)

Simplified HL-LHC Trigger

CMS-TDR-021

Simplified HL-LHC Trigger

Trigger	Threshold [GeV]

Thresholds set by backgrounds, limited resolution @ L1, and rate budget

[CMS-TDR-021](CMS-TDR-021)

Simplified HL-LHC Trigger

Trigger	Threshold [GeV]

- Single/double/triple muons/electrons

Thresholds set by backgrounds, limited resolution @ L1, and rate budget

CMS-TDR-021

Simplified HL-LHC Trigger

- Single/double/triple muons/electrons

Trigger	Threshold [GeV]
1 μ	22
2 μ	15, 7
3 μ	5, 3, 3
1 e	36
2 e	25, 12

Thresholds set by backgrounds, limited resolution @ L1, and rate budget

CMS-TDR-021

Simplified HL-LHC Trigger

- Single/double/triple muons/electrons
- Photons

Trigger	Threshold [GeV]
1 μ	22
2 μ	15, 7
3 μ	5, 3, 3
1 e	36
2 e	25, 12
1 γ	36
2 γ	22, 12

Thresholds set by backgrounds, limited resolution @ L1, and rate budget

CMS-TDR-021

Simplified HL-LHC Trigger

- Single/double/triple muons/electrons
- Photons
- Taus

Trigger	Threshold [GeV]
1 μ	22
2 μ	15, 7
3 μ	5, 3, 3
1 e	36
2 e	25, 12
1 γ	36
2 γ	22, 12
1 τ	150
2 τ	90, 90

Thresholds set by backgrounds, limited resolution @ L1, and rate budget

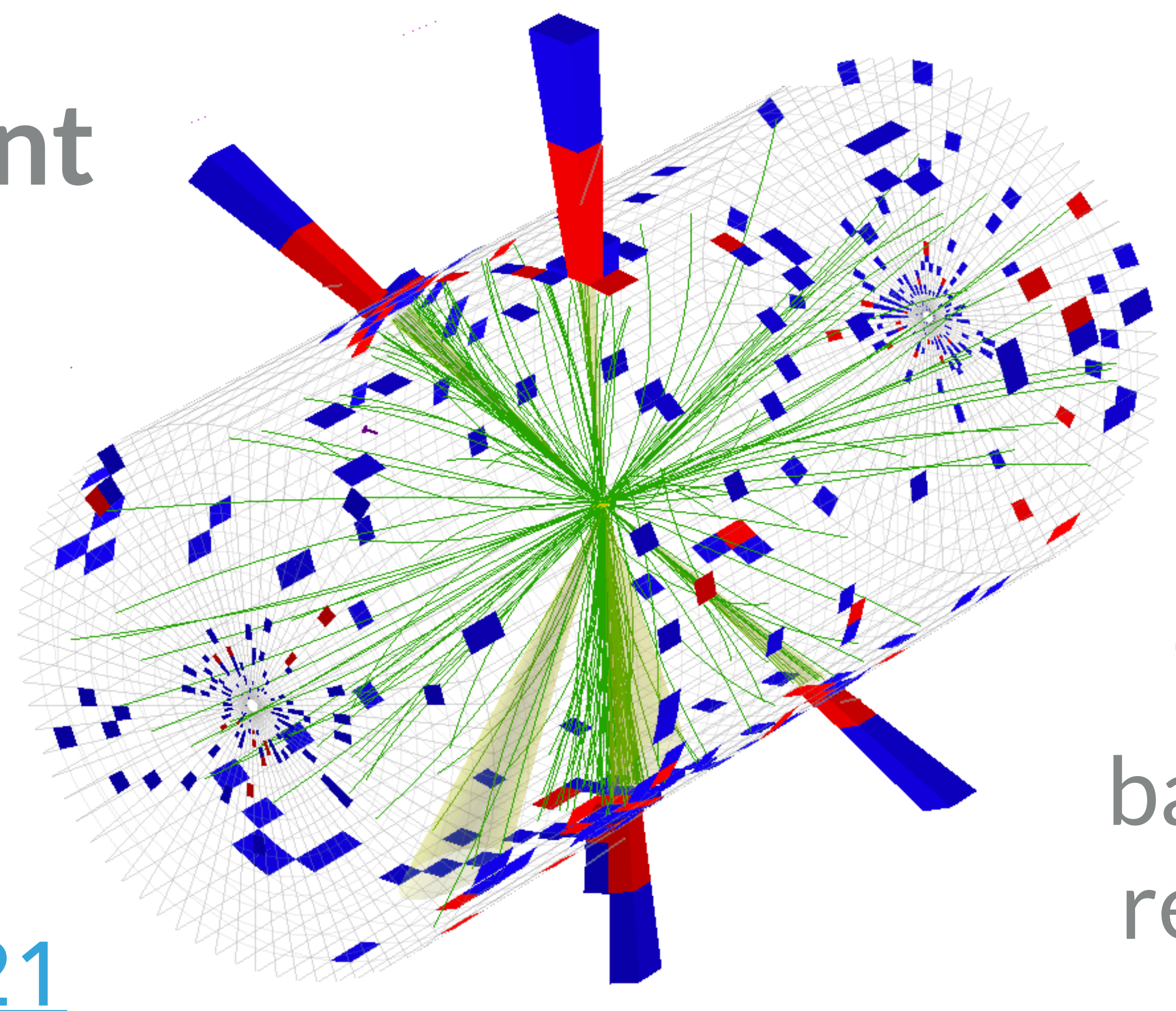
CMS-TDR-021

Simplified HL-LHC Trigger

- Single/double/triple muons/electrons
- Photons
- Taus
- Hadronic

Trigger	Threshold [GeV]
1 μ	22
2 μ	15, 7
3 μ	5, 3, 3
1 e	36
2 e	25, 12
1 γ	36
2 γ	22, 12
1 τ	150
2 τ	90, 90
1 jet	180
2 jet	112, 112
H_T	450
4 jet + H_T	75, 55, 40, 40, 400

4-jet event



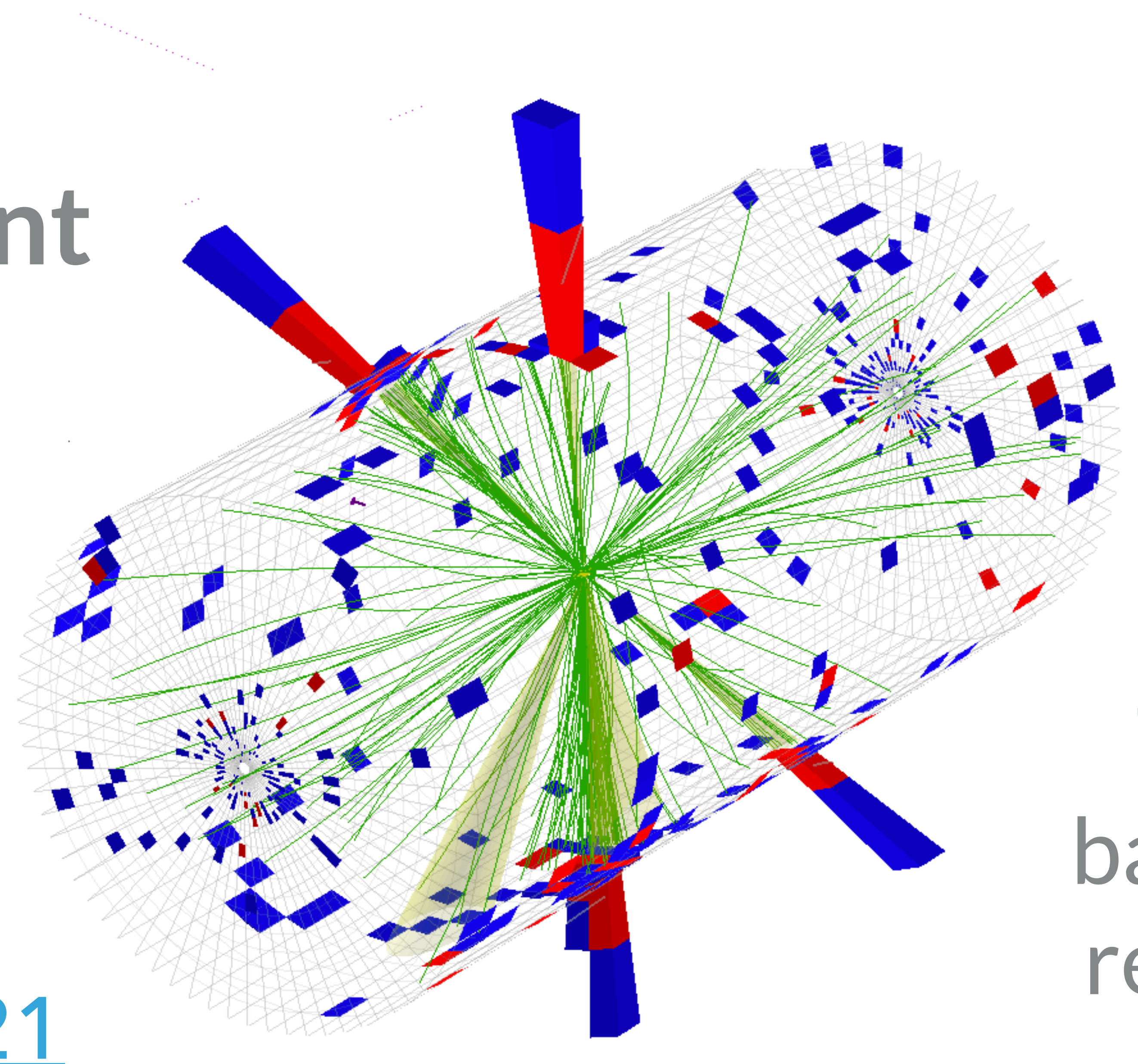
CMS-TDR-021

Thresholds set by backgrounds, limited resolution @ L1, and rate budget

Simplified HL-LHC Trigger

- Single/double/triple muons/electrons
- Photons
- Taus
- Hadronic
- Missing transverse energy

4-jet event



[CMS-TDR-021](CMS-TDR-021)

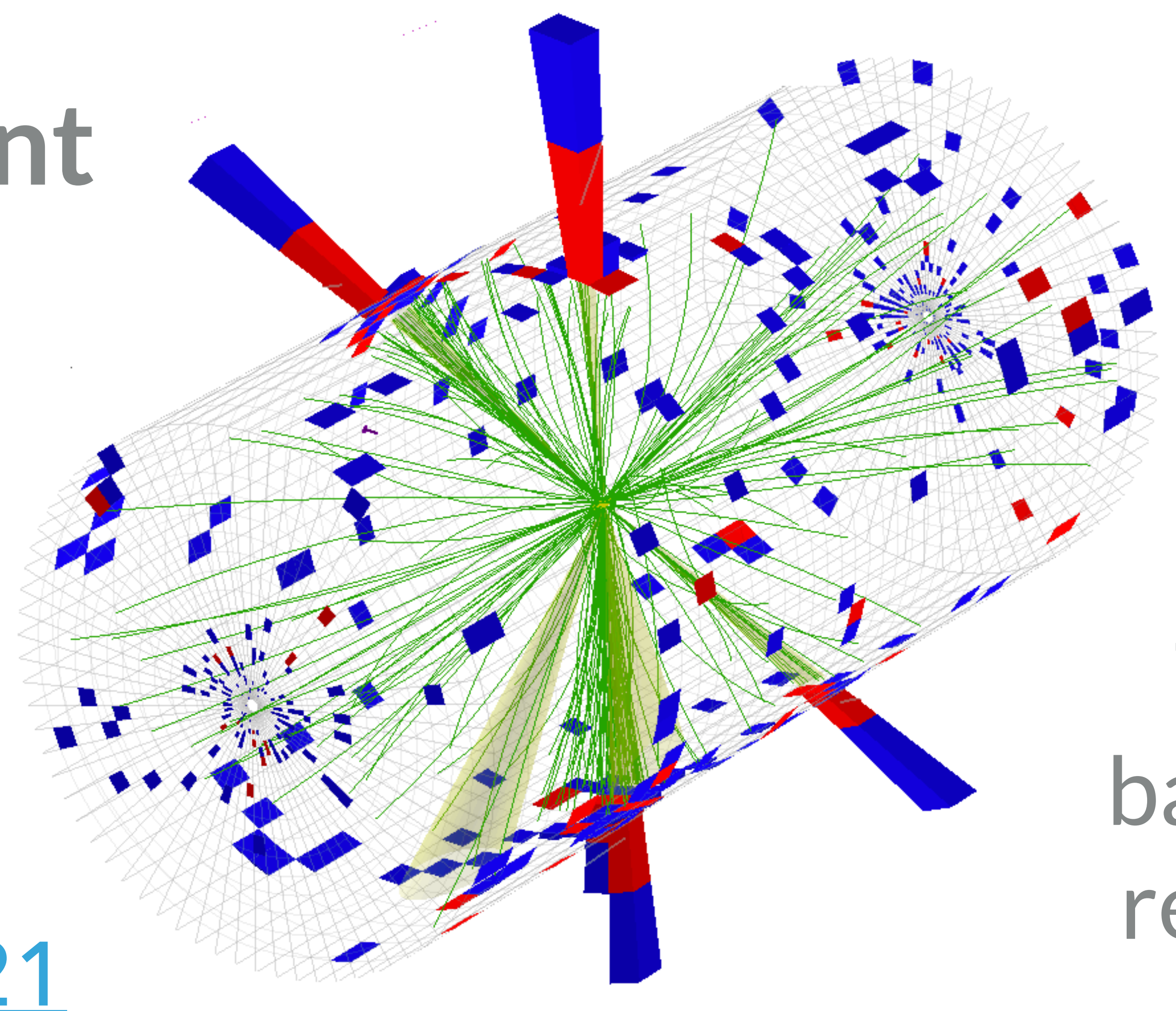
Thresholds set by backgrounds, limited resolution @ L1, and rate budget

Trigger	Threshold [GeV]
1 μ	22
2 μ	15, 7
3 μ	5, 3, 3
1 e	36
2 e	25, 12
1 γ	36
2 γ	22, 12
1 τ	150
2 τ	90, 90
1 jet	180
2 jet	112, 112
H_T	450
4 jet + H_T	75, 55, 40, 40, 400
p_T^{miss}	200

Simplified HL-LHC Trigger

- Single/double/triple muons/electrons
- Photons
- Taus
- Hadronic
- Missing transverse energy
- "Cross" triggers (not shown)

4-jet event



CMS-TDR-021

Thresholds set by backgrounds, limited resolution @ L1, and rate budget

Trigger	Threshold [GeV]
1 μ	22
2 μ	15, 7
3 μ	5, 3, 3
1 e	36
2 e	25, 12
1 γ	36
2 γ	22, 12
1 τ	150
2 τ	90, 90
1 jet	180
2 jet	112, 112
H_T	450
4 jet + H_T	75, 55, 40, 40, 400
p_T^{miss}	200

What could be missing?

What could be missing?

- How can we trigger on more complex low-energy hadronic signatures? Long-lived/displaced particles?

SUEP

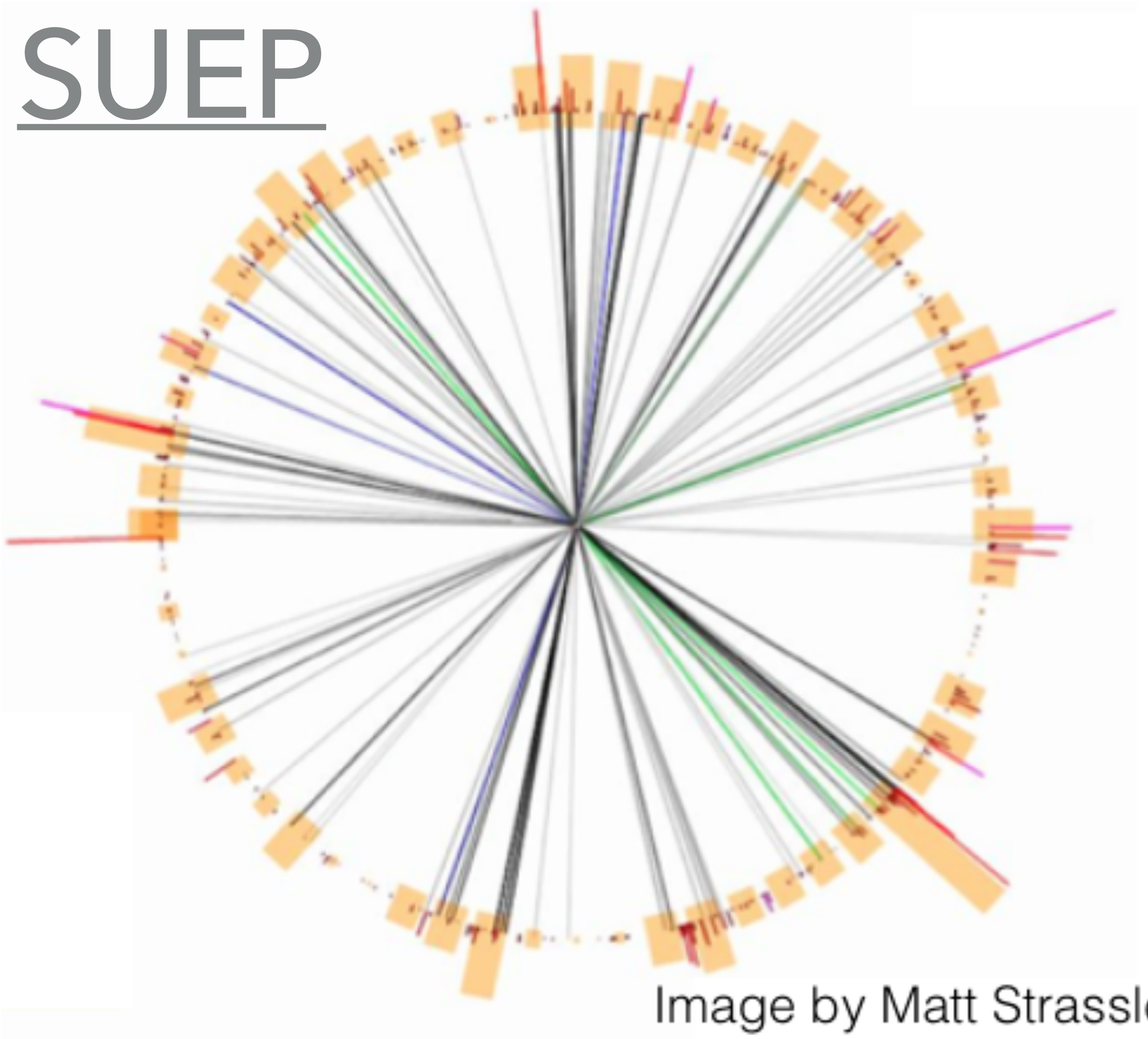


Image by Matt Strassler

What could be missing?

- How can we trigger on more complex low-energy hadronic signatures? Long-lived/displaced particles?
- What if we don't know exactly what to look for?

SUEP

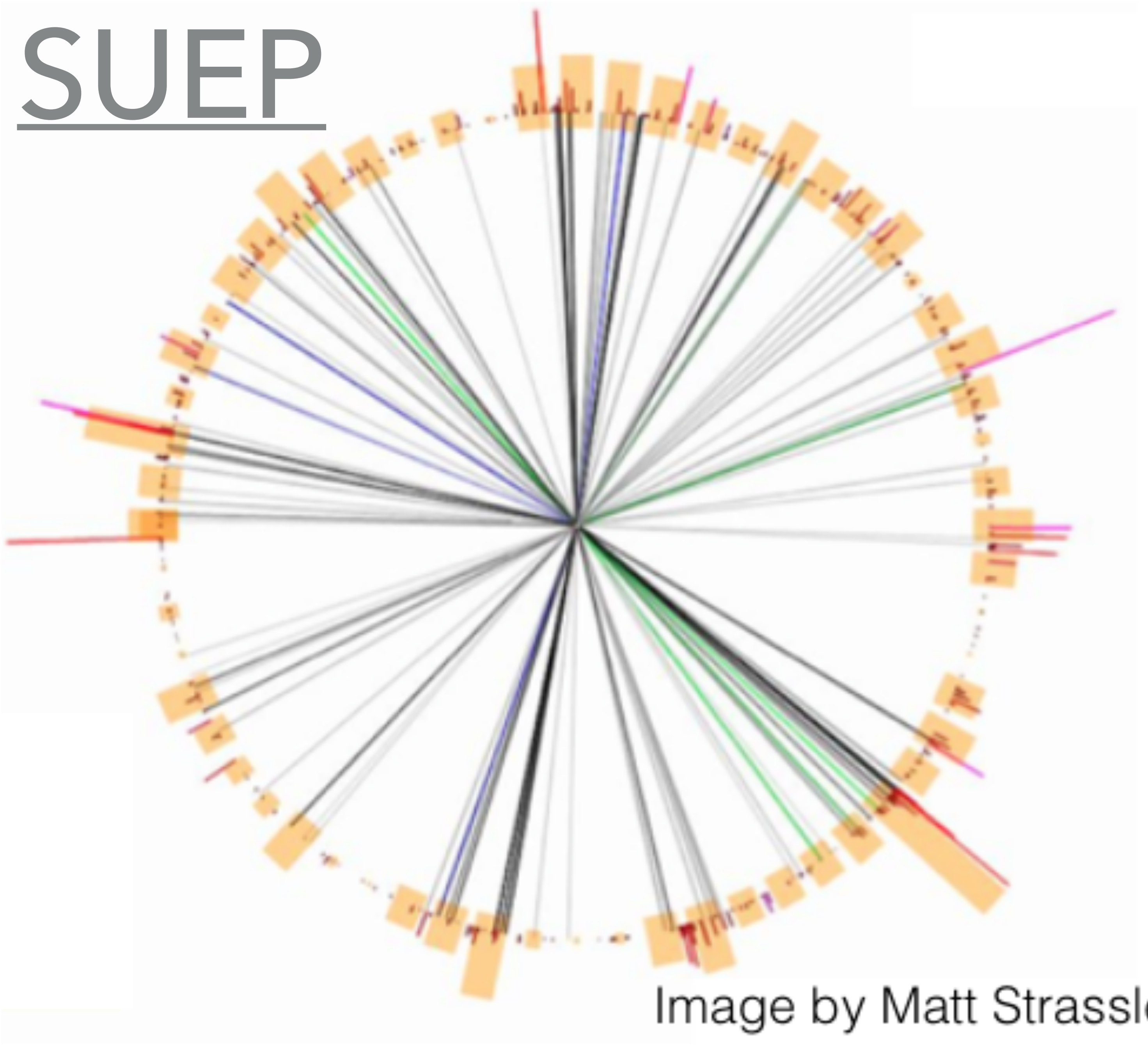
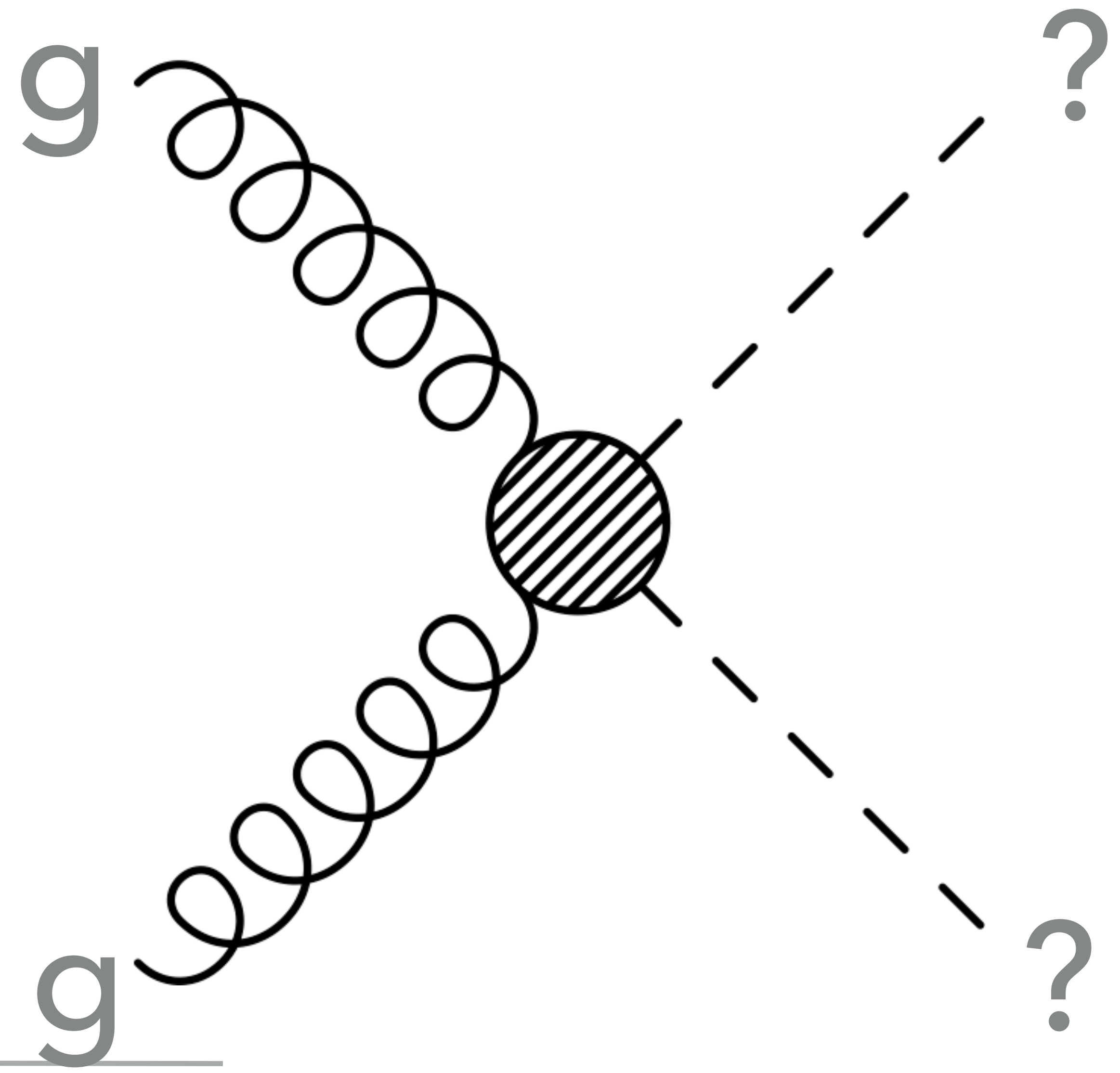


Image by Matt Strassler



What could be missing?

- How can we trigger on more complex low-energy hadronic signatures? Long-lived/displaced particles?
- What if we don't know exactly what to look for?
- What if our signatures require complex multivariate algorithms (e.g. b tagging)?

SUEP

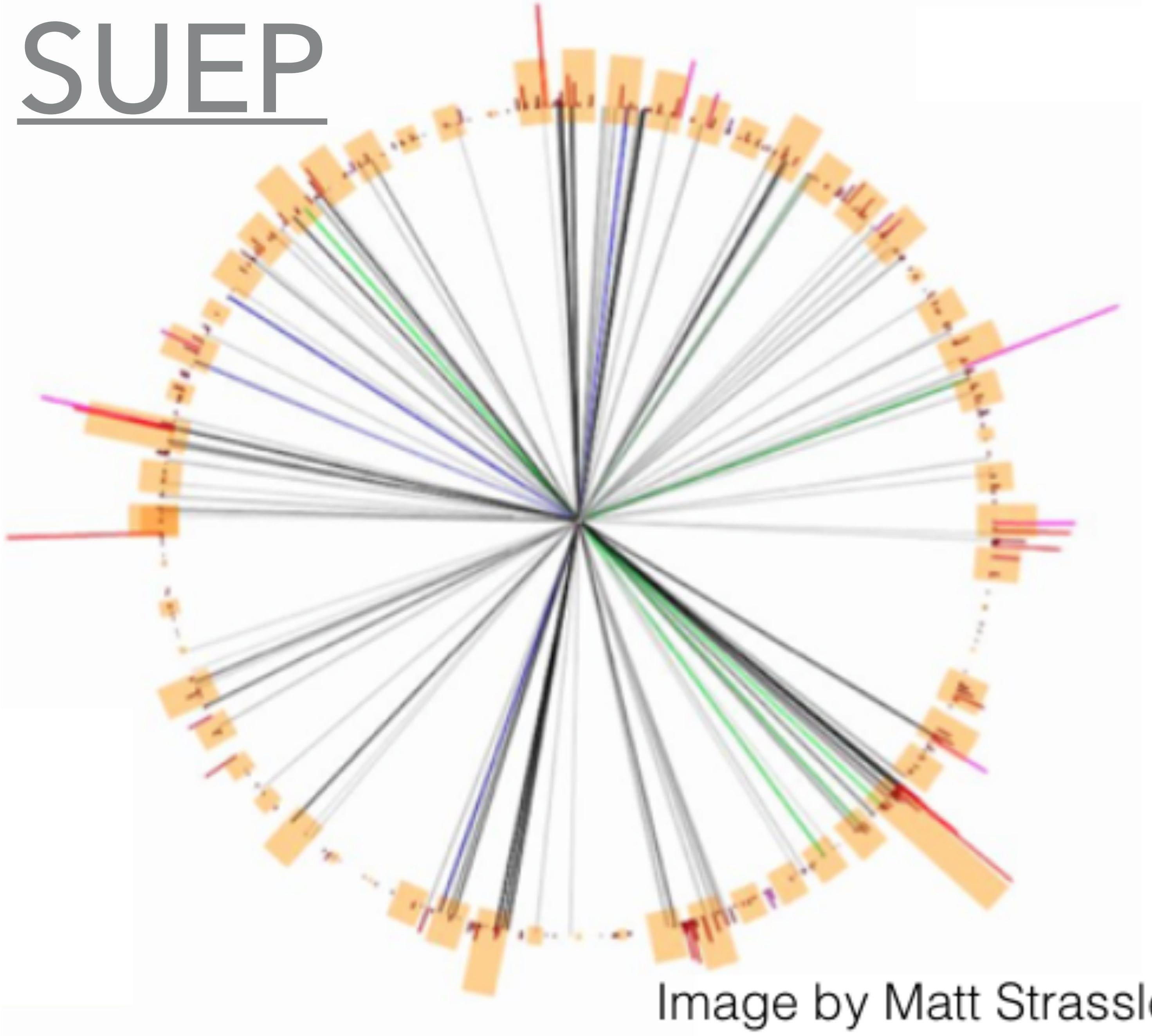
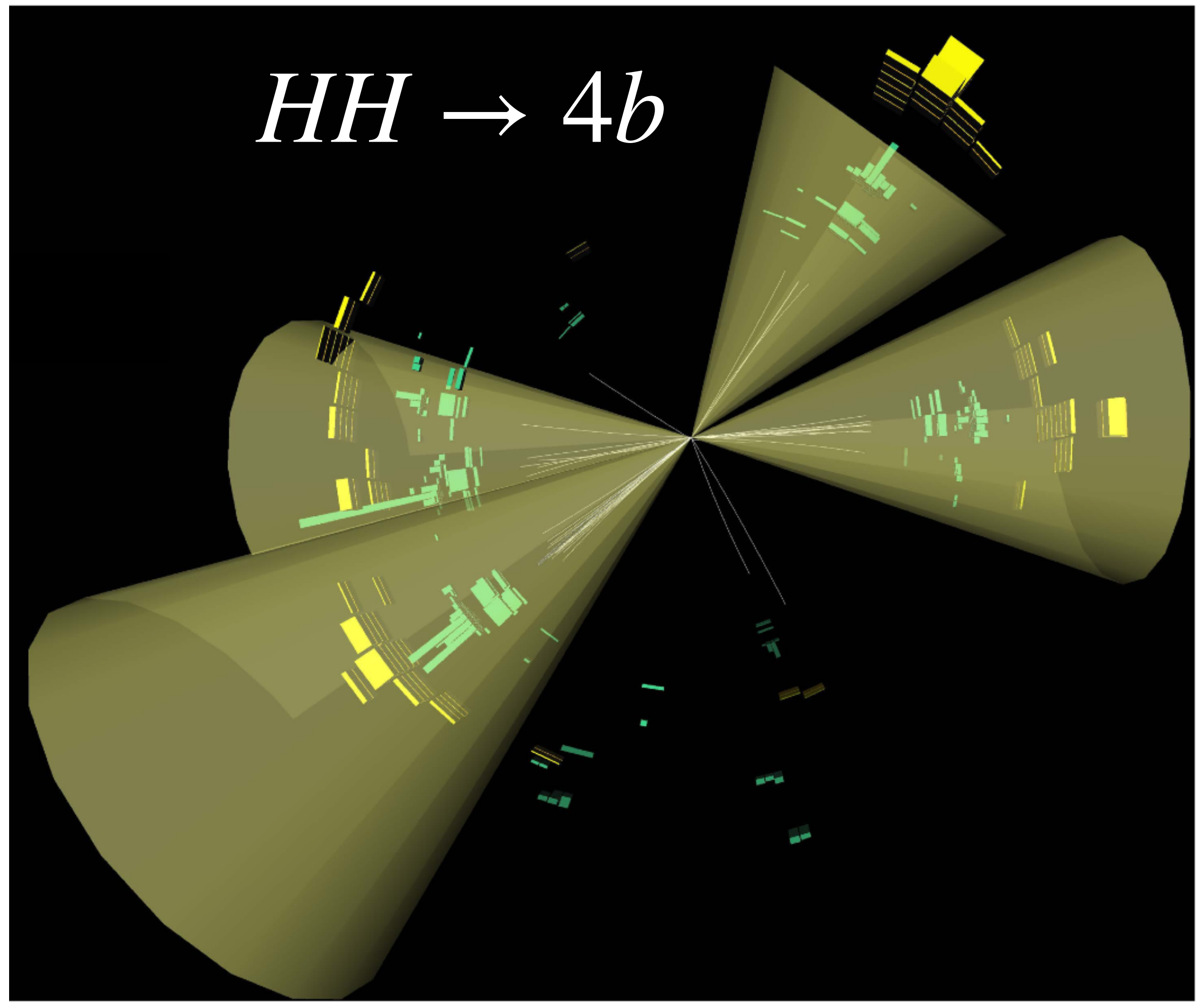
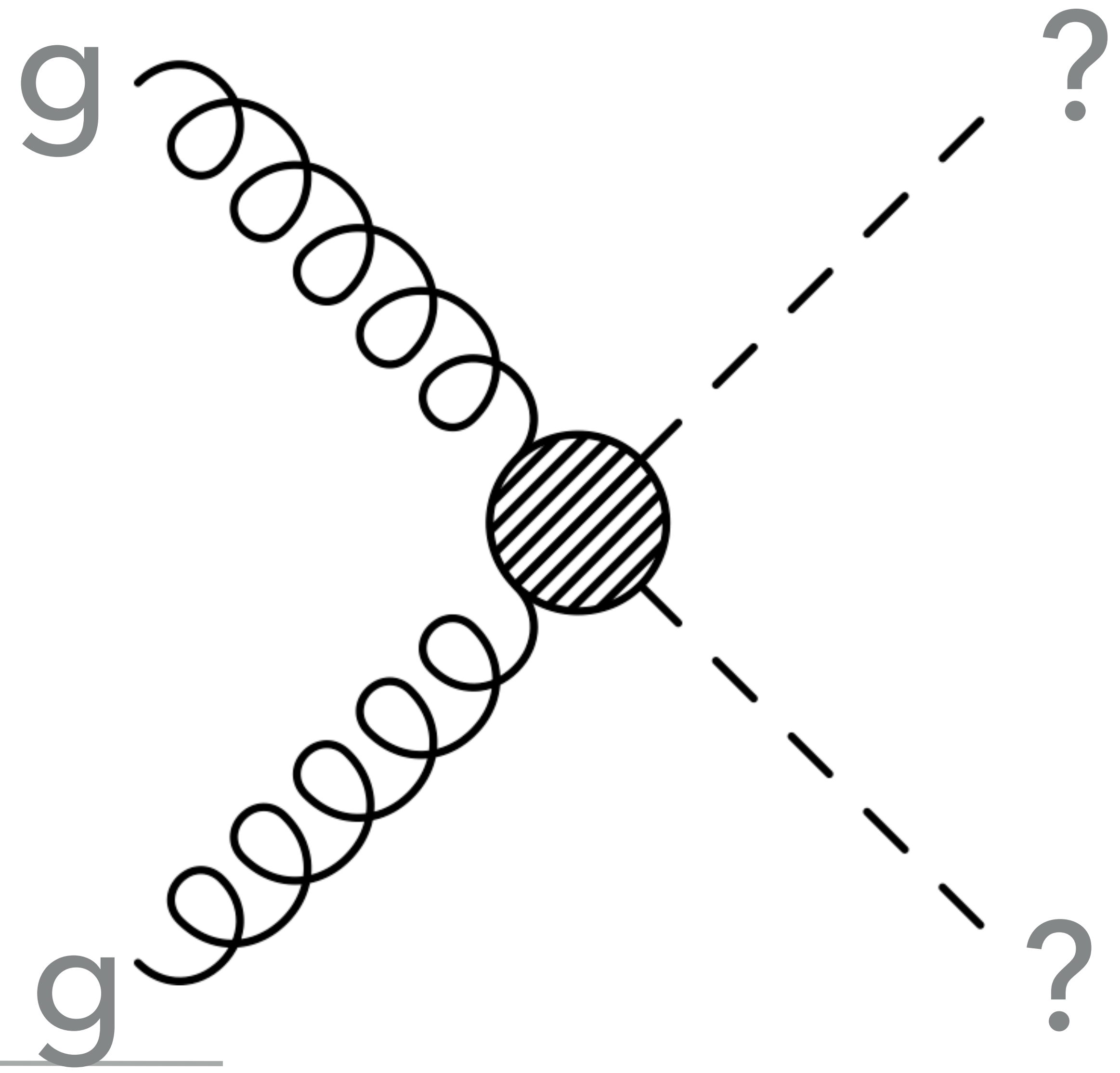


Image by Matt Strassler





What could be missing?

- How can we trigger on more complex low-energy hadronic signatures? Long-lived/displaced particles?
- What if we don’t know exactly what to look for?
- What if our signatures require complex multivariate algorithms (e.g. b tagging)?
- How can we improve on our traditional (often slow) reconstruction algorithms?

