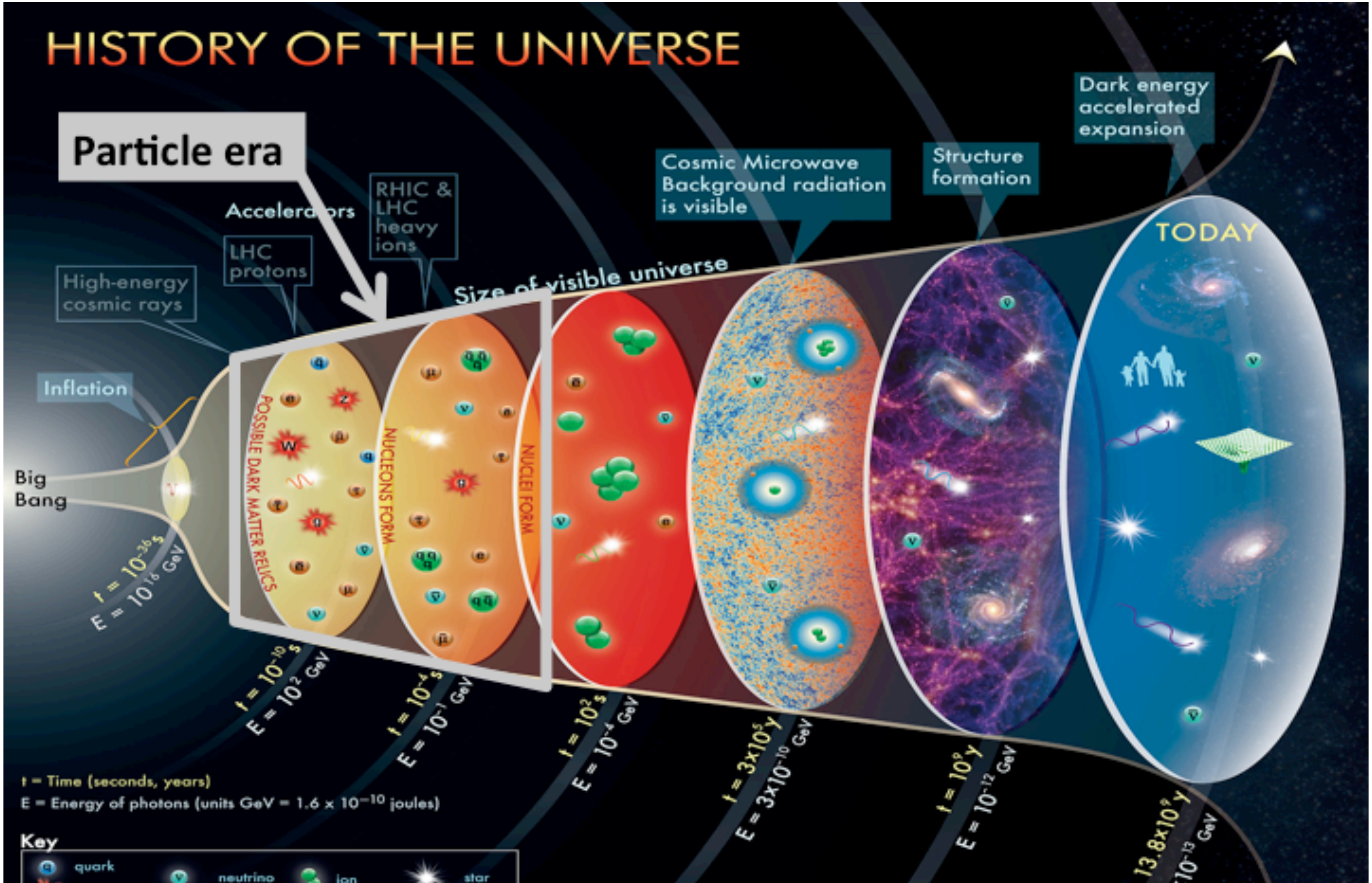
Artificial Intelligence Accelerated Discoveries At the Large Hadron Collider

Mia Liu
Purdue University

Oct 14, 2021
University of Warwick

Accelerated AI Algorithms for Data-Driven Discovery

The Standard Model of Fundamental Particles



The Standard Model of Fundamental Particles

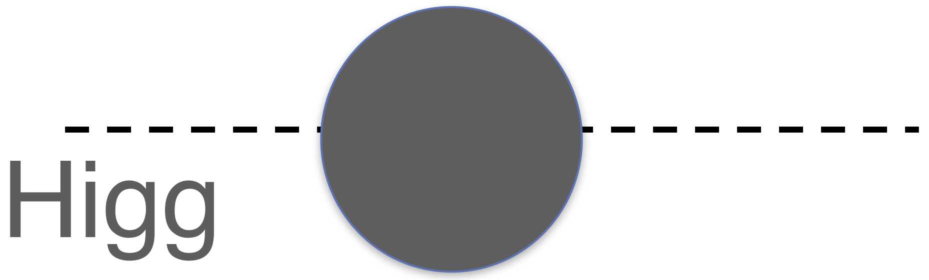
LAST MISSING PIECE
2012: HIGGS DISCOVERED AT THE LHC!

Physicists Find Elusive Particle Seen as Key to Universe

By DENNIS OVERBYE JULY 4, 2012

Remaining puzzles

SM particles



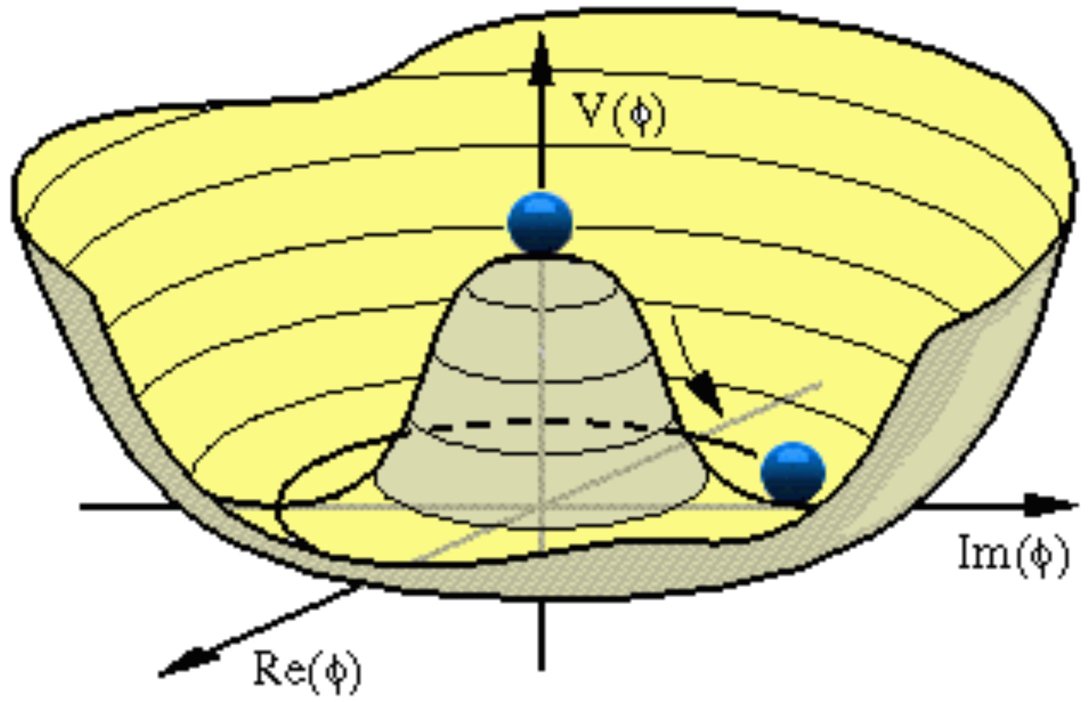
Higg

$$m^2_{obs} = m^2_{bare} - \frac{\lambda_f^2}{8\pi^2}\Lambda^2$$

125 GeV

$\varepsilon*10^{19}$ GeV?

10^{19} GeV



Fine tuning?

Dynamical origin?

…

Experimental:

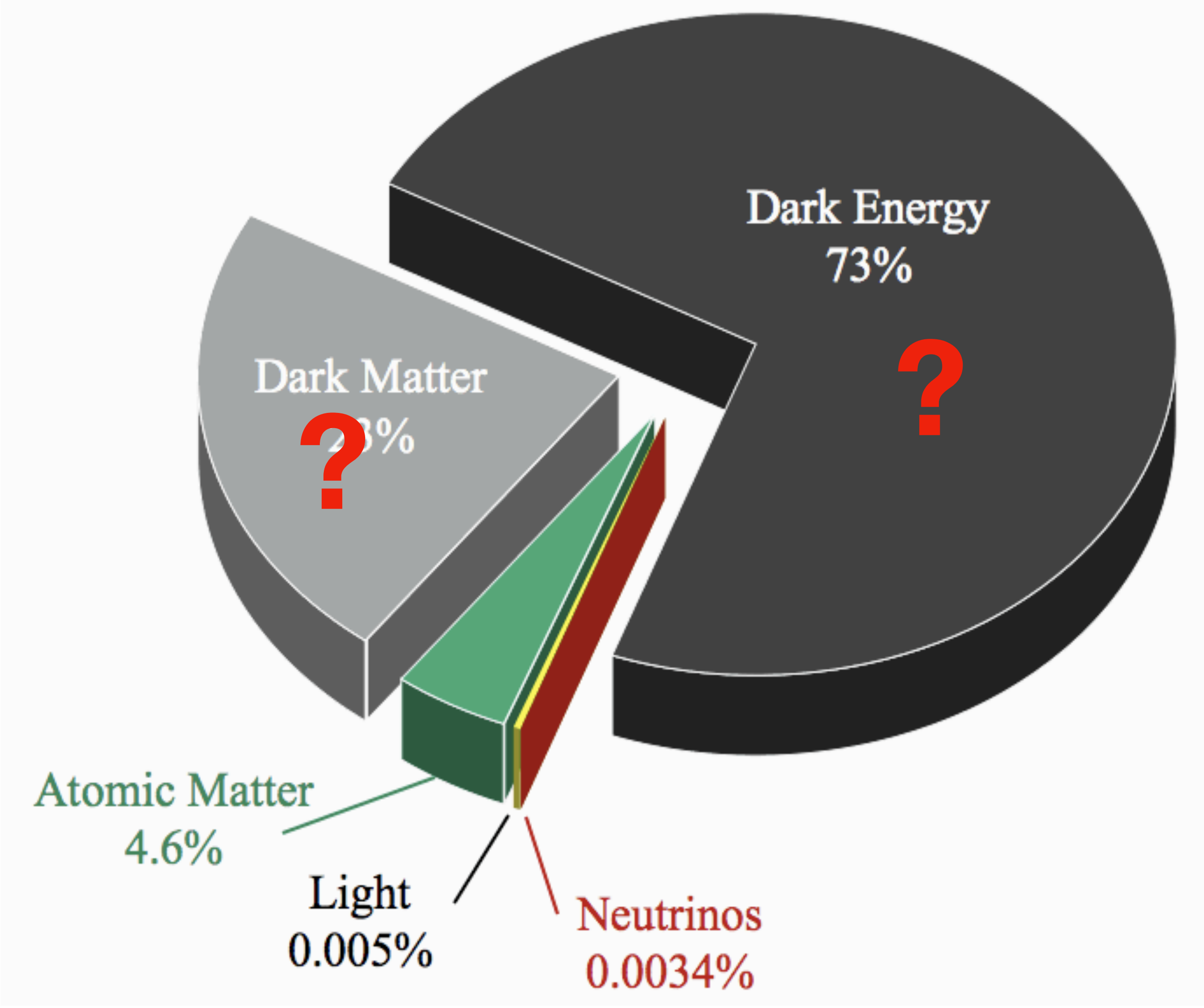
Dark matter/dark energy

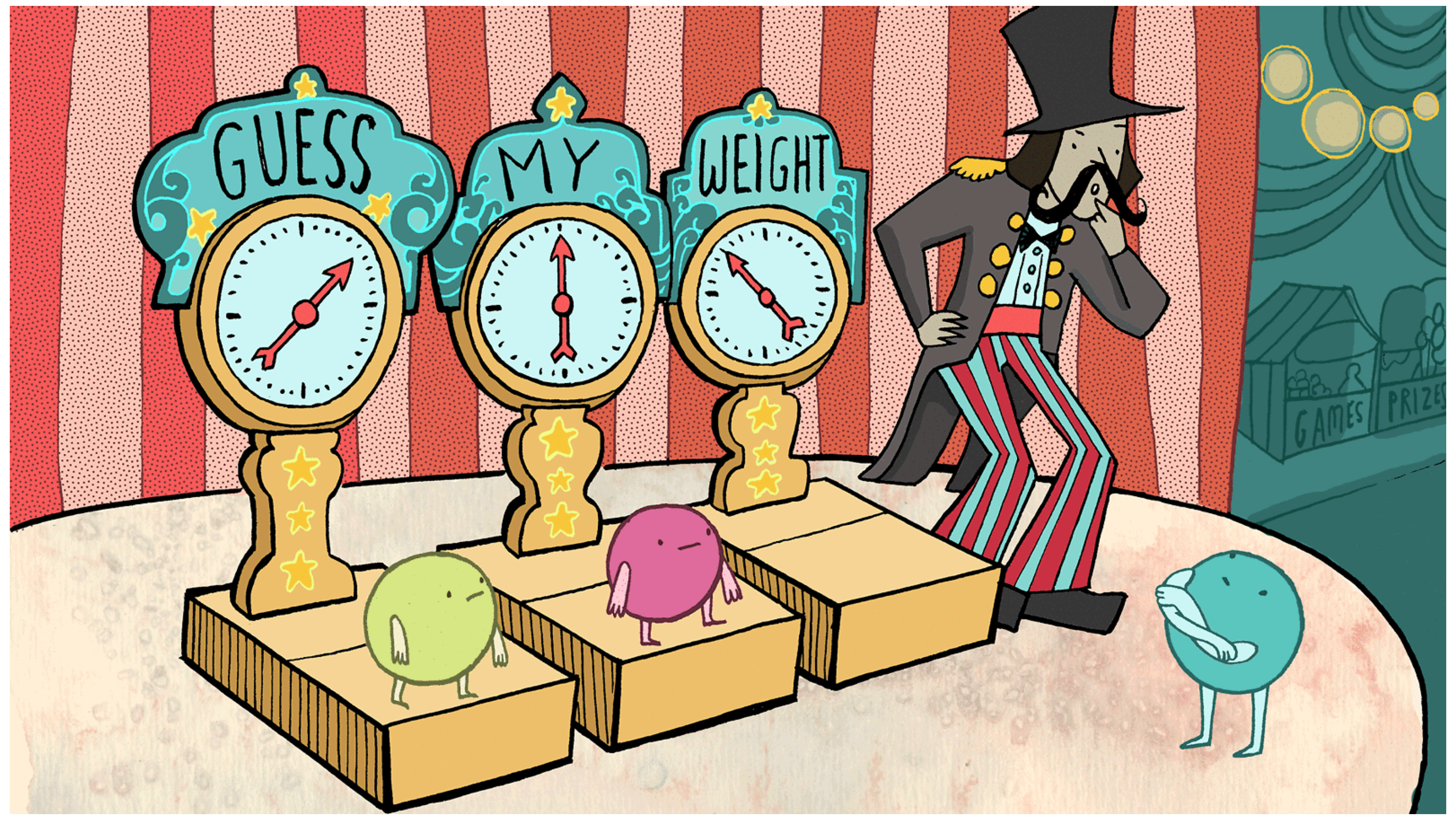
Not in SM!

Neutrinos in SM,

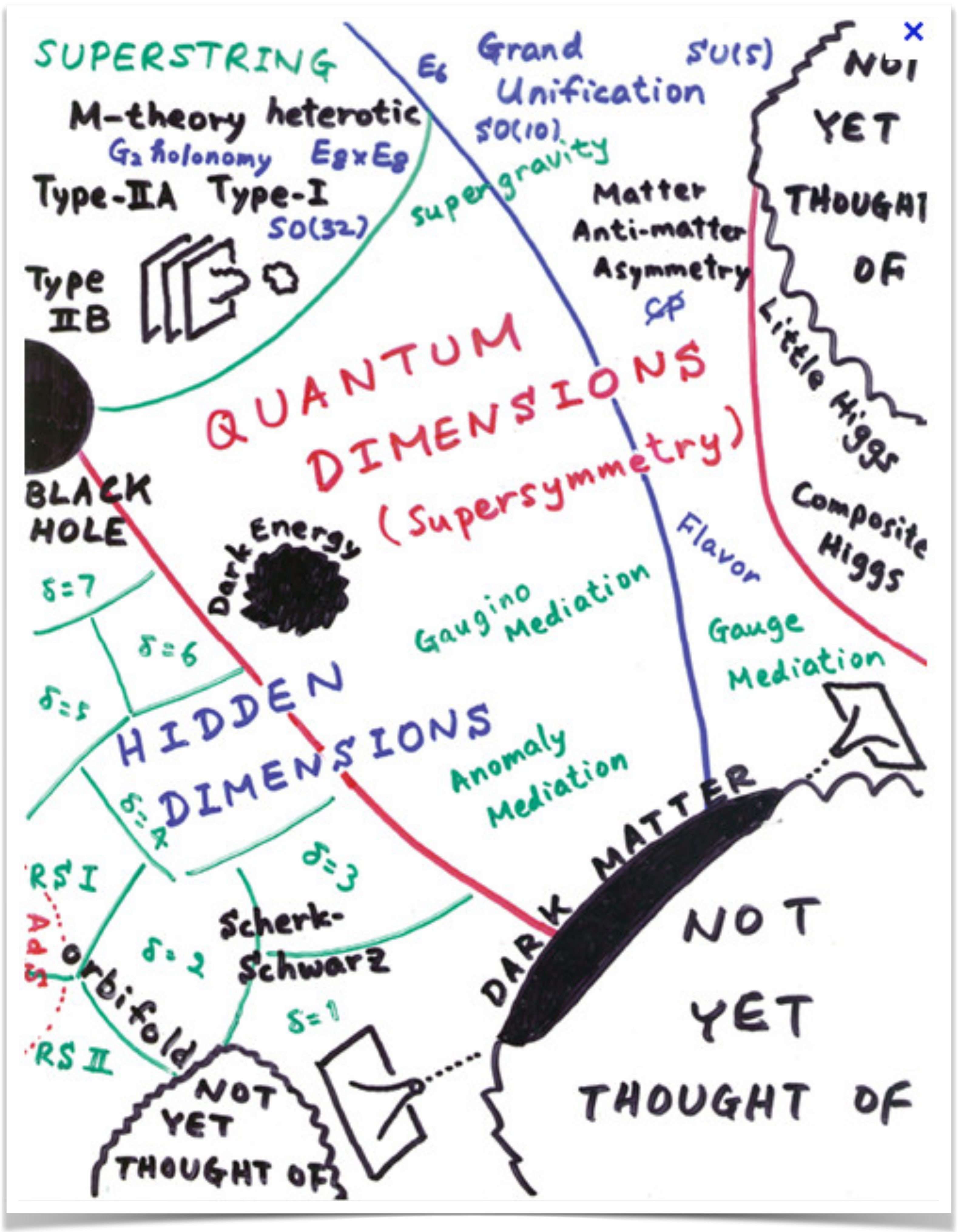
masses?…

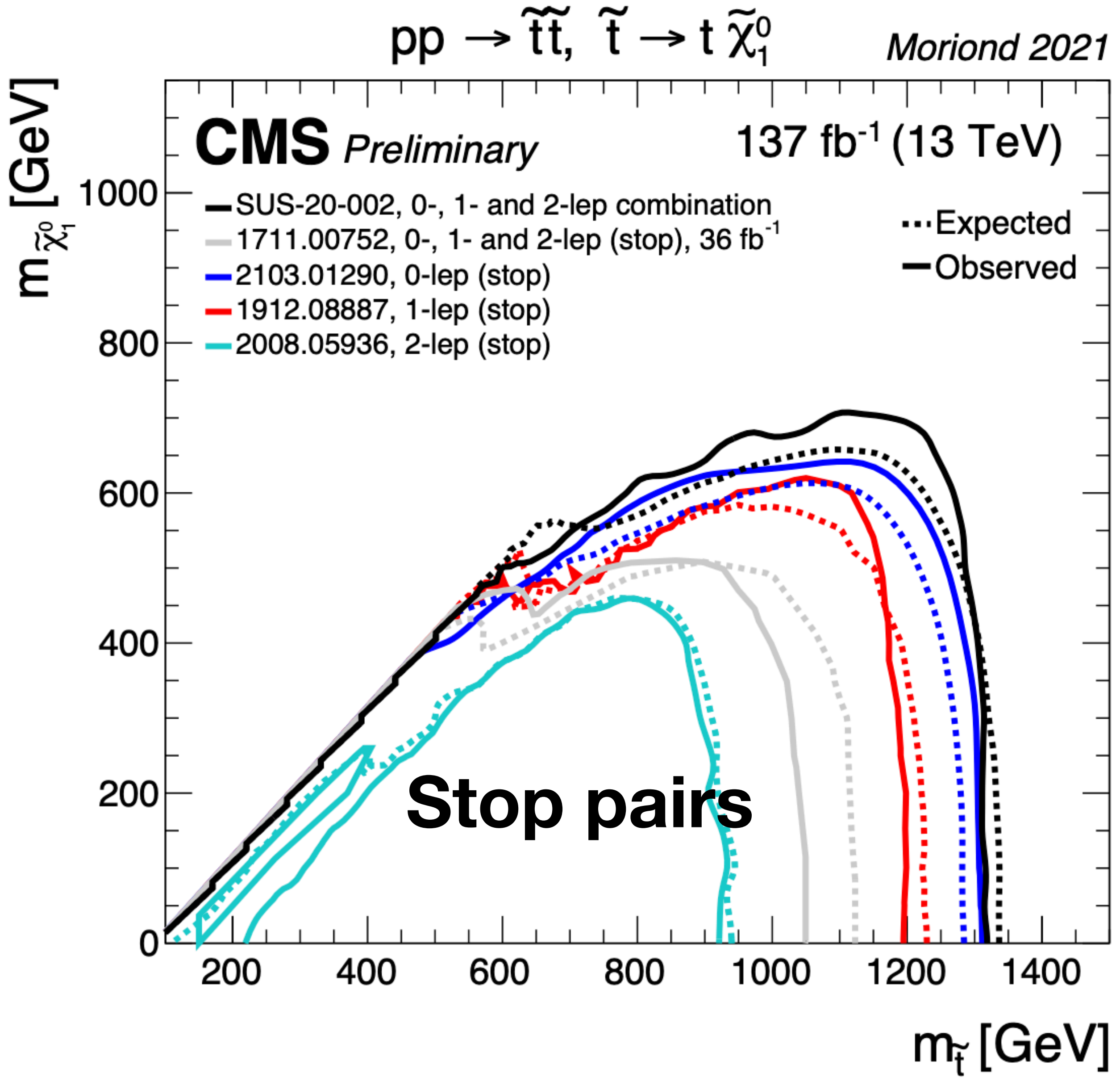
Anomalies: Muon g-2, LHCb lepton flavour universality?

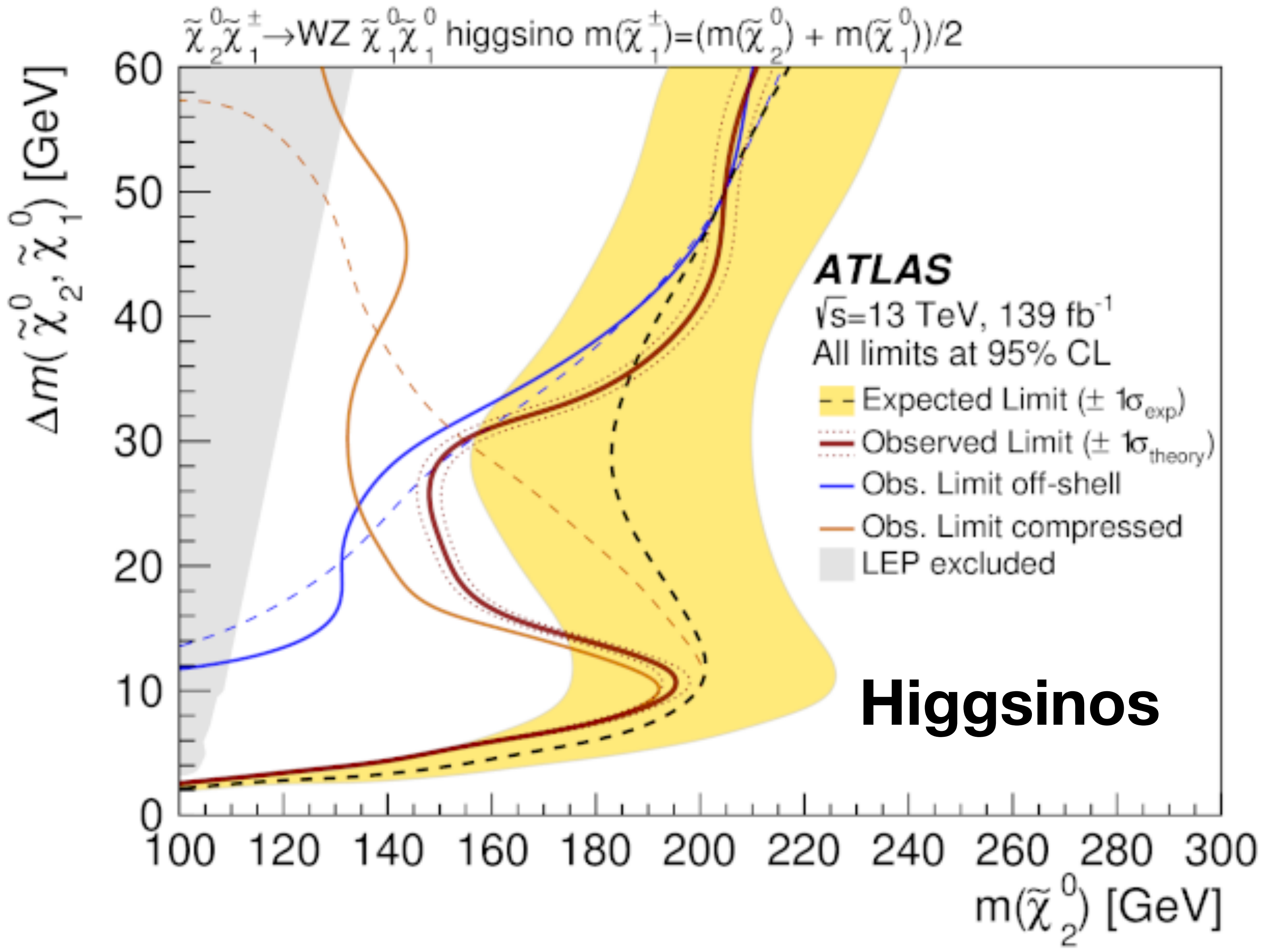


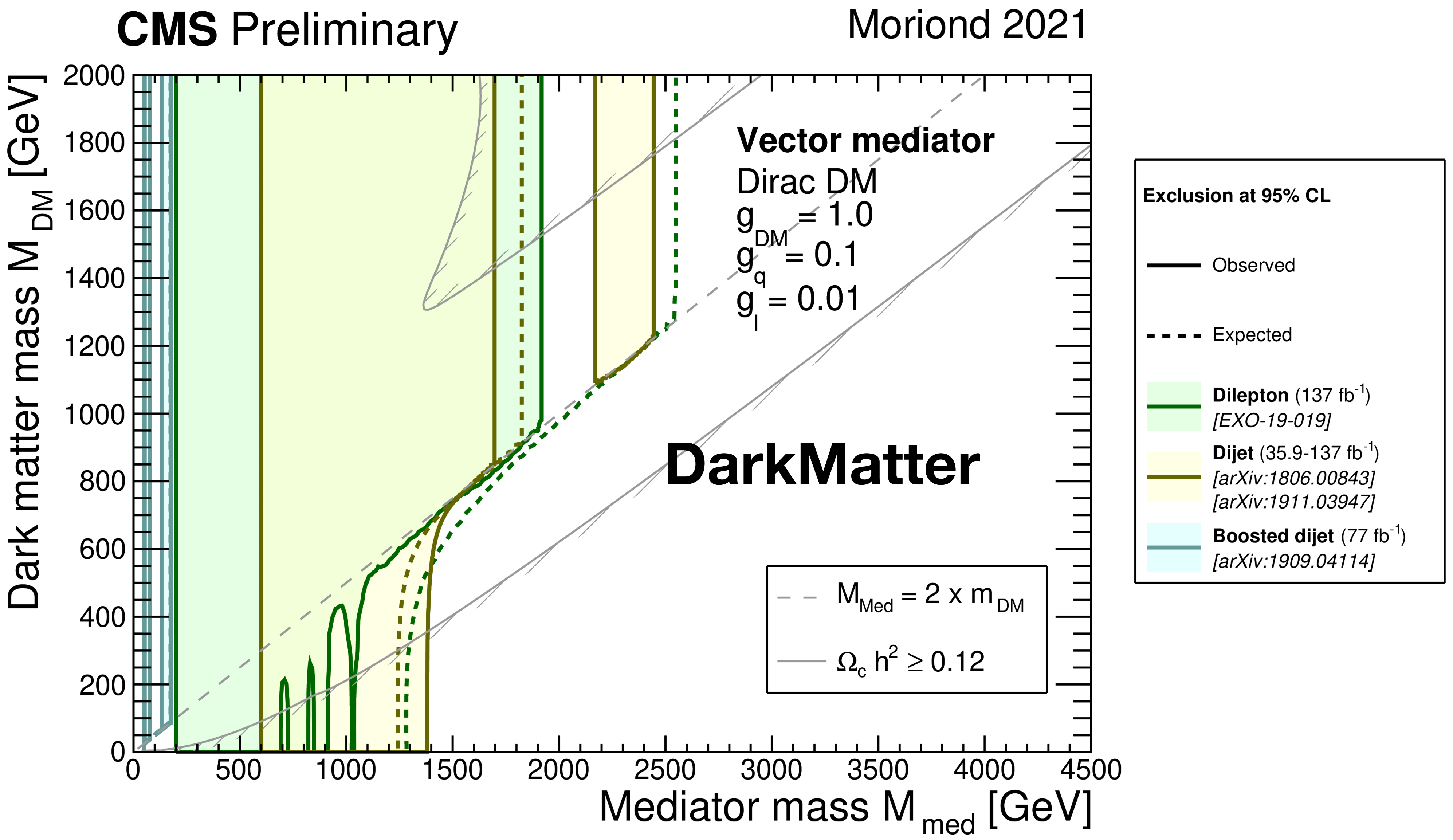


Extensive searches performed at the LHC

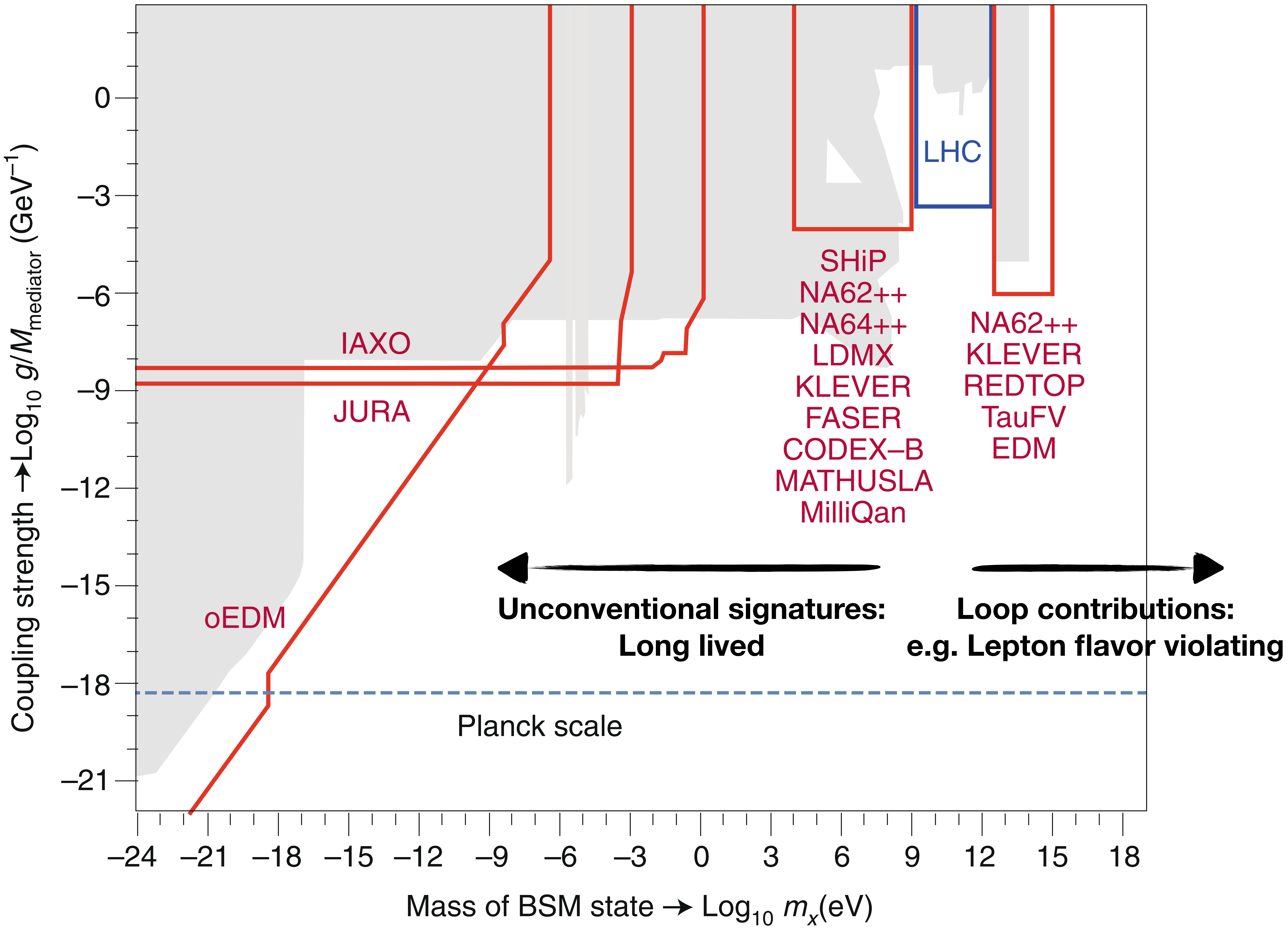








Can we do better at the LHC?



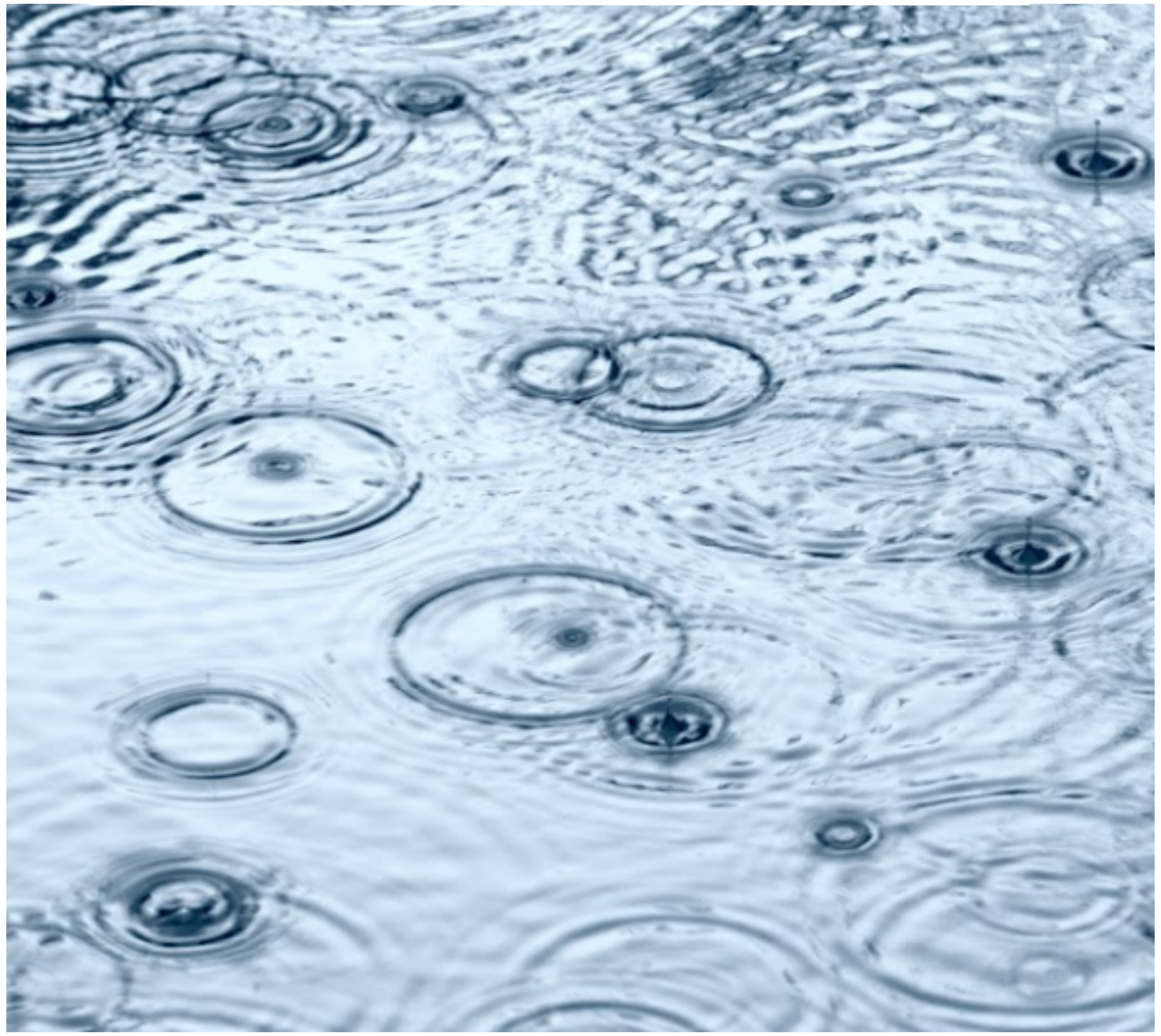
The Fast and Furious

Extreme data volume & rate from LHC collisions.

Multiple pp collisions in the same beam crossing:
LHC: 20-50. HL-LHC: 140-200

How to be prepared for LHC Run-3/4?

Now: The LHC



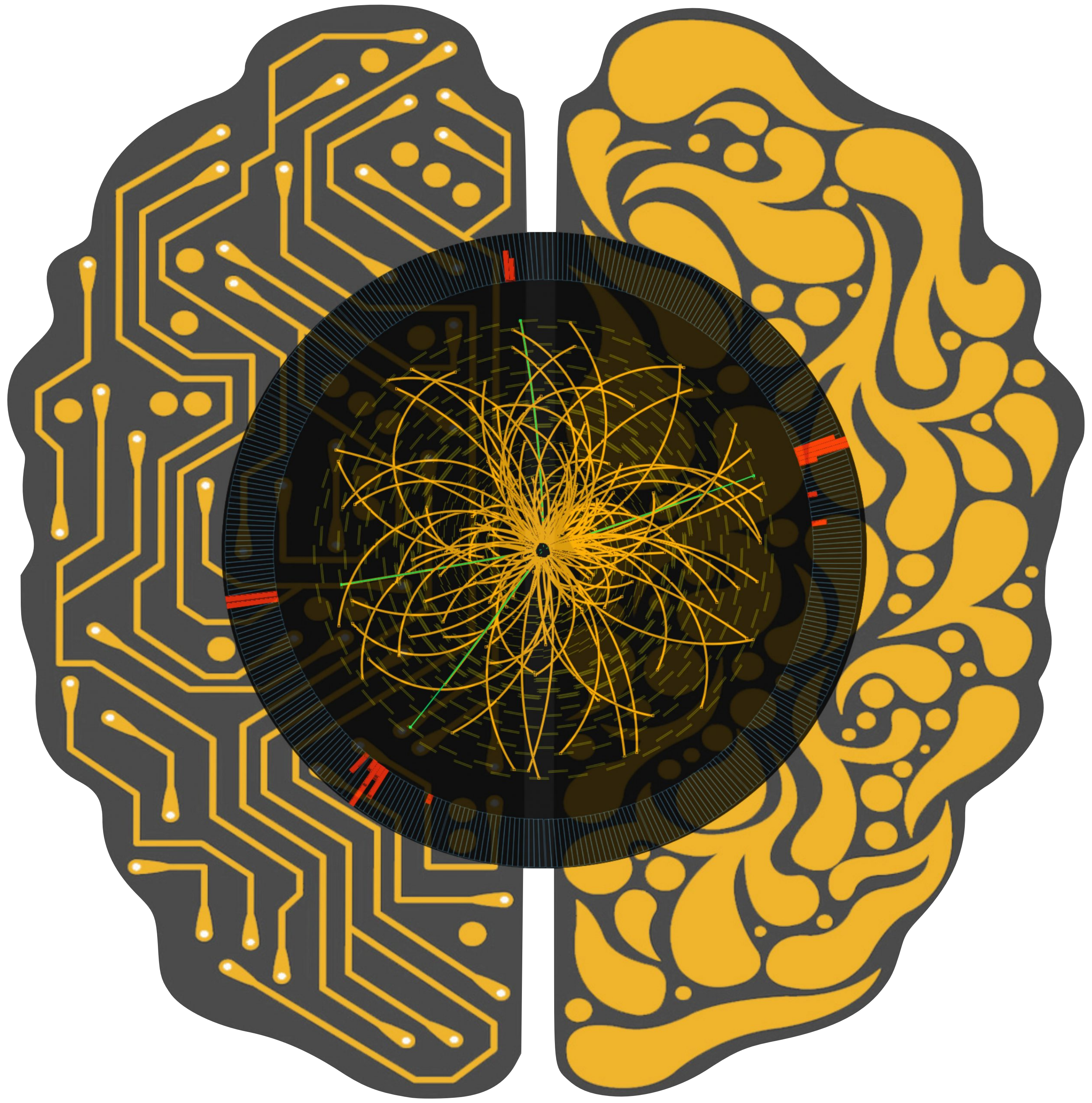
LHC Run-4

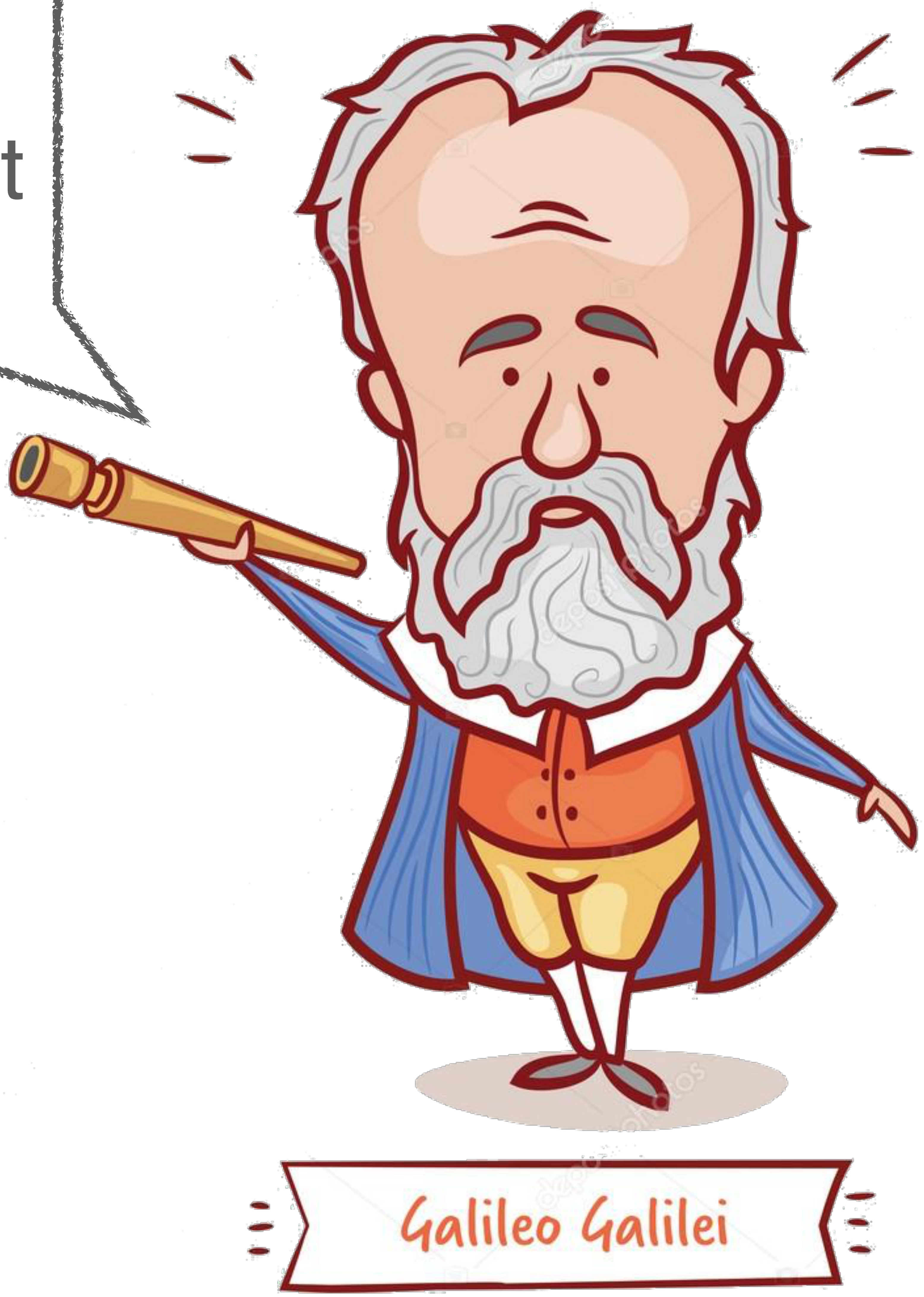
Higher pileup, fine granularity detectors.

