UNIVERSITY OF MICHIGAN

Current applications of ML in ASICs and opportunities / challenges for future facilities

Christian Herwig
University of Michigan
ML4FE workshop, University of Hawai'i Manoa
May 19, 2025

Why ASICs?

1. …

Why not ASICs?

Are we sure this is a good idea….??

Why not ASICs?

Are we sure this is a good idea….??

1. **No room for mistakes!**
   - Additional complexity increases the chance to corrupt data, brick chip…
   - Should one design in a “fall-back” plan?

Why not ASICs?

Are we sure this is a good idea….??

1. **No room for mistakes!**
   - Additional complexity increases the chance to corrupt data, brick chip…
   - Should one design in a “fall-back” plan?
2. Environment introduces many **new constraints.**
   - Power, size, potentially radiation-hardness.

Why not ASICs?

Are we sure this is a good idea….??

1. **No room for mistakes!**
   - Additional complexity increases the chance to corrupt data, brick chip…
   - Should one design in a “fall-back” plan?
2. Environment introduces many **new constraints.**
   - Power, size, potentially radiation-hardness.
3. Requires a **specialized knowledge and tools** for circuit design.
   - Higher barrier (resources, expertise) to prototype new efforts.

Why not ASICs?

Are we sure this is a good idea….??

1. **No room for mistakes!**
   - Additional complexity increases the chance to corrupt data, brick chip…
   - Should one design in a “fall-back” plan?
2. Environment introduces many **new constraints.**
   - Power, size, potentially radiation-hardness.
3. Requires a **specialized knowledge and tools** for circuit design.
   - Higher barrier (resources, expertise) to prototype new efforts.
4. Challenging to complete a **full design verification** before fab.
   - How to achieve full coverage of all paths for all ML algo configs?

Why not ASICs?

Are we sure this is a good idea….??

1. **No room for mistakes!**
   - Additional complexity increases the chance to corrupt data, brick chip…
   - Should one design in a “fall-back” plan?
2. Environment introduces many **new constraints.**
   - Power, size, potentially radiation-hardness.
3. Requires a **specialized knowledge and tools** for circuit design.
   - Higher barrier (resources, expertise) to prototype new efforts.
4. Challenging to complete a **full design verification** before fab.
   - How to achieve full coverage of all paths for all ML algo configs?
5. Cannot take advantage of **latest technology nodes.**
   - Versus the latest from AMD Xilinx + friends (never mind your iPhone)

Why ASICs?

1. **No other option** is capable of addressing our modern scientific data processing challenges.

FastML for Science

https://arxiv.org/abs/2207.07958

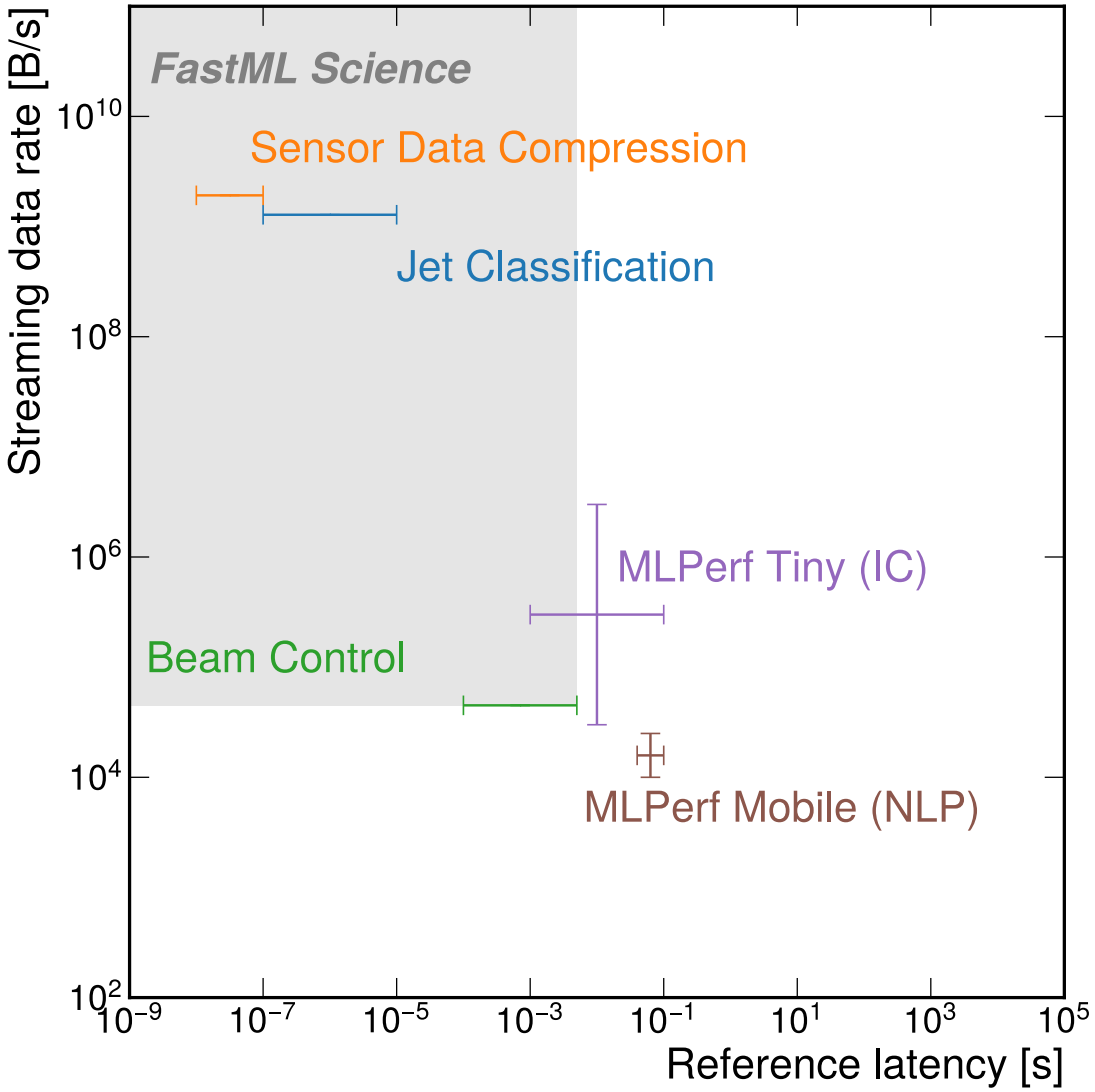
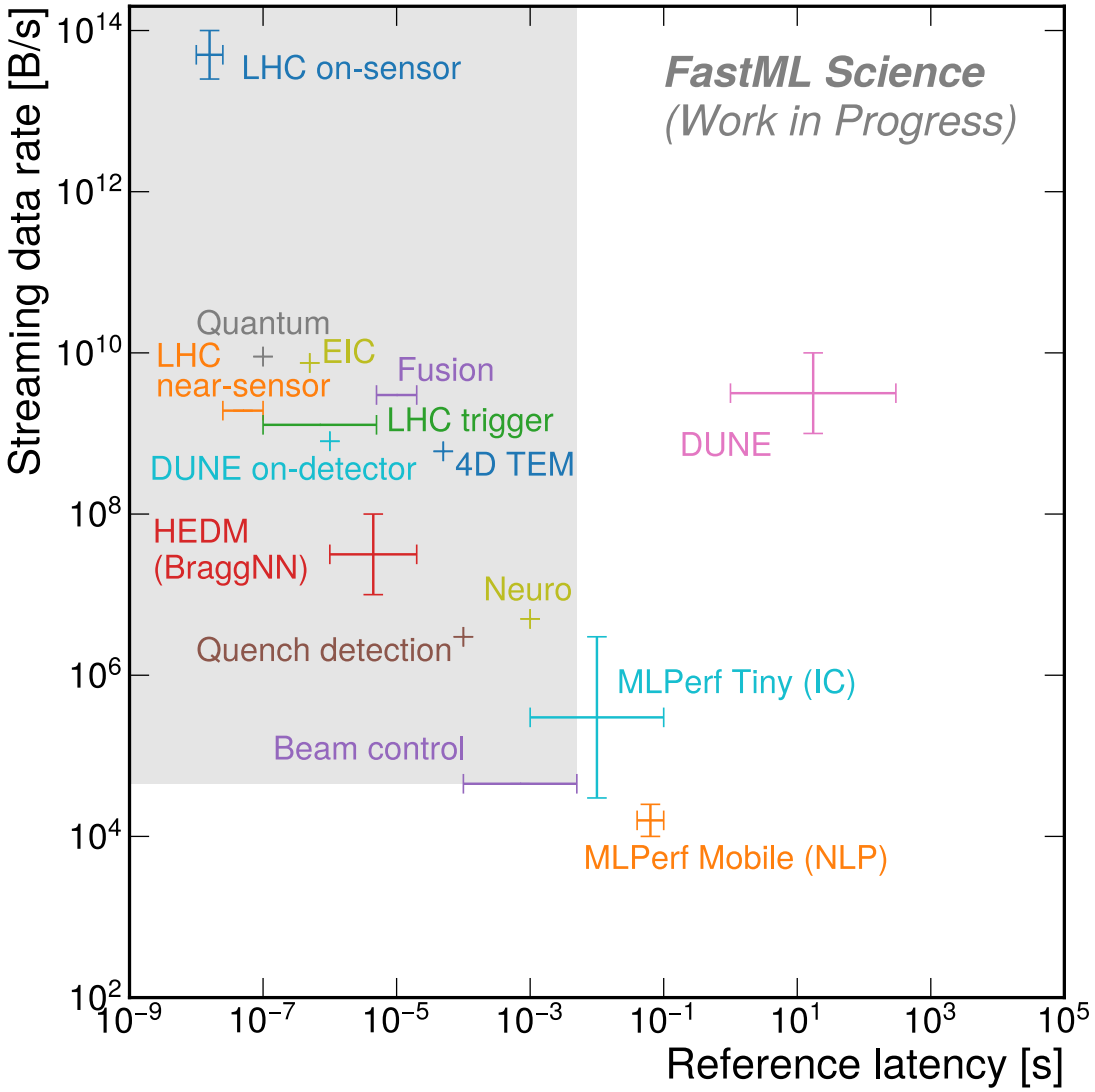


Figure: Reference latencies and streaming input data rates for common industry benchmarks and FastML Science.

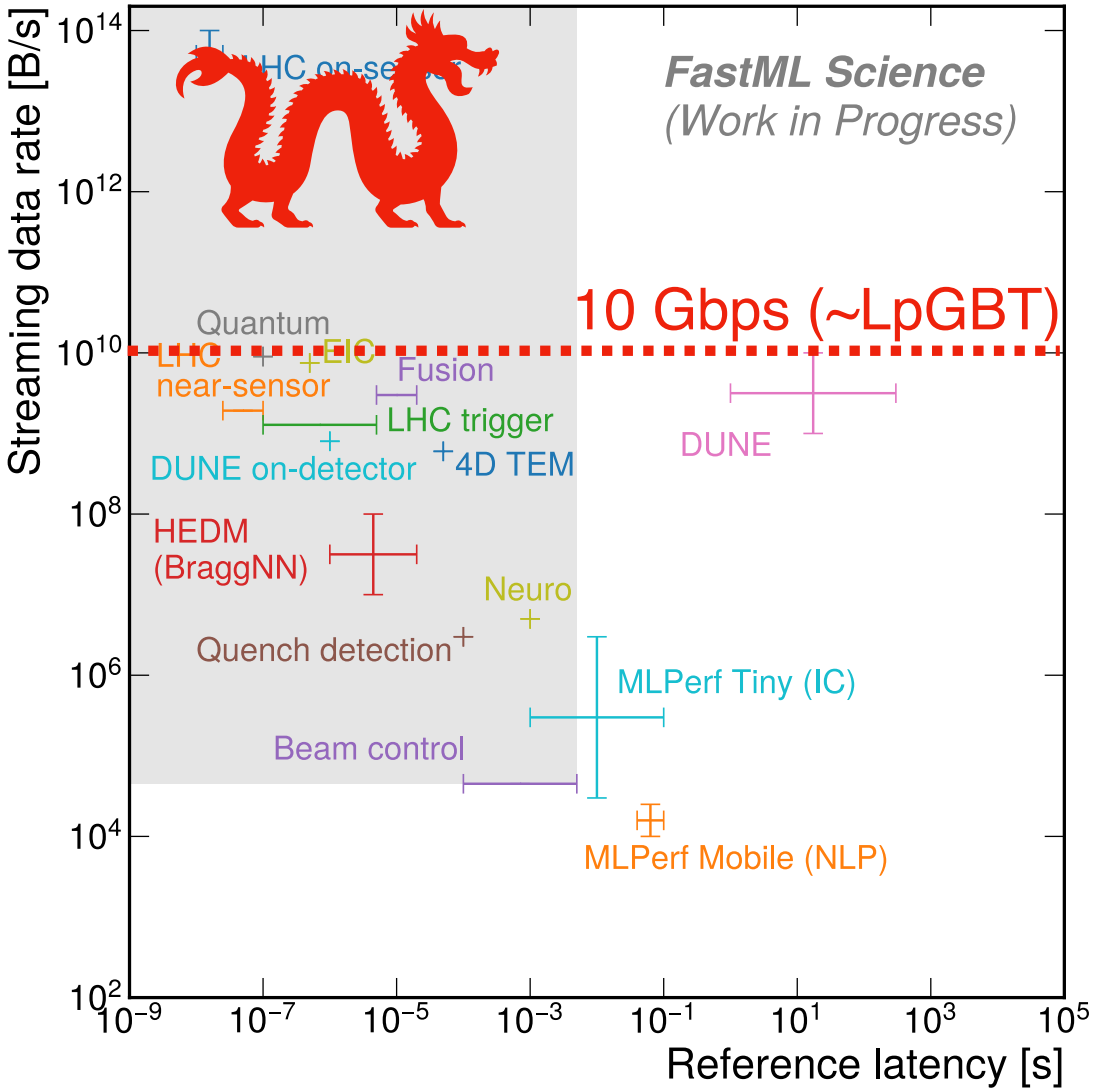
- Collection of benchmark tasks for FastML for Science
- Spans many scientific domains and task categories
  - Supervised ML
  - Unsupervised ML
  - RL for realtime control
- Aside: our tasks are not well-represented by industry-driven benchmarks like MLPerf!