SUEP

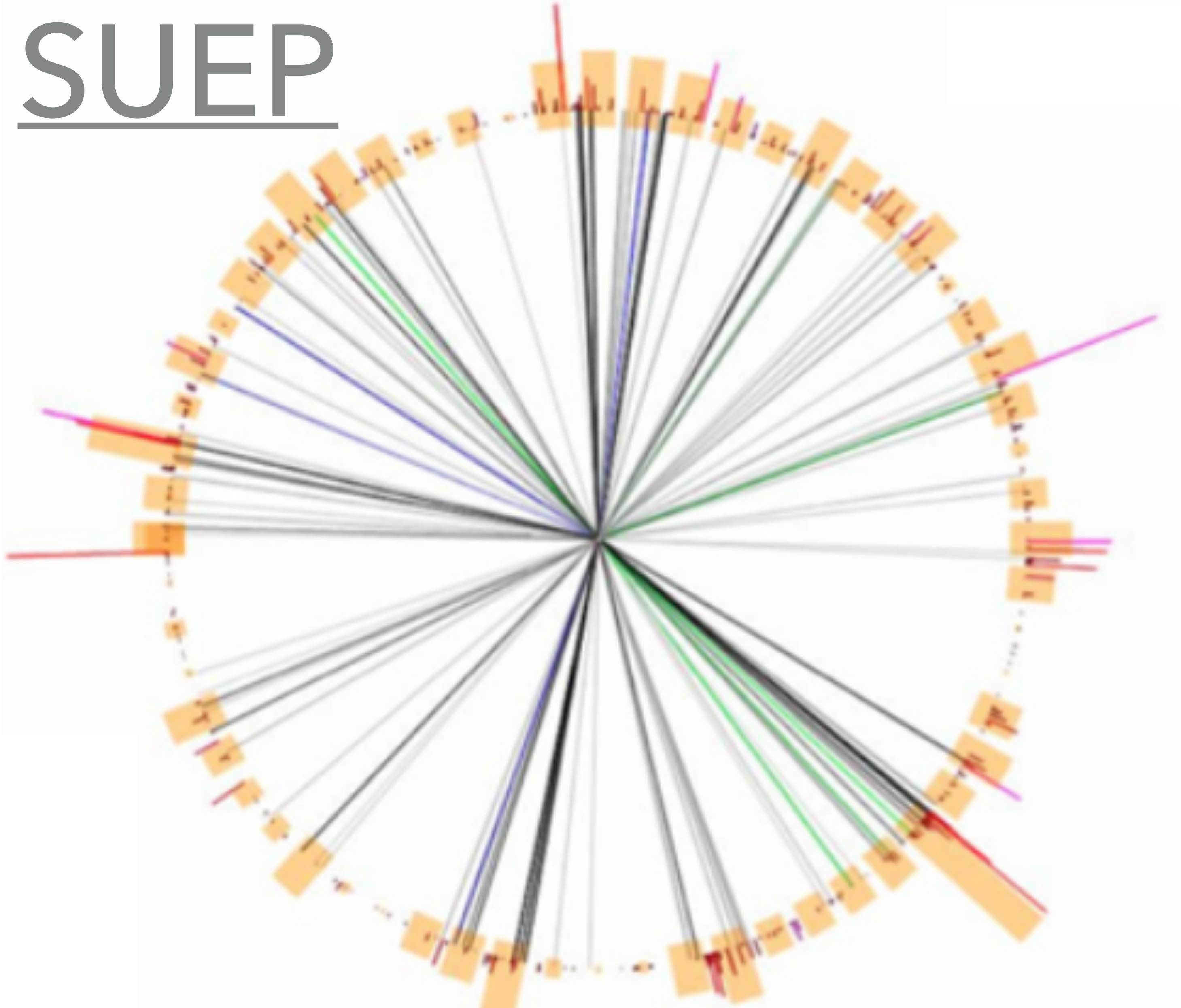
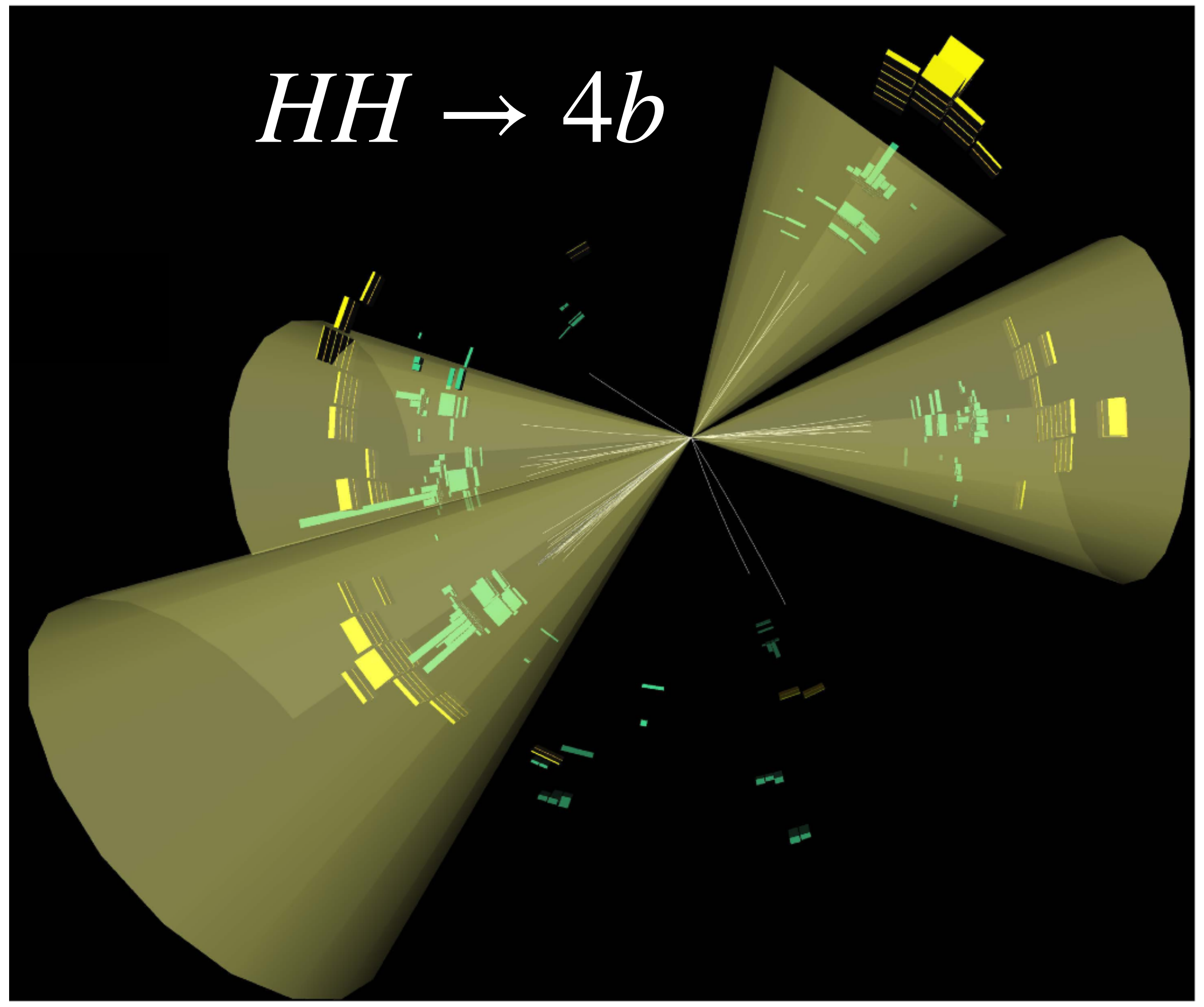
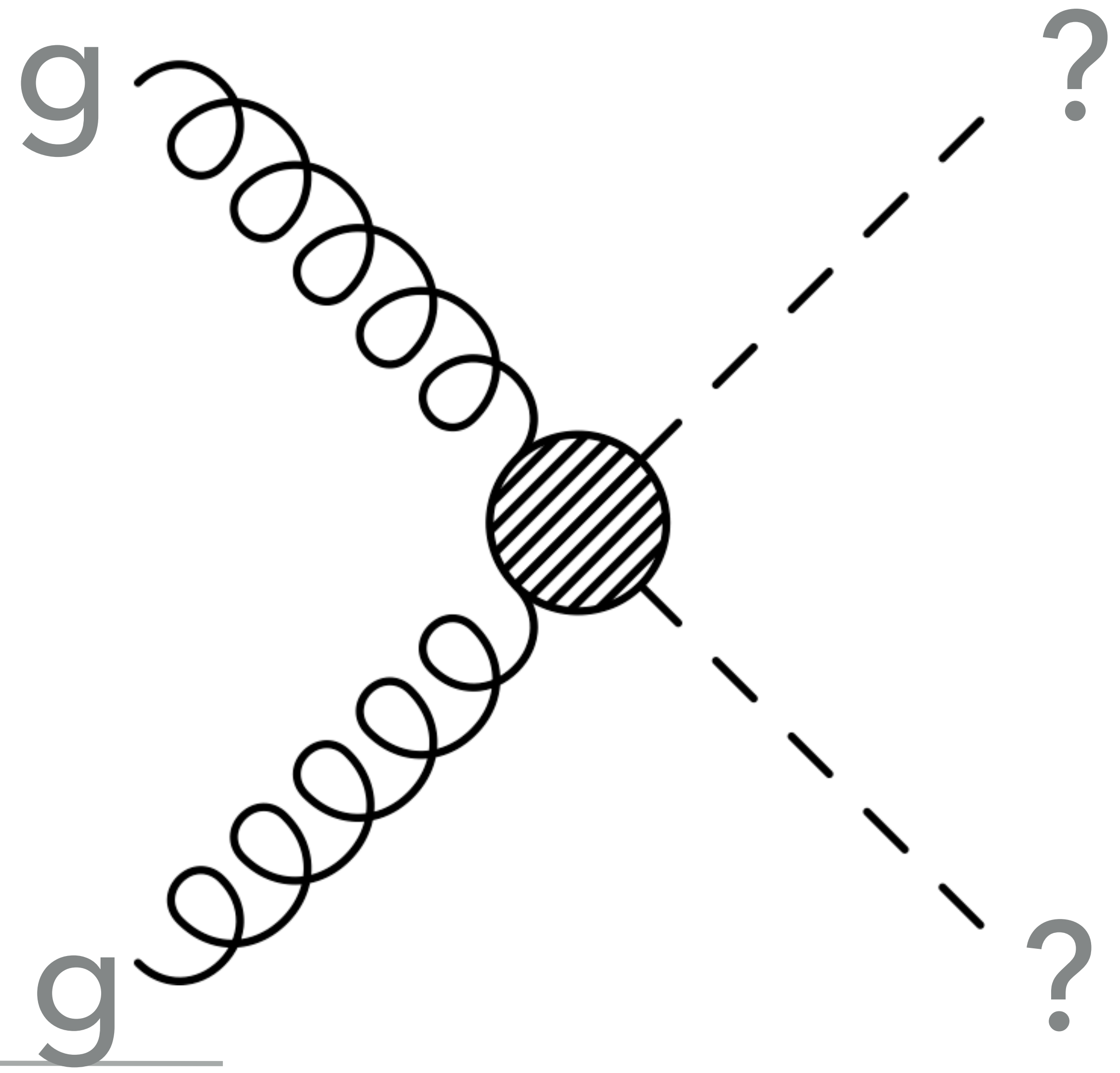
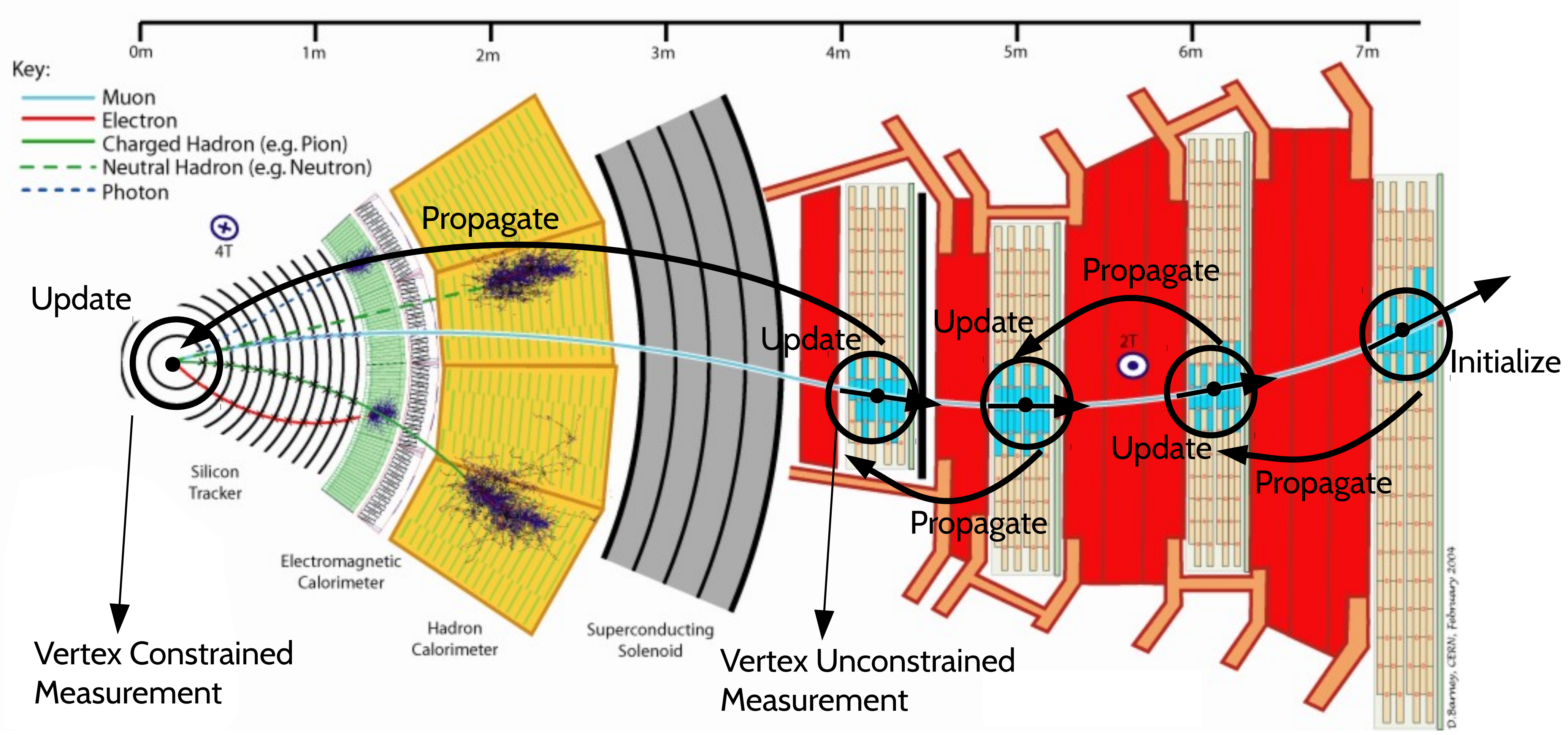


Image by Matt Strassler





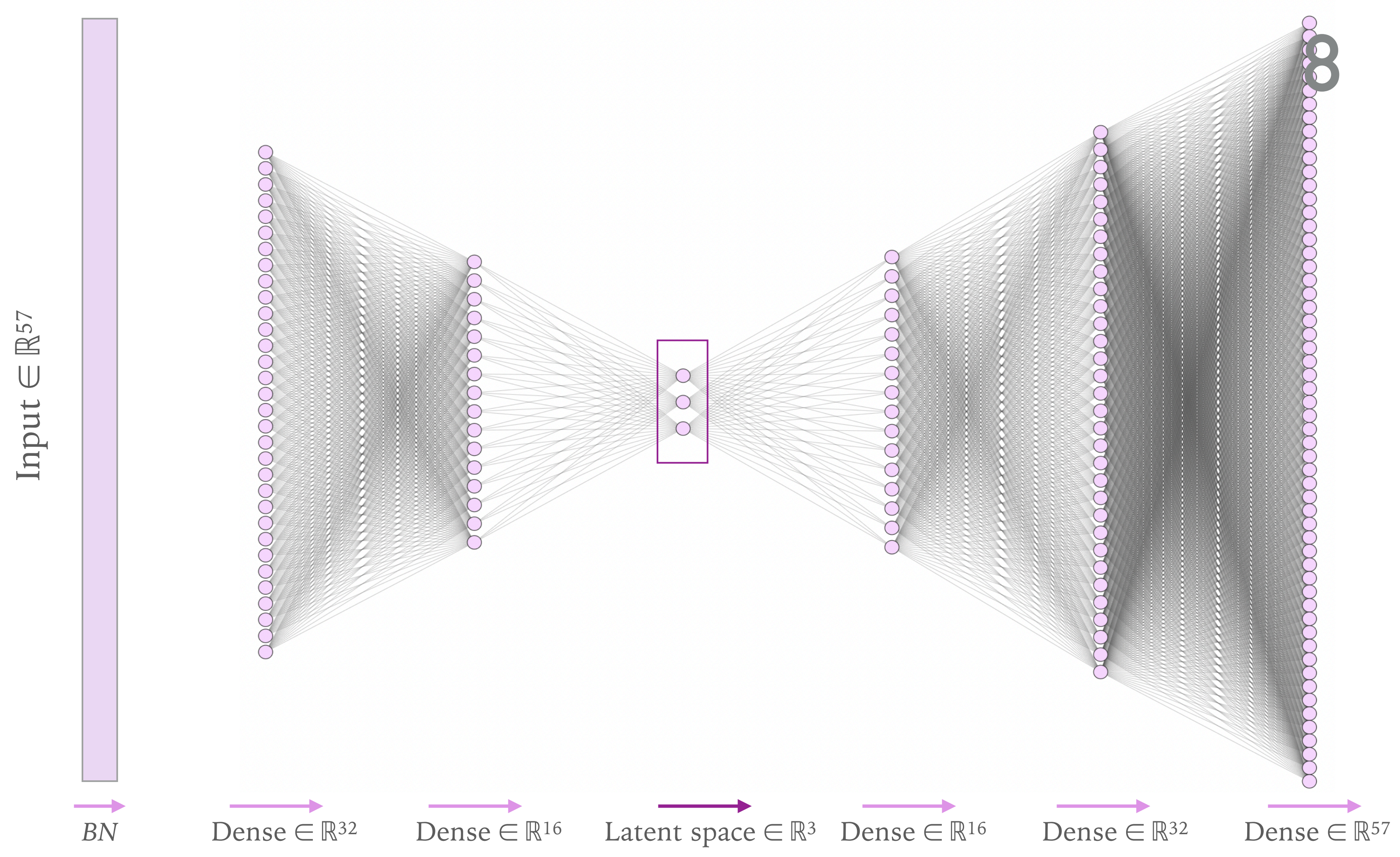


ML in Trigger

ML in Trigger

- (Variational) autoencoders for anomaly detection

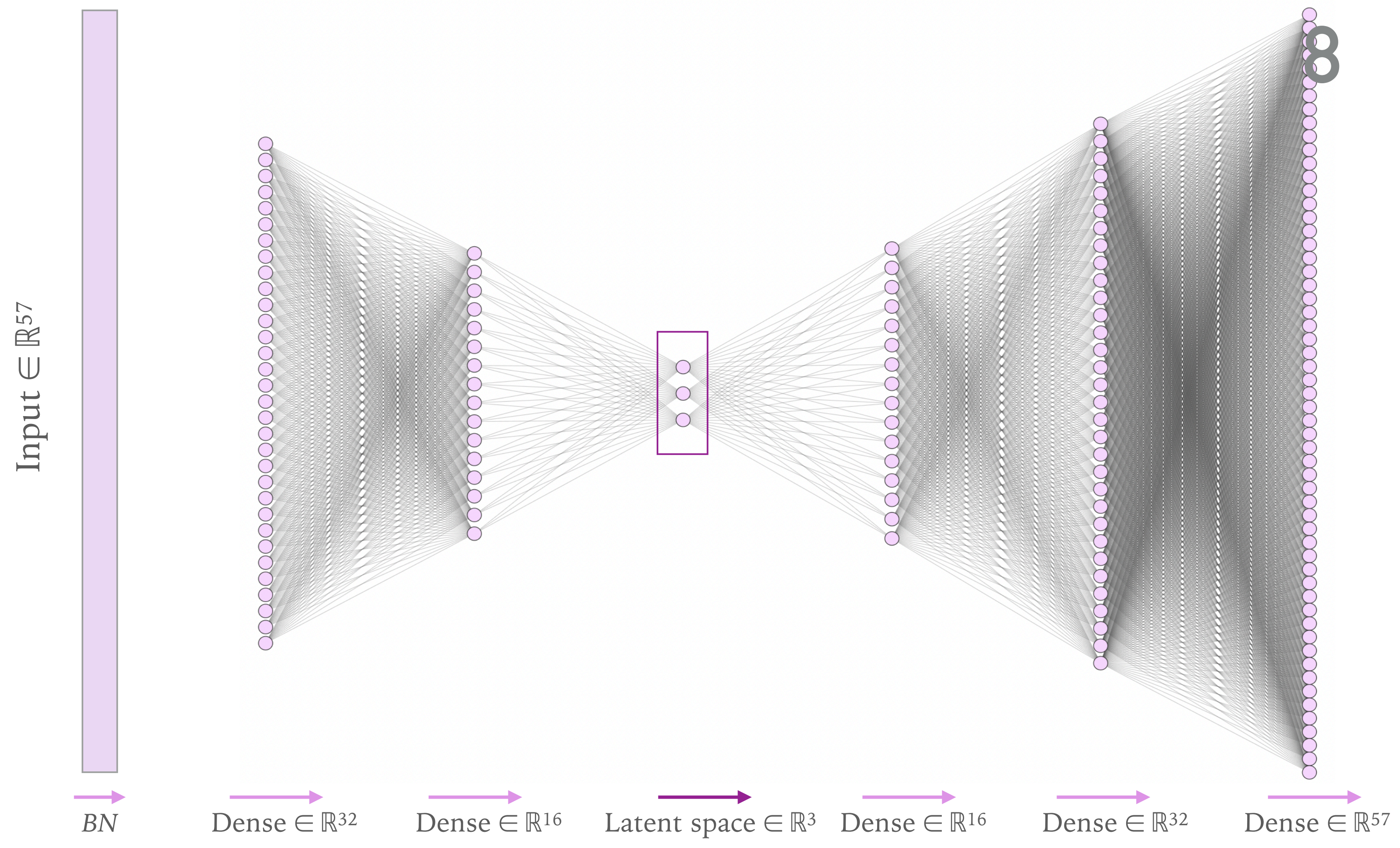
Nat. Mach. Intell. 4, 154 (2022)

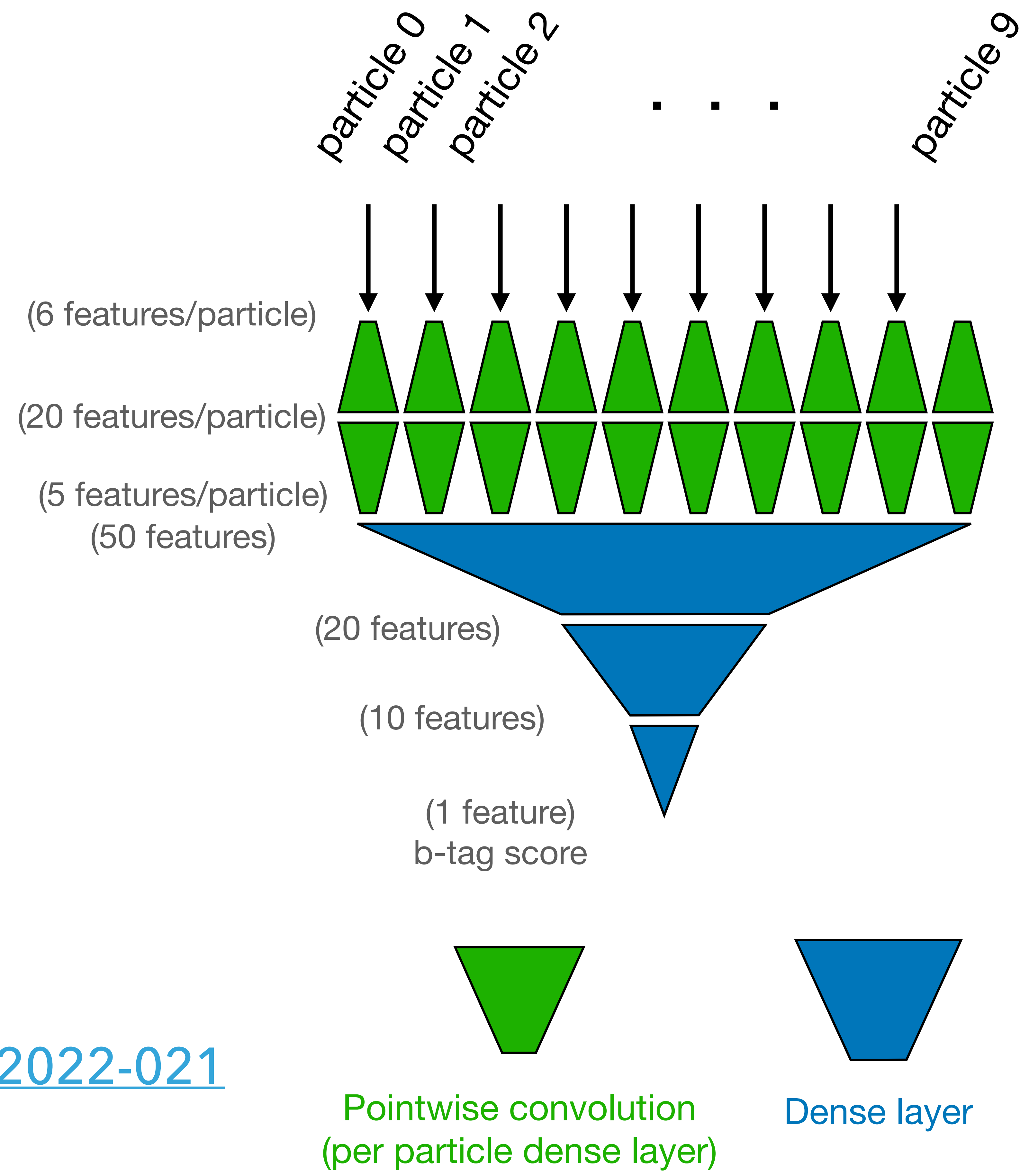


ML in Trigger

[Nat. Mach. Intell. 4, 154 (2022)]

- (Variational) autoencoders for anomaly detection
- 1D convolutional neural networks for b-tagging



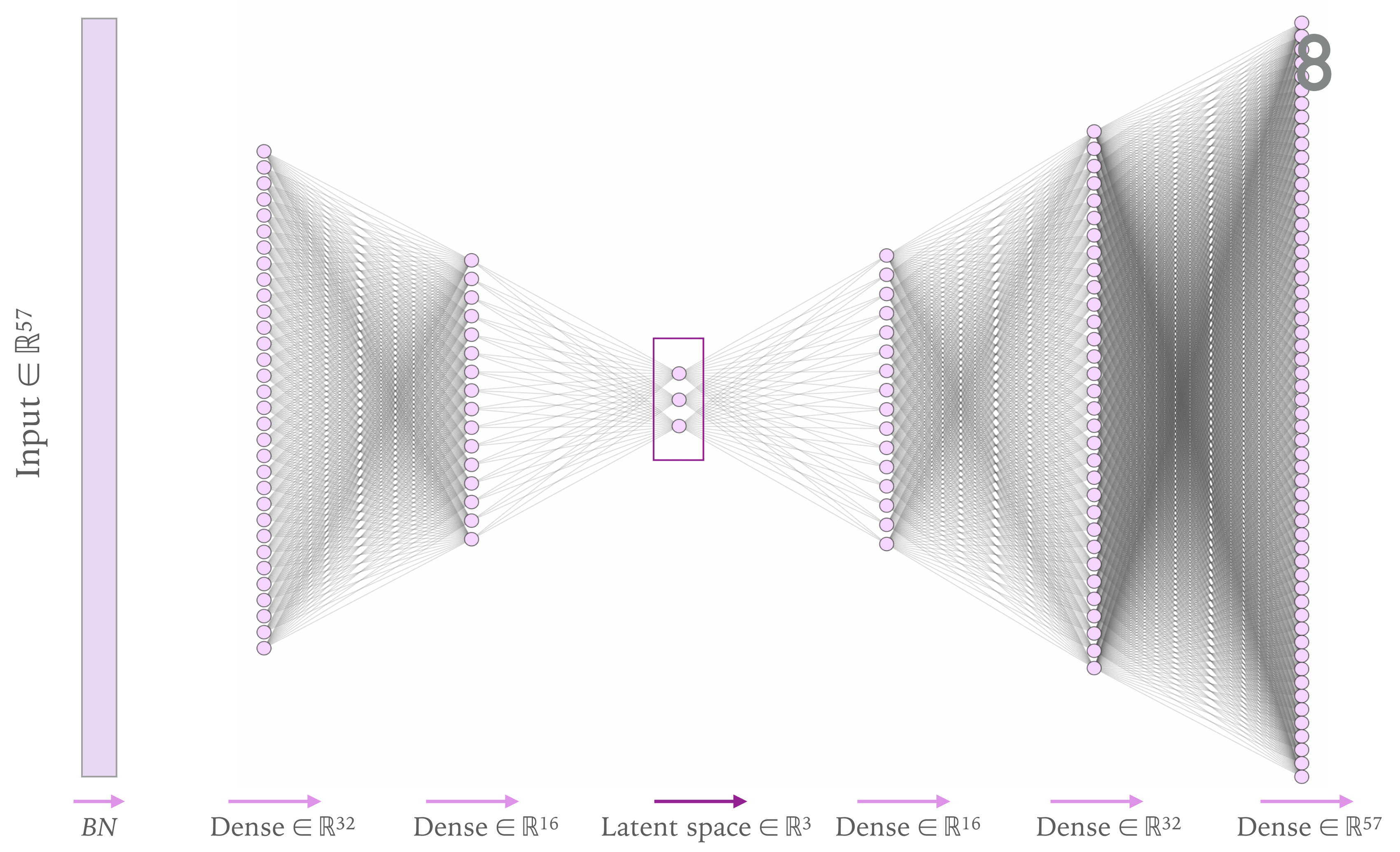


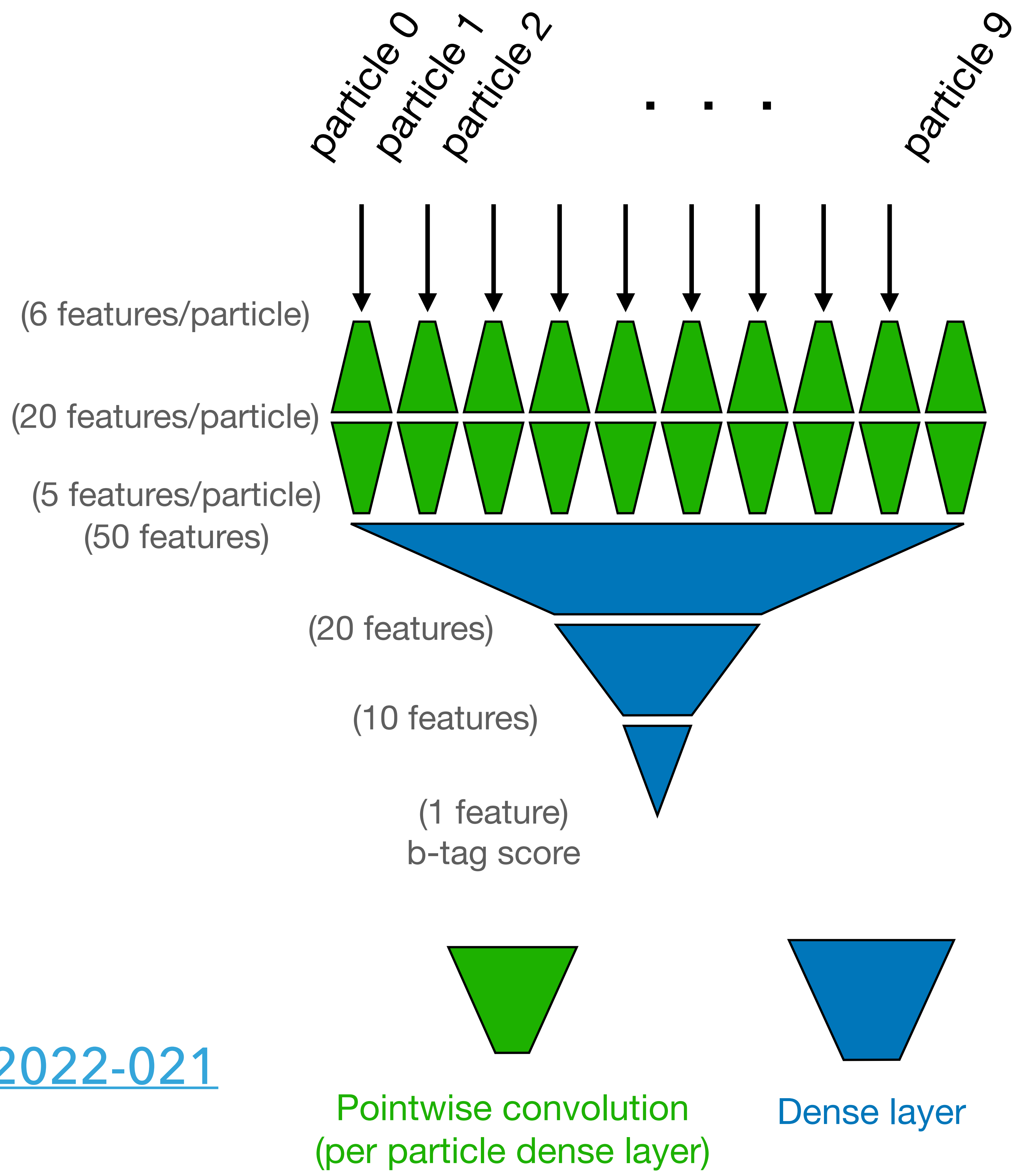
[CMS-DP-2022-021]

ML in Trigger

- (Variational) autoencoders for anomaly detection
- 1D convolutional neural networks for b-tagging
- Graph neural networks for tracking

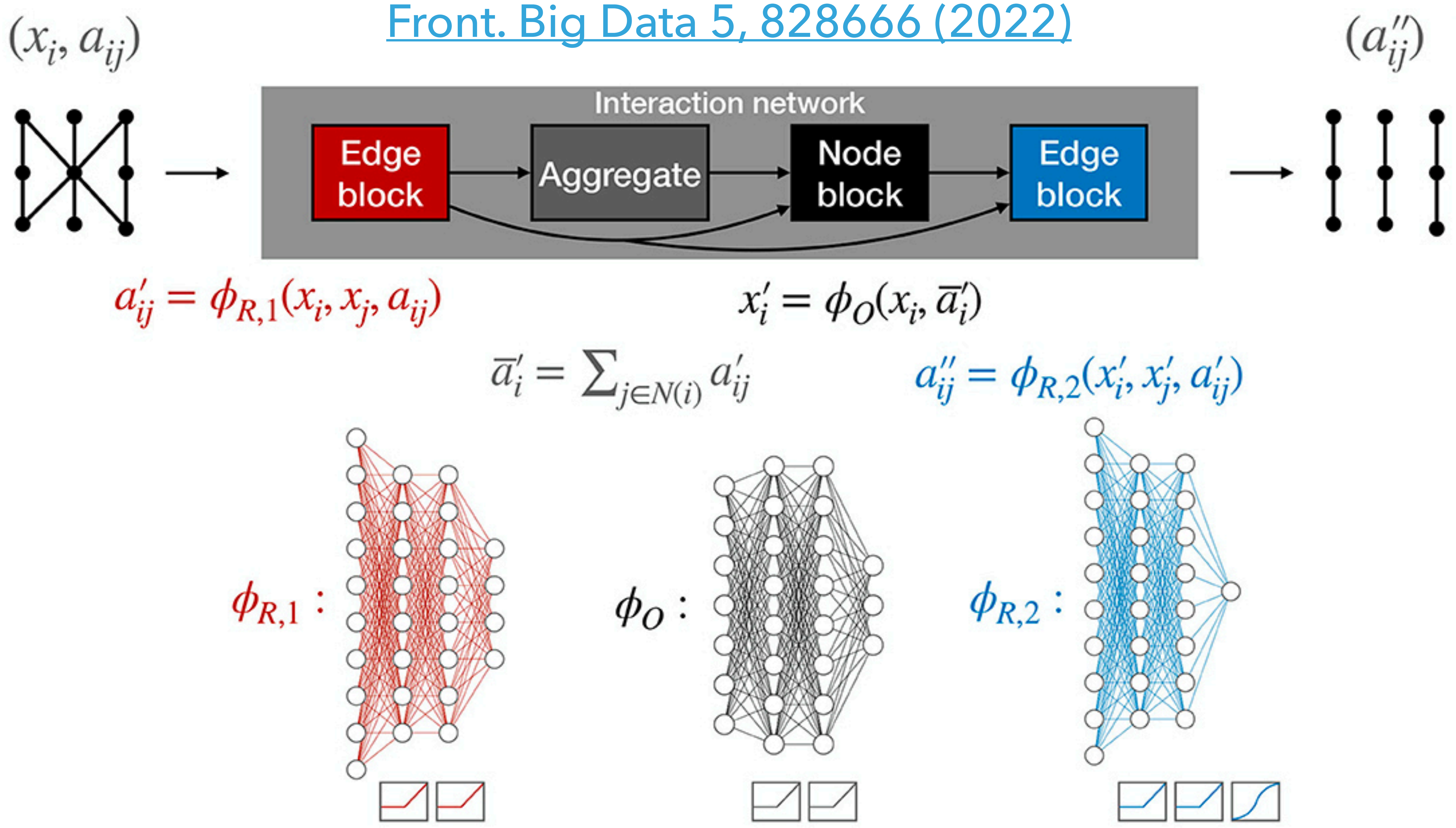
Nat. Mach. Intell. 4, 154 (2022)





CMS-DP-2022-021

Front. Big Data 5, 828666 (2022)

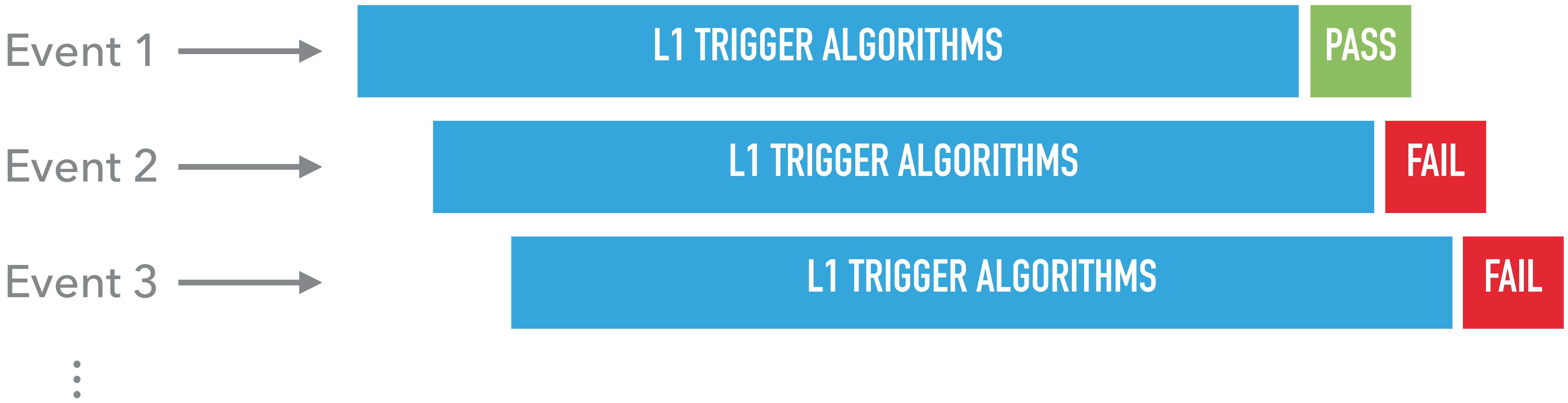


What makes this Hard?