Advanced algorithms to maintain/improve the acceptance of (un)conventional signatures

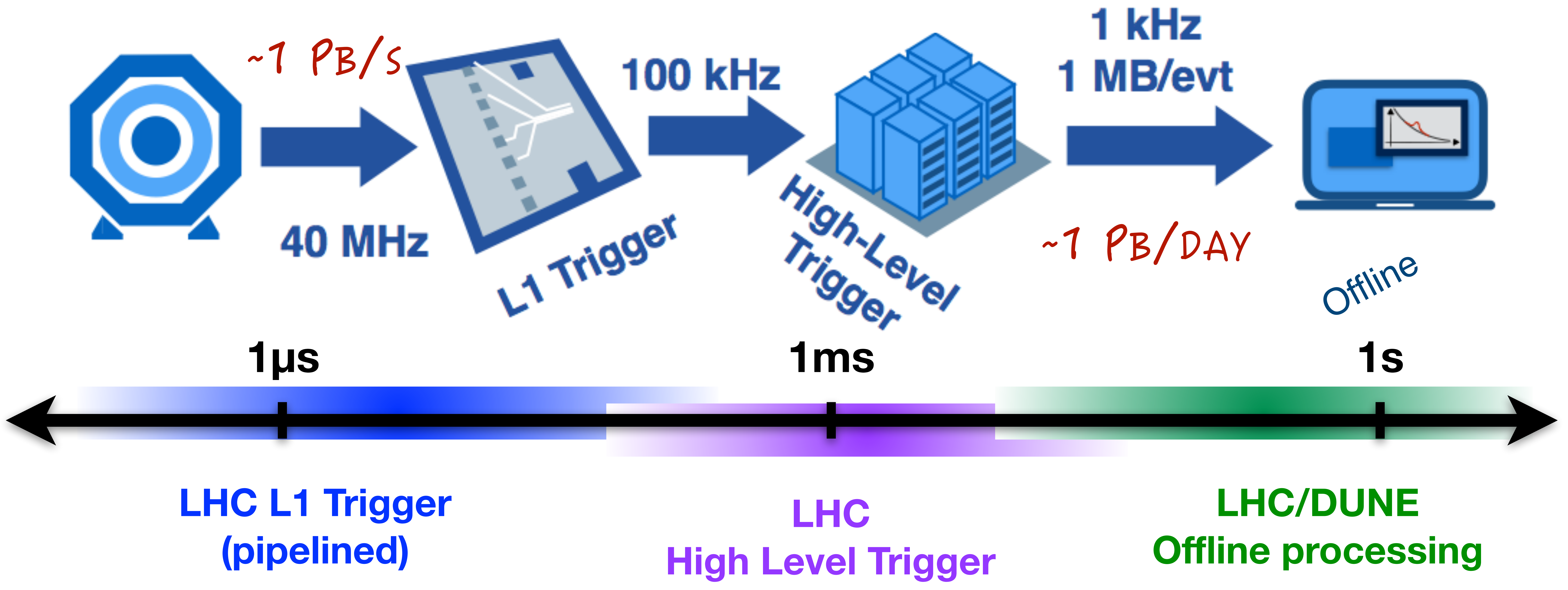
Ancient Wisdom With Modern Twist

Measure what is measurable
and make measurable what is not
so **with Artificial Intelligence**?





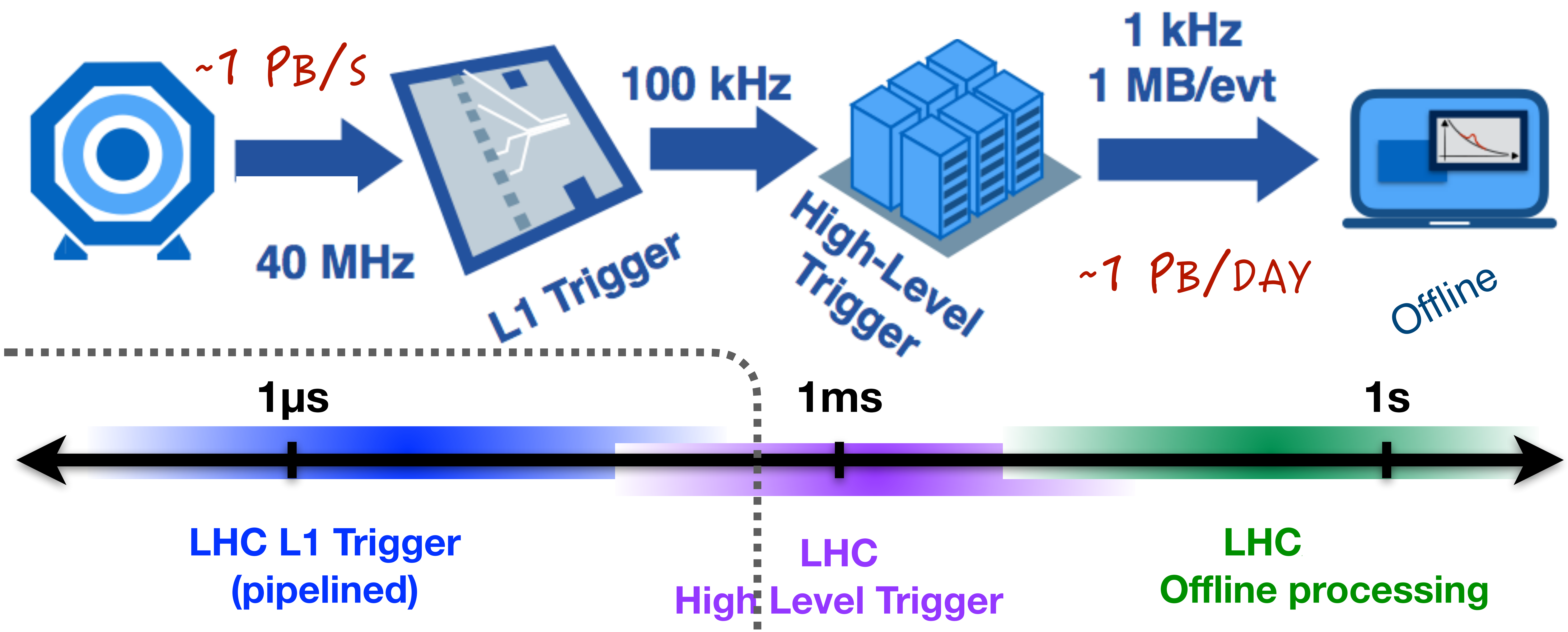
From Collisions to Discoveries



Fermilab

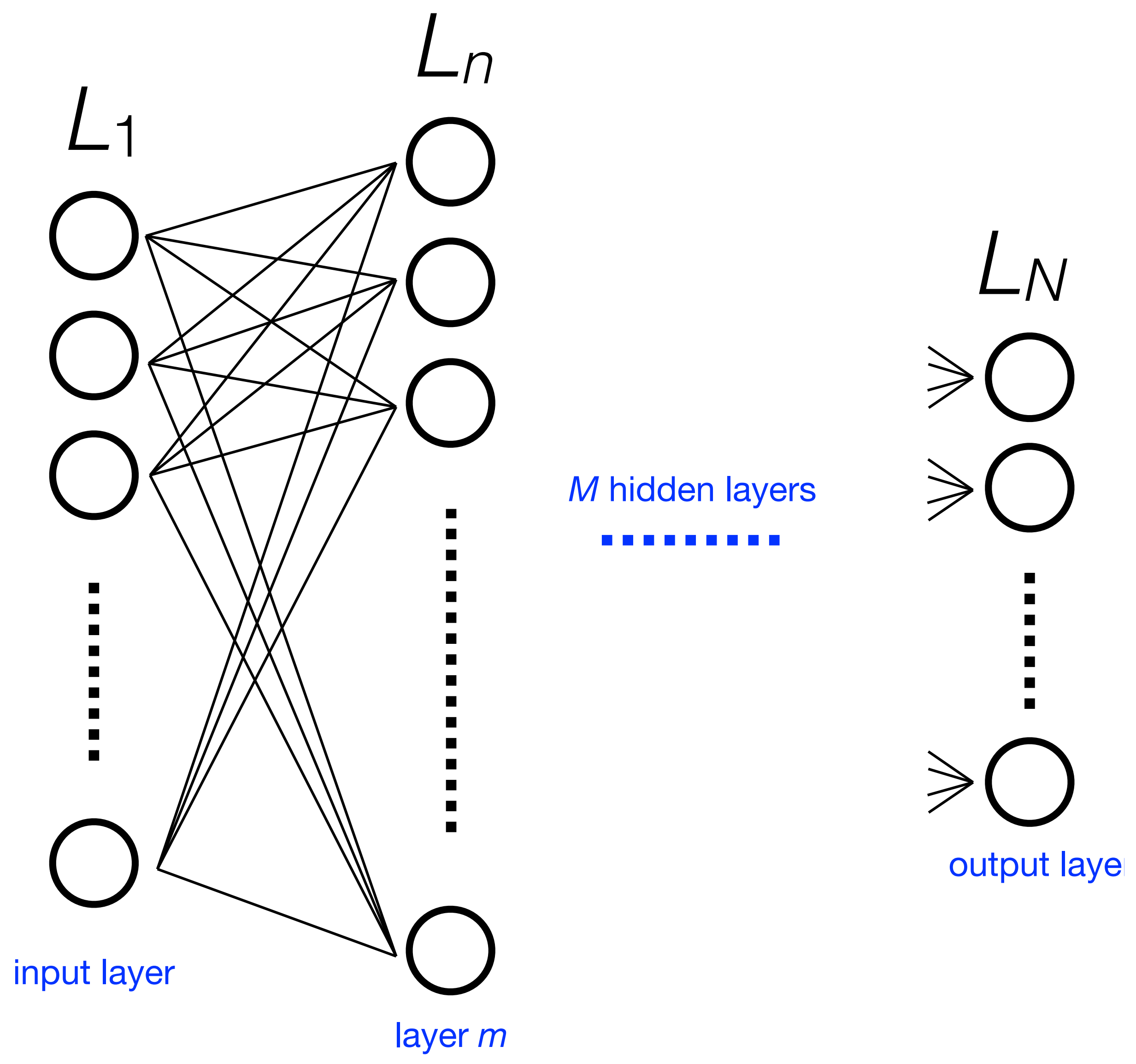
M.LIU

Real-time ML… @ Level-1

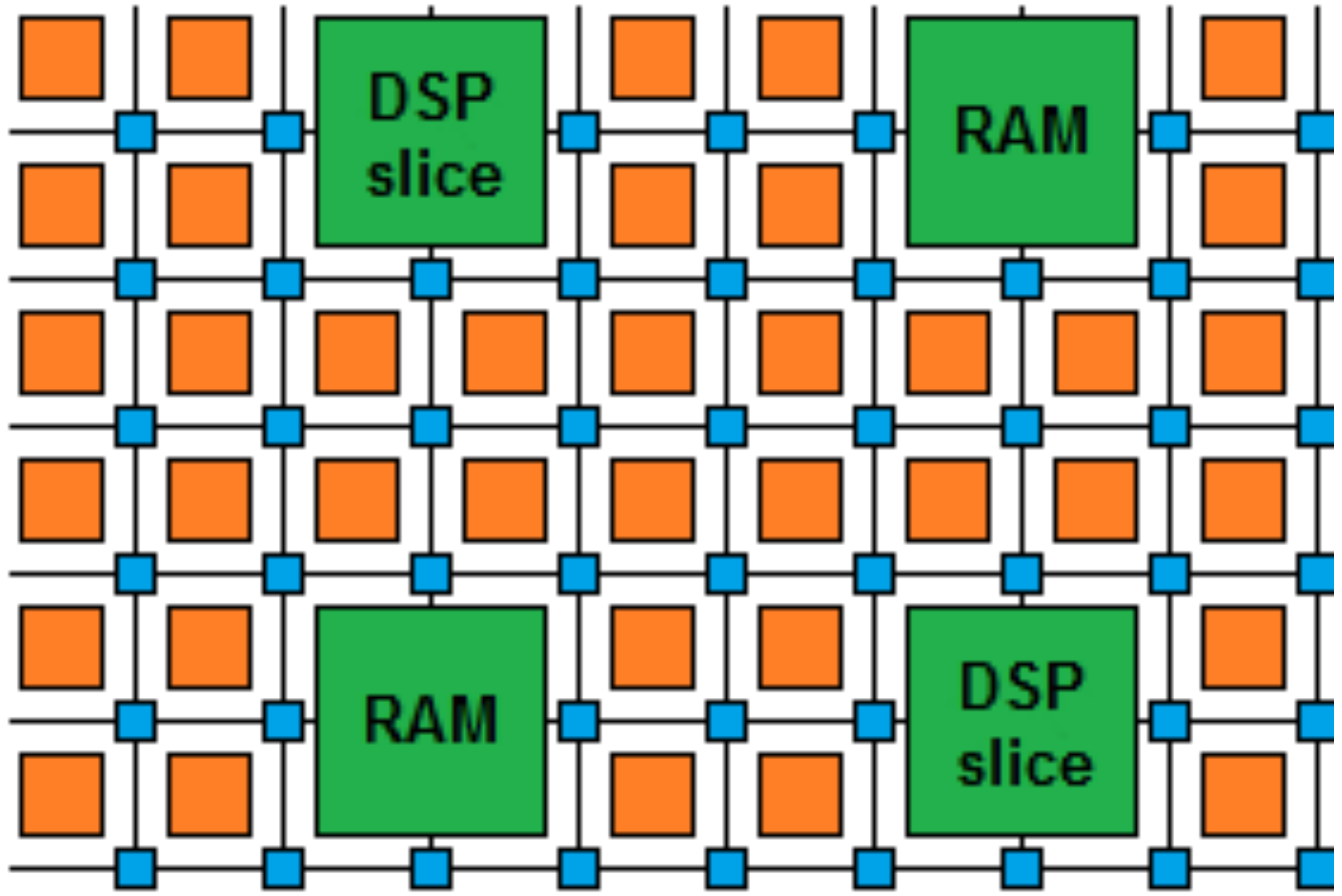


FPGAs/ASICs - high bandwidth low latency specialized compute hardware

NN on FPGAs



FPGA diagram



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

Activation functions

Precomputed, and stored in BRAMs

Multiplications

Digital Signal Processing DSPs

Addition

Logic cells

Natural fit for FPGAs… limited resources

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

Activation functions

Precomputed, and stored in BRAMs

Multiplications

DSPs

Addition

Logic cells

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

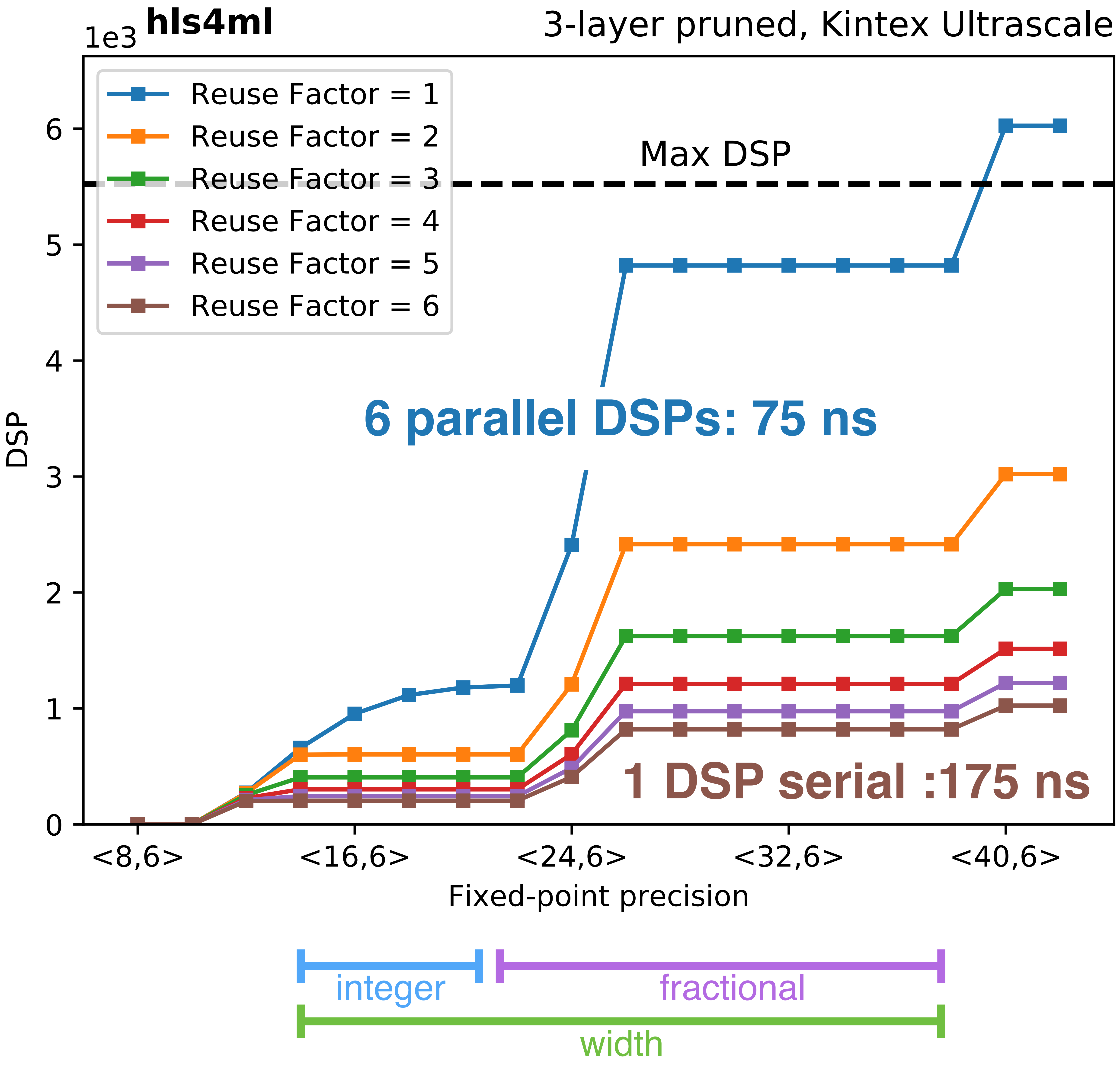
Virtex Ultrascale+ VU9P

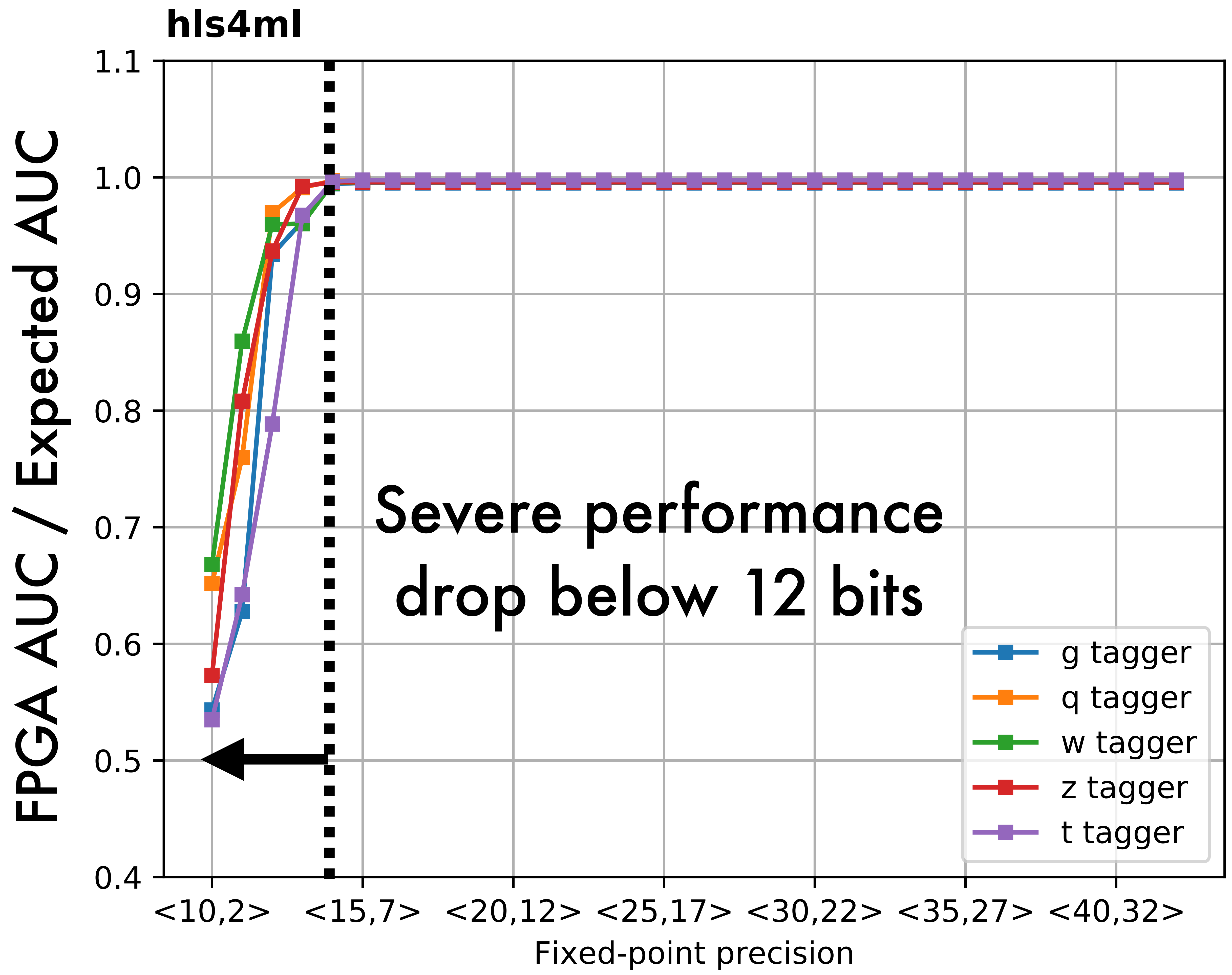
6800 DSPs
1M LUTs
2M FFs
75 Mb BRAM

Small network: thousands of connections

Limitation: Number of DSPs

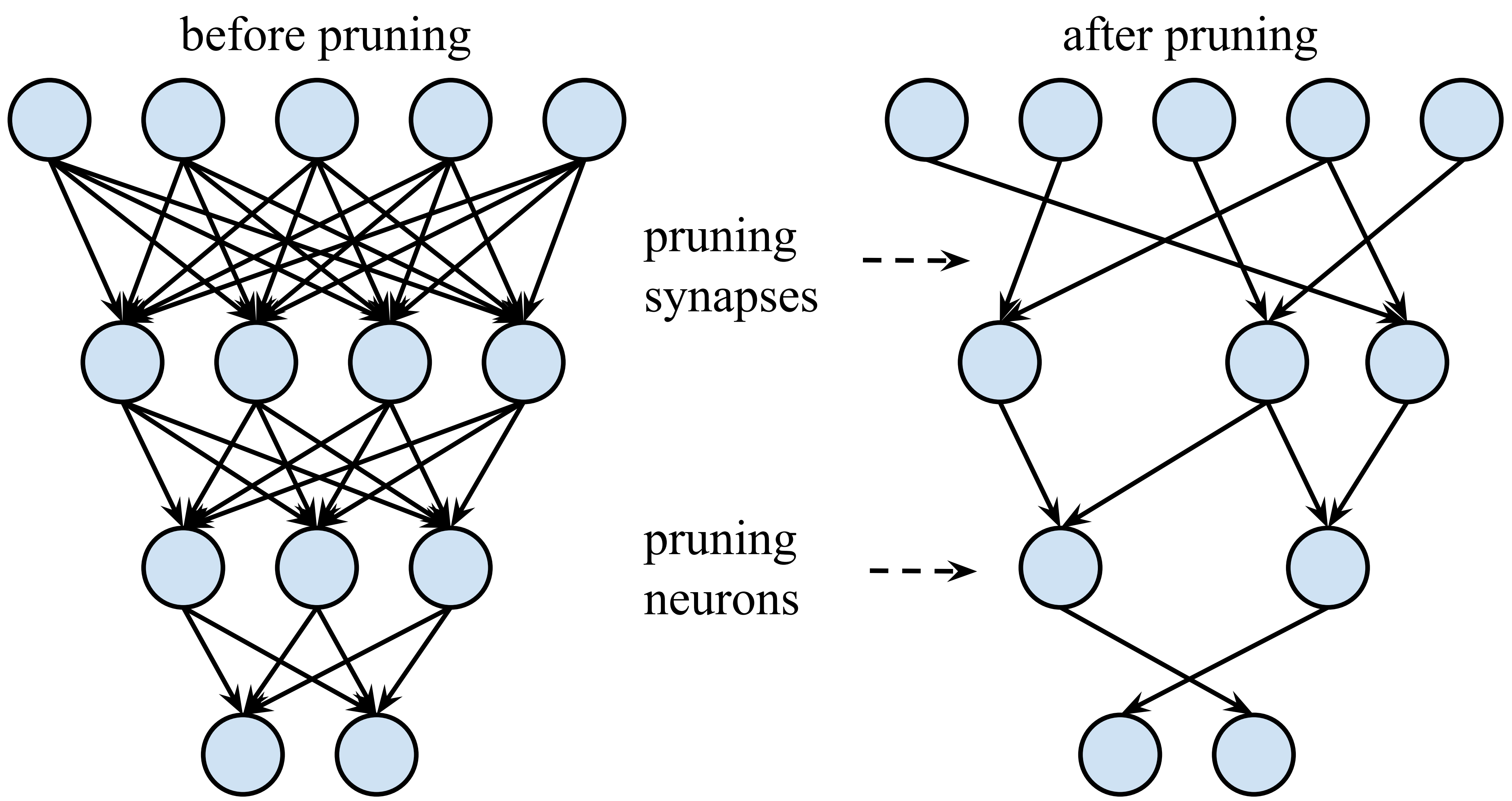
Fit NN on FPGA: Quantization & Reuse

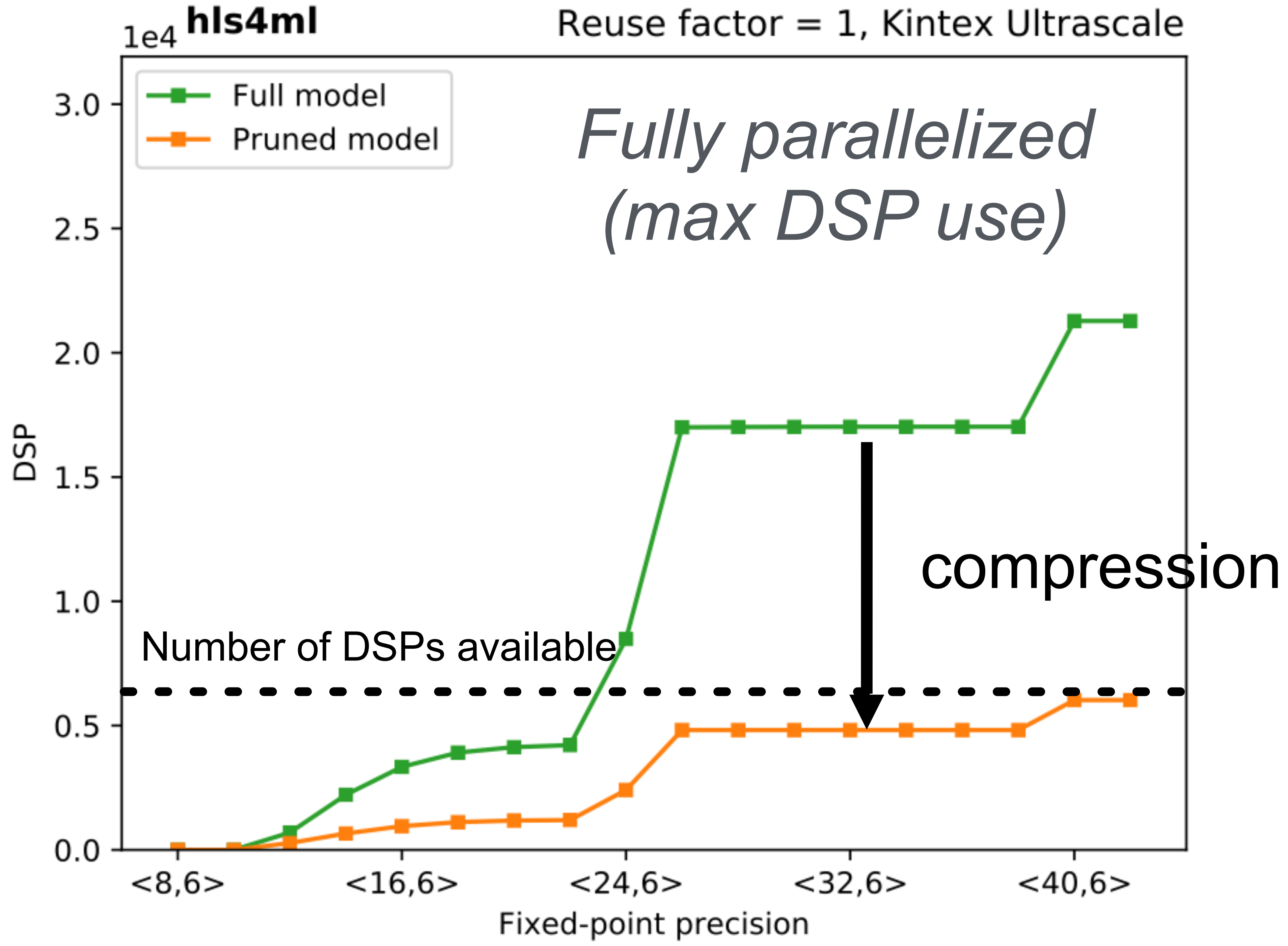




Efficient NN design: compression

[https://arxiv.org/abs/1804.06913]



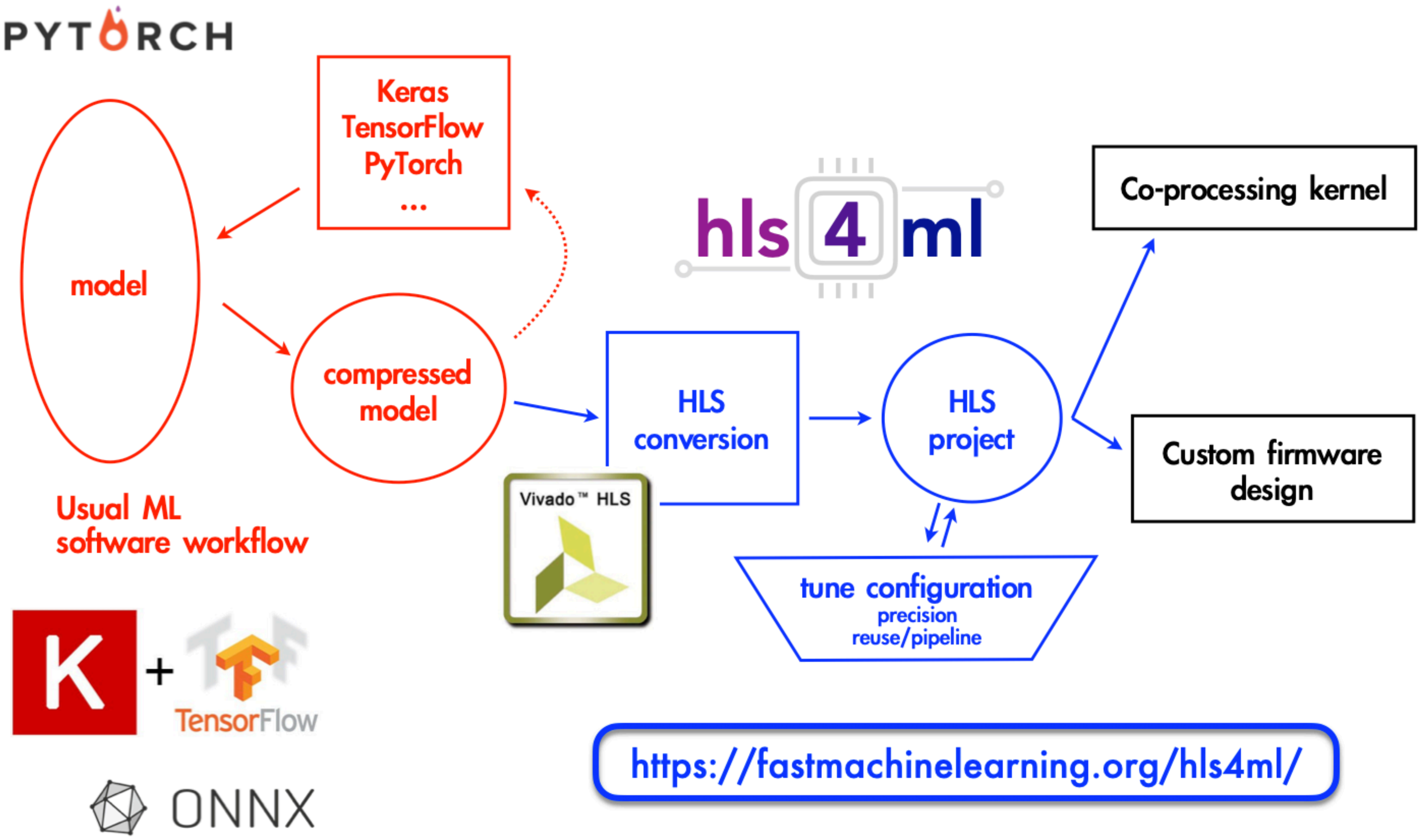


Neural Network compression is a widespread technique to reduce the size, energy consumption, and overtraining of deep neural networks

Several approaches in literature

arxiv.1510.00149, arxiv.1712.01312, arxiv.1405.3866, arxiv.1602.07576, doi:10.1145/1150402.1150464

High Level Synthesis 4 Machine Learning



https://fastmachinelearning.org/hls4ml/

Recent developments

First paper demonstrated a fully connected NN in 100 ns.

HLS4ML in CMS

Run 3: muon momentum regression in CMS

More models demonstrated for Phase-2 trigger upgrade TDR

Advanced models:

binary/ternary, CNNs, RNNs, auto-encoders. Support for Graph Neural Network Models

Advanced Pruning/quantization:

Quantization-aware training with QKeras/Quantization-aware pruning

On ASICs and Low power devices.

For latest status: please check hls4ml website, CPAD 2021 talk,

Try it out: hls4ml tutorials

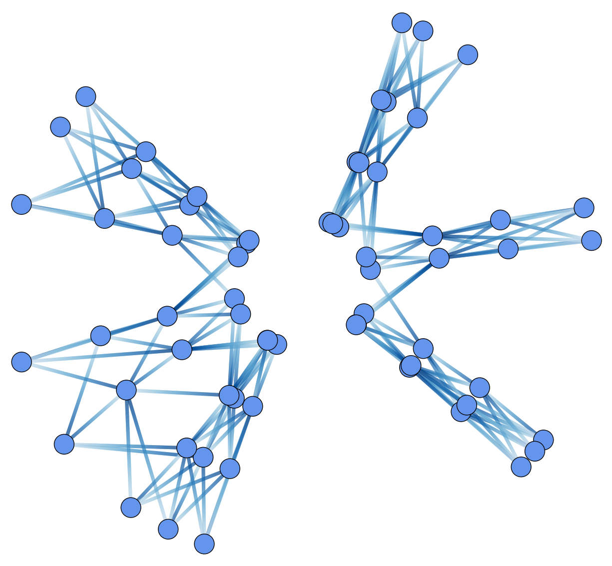
Including Graph Neural Networks

Application Algorithms drive HLS4ML developments.

Graph Neural Networks (GNNs):

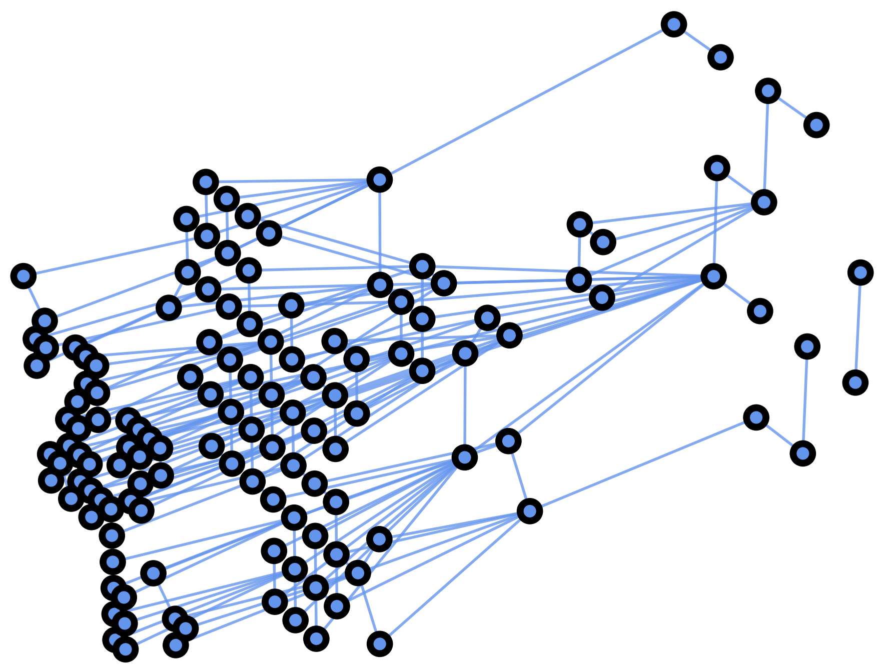
Represent data as nodes and edges

appropriate representation of particle physics data: irregular, structural, relational

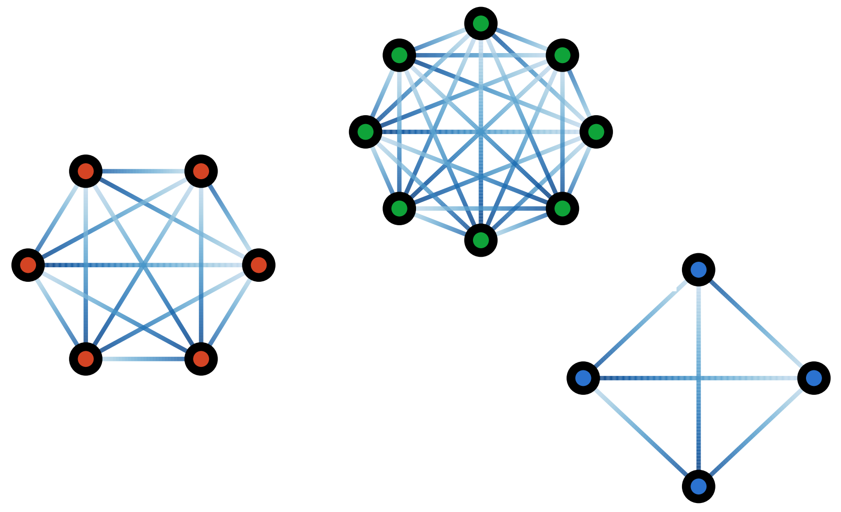


GNN developments driven by social media.