FastML for Science



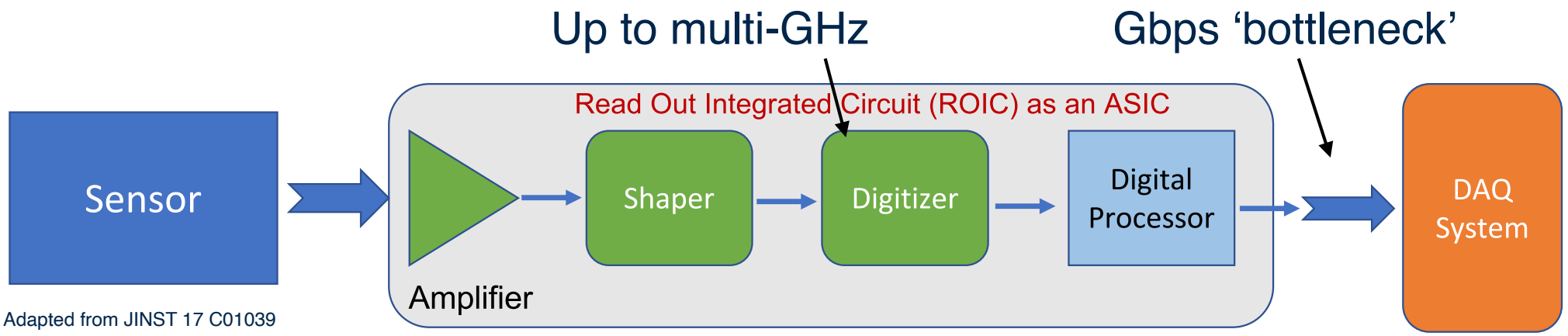
- Collection of benchmark tasks for FastML for Science
- Spans many scientific domains and task categories
  - Supervised ML
  - Unsupervised ML
  - RL for realtime control
- More folks got interested…

FastML for Science



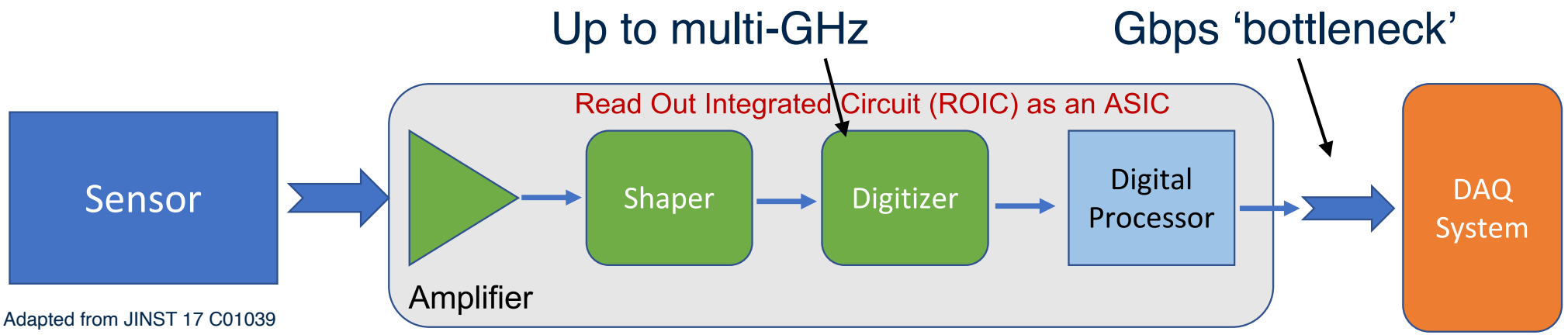
- Collection of benchmark tasks for FastML for Science
- Spans many scientific domains and task categories
  - Supervised ML
  - Unsupervised ML
  - RL for realtime control
- More folks got interested…

A typical readout chain



Adapted from JINST 17 C01039

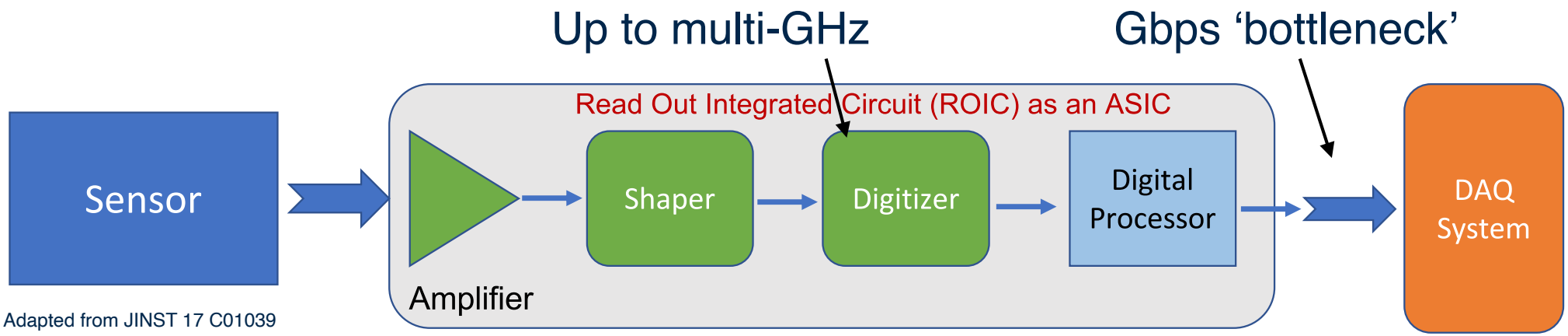
A typical readout chain



Adapted from JINST 17 C01039

A modest example: 100^2 array of 10b pixel data @MHz readout ⇒ 100 Gbps

A typical readout chain

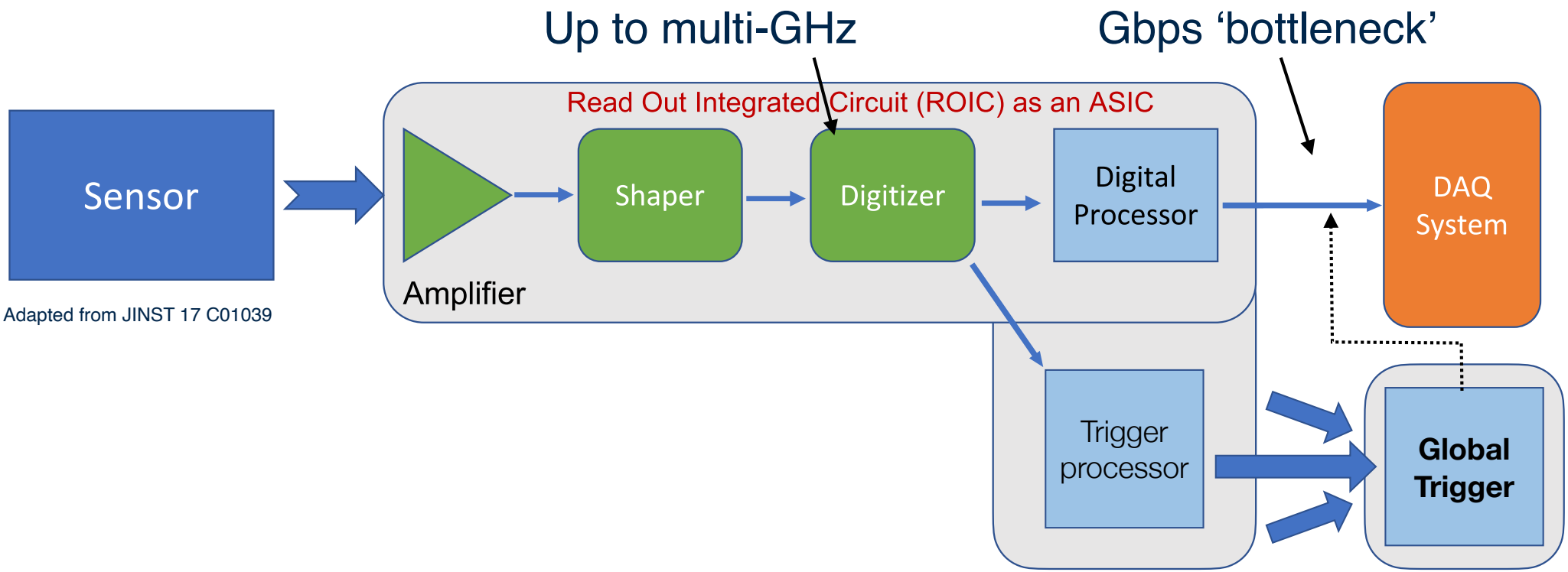


A modest example: 100^2 array of 10b pixel data @MHz readout ⇒ 100 Gbps

For complex sensor data, limited output bandwidth implies

1. Compromise on the data 'quality' or transmission rate (ext trigger),

A typical readout chain

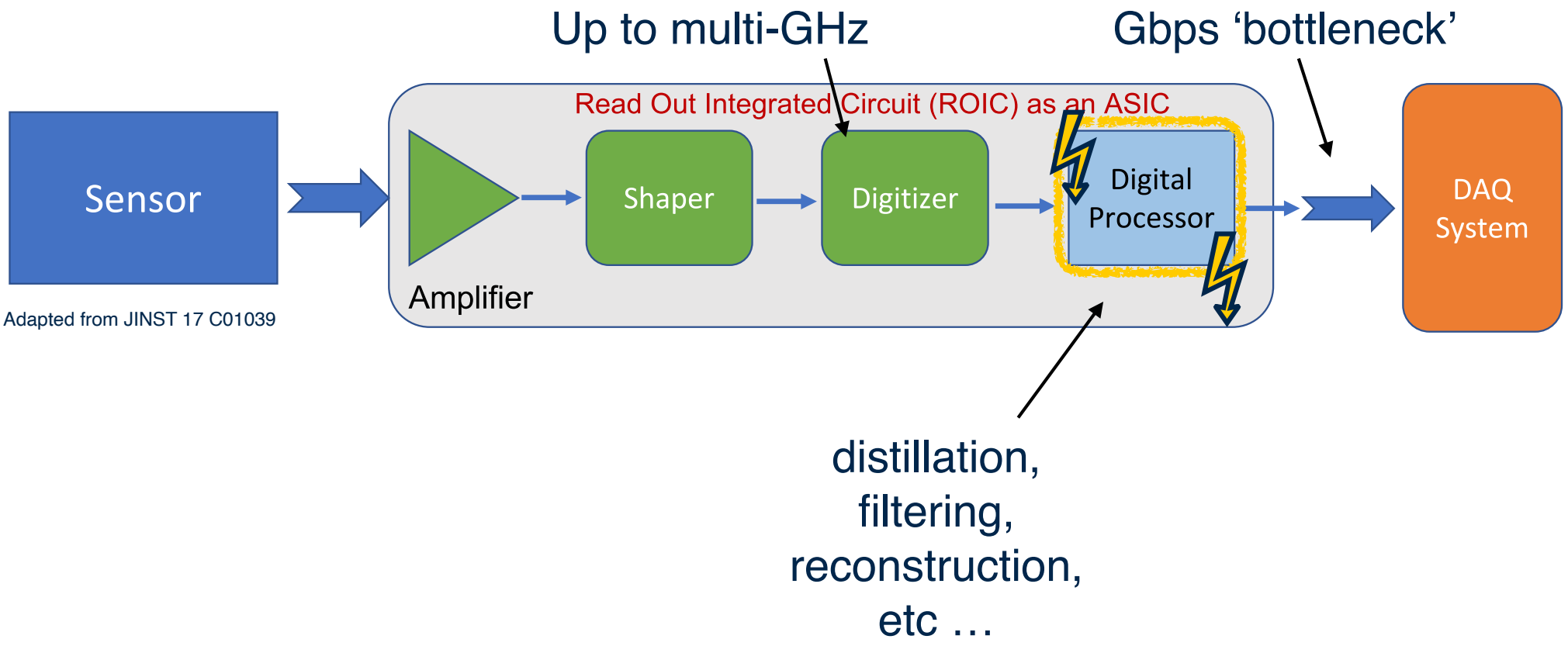


Adapted from JINST 17 C01039

For complex sensor data, limited output bandwidth implies

1. Compromise on the data 'quality' or transmission rate (ext trigger),

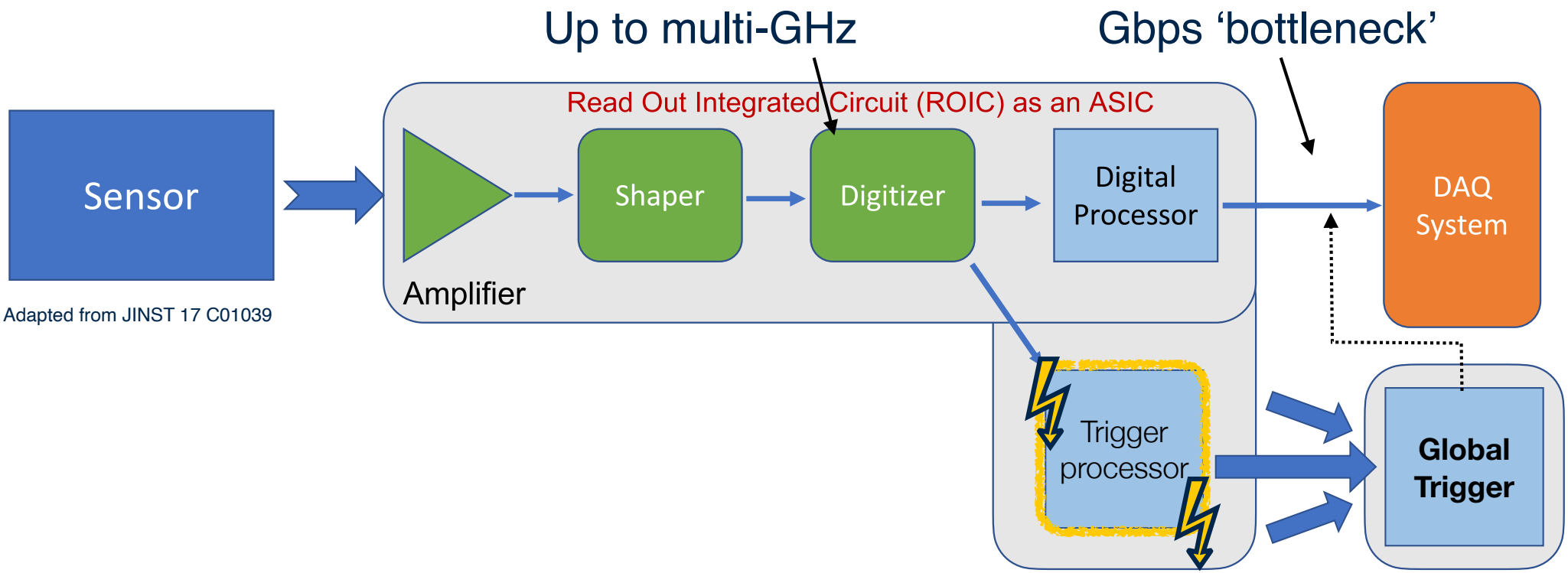
A typical readout chain



For complex sensor data, limited output bandwidth implies

1. Compromise on the data 'quality' or transmission rate (ext trigger), or
2. **On-device data reduction**.

A typical readout chain



Adapted from JINST 17 C01039

For complex sensor data, limited output bandwidth implies

1. Compromise on the data 'quality' or transmission rate (ext trigger), or
2. **On-device data reduction**.

A quick tour of past & future work

Applications to **digitization** and **data processing**:

- Waveform analysis for LGAD readout
- Data reduction for x-ray ptychography

Challenges at **near-future** particle detector experiments:

- High-granularity calorimetry
- Fast charged-particle tracking

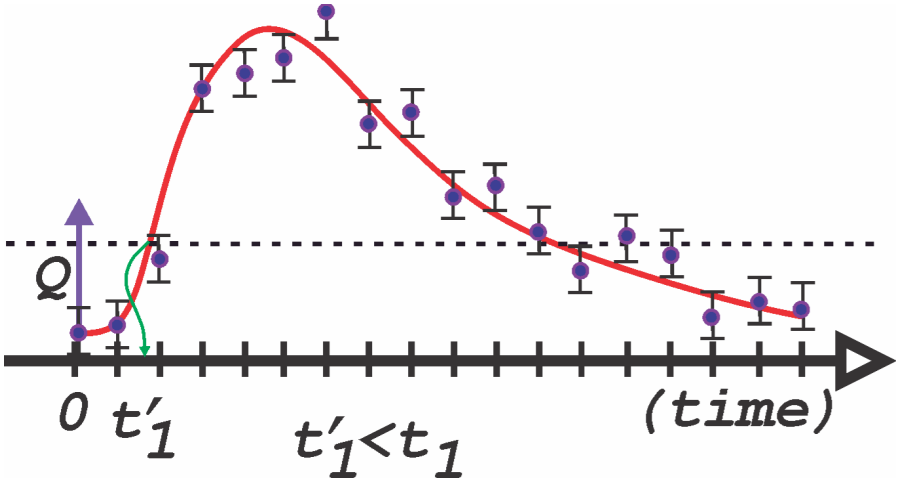
What challenges will next-gen experiments (pp, e^+e^-, μCol) introduce?

- Where might they most benefit from ML at the front end?

Stay tuned for much more information in the remainder of this workshop!

Waveform processing

ML could improve single-channel measurements already at the digi stage.



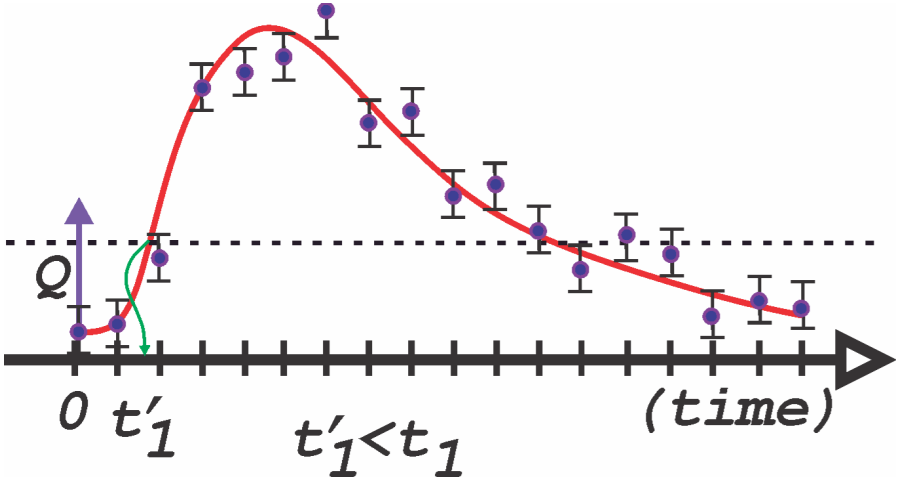
From **sparse samples** to:

∫charge, amplitude, t_0, noise, …

Miryala+, JINST 17 C01039
Miryala+, "Peak Prediction Using Multi Layer Perceptron (MLP) for Edge Computing ASICs Targeting Scientific Applications," 2022 23rd International Symposium on Quality Electronic Design (ISQED), 2022

Waveform processing

ML could improve single-channel measurements already at the digi stage.

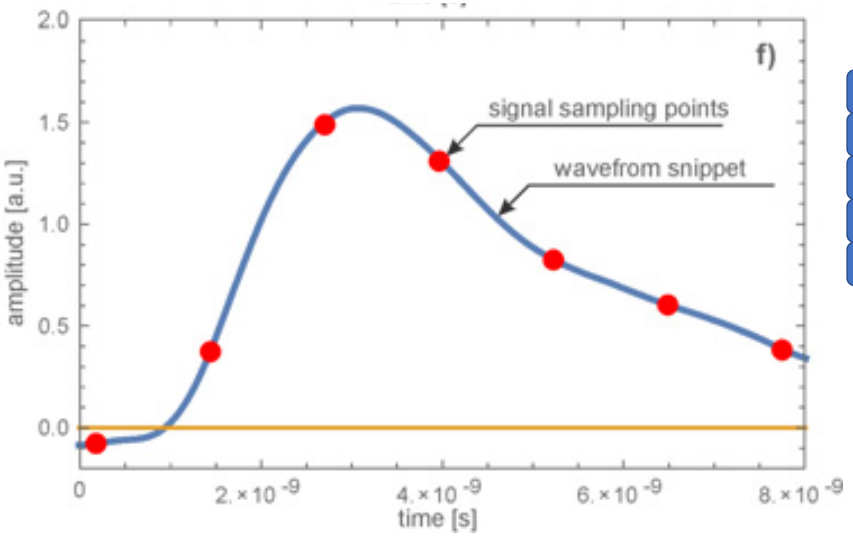


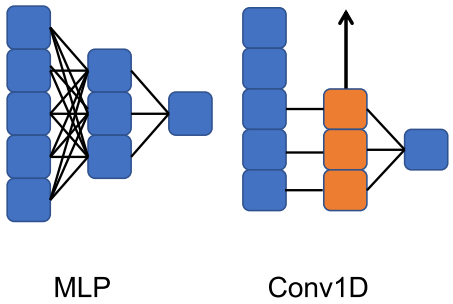
From **sparse samples** to:

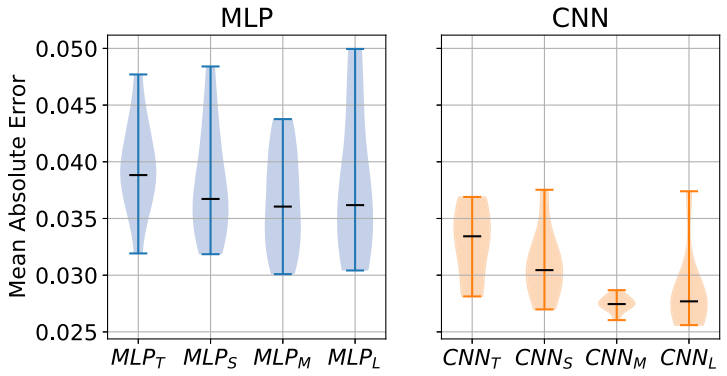
∫charge, amplitude, t_0, noise, …

Miryala+, JINST 17 C01039
Miryala+, "Peak Prediction Using Multi Layer Perceptron (MLP) for Edge Computing ASICs Targeting Scientific Applications," 2022 23rd International Symposium on Quality Electronic Design (ISQED), 2022

BNL group studied ps-level simulation of 50µm LGAD; 200x sub-samples.



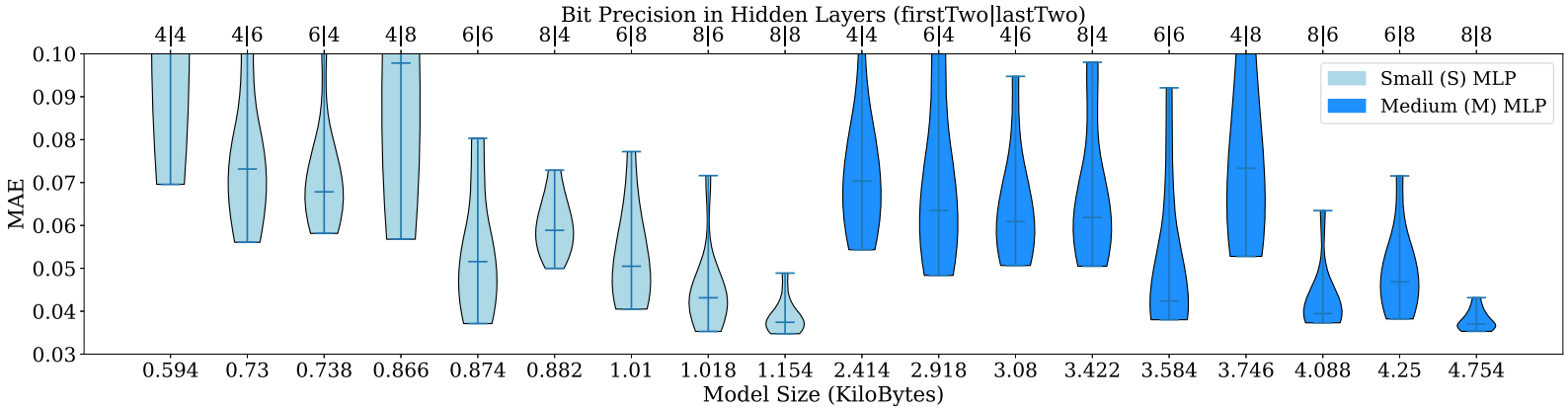




NN improve amplitudes by >4x (by MAE).

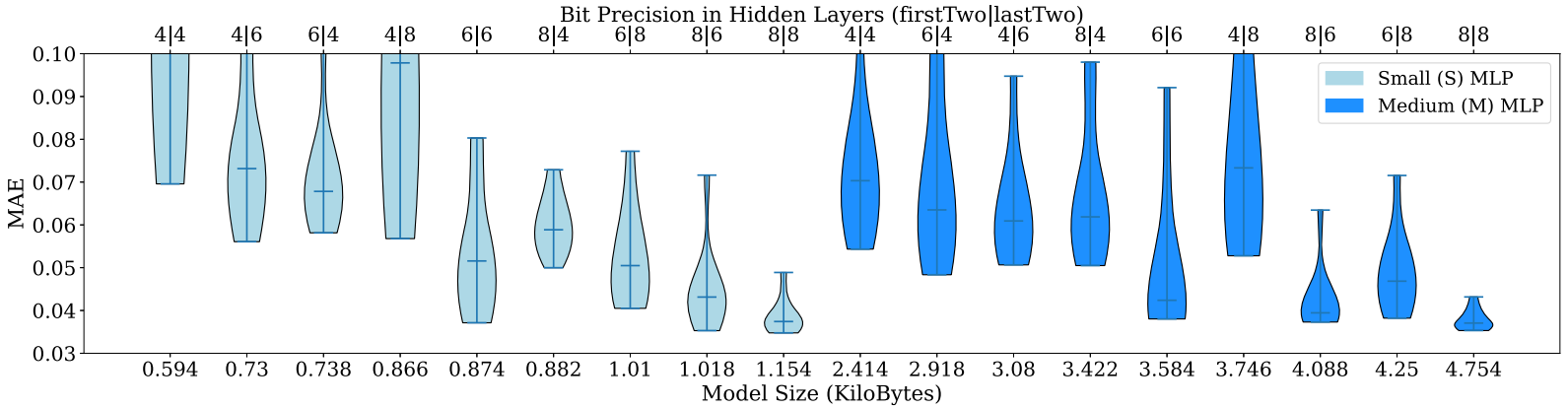
Waveform processing

Python: compare NN architectures, pruning, quantization, …

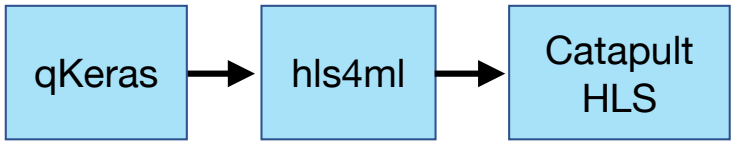


Waveform processing

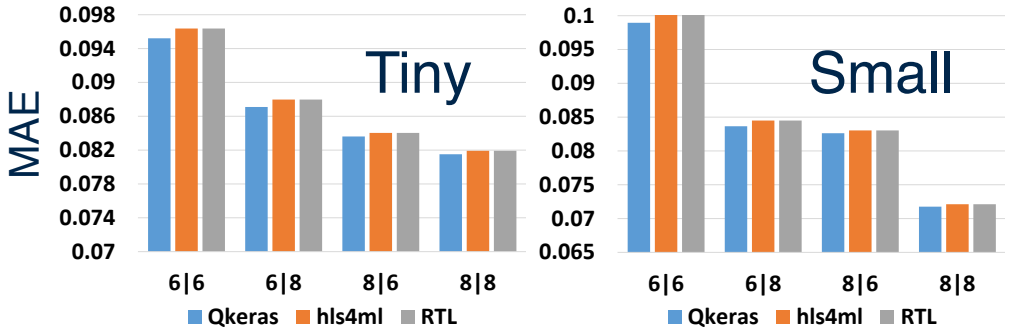
Python: compare NN architectures, pruning, quantization, …