Event 1 → L1 TRIGGER ALGORITHMS | PASS

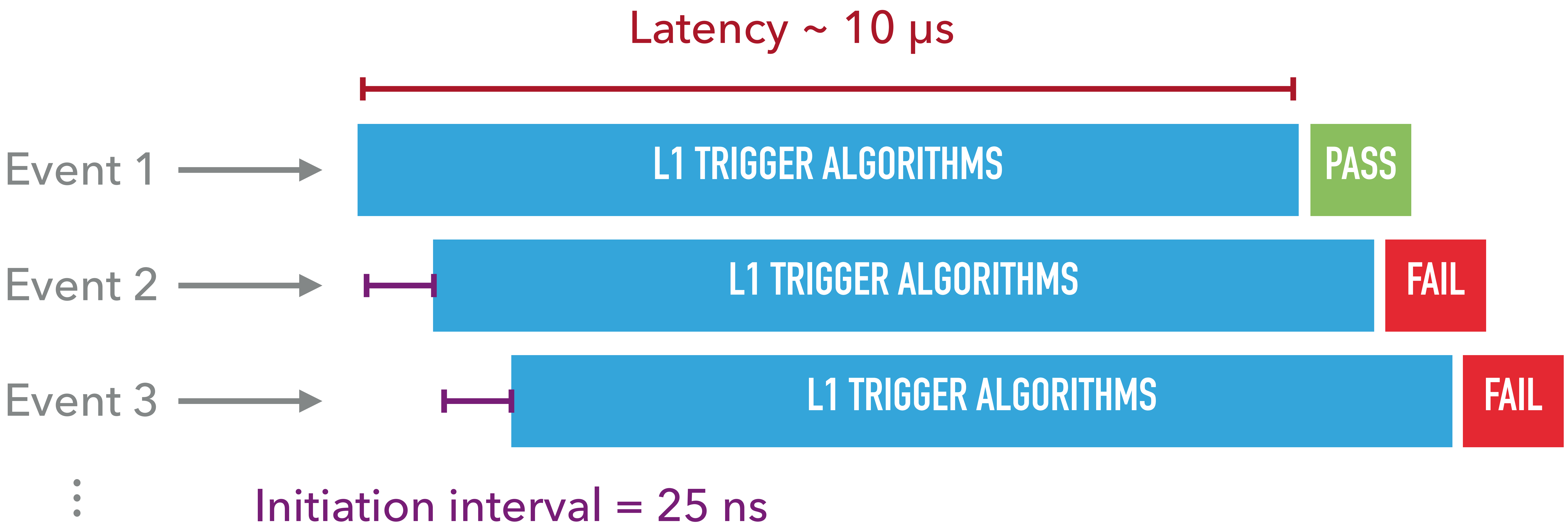
What makes this Hard?

- Reconstruct all events and reject 98% of them in ~10 µs



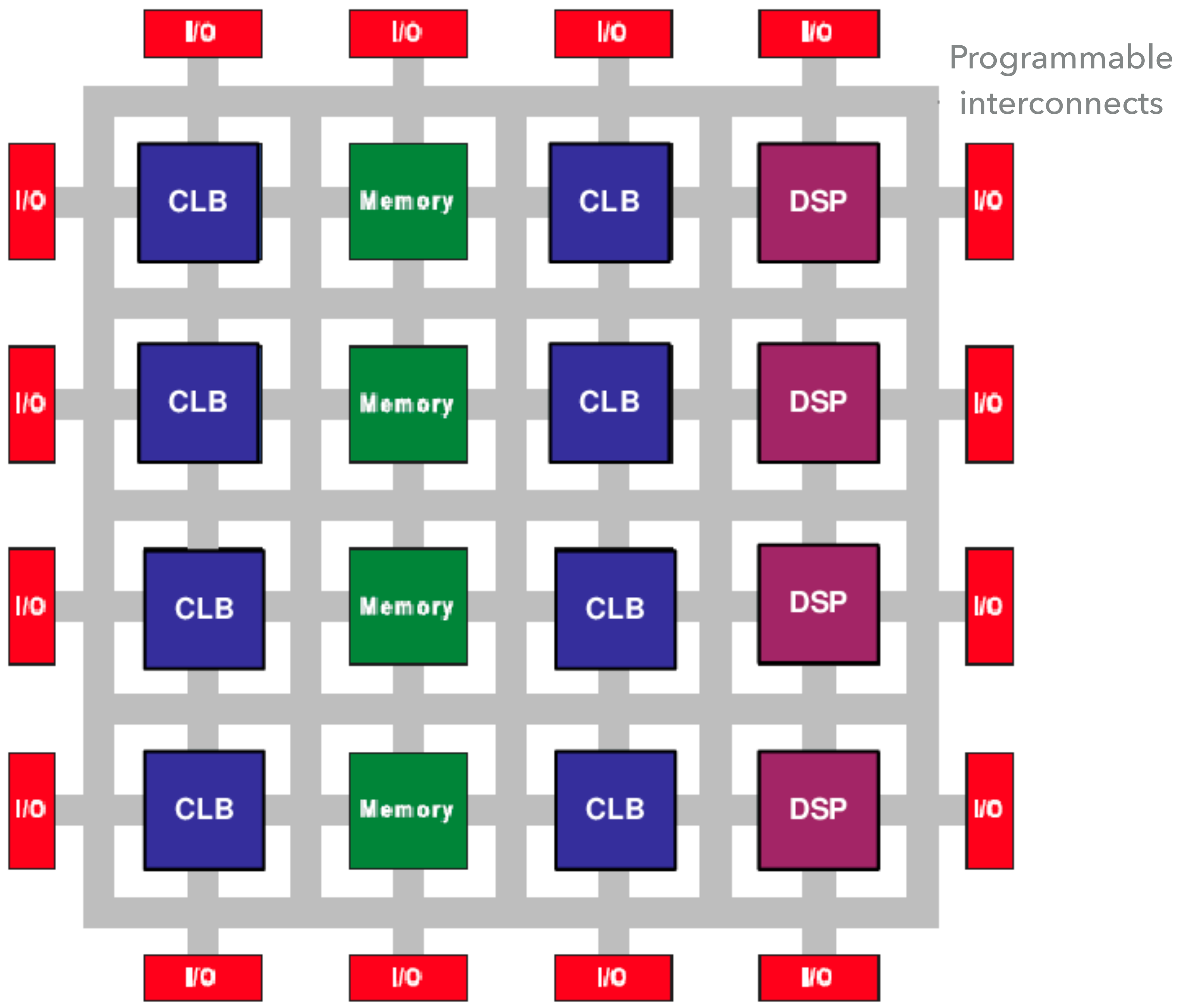
What makes this Hard?

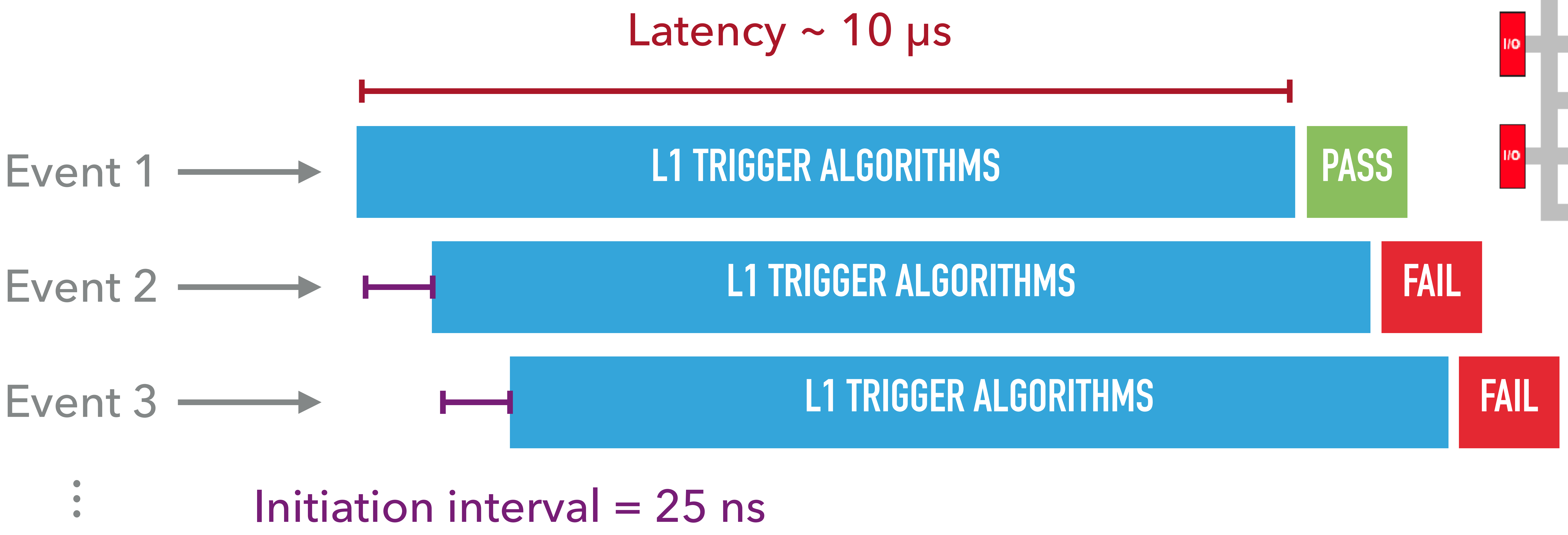
- Reconstruct all events and reject 98% of them in ~10 μs
  - Algorithms have to be <1 μs and process new events every (25 ns) × N_{tmux}



What makes this Hard?

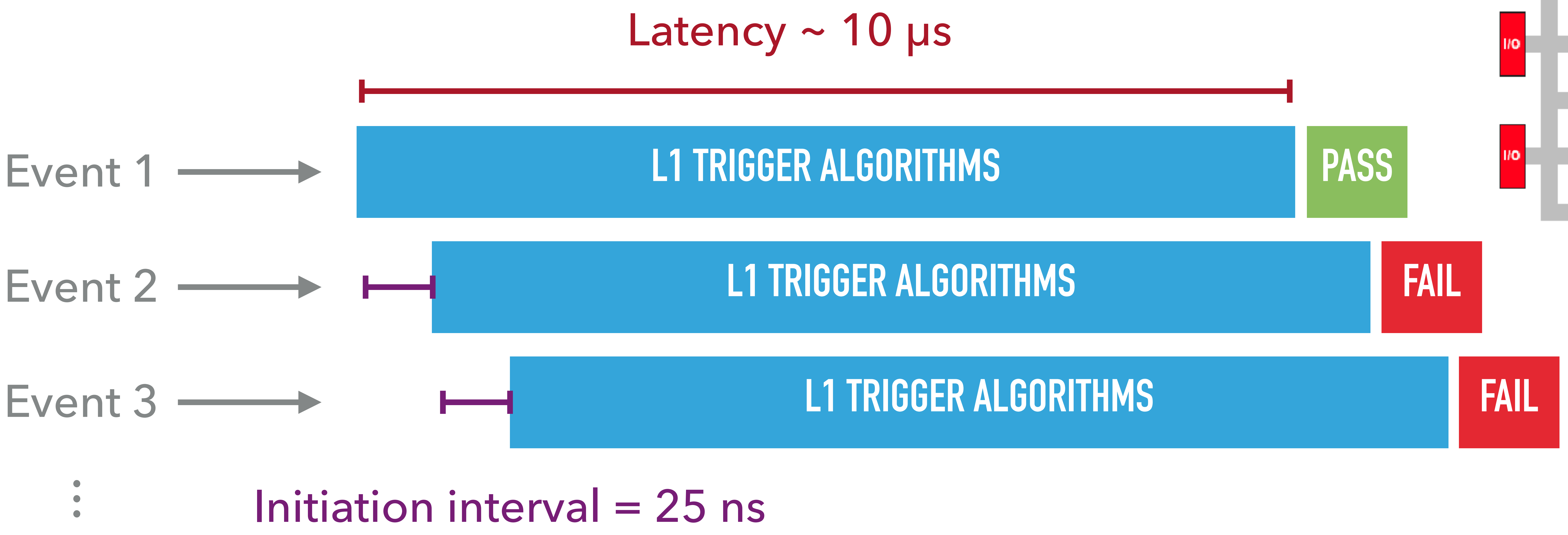
- Reconstruct all events and reject 98% of them in ~10 µs
    - Algorithms have to be <1 µs and process new events every (25 ns) × N_{tmux}
- Latency necessitates all **FPGA** design

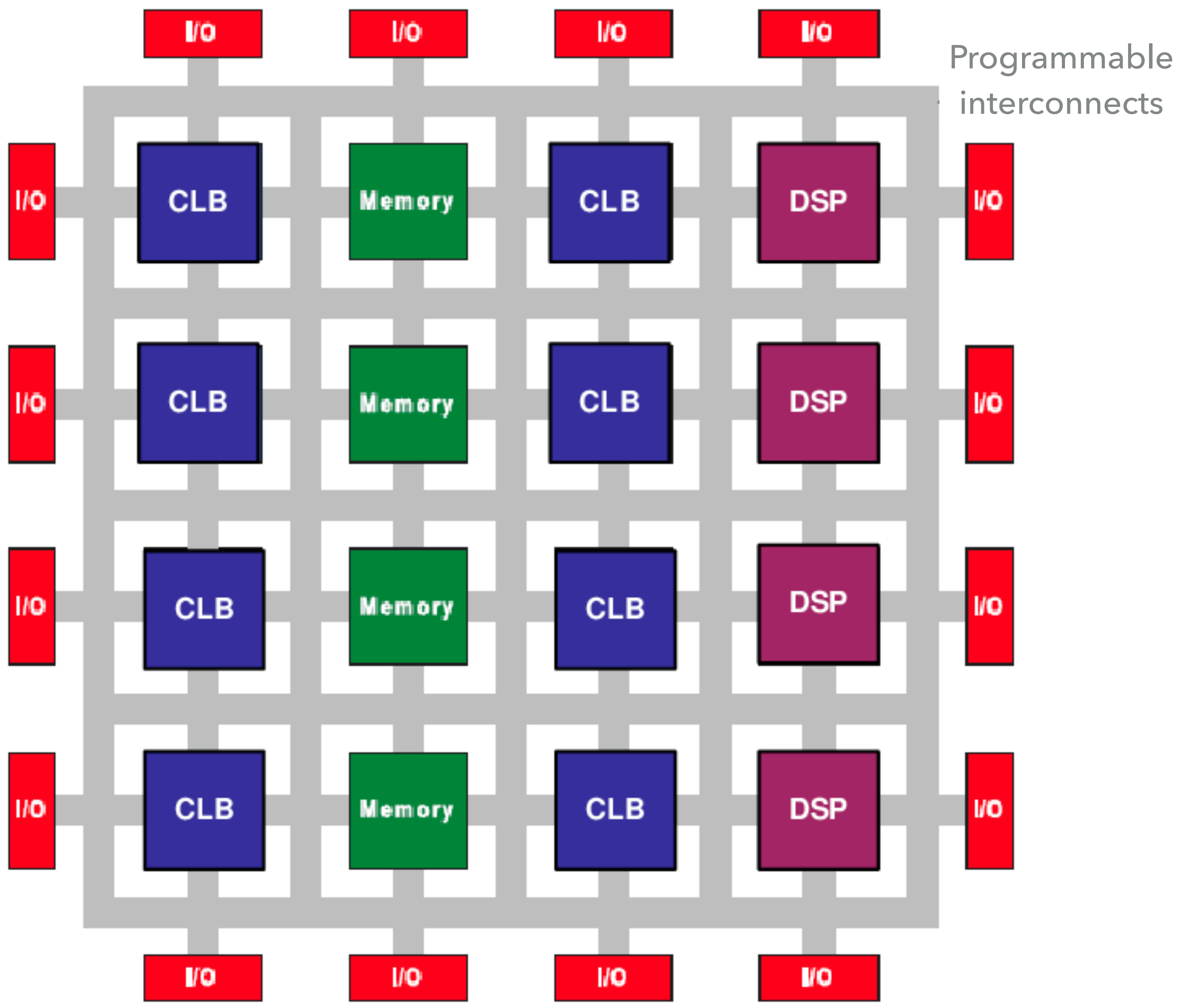




What makes this Hard?

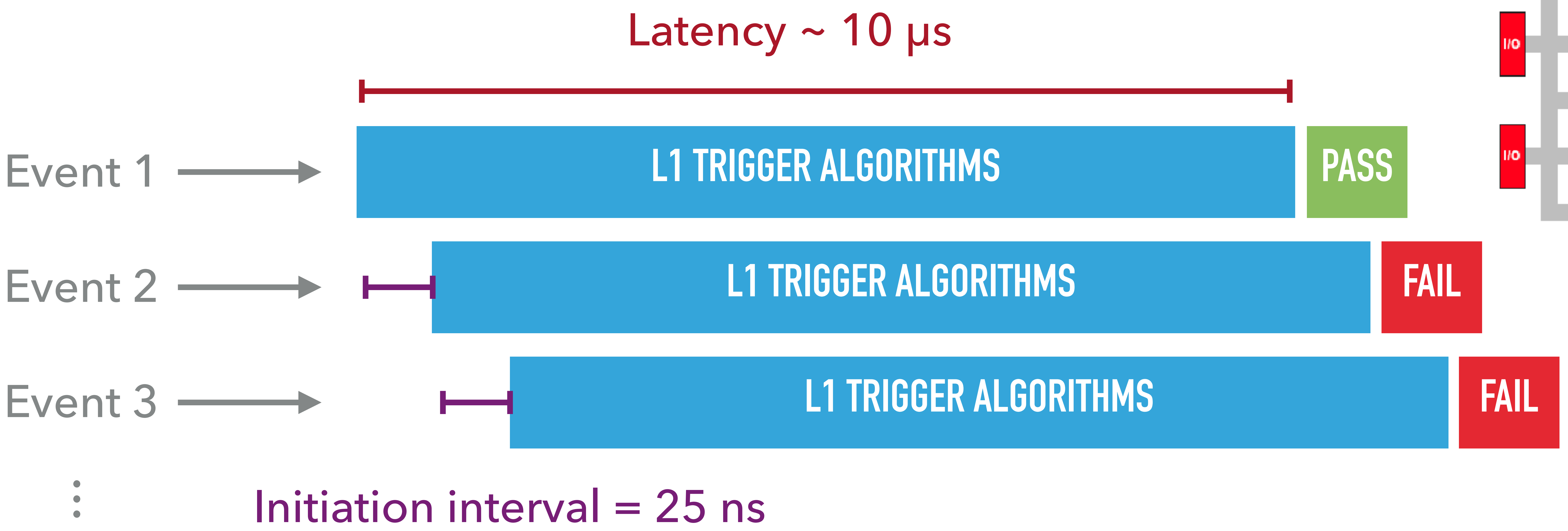
- Reconstruct all events and reject 98% of them in ~10 µs
  - Algorithms have to be <1 µs and process new events every (25 ns) × N_{tmux}
- Latency necessitates all **FPGA** design
  - Algorithms have to fit on <1 FPGA

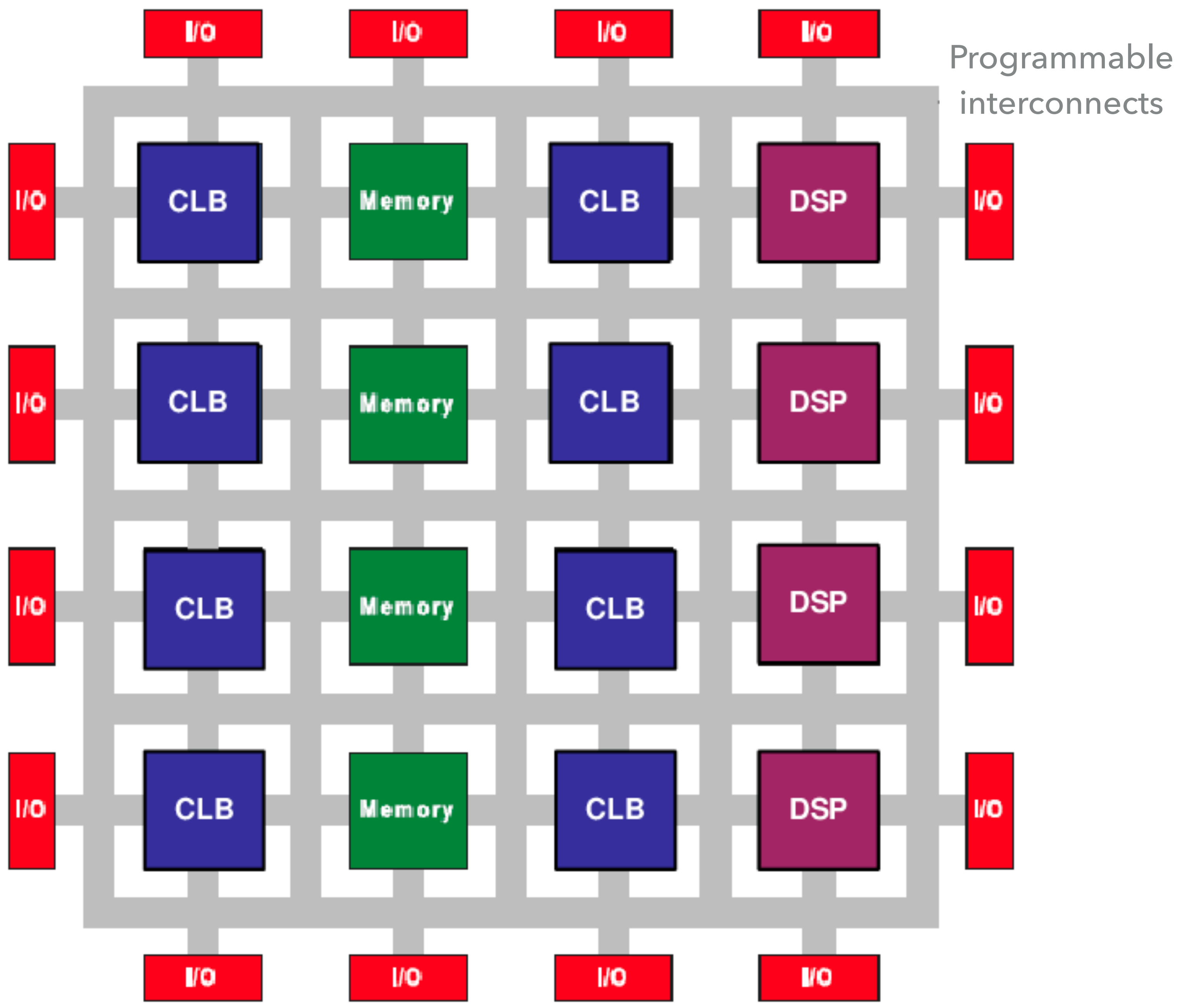




What makes this Hard?

- Reconstruct all events and reject 98% of them in ~10 μs
  - Algorithms have to be <1 μs and process new events every (25 ns) × N_{tmux}
- Latency necessitates all **FPGA** design
  - Algorithms have to fit on <1 FPGA
- How can we satisfy these constraints?

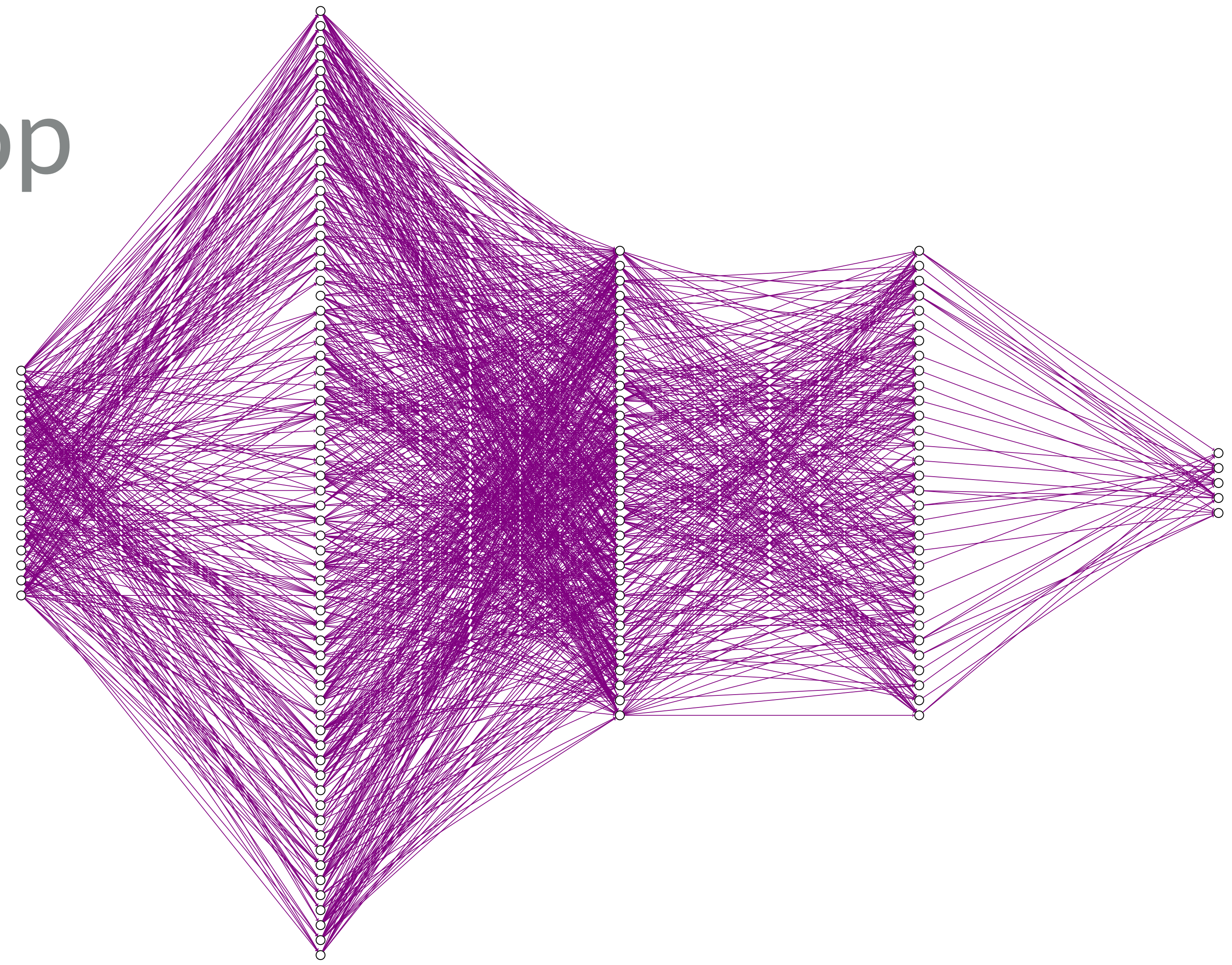




Codesign

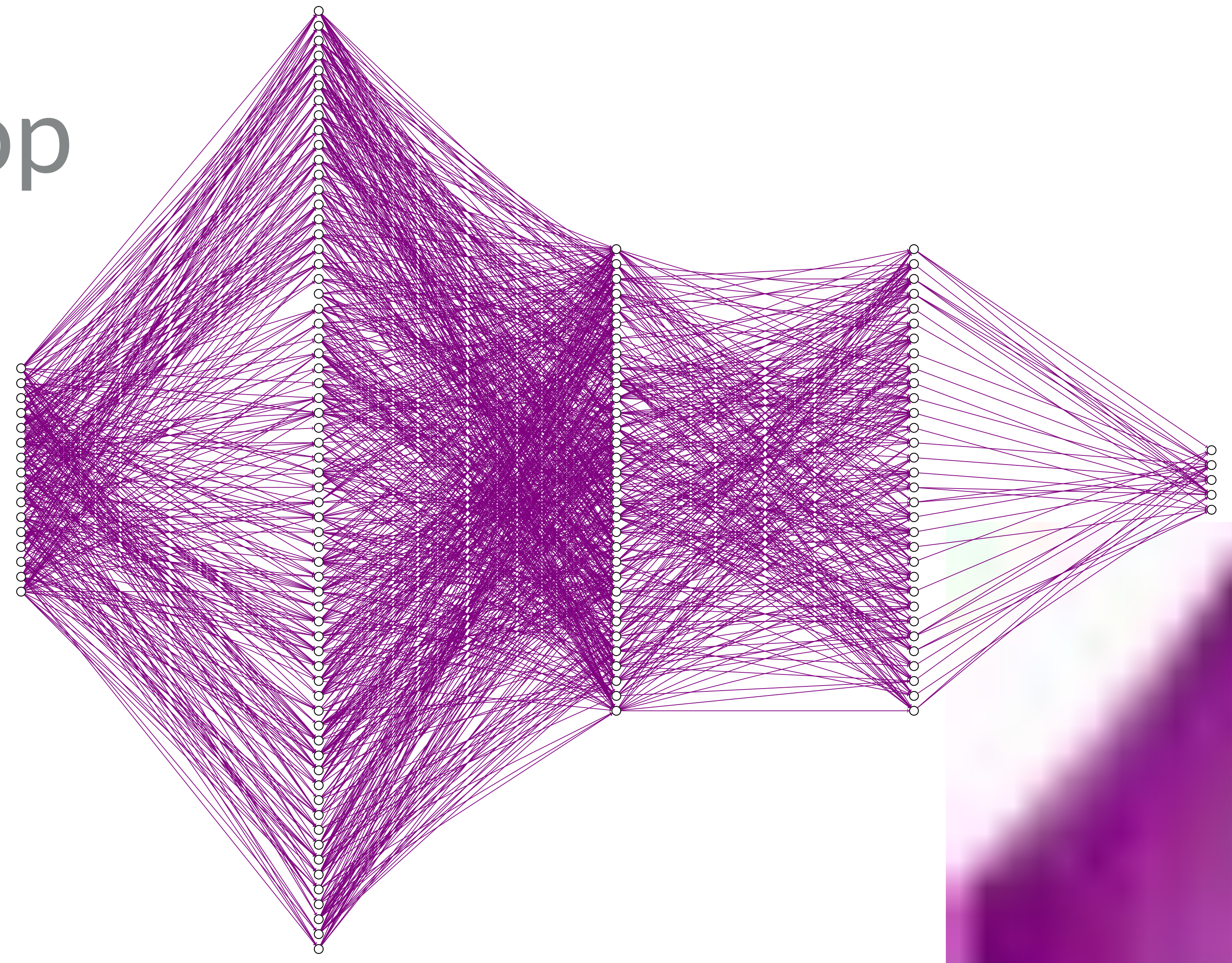
Codesign

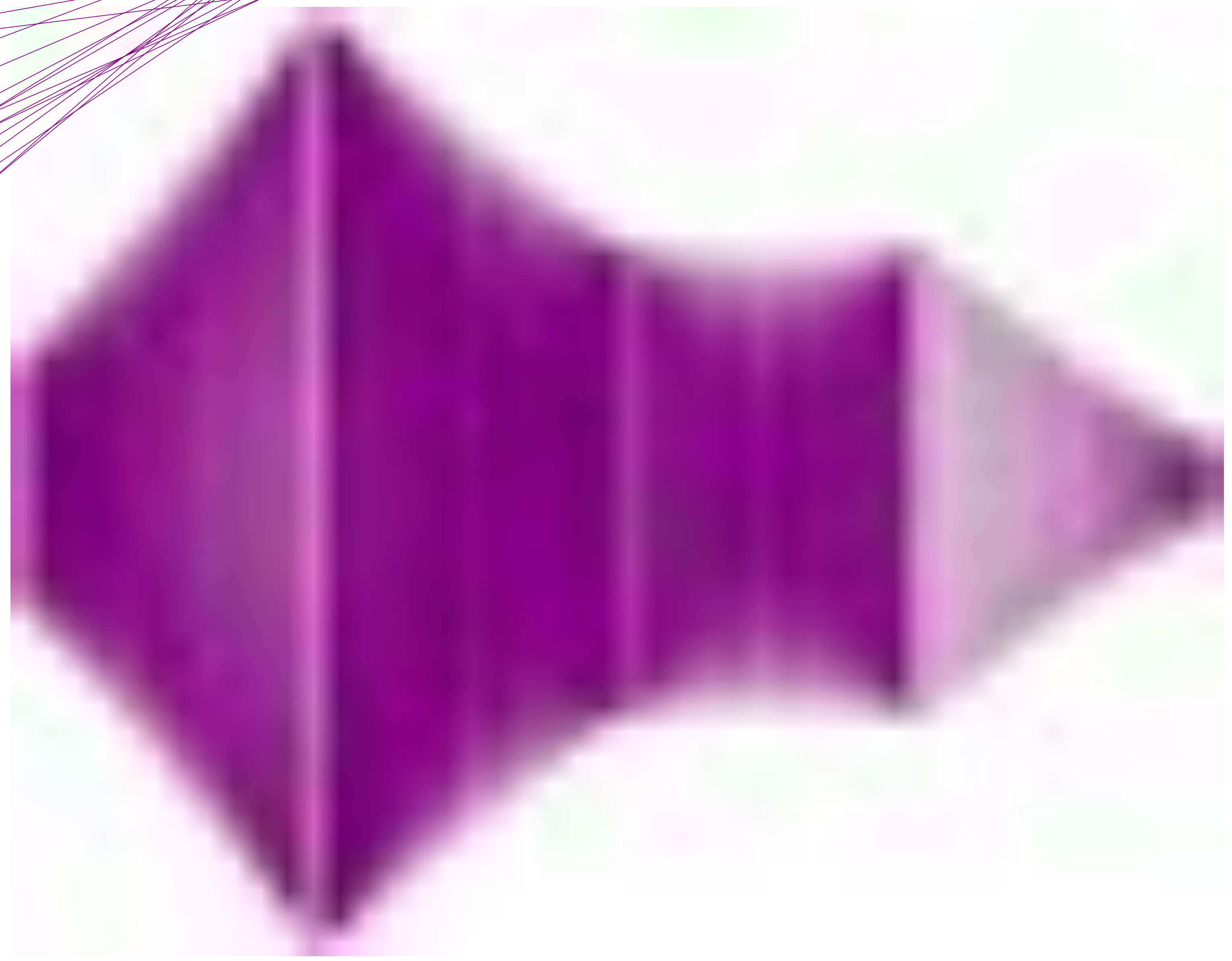
- **Codesign**: intrinsic development loop between ML design, training, and implementation
- Pruning
  - Maintain high performance while removing redundant operations



Codesign

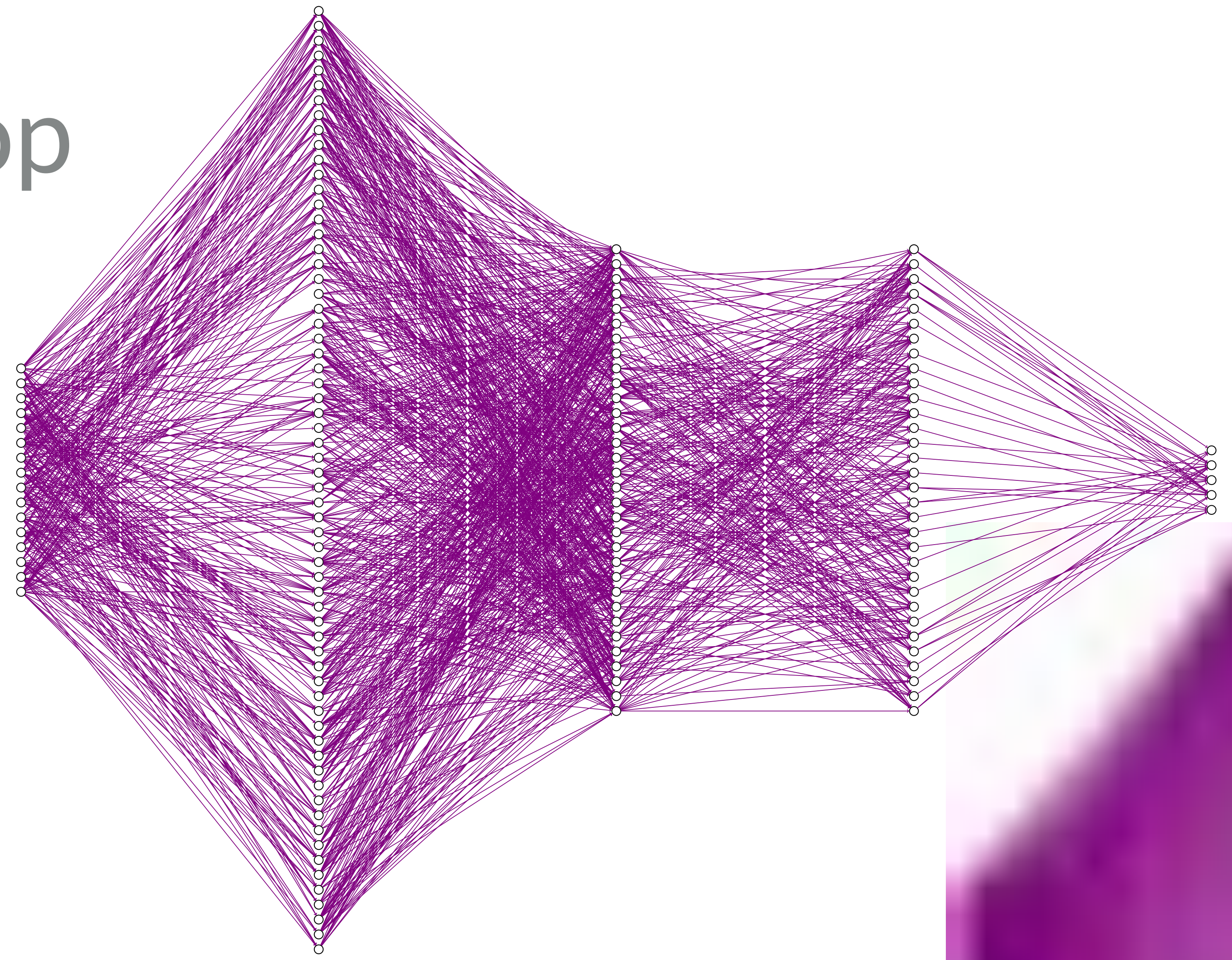
- **Codesign**: intrinsic development loop between ML design, training, and implementation
- Pruning
  - Maintain high performance while removing redundant operations
- Quantization
  - Reduce precision from 32-bit floating point to 16-bit, 8-bit, …