Active development and applications in science domains: how to adapt to domain knowledge & applications

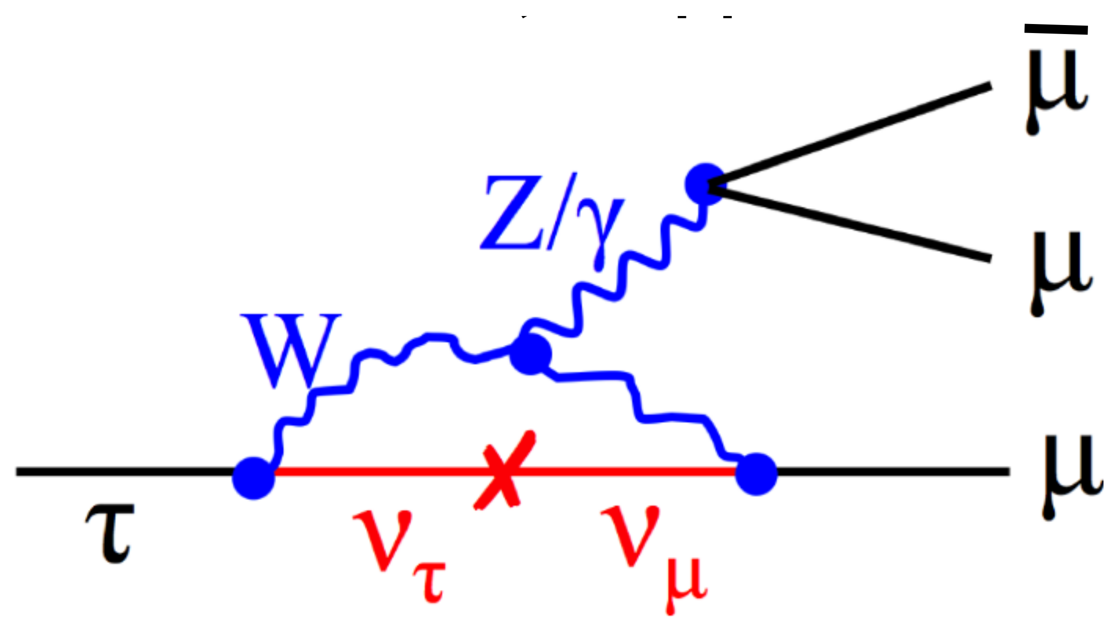


Many variations of GNNs. Problem formulation is important: Node/edge classification, graph classification, identifying subgraphs



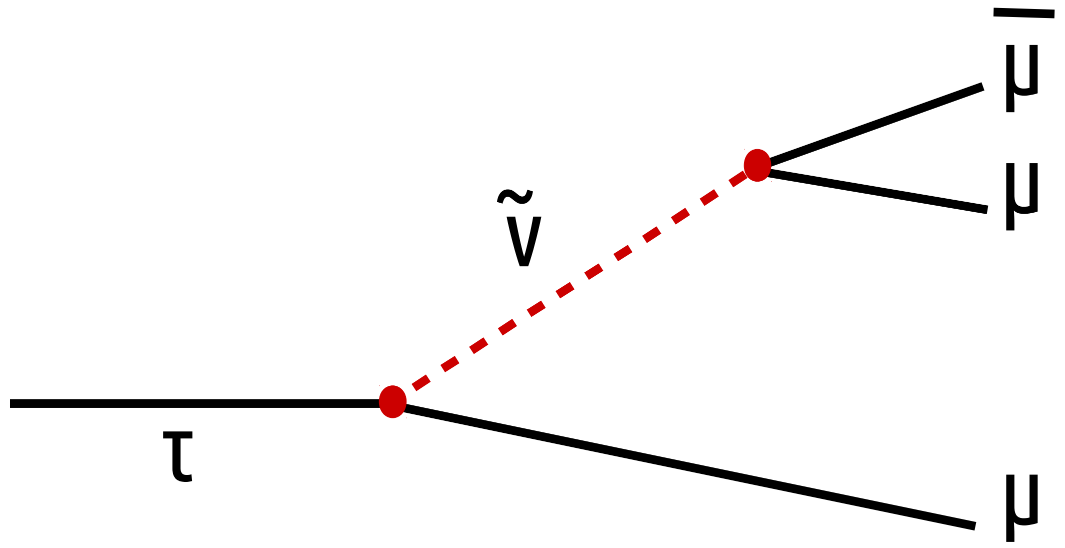
Will show two GNN studies at Purdue.

$\tau \to 3\mu$: motivation and current trigger

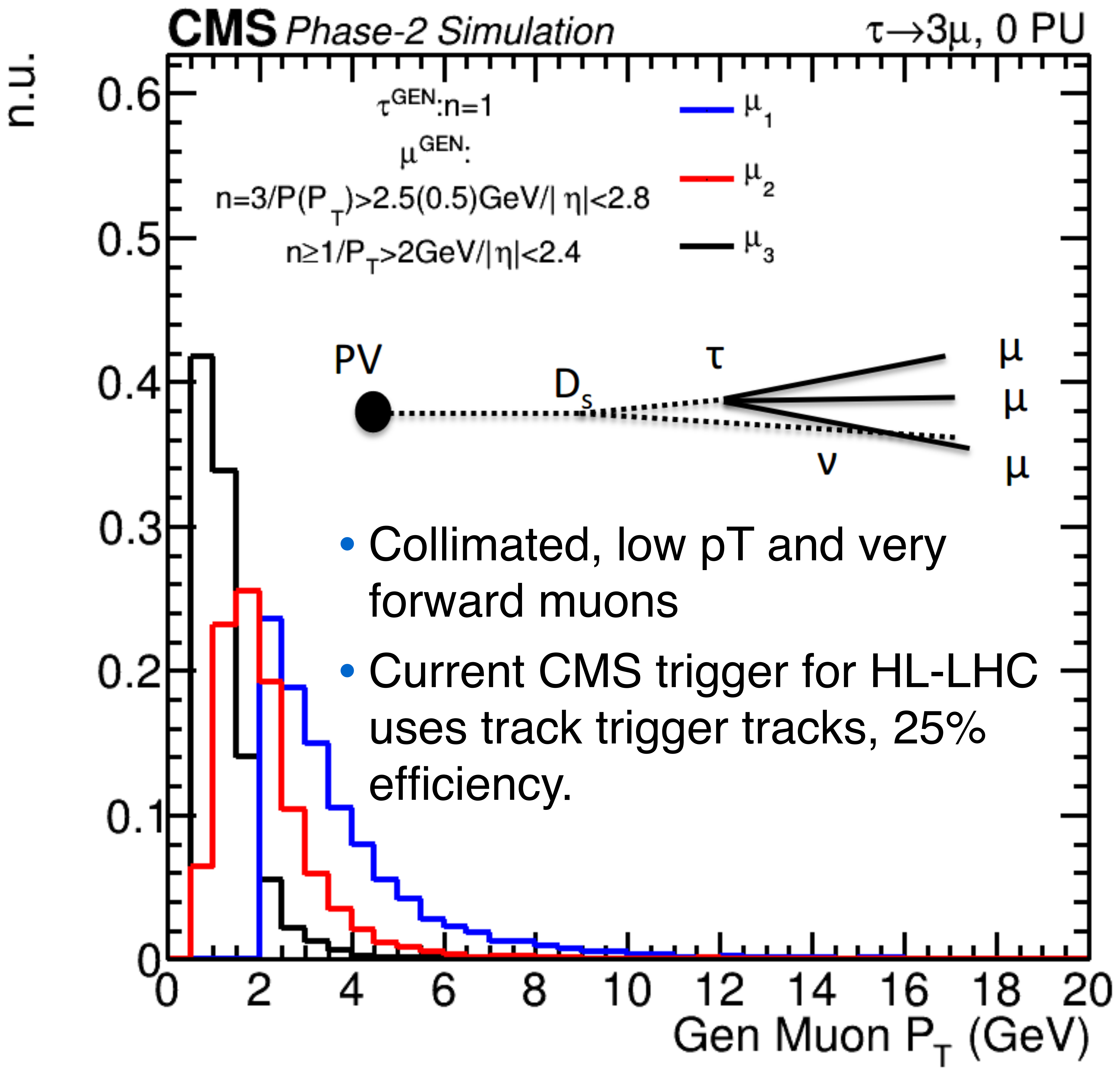


Very rare in the SM, Neutrino oscillations:

BR ~O(10^{-14})



R-parity violating SUSY



- Collimated, low pT and very forward muons
- Current CMS trigger for HL-LHC uses track trigger tracks, 25% efficiency.

GNN τ->3μ classifier

Graph Inputs: Muons hits (Coordinates and bending angle) from L1 primitives, represented as nodes in graph.

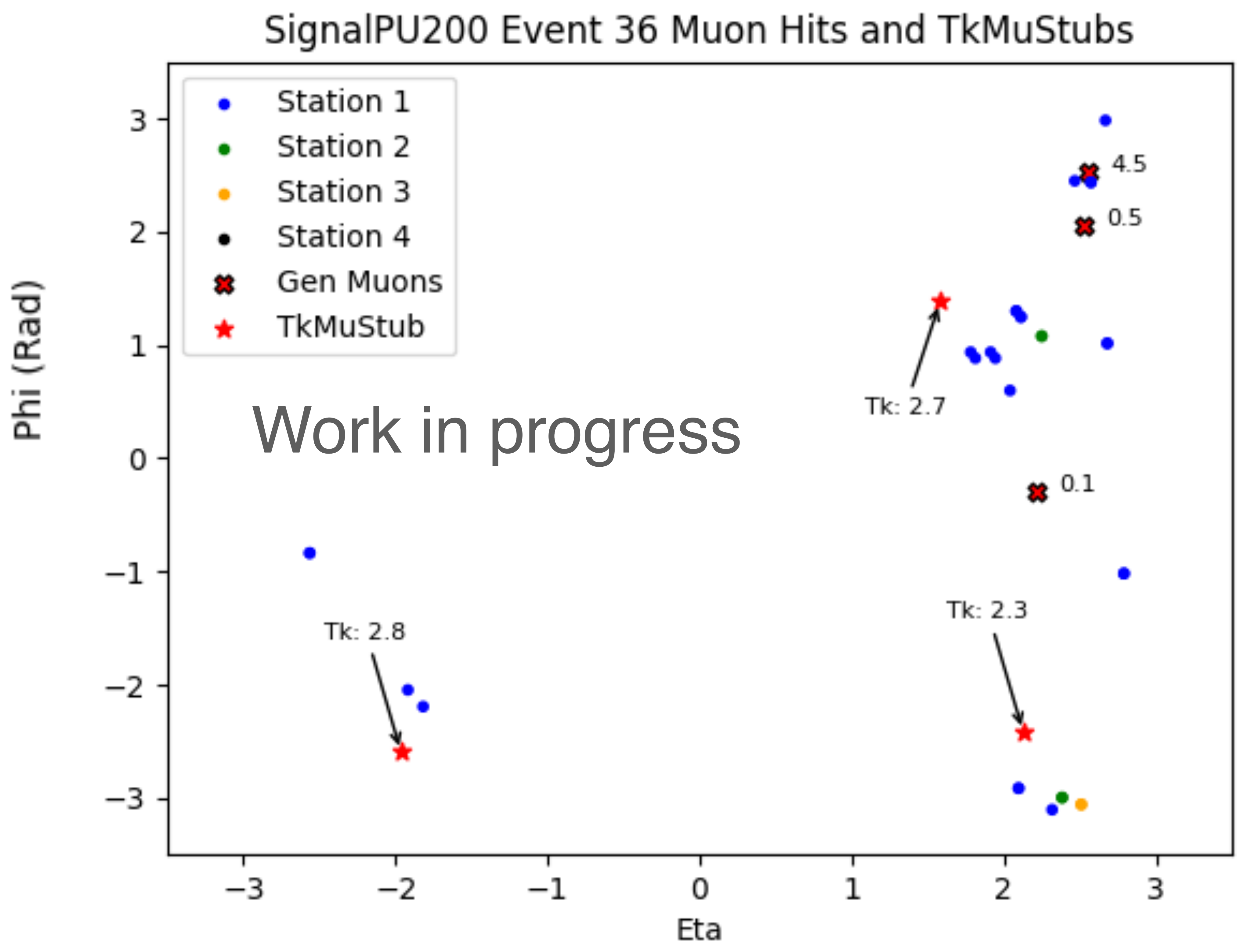
Attention mechanism for information aggregation between local nodes and global node

Training setting:

τ->3μ signals mixed with PU backgrounds vs pure PU sample

> 90% efficiency for 30 kHZ trigger bandwidth

Fully connected network: ~26%



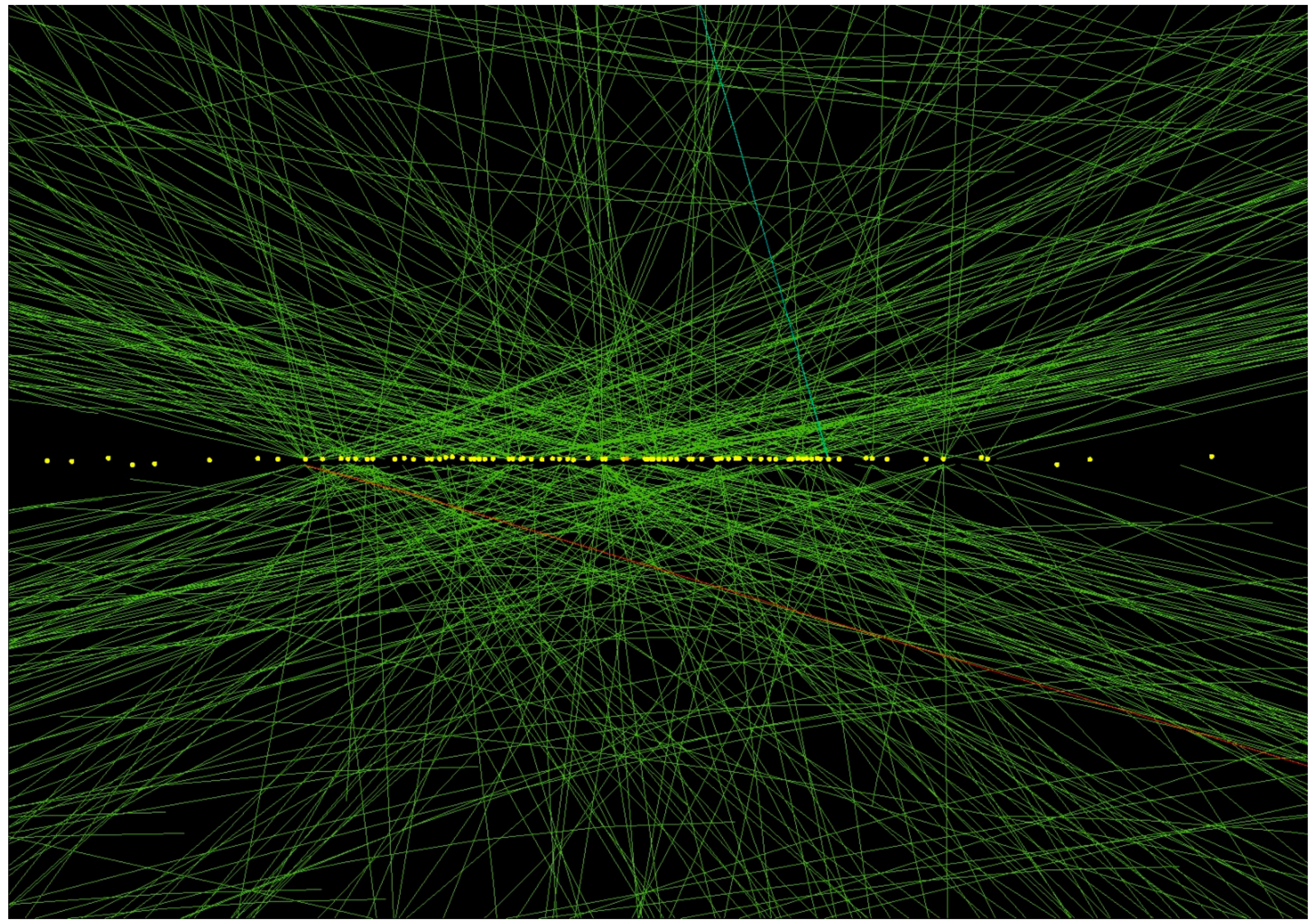
On-going: Model interpretations, Data augmentation. Regression of muon kinematics, Model adaption (to other signals & other experiments), implementation with HLS4ML.

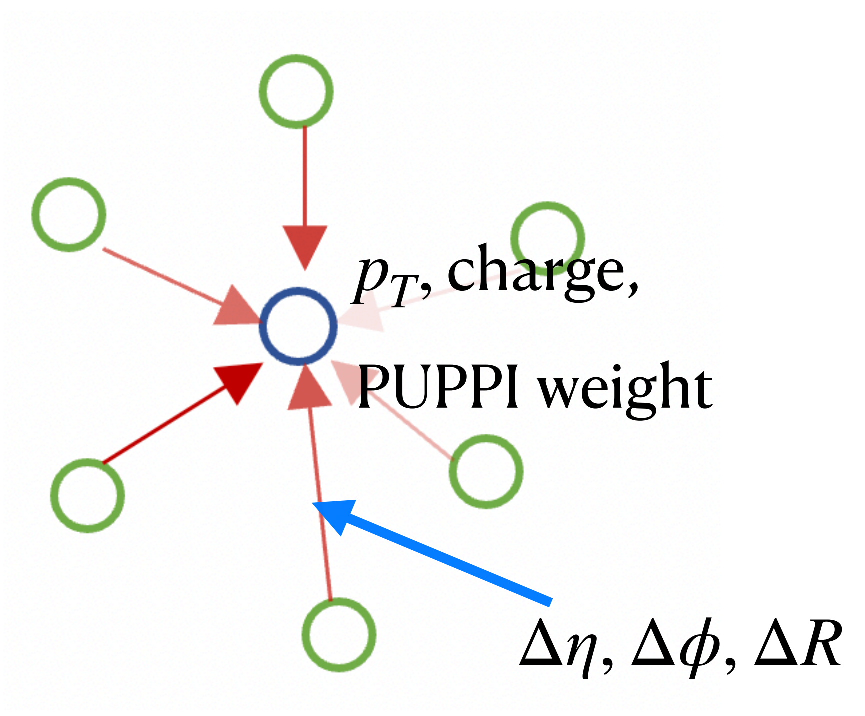
The Fast and Furious

Semi-supervised Pileup mitigation with GNN

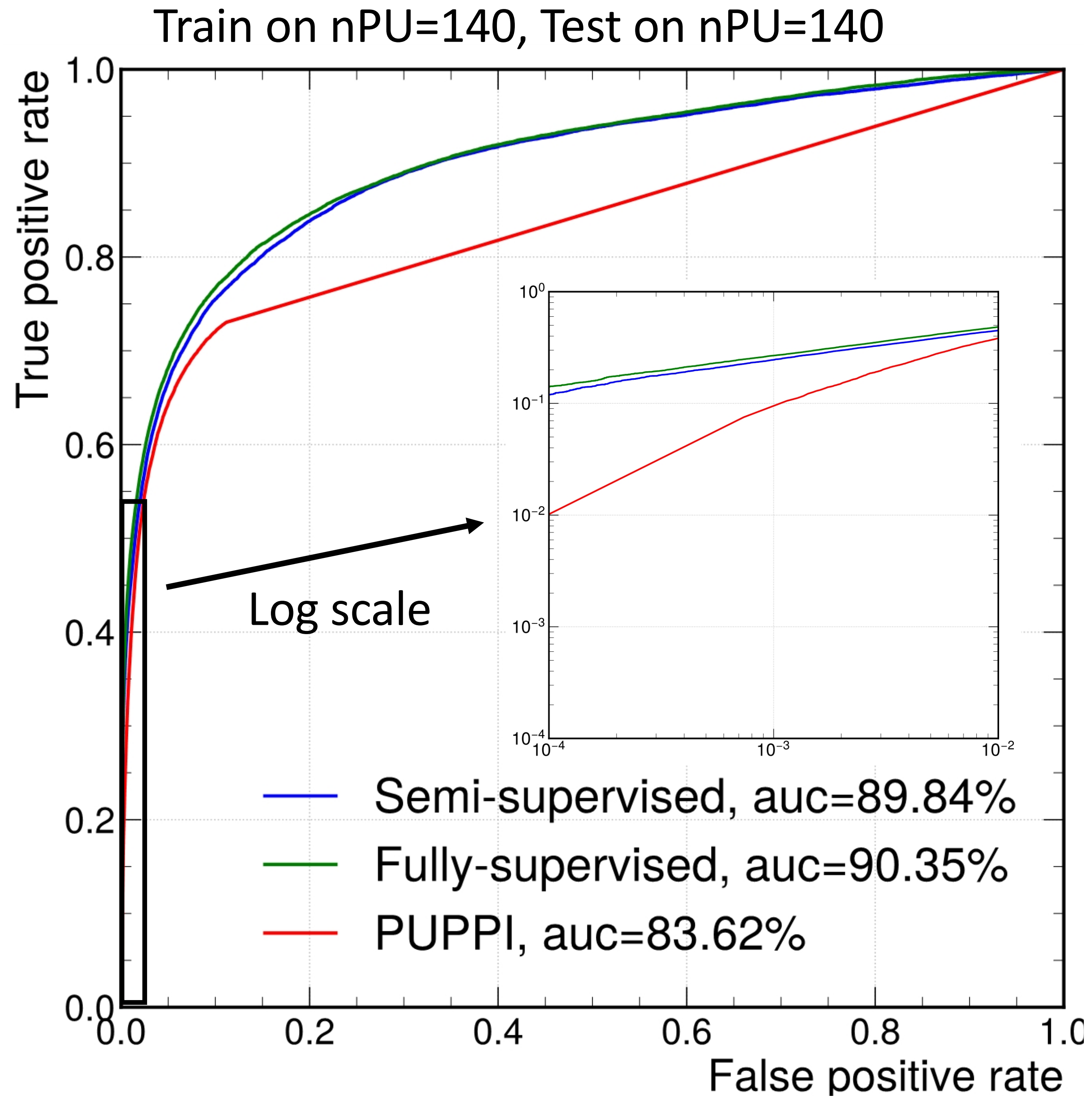
Improve Per-Particle Pileup Mitigations With better

- Trained on charged particles and applied to neutral ones —> can learn from data.
- Outperforms Puppi, comparable to fully supervised method.
- Presented at BOOST 2021, Short paper submitted to NeurIPS 2021 AI for Science Workshop. Long paper targeting PRD in preparation.
- Next: Apply to CMS simulation & data. Neutral particle vertex association in for the forward region.

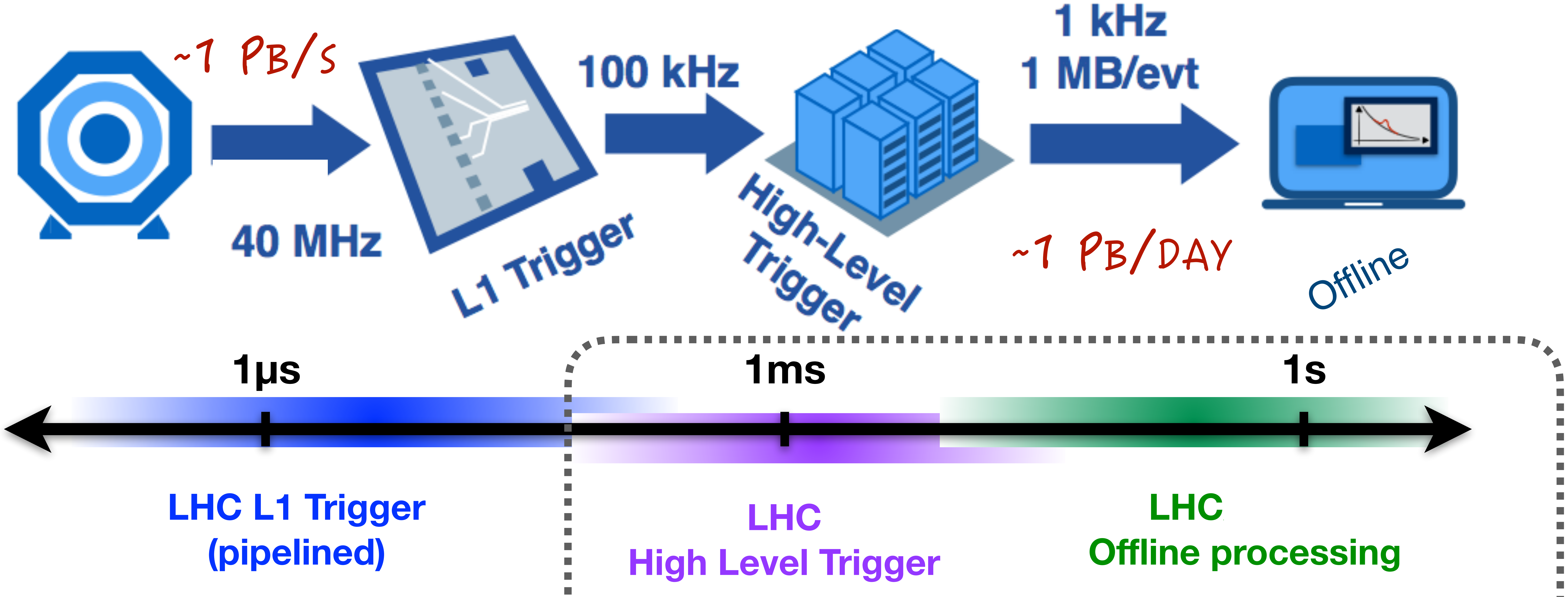




Neighboring structure



Speeding up HLT & Offline



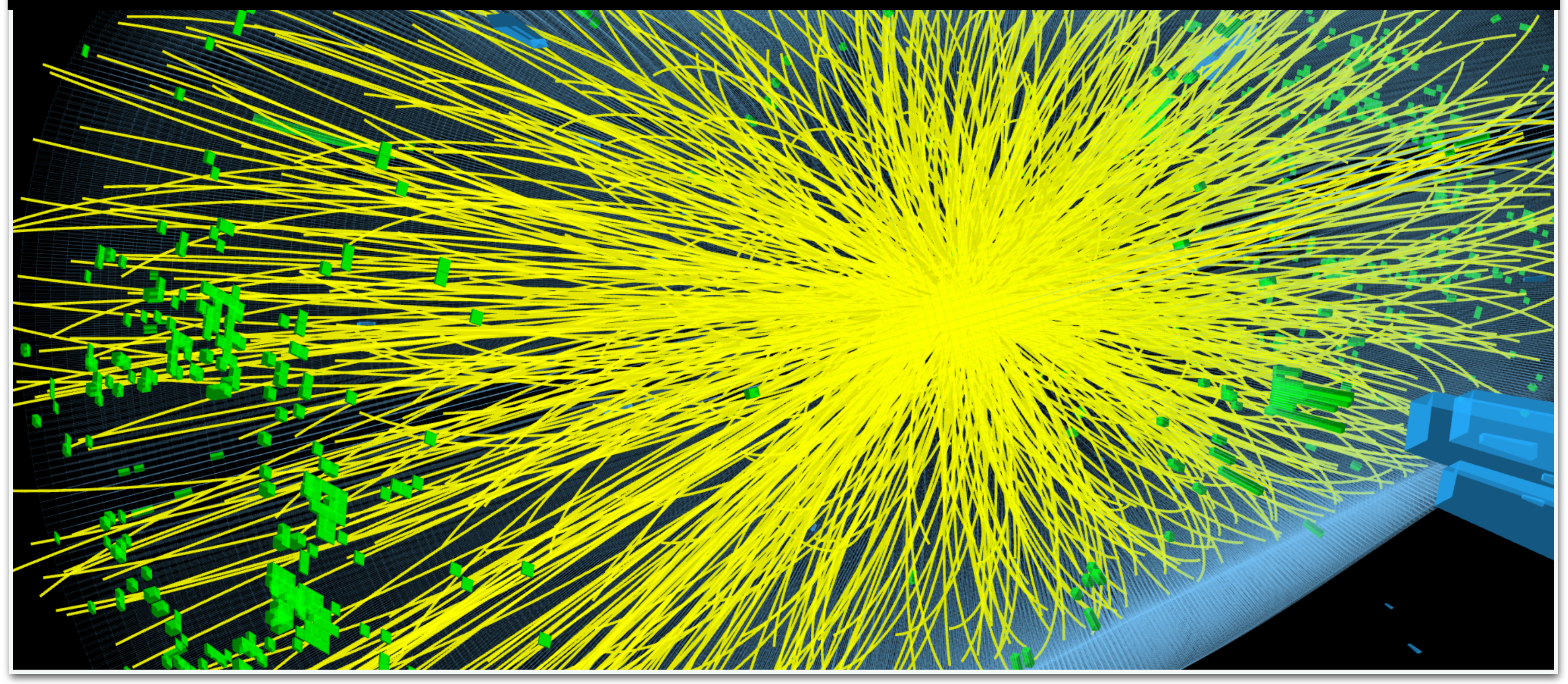
Traditionally CPU; Moving towards heterogeneous computing

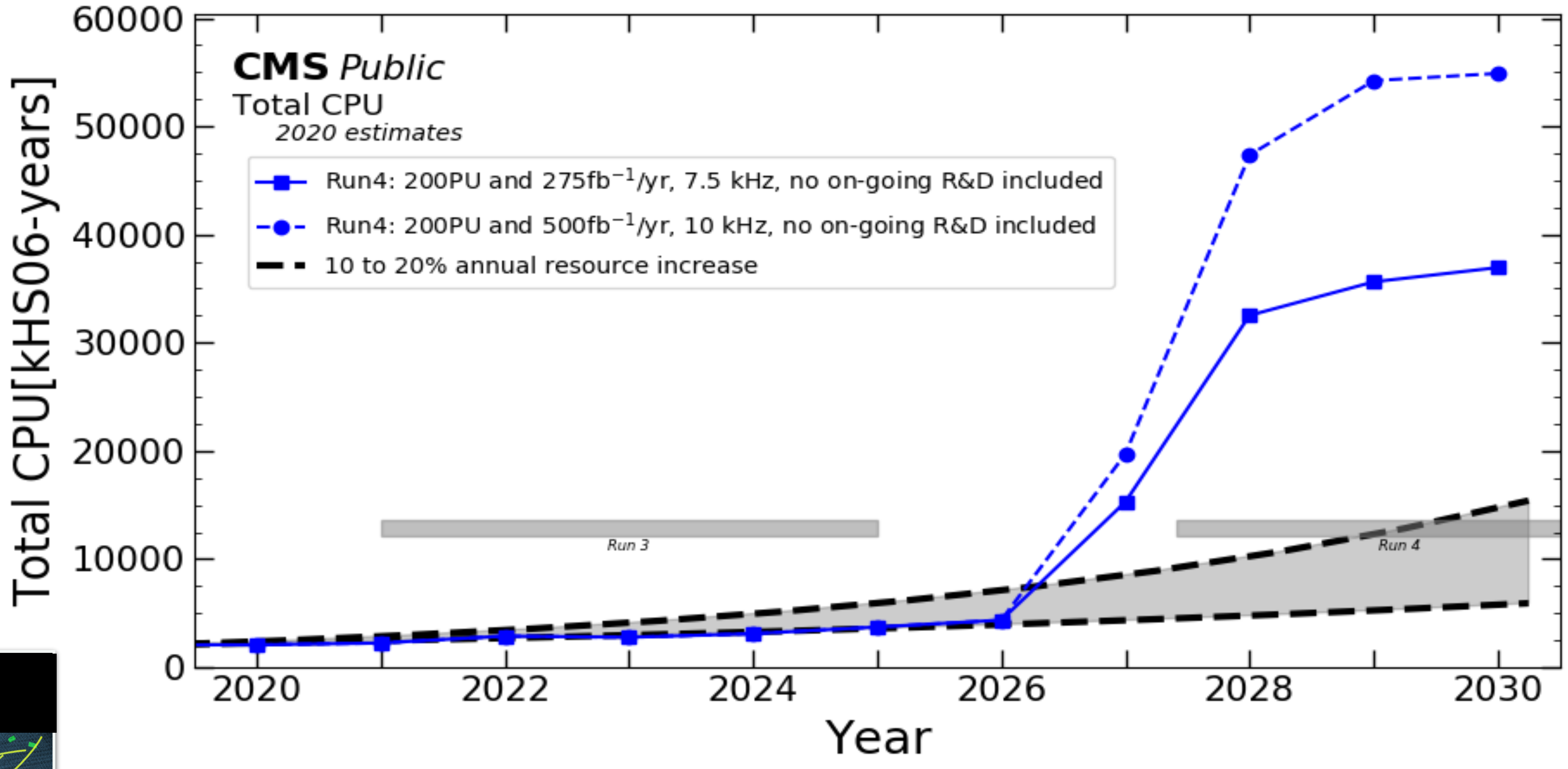
HL-LHC:Big Data Challenge

>5x Event complexity

>10x Data Volume

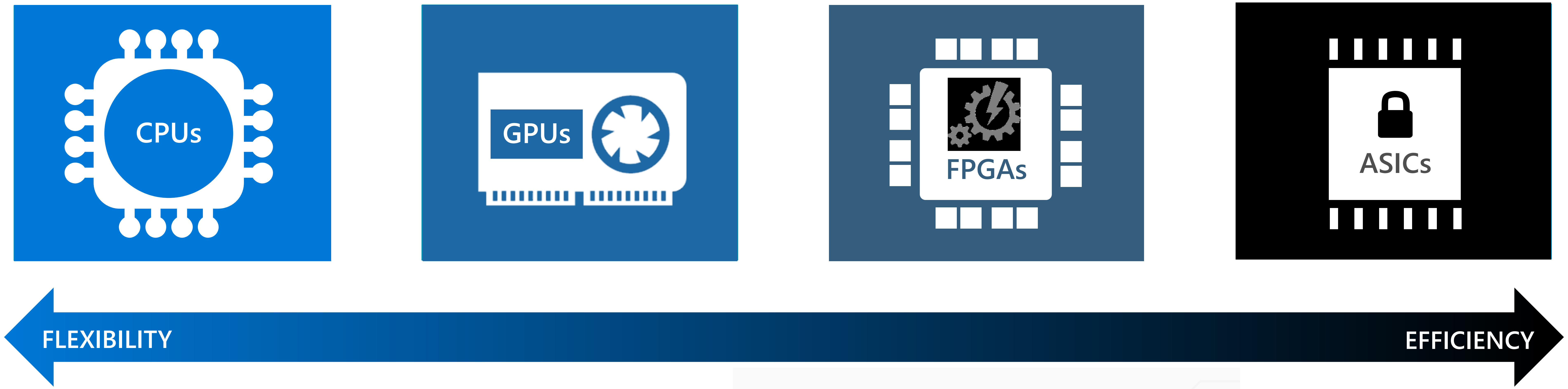
Current: ~5 minutes per HL-LHC event

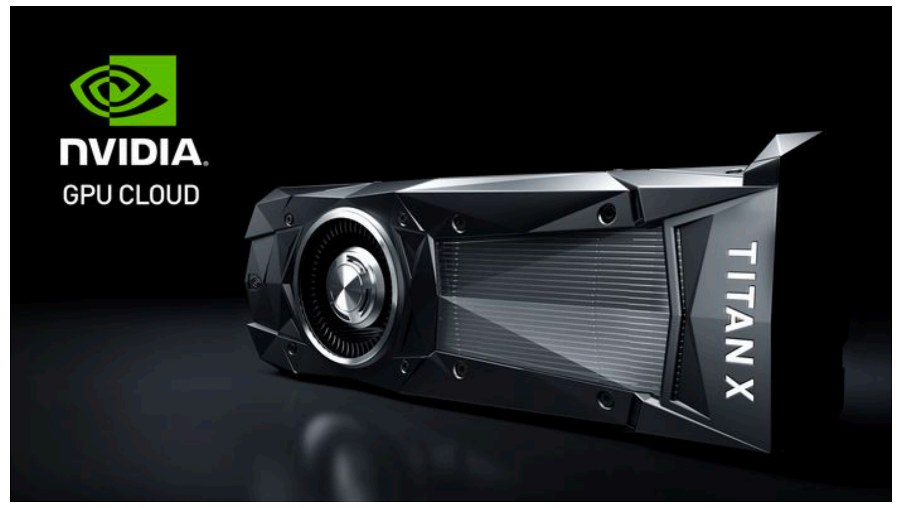


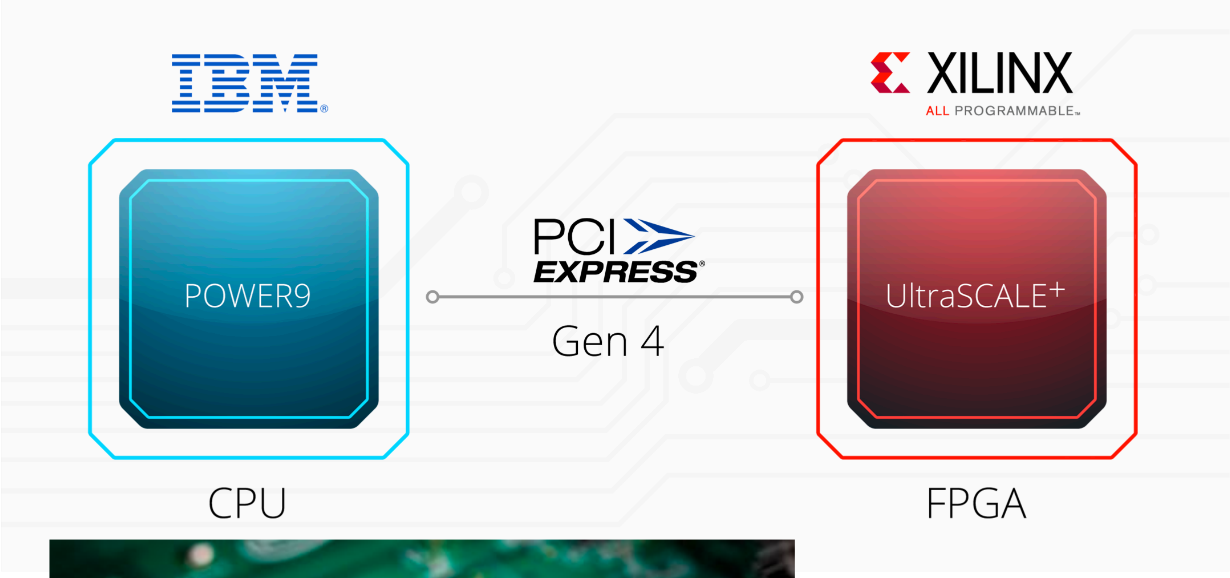


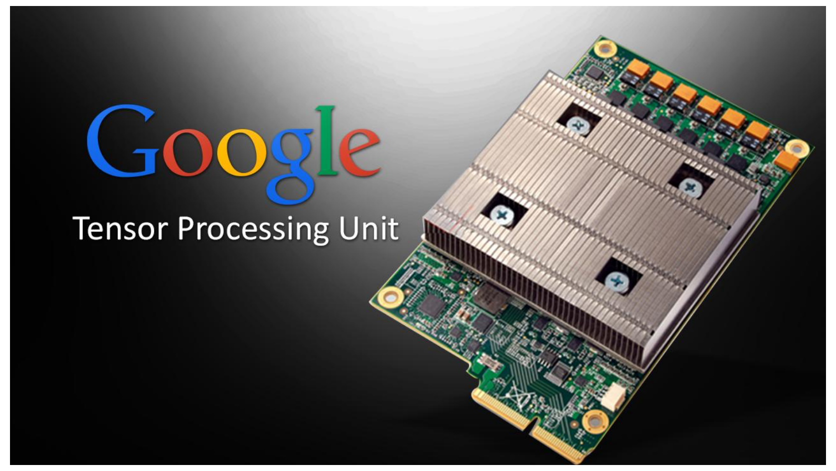
Moore's Law continues
…but Dennard Scaling fails.
No faster computers for free!