Implemented in 65nm TSMC

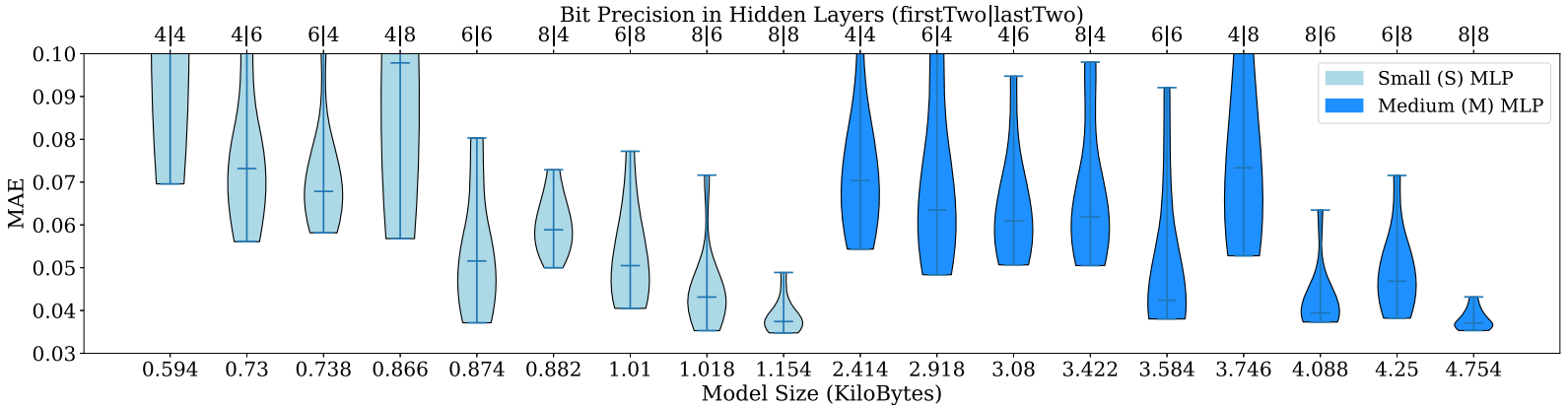


Good agreement found at each stage



Waveform processing

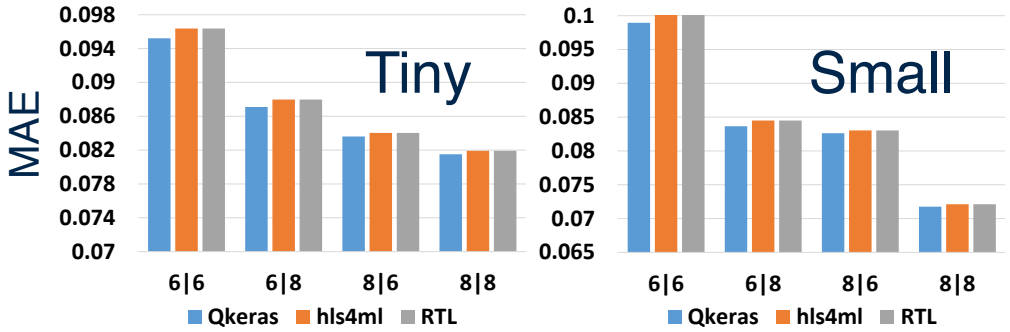
Python: compare NN architectures, pruning, quantization, …



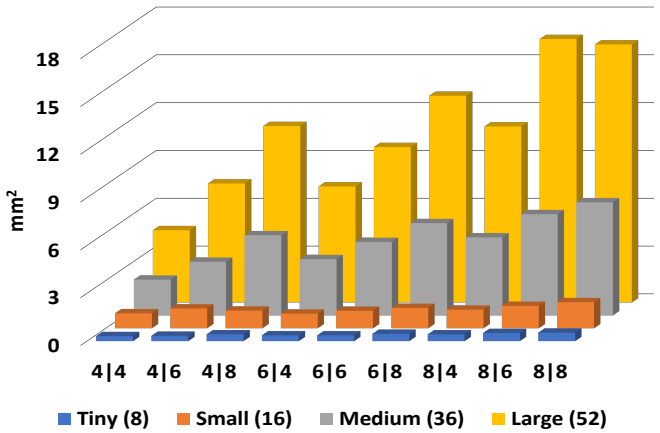
Implemented in 65nm TSMC

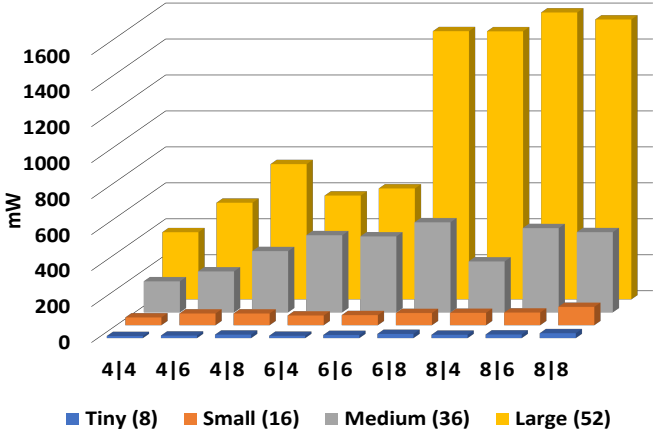
qKeras → hls4ml → Catapult HLS

Good agreement found at each stage

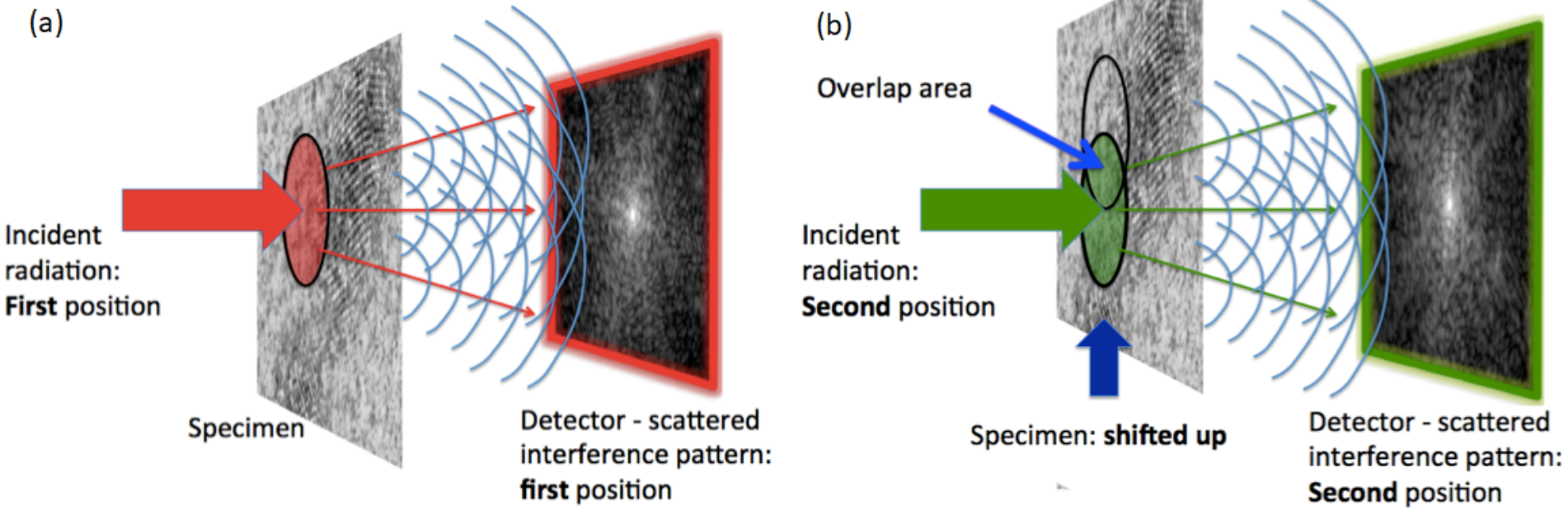


Tradeoff model performance for area, power consumption

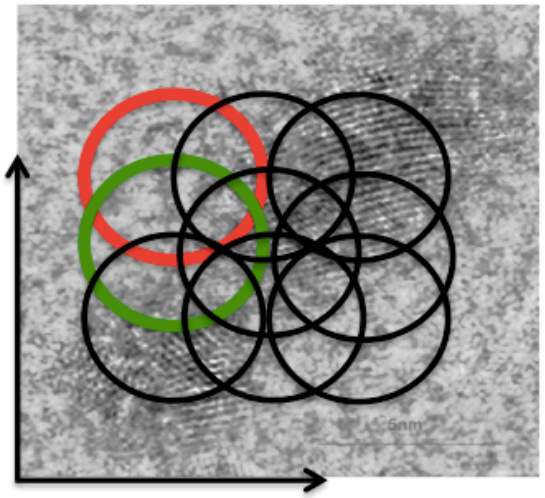




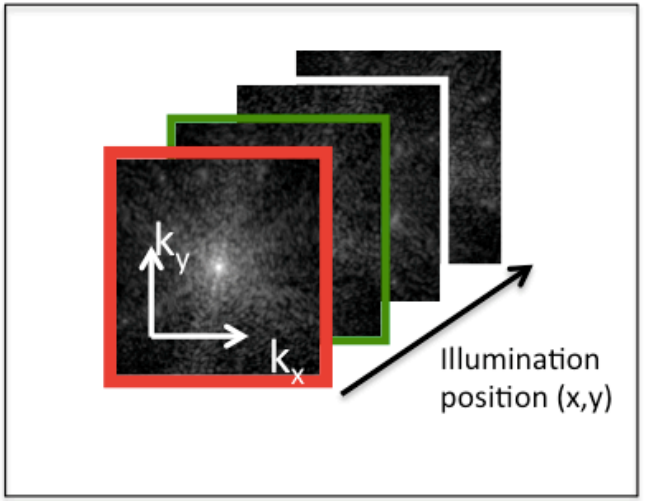
X-ray ptychography



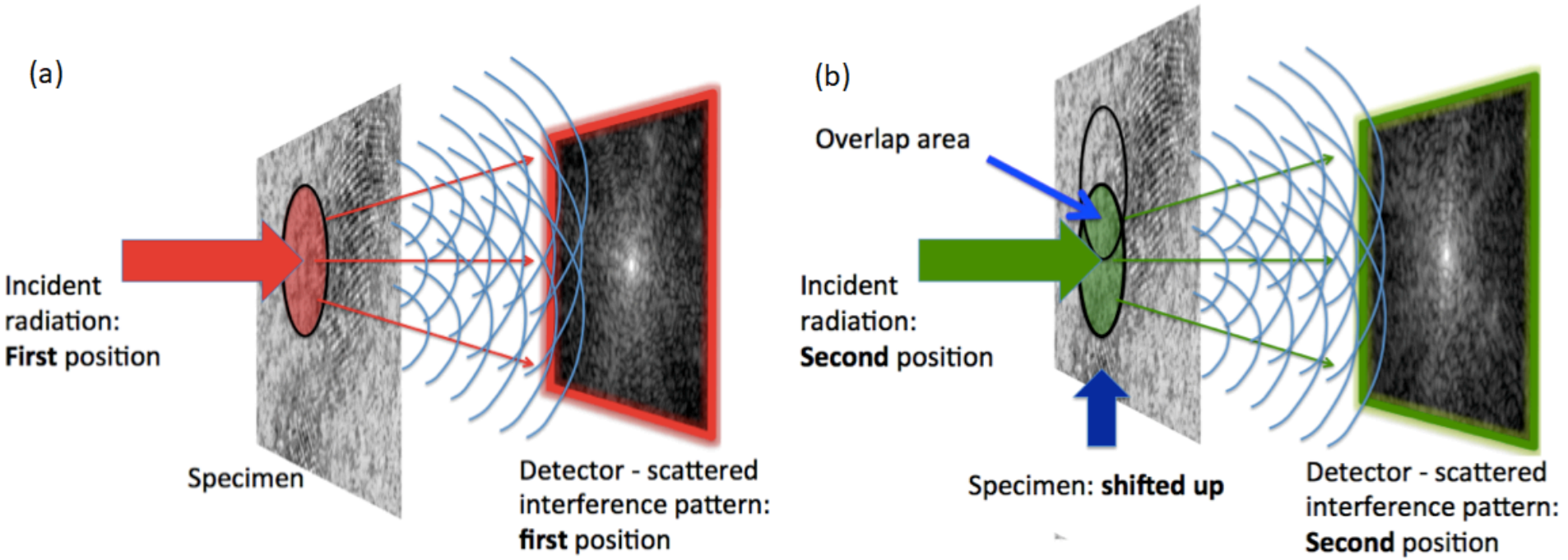
(c)



(d)



X-ray ptychography



Typical sensors: (few-hundred)^2 pixel array with 10-12b ADCs

Real data has high occupancy → zero suppression is ineffective.

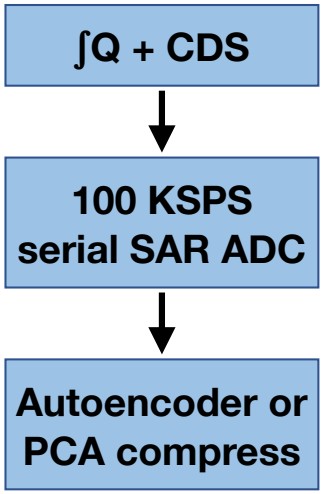
Readout limited: Data reduction of 50-100x needed to enable Mfps.