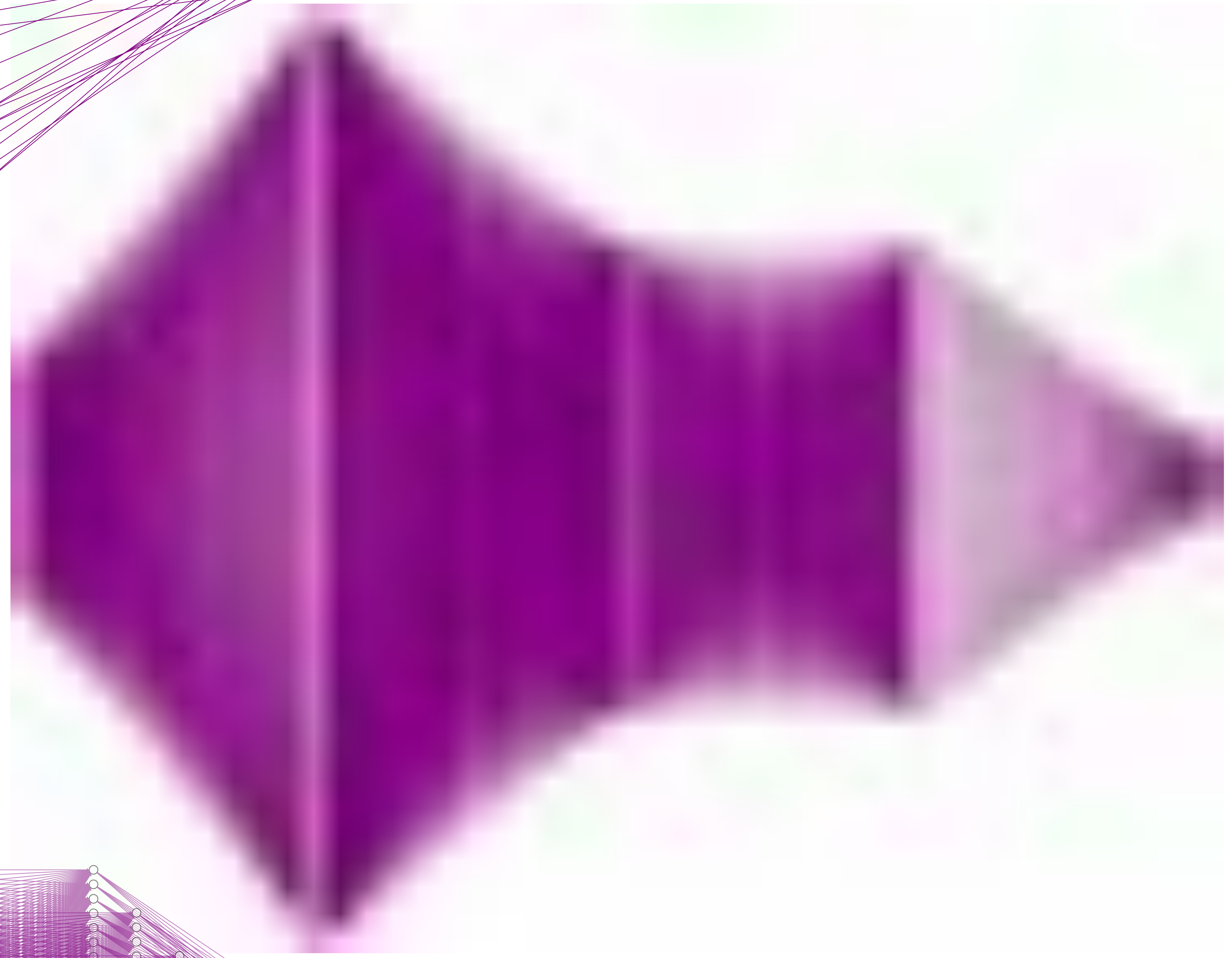


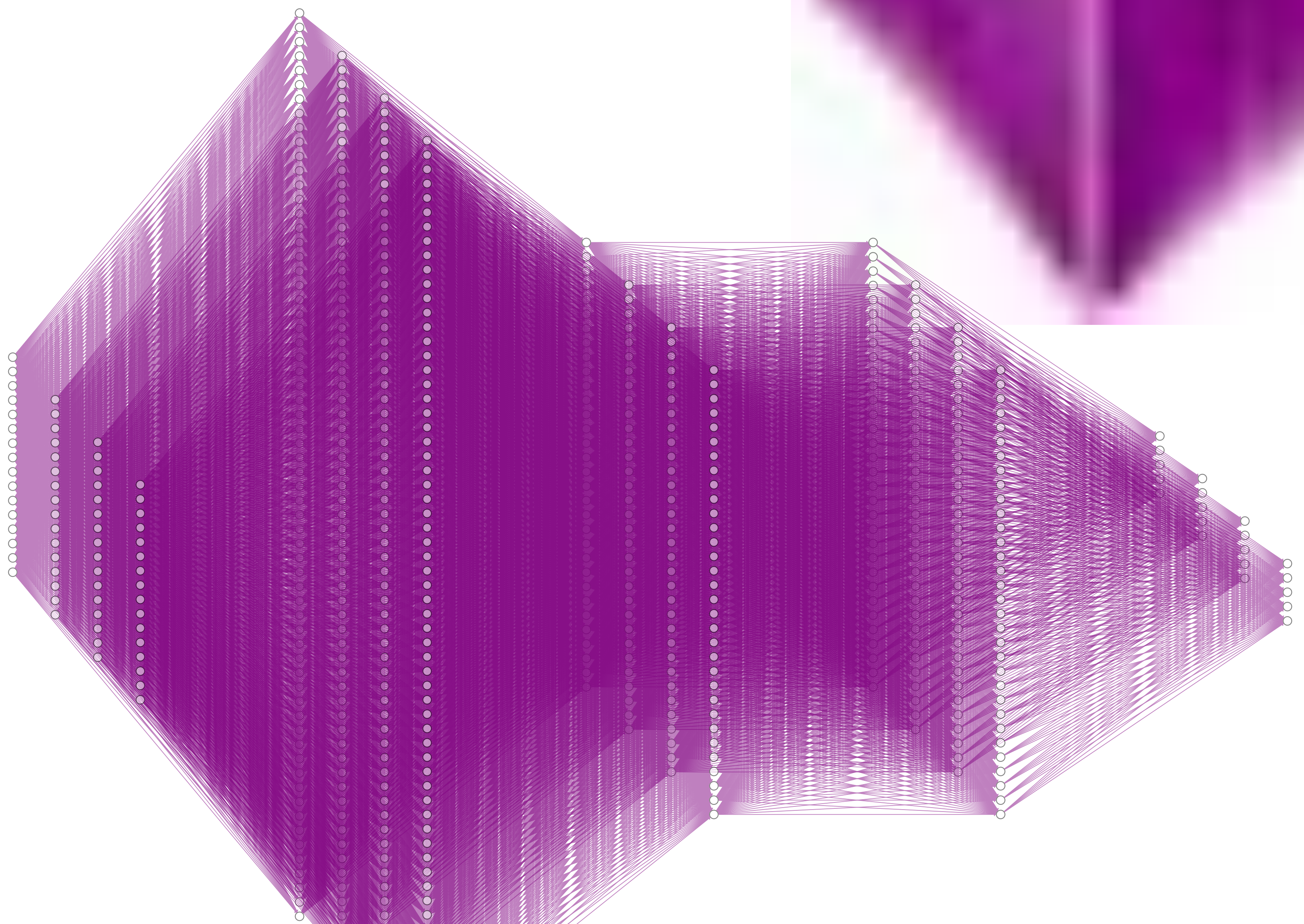


Codesign

- **Codesign**: intrinsic development loop between ML design, training, and implementation
- Pruning
  - Maintain high performance while removing redundant operations
- Quantization
  - Reduce precision from 32-bit floating point to 16-bit, 8-bit, …
- Parallelization
  - Balance parallelization (how fast) with resources needed (how costly)







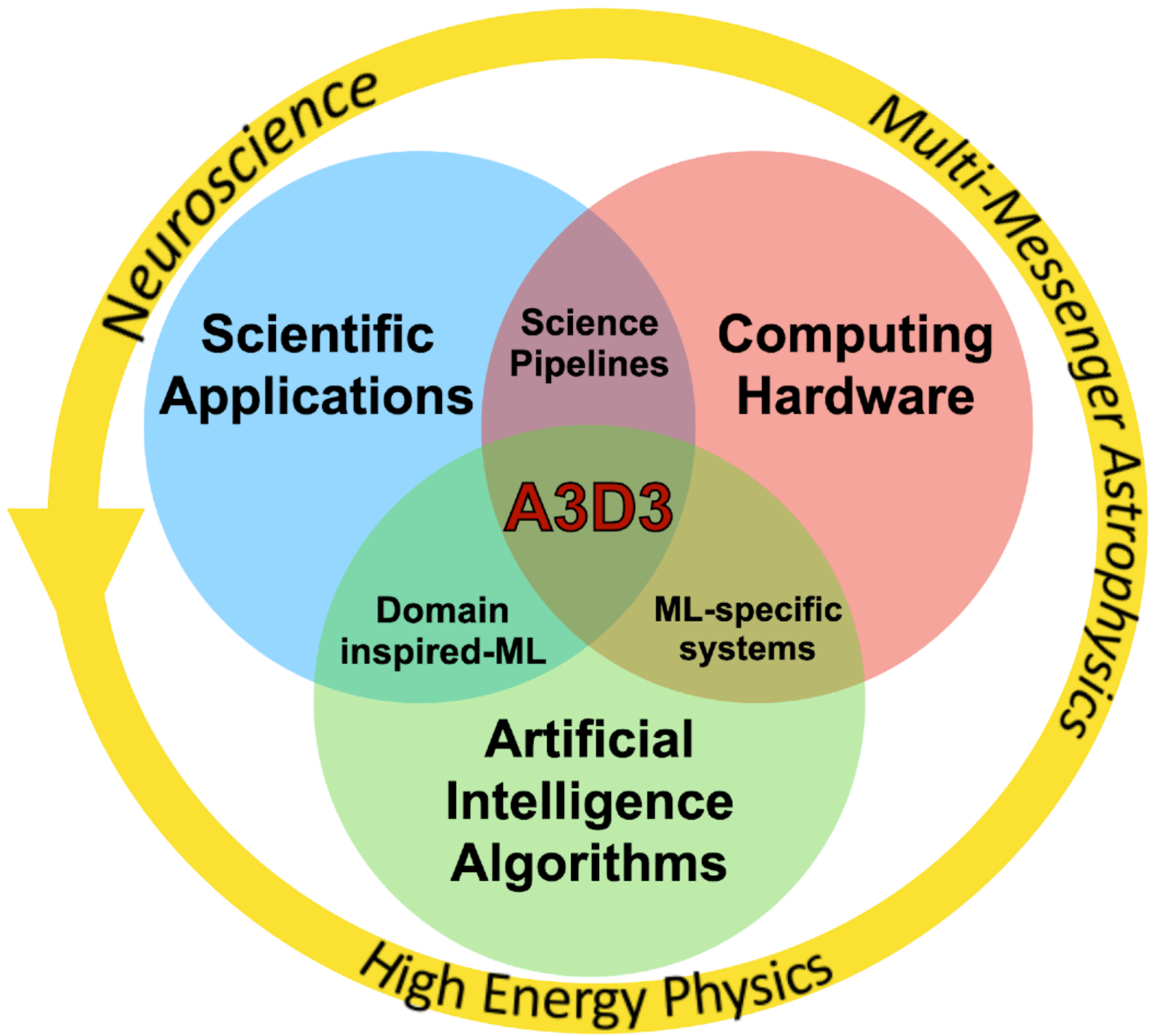
NSF A3D3 Institute

Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery

Our Mission is to enable real-time AI techniques for scientific and engineering discovery by uniting three core components: Scientific Applications, Artificial Intelligence Algorithms, and Computing Hardware.

Collaborators welcome! Check the a3d3.ai for events

OAC-2117997

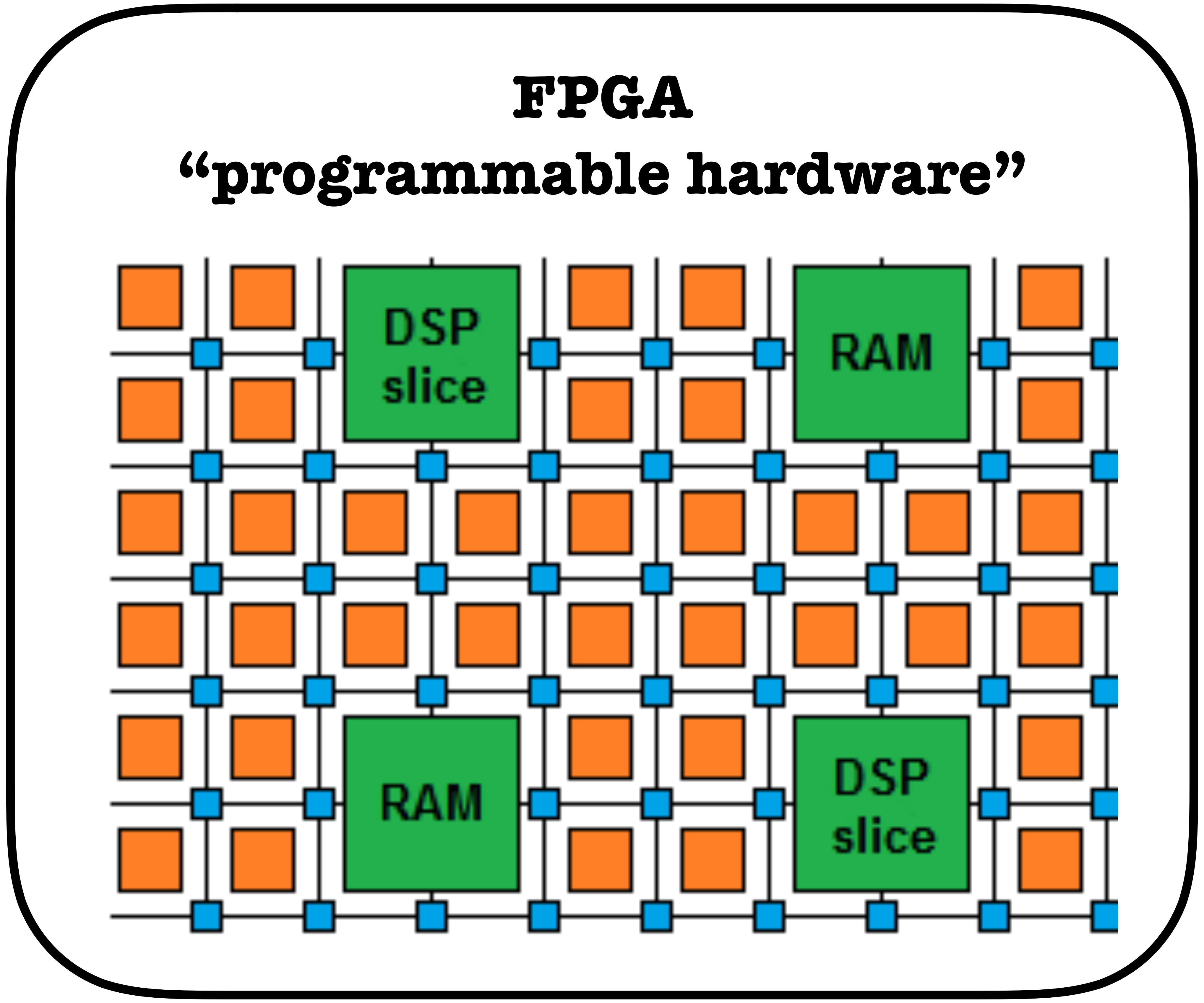


Accelerated AI Algorithms for Data-Driven Discovery

Modern FPGAs

Pros:

- Reprogrammable interconnects between embedded components that perform multiplication (DSPs), apply logical functions (LUTs), or store memory (BRAM)
- High throughput I/O: O(100) optical transceivers running at O(15) Gbps
- Massively parallel
- Low power



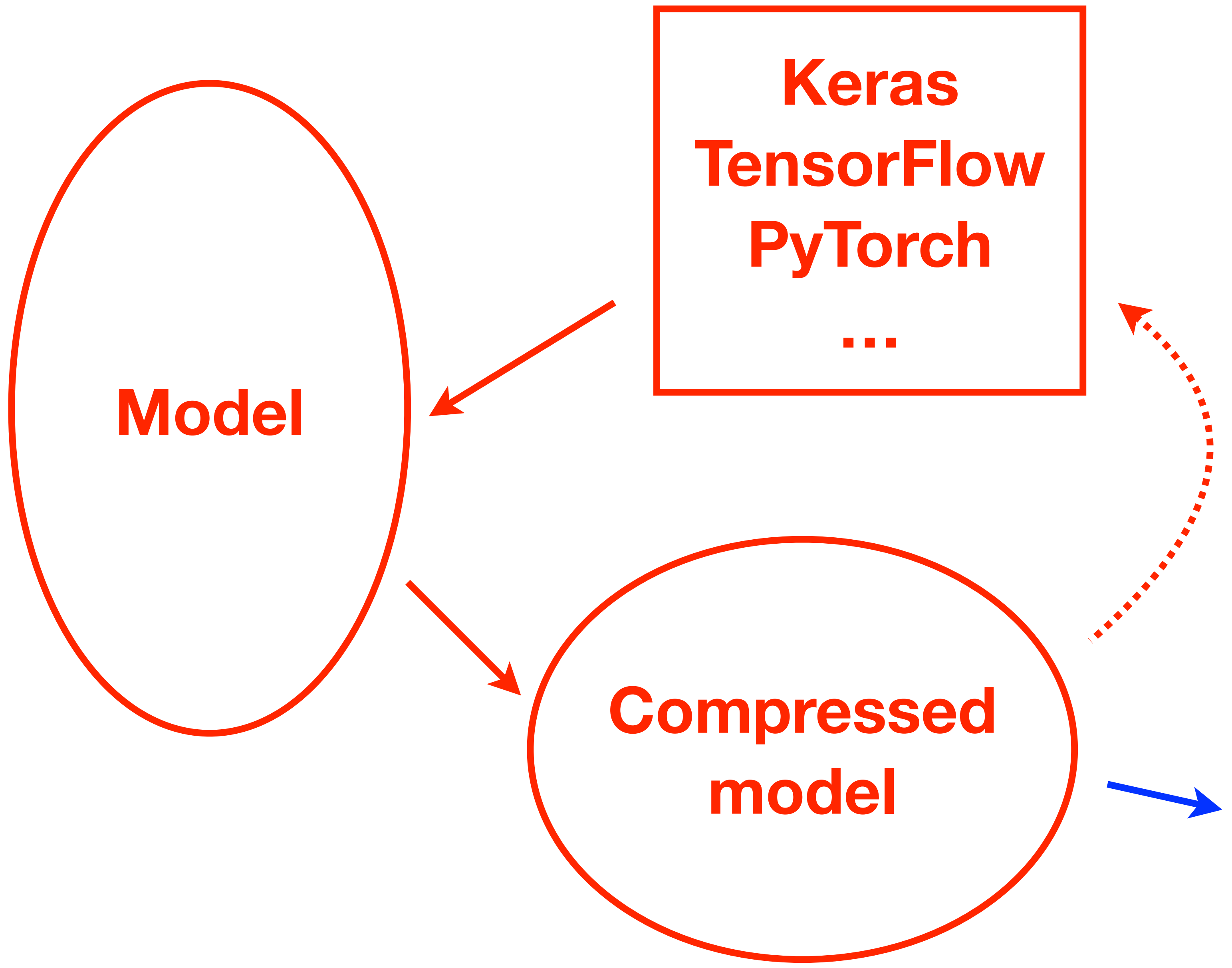
Cons:

- Requires domain knowledge to program (using VHDL/Verilog)

Design Exploration with hls4ml

JINST 13, P07027 (2018)

- hls4ml for scientists or ML experts to translate ML algorithms into RTL firmware

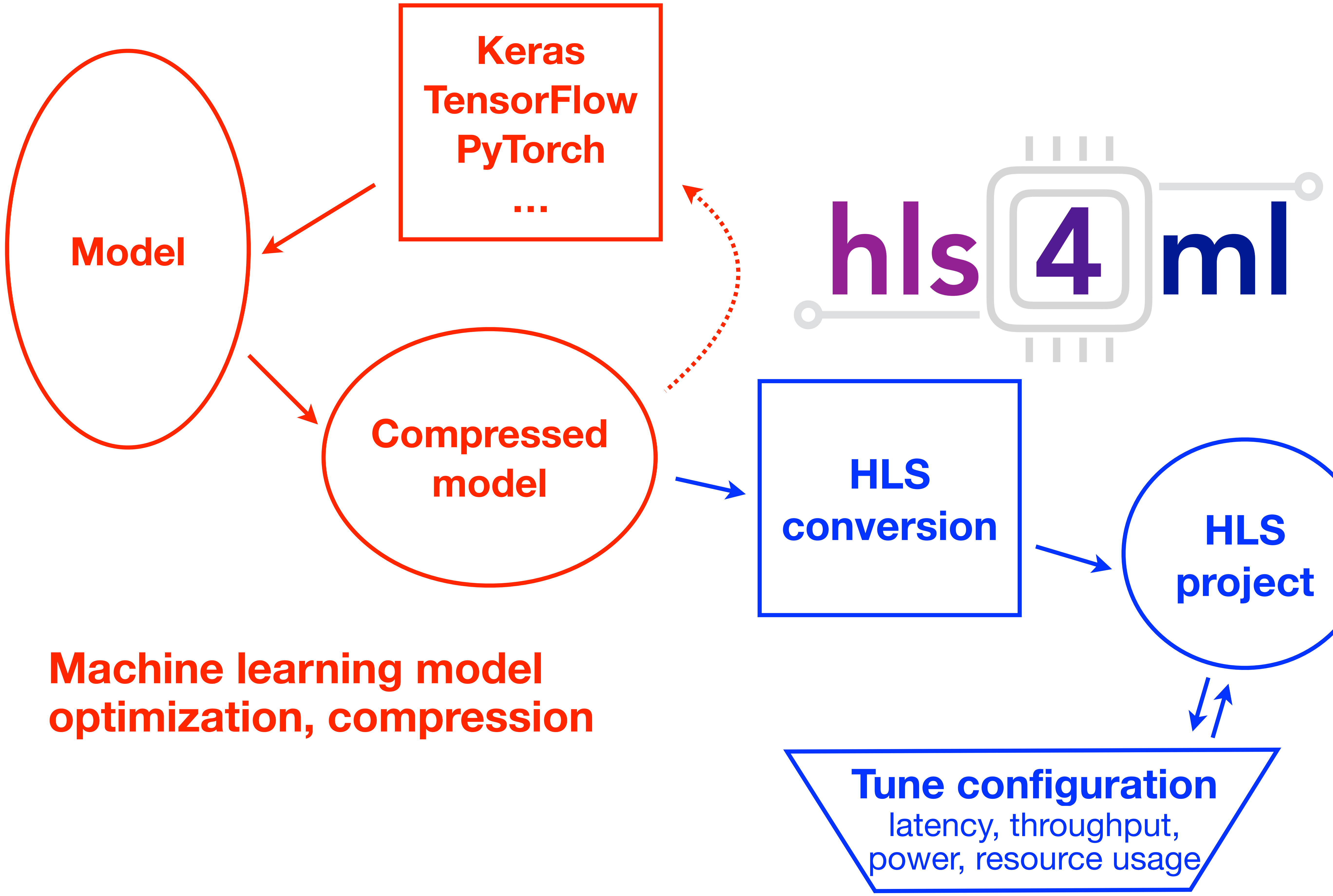


Machine learning model optimization, compression

Design Exploration with hls4ml

JINST 13, P07027 (2018)

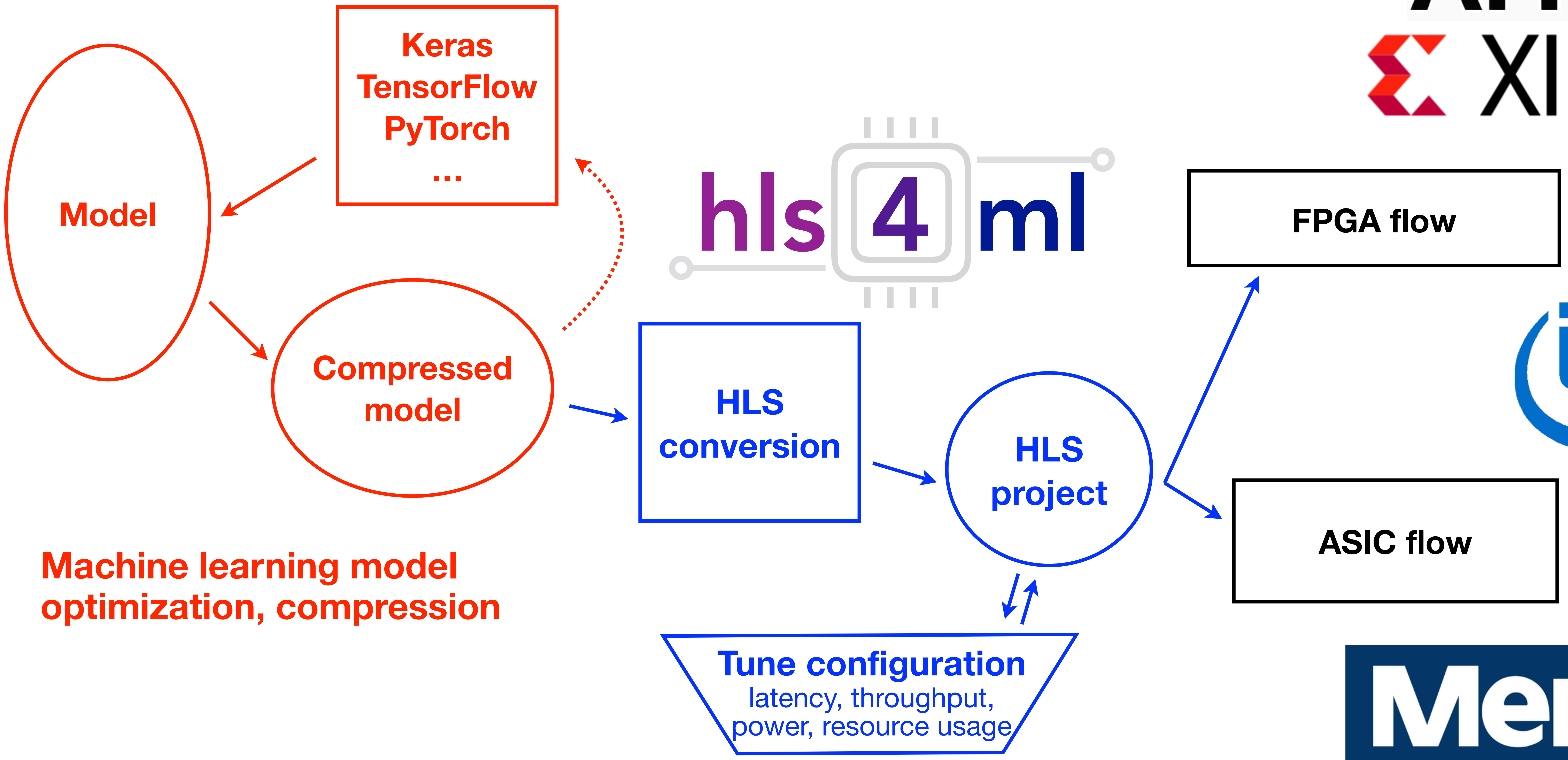
- hls4ml for scientists or ML experts to translate ML algorithms into RTL firmware



Design Exploration with hls4ml

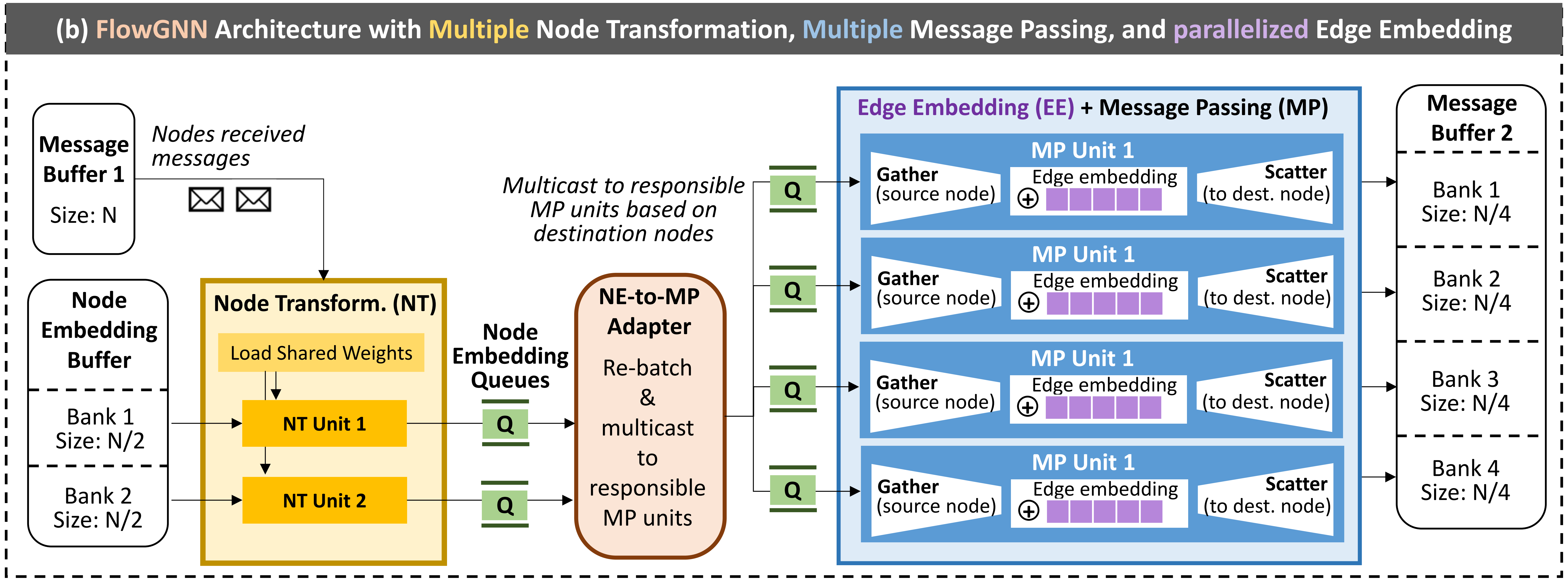
[JINST 13, P07027 (2018)](#)

- [hls4ml](#) for scientists or ML experts to translate ML algorithms into RTL firmware



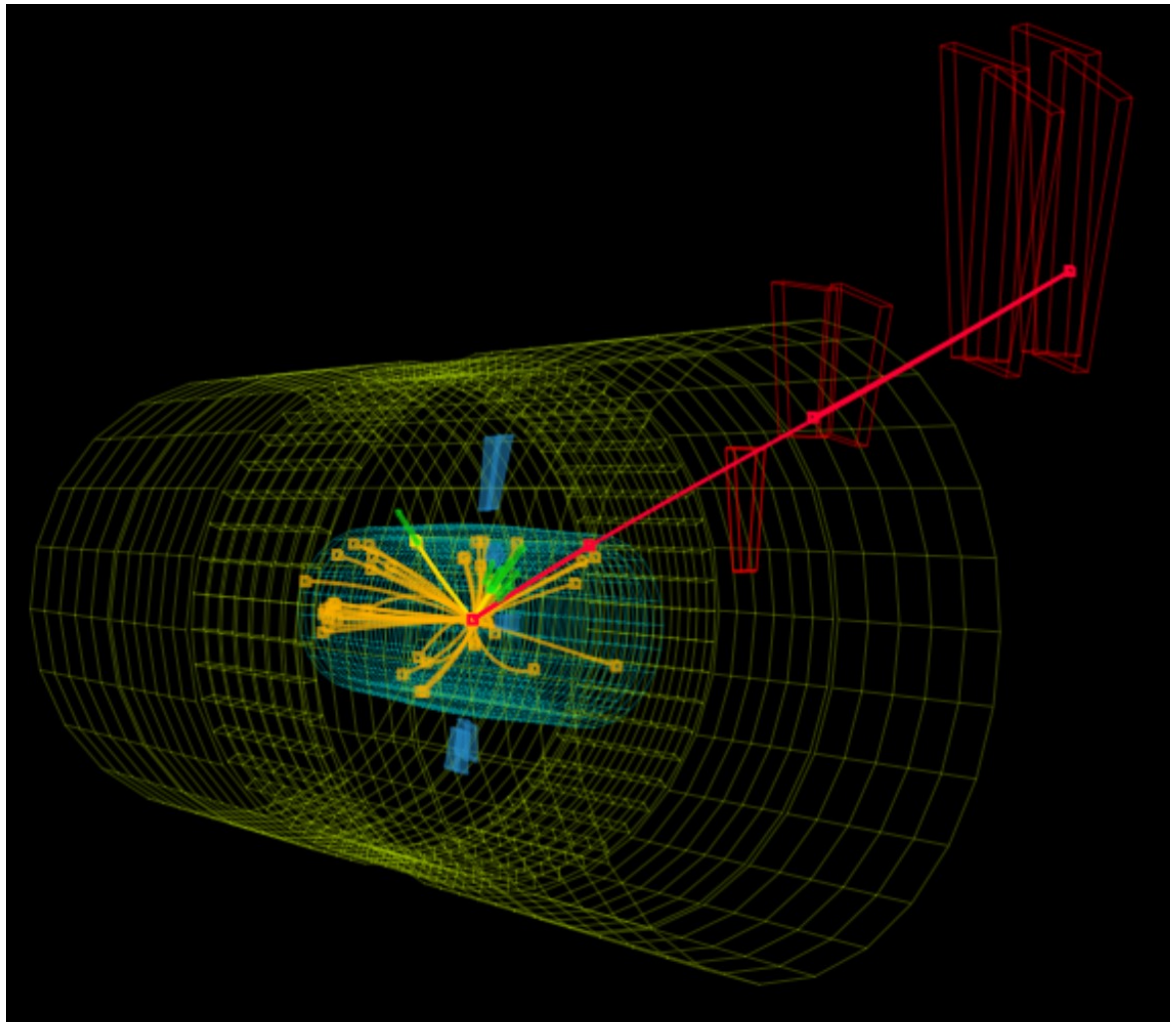
Many tools with different strengths

- FINN (NNs): https://finn.readthedocs.io/en/latest/
- Conifer (BDTs): https://github.com/thesps/conifer
- fwXMachina (BDTs): http://fwx.pitt.edu/
- FlowGNN: https://github.com/sharc-lab/flowgnn

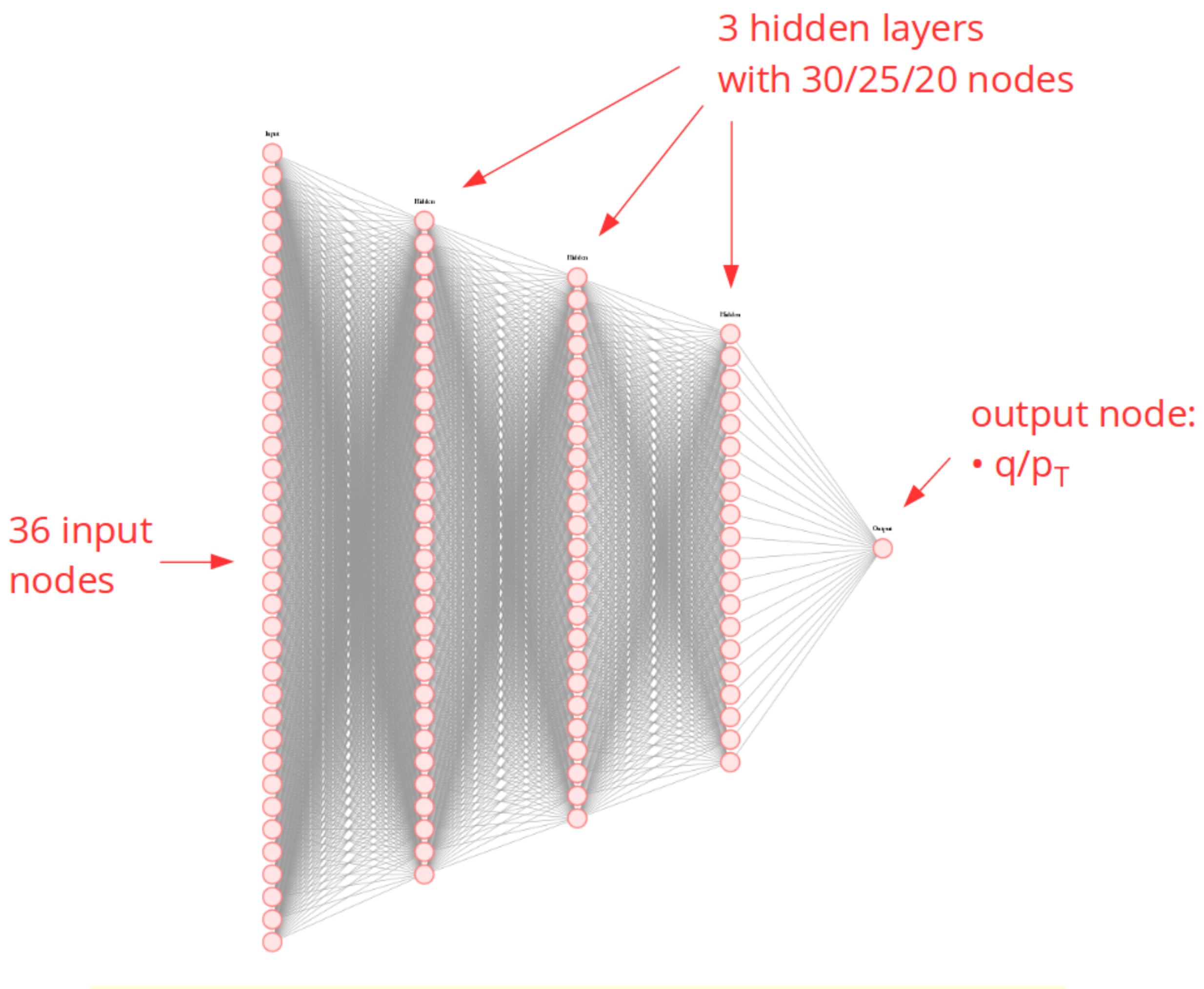


(b) FlowGNN Architecture with Multiple Node Transformation, Multiple Message Passing, and parallelized Edge Embedding