#Trending in Industry: Heterogeneous Computing

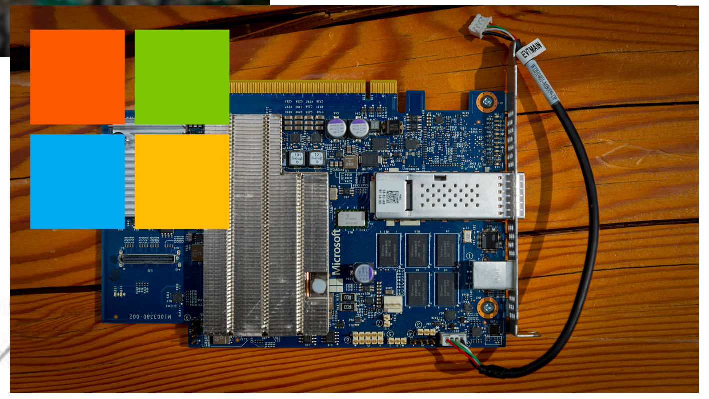






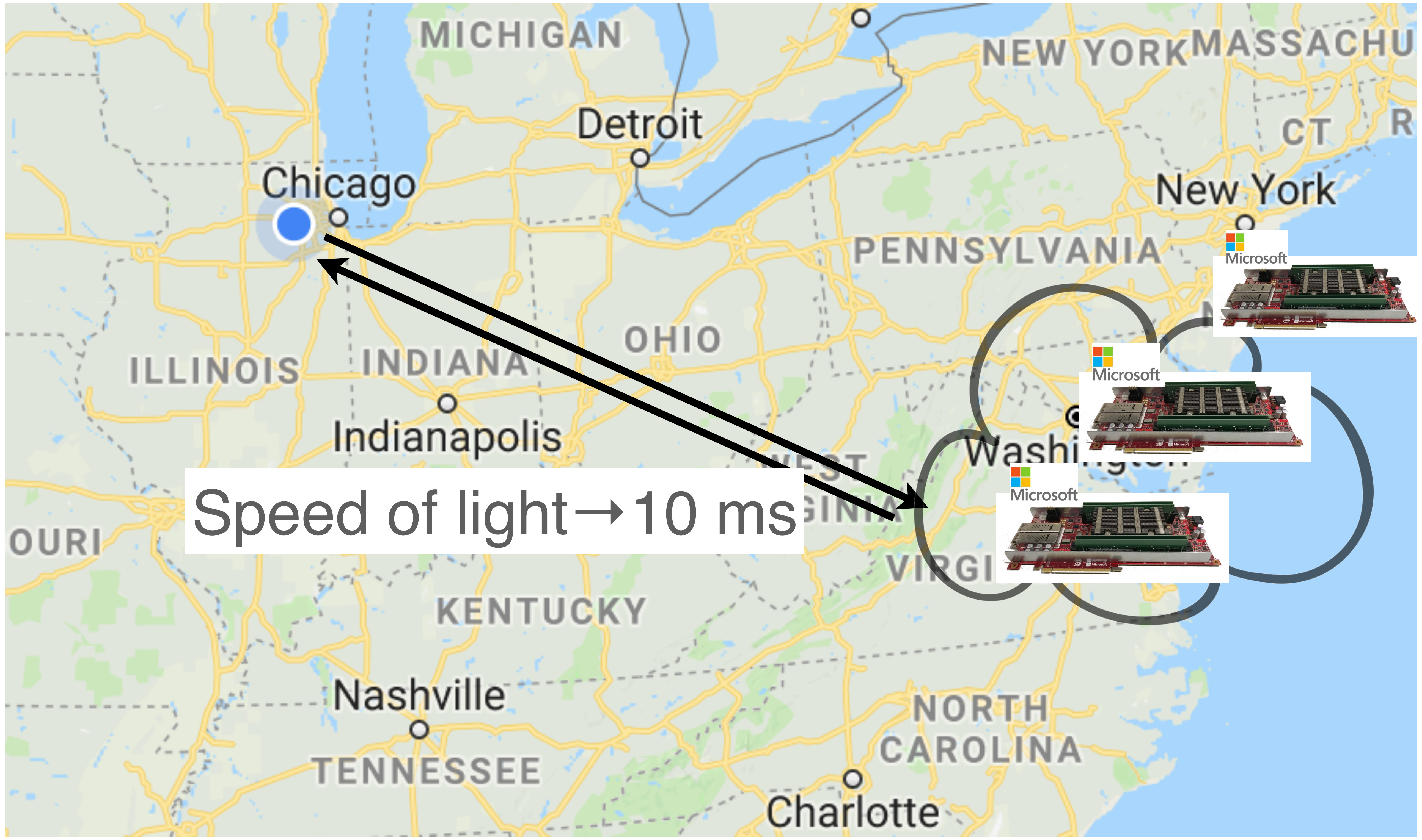


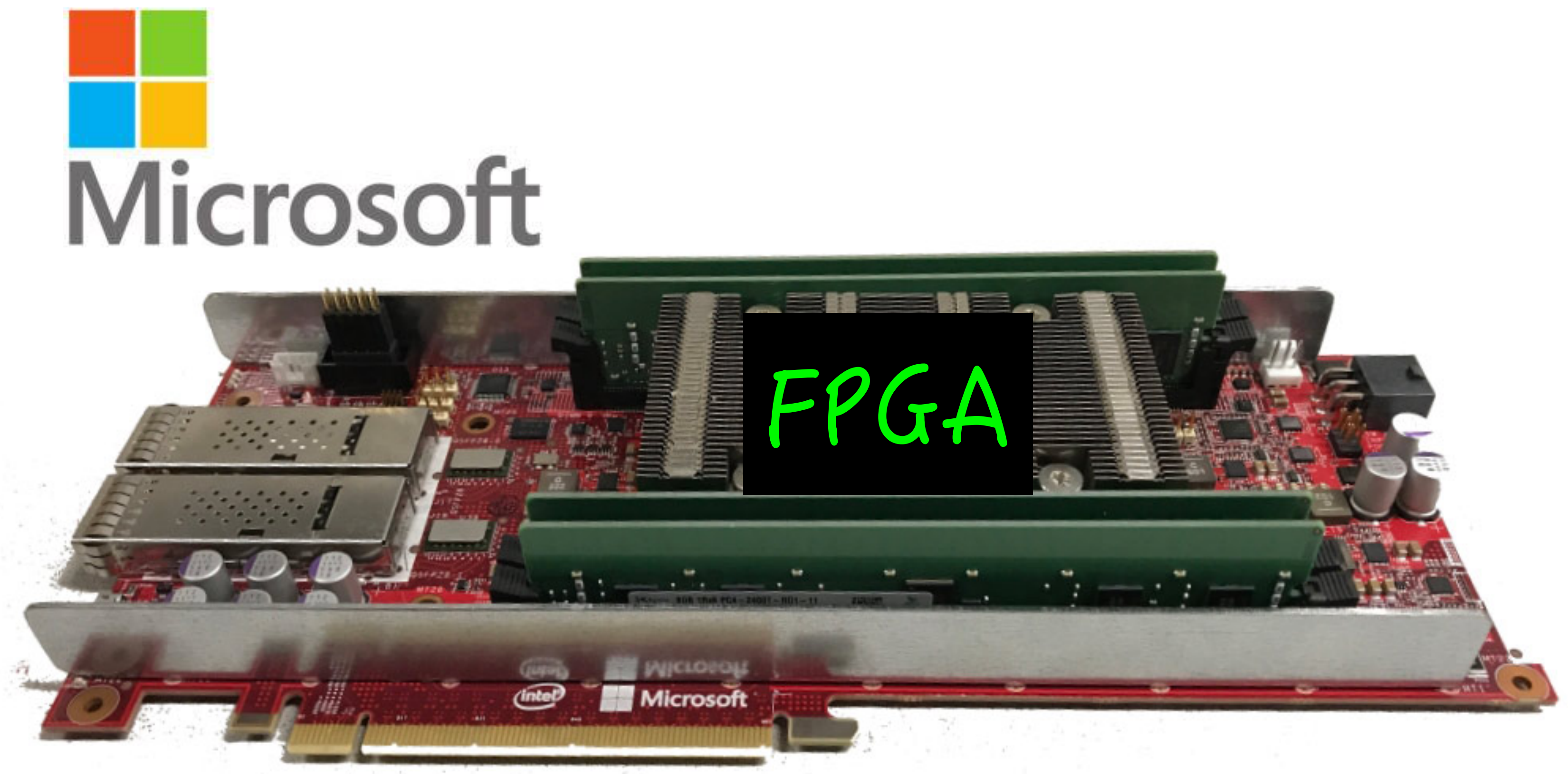
Advances driven by
big data explosion
& machine learning

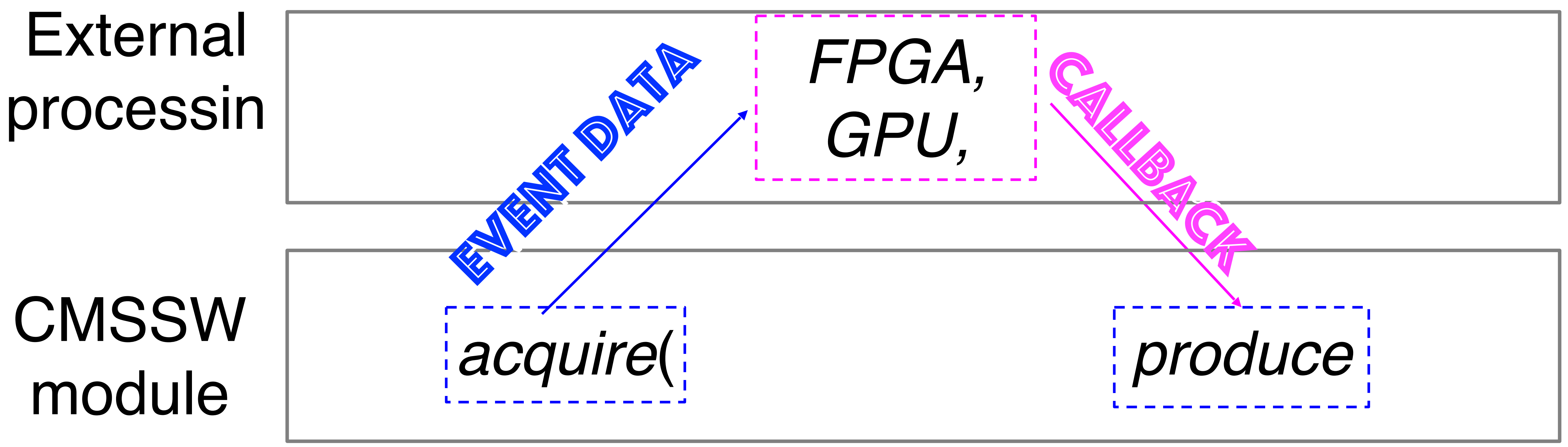


Machine Learning Inference as-a-service

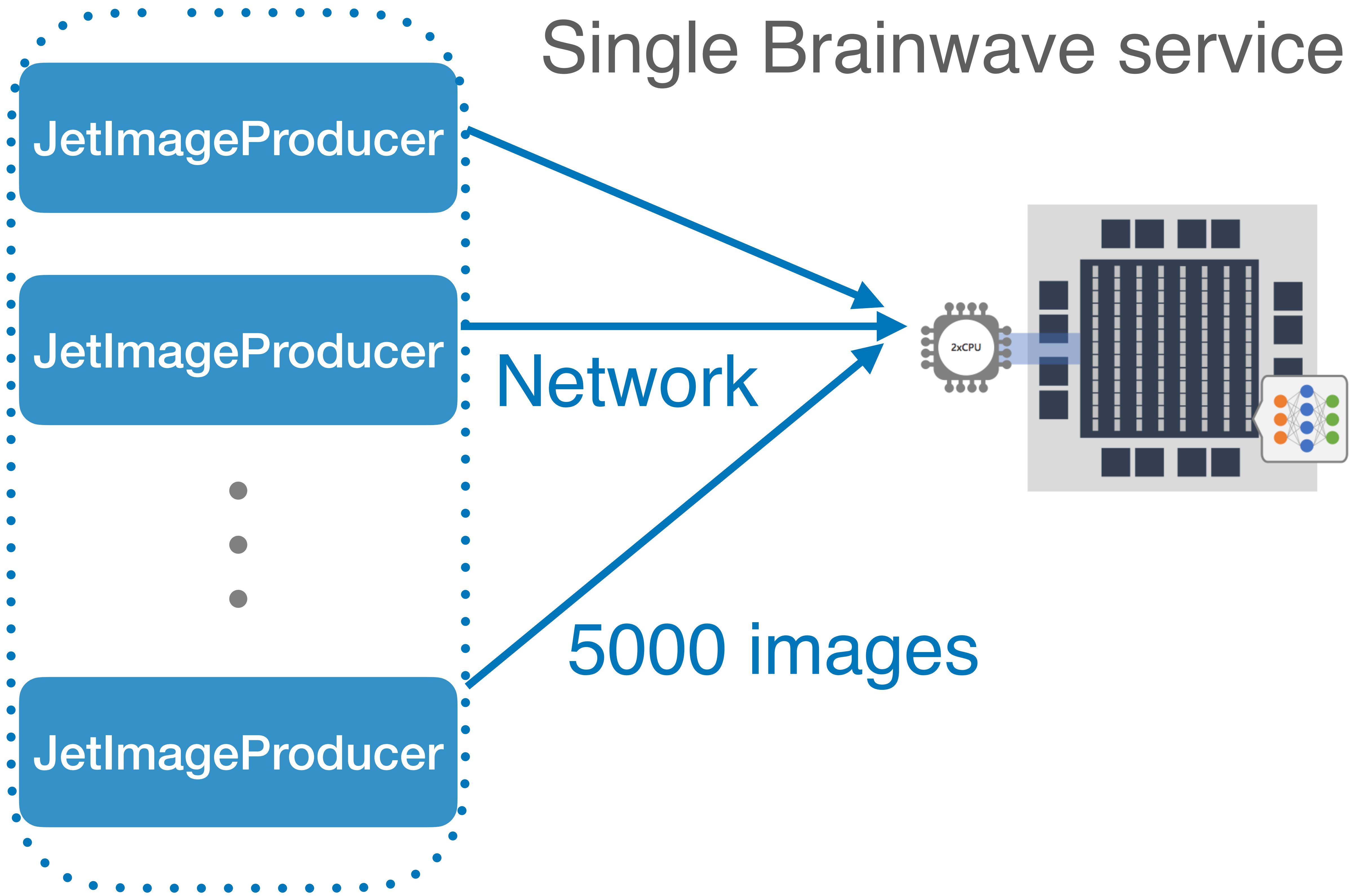
Services for Optimized Network Inference on Co-processors
(SONIC paper)

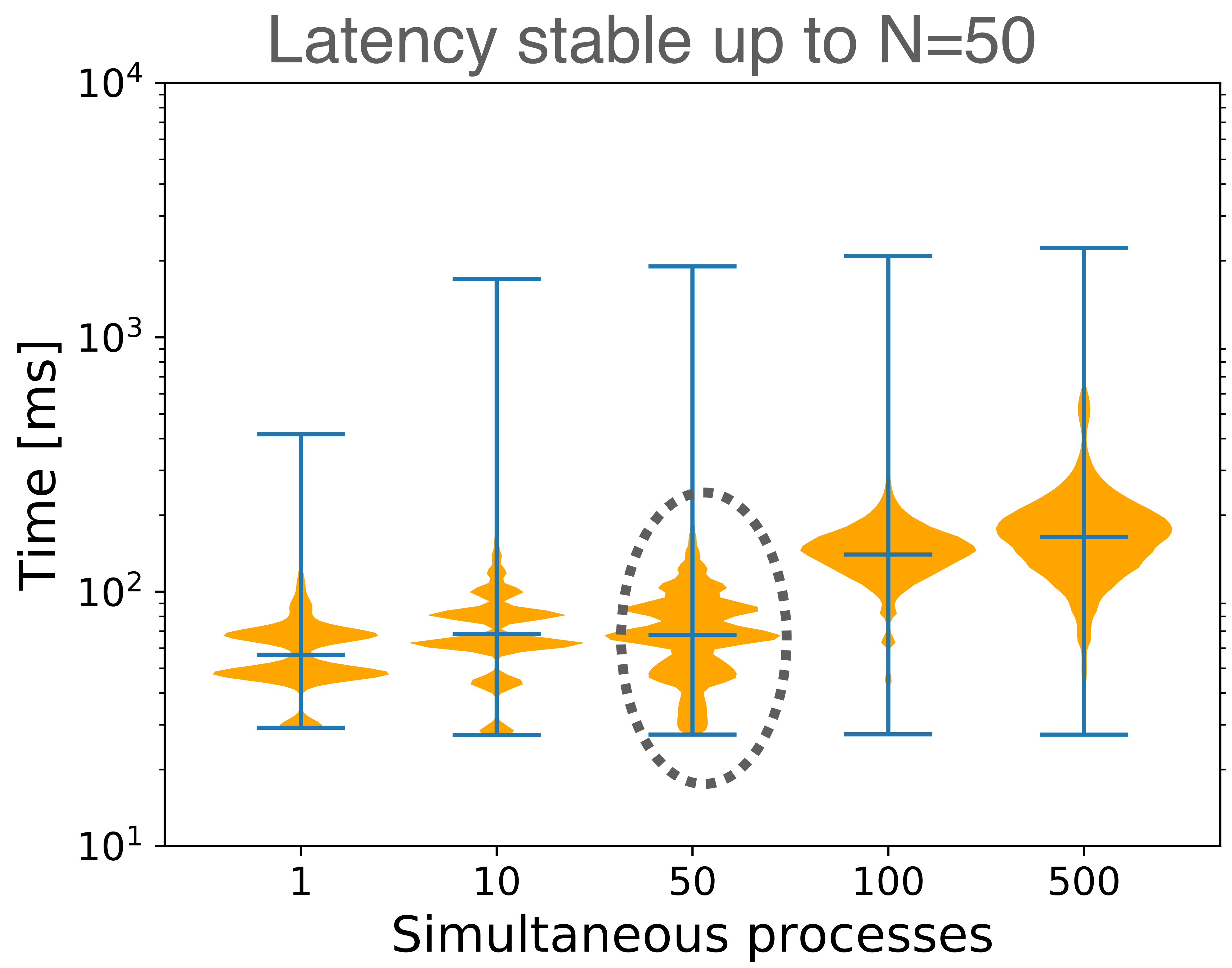






Performance: latency & throughput





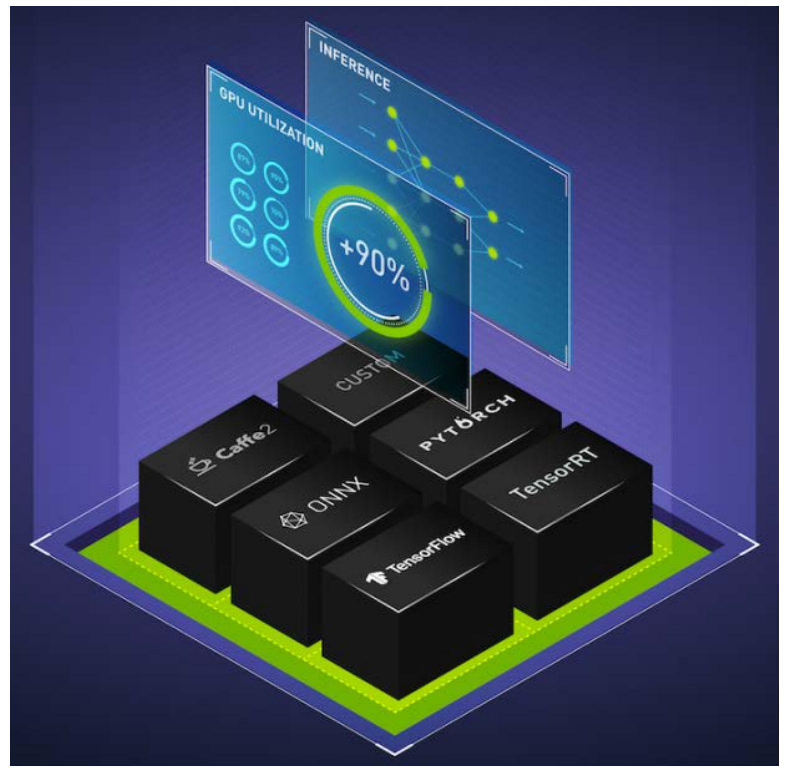
Latency: 10 ms (60 ms) for local (remote cloud) server, (10/100) faster than CPU-only

Max data throughout: 600-700 images/sec

SONIC: recent explorations

GPU-as-a-service
https://arxiv.org/abs/2007.10359

GPU-as-a-service for DUNE
https://arxiv.org/pdf/2009.04509.pdf



NVIDIA Triton

Hardware platforms

FPGA-as-a-service Toolkit

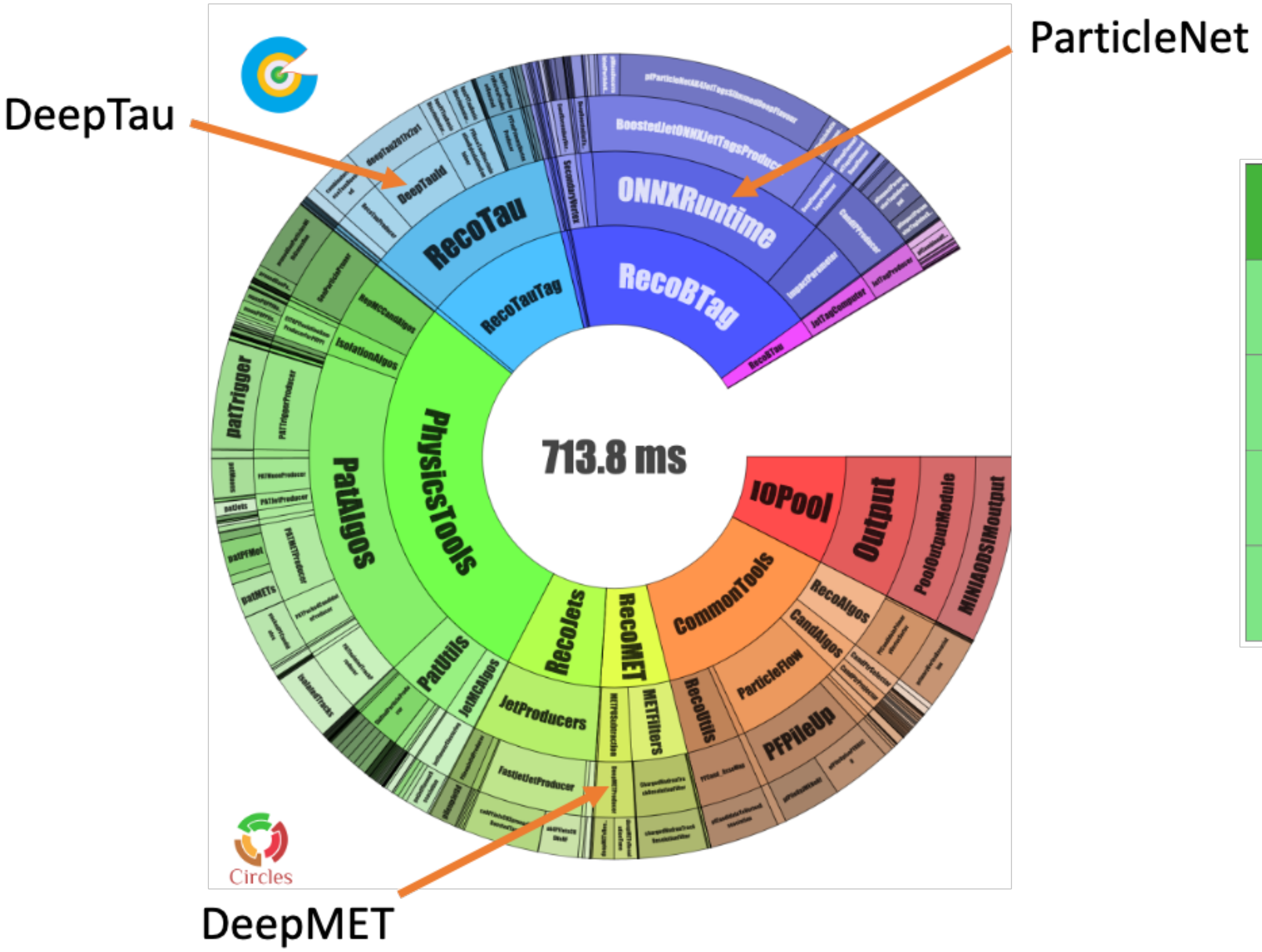
hls4ml

Open source tools: flexibility

Algorithm complexity

More benchmarks driven by use cases to test scaling for HLT/offline: 2k- 10M parameters

SONIC in CMS miniAOD

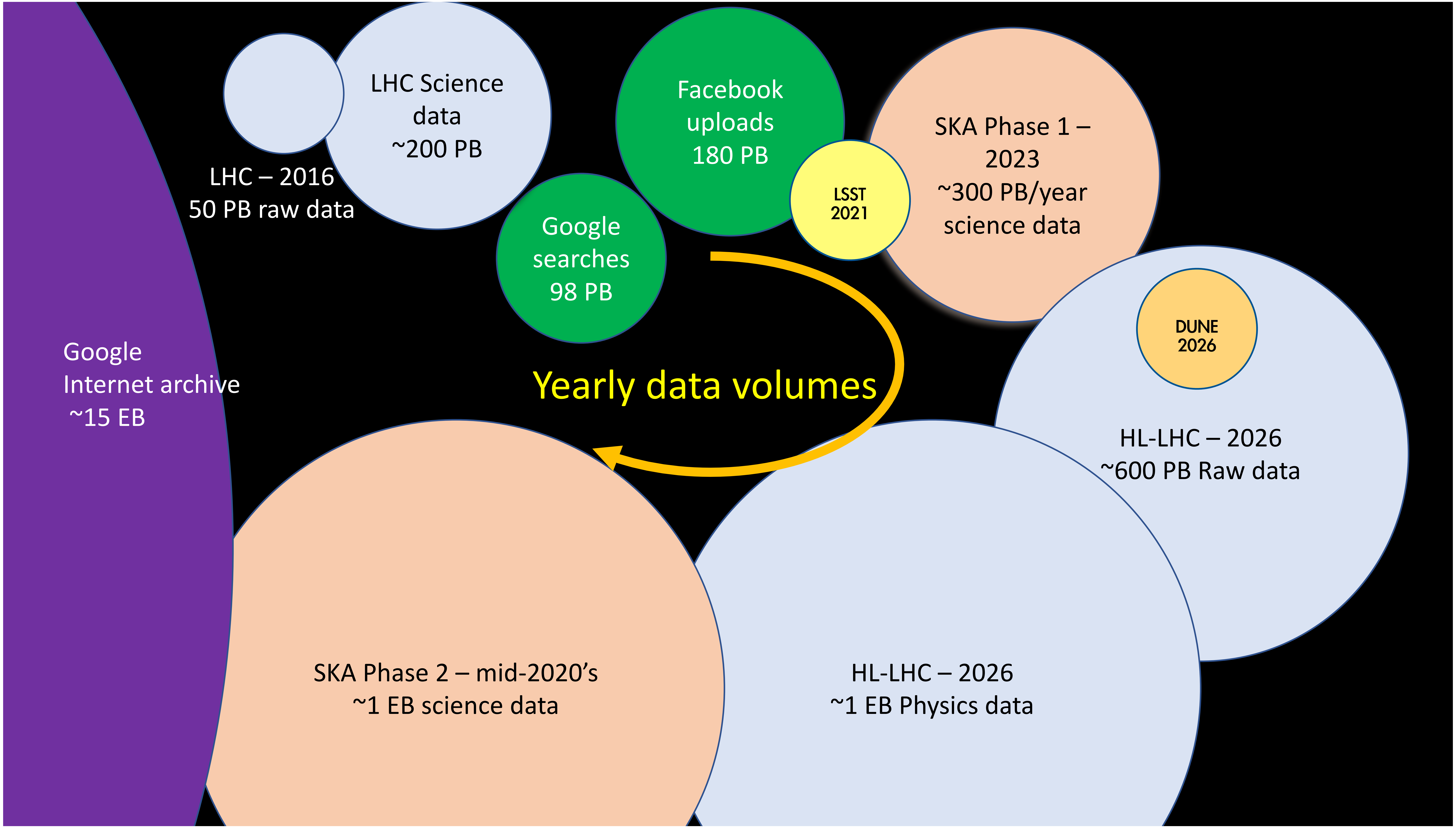


	Latency (ms)	Fraction (%)
Total	718.3	100
ParticleNet	90.3	12.7
DeepTau	24.6	3.4
DeepMET	12.8	1.8

Please see slides and interactive pie chart

- SONIC miniAOD workflow has been developed, machine learning algorithms offloaded to GPU servers.
- Testing at Purdue T2 with local GPUs/GPUs in google cloud.
- Infrastructure development in plan such as server management, authentication etc
- A HLT workflow has also been developed for non-ML algorithms (patatrk, tracking on GPUs)

Big Data Era


LHC Science data ~200 PB
LHC – 2016
50 PB raw data
Facebook uploads 180 PB
Google searches 98 PB
LSST 2021
SKA Phase 1 – 2023 ~300 PB/year science data
Google Internet archive ~15 EB
Yearly data volumes
DUNE 2026
HL-LHC – 2026 ~600 PB Raw data
SKA Phase 2 – mid-2020's ~1 EB science data
HL-LHC – 2026 ~1 EB Physics data

Fast Machine Learning Community

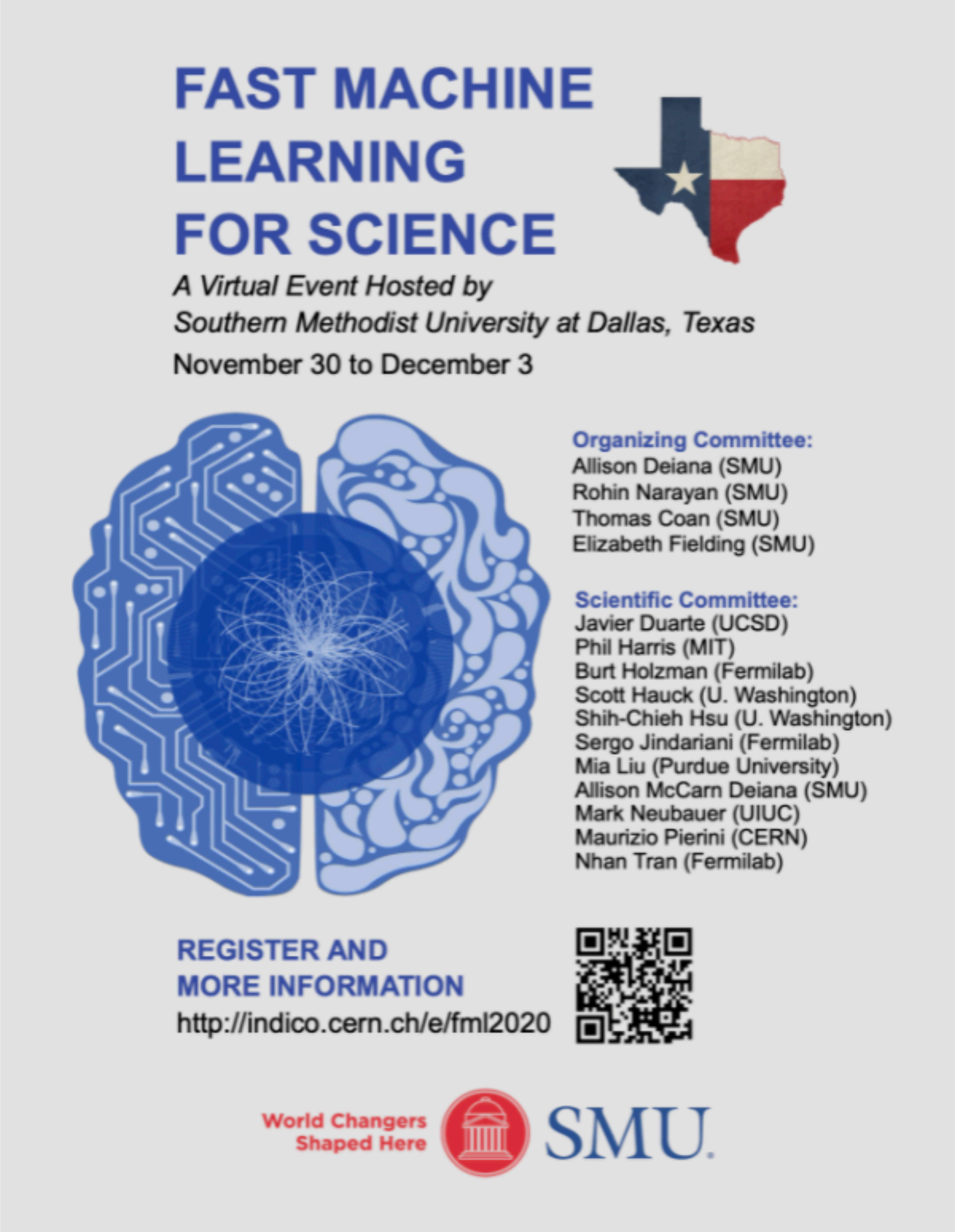
Community built upon hls4ml & sonic effort: monthly general meetings, alternating hls4ml & co-processor meetings.

Workshops at Fermilab/SMU (virtual)

Fruitful discussions on common challenges across science domains & interesting intersections with industry and other fields

: HEP, neutrino, astrophysics, plasma physics (fusion control), material science, Xilinx, Nvidia, Neuromophic compute.

White papers: 2019, 2020 submitted to frontier in big data.

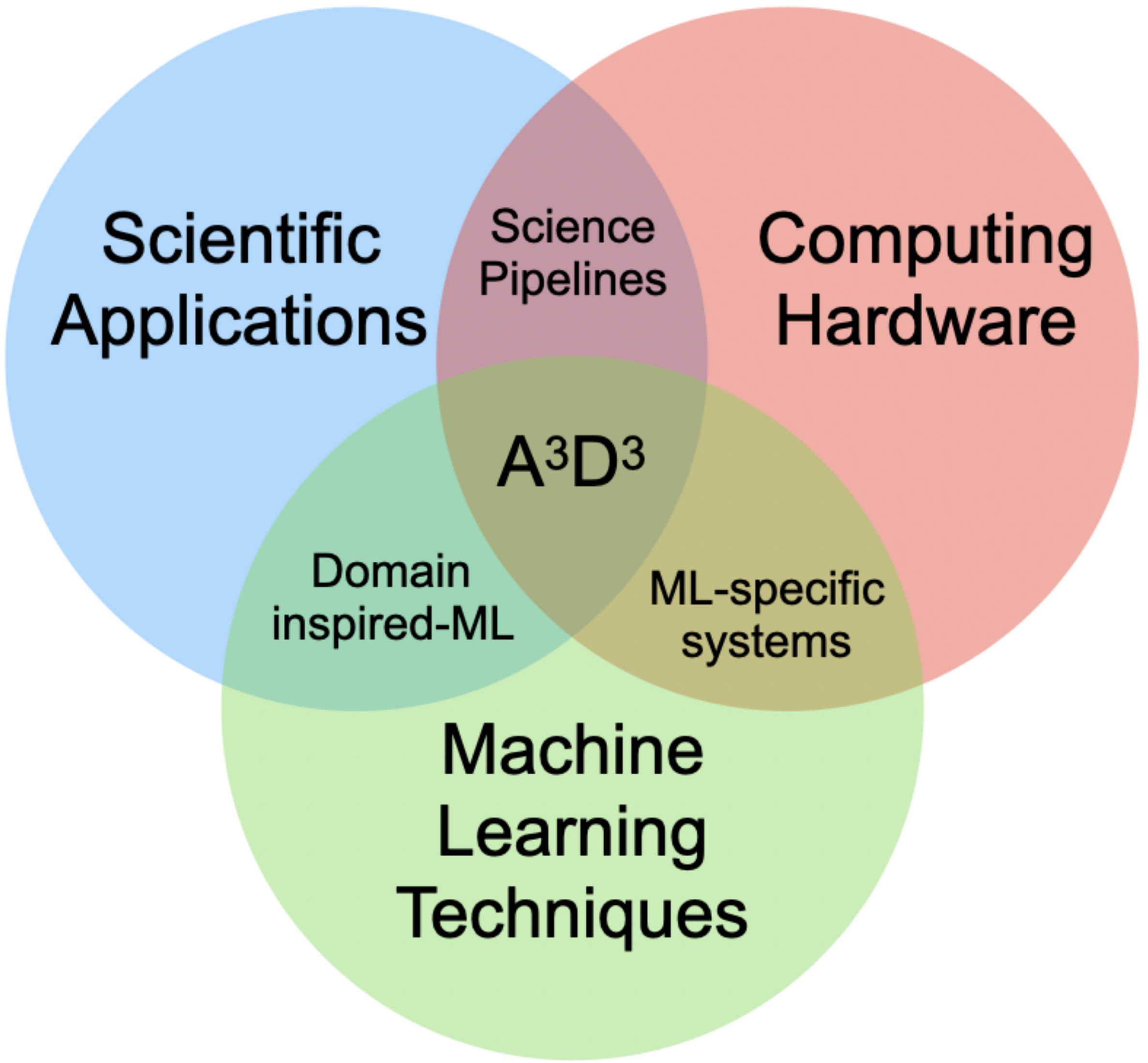


Next (mini) Fast ML workshop in Spring 2022

A3D3 Institute

Harnessing the data revolution Institute grant awarded by National Science Foundation (NSF)

Accelerated Artificial Intelligence For Data-Driven Discoveries.

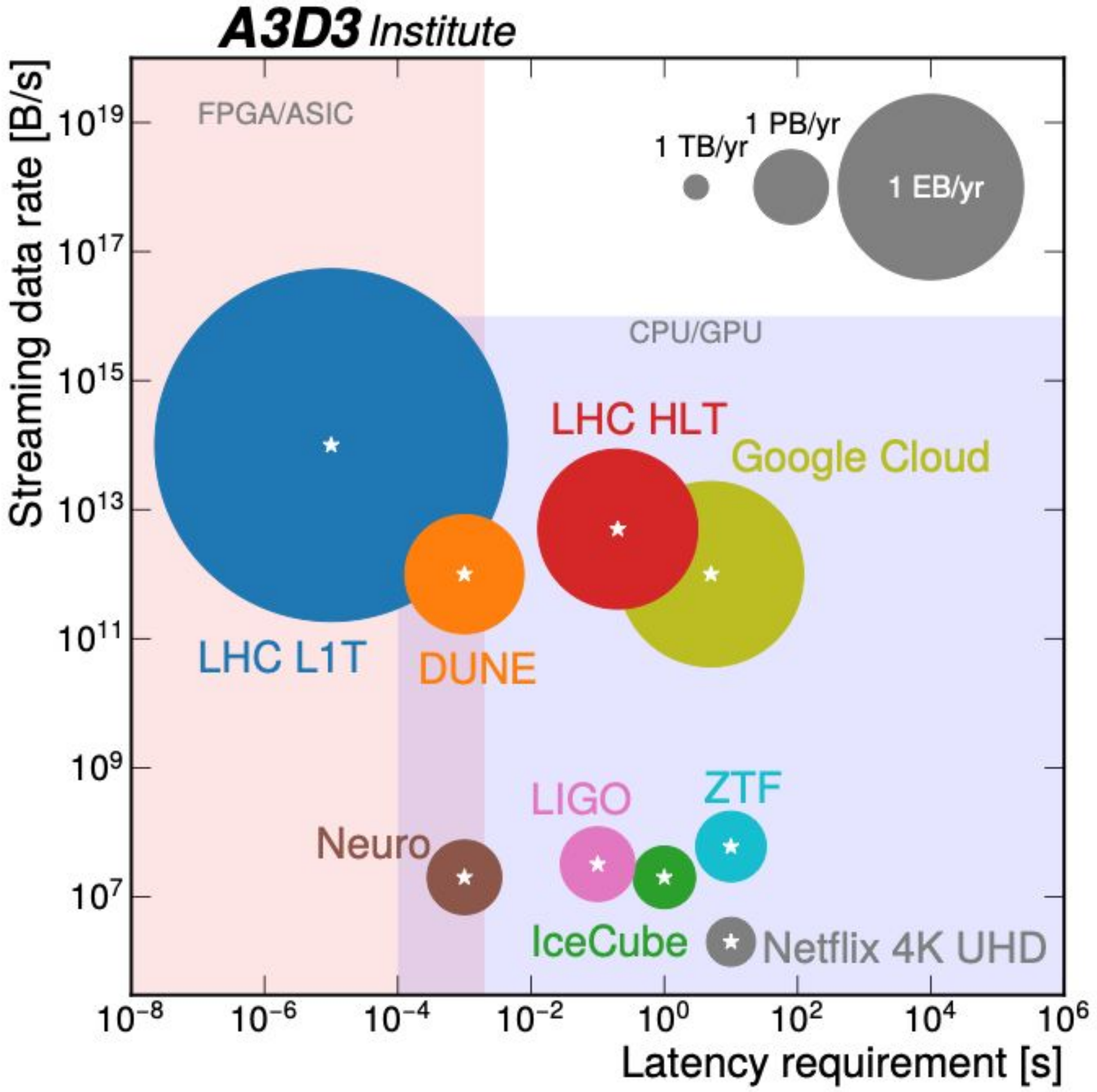


Accelerated AI Algorithms for Data-Driven Discovery

The challenge for domain scientists is that a broad range of expertise is required to arrive at full ML device implementations.

15 M grant, 9 Institutions. HEP, astrophysics, neural science, AI algorithm, Hardware acceleration.

Interface with frontier algorithm & engineering research (machine learning compiler)



Summary

Machine learning methods offer opportunities to significantly boost the discovery potential at the LHC (e.g $\tau \to 3\mu$).

Accelerated machine learning inference in online & offline processing.

User-friendly prototype tools for domain experts.

Multidisciplinary teams to realize optimal ML on targeted (e.g. CMS L1 trigger) or heterogeneous systems (e.g. CMS HLT & offline).

Look forward to the visions unfold in the next few years!

Adaptive algorithms and tools

Dark sector searches at SeaQuest @ Fermilab

CNNs, Graphs, RNNs, auto-encoders, binary/ternary

e.g. Lepton flavor violation: τ->3μ

Measuring muon EDM with frozen spin techniques.

As-a-service Computing Model

How many CPU can this GPU serve?