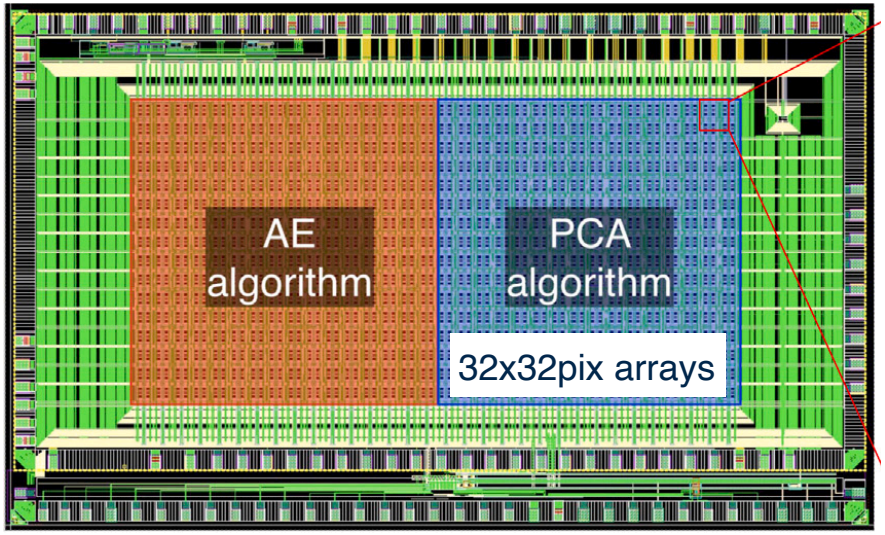
Enable quicker scanes / more samples (beam-time is in high demand!)

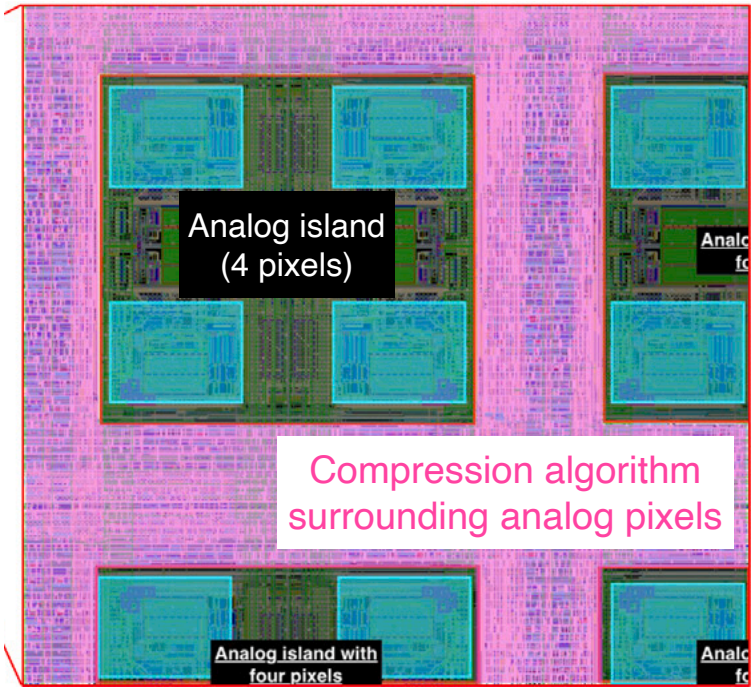
X-ray ptychography

Valentin+, NIM A 1057 (2023) 168665

AI-In-Pixel-65 (FNAL/Northwestern)
demonstrator in 65nm Low Power CMOS



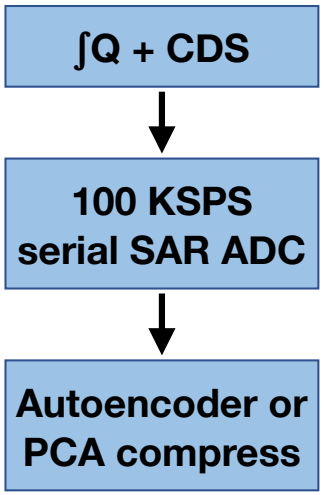


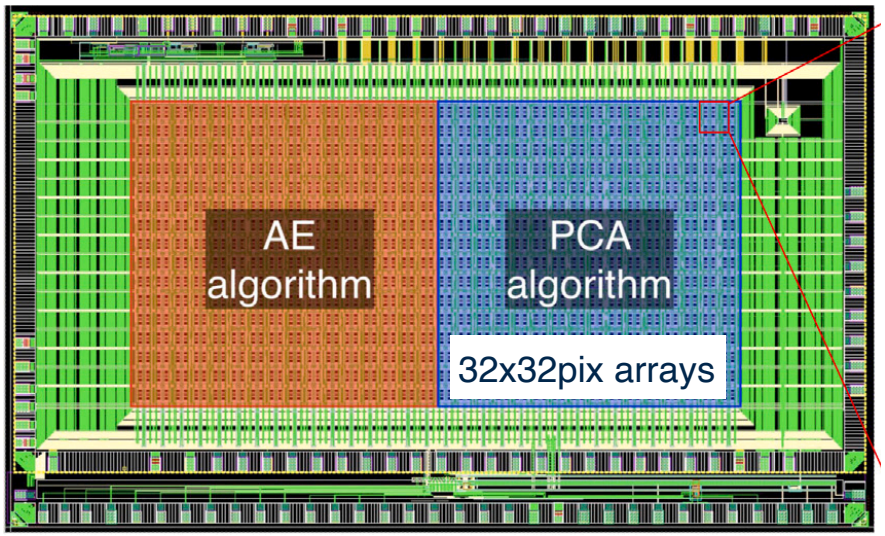


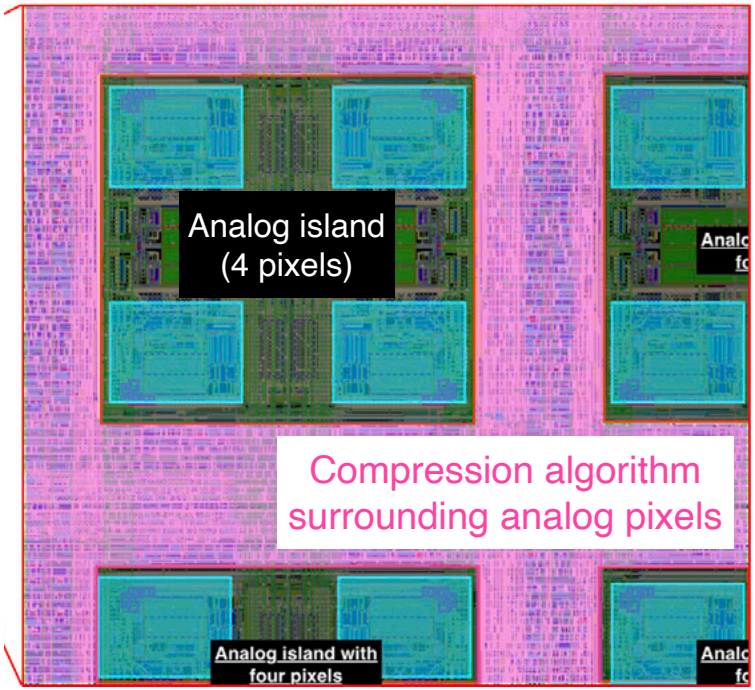
X-ray ptychography

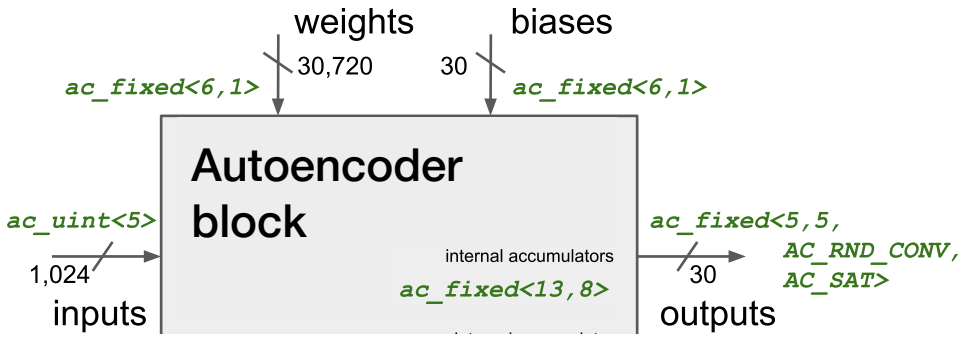
Valentin+, NIM A 1057 (2023) 168665

AI-In-Pixel-65 (FNAL/Northwestern)
demonstrator in 65nm Low Power CMOS





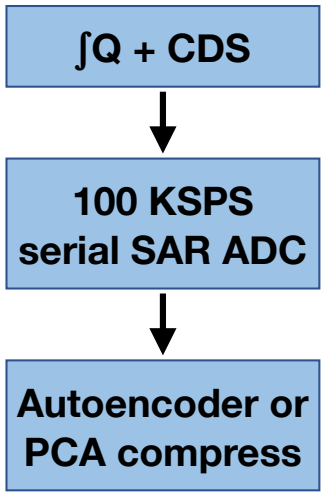


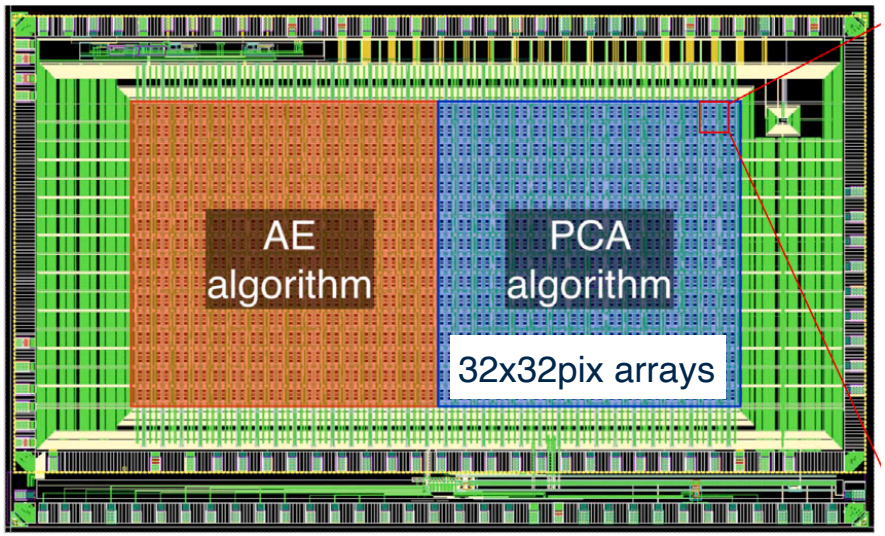


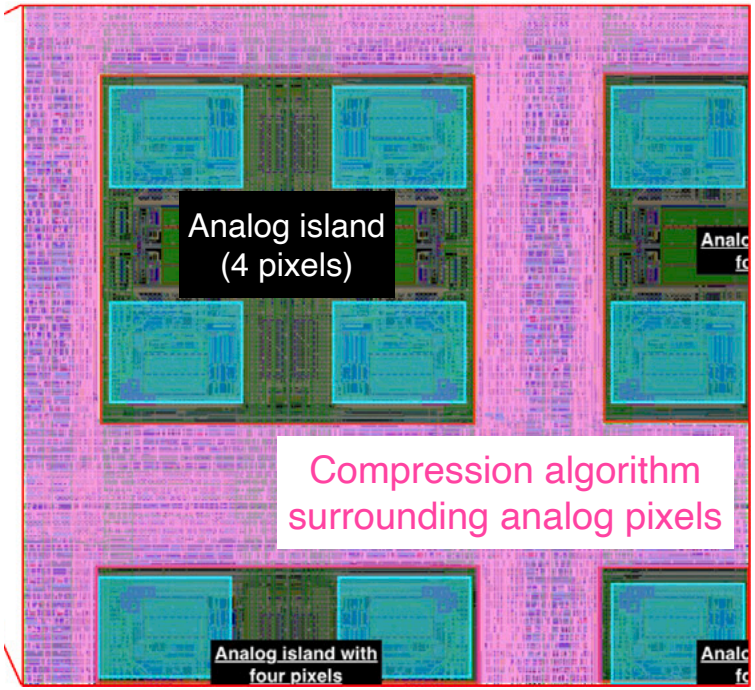
X-ray ptychography

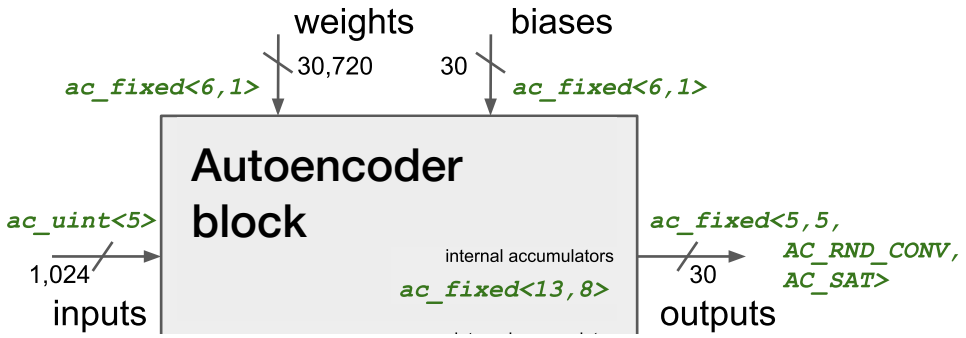
Valentin+, NIM A 1057 (2023) 168665

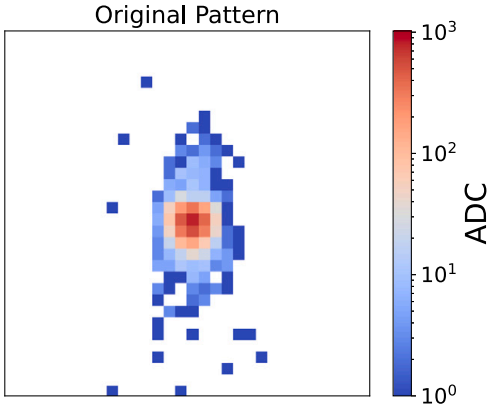
AI-In-Pixel-65 (FNAL/Northwestern)
demonstrator in 65nm Low Power CMOS

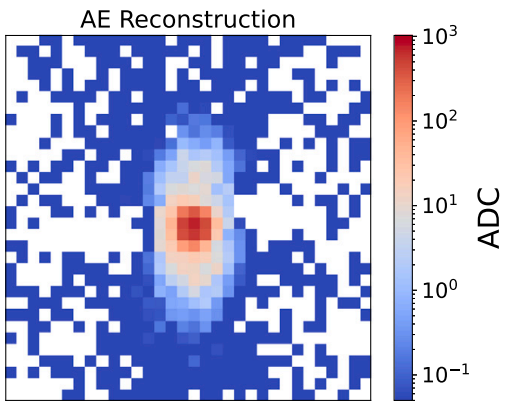












Tool flow: qKeras → hls4ml → Catapult HLS

HLS Solution	Latency	Area
modular	30	0.549
inlined	1	1.700

~20% of total chip area

Looking to the (near) future

Next-generation particle detectors will generate data at PB/s.