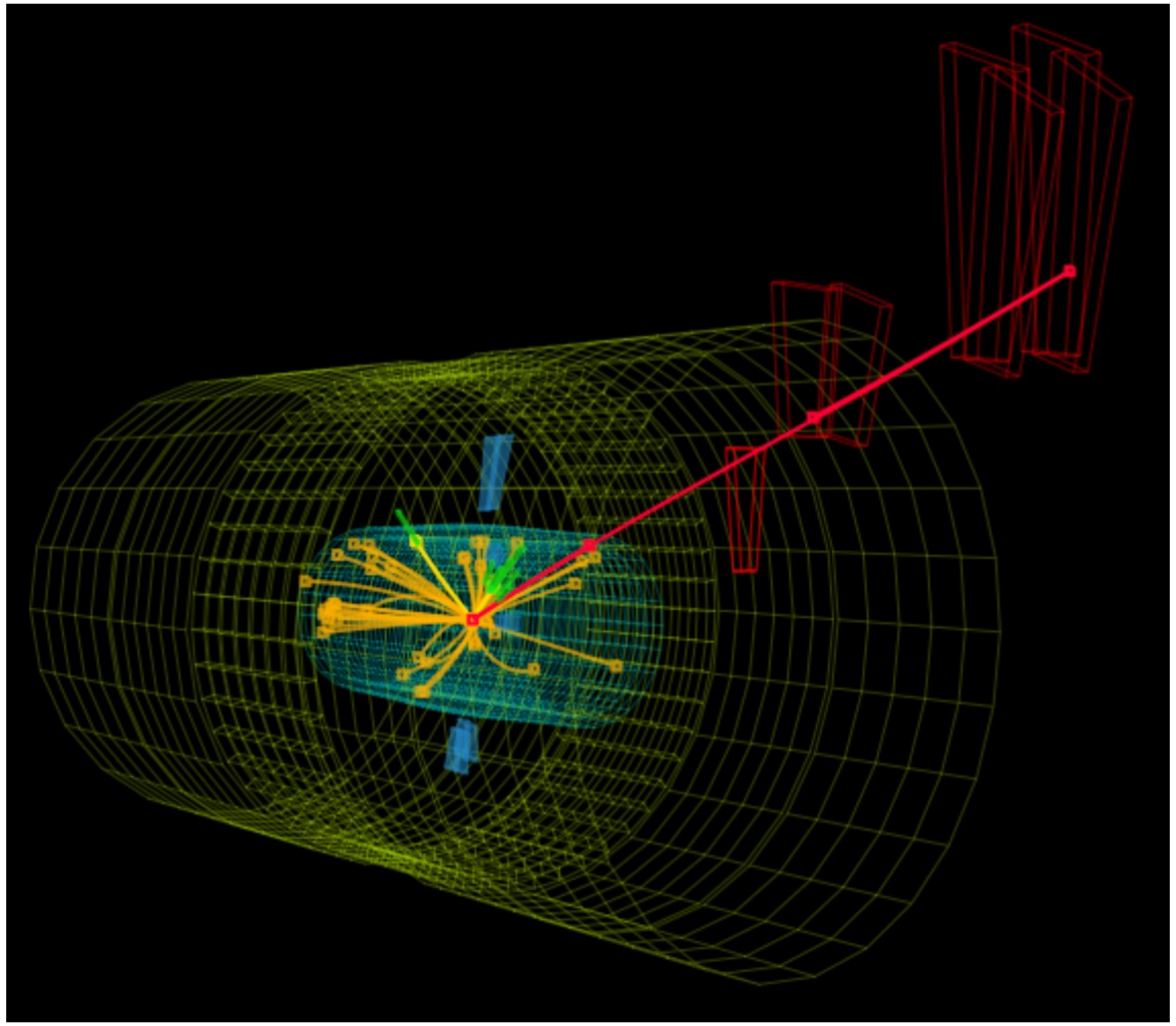
Application: Measure Muon p_T at 40 MHz



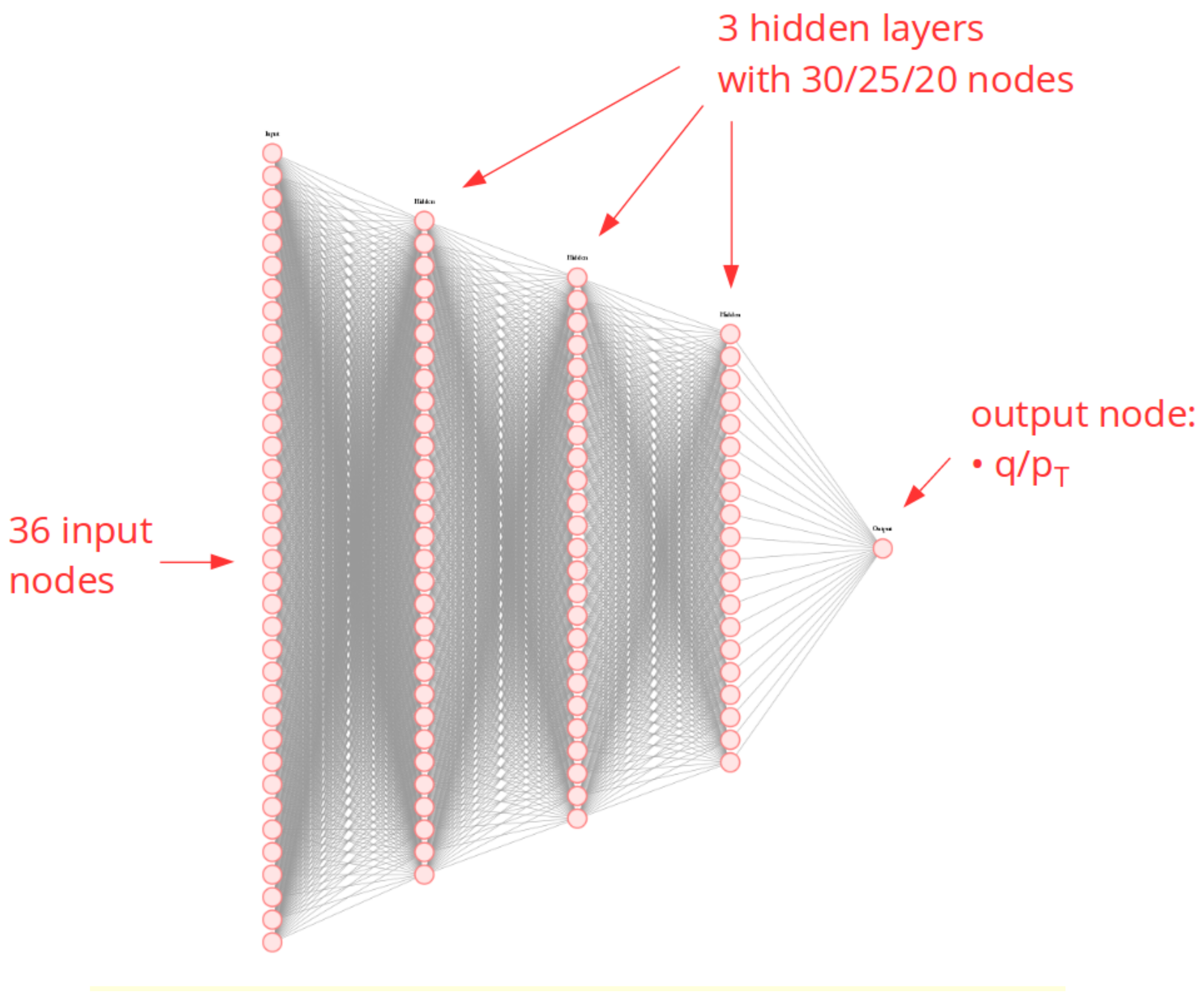
CMS-TDR-021



Application: Measure Muon p_T at 40 MHz

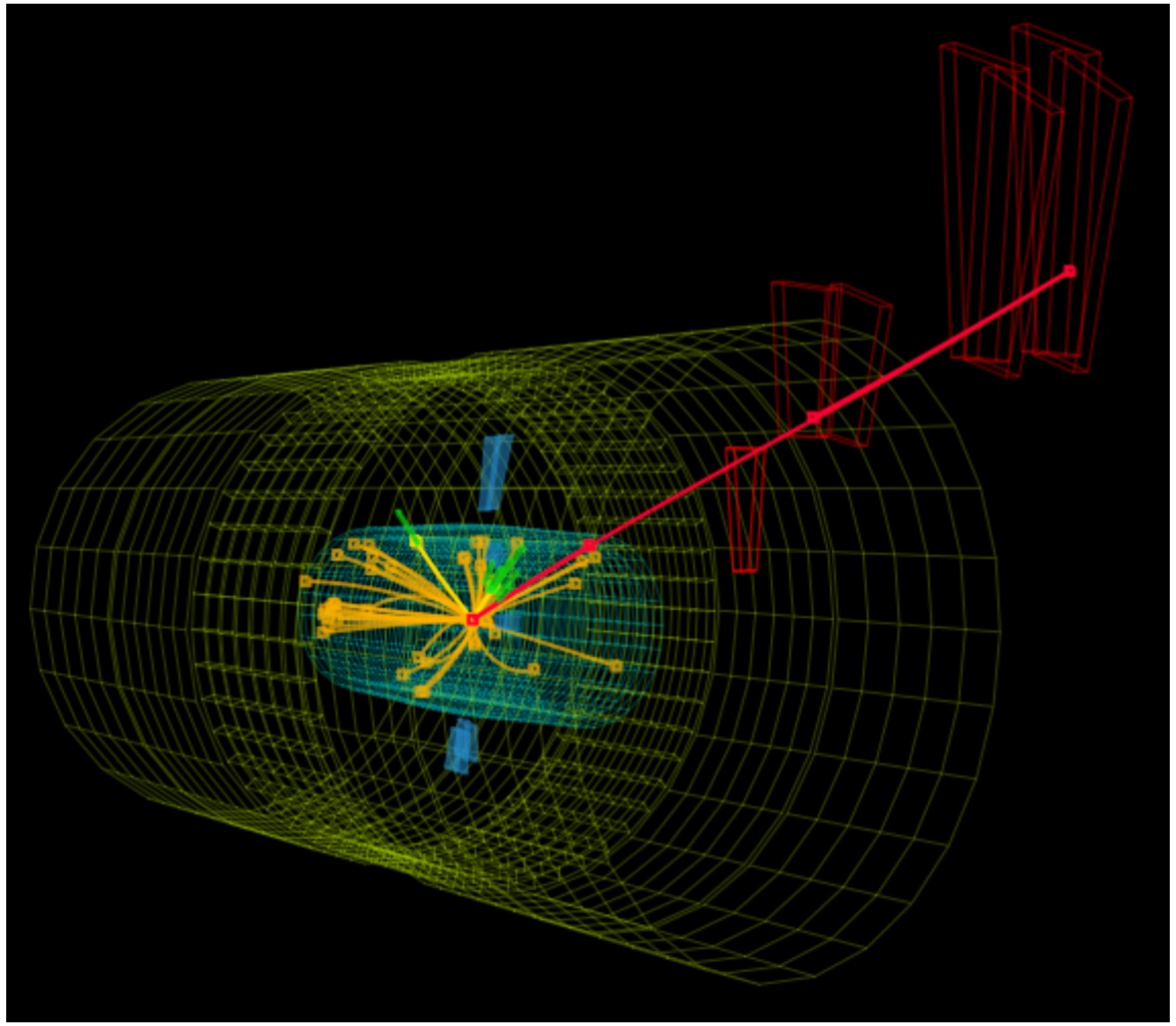


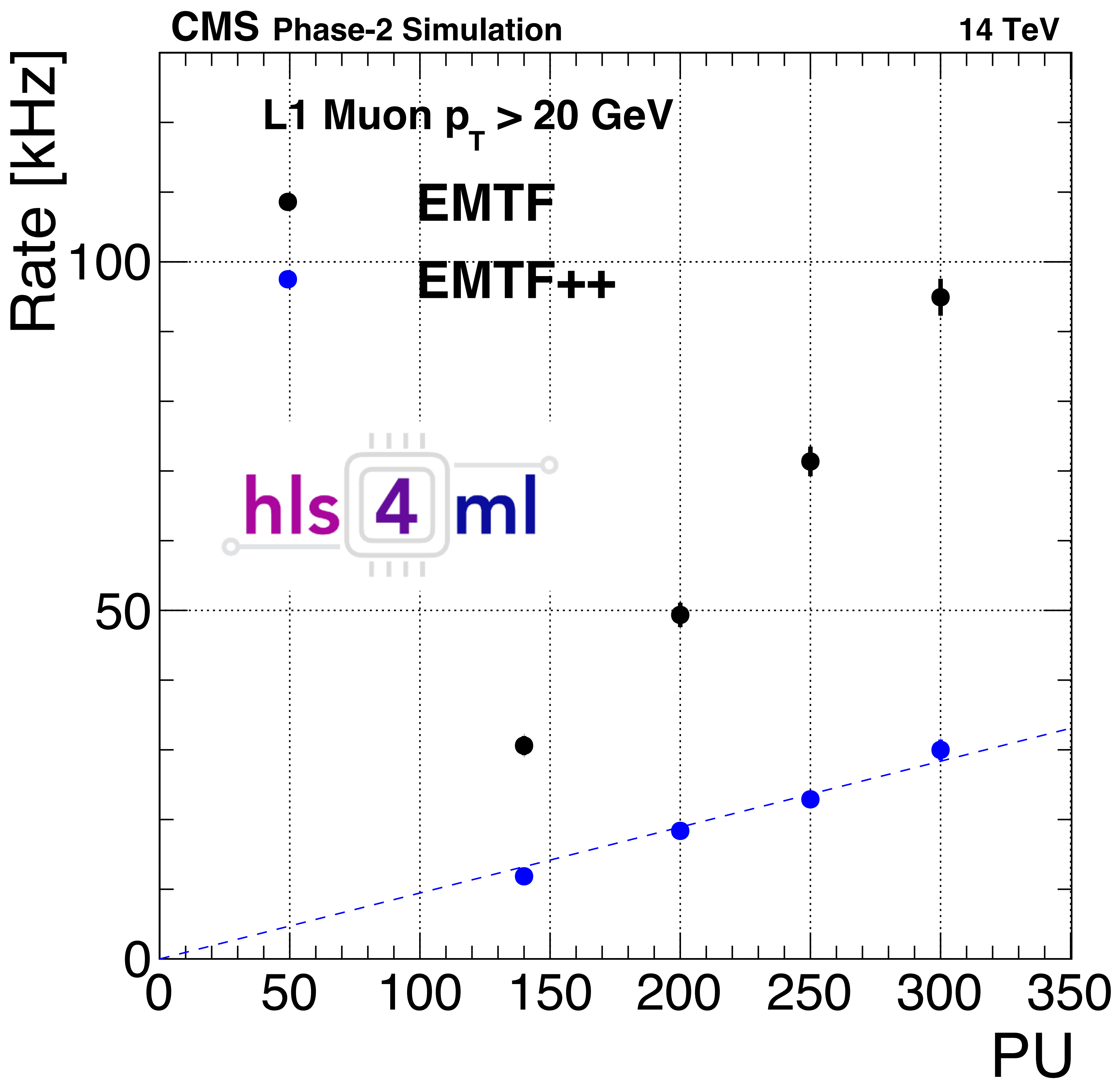
CMS-TDR-021



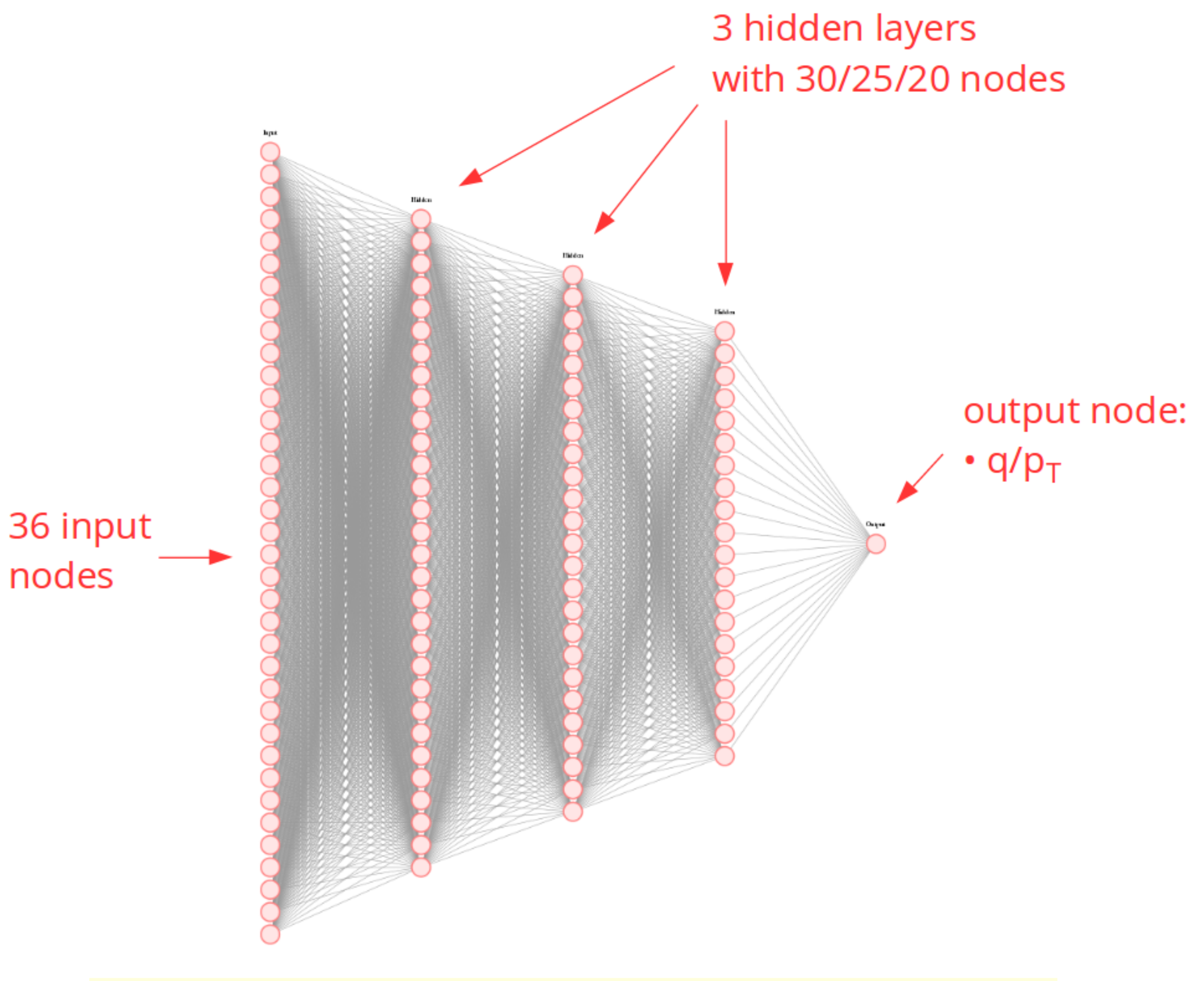
- NN measures muon momentum

Application: Measure Muon p_T at 40 MHz



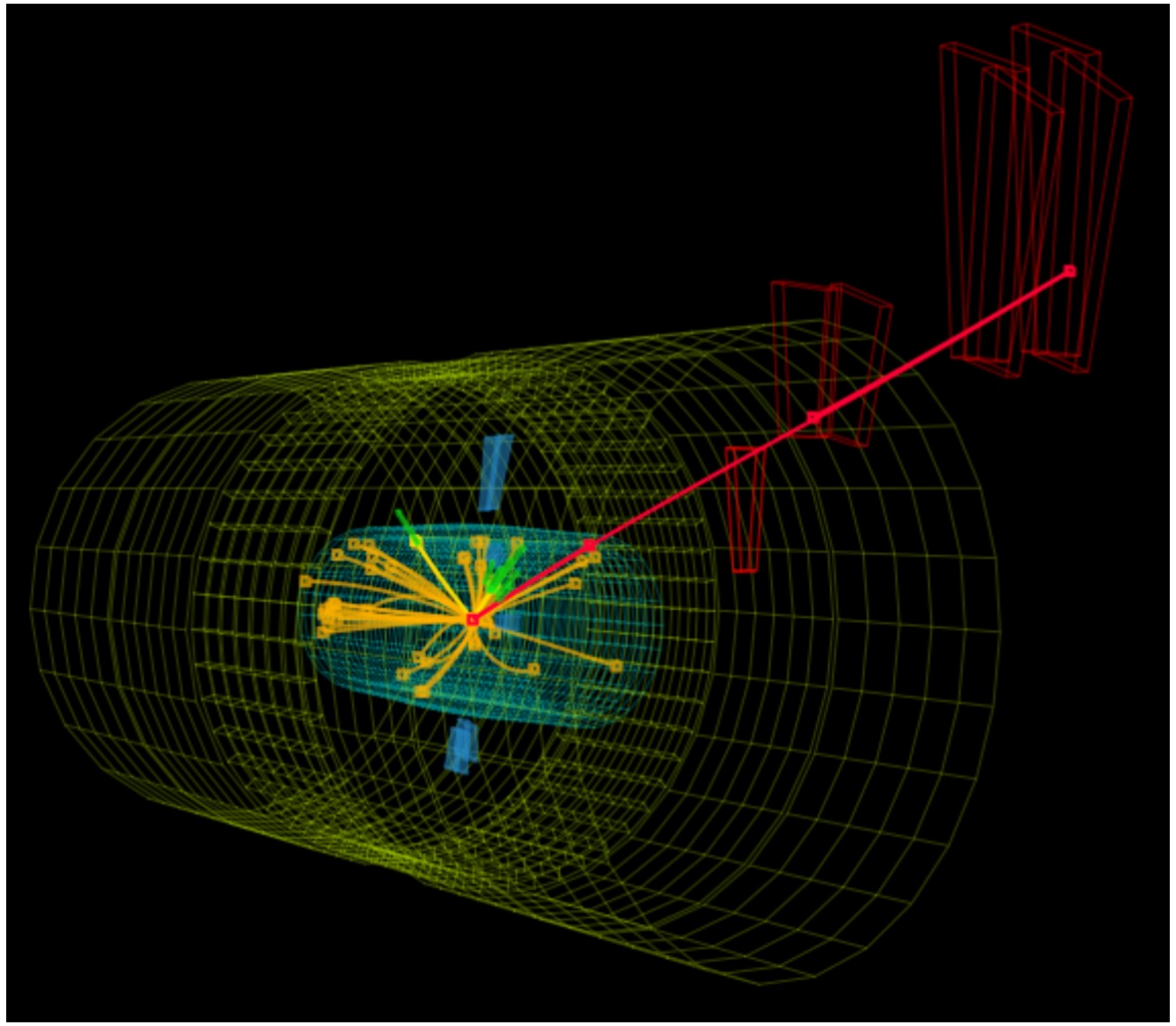


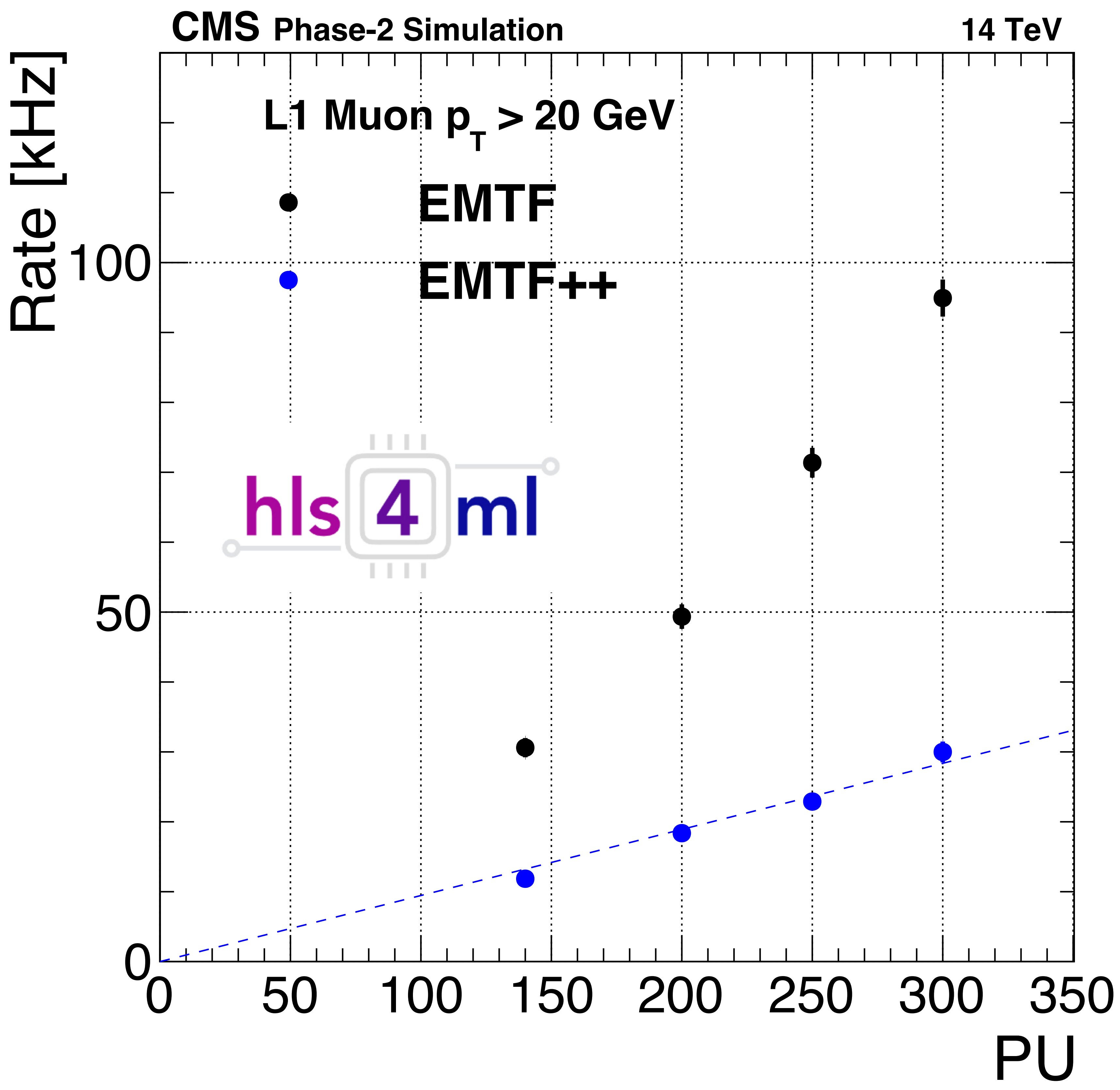
CMS-TDR-021



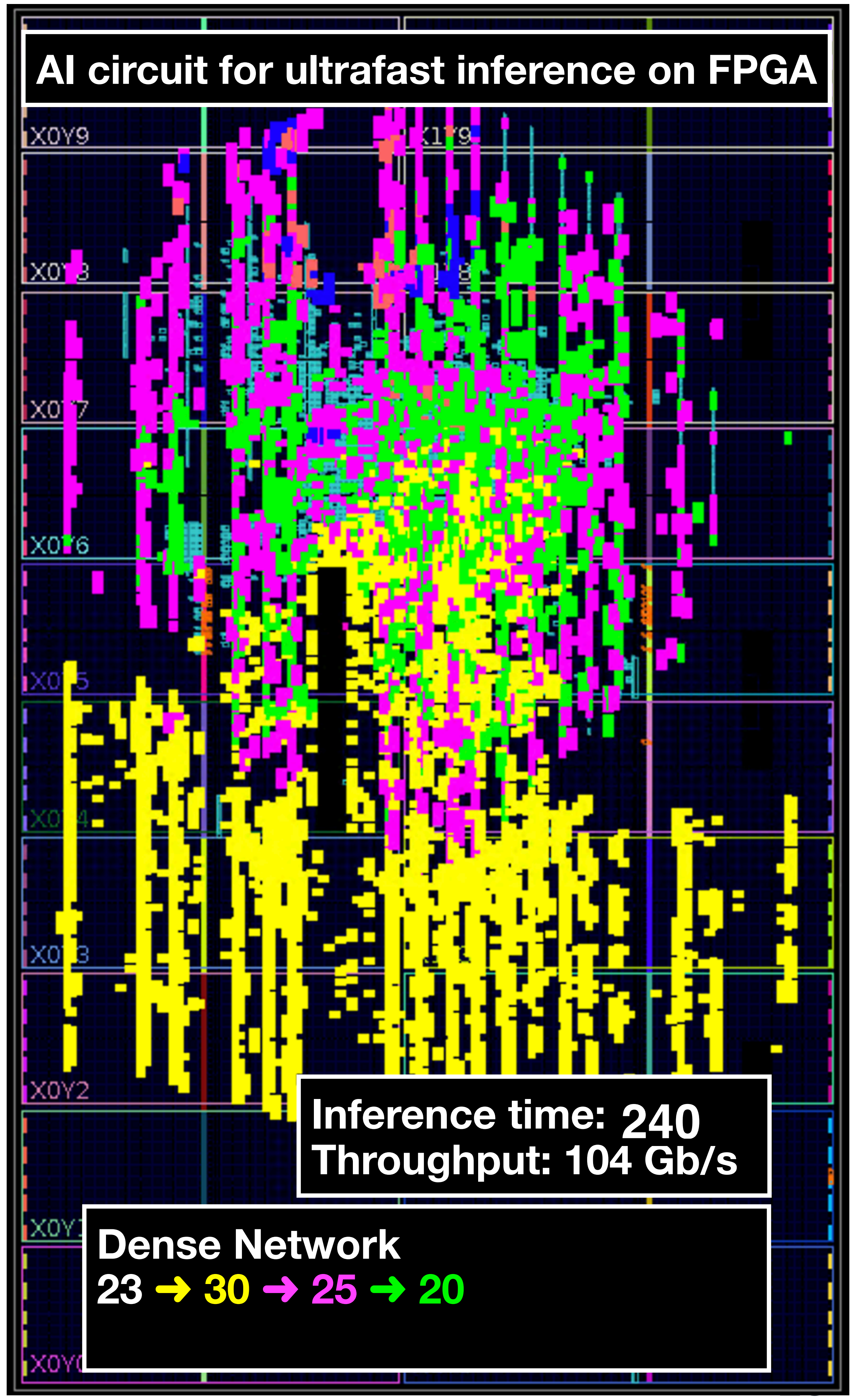
- NN measures muon momentum
  - 3× reduction in the trigger rate for NN!

Application: Measure Muon p_T at 40 MHz





[CMS-TDR-021](CMS-TDR-021)



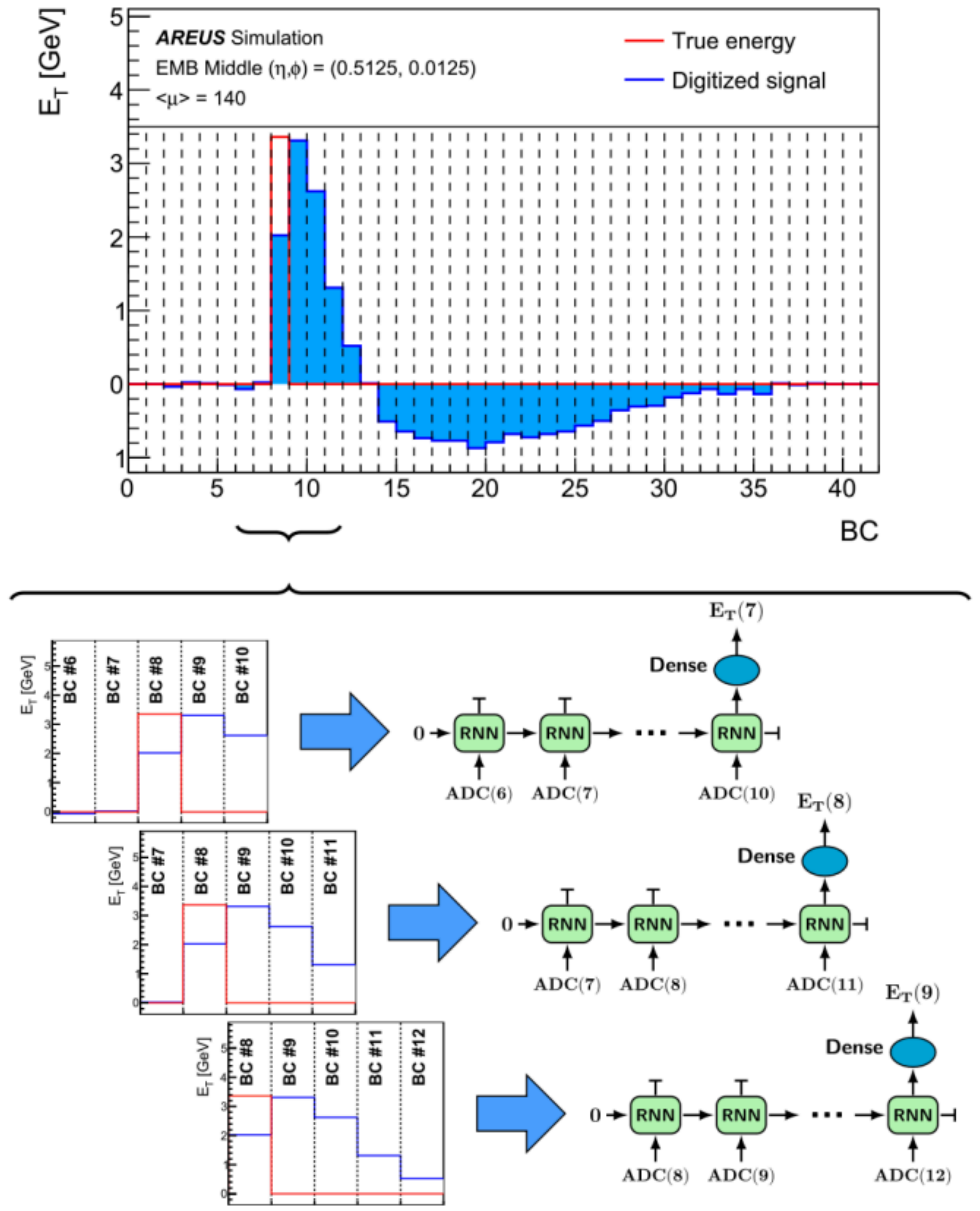
- NN measures muon momentum
  - 3× reduction in the trigger rate for NN!
- Fits within L1 trigger latency (240 ns!) and FPGA resource requirements (less then 30%)

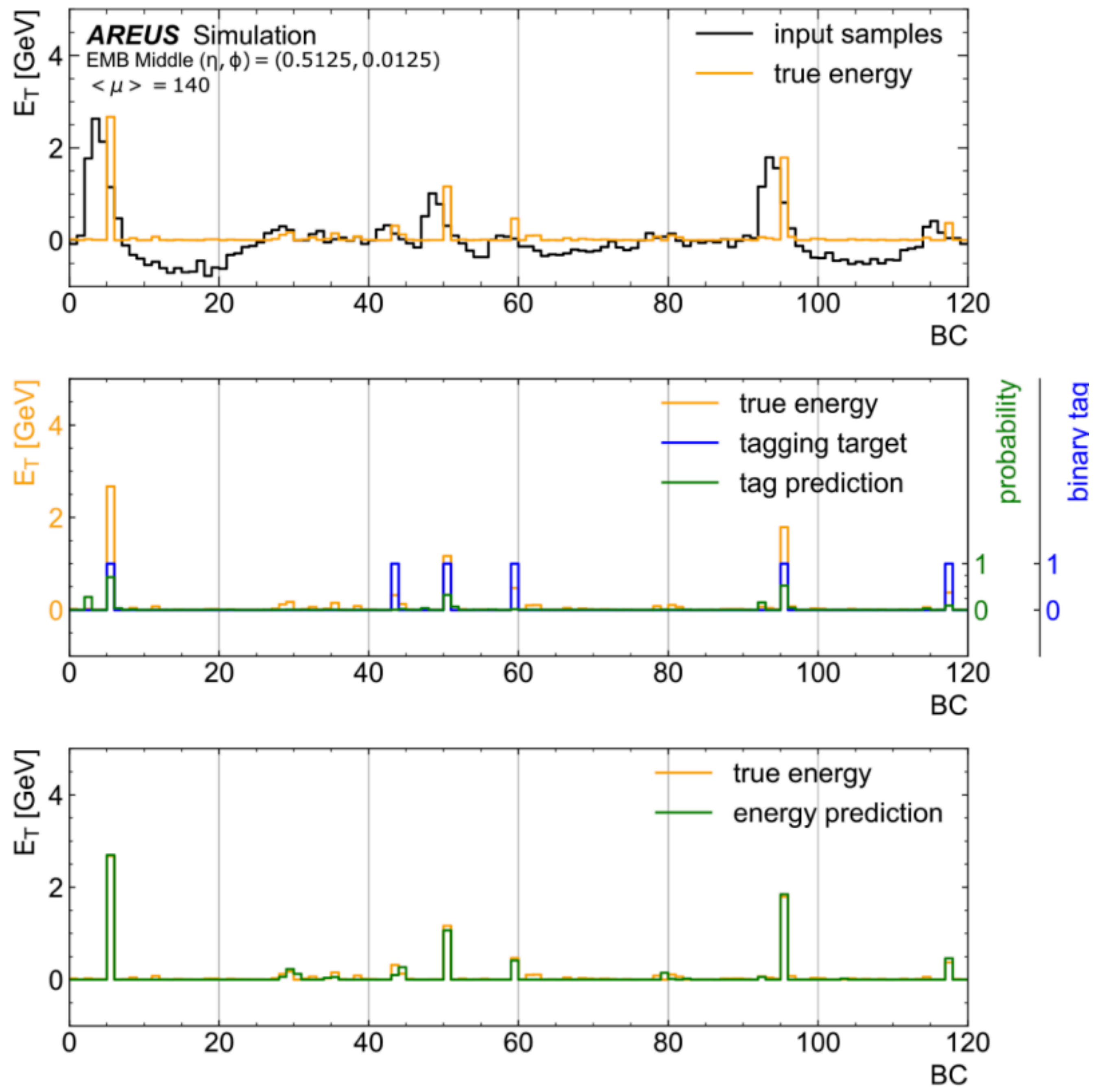
Application: ATLAS LAr Calorimeter

Convolutional and **Recurrent Neural Networks**
for real-time energy reconstruction of ATLAS LAr Calorimeter for Phase 2

- Up to around 600 calorimeter channels processed by on device
- 200 ns latency of predictions

- Implemented on Intel FPGAs (previous examples are all AMD)
  - Team contributed majorly to RNN and Intel implementations of hls4ml





[10.1007/s41781-021-00066-y](https://doi.org/10.1007/s41781-021-00066-y)

Application: Anomaly Detection

Nat. Mach. Intell. 4, 154 (2022)
Data challenge: mpp-hep.github.io/ADC2021

Application: Anomaly Detection

Nat. Mach. Intell. 4, 154 (2022)

Data challenge: mpp-hep.github.io/ADC2021

- **Challenge:** if new physics has an unexpected signature that doesn't align with existing triggers, precious BSM events may be discarded at trigger level

Application: Anomaly Detection

Nat. Mach. Intell. 4, 154 (2022)

Data challenge: mpp-hep.github.io/ADC2021

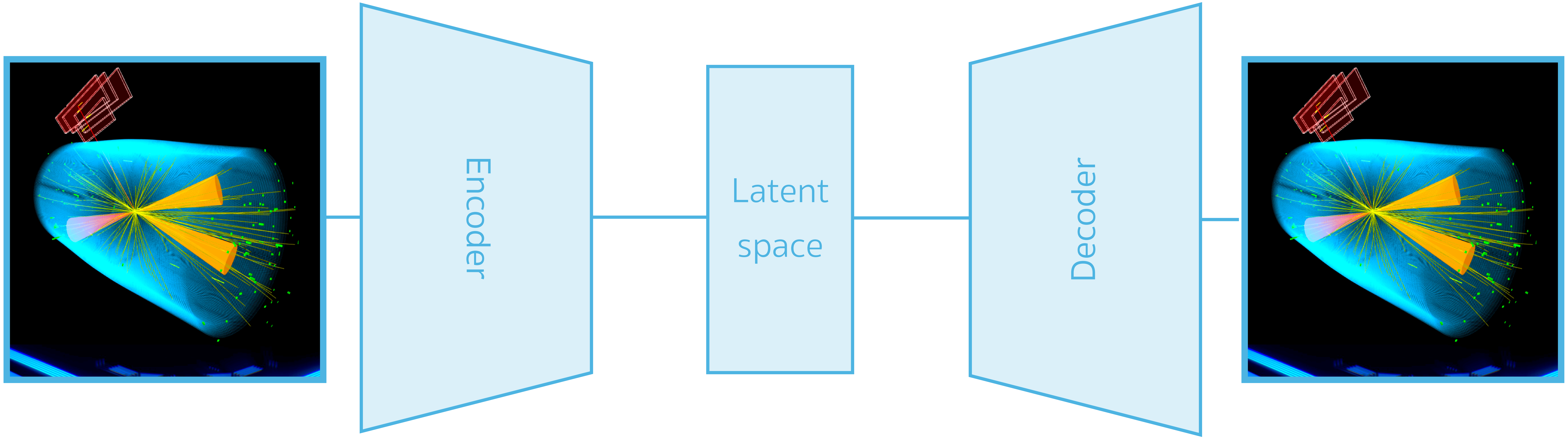
- **Challenge:** if new physics has an unexpected signature that doesn't align with existing triggers, precious BSM events may be discarded at trigger level
- Can we use unsupervised algorithms to detect non-SM-like anomalies?