CPU-to-GPU ratio:

t_total_cpu - t_othercpu = t_ml

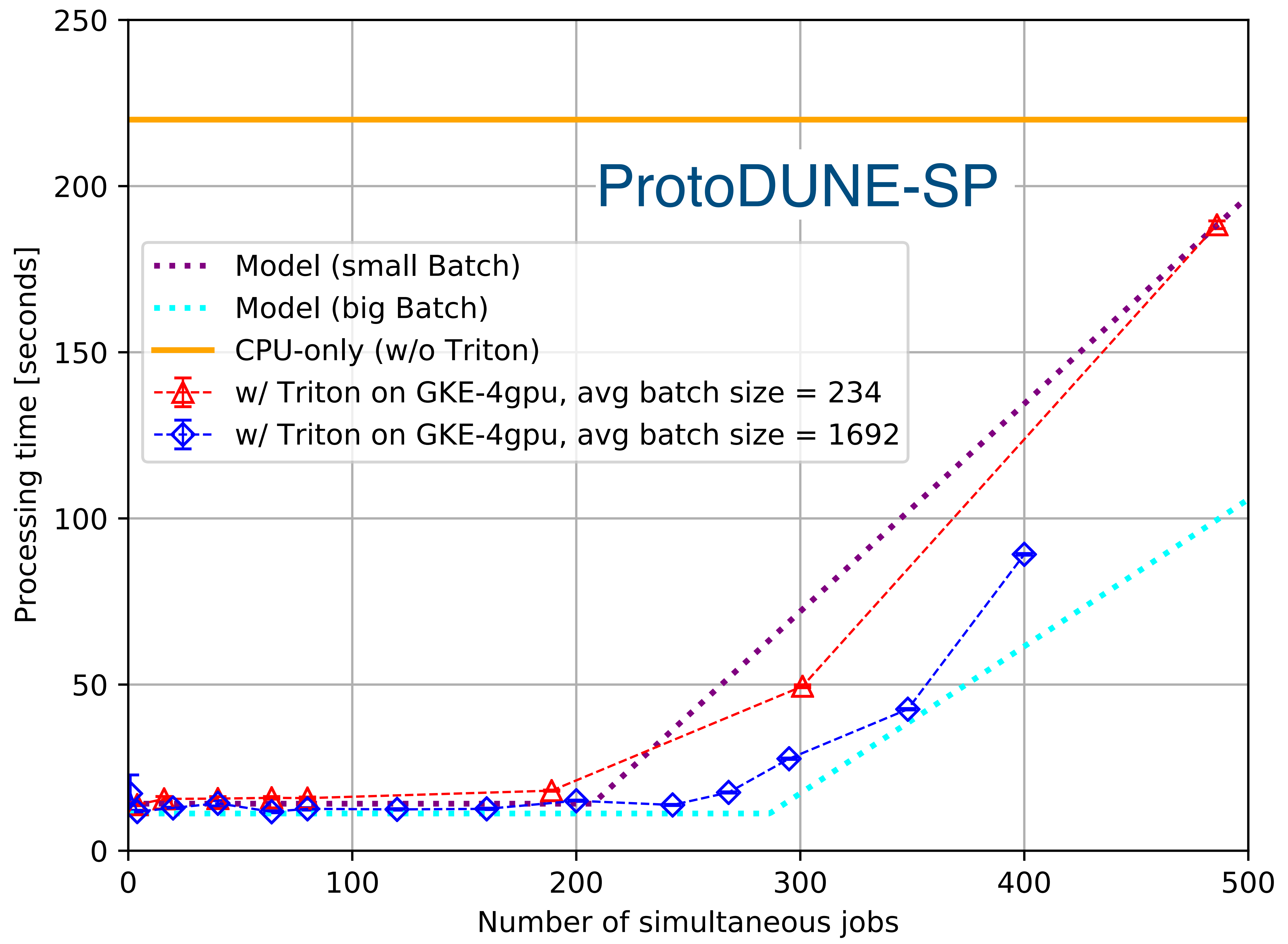
t_total_sonic - t_othercpu - t_sonic_cpu = t_transfer + t_scheduling + t_sonic_gpu

t_ml/ (t_sonic_cpu + t_transfer + t_scheduling + t_sonic_gpu) =

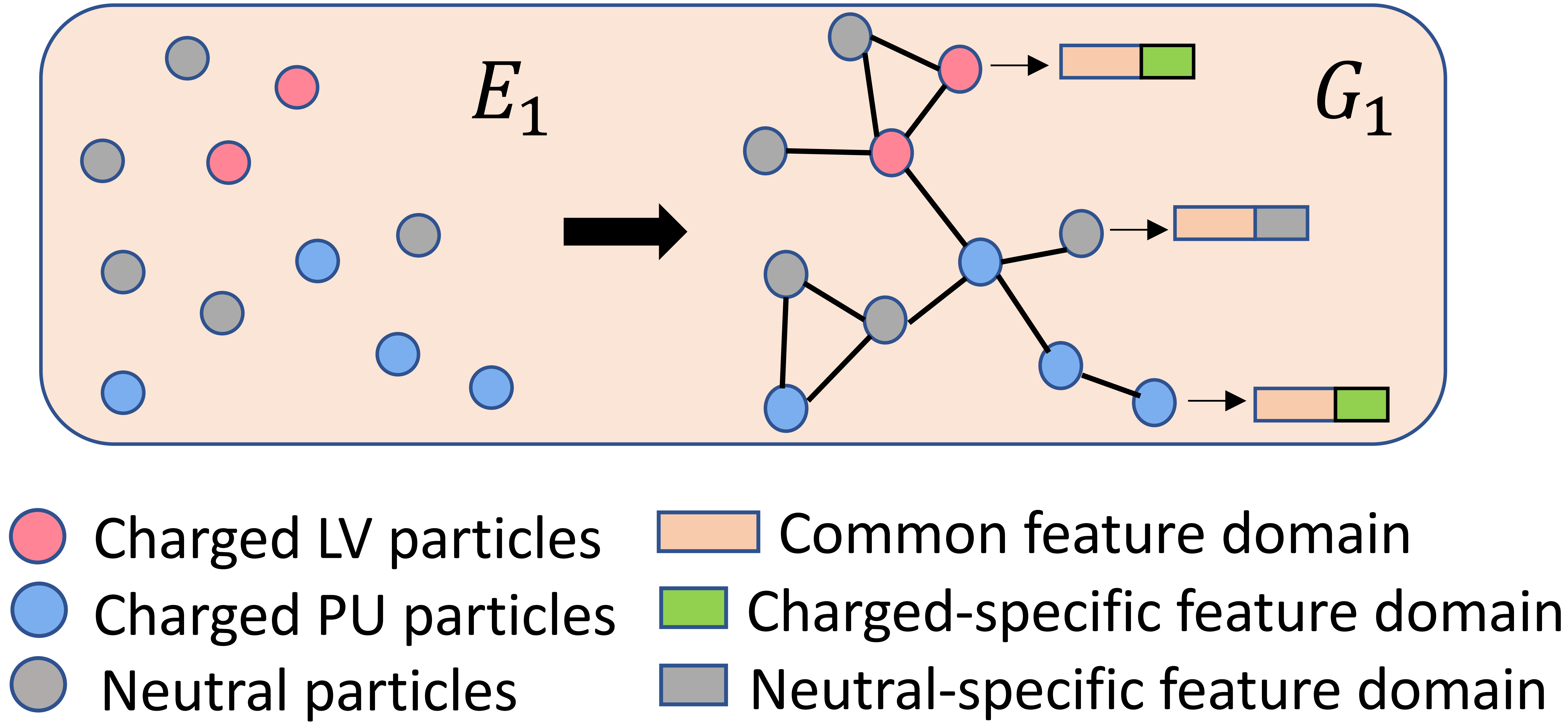
tsonic_cpu_part = t_sonic - t_transfer - t_scheduling

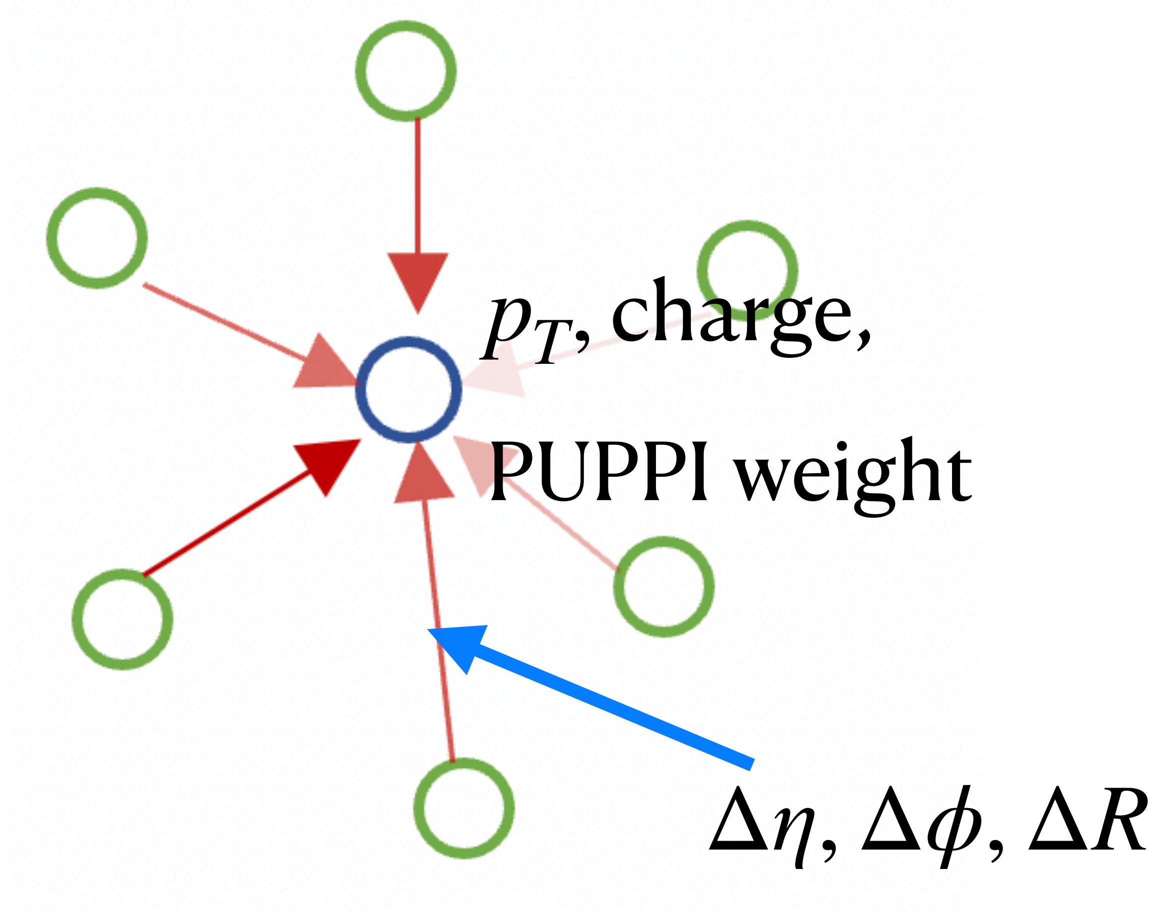
ratio = t_sonic_cpu_part/t_gpu

Passes this ratio, GPU saturates and average processing time increase.

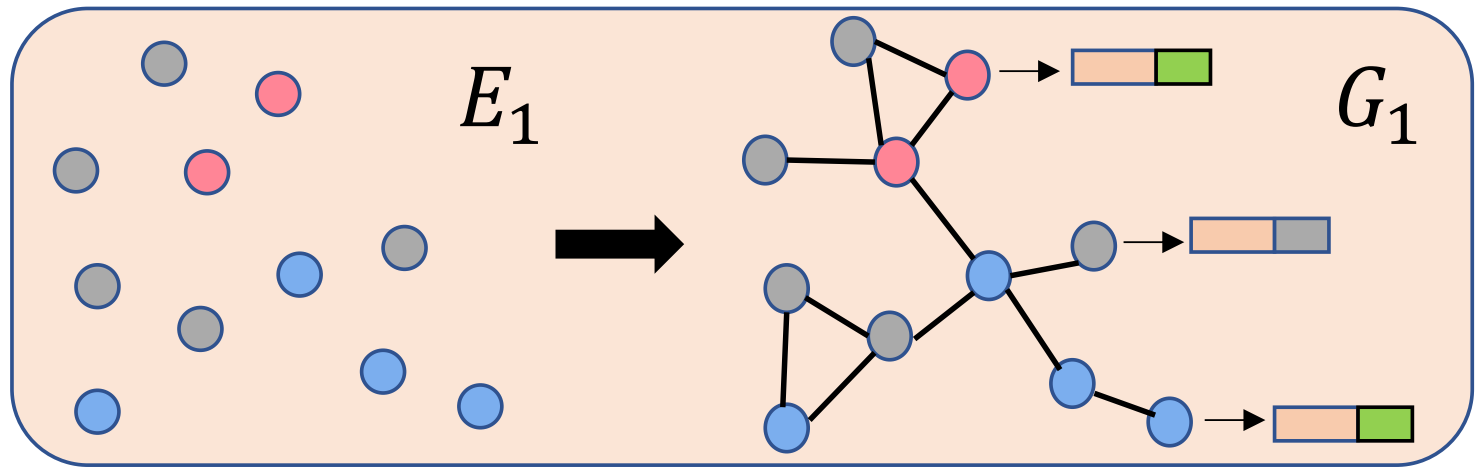


Semi-supervised Graph Puppi

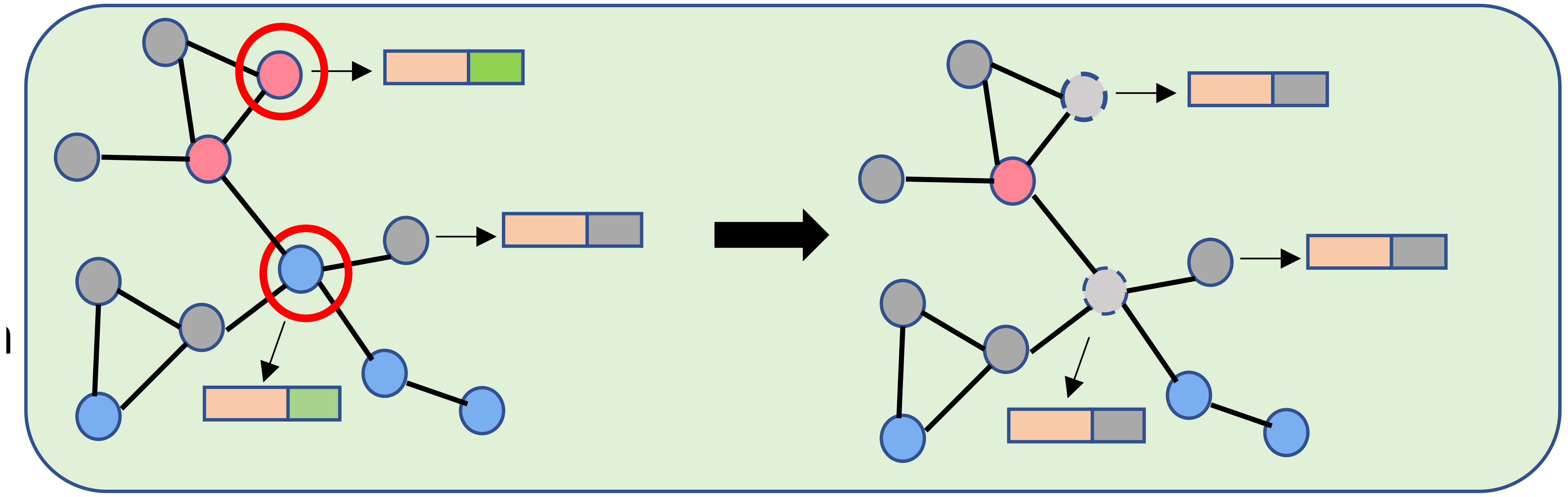


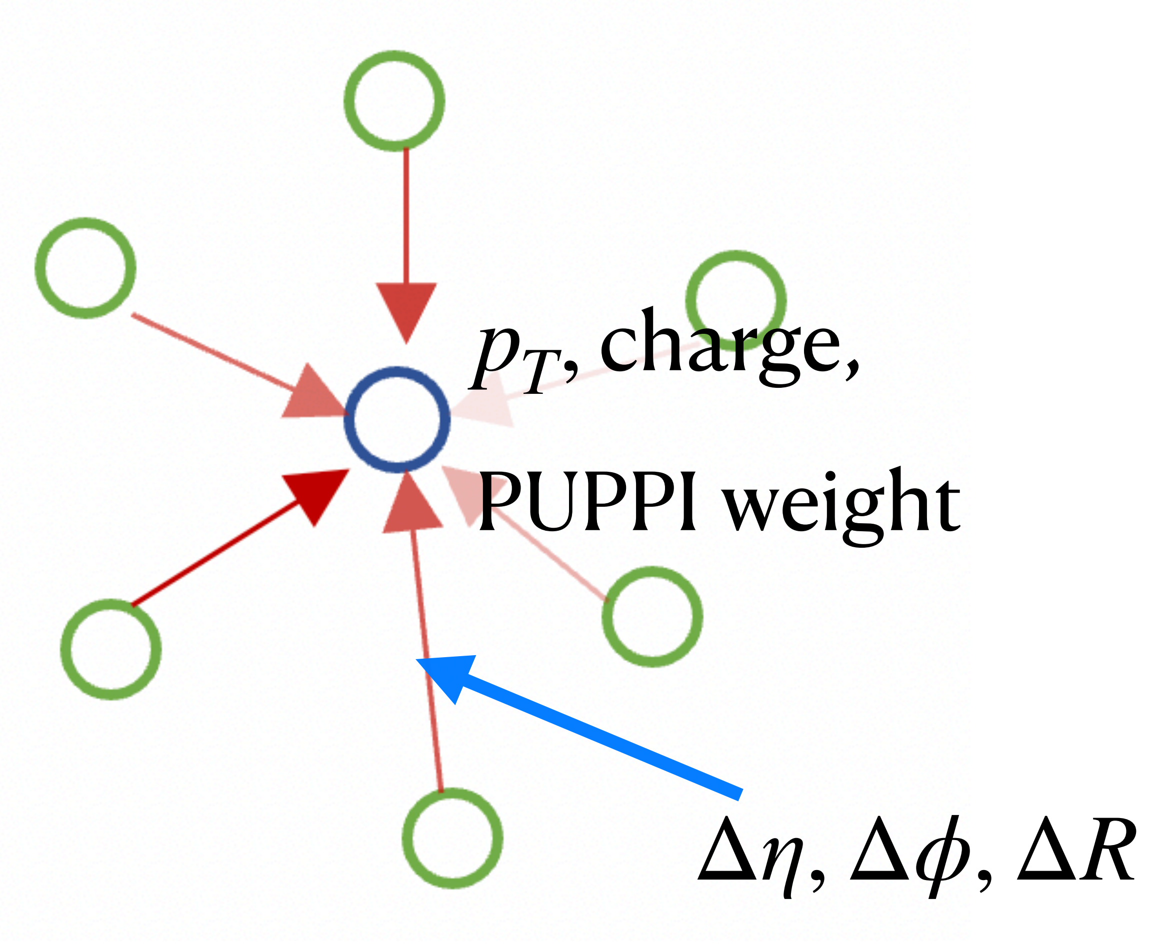


Semi-supervised Graph Puppi



(b). Randomly select charged LV/PU particles, and mask charged-specific feature domain for training

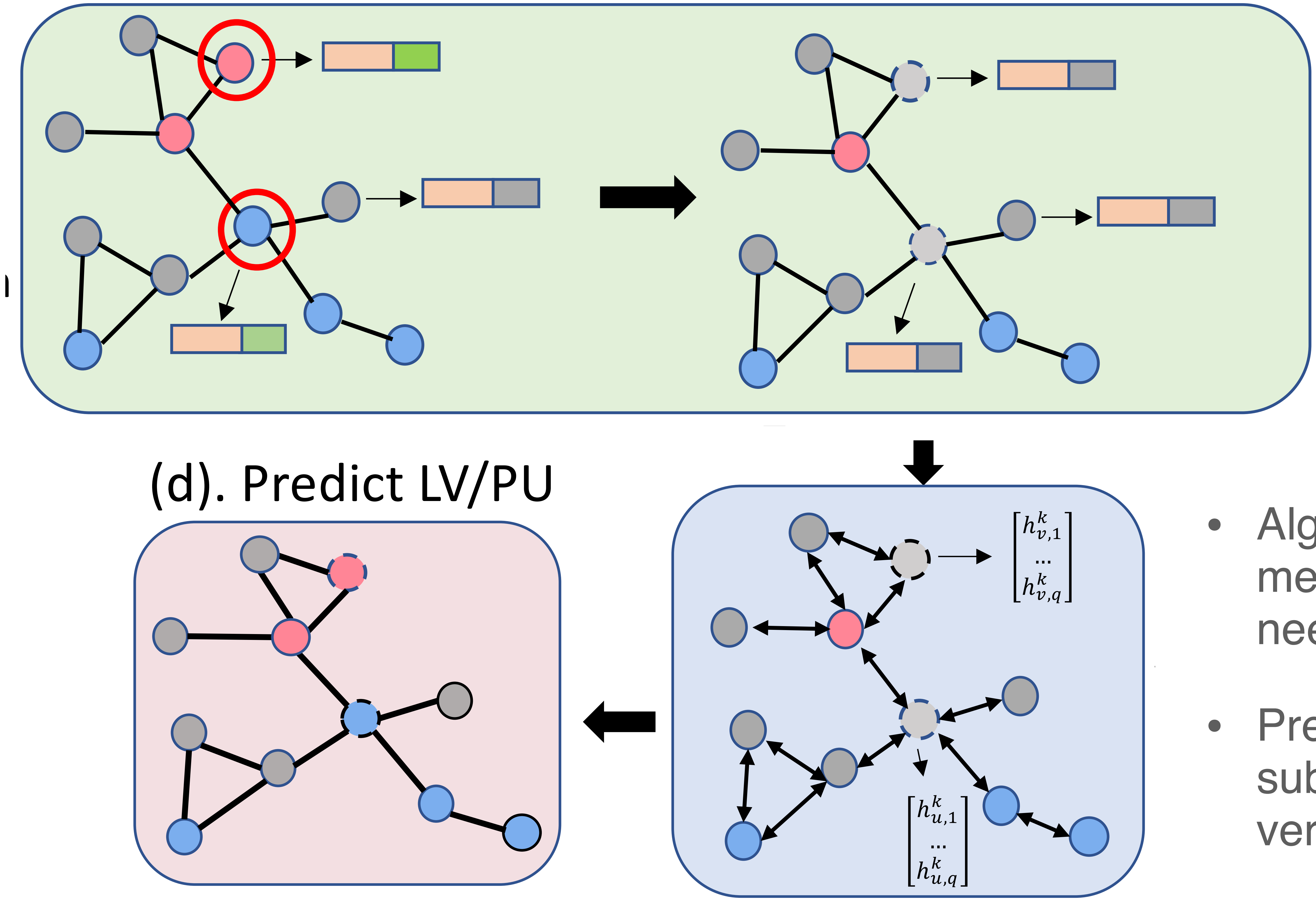


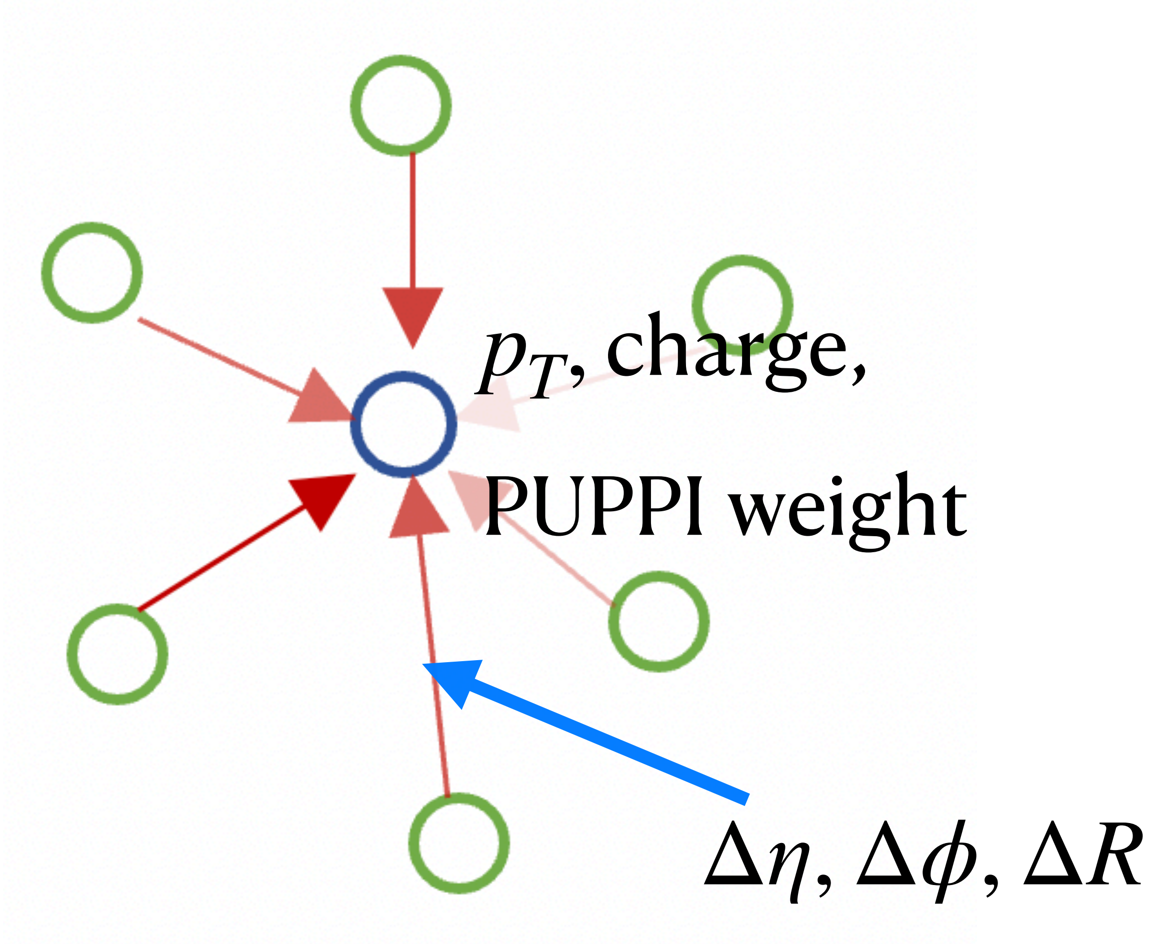


Charged LV particles
Charged PU particles
Neutral particles
Common feature domain
Charged-specific feature domain
Neutral-specific feature domain

Semi-supervised Graph Puppi

(b). Randomly select charged LV/PU particles, and mask charged-specific feature domain for training





- Algorithm outperforms Puppi, comparable to fully supervised method. Can be adapted to different pile up conditions. No need for tuning as the particle itself is represented as a node.
- Presented at BOOST 2021, Short version of the paper submitted to NeurIPS 2021 AI for Science Workshop. Long version paper targeting PRD in preparation.
- Next: Apply to CMS simulation & data. Neutral particle vertex association in for the forward region.

Summary and outlooks

LHC data let us probe the new physics scale at the LHC

SUSY searches and examine SM’s description of triboson processes

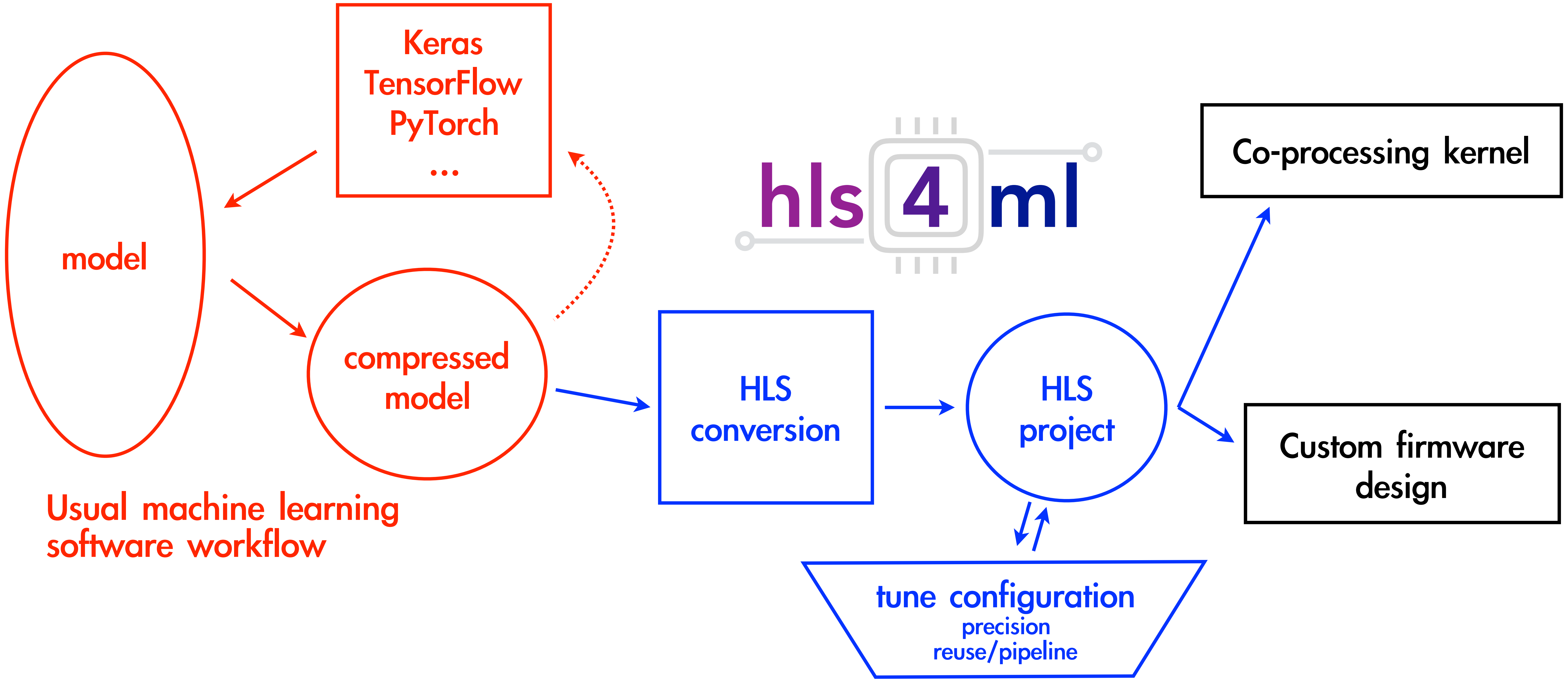
Accelerated discovery potential with ML

Fast machine learning inference for CMS data processing

Crucial in maximizing the HL-LHC physics potential

Look forward to continue with this exciting journey at UCR!

High-Level Synthesis 4 Machine learning



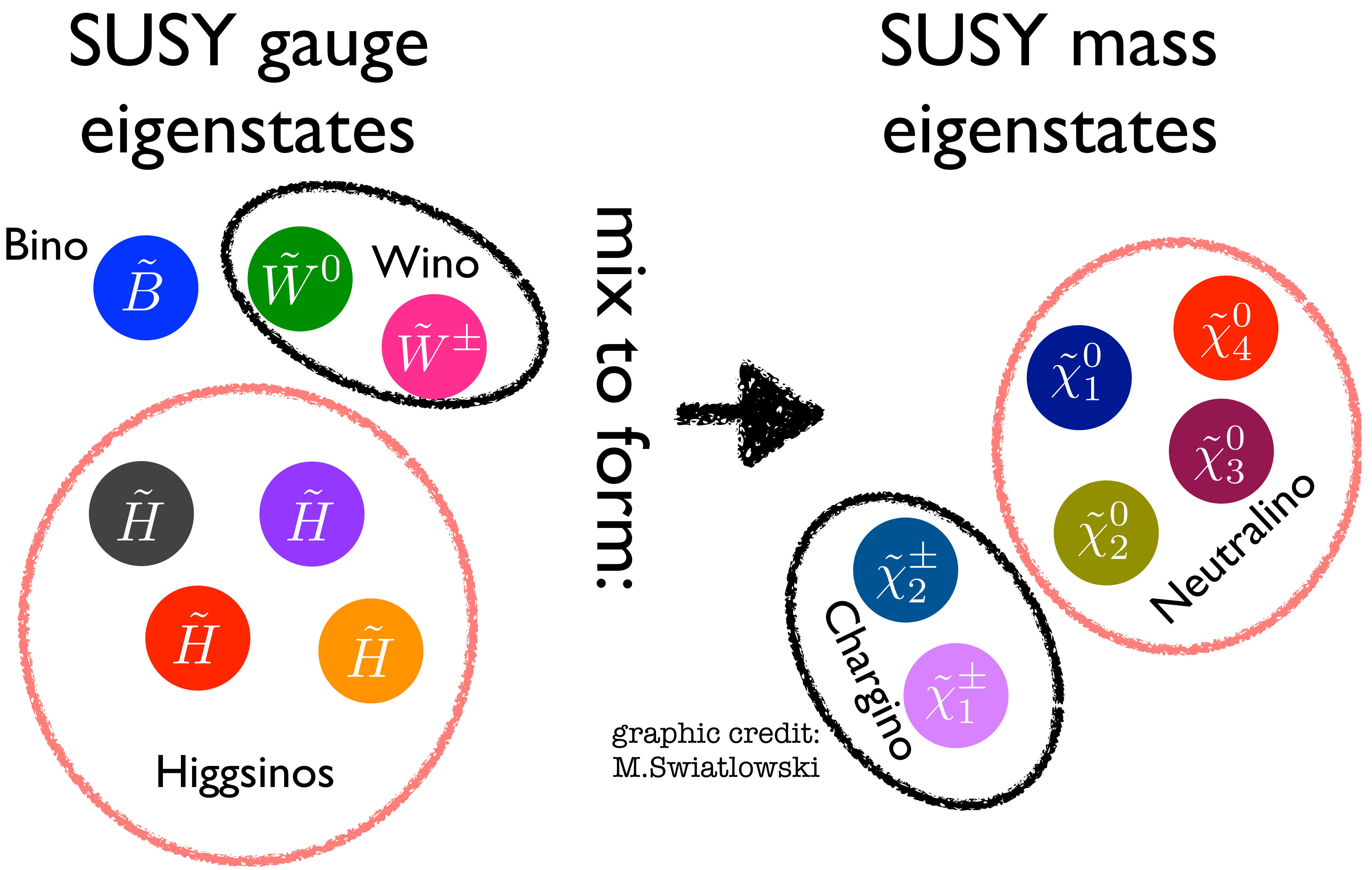
HLS - High Level Synthesis - compiler for C, C++, SystemC into FPGA IP cores

HLS 4 Machine learning :Prototype ML algorithms for FPGA WITHOUT Verilog/VHDL: firmware in a few hours

Electroweakinos 101

- SUSY partners of the SM electroweak sector:
  - U(1)-> Bino, SU(2)->Winos
  - Higgs-> Higgsinos
  - Leptons-> sleptons
- Could be light and accessible at the LHC

Important to search for them at the LHC!

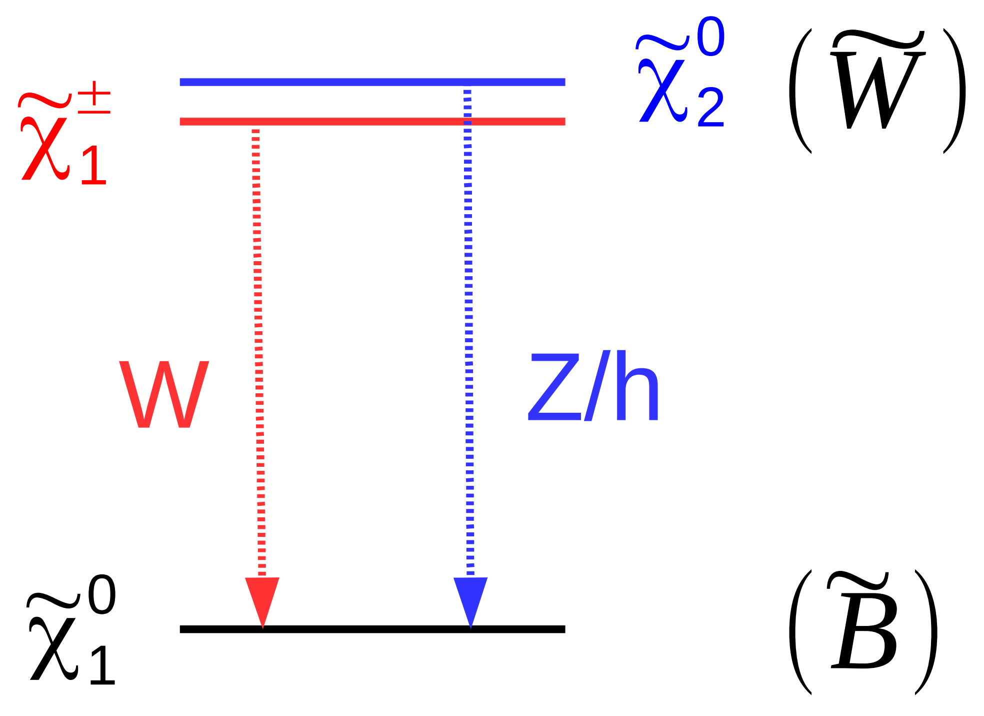


Typical mass spectrums of chargino-neutralinos

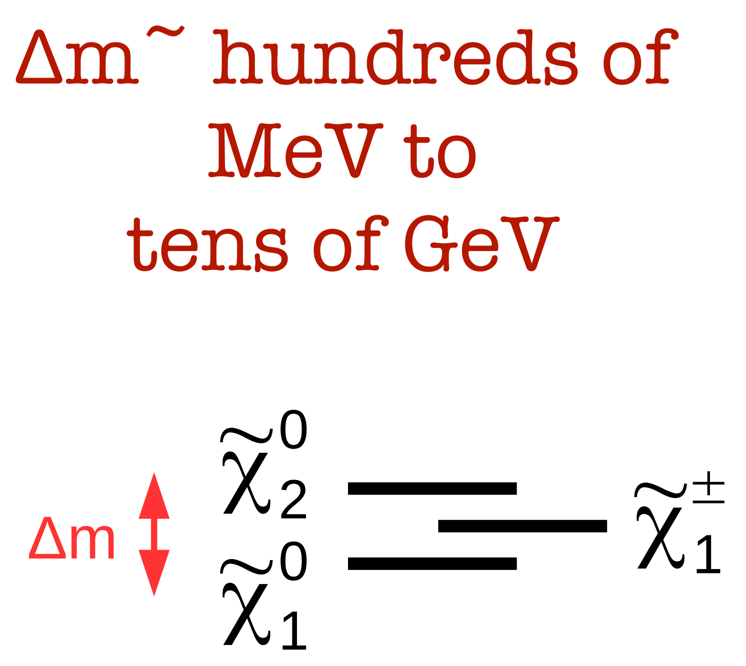
- Depending on the mass scales of Bino/ Winos/Higgsinos:

  lightest chargino/ neutralinos form different mass spectrums
- Two main mass spectrums explored at the LHC: Wino-like, Higgsino-like

- My focus: search for Wino-like production with WH events:
  - Larger cross section.
  - Loosely constrained in 8 TeV searches compared to WZ



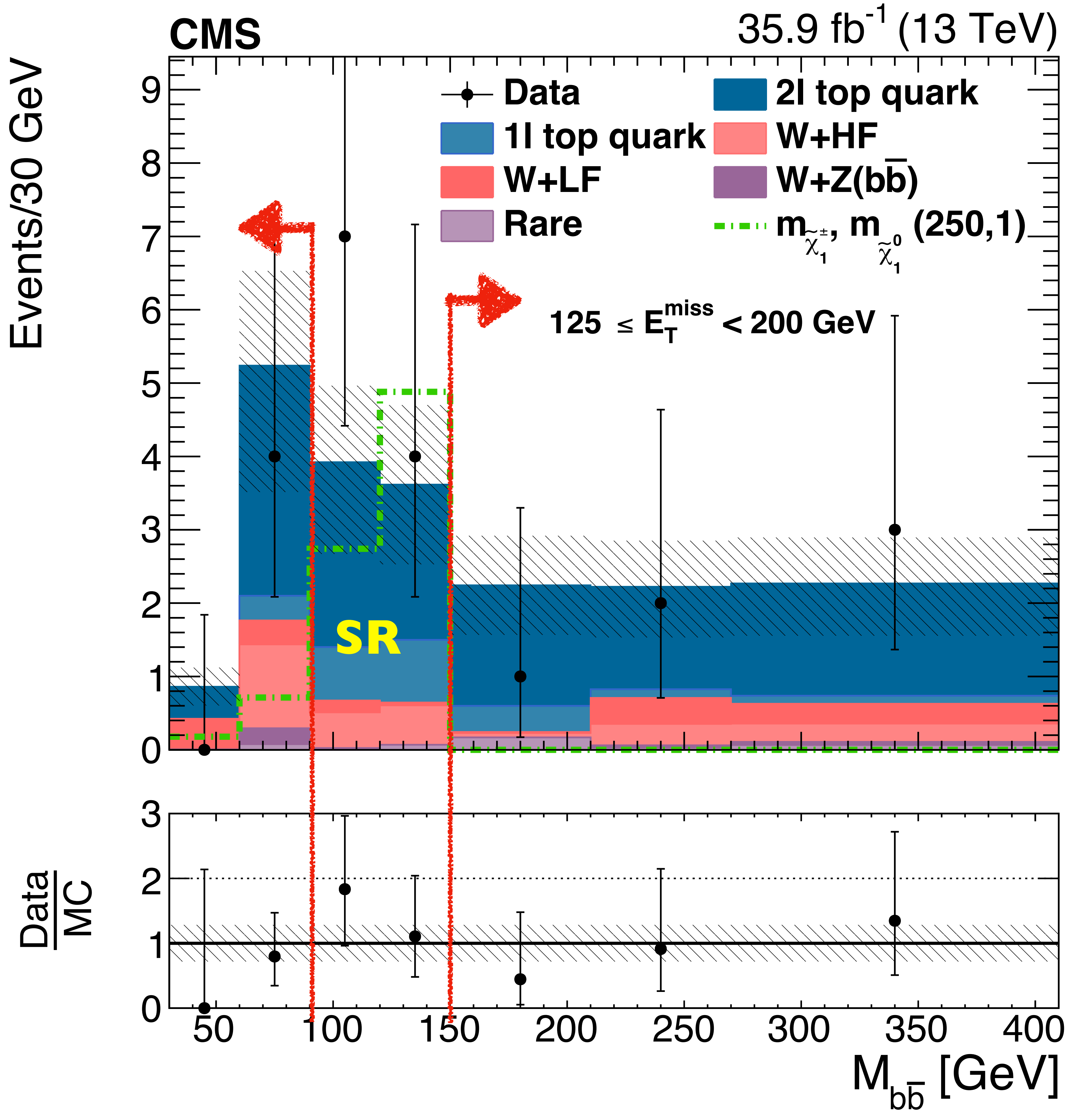
Wino-like
45 fb*



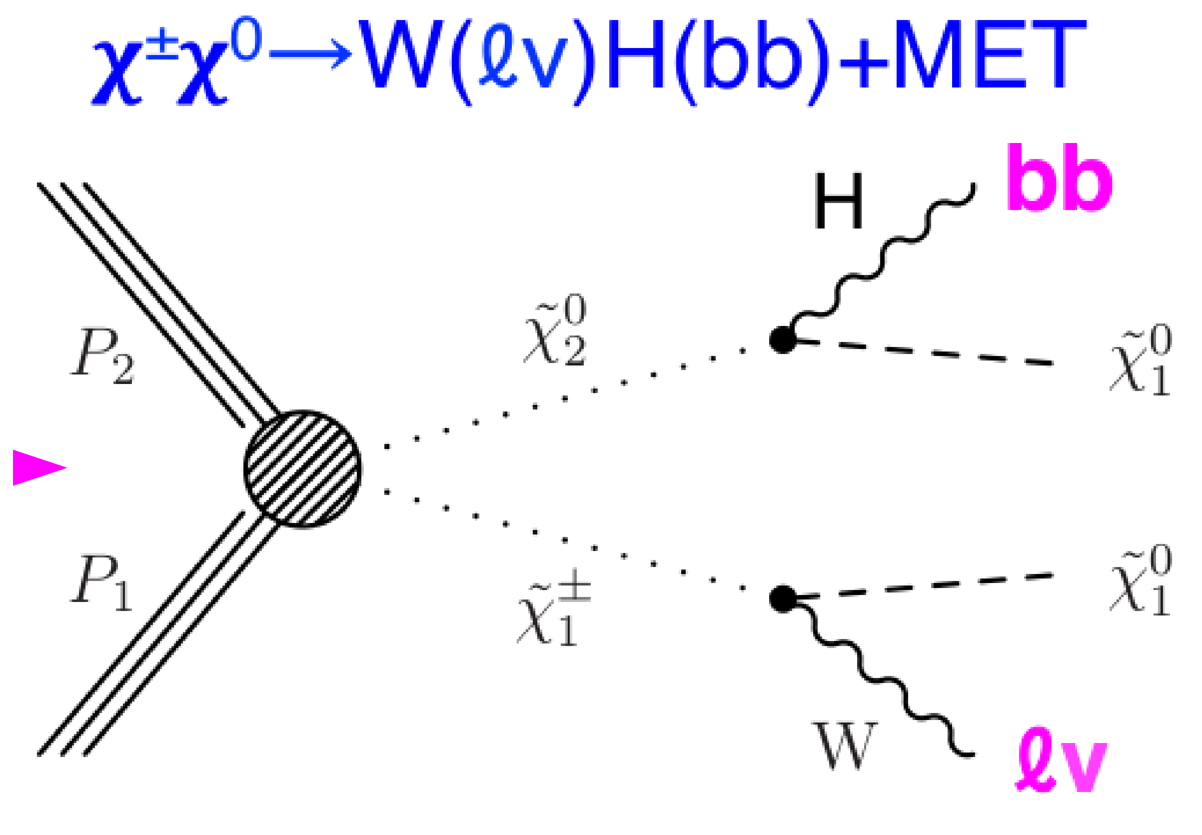
Higgsino-like
11 fb*

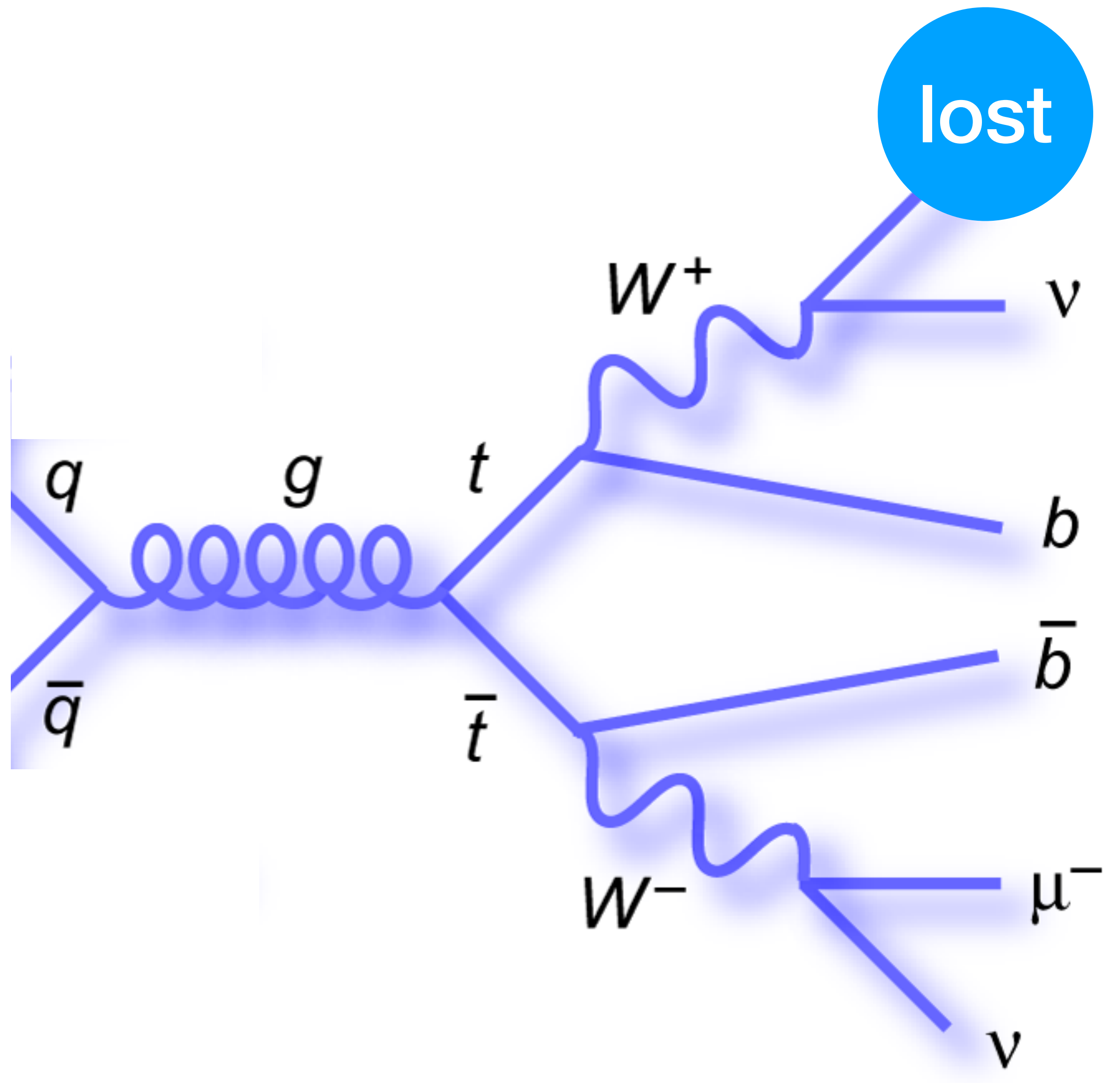
*Cross-sections for 500 GeV sparticles @ 13 TeV($\tilde{\chi}_2^0\,\tilde{\chi}_1^{\pm}$ only)