With high-luminosity LHC detectors, we are ~already there.

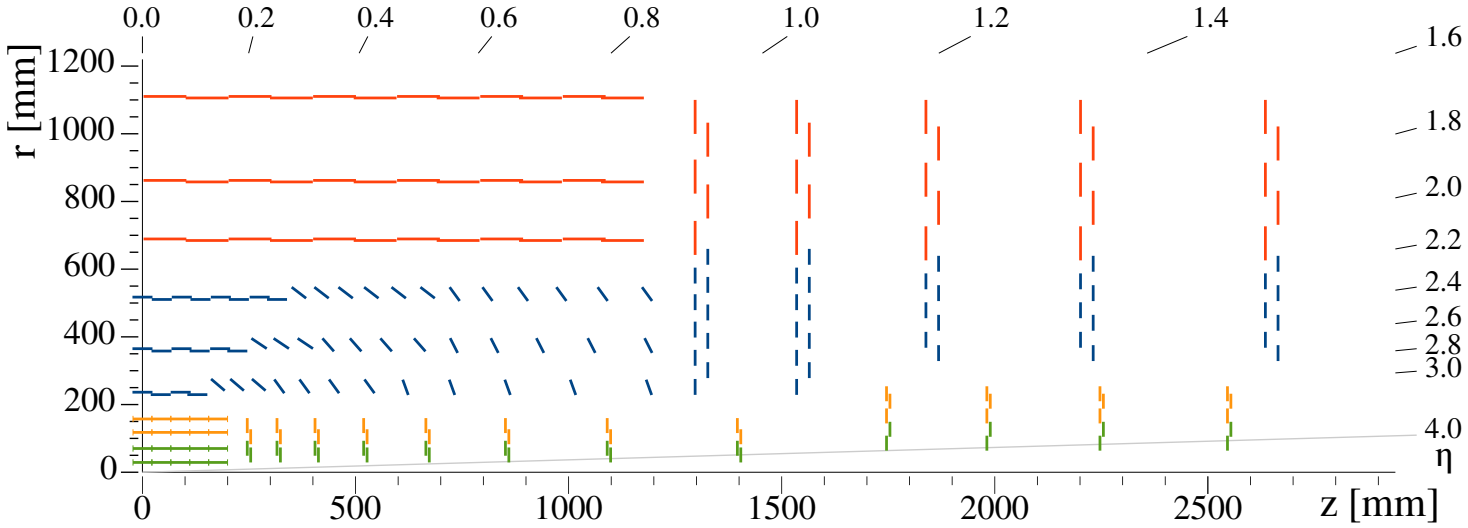
Looking to the (near) future

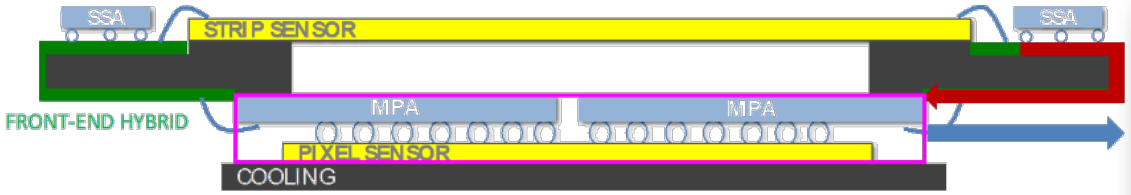
Next-generation particle detectors will generate data at PB/s.

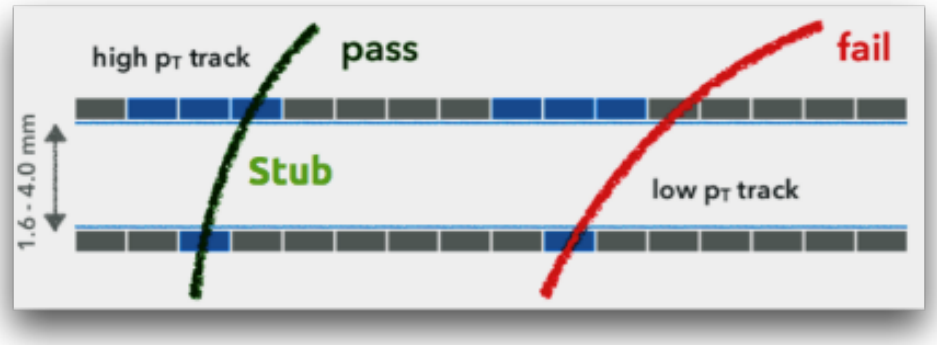
With high-luminosity LHC detectors, we are ~already there.

CMS outer tracker:
9.6M→213M channels
(Partial 40 MHz)

Inner tracker:
124M→1.9B channels
(.75 MHz readout)







'On-detector intelligence' filters hits from straight tracks (high-p_T)

Looking to the (near) future

Next-generation particle detectors will generate data at PB/s.

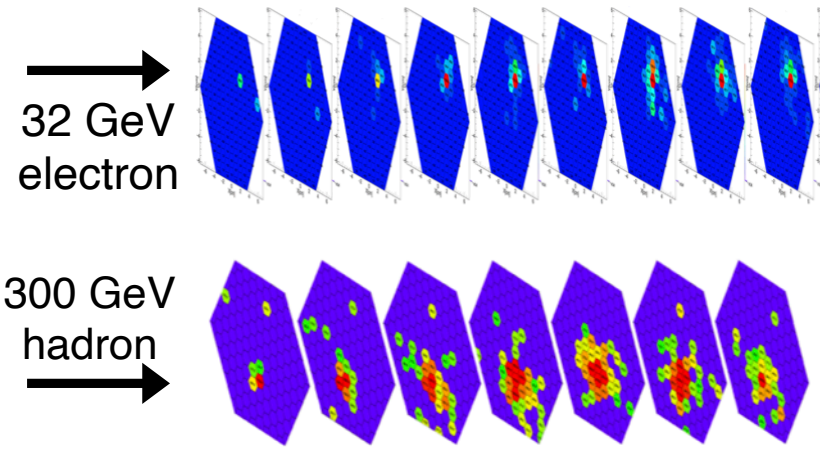
With high-luminosity LHC detectors, we are ~already there.

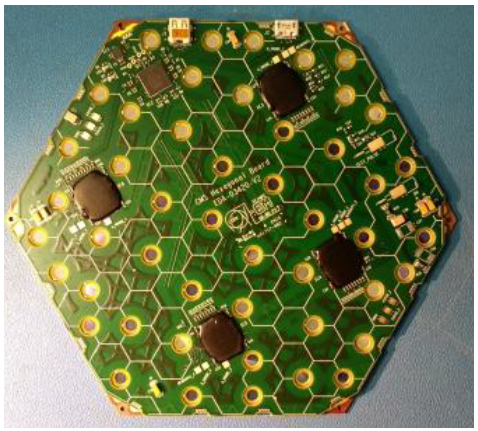
High-granularity endcap calorimeter (CMS HGCal)

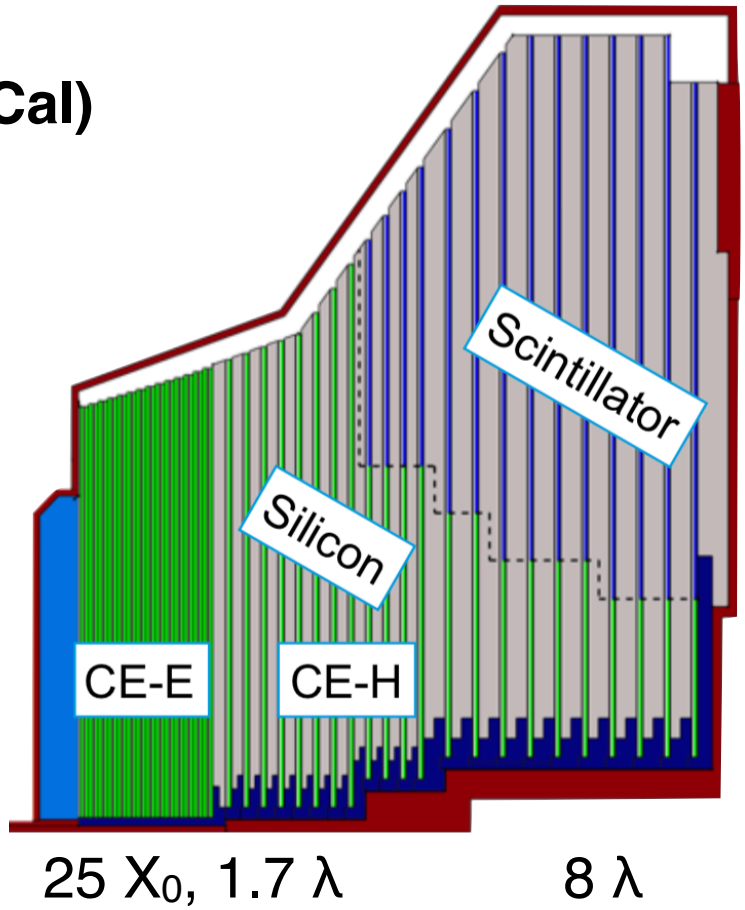
"5D Imaging calorimeter"

6.3M channels, 0.5-2 cm^2, σ_t=50ps (was 50k)

High dynamic range (1-100k MIPs)



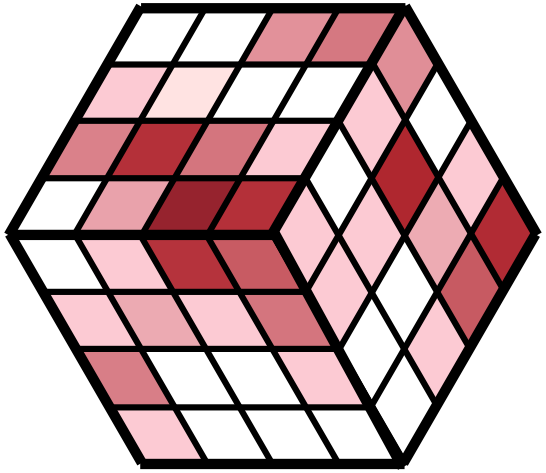




Data challenge in an ultra-high radiation environment!