Application: Anomaly Detection

Nat. Mach. Intell. 4, 154 (2022)

Data challenge: mpp-hep.github.io/ADC2021

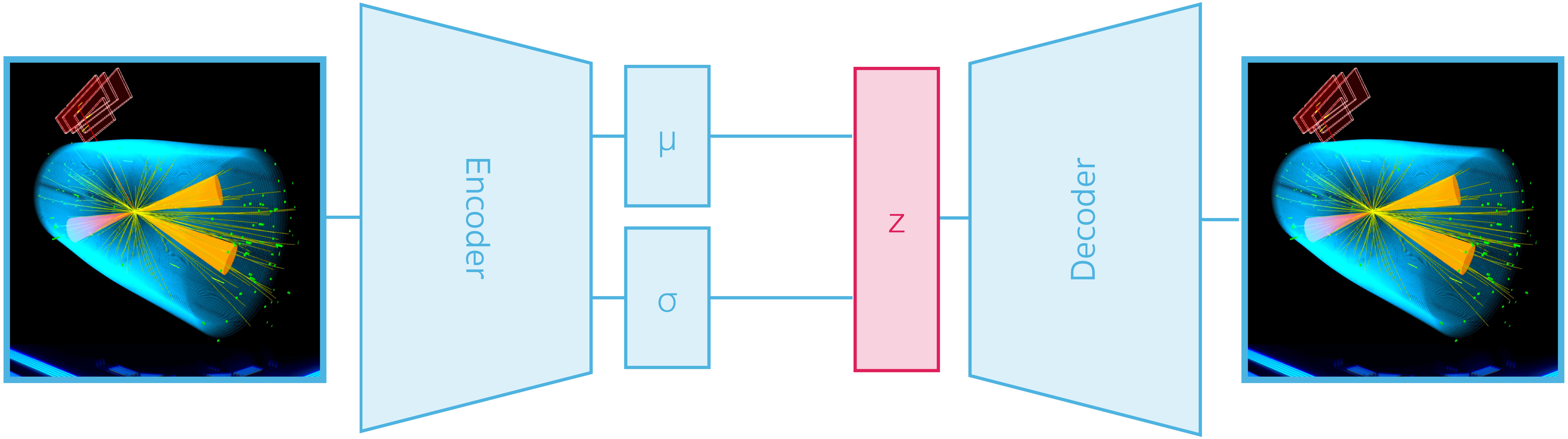
- **Challenge:** if new physics has an unexpected signature that doesn't align with existing triggers, precious BSM events may be discarded at trigger level
- Can we use unsupervised algorithms to detect non-SM-like anomalies?
  - **Autoencoders (AEs):** compress input to a smaller dimensional latent space then decompress and calculate difference



Application: Anomaly Detection

Nat. Mach. Intell. 4, 154 (2022)
Data challenge: mpp-hep.github.io/ADC2021

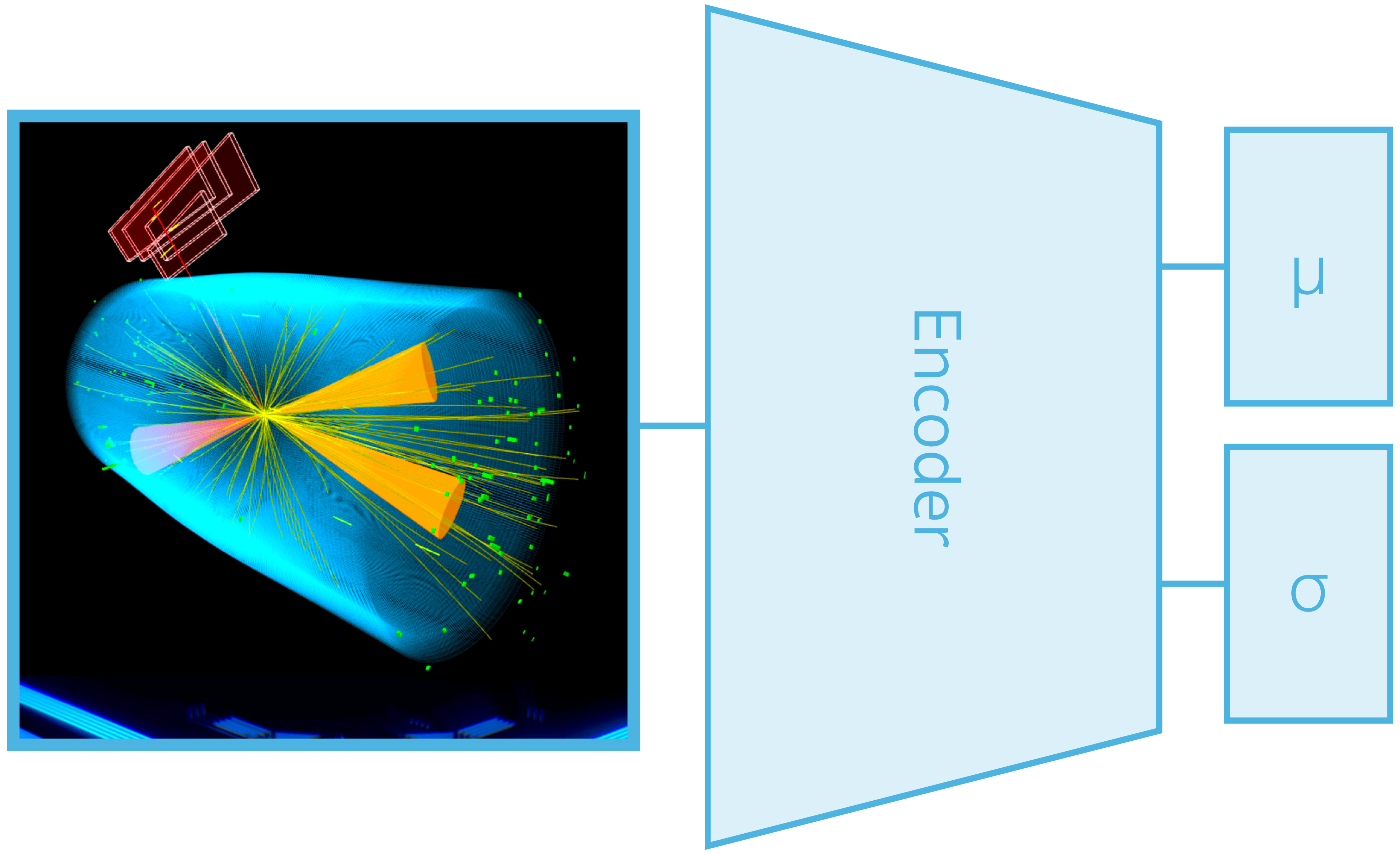
- **Challenge:** if new physics has an unexpected signature that doesn't align with existing triggers, precious BSM events may be discarded at trigger level
- Can we use unsupervised algorithms to detect non-SM-like anomalies?
  - **Autoencoders (AEs):** compress input to a smaller dimensional latent space then decompress and calculate difference
  - **Variational autoencoders (VAEs):** model the latent space as a probability distribution; possible to detect anomalies purely with latent space variables



Application: Anomaly Detection

Nat. Mach. Intell. 4, 154 (2022)
Data challenge: mpp-hep.github.io/ADC2021

- **Challenge:** if new physics has an unexpected signature that doesn't align with existing triggers, precious BSM events may be discarded at trigger level
- Can we use unsupervised algorithms to detect non-SM-like anomalies?
  - **Autoencoders (AEs):** compress input to a smaller dimensional latent space then decompress and calculate difference
  - **Variational autoencoders (VAEs):** model the latent space as a probability distribution; possible to detect anomalies purely with latent space variables



Key observation: Can build an anomaly score from the latent space of VAE directly! No need to run decoder!

$$R_z = \sum_i \frac{\mu_i^2}{\sigma_i^2}$$

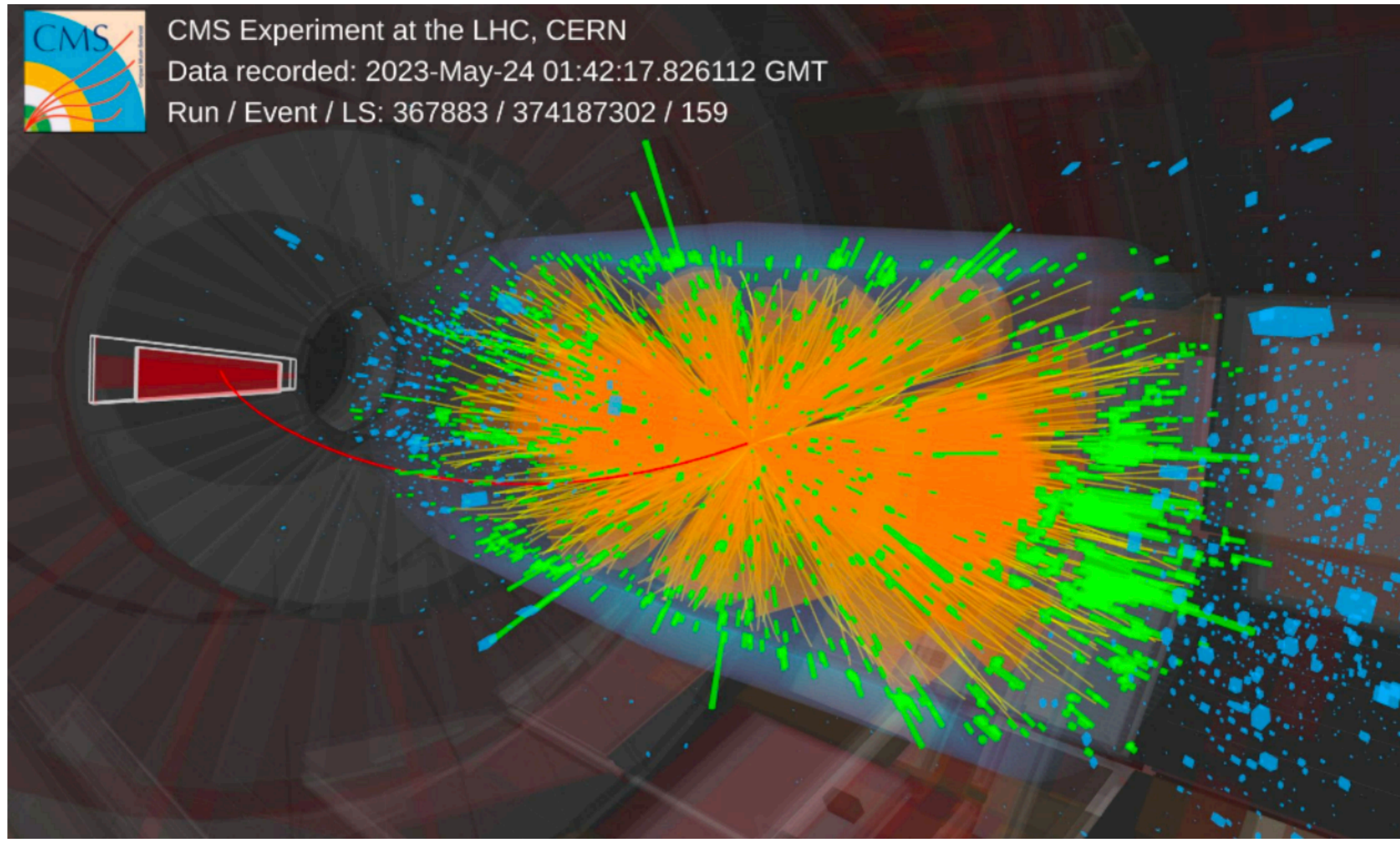
Application: CMS Anomaly Trigger

CMS has implemented a similar idea: **AXOL1TL**

- L1 Hardware implemented VAE-based AD trigger (based on https://arxiv.org/abs/2108.03986)
- Trained on 2018 zerobias data, ran in 2023 Global Trigger Test Crate
- CMS is also developing CICADA, a calorimeter only AD trigger

Similar effort is ongoing in ATLAS

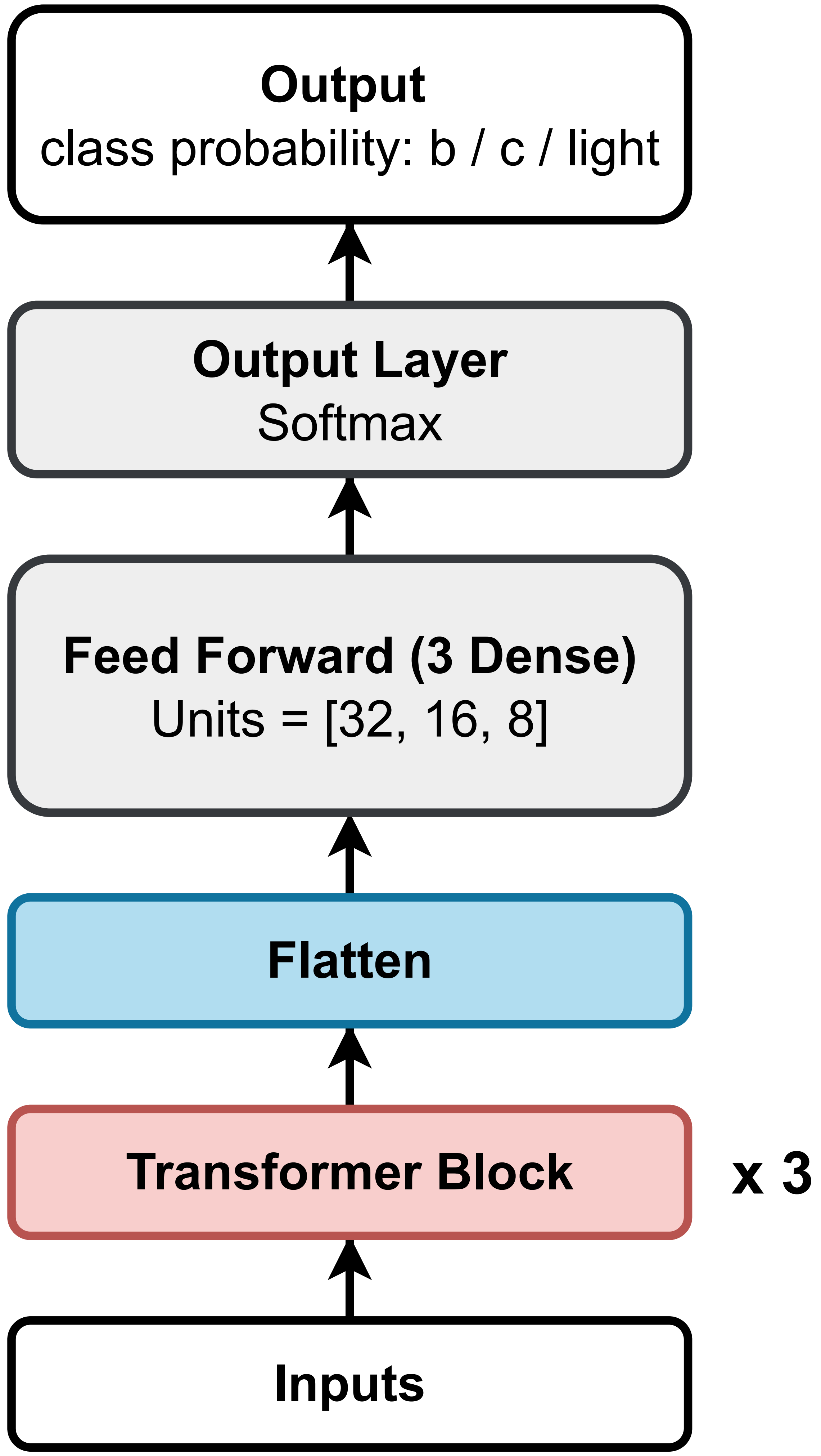
Event display of the highest anomaly score



CMS-DP-2023-079

Low-latency Transformers

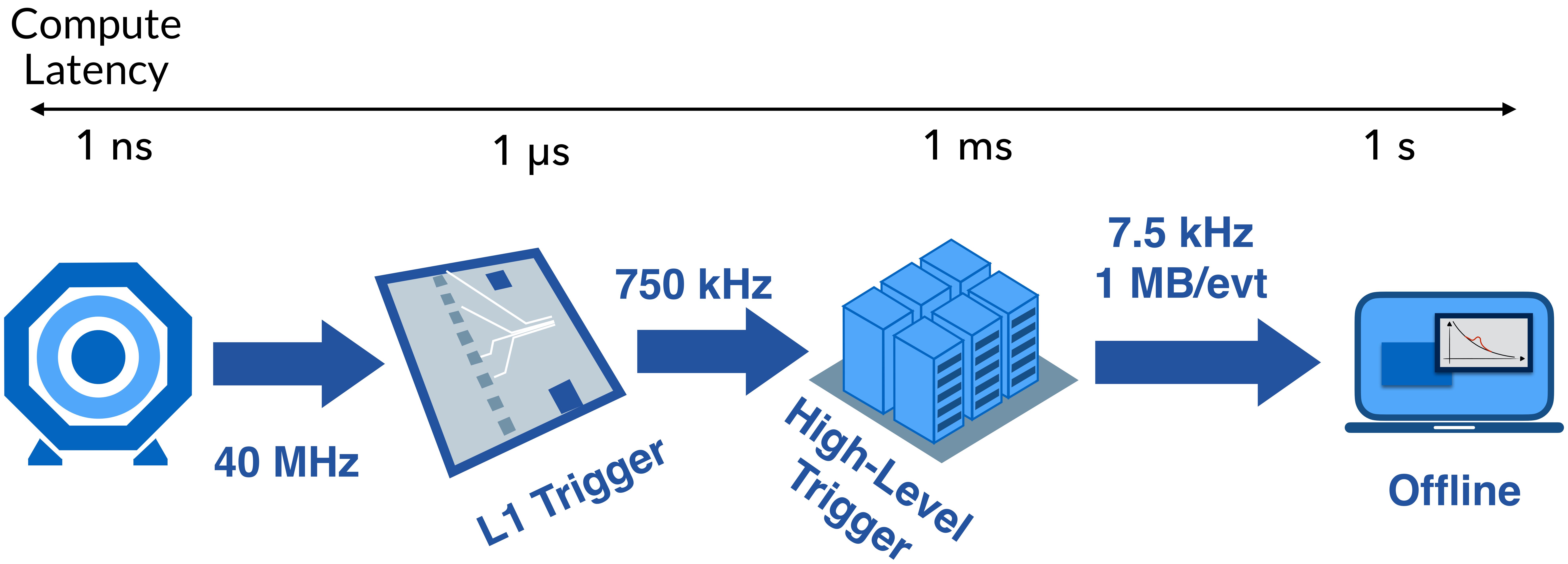
arXiv:2402.01047



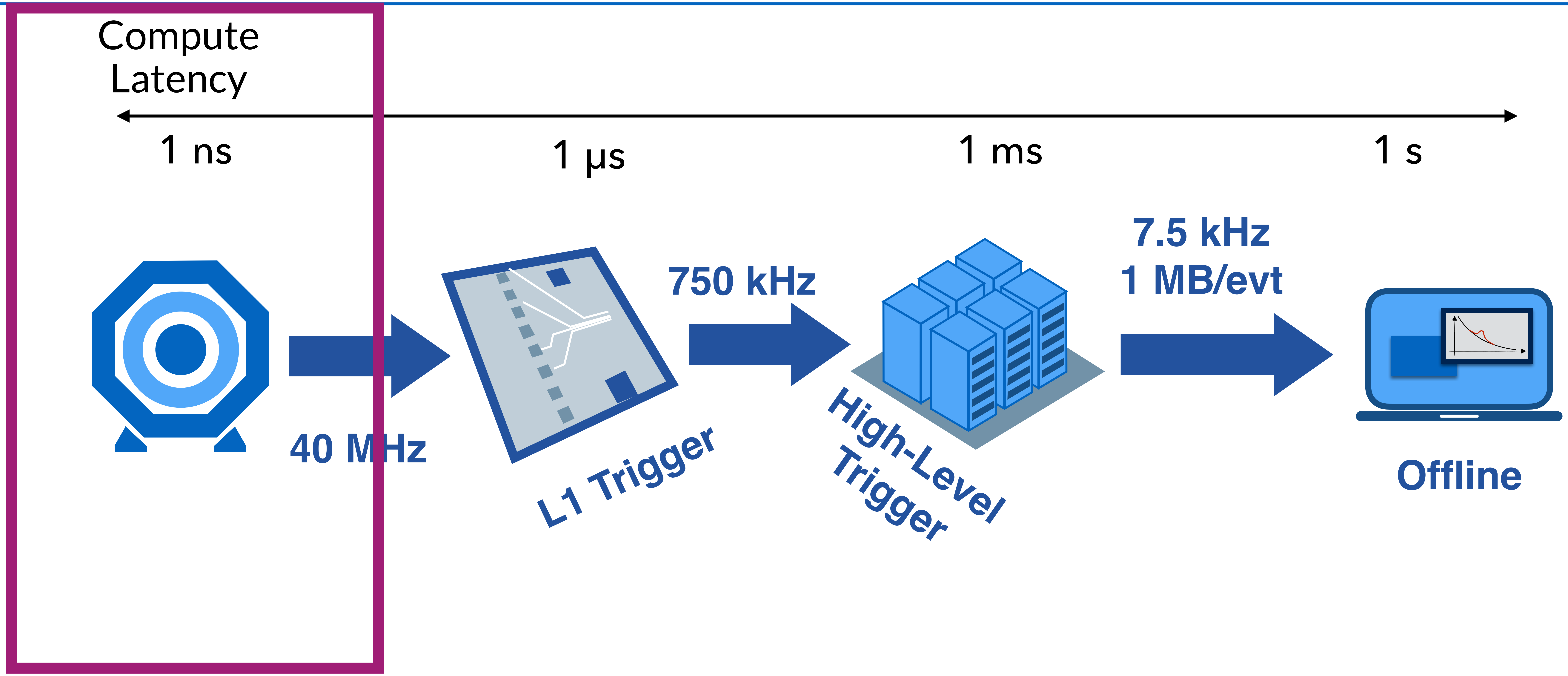
Observed Inference Latency ~ 2-6 μs

Reuse and clk	Interval (cycle)	Latency (cycles)	Latency(time)
R1 (6.577 ns)	49	269	2.077 us
R2 (6.215 ns)	65	449	3.467 us
R4 (4.723 ns)	100	768	5.853 us

HL-LHC Data Processing

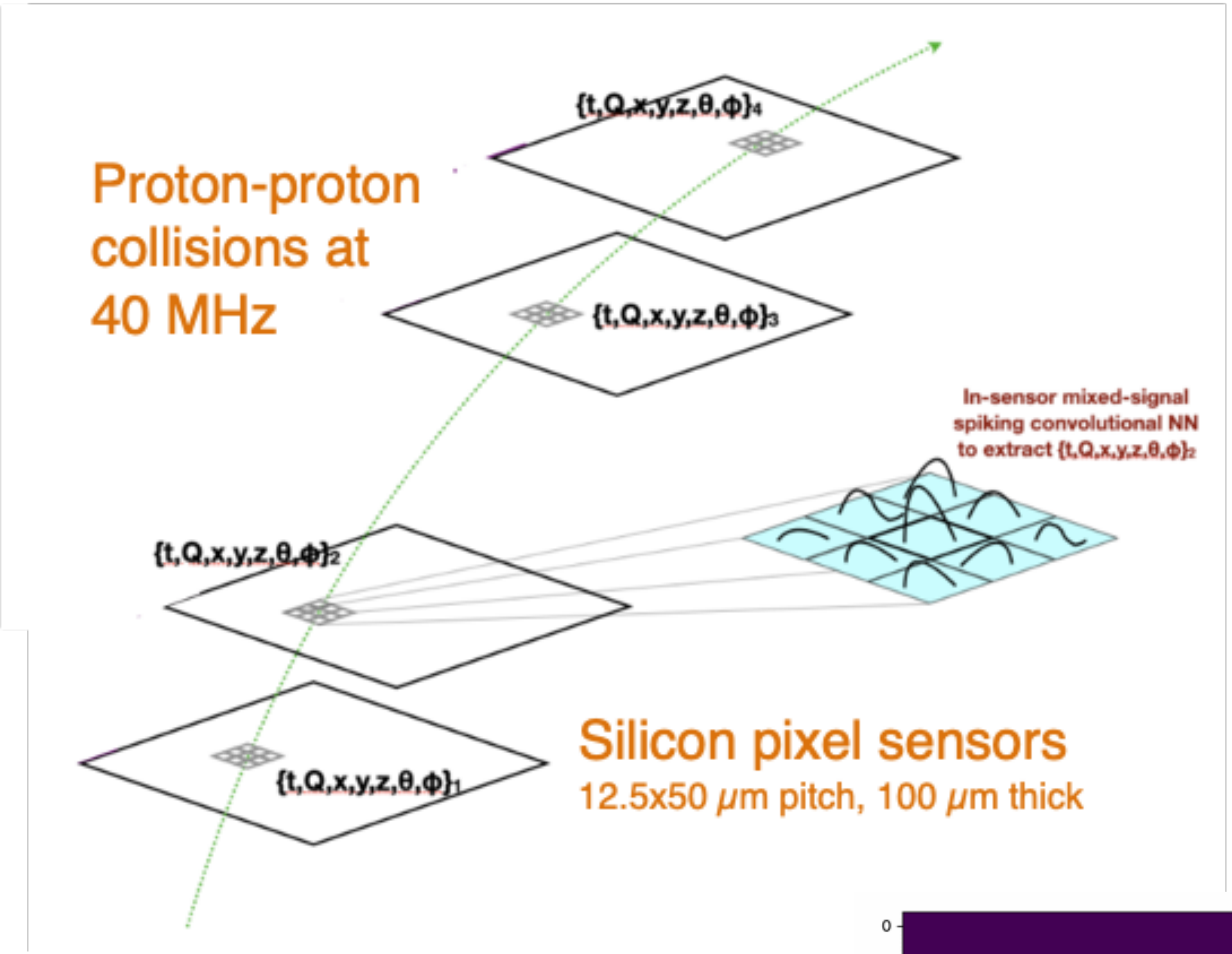


HL-LHC Data Processing



Smart Pixel

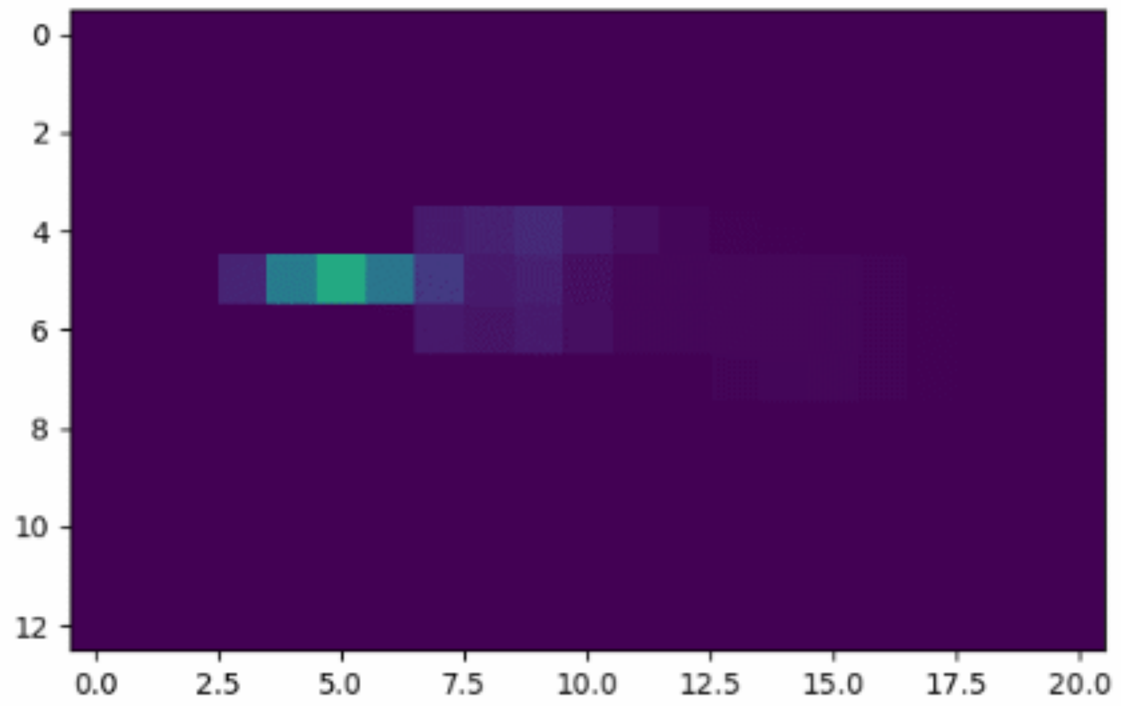
arXiv:2310.02474
arXiv:2312.11676



Data reduction and reconstruction on sensor for silicon pixel detectors

We can reduce the data rate read out by a futuristic pixel detector using AI on-chip

- Factor of ~20 from pT filter
- Additional savings from compression

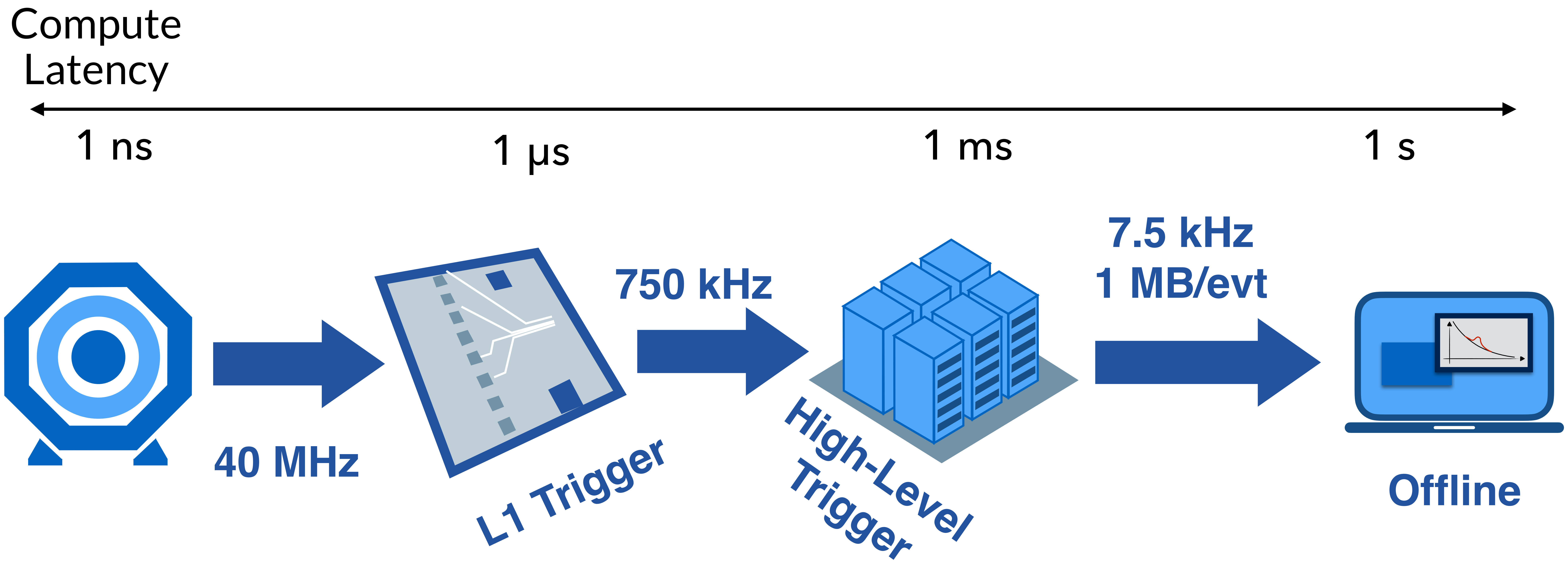


State-of-the-art dataset for developing algorithms for implementation on-ASIC

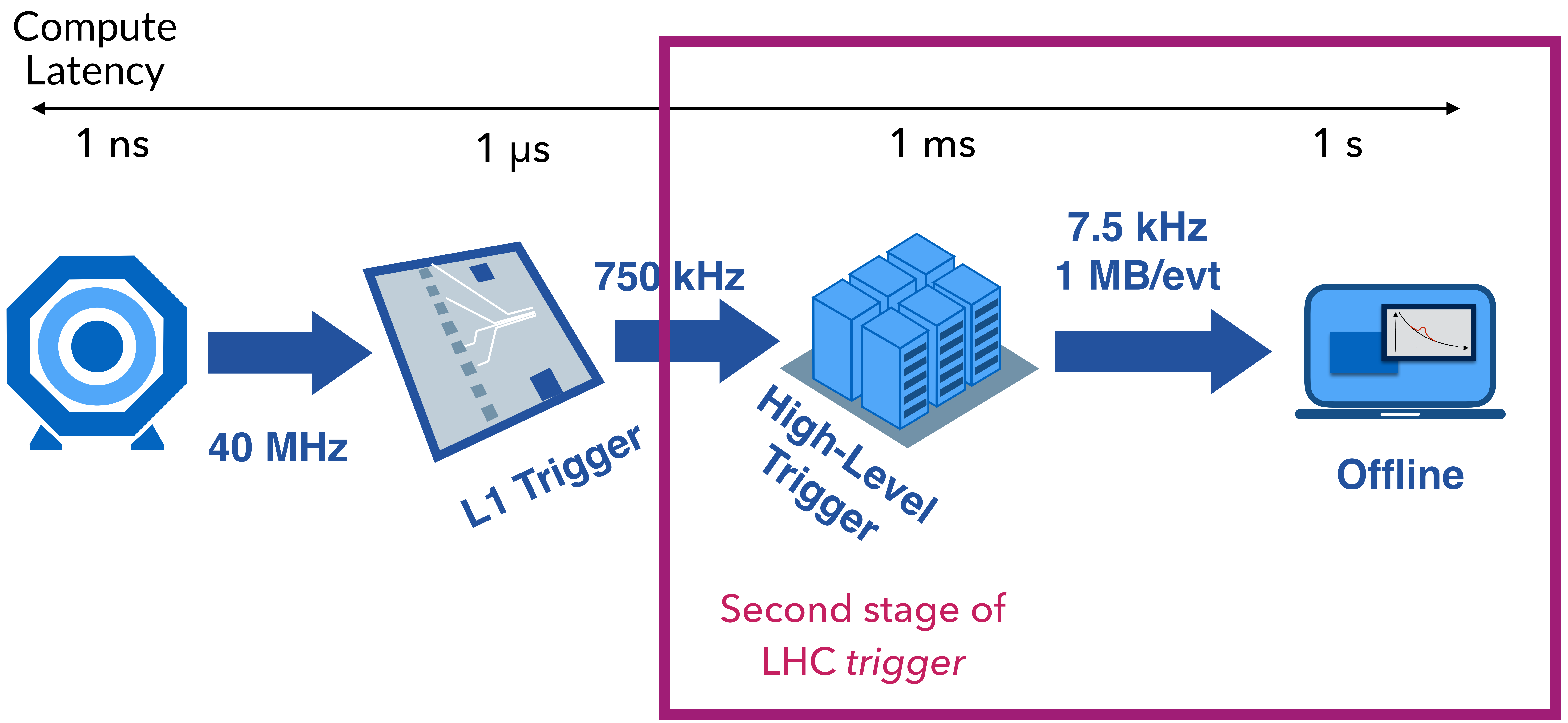
- Simulated MIP interactions in a futuristic pixel detector