Search for electroweakinos with WH events



10.1007/JHEP11(2017)029

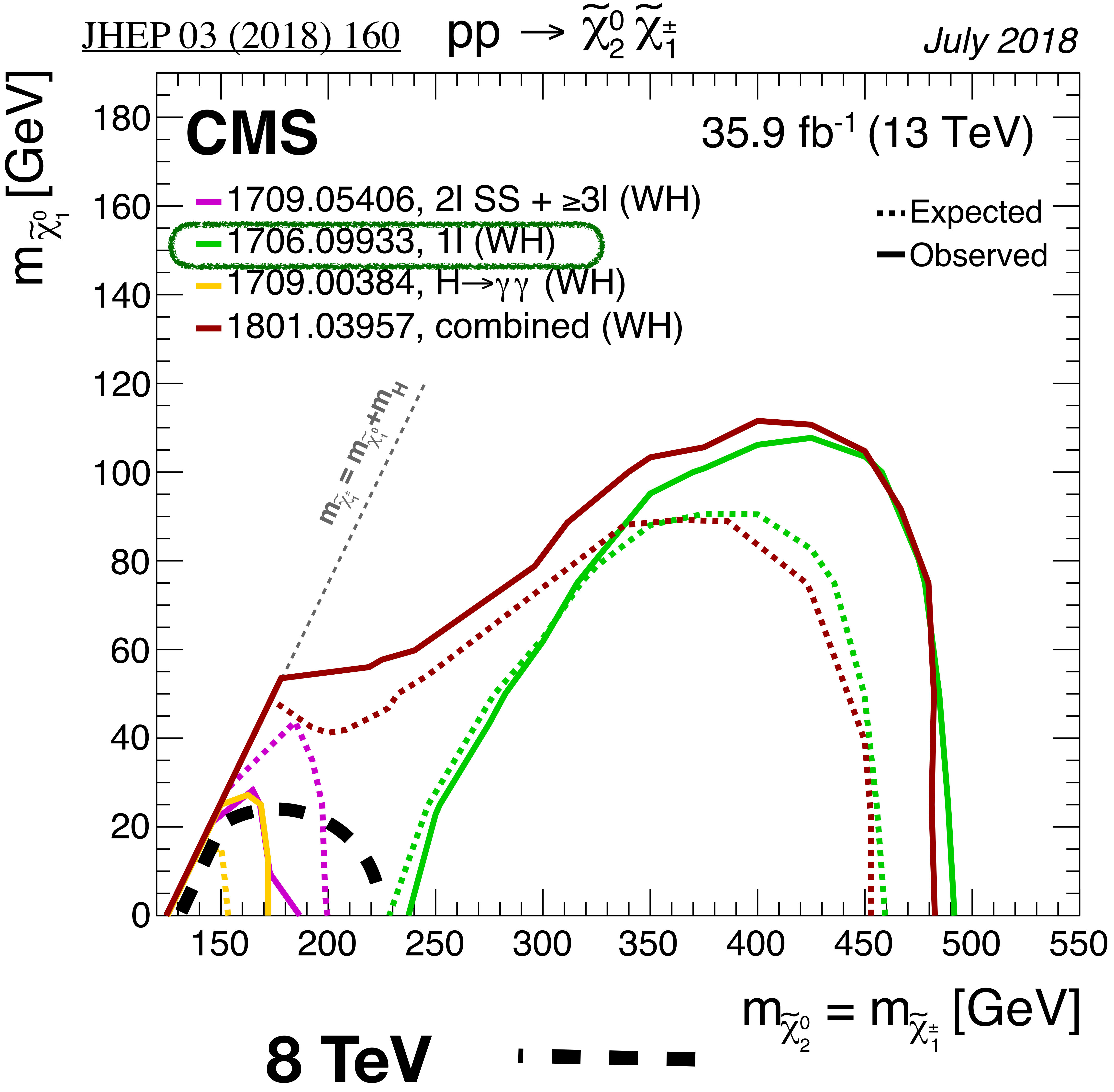




- 1 lepton (e/μ)+bb + large missing transverse energy:
  - leptons: trigger and handle against backgrounds
  - Higgs-> bb (60%). Mass peak in the SM kinematic tails. ttbar-> 2L background directly controlled in the mass sideband.
- First result to probe chargino mass to 500 GeV in WH decay

WH(lvbb) + MET: pushing Wino limits

JHEP 03 (2018) 160



- Probes chargino mass up to 500 GeV in the WH topology
- 300 GeV improvement wrt 8 TeV reach
- Dominates the sensitivity in the bulk.

Same sign selection

Table 1: Event selection criteria for the *SS category*, which contains events with two same-sign leptons and at least two hadronic jets

Variable	$e^{\pm}e^{\pm}$	$e^{\pm}\mu^{\pm}$	$\mu^{\pm}\mu^{\pm}$
Signal leptons	exactly 2 tight equally-charged leptons with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$		
Additional leptons	no additional rejection lepton		
Isolated tracks	no (additional) isolated tracks		
Jets	$\geq$ 2jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$, $\lvert\eta\rvert < 2.5$		
b-tagged jets	no b-tagged jet		
Dijet mass (closest ΔR)	$65 < M_{\mathrm{jj}} < 95\,\mathrm{GeV}$ (M_{jj}-in) **OR** $\lvert M_{\mathrm{jj}} - 80\,\mathrm{GeV}\rvert \geq 15\,\mathrm{GeV}$ (M_{jj}-out)		
Dijet mass (leading jets)	$< 400\,\mathrm{GeV}$		
$\Delta\eta$ of two leading jets	< 1.5		
$p_{\mathrm{T}}^{\mathrm{miss}}$	$> 60\,\mathrm{GeV}$		$> 60\,\mathrm{GeV}$ if M_{jj}-out
$M_{\ell\ell}$	$> 40\,\mathrm{GeV}$	$> 30\,\mathrm{GeV}$	$> 40\,\mathrm{GeV}$
$M_{\ell\ell}$	$\lvert M_{\ell\ell} - M_{\mathrm{Z}}\rvert > 10\,\mathrm{GeV}$	—	
$M_{\mathrm{T}}^{\mathrm{max}}$	—	$> 90\,\mathrm{GeV}$	—

Three leptons

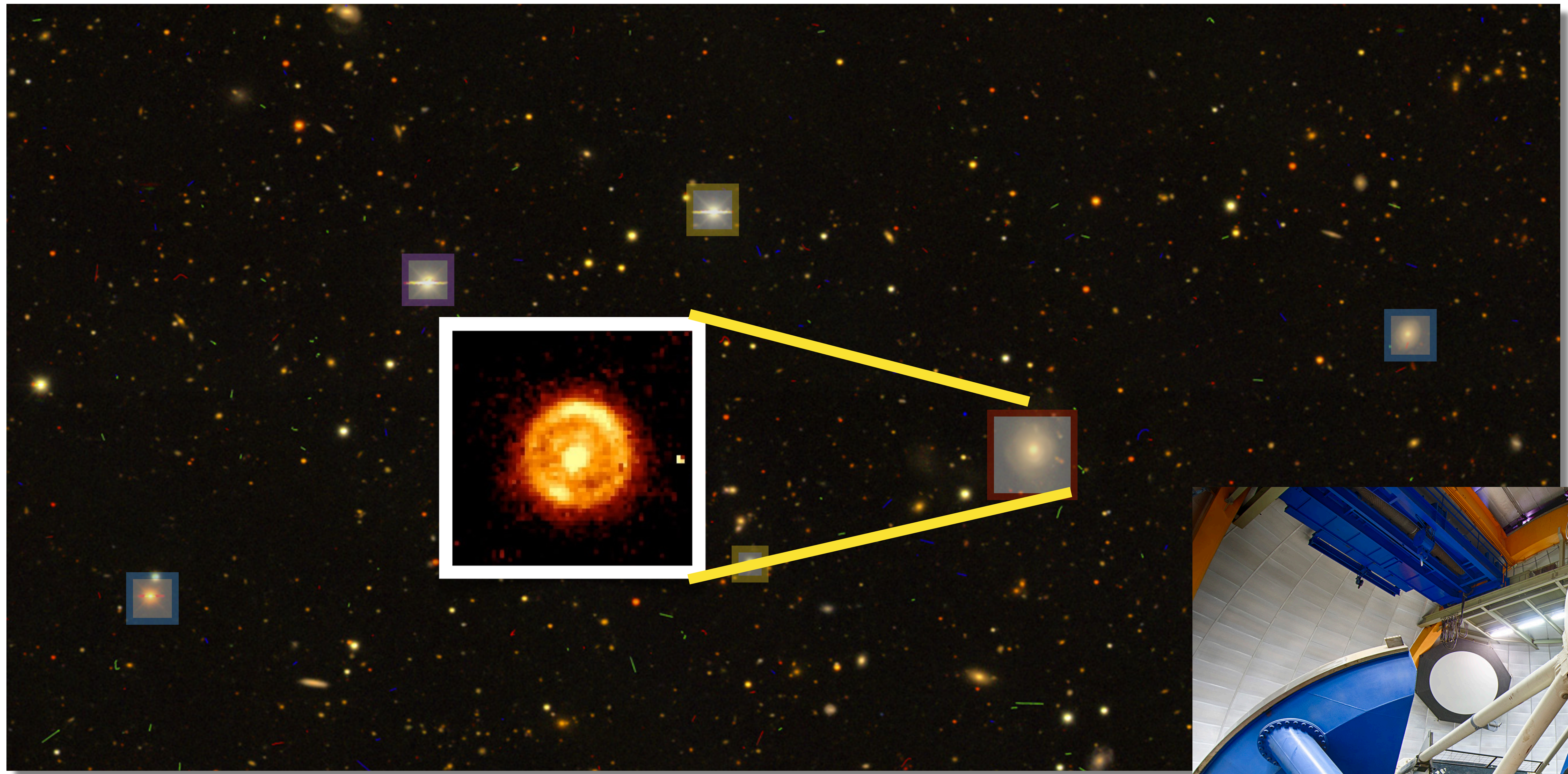
Table 2: Event selection criteria for the 3ℓ *category*, which contains events with exactly three leptons

Variable	0 SFOS	1 SFOS	2 SFOS
Signal leptons	exactly 3 tight charged leptons with $p_{\mathrm{T}} > 25/20/20\,\mathrm{GeV}$ and charge sum = ± 1e		
Additional leptons	no additional rejection lepton		
Jets	$\leq$ 1jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$, $\lvert\eta\rvert < 5$		
b-tagged jets	no b-tagged jet		
$p_{\mathrm{T}}(\ell\ell\ell)$	—	$> 60\,\mathrm{GeV}$	
$\Delta\phi\left(\vec{p}_{\mathrm{T}}(\ell\ell\ell), \vec{p}_{\mathrm{T}}^{\mathrm{miss}}\right)$	> 2.5		
$p_{\mathrm{T}}^{\mathrm{miss}}$	$> 30\,\mathrm{GeV}$	$> 45\,\mathrm{GeV}$	$> 55\,\mathrm{GeV}$
$M_{\mathrm{T}}^{\mathrm{max}}$	$> 90\,\mathrm{GeV}$	—	
$M_{\mathrm{T}}^{\mathrm{3rd}}$	—	$> 90\,\mathrm{GeV}$	—
SF lepton mass	$> 20\,\mathrm{GeV}$	—	
Di-electron mass	$\lvert M_{\mathrm{ee}} - M_{\mathrm{Z}}\rvert > 15\,\mathrm{GeV}$	—	
M_{SFOS}	—	$\lvert M_{\mathrm{SFOS}} - M_{\mathrm{Z}}\rvert > 20\,\mathrm{GeV}$ and $M_{\mathrm{SFOS}} > 20\,\mathrm{GeV}$	
$M_{\ell\ell\ell}$	$\lvert M_{\ell\ell\ell} - M_{\mathrm{Z}}\rvert > 10\,\mathrm{GeV}$		

Fermilab

M.LIU

ML in the Sky: it's full of stars

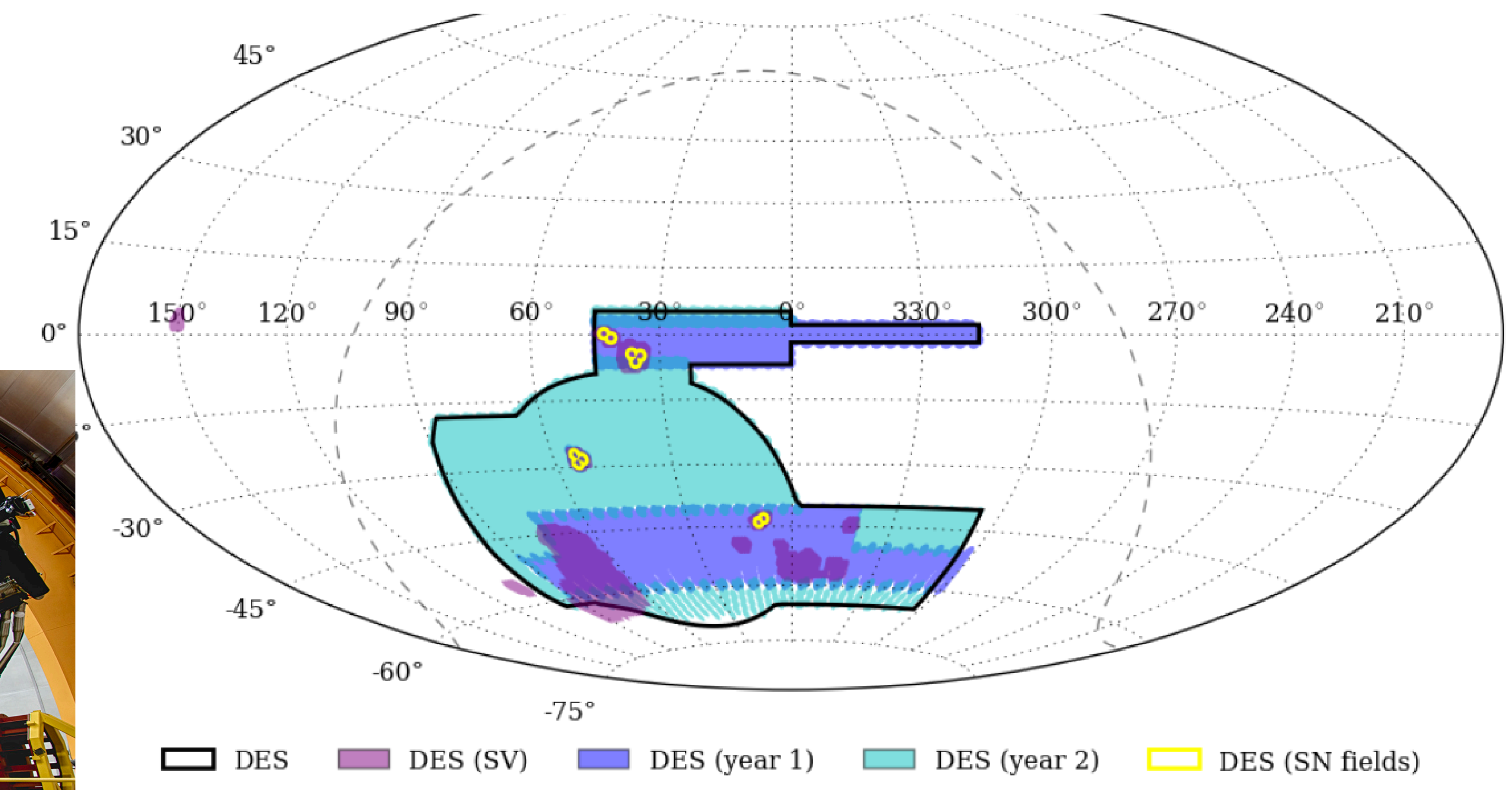


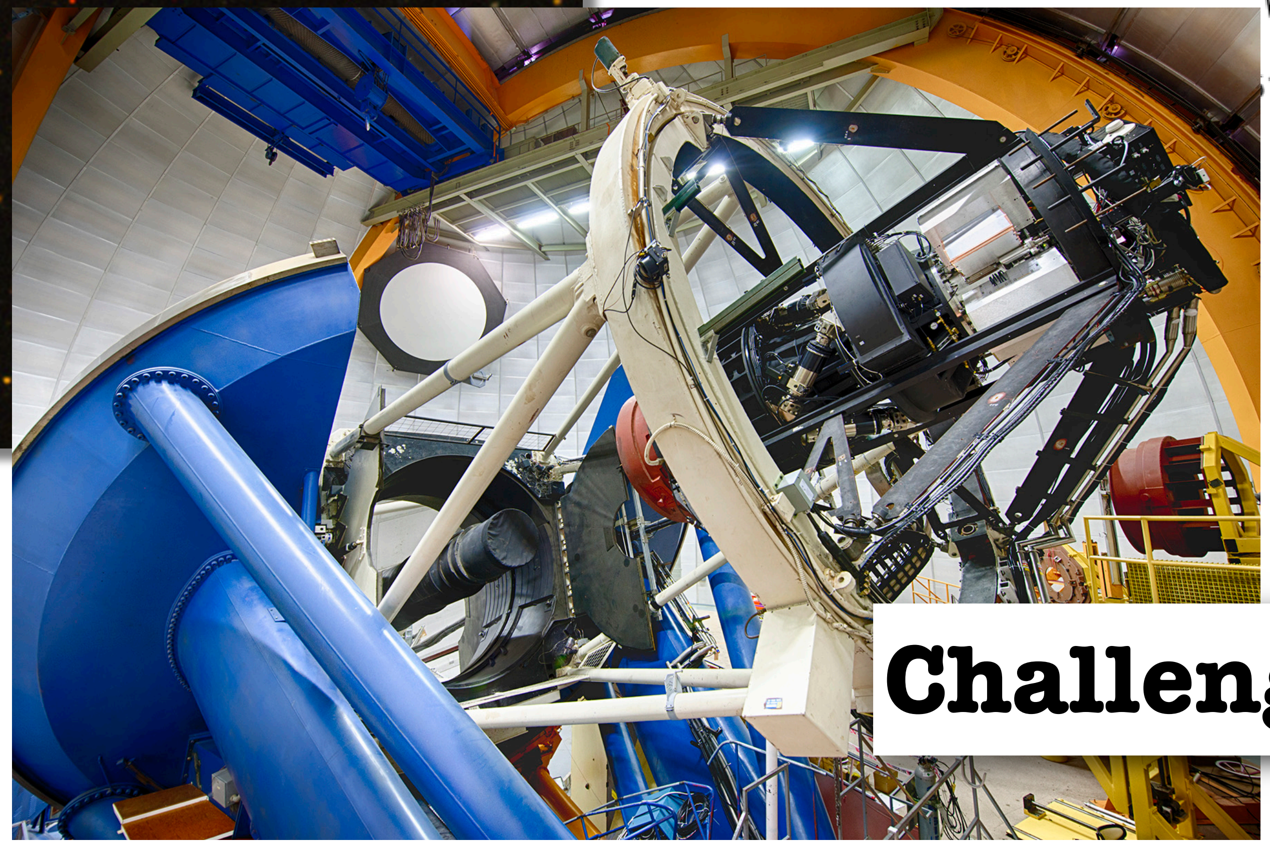
Populations of objects show **dark matter, dark energy**

Region-based CNNs on heterogeneous compute devices

- LSST: 20 Tb / night
- 1 Billion transient alerts /night

Dark Energy Survey: Sky Footprint of Observations





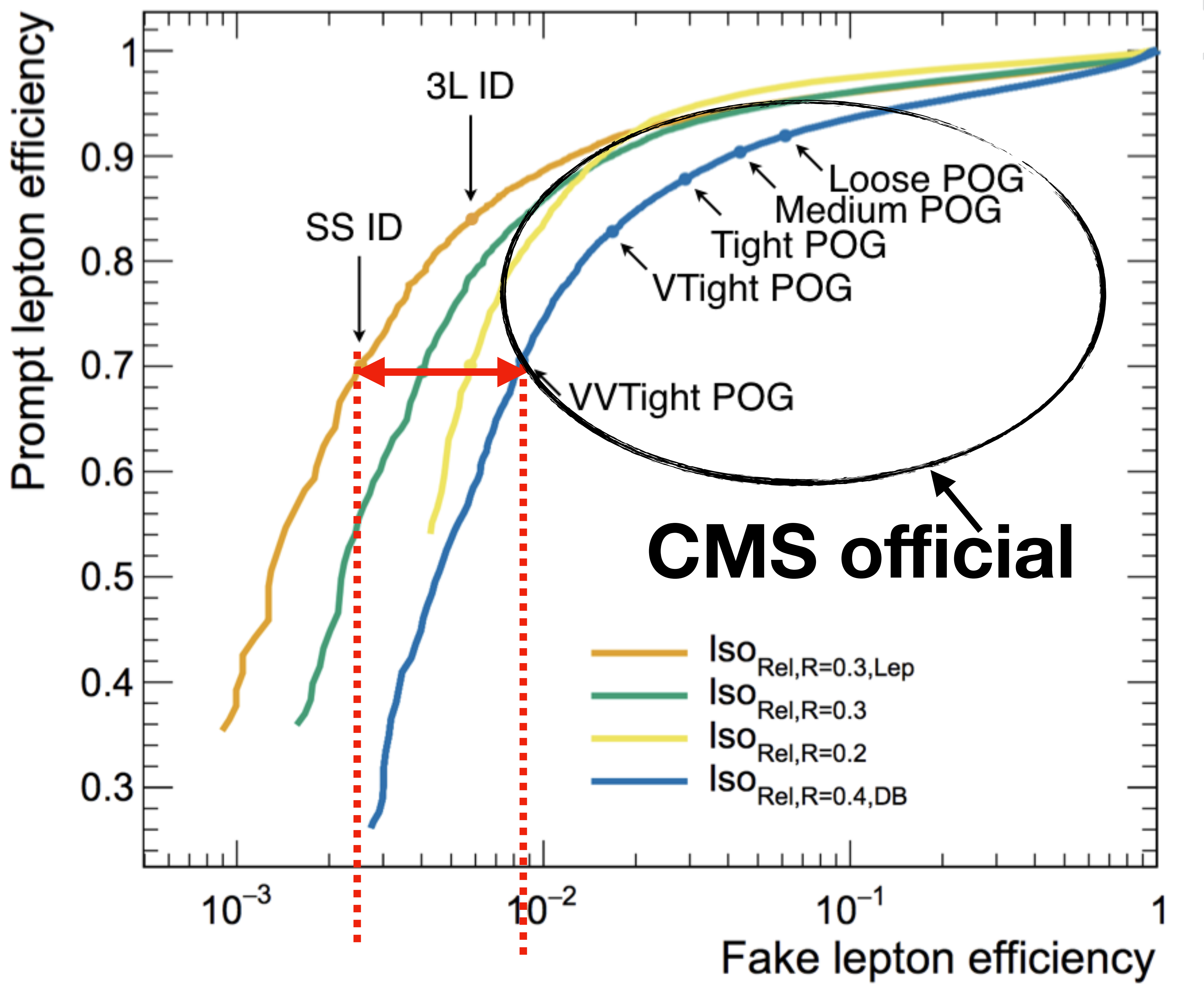
Challenge of scheduling on multiple time

Long: competition between faint galaxies, transient objects

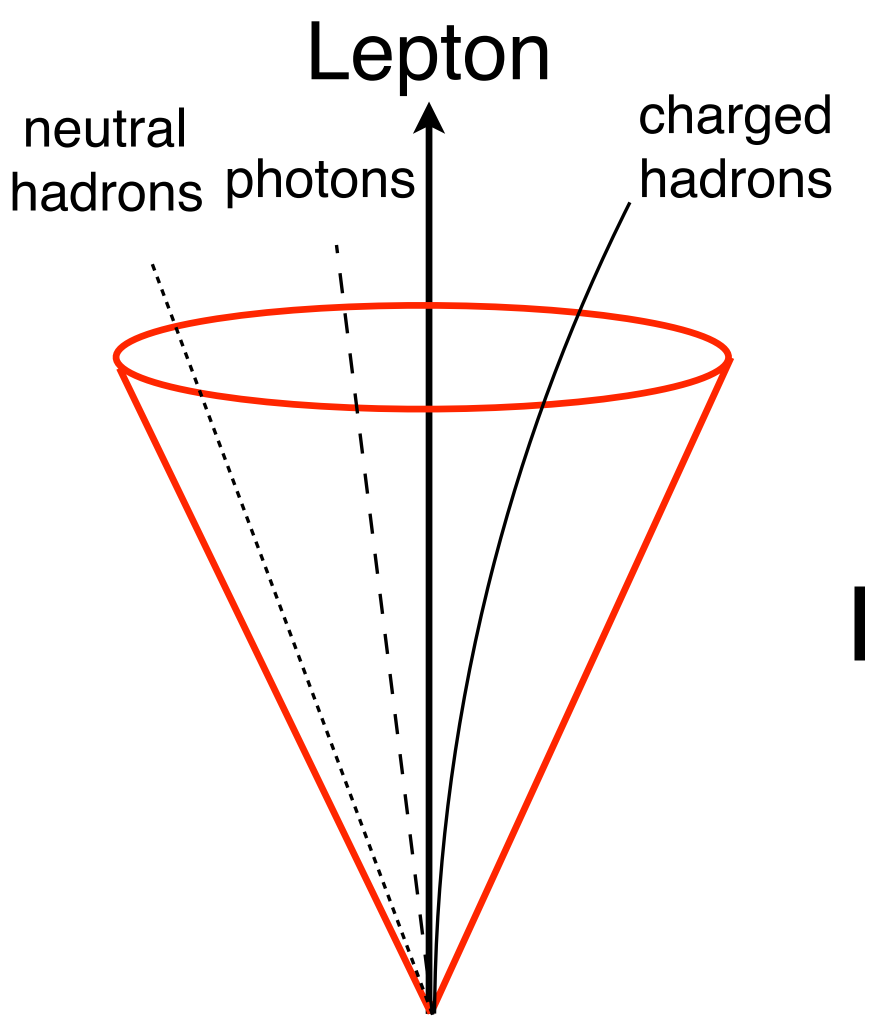
Short: Weather, annual modulation of sky positions

Smart telescopes: reinforcement learning for optimal scheduling and control

Improved isolation definition



3.5 X background rejection for muons
@ 70% efficiency

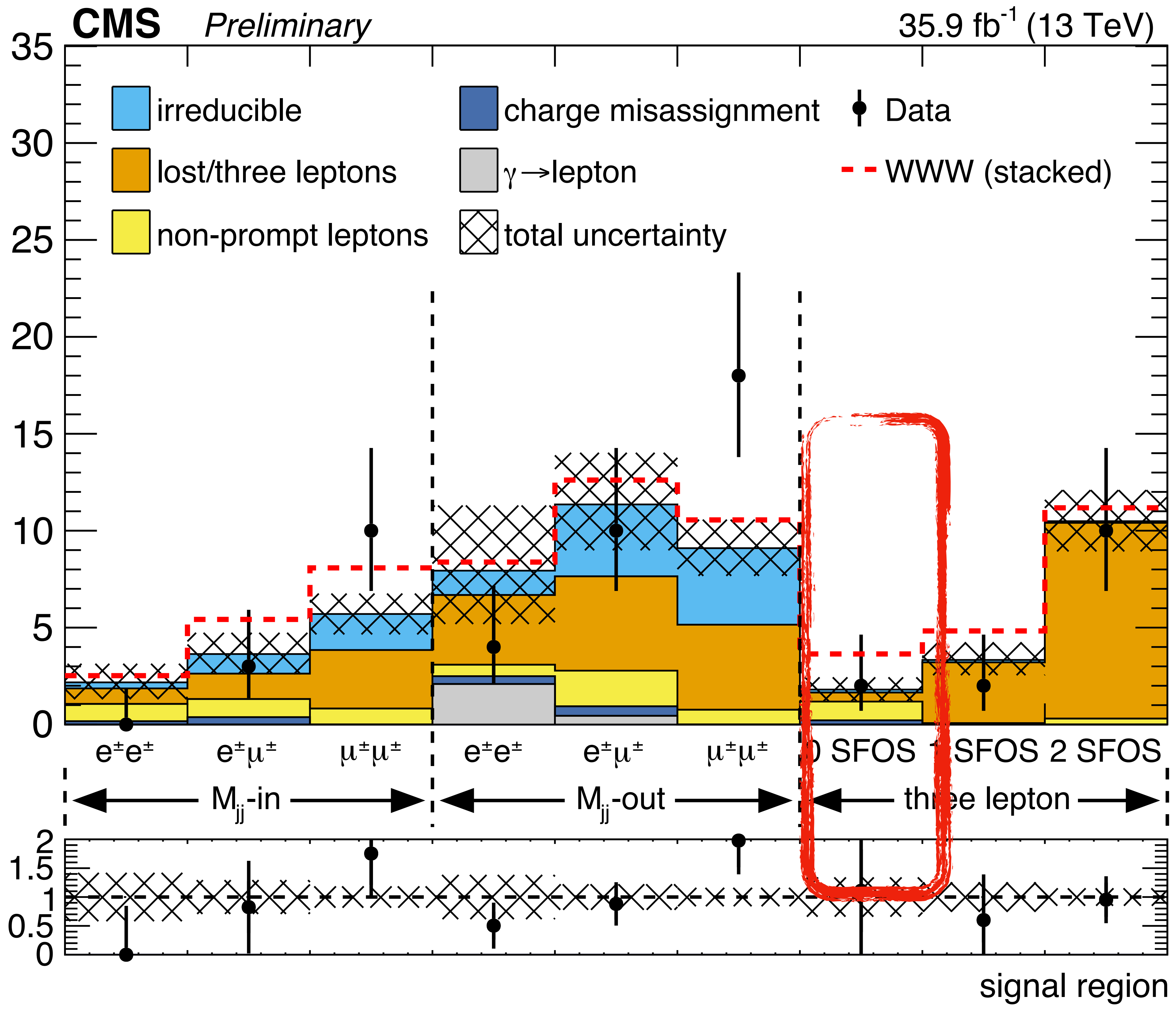


$$Iso_{Rel} = \frac{\Sigma \text{PF cand's } P_T \text{ in the cone} - \text{PU}}{\text{Lepton } P_T}$$

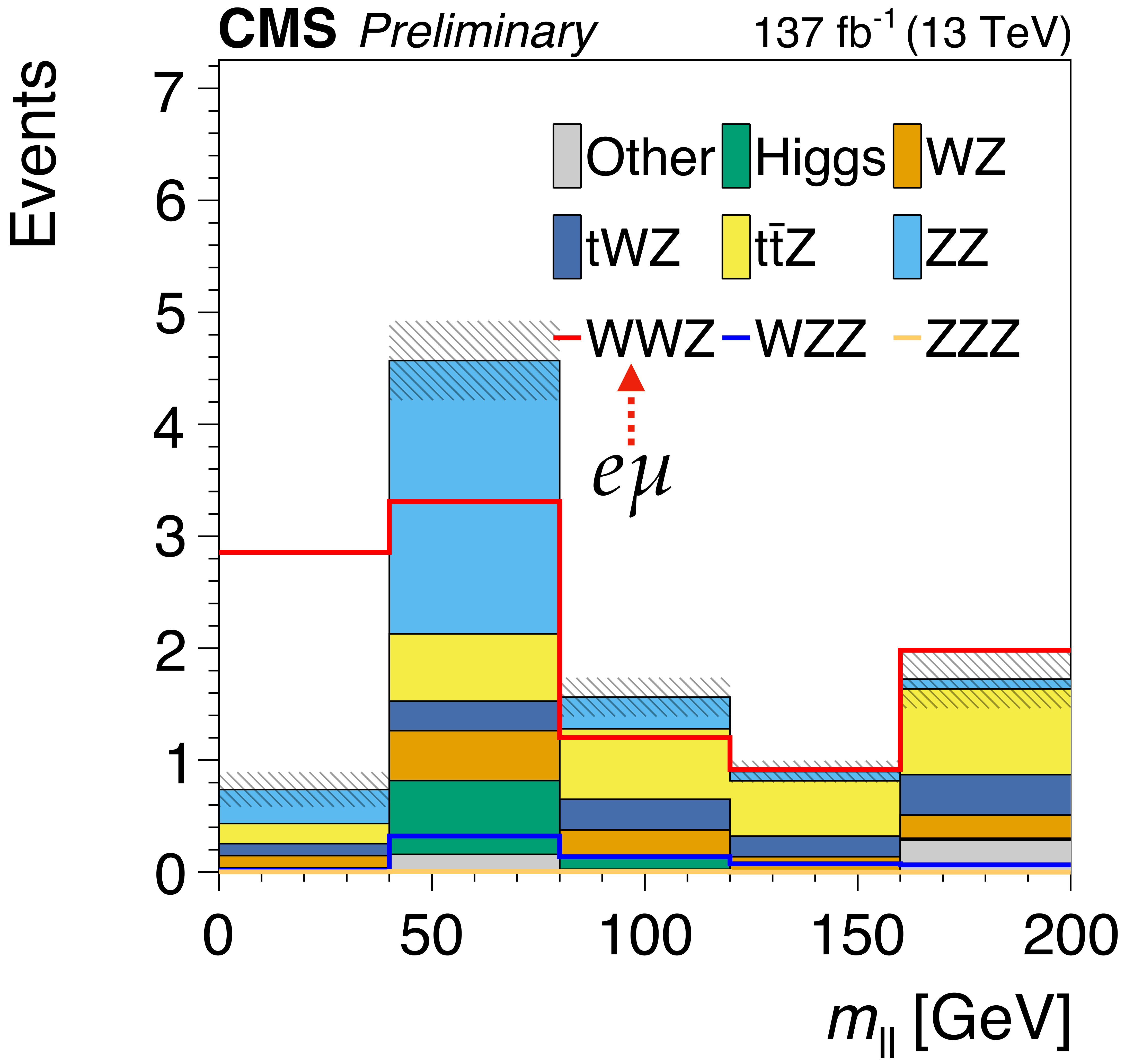
- Smaller cone-size: 0.4—>0.3
- Add lepton candidates to Isolation: improves rejection against heavy flavor decay (B→D→2 leptons + X), one of the leptons is selected as good lepton.

Towards Run-2 Result

- Improve cut-based analysis:
  - e.g. OSFOS (e+/-e+/-m-/+,m+/-m+/-e-/+)
  - Large fraction of fake leptons in 2016 analysis:
    - Dilepton ttbar
  - Low event yield:
    - susceptible to statistical fluctuations.
  - New selection features:
    - Customized IDs for e+/-e+/-m-/+,m+/-m+/-e-/+
    - Soft b jet veto : 30% fake rejection, no signal loss
    - Lifted kinematic selections
    - Overall improvement >50%.
- In parallel, exploring MVA based analysis.



Covering all VVV processes



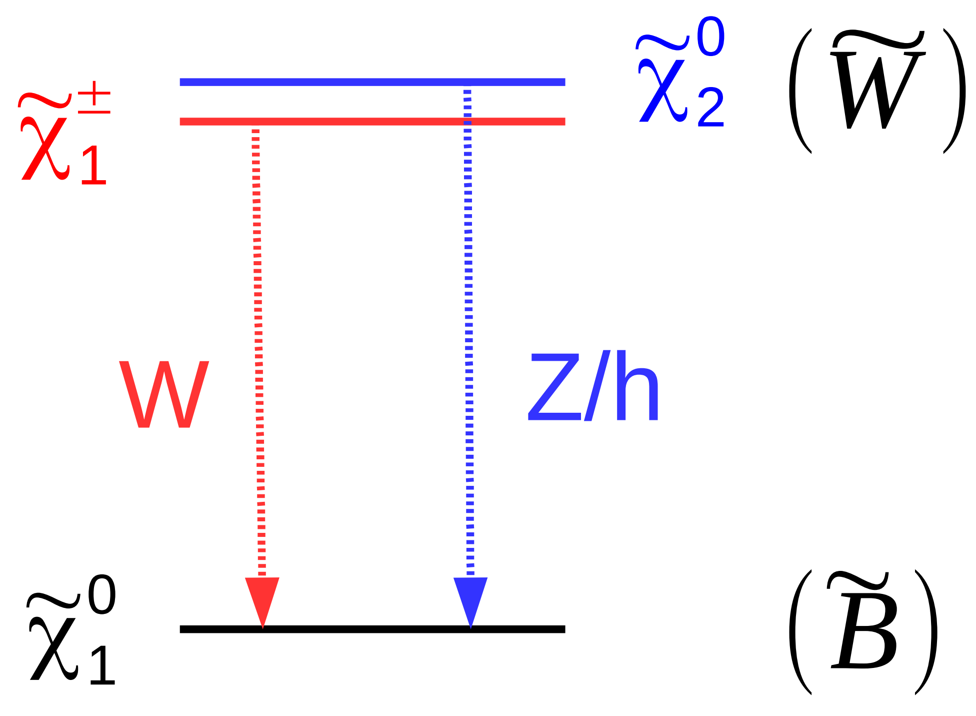
Quantities	WWZ	WZZ	ZZZ
$\sigma_{\text{total}} \times \mathcal{B}_{VVV\to 4\ell}$ (fb)	4.12	0.74	1.19
$\sigma_{\text{total}} \times \mathcal{B}_{VVV\to 5\ell}$ (fb)	-	0.36	-
$\sigma_{\text{total}} \times \mathcal{B}_{VVV\to 6\ell}$ (fb)	-	-	0.05
$\sigma_{\text{total}} \times \mathcal{B}_{VVV\to 4\ell} \times 137\,\text{fb}^{-1}$ (N evt.)	564	101	163
$\sigma_{\text{total}} \times \mathcal{B}_{VVV\to 5\ell} \times 137\,\text{fb}^{-1}$ (N evt.)	-	49.3	-
$\sigma_{\text{total}} \times \mathcal{B}_{VVV\to 6\ell} \times 137\,\text{fb}^{-1}$ (N evt.)	-	-	6.85

- Handful of other VVV events produced with full Run-2 data (Leptonic channels).
- WWZ(4l) has the best expected sensitivity: 3.5 σ
  - Events categorized with W leptons(ee/mm vs em).