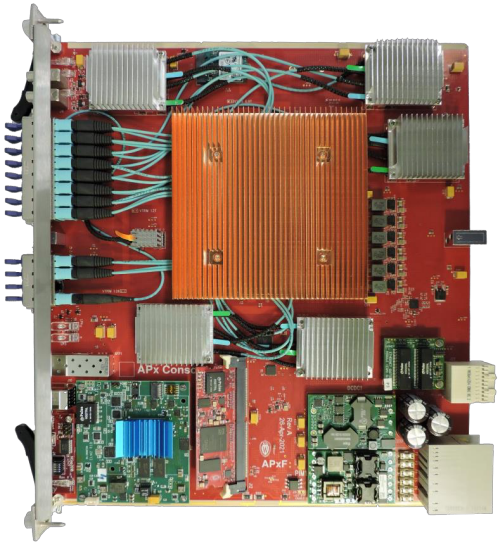
CMS HGCal data concentrator

High-granularity detector data

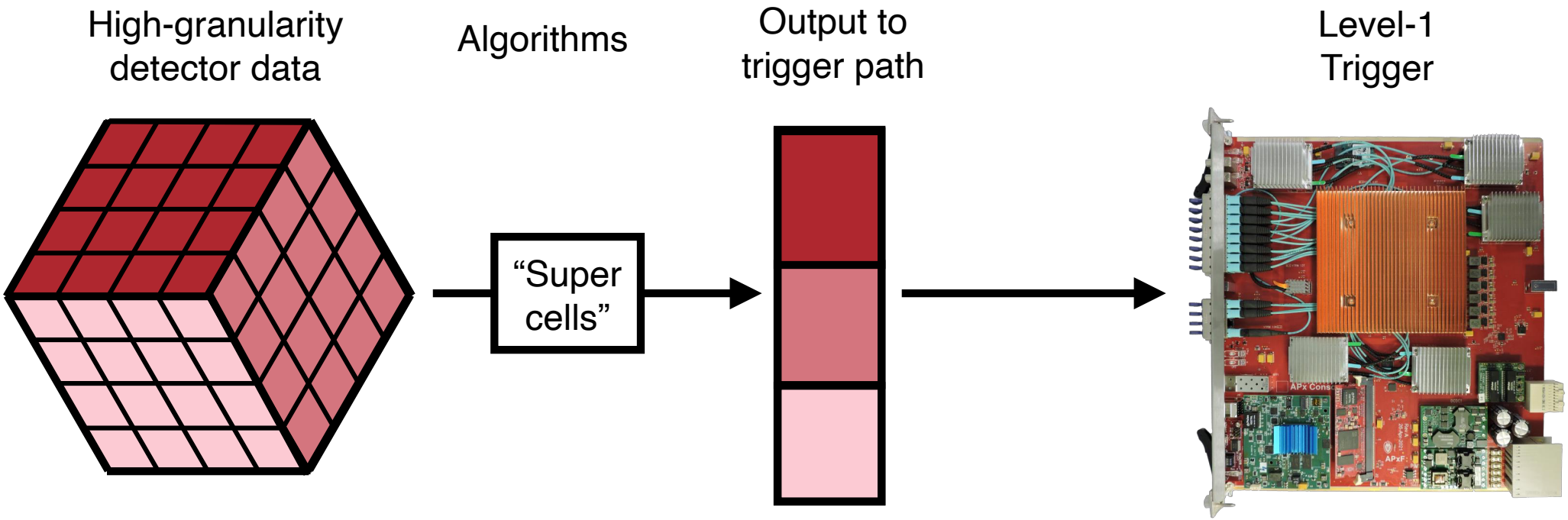


Output to trigger path

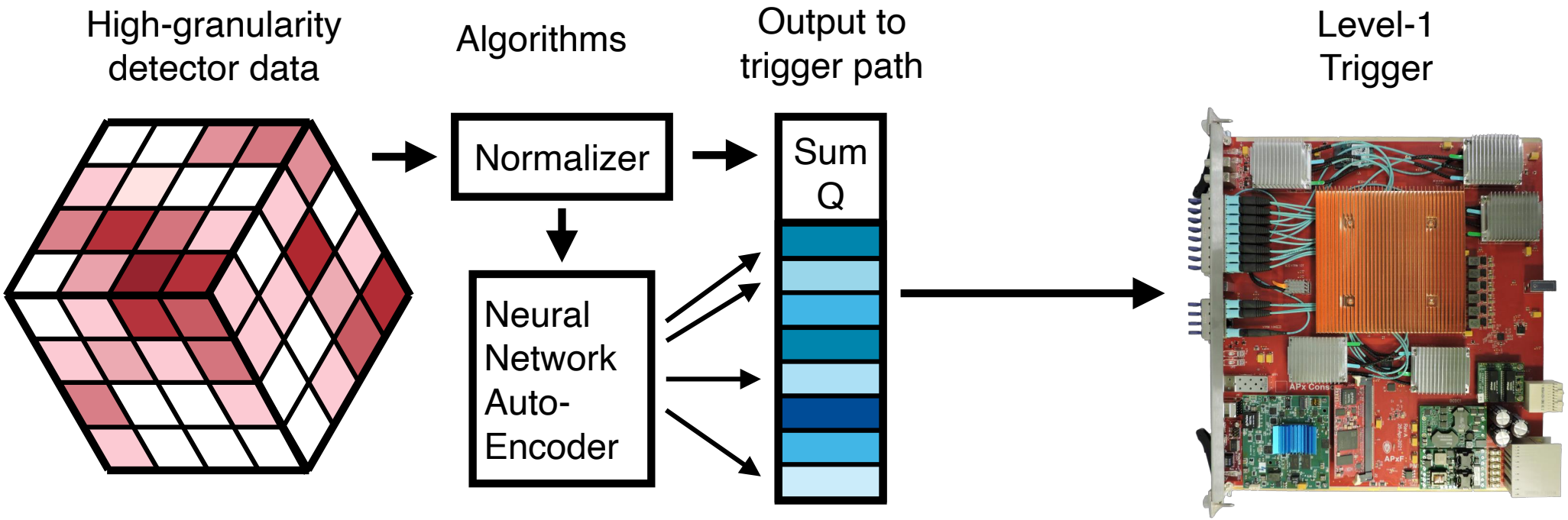
Level-1 Trigger



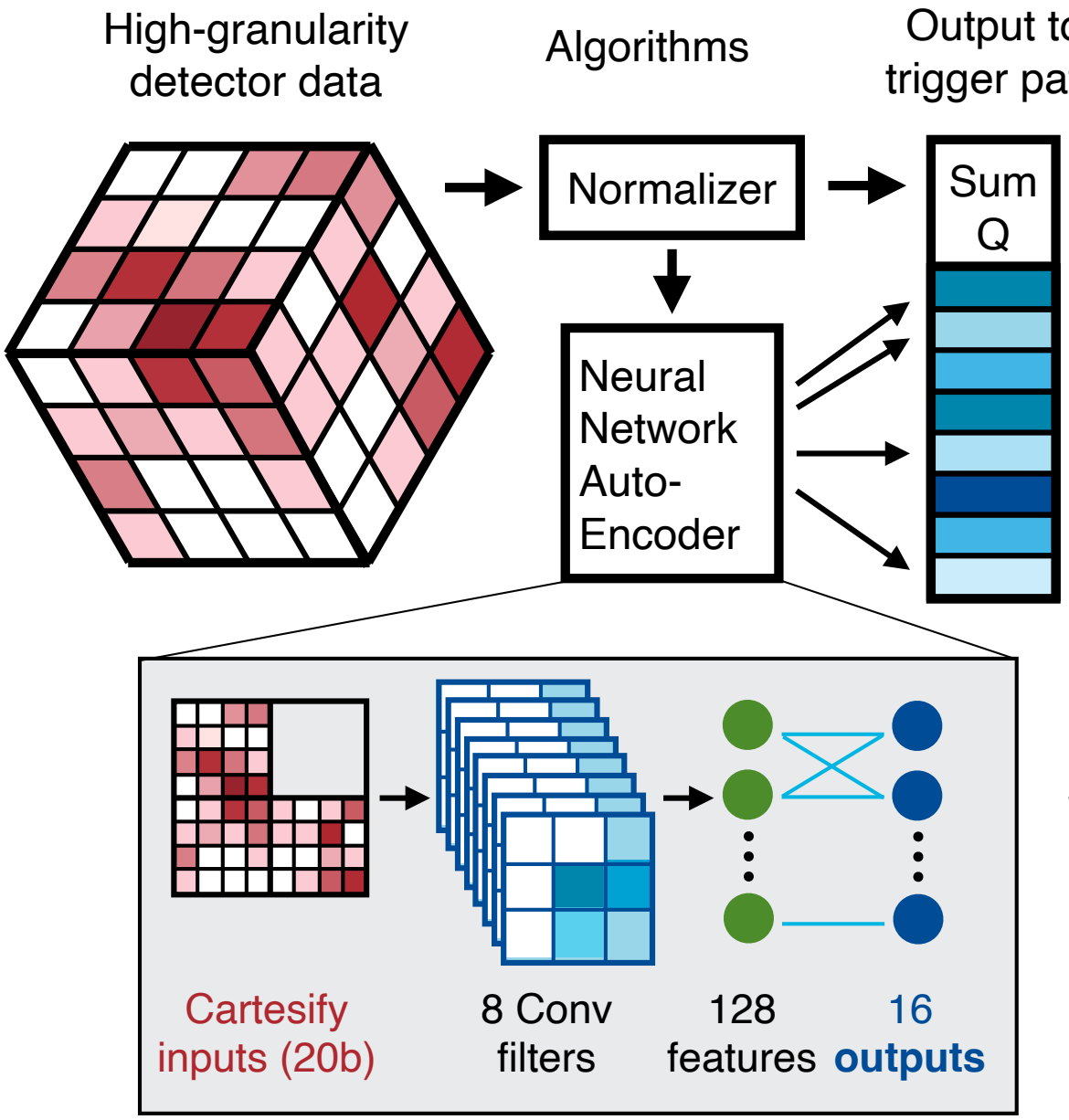
CMS HGCal data concentrator

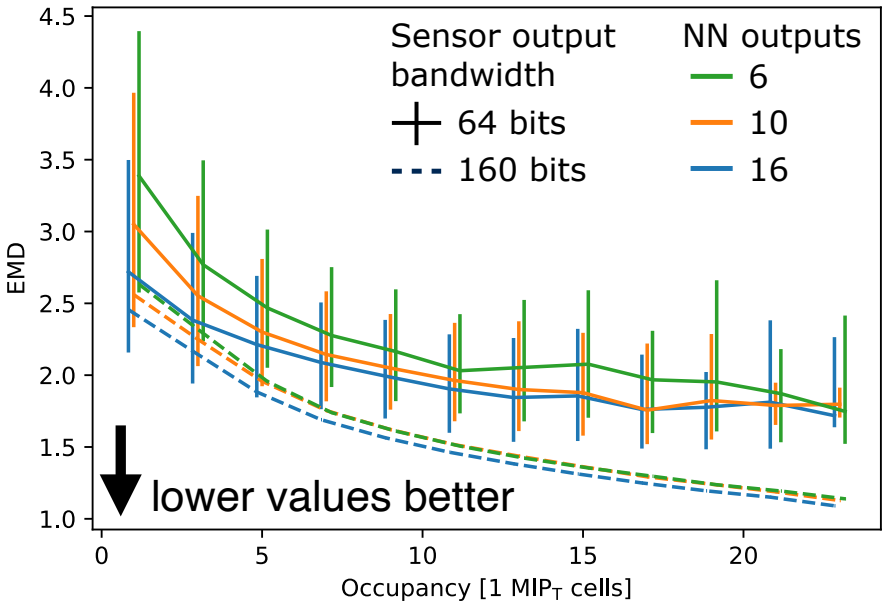


CMS HGCal data concentrator



CMS HGCal data concentrator



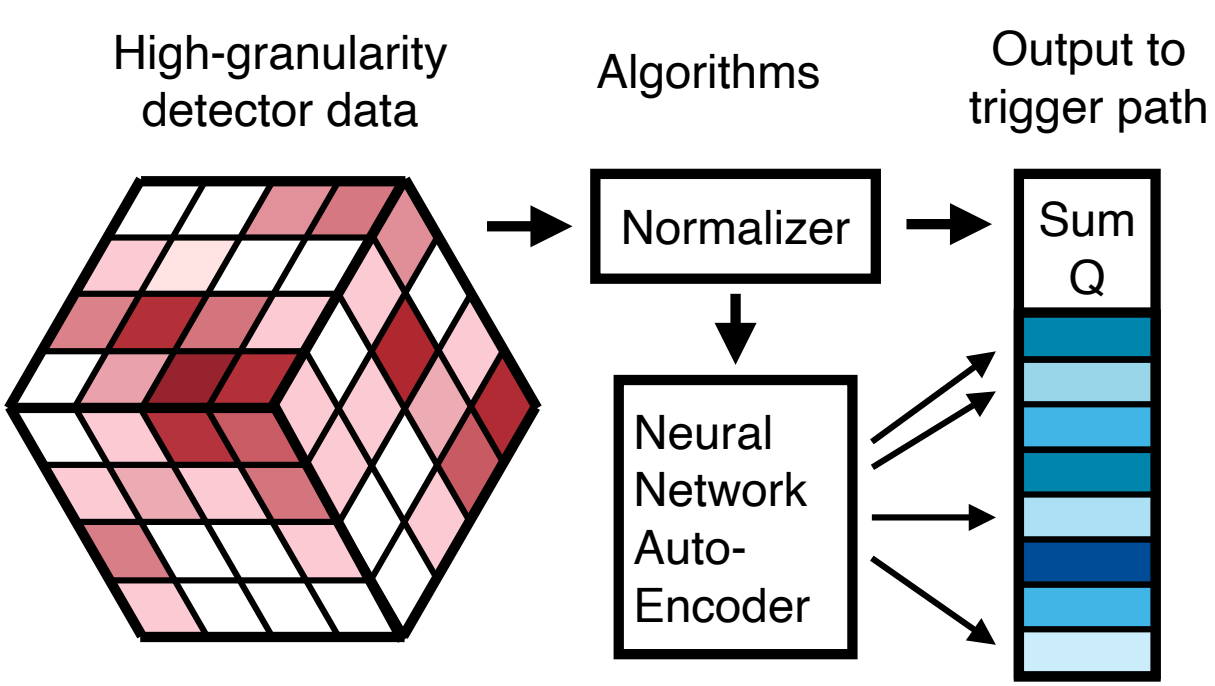


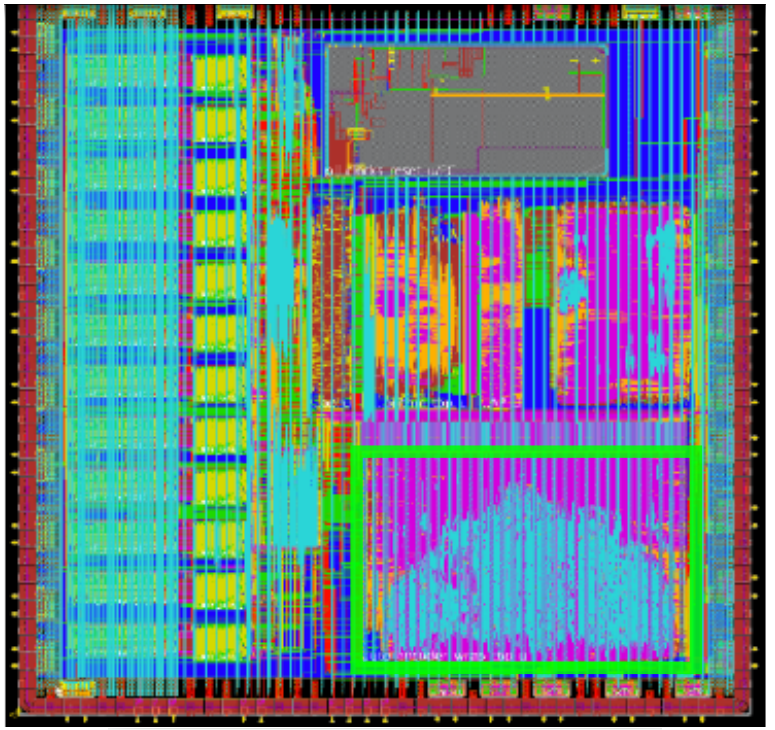
Search for NN architectures with

- Best decoded image fidelity
- Minimal logic (power, area)

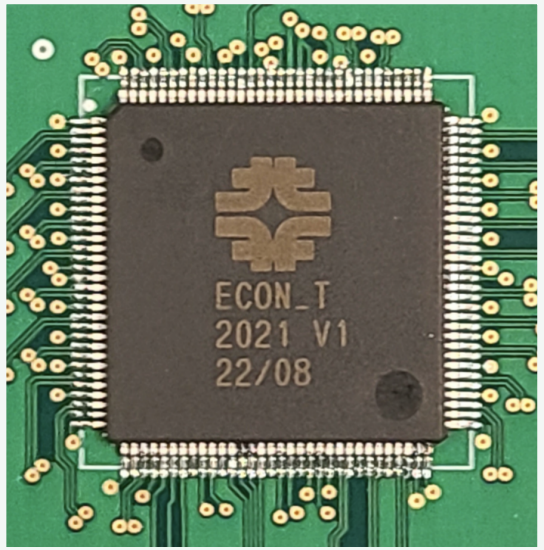
Minimal bit-widths prevent overflow

CMS HGCal data concentrator





Metric	Simulation	Target
Power	48 mW	<100 mW
Energy / inference	1.2 nJ	N/A
Area	2.88 mm^2	<4 mm^2
Gates	780k	N/A
Latency	50 ns	<100 ns



Di Guglielmo+, IEEE TNS 68.8 (2021) 2179

In-pixel data filtering

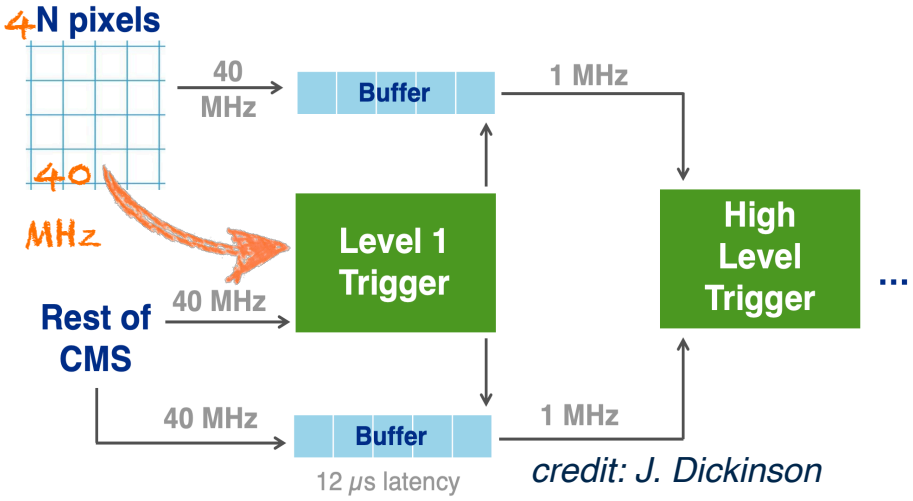
A core HL-LHC motivation is (di-)Higgs production. Main decay: h→bb.

- Pixel tracker critical to identify B decays with O(mm) displacements.

In-pixel data filtering

A core HL-LHC motivation is (di-)Higgs production. Main decay: h→bb.

- Pixel tracker critical to identify B decays with O(mm) displacements.



credit: J. Dickinson

SmartPixels concept could (e.g.) upgrade inner CMS layers for 50x readout rate.

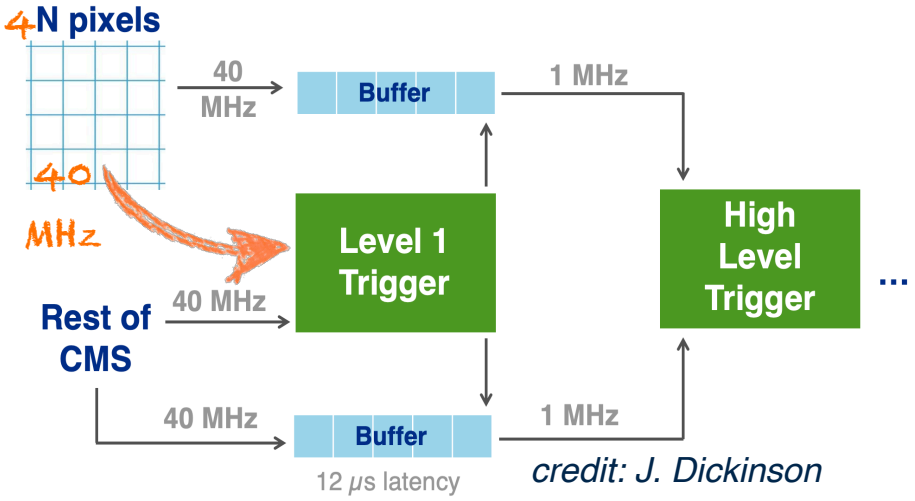
Demonstrator: 50 x 12.5 x 100 μm pixels

Consider size, shape, + time structures to remove sub-2 GeV track data (95% hits).

In-pixel data filtering

A core HL-LHC motivation is (di-)Higgs production. Main decay: h→bb.

- Pixel tracker critical to identify B decays with O(mm) displacements.



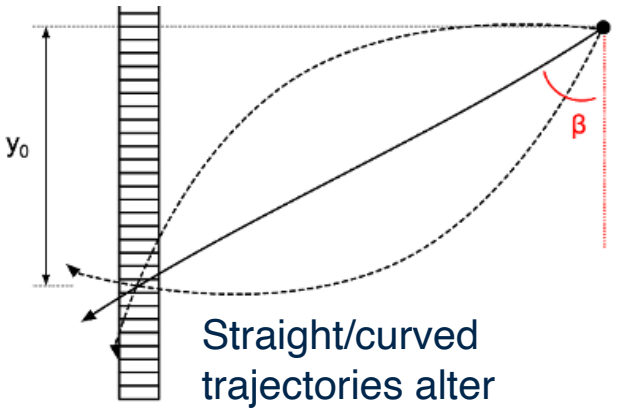
credit: J. Dickinson

SmartPixels concept could (e.g.) upgrade inner CMS layers for 50x readout rate.

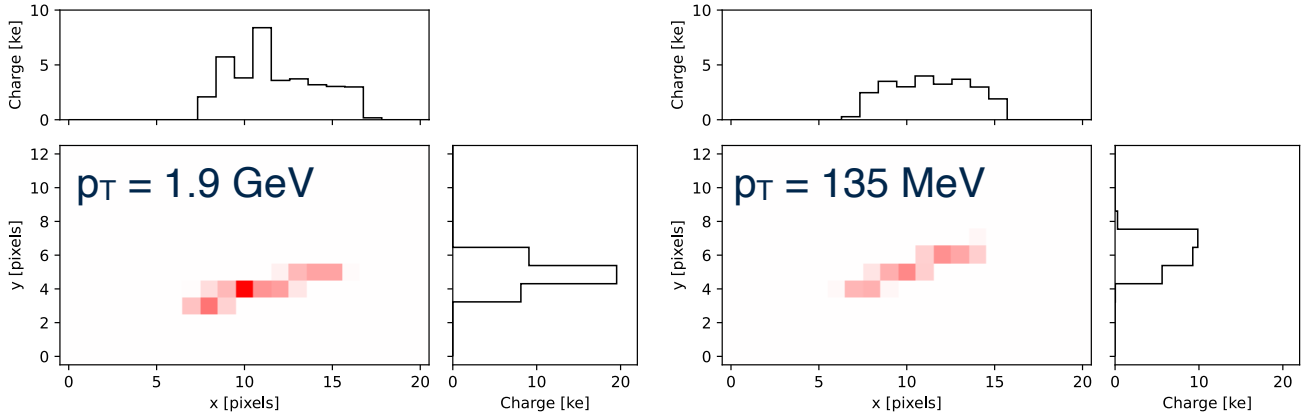
Demonstrator: 50 x 12.5 x 100 μm pixels

Consider size, shape, + time structures to remove sub-2 GeV track data (95% hits).

Yoo+, Mach.Learn.Sci.Tech. 5 (2024) 3, 035047



Straight/curved trajectories alter pixel cluster shapes

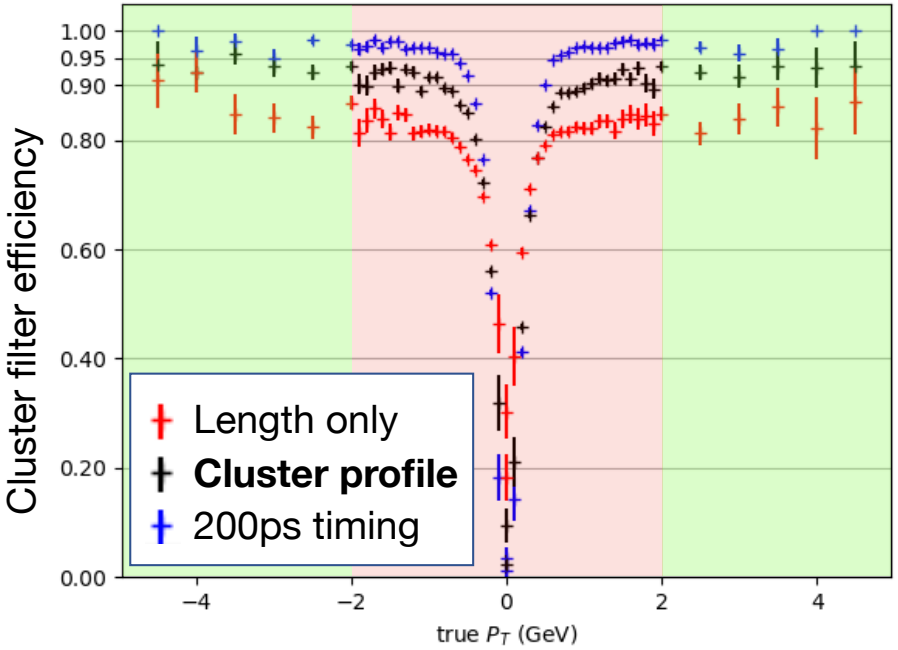


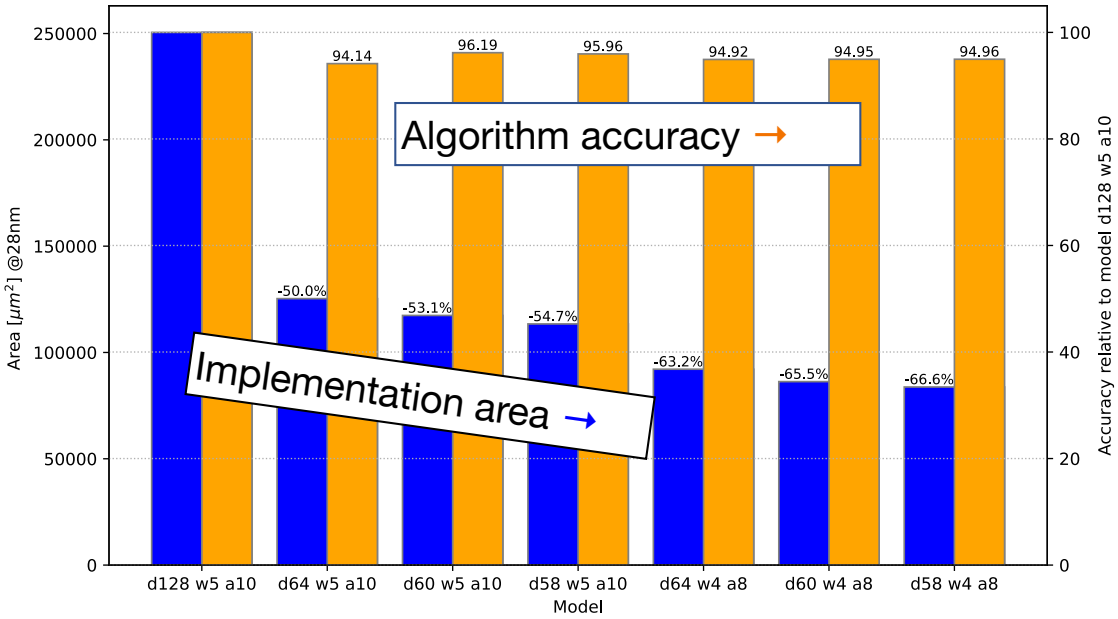
Differences can be subtle!

In-pixel data filtering

NN of varying complexity were optimized, and implemented in 28nm CMOS

Cluster profile model balances performance, complexity.

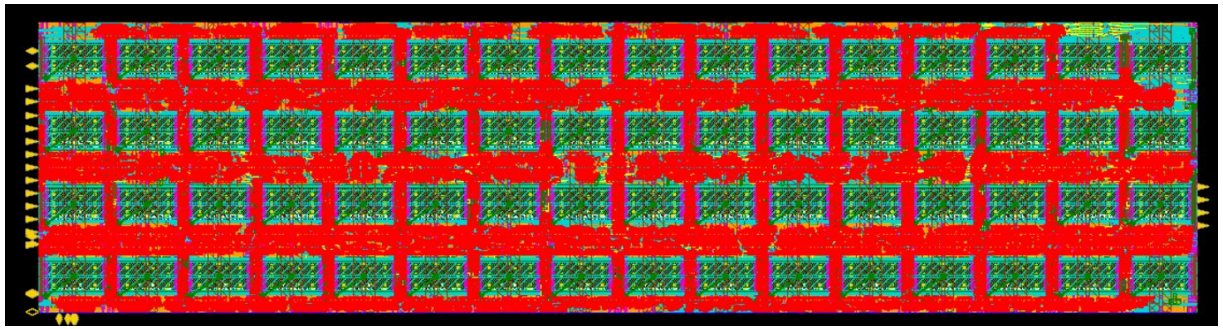




Fully parallel implementation minimizes latency

Familiar chain for co-design (qKeras→hls4ml→Catapult)

Second prototype ROIC now being characterized.



Next-gen tracking detectors