Dataset available on **zenodo**

HL-LHC Data Processing



HL-LHC Data Processing



Computing Hardware

Second stage of LHC *trigger*

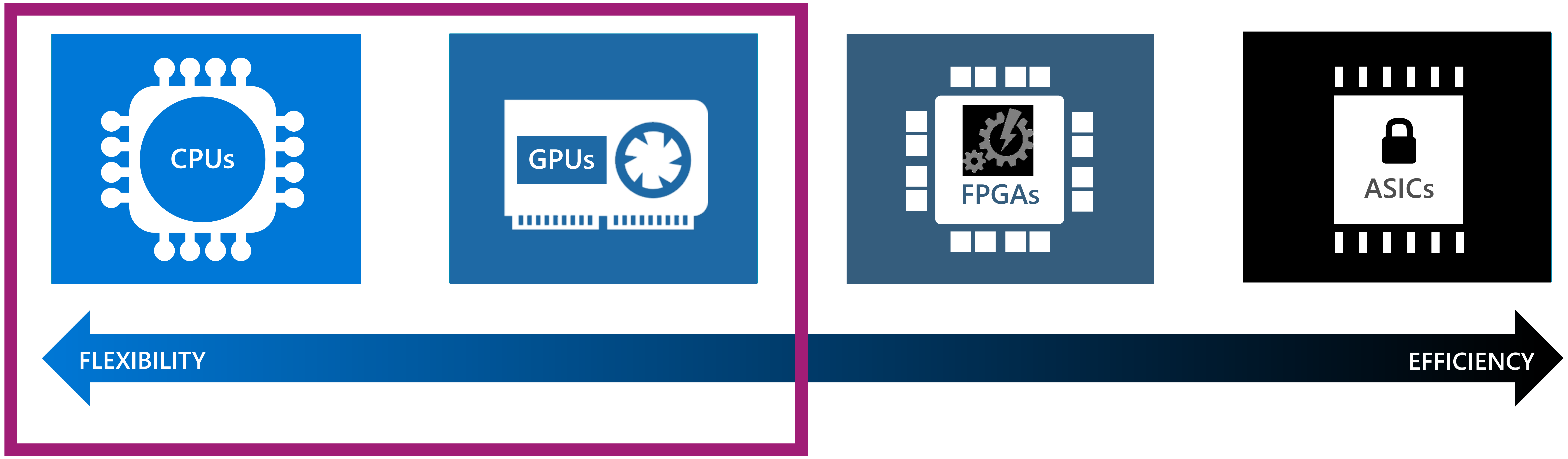
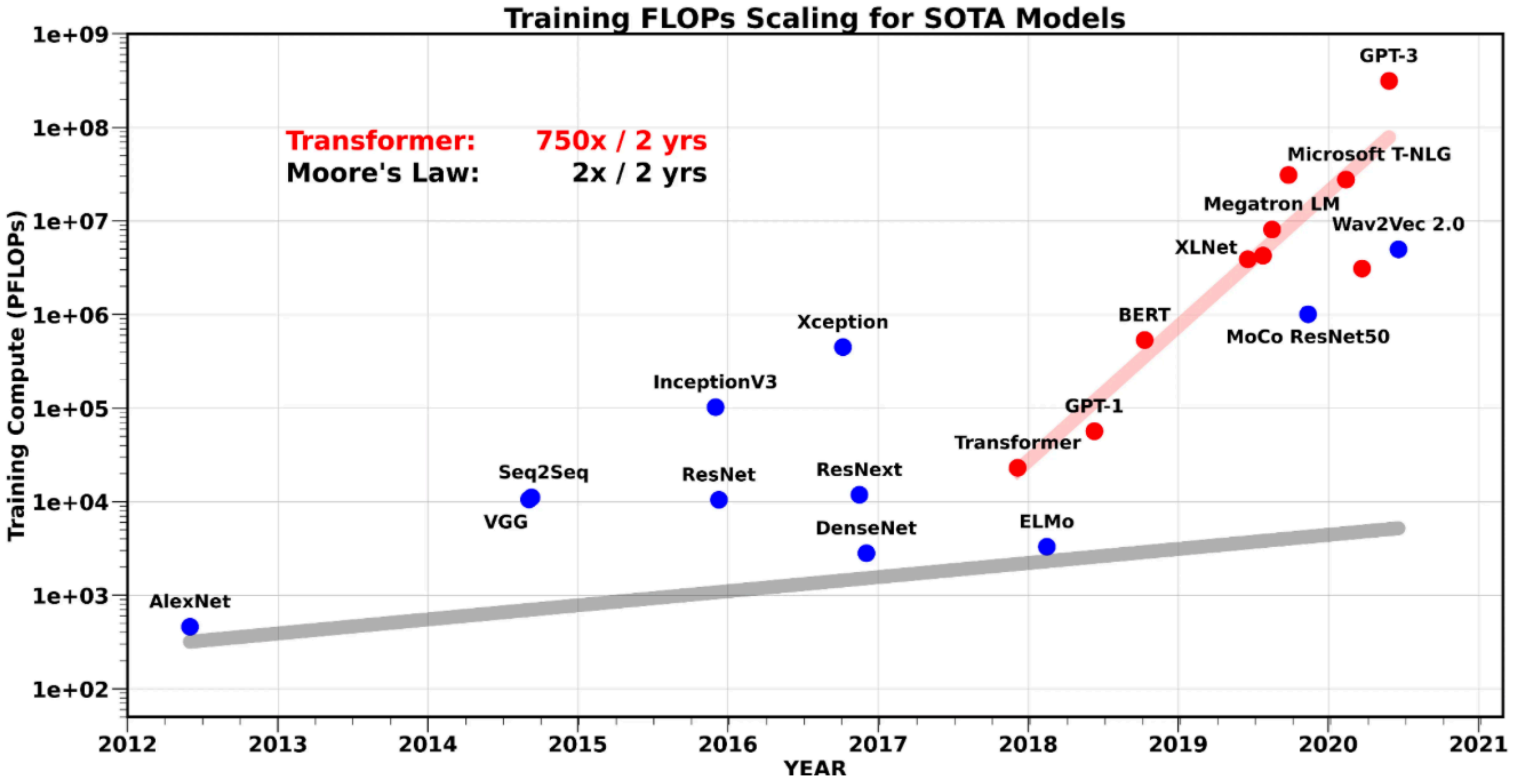


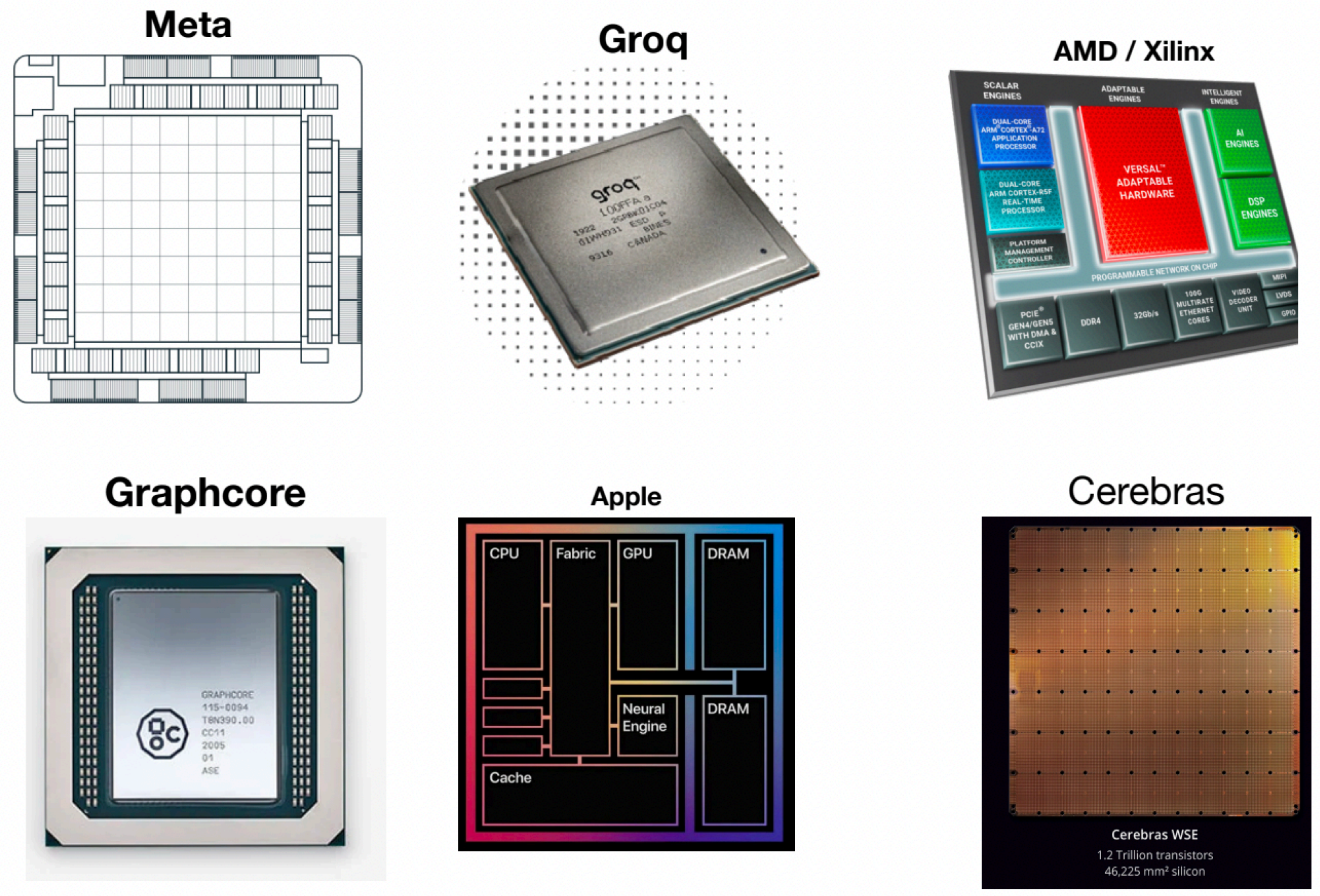
Image: [Microsoft]

Exponential trend in computational need of AI



A. Gholami

AI Chips in 2023



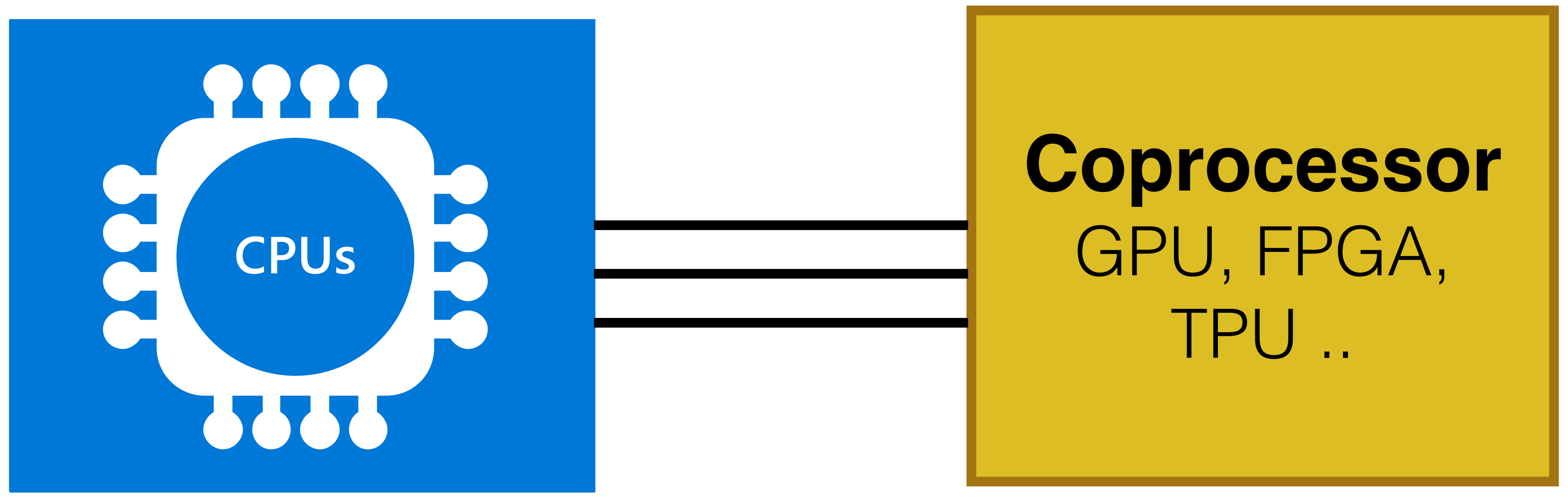
Who to include these different processors into our computing system?

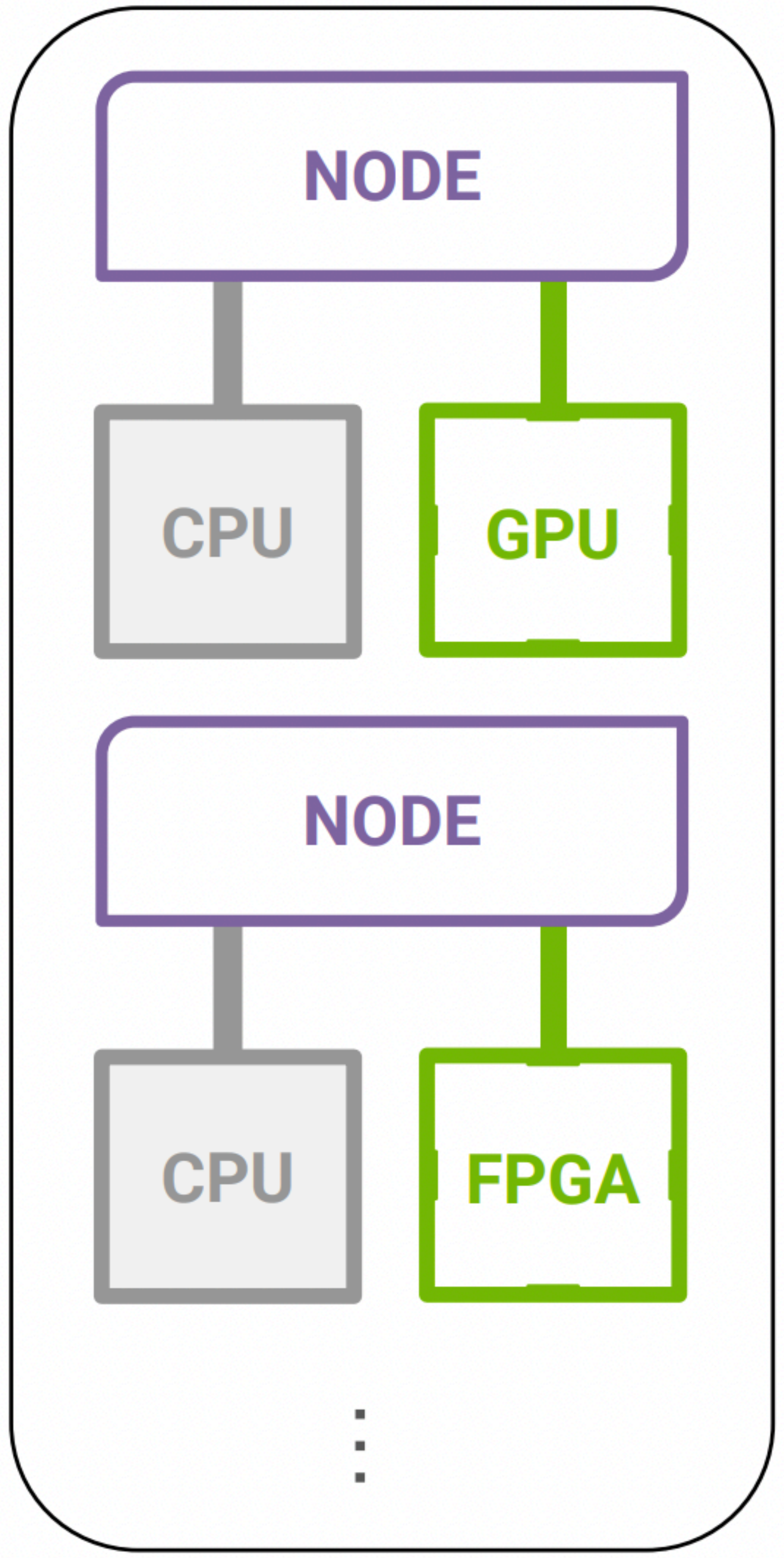
Heterogeneous computing platform

Coprocessors: specialized processors like GPU, FPGA, TPU, GraphCore, other AI chips, etc

Increased usage of specialized processors in the future

Direct Connection: Different heterogeneous systems are directly connected to each other



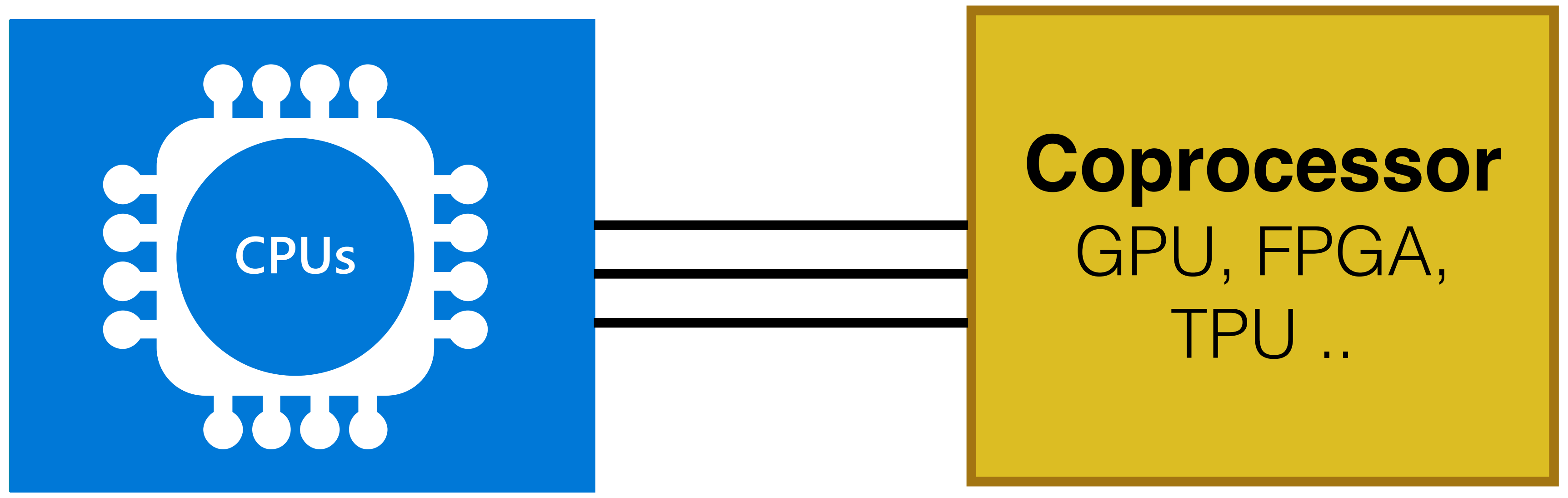


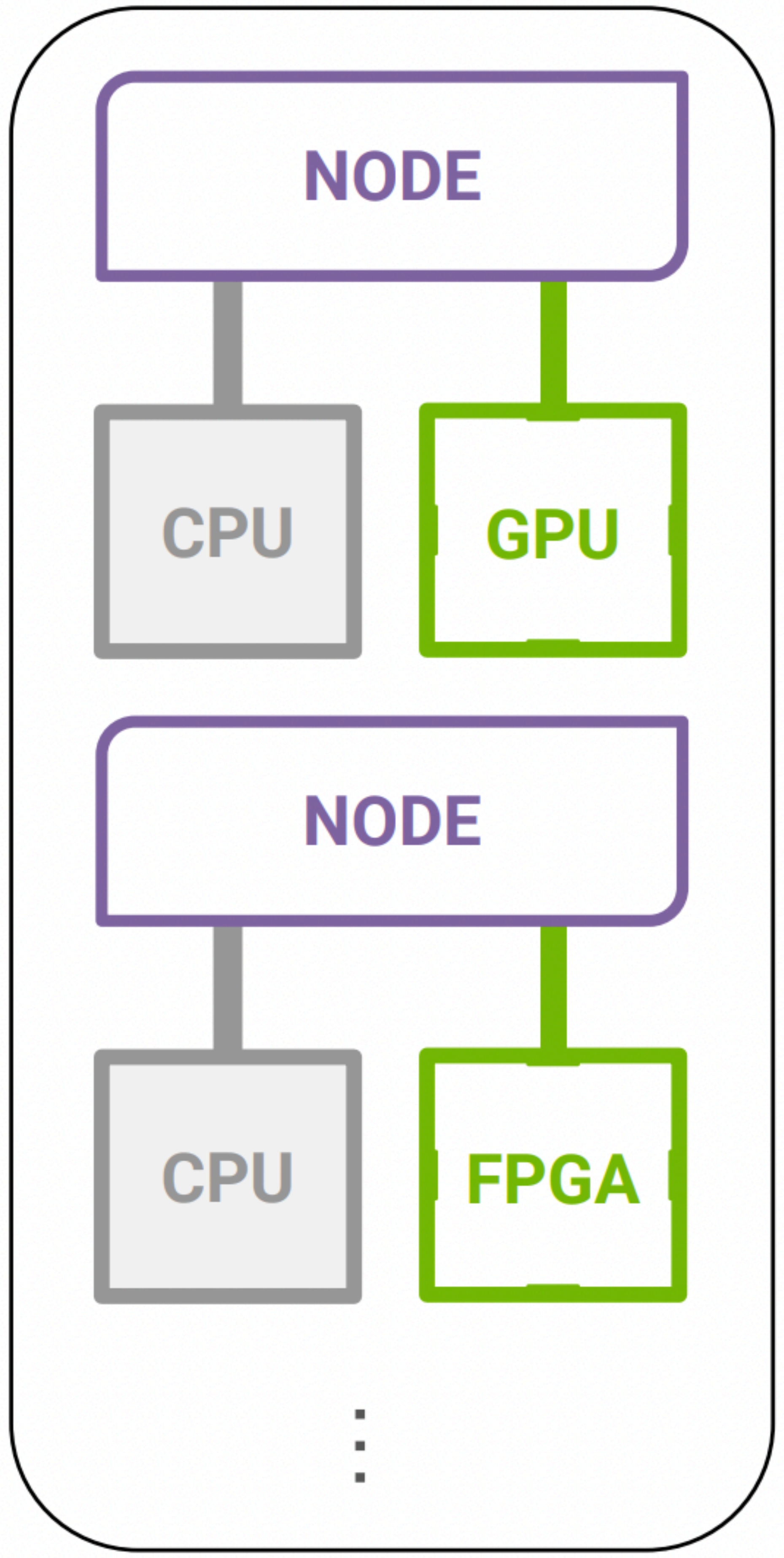
Heterogeneous computing platform

Coprocessors: specialized processors like GPU, FPGA, TPU, GraphCore, other AI chips, etc

Increased usage of specialized processors in the future

Direct Connection: Different heterogeneous systems are directly connected to each other



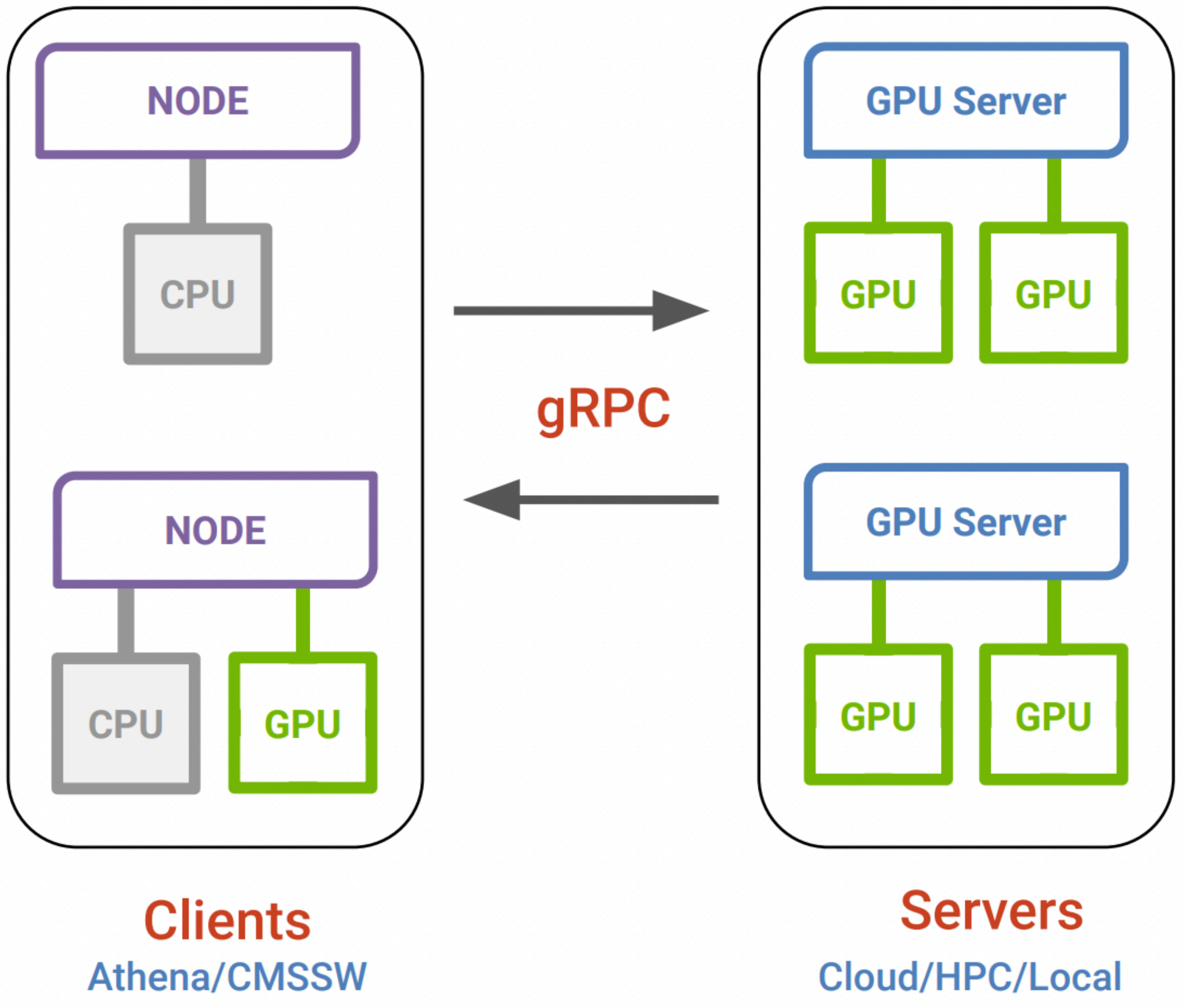


Advantage: fast and stable
Disadvantage: not flexible and not fully utilized due to inferences' complexity varies.

Inference as-a-Service

Client - Server connections are made through network

- Server running on single / multiple GPUs
- Single server can process multiple client requests



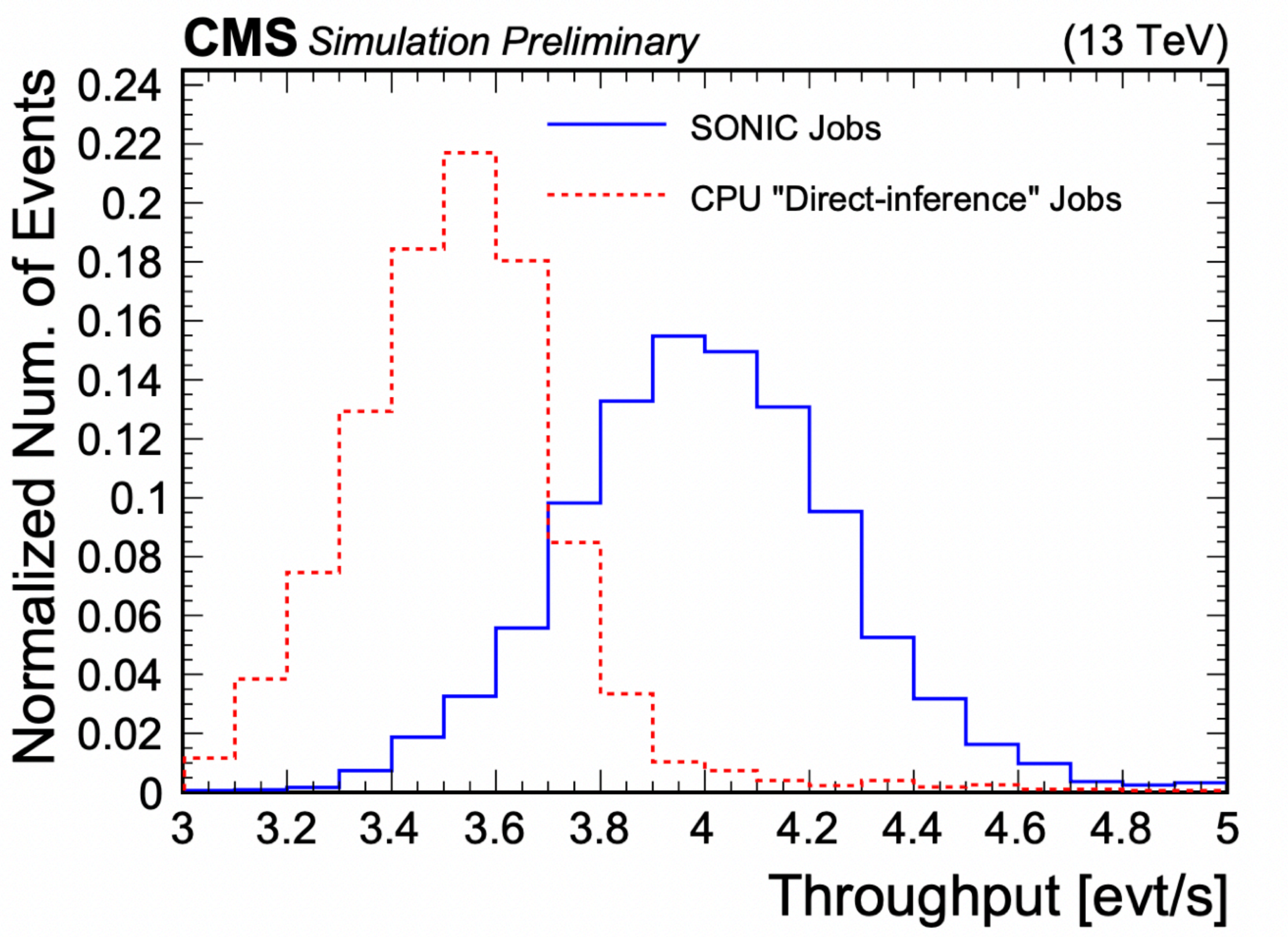
Advantage: flexible and CPU-coprocessor ratio can be optimized
Disadvantage: network topology and stability affect the inference throughput and latency

Inference as-a-Service

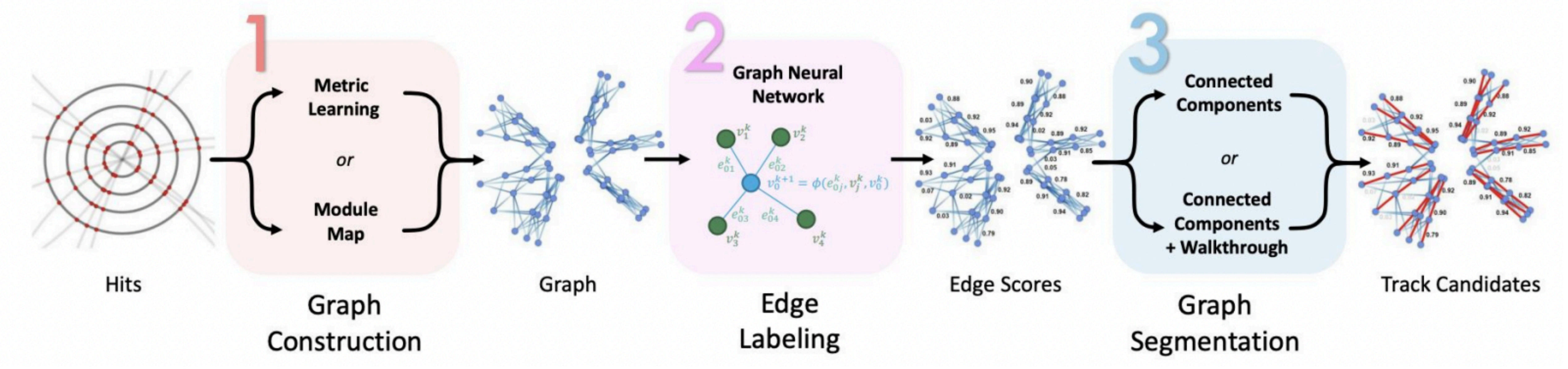
CMS-PAS-MLG-23-001

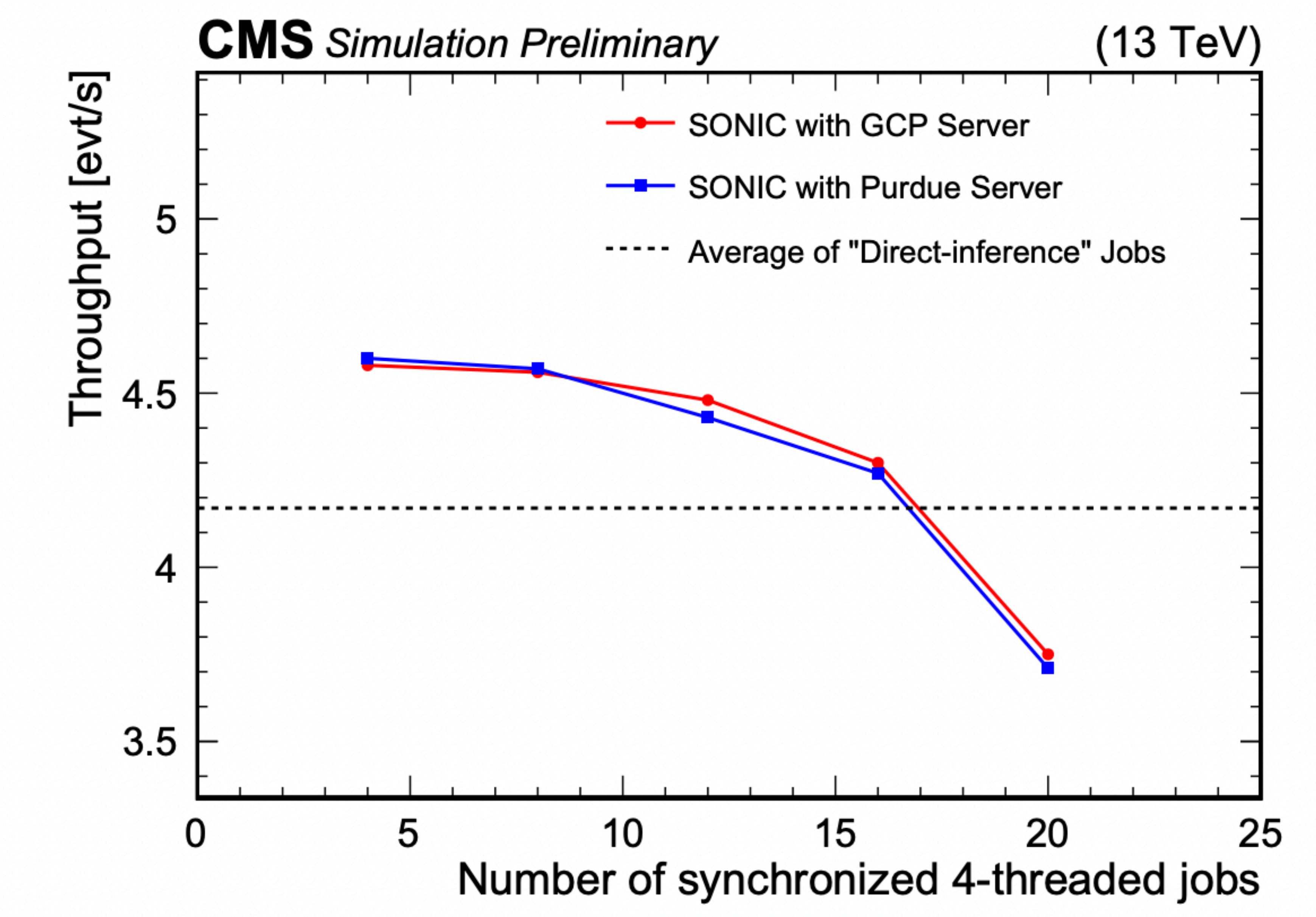
Demonstration on how it would work in the 'CMS offline computing' reality, and how much do we gain

- Roughly 13% gain in throughput
- The distance between the client and server does not impact the latency



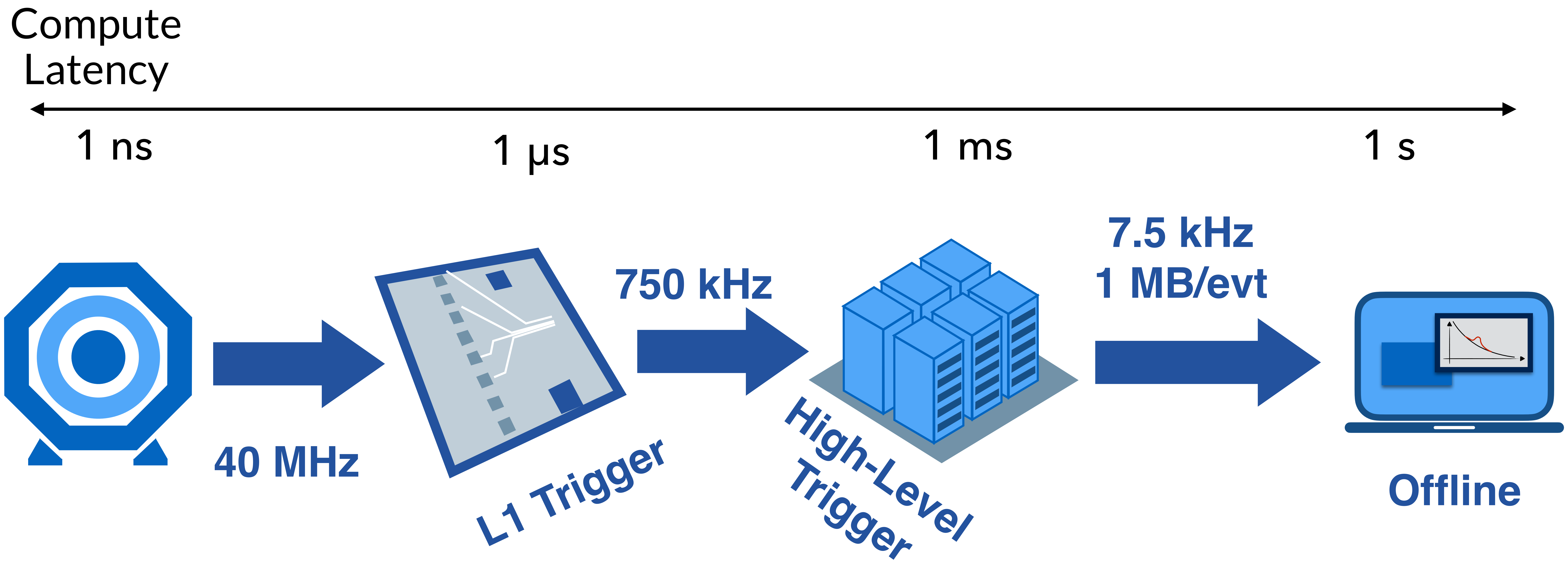
ATLAS is currently working on making GNN-based tracking as-a-service