Expect the first evidences of WWW and WWZ production with full Run-2 dataset

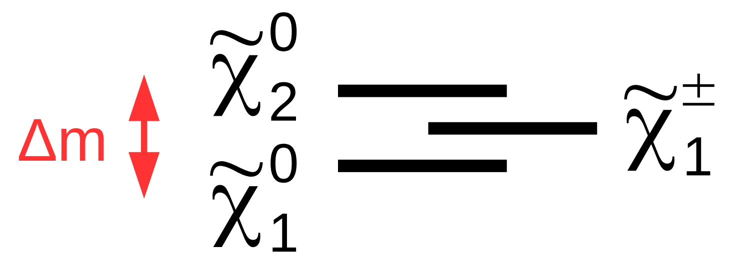
Some topics need HL-LHC dataset

- 3000 fb-1 data expected at the HL-LHC
- e.g. Higgsinos: Low cross section, challenging signatures
    - Δm ~ tens of GeV : Soft decay products
    - Δm ~ hundreds of MeV : Long-lived signatures



Wino-like
45 fb*

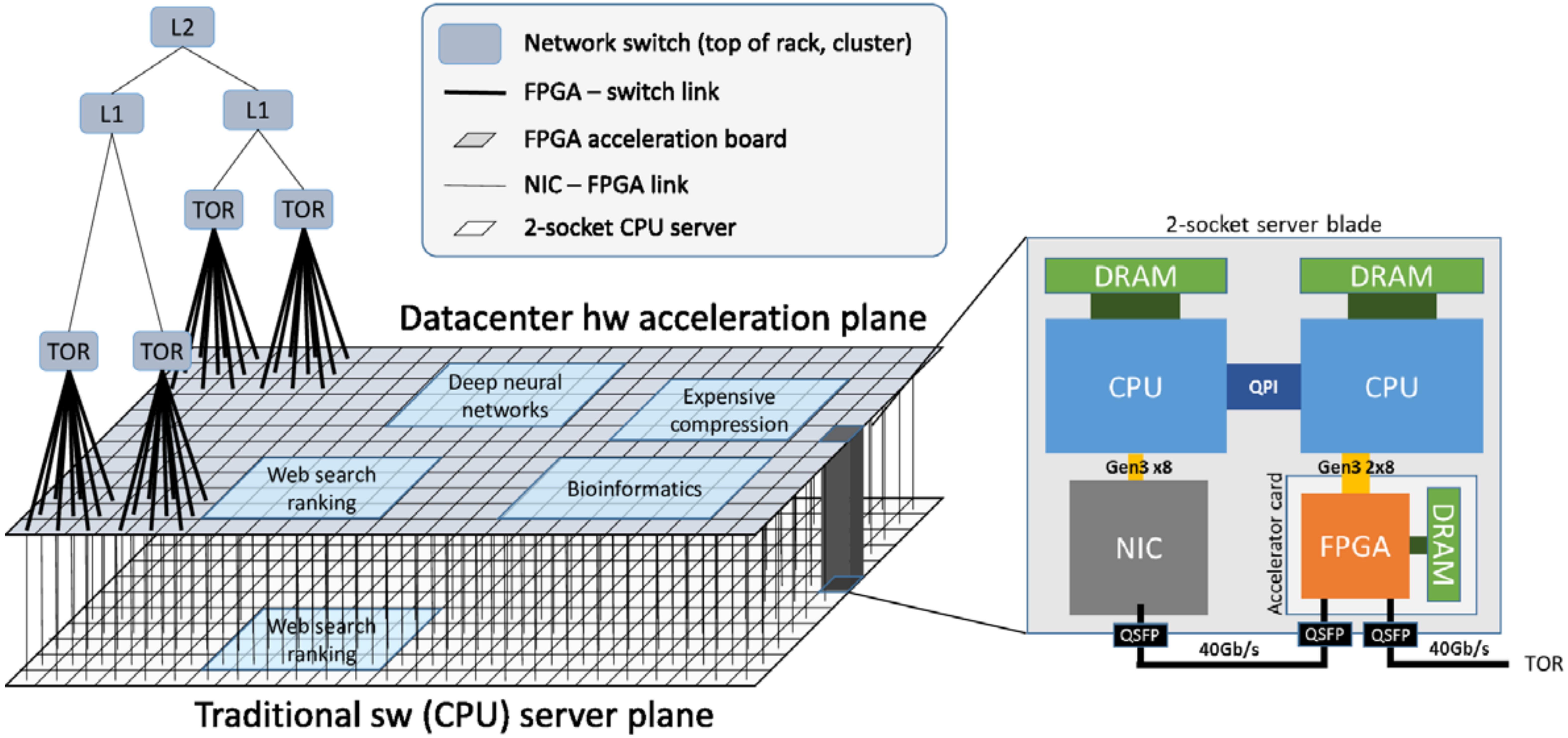
Δm ~ hundreds of MeV to tens of GeV



Higgsino-like
11 fb*

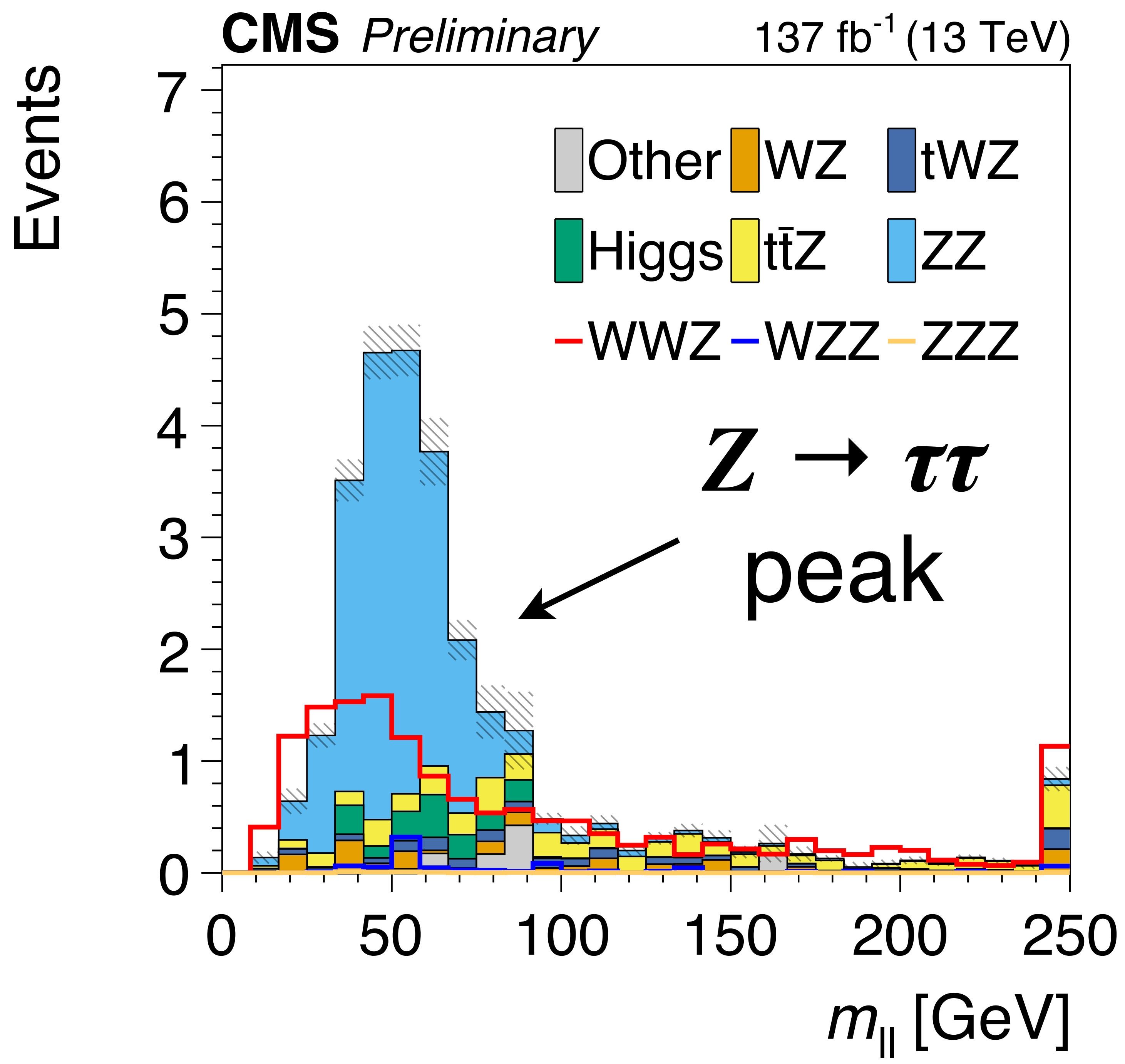
*Cross-sections for 500 GeV sparticles @ 13 TeV($\tilde{\chi}_2^0\tilde{\chi}_1^{\pm}$ only)

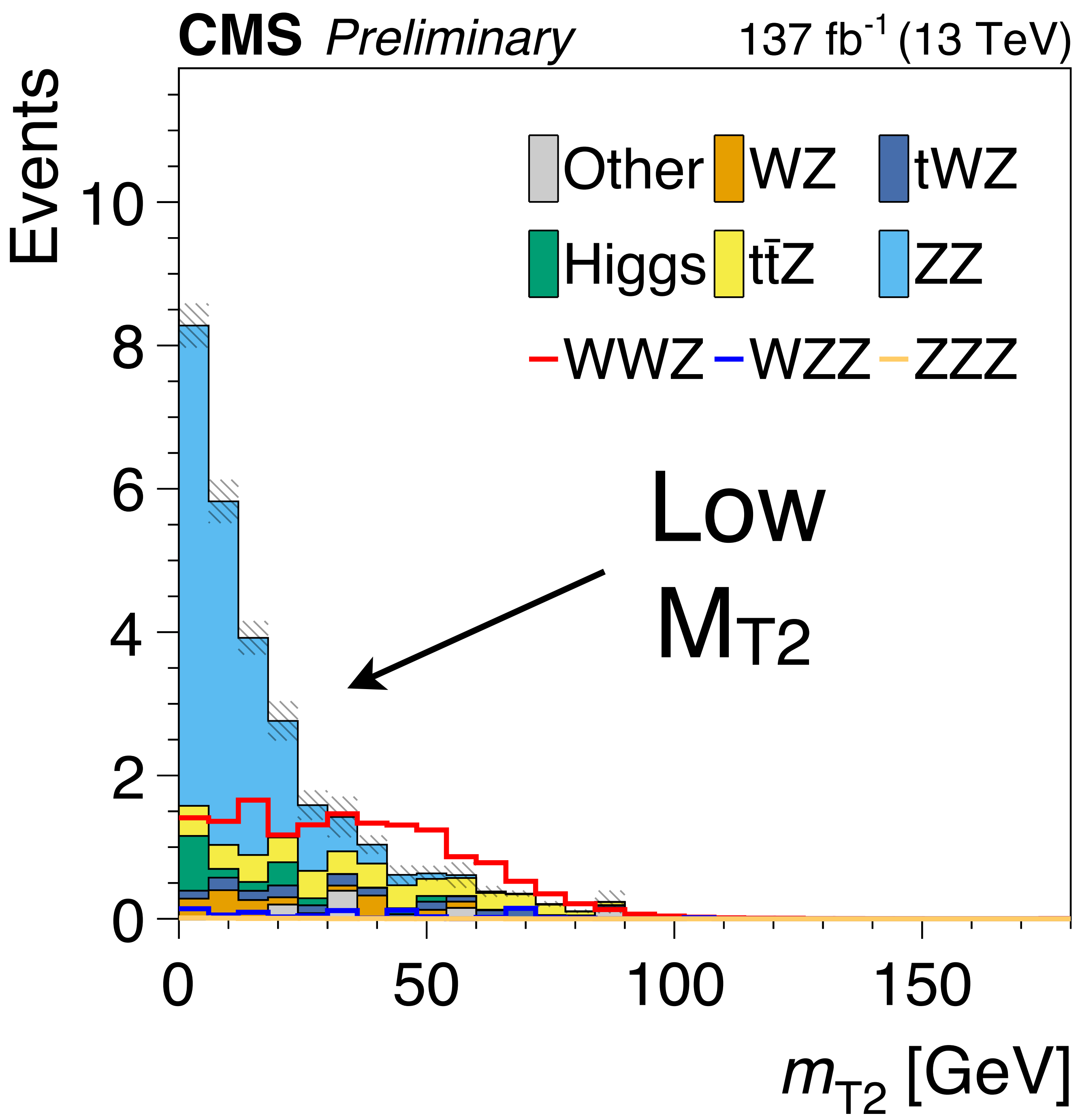
Microsoft Brainwave



- **Mature service** at scale (more than just a single co-processor)
- Multi-FPGA/CPU fabric accelerates *both* **computing** and **network**
- Models supported:
  - ResNet50, ResNet152, DenseNet121 ,VGGNet16…
  - Partially fixed neural network architecture. weights can be retuned.

WWZ: smaller rate but clean

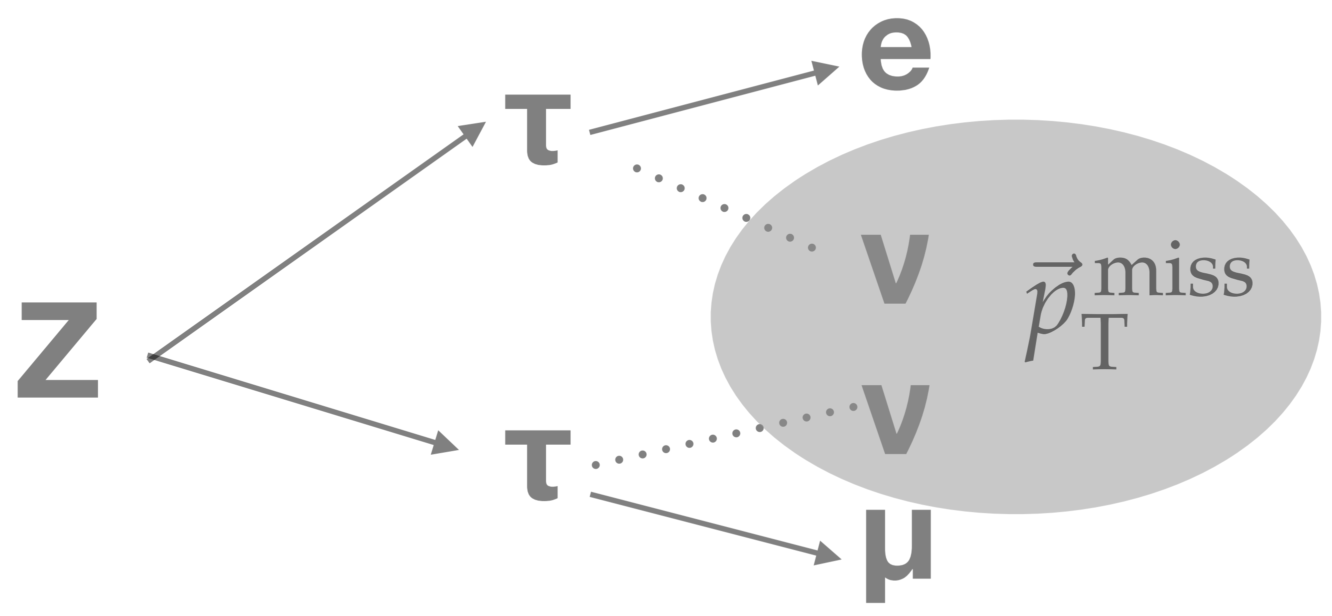




40% smaller than WWW , leptonic decays

4ℓ: Tag Z->$\ell\ell$, WW-> (eμ/ee/μμ)

- eμ shown as example: kinematic selections against ZZ



$$m_{\mathrm{T2}} = \min_{\vec{p}_{\mathrm{T}}^{\,\nu(1)}+\vec{p}_{\mathrm{T}}^{\,\nu(2)}=\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}} \left[\max\left(m_{\mathrm{T}}^{(1)}(\vec{p}_{\mathrm{T}}^{\,\nu(1)},\vec{p}_{\mathrm{T}}^{\,\mathrm{e}}), m_{\mathrm{T}}^{(2)}(\vec{p}_{\mathrm{T}}^{\,\nu(2)},\vec{p}_{\mathrm{T}}^{\,\mu})\right)\right]$$

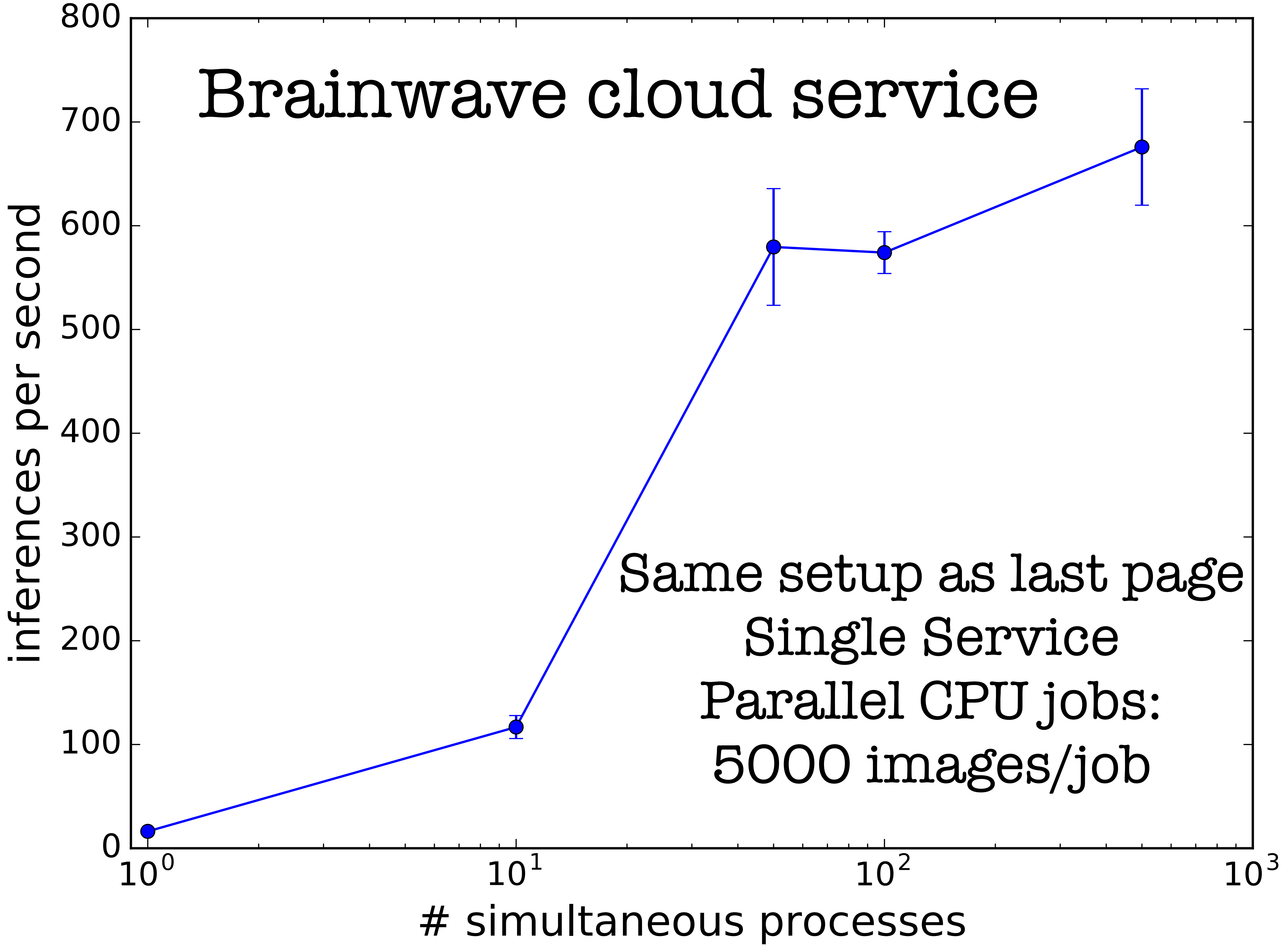
EDGE DATA BOX @ Fermilab

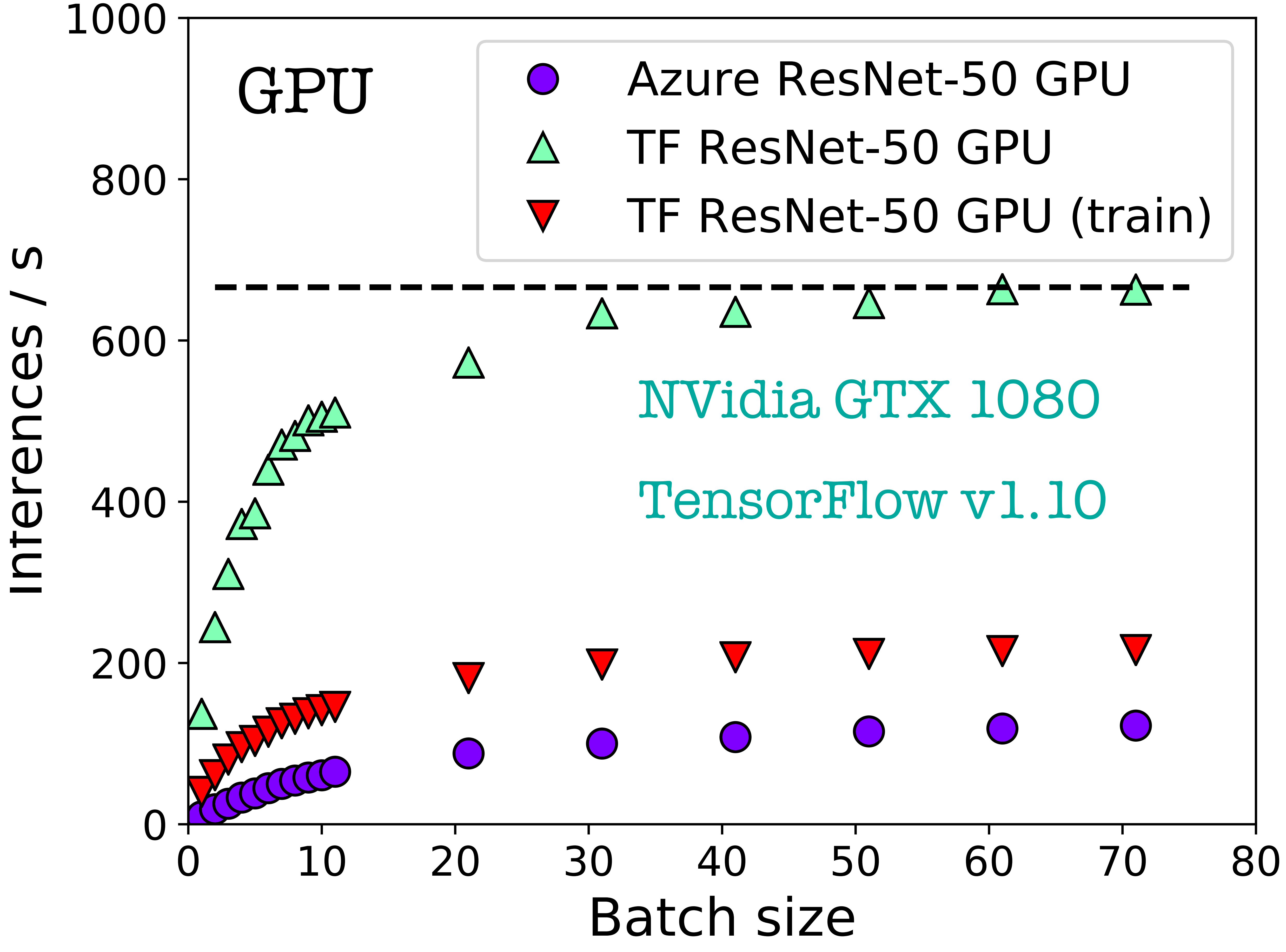
Edge data box at Feymann computing center

Docker container on server (PCIe connection):	14 ± 25 ms
Fermilab computing cluster:	20 ± 30 ms
Local laptop:	68 ± 27 ms
CERN (Geneva):	168 ± 62 ms

- Gain experience in deploying co-processors in local clusters with cloud native tools: docker image, kubernetes
- Benchmark latency and scaling performance, compare with previous studies
- Can be used for neutrino and cosmology experiments as of ~today—> next slide

data throughout compared to





Comparable max data throughout: 600-700 images/sec

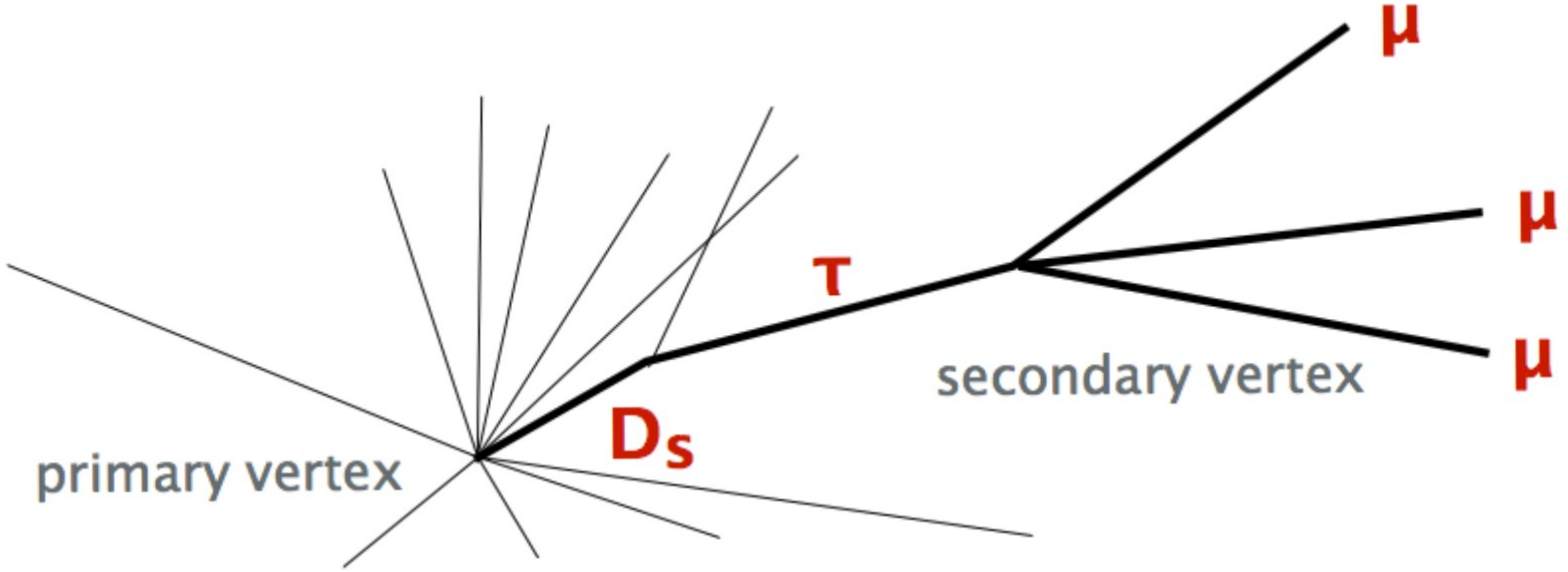
Beyond fully connected neural networks

Neutrinos oscillate:

Lepton number not conserved

At the LHC

What about in charged leptons? τ->3μ

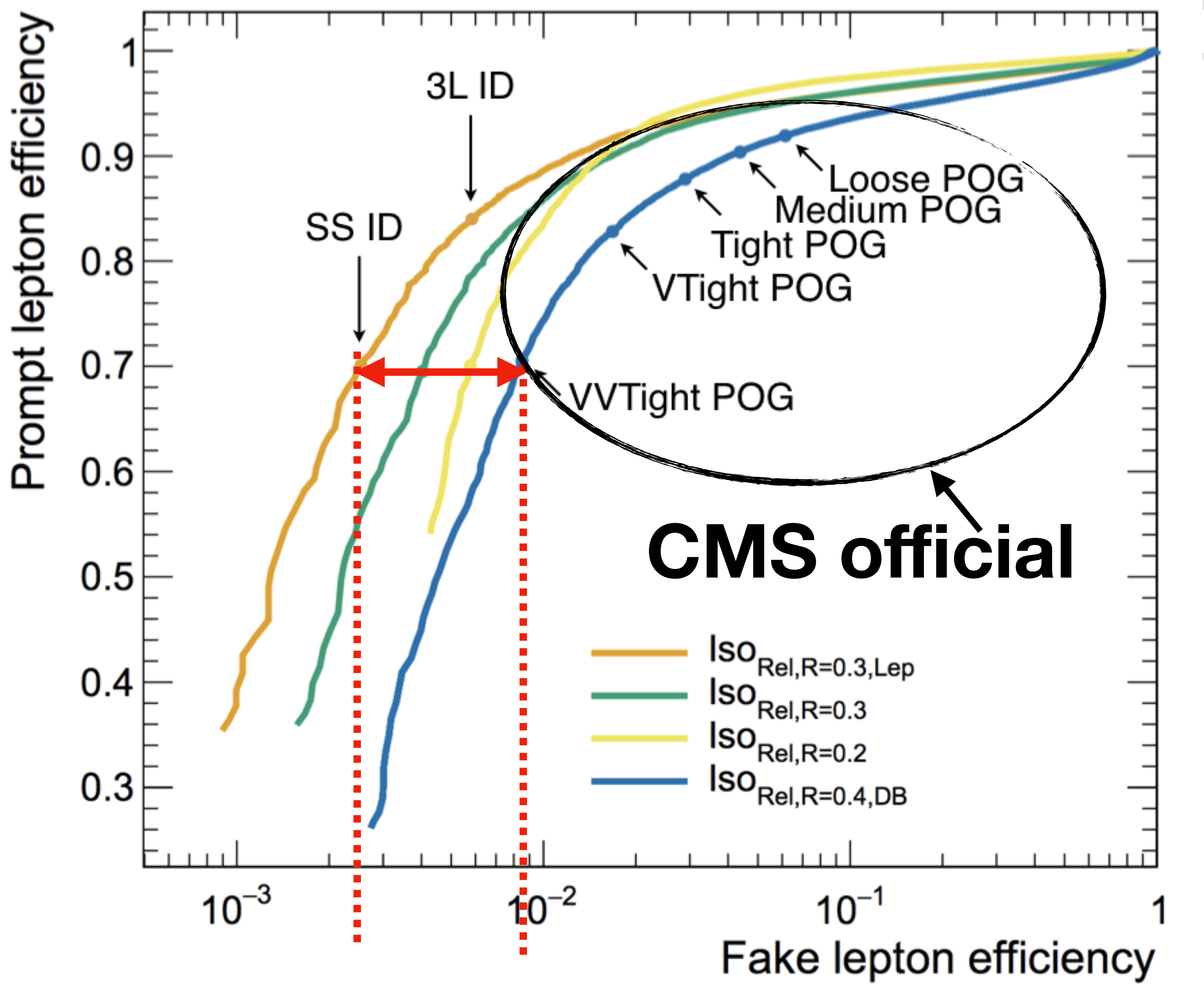


SM: 10^{-25}

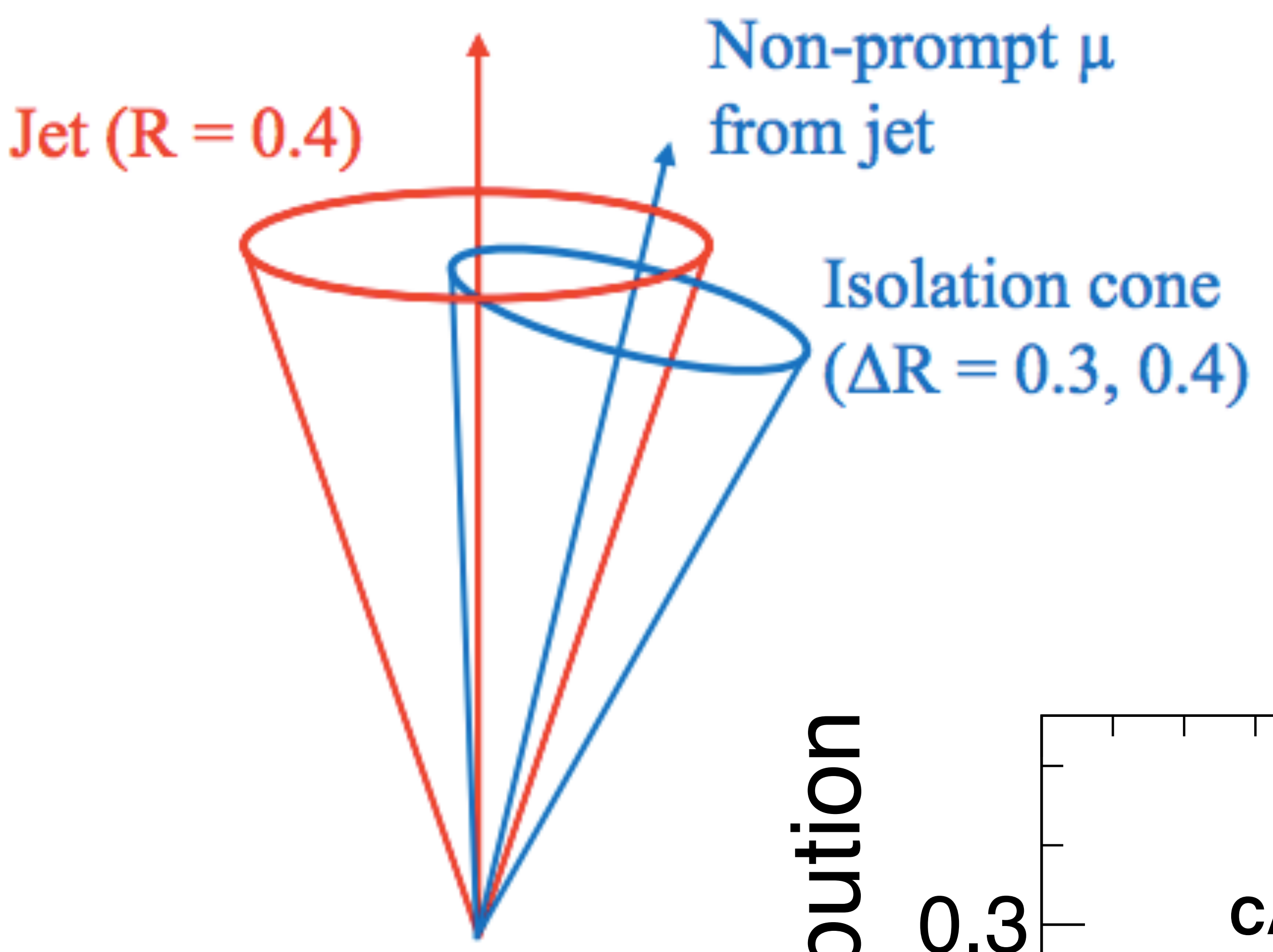
Currently:< 10^{-8}

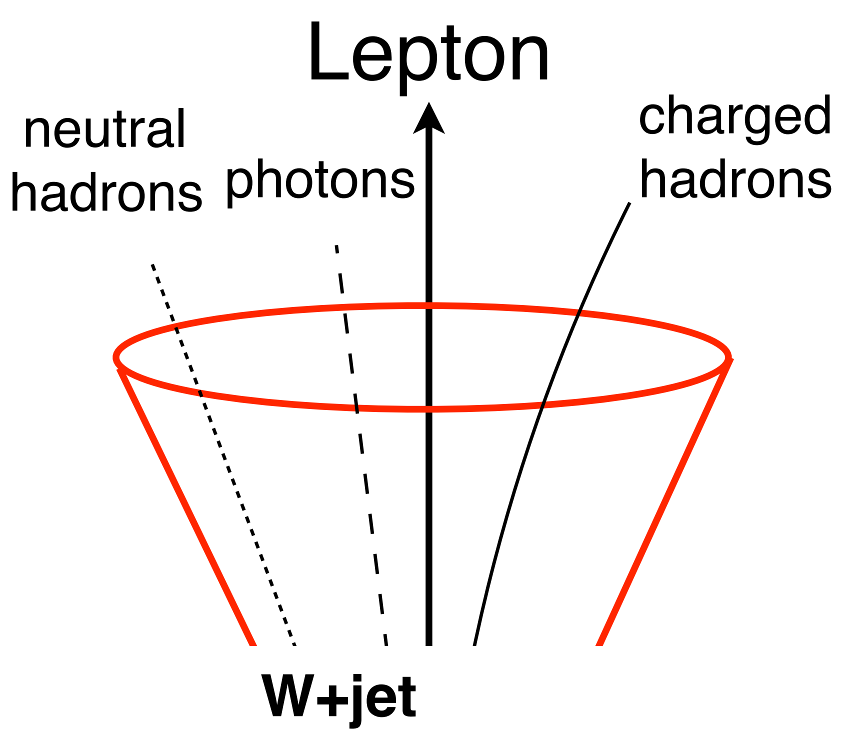
M.LIU

Improved lepton isolation definition



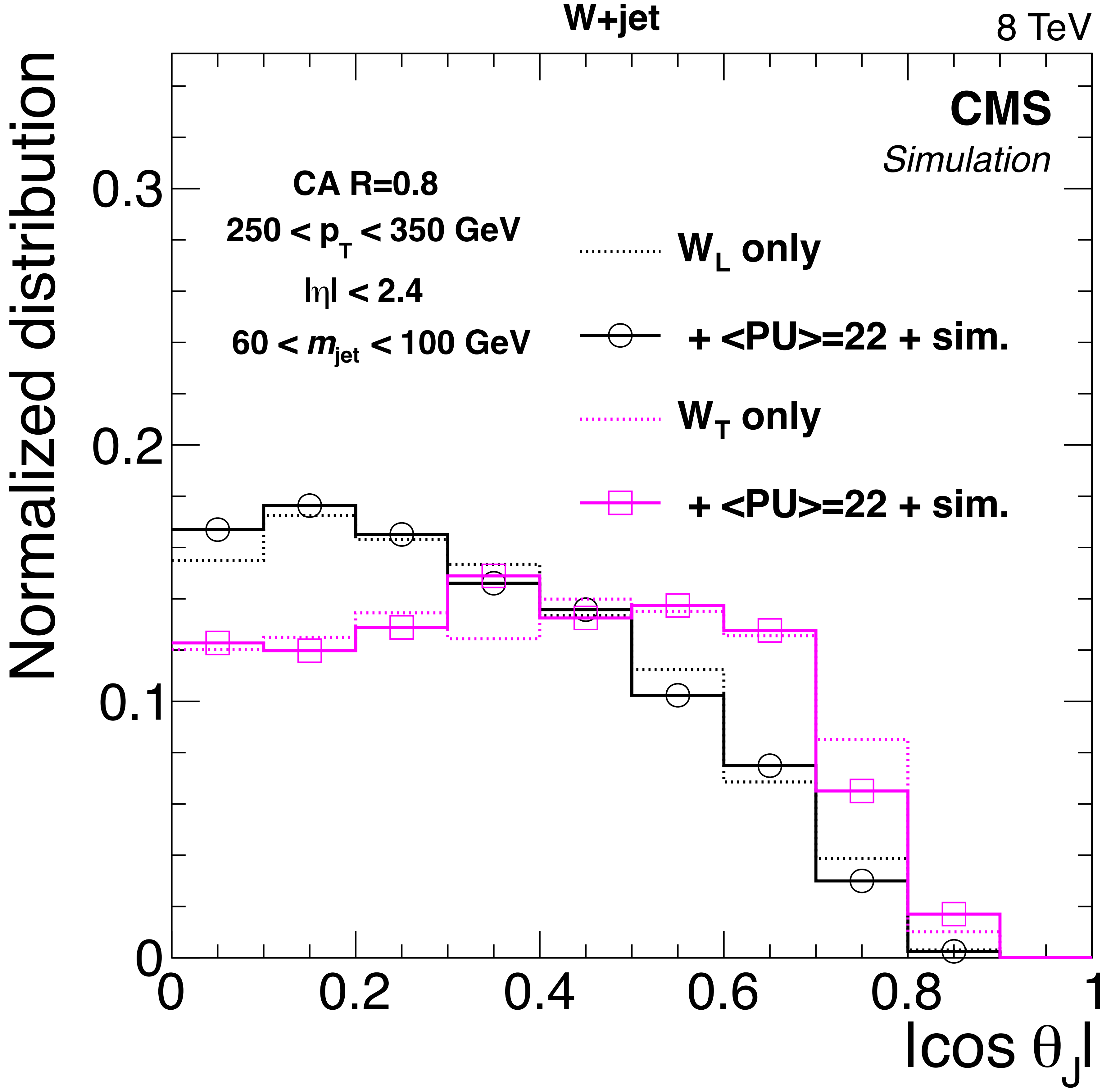
3.5 X background rejection for muons





Iso_{Rel} :

- Smaller co
- Add lepton
improves re
decay (B→l
leptons is s



Pixel detector of CMS at the LHC

Need to cope with more challenging LHC environment in Run 2 & Run 3 (300 fb^{-1}) until HL-LHC upgrade (2023).

Module designed to reduce dynamic inefficiency

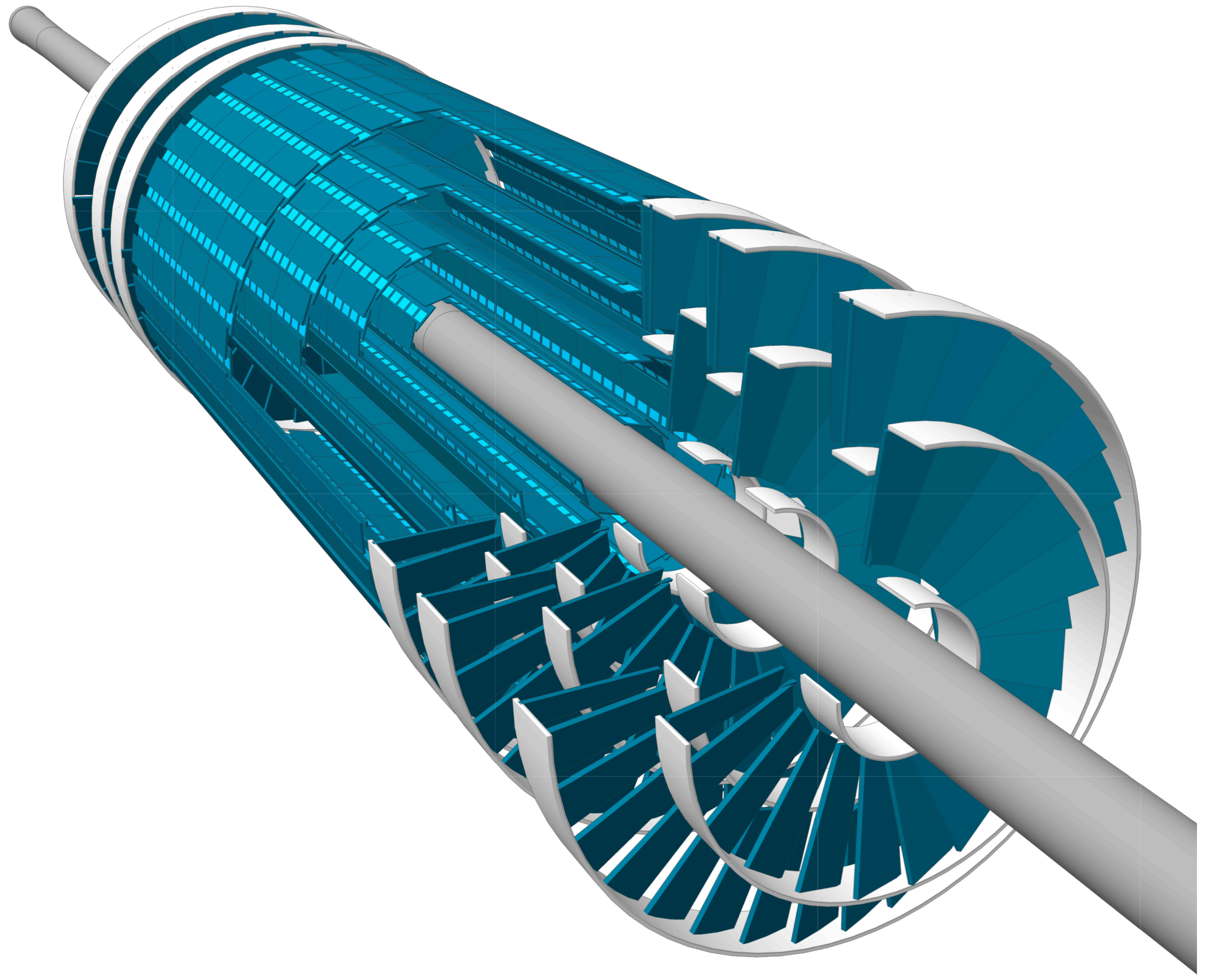
Digital readout chip (ROC). Faster readout.

Geometry design: ensure tracking and vertex quality

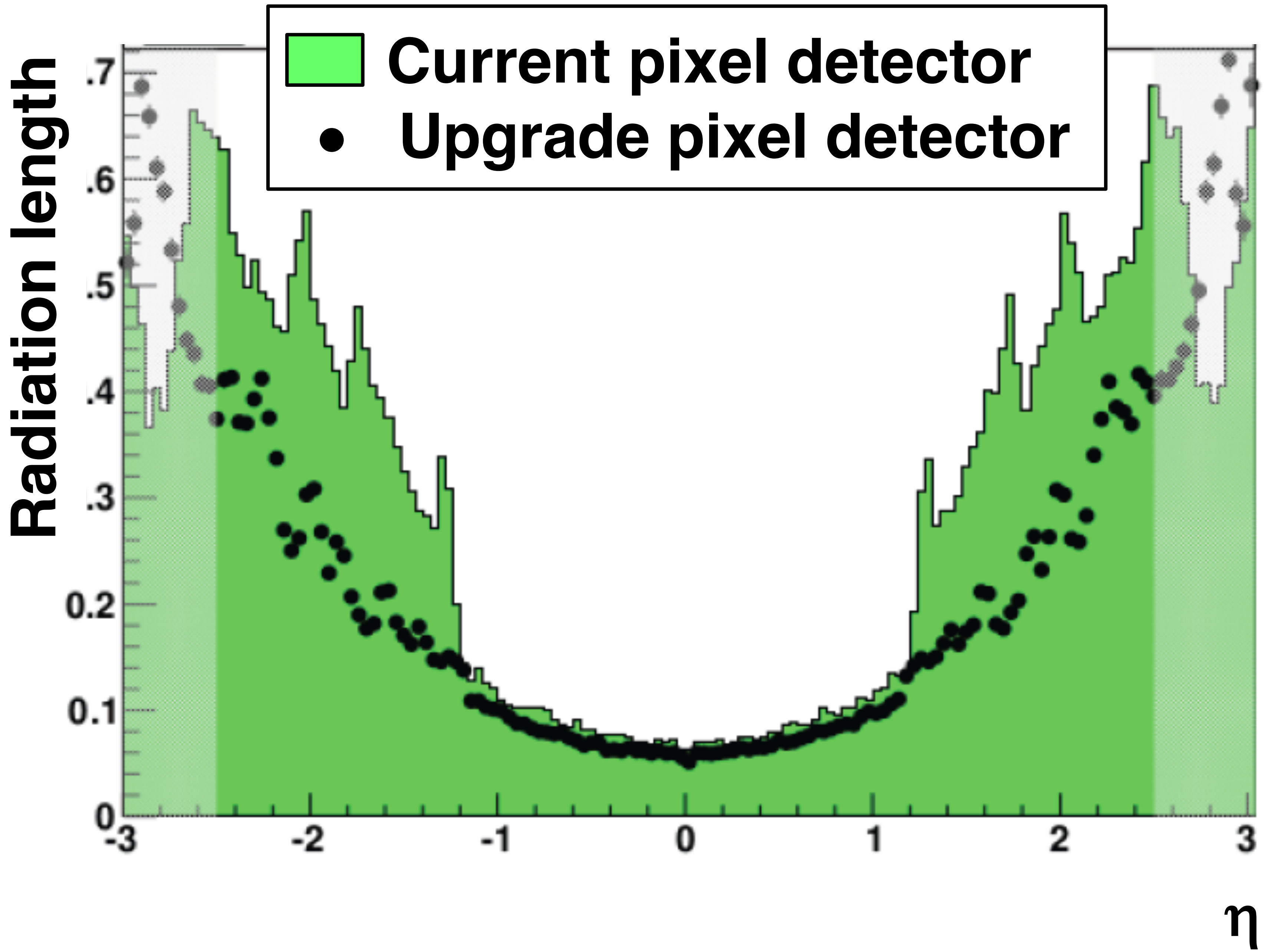
Added layers, channels doubled

Services: reduce material budget

CO_2 cooling, DCDC powering, Service electronics out of tracker volume.

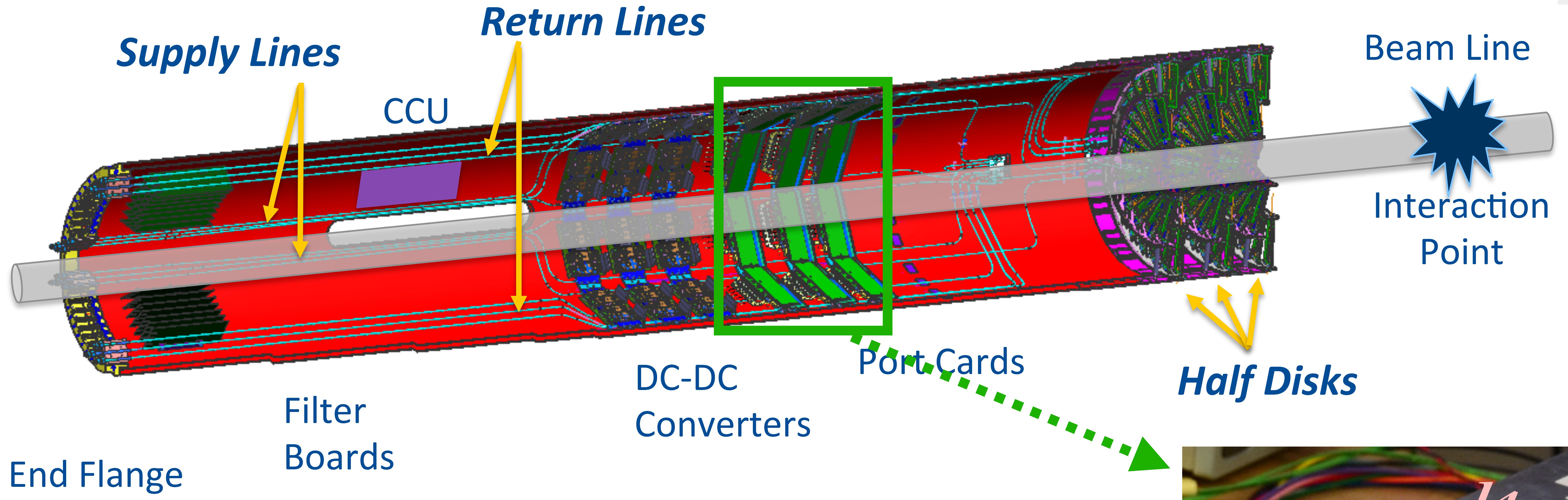


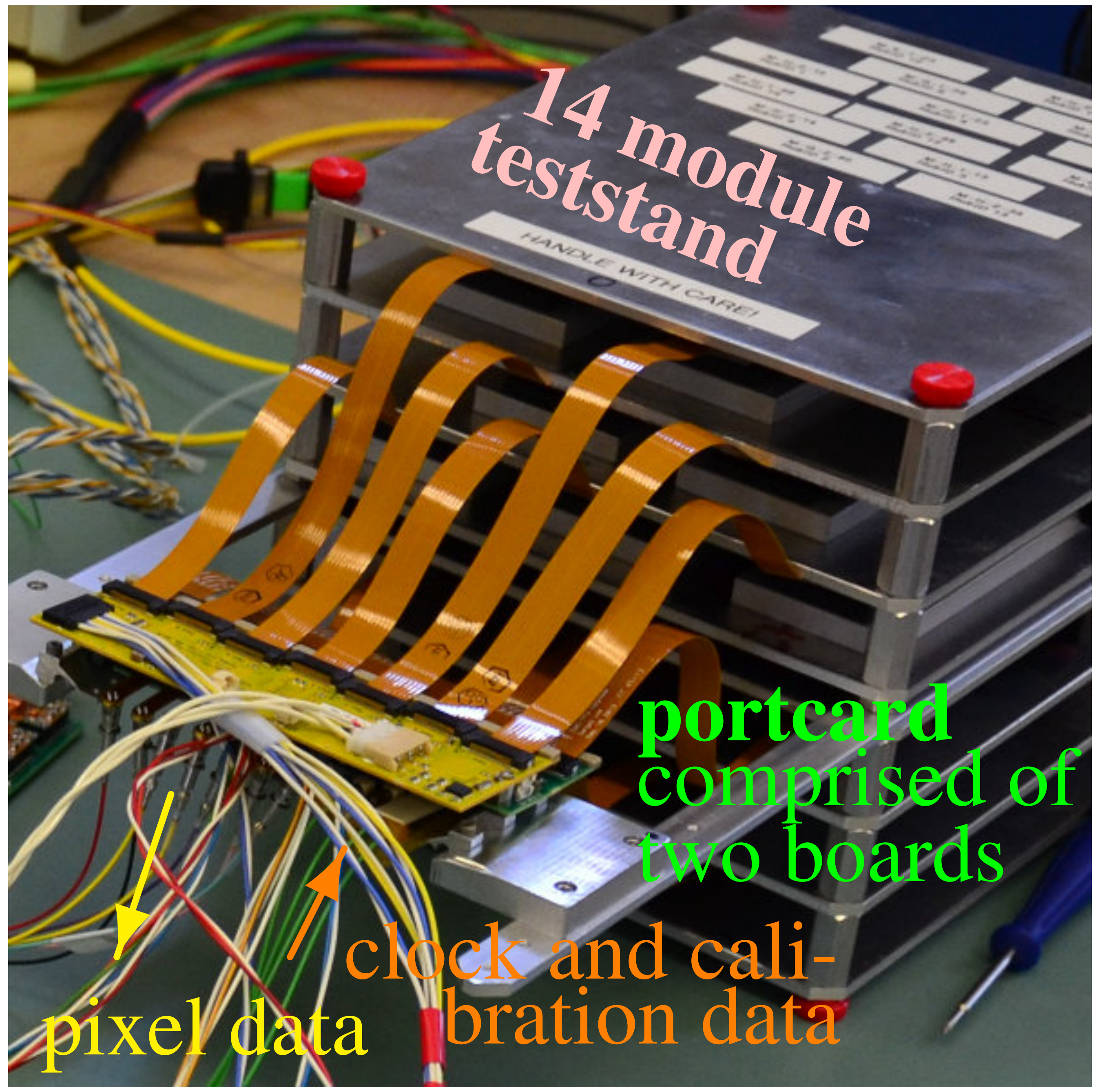
Material budget

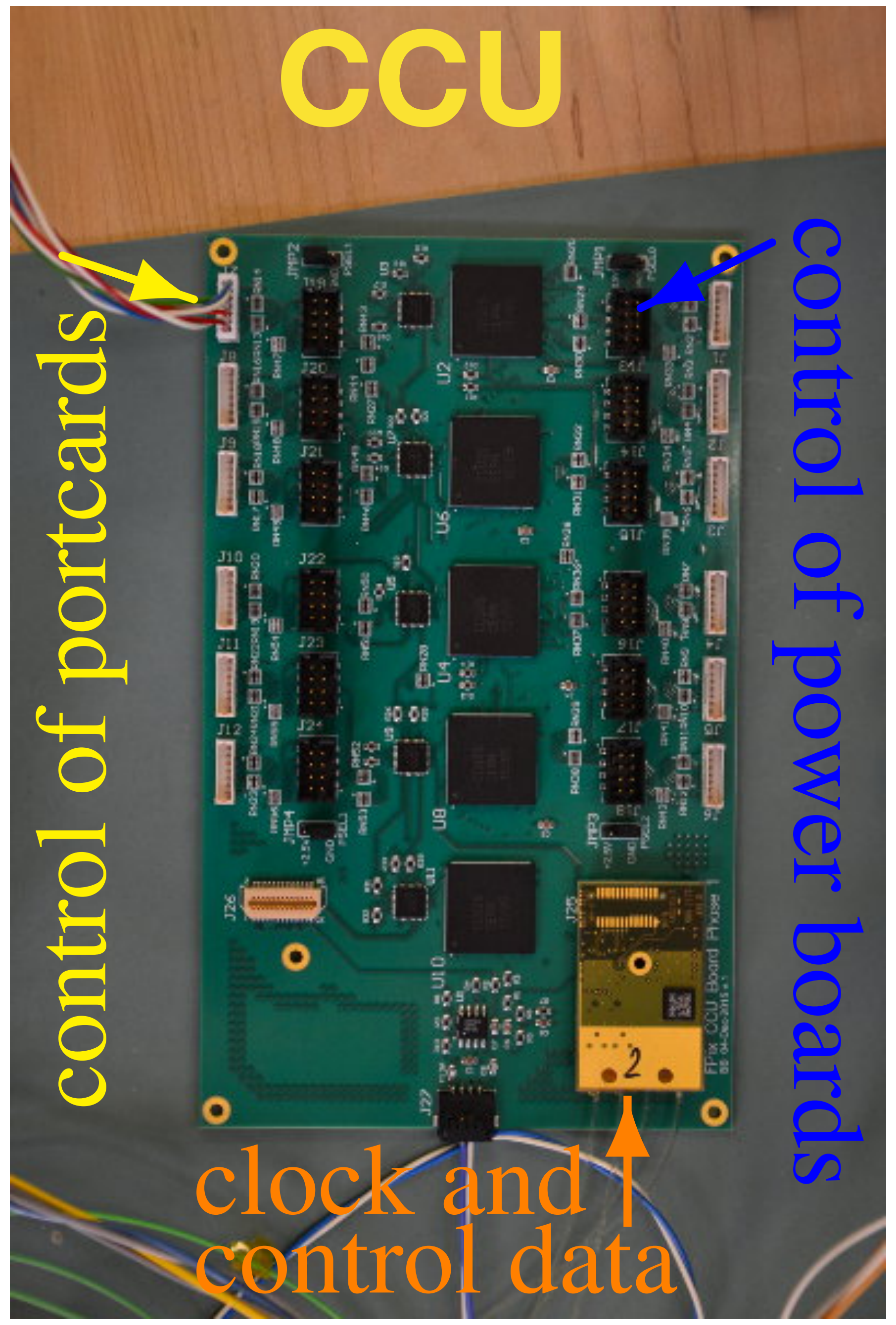


Half cylinders and service electronics

M.Liu

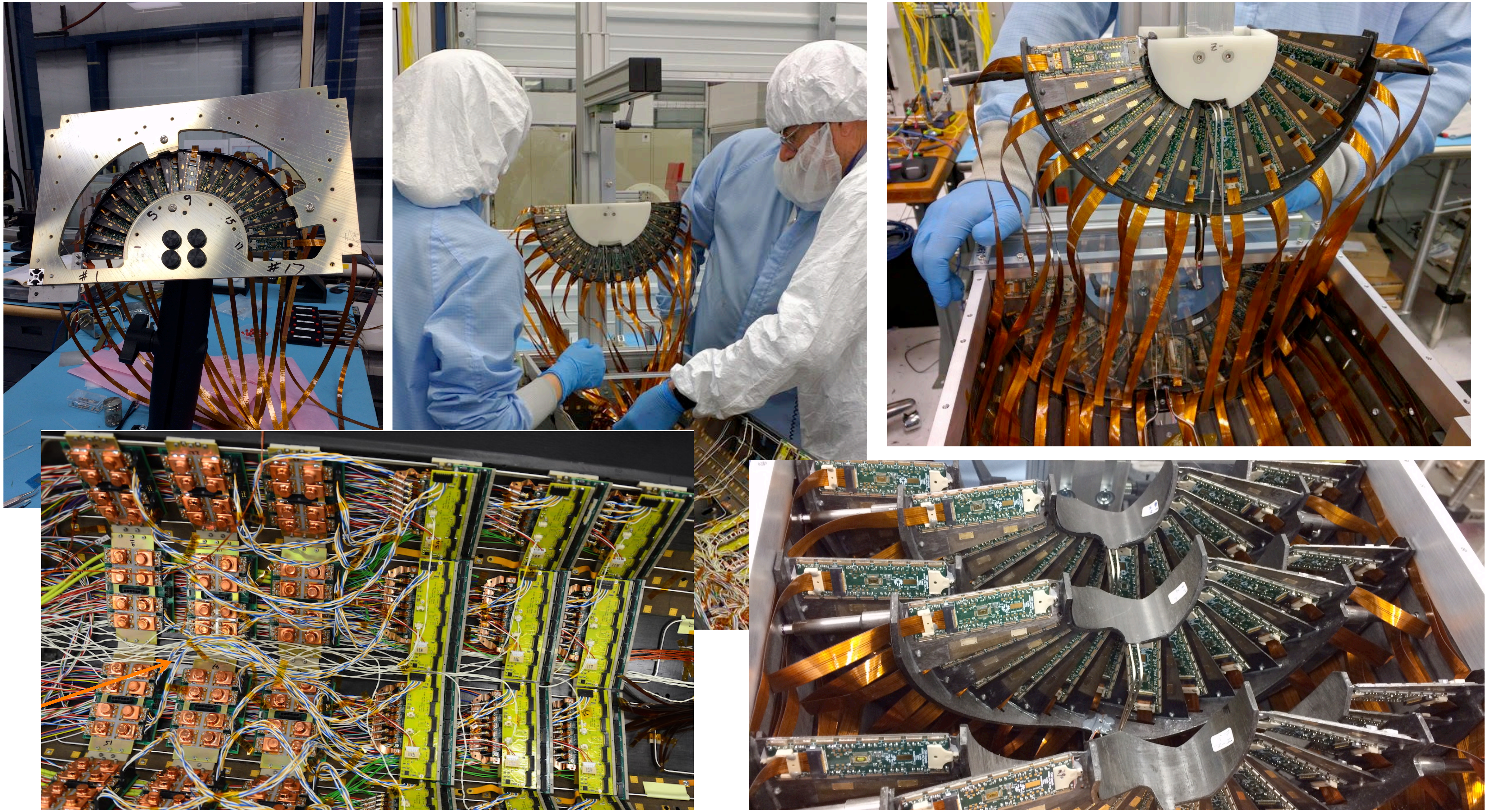






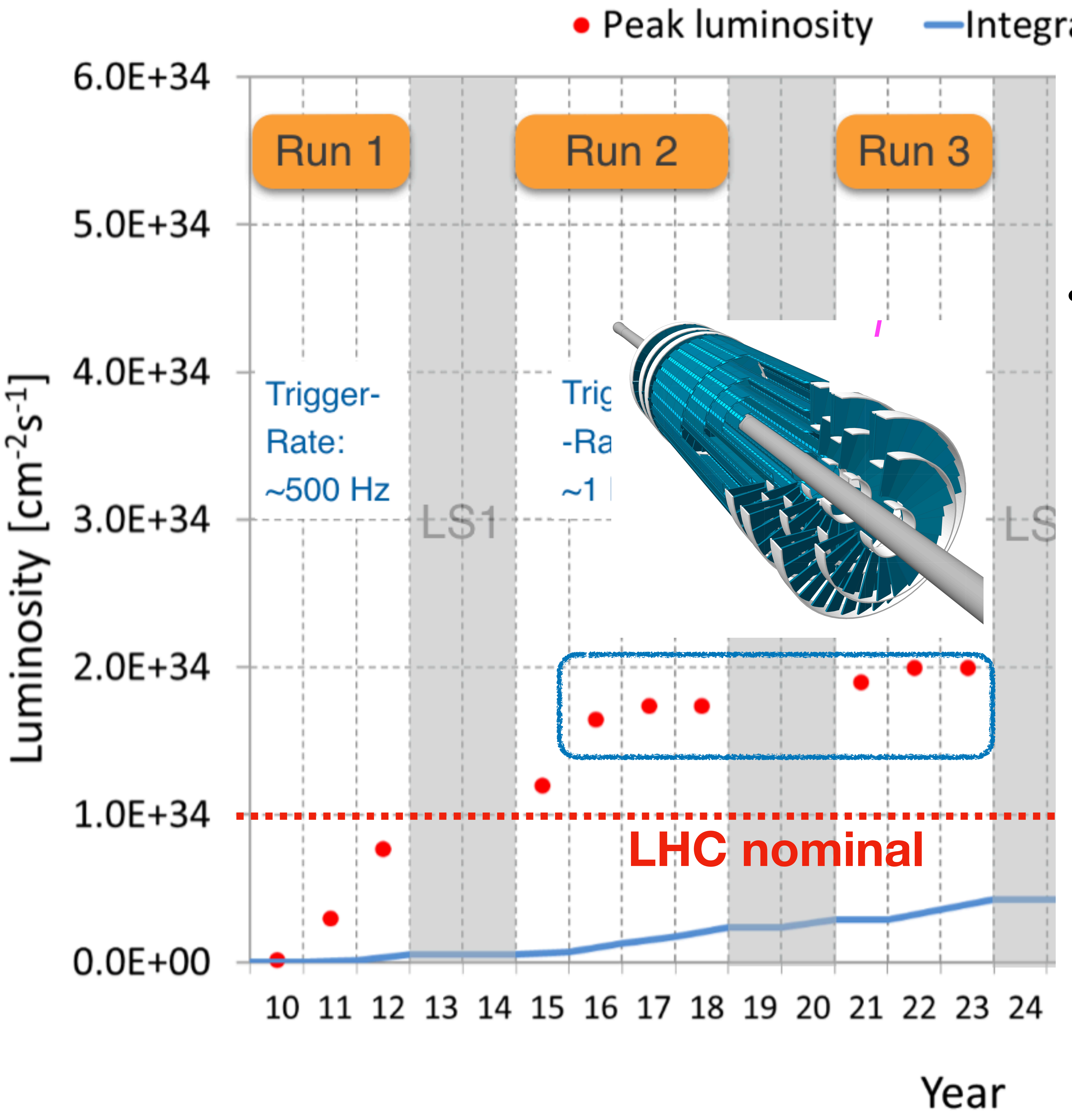
- **Portcard:**
  - Distributes power and bias voltages, clock, trigger and calibration signals to modules. Programs Modules (TBM and ROCs)
  - **Electric/optical Converters mounted**
    - Digital opto-hybrid (DOH): Optical—>Electrical
    - Pixel opto-hybrid (POH): Electrical—>Optical
- **CCU**:Communication & Control Unit
- **uTCA crate hosting front-end controller/drivers.**

FPIX assembly at Fermilab



All four half-cylinders tested with full DAQ readout chain at Fermilab

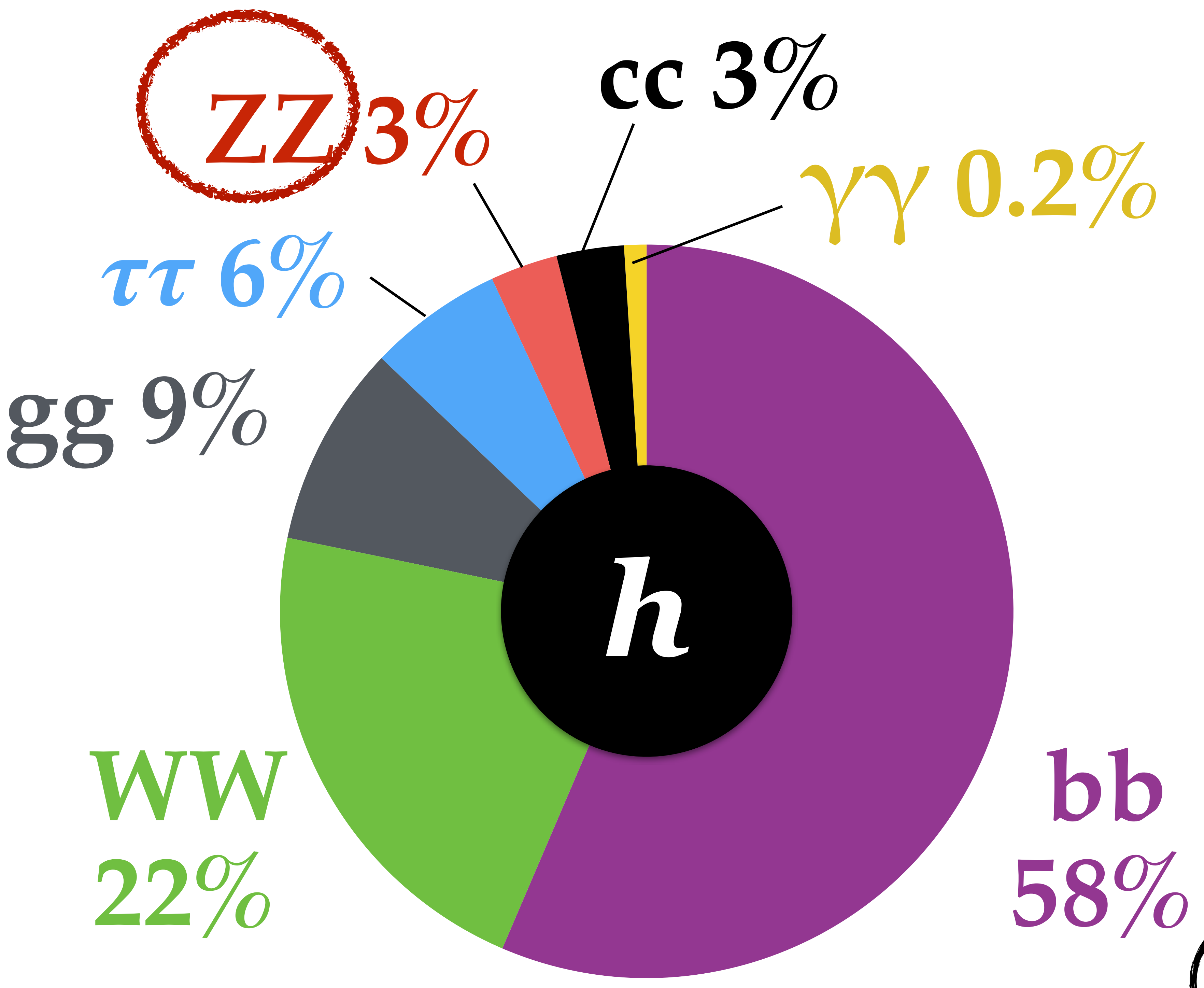
challenge: build the detectors!

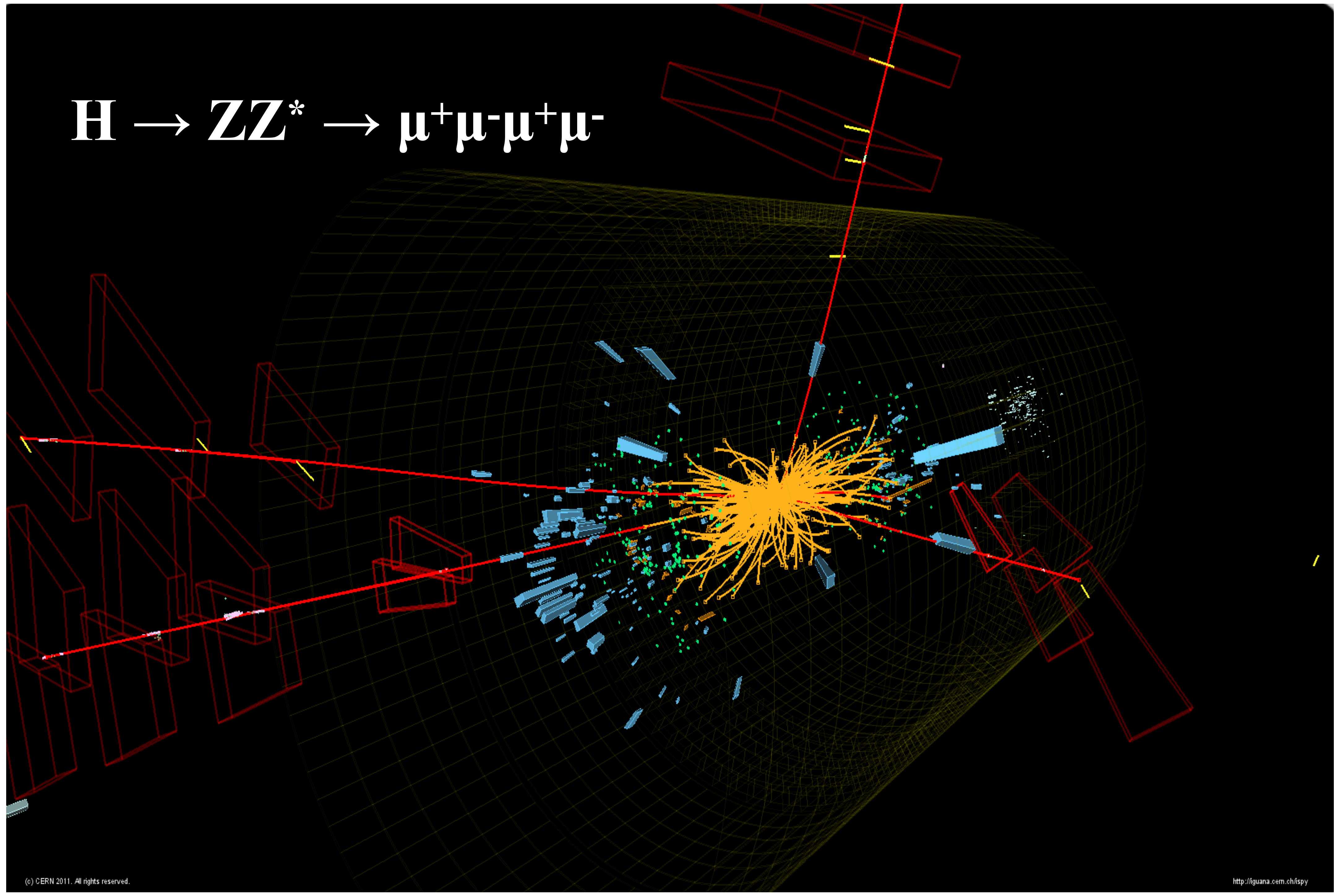


- CMS Phase 0 detector designed for LHC nominal luminosity
  - Tracking efficiency drops to 80% at PU=40
- Phase 1 pixel : designed for LHC Run 2 & Run 3 data-taking(300 fb^{-1}) until HL-LHC upgrade (2024).
  - Improved module design, geometry, material budget.

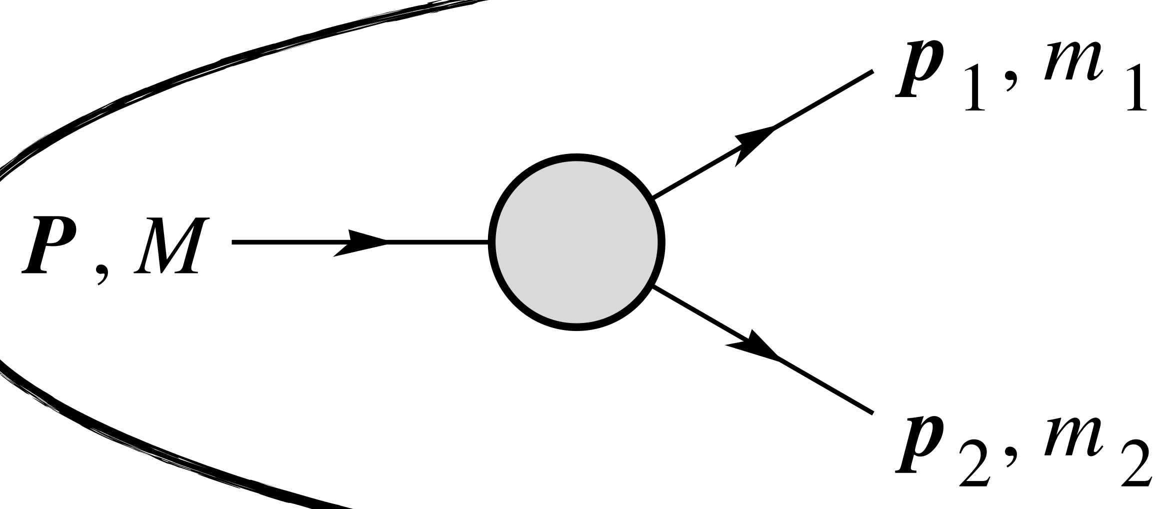
Higgs boson detection in CMS

Higgs boson decays





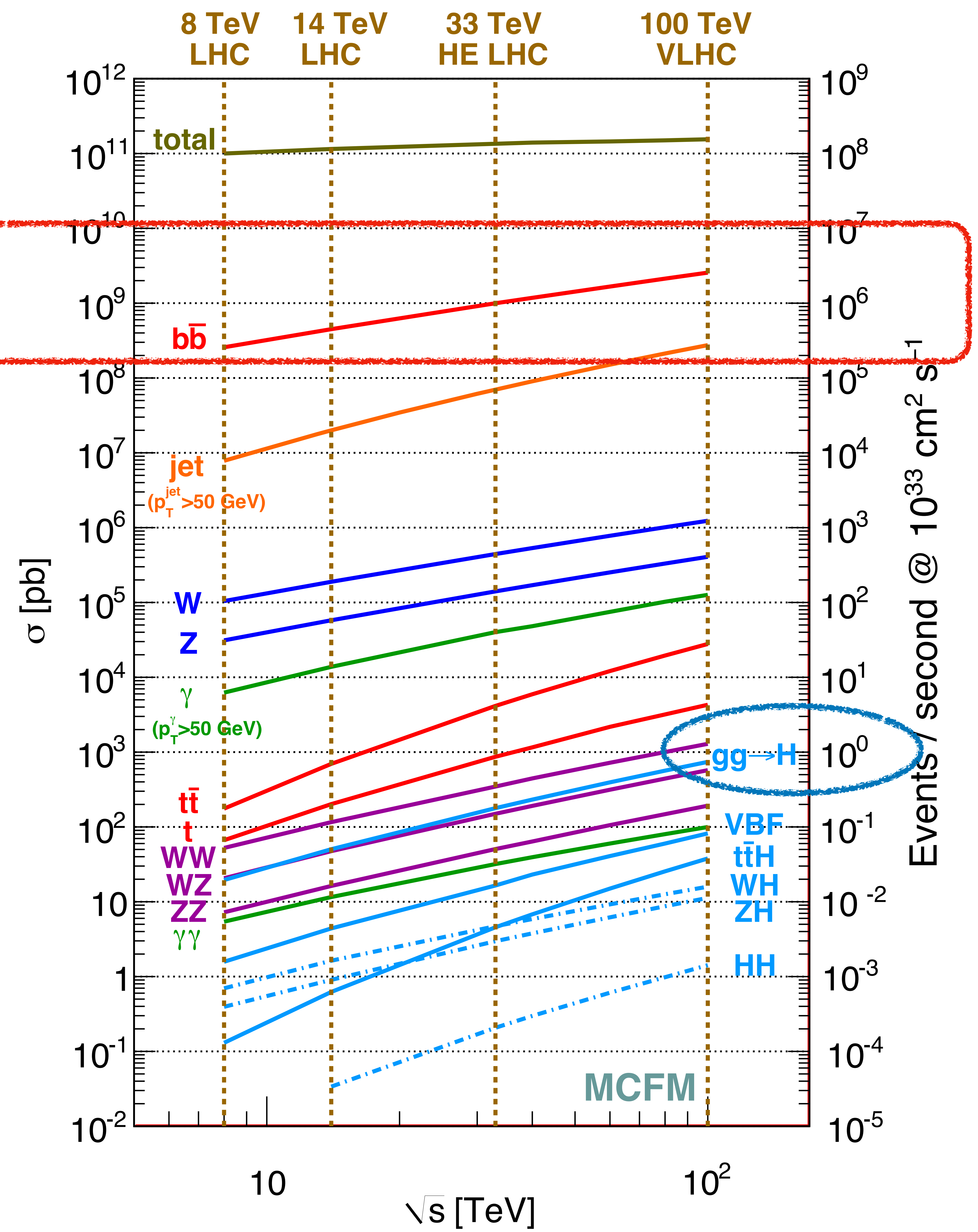
Construct invariant mass from decay products

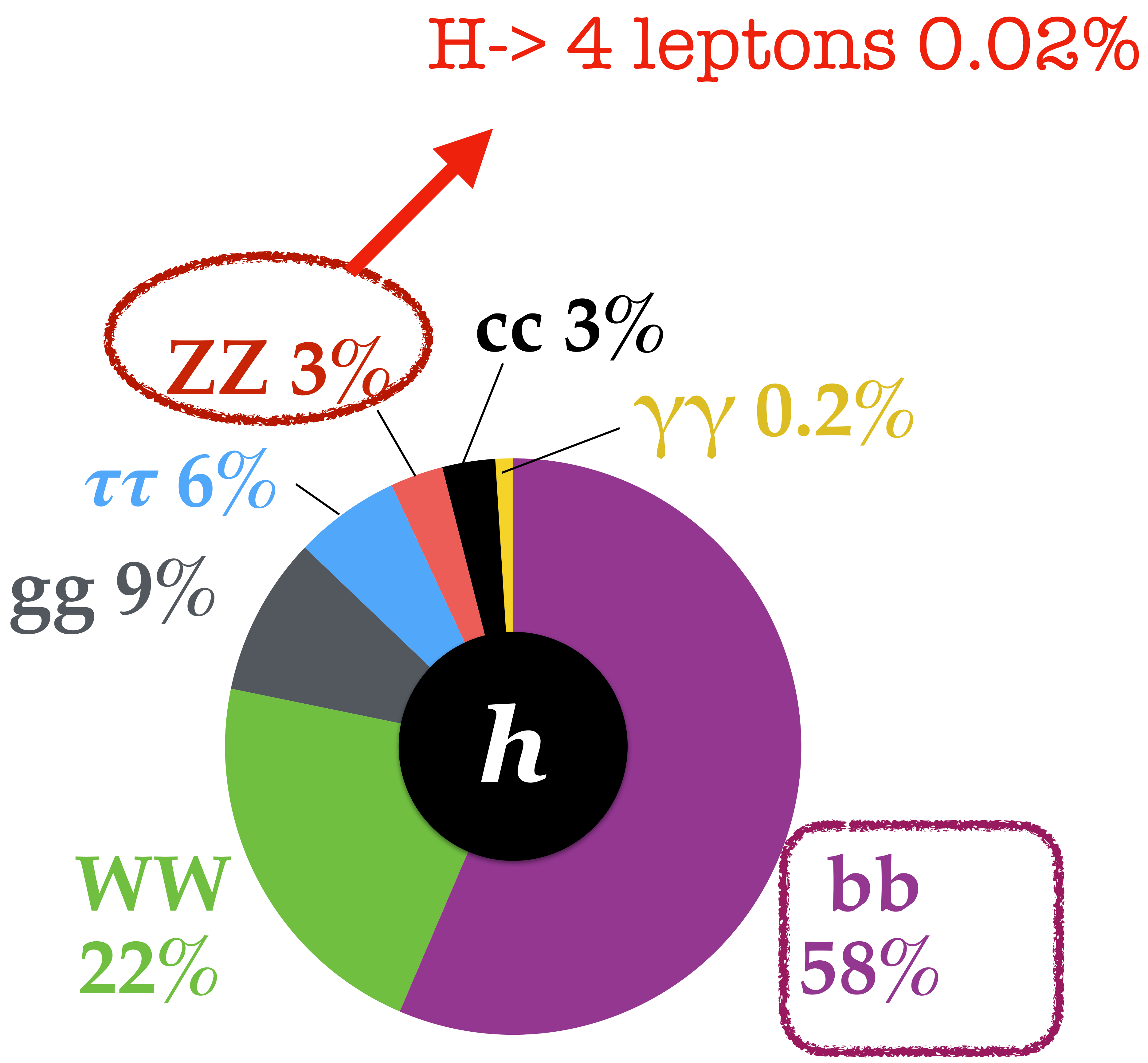


$$E_{\mathrm{cm}} = \left[(E_1 + E_2)^2 - (\boldsymbol{p}_1 + \boldsymbol{p}_2)^2\right]^{1/2}$$

Signal vs background

M.LIU





Higgs boson discovery: decay modes of lower backgrounds (WW/ZZ/γγ).

Re-train Res-Net 50 to tag top jets

Quantized model Brainwave's implementation of ResNet50 on FPGA

State of art performance achieved with quantized ResNet 50 on BrainWave service