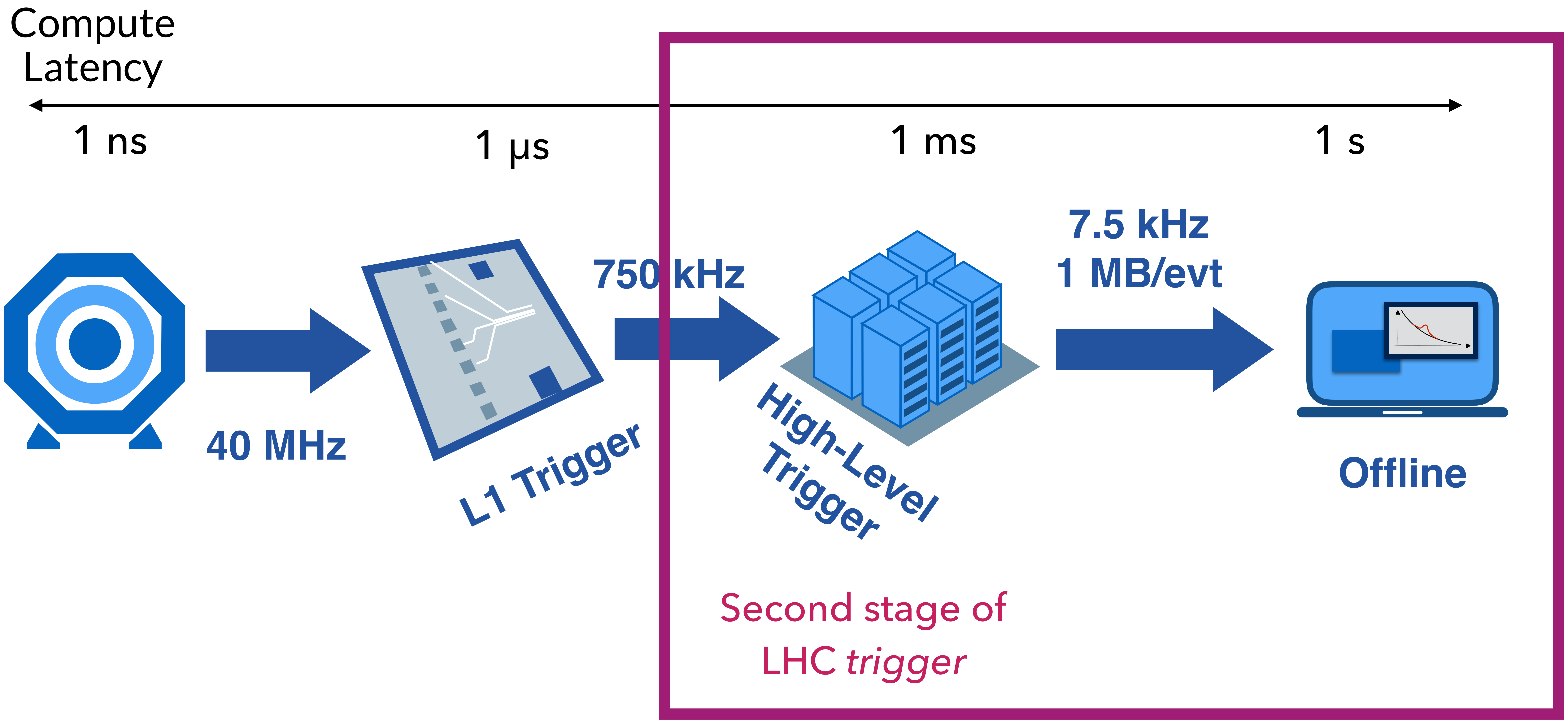


ACAT 2024 talk

HL-LHC Data Processing



HL-LHC Data Processing



ML-based Particle Flow

doi:10.5281/zenodo.8260741

- Gen. particles, reco. tracks and calorimete hits, reco. Pandora PF particles in EDM4HEP format
- CLIC detector (CLIC_o3_v14) simulation with Geant4, reco. with Marlin interfaced via Key4HEP including Pandora PF reco.
- Processes generated with Pythia8 at $\sqrt{s} = 380\,\mathrm{GeV}$
  - $e^+e^- \to t\bar{t}, q\bar{q}, ZH(\tau\tau), WW, t\bar{t} + \mathrm{PU10}$
  - Single-particle: $e^{\pm}, \mu^{\pm}, K_L^0, n, \pi^{\pm}, \gamma$ between $[1{,}100]\,\mathrm{GeV}$
- 2.5 TB, 6 million events in total

Particle Flow Reconstruction

Scalable Neural Network Models and Terascale Datasets

https://www.coe-raise.eu/od-pfr

ML-based Particle Flow

doi:10.5281/zenodo.8260741

- Gen. particles, reco. tracks and calorimete hits, reco. Pandora PF particles in EDM4HEP format
- CLIC detector (CLIC_o3_v14) simulation with Geant4, reco. with Marlin interfaced via Key4HEP including Pandora PF reco.
- Processes generated with Pythia8 at $\sqrt{s} = 380\,\mathrm{GeV}$
  - $e^+e^- \rightarrow t\bar{t}, q\bar{q}, ZH(\tau\tau), WW, t\bar{t} + \mathrm{PU10}$
  - Single-particle: $e^\pm, \mu^\pm, K_L^0, n, \pi^\pm, \gamma$ between $[1{,}100]\,\mathrm{GeV}$
- 2.5 TB, 6 million events in total

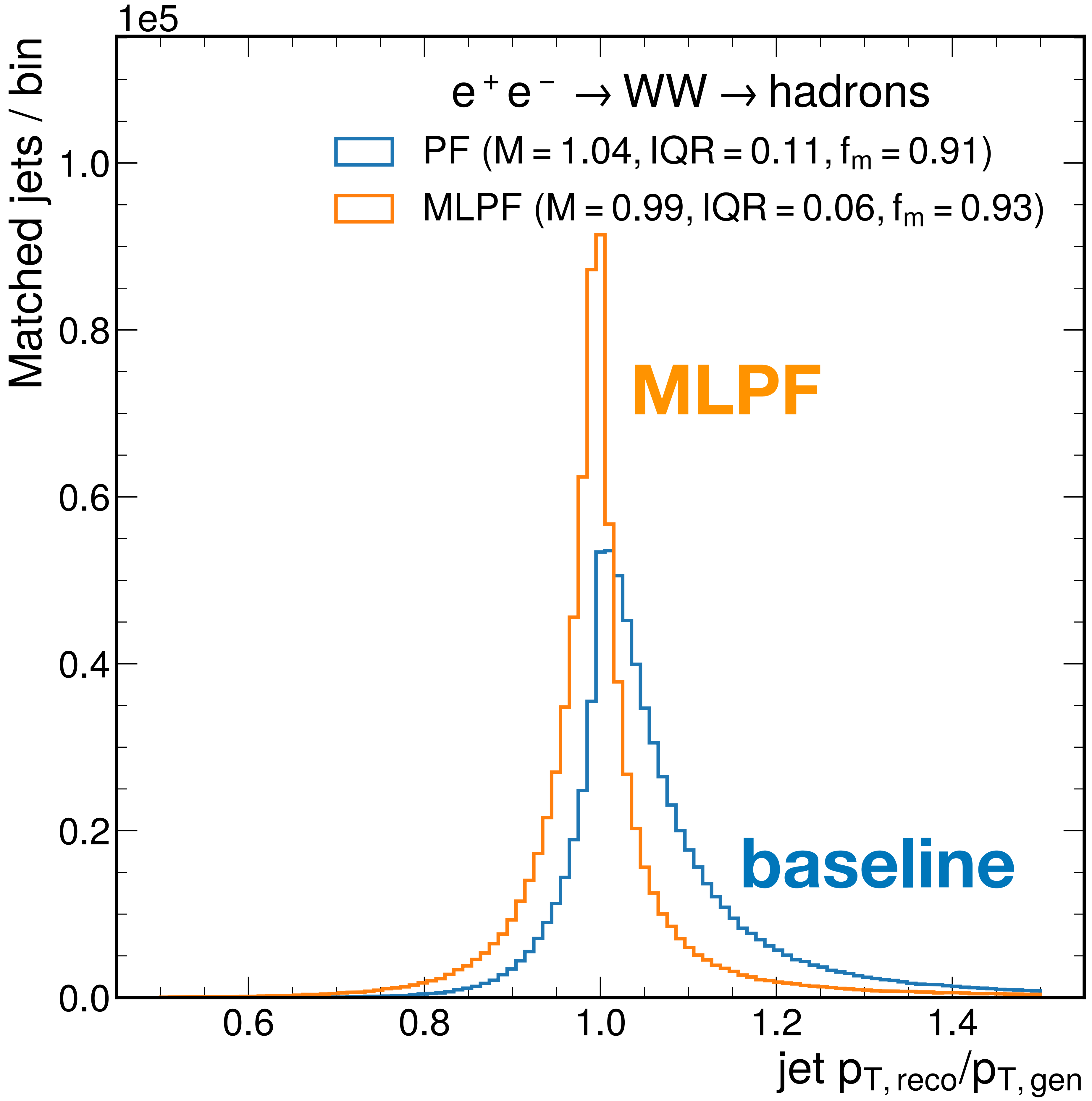
Particle Flow Reconstruction

Scalable Neural Network Models and Terascale Datasets

https://www.coe-raise.eu/od-pfr

MLPF Performance

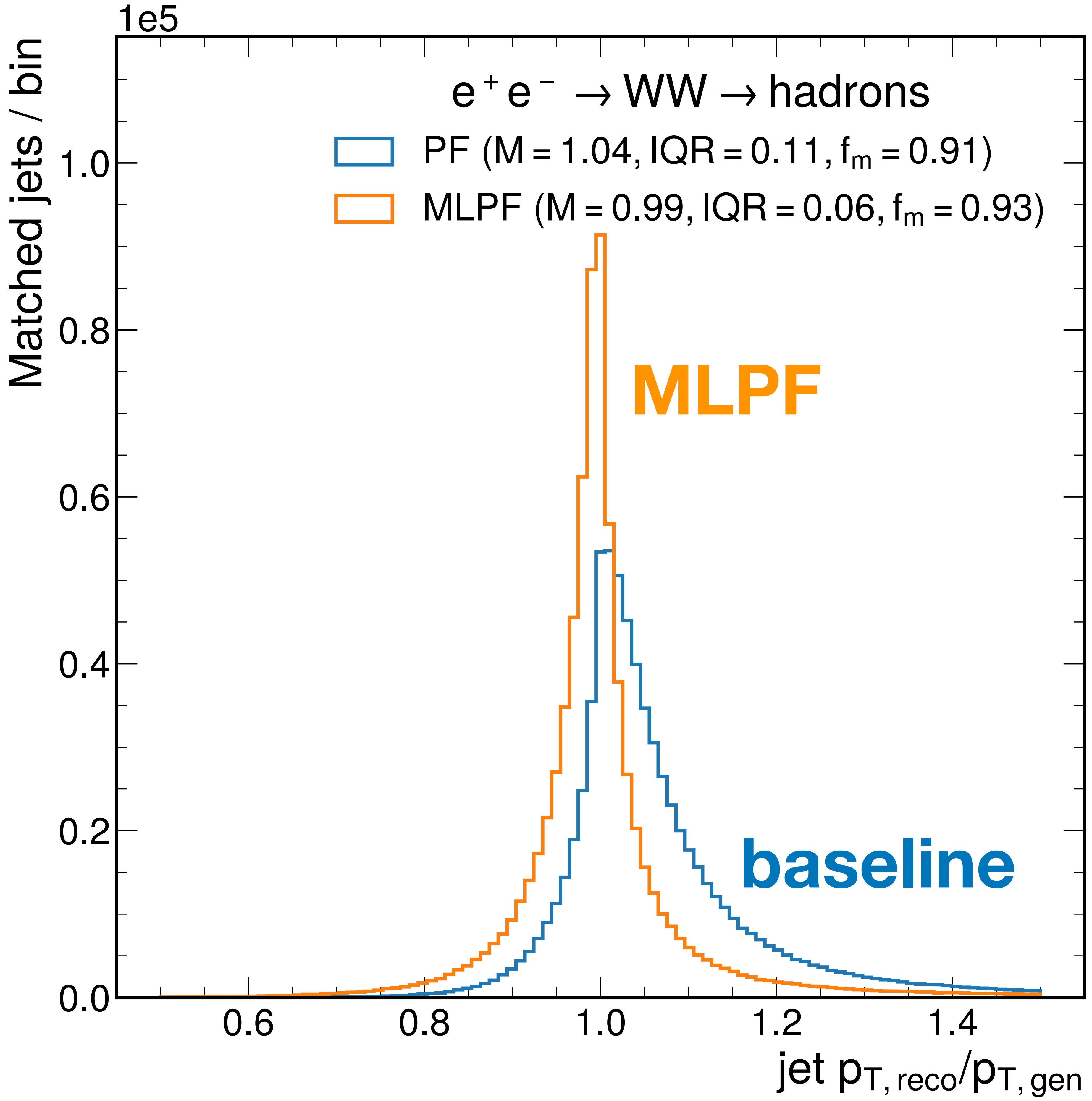
arXiv:2309.06782



MLPF Performance

arXiv:2309.06782

- Generalizes to samples (e.g., $e^+e^- \rightarrow WW \rightarrow$ hadrons) never used in training

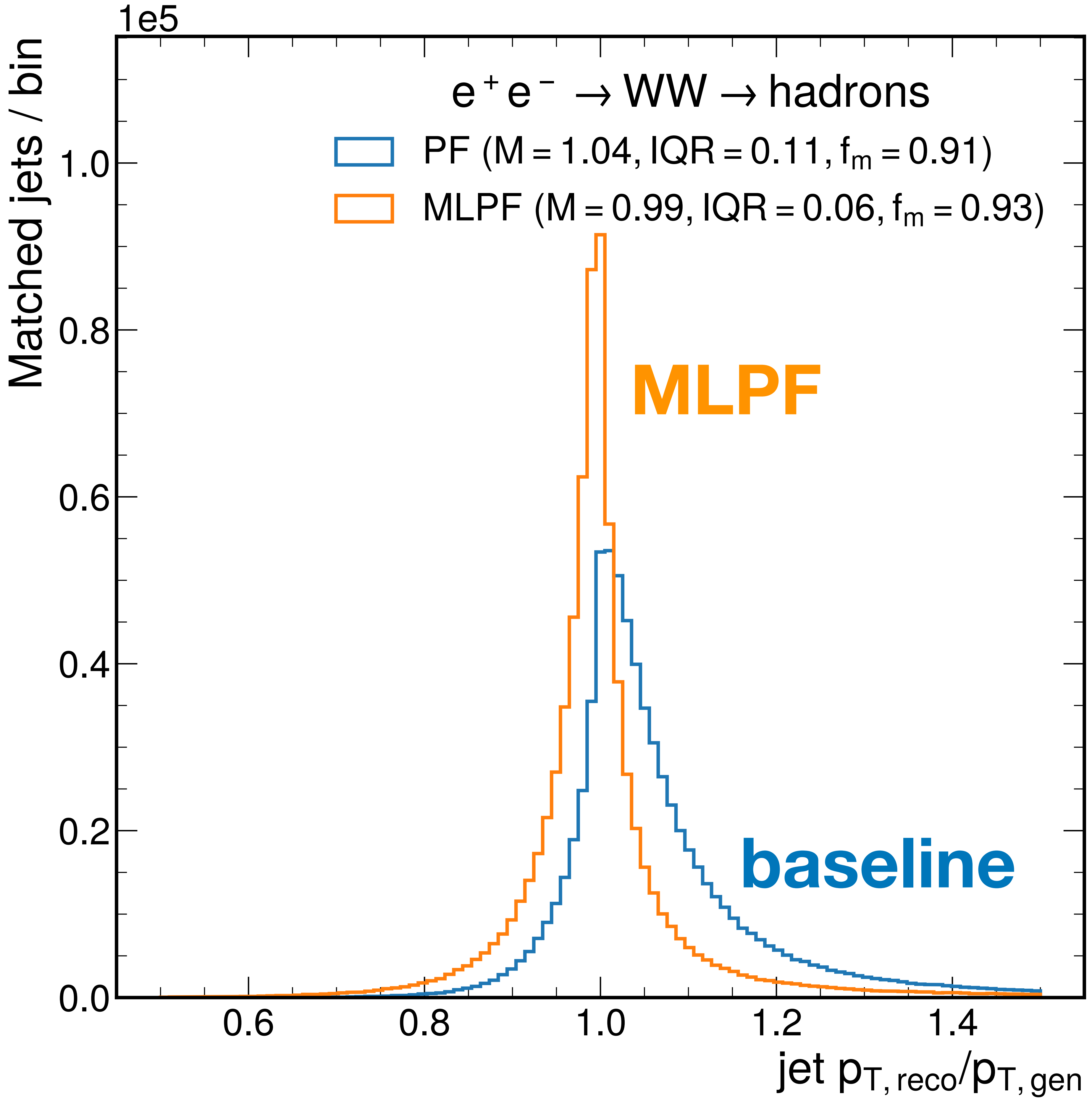


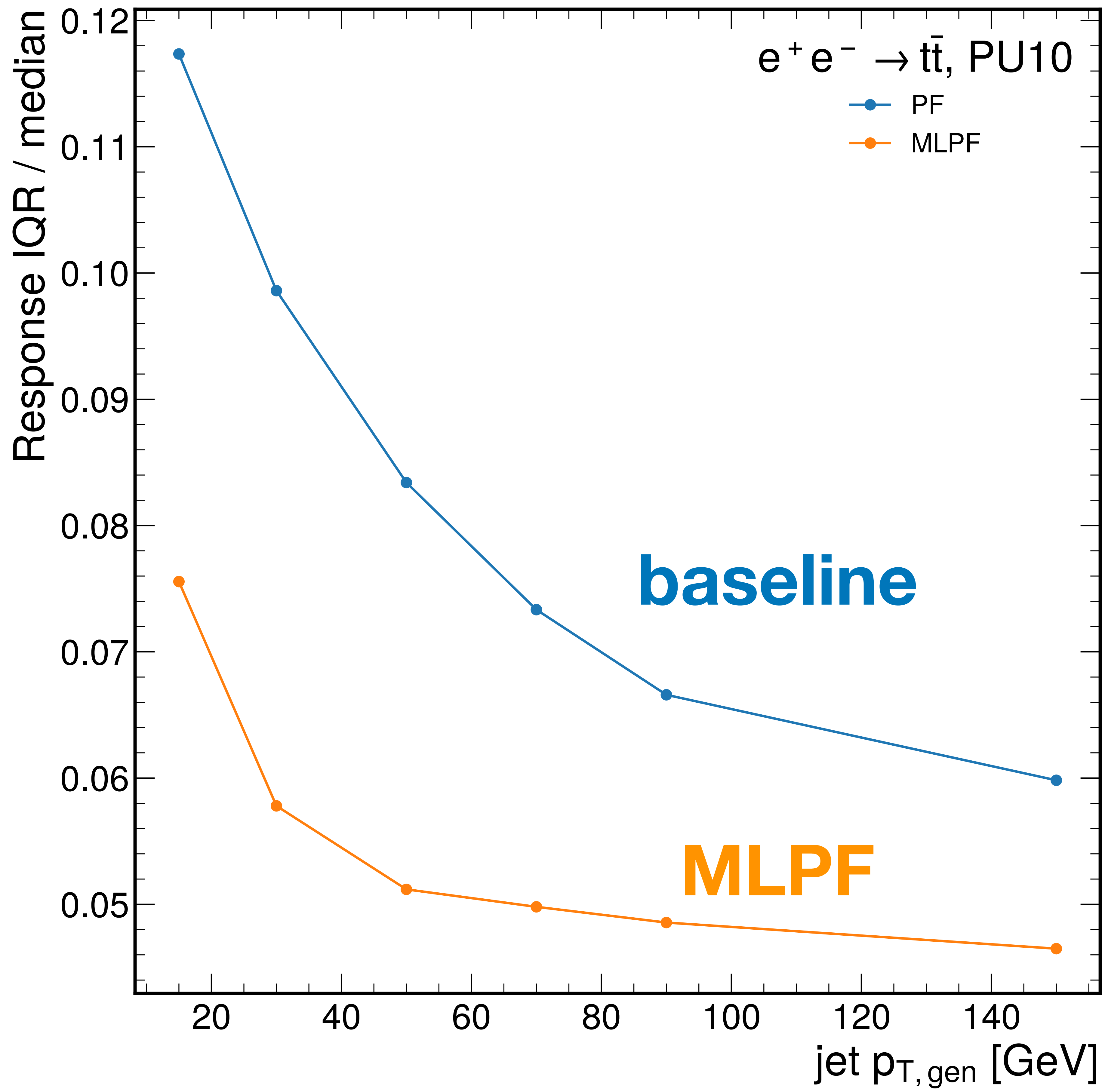
MLPF Performance

arXiv:2309.06782

- Generalizes to samples (e.g., $e^+e^- \rightarrow WW \rightarrow$ hadrons) never used in training
- ~50% improvement in jet response width over the baseline*

*Defined with gen. particle status = 1





Summary and Outlook

Summary and Outlook

- ML allows us to better reconstruct our data and save potentially overlooked data

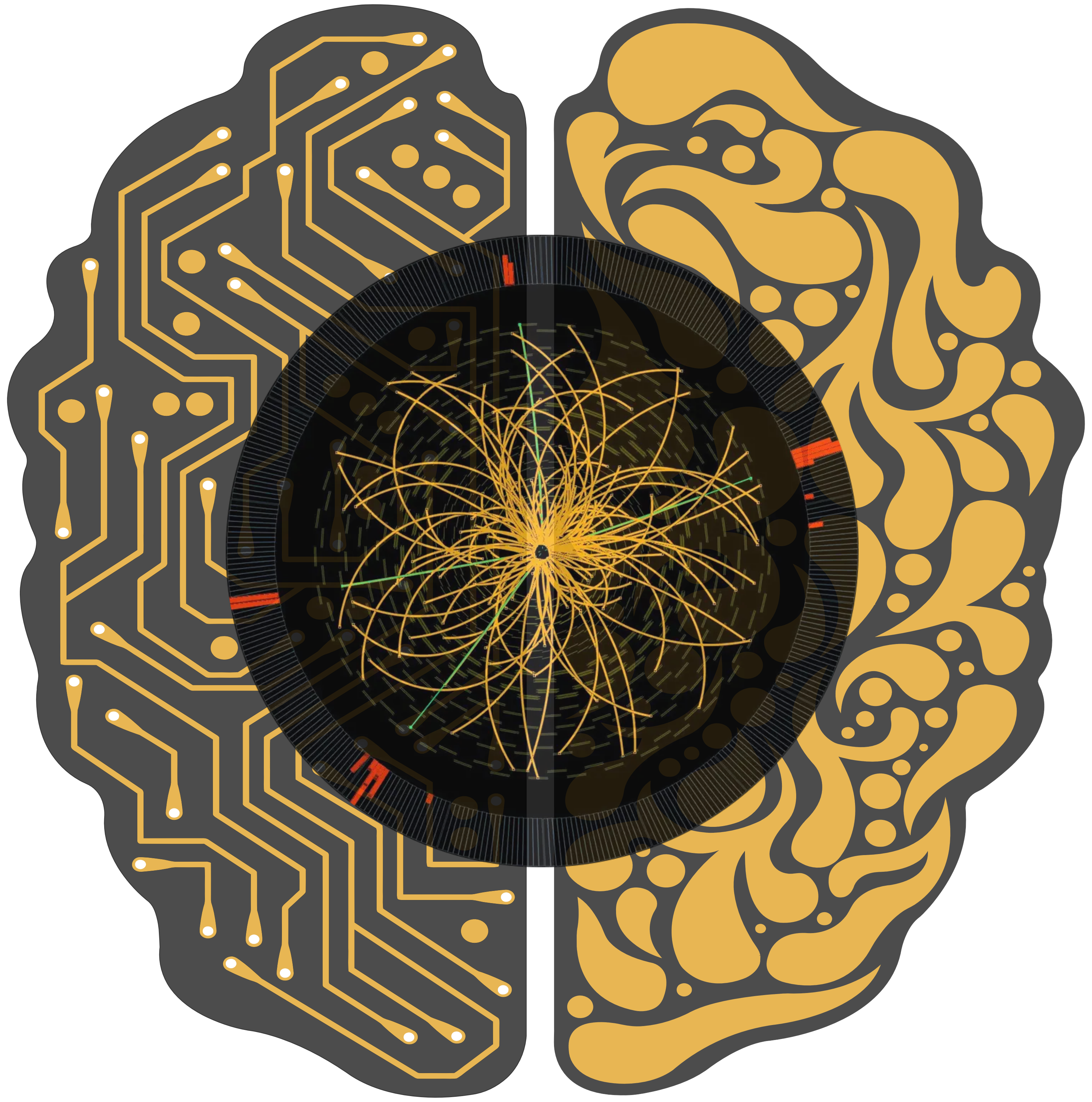
Summary and Outlook

- ML allows us to better reconstruct our data and save potentially overlooked data
- ***Codesign*** principles can enable ML on hardware with stringent constraints

Summary and Outlook

- ML allows us to better reconstruct our data and save potentially overlooked data
- ***Codesign*** principles can enable ML on hardware with stringent constraints

Accelerated AI Algorithms for Data-Driven Discovery

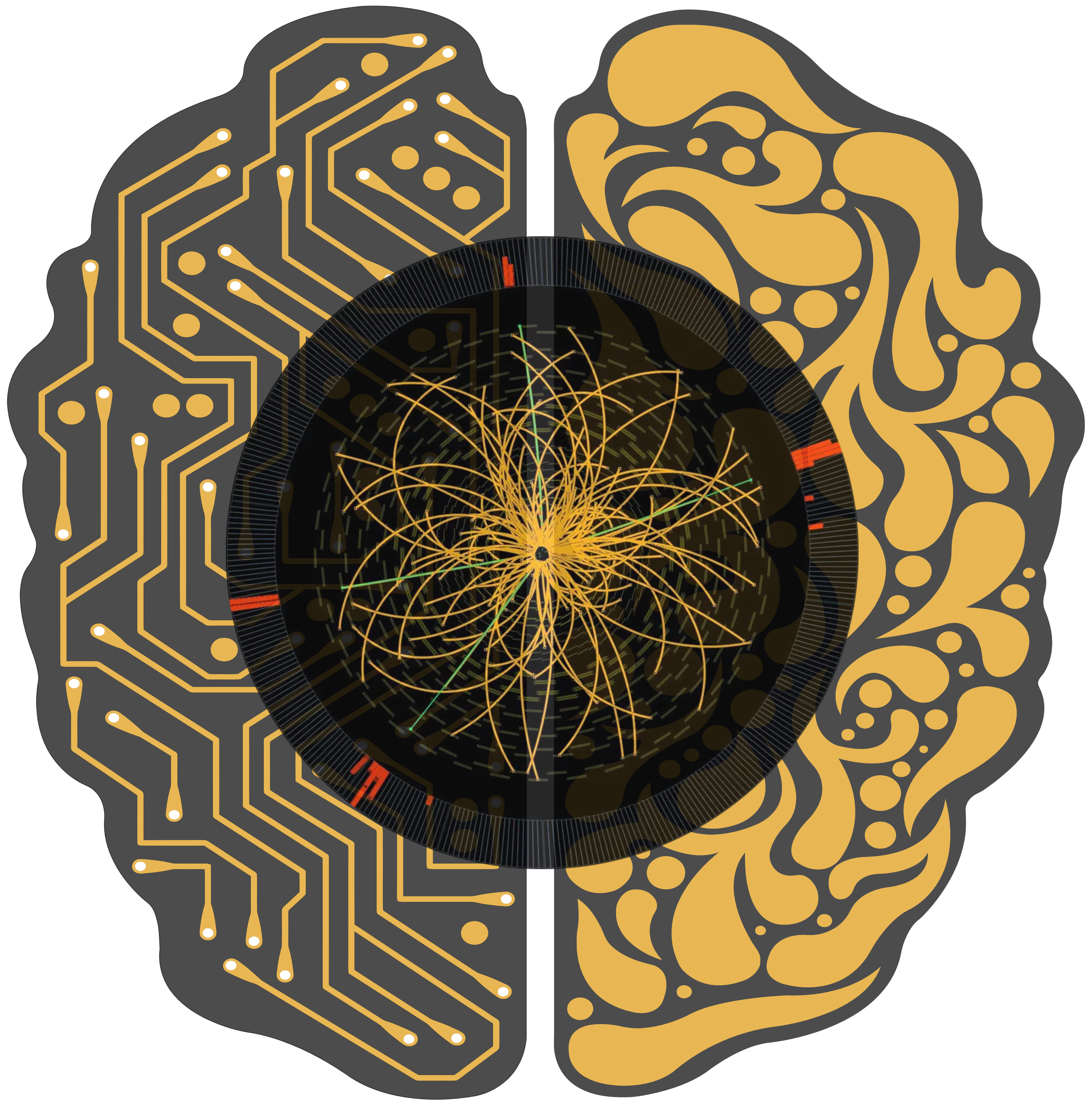


Summary and Outlook

- ML allows us to better reconstruct our data and save potentially overlooked data
- ***Codesign*** principles can enable ML on hardware with stringent constraints

- Alternative computing solutions like ***as a service*** approach will help us adopt to the growing discovery of computing hardware

Accelerated AI Algorithms for Data-Driven Discovery



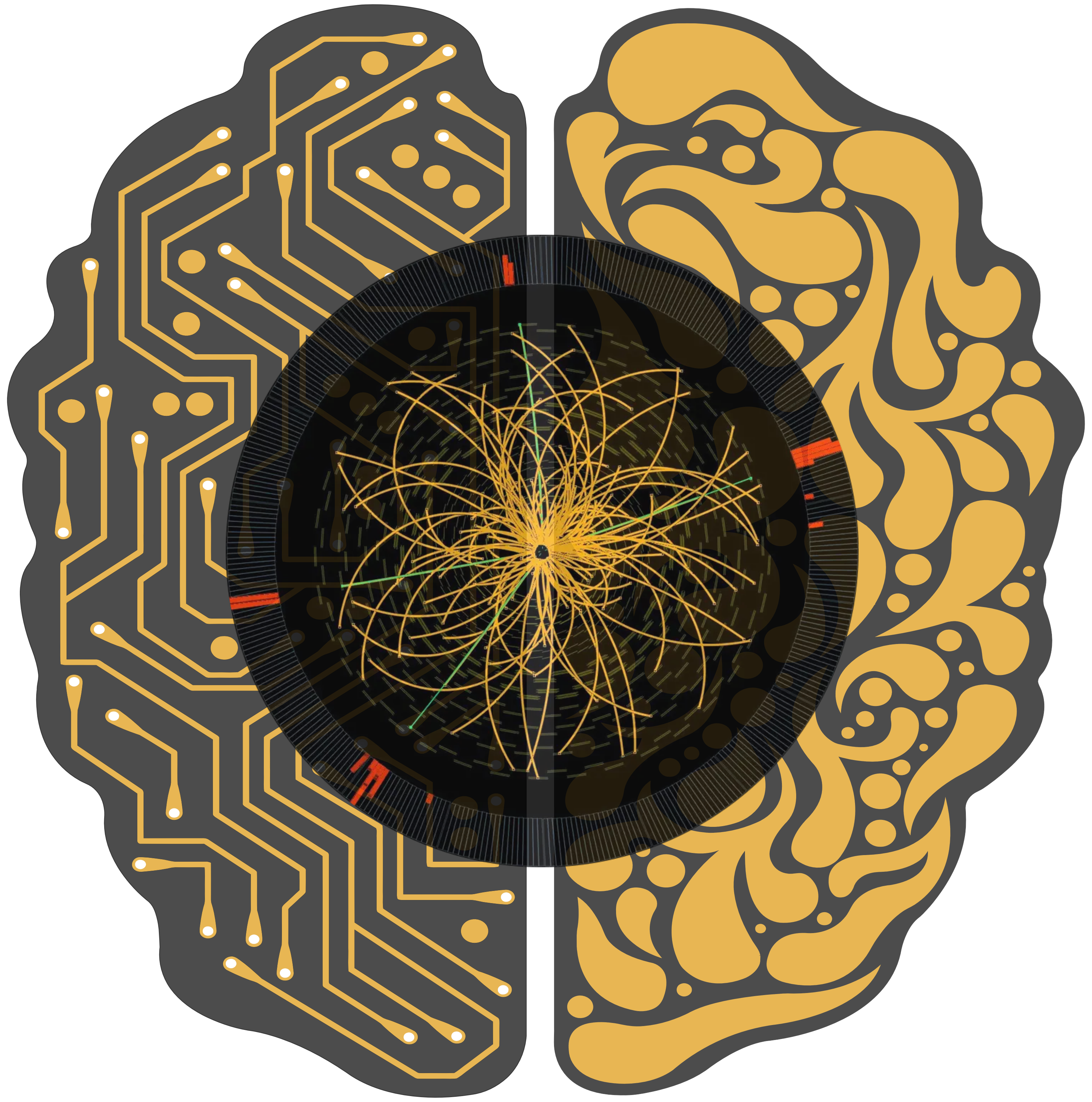
Summary and Outlook

- Community (fastmachinelearning.org, e-group hls-fml@cern.ch) and Institute (a3d3.ai) developing open-source tools and techniques to enable this
  - hls4ml: expanding open-source toolkit for translating ML into hardware aimed at trigger applications and more…
- Applications range from momentum regression, to b-tagging, tracking, and more!
  - Enhance **future particle physics program**

Summary and Outlook

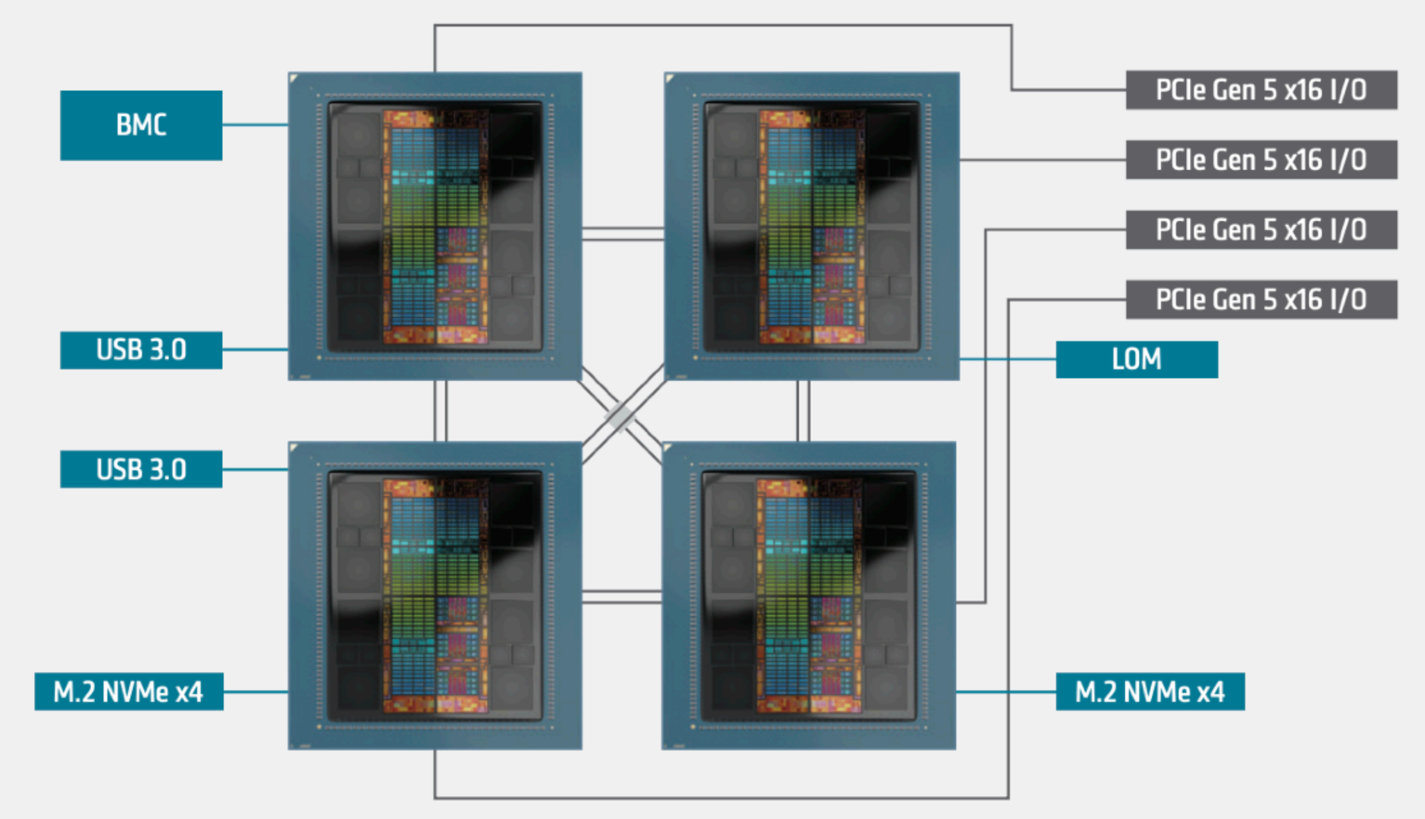
Accelerated AI Algorithms for Data-Driven Discovery

- Community (fastmachinelearning.org, e-group hls-fml@cern.ch) and Institute (a3d3.ai) developing open-source tools and techniques to enable this
  - hls4ml: expanding open-source toolkit for translating ML into hardware aimed at trigger applications and more…
- Applications range from momentum regression, to b-tagging, tracking, and more!
  - Enhance **future particle physics program**



Towards Future Collider

AMD MI300A APU



Example server architecture with four interconnected APUs

As the computing developments are very dynamic it is very difficult to guess the future solutions

- Larger ML models are becoming common
- Faster hardware are emerging

HL-LHC is a good checkpoint for upgrading our software / hardware infrastructure for Fast Inference (with heterogeneous computing)

- Integrate more AI/ML into wide range of activities

As a community we need to continue pushing the frontier and stay at the front of this rapid development

Thank You

BACKUP

Small NN benchmark correctly identifies particle “jets” 70-80% of the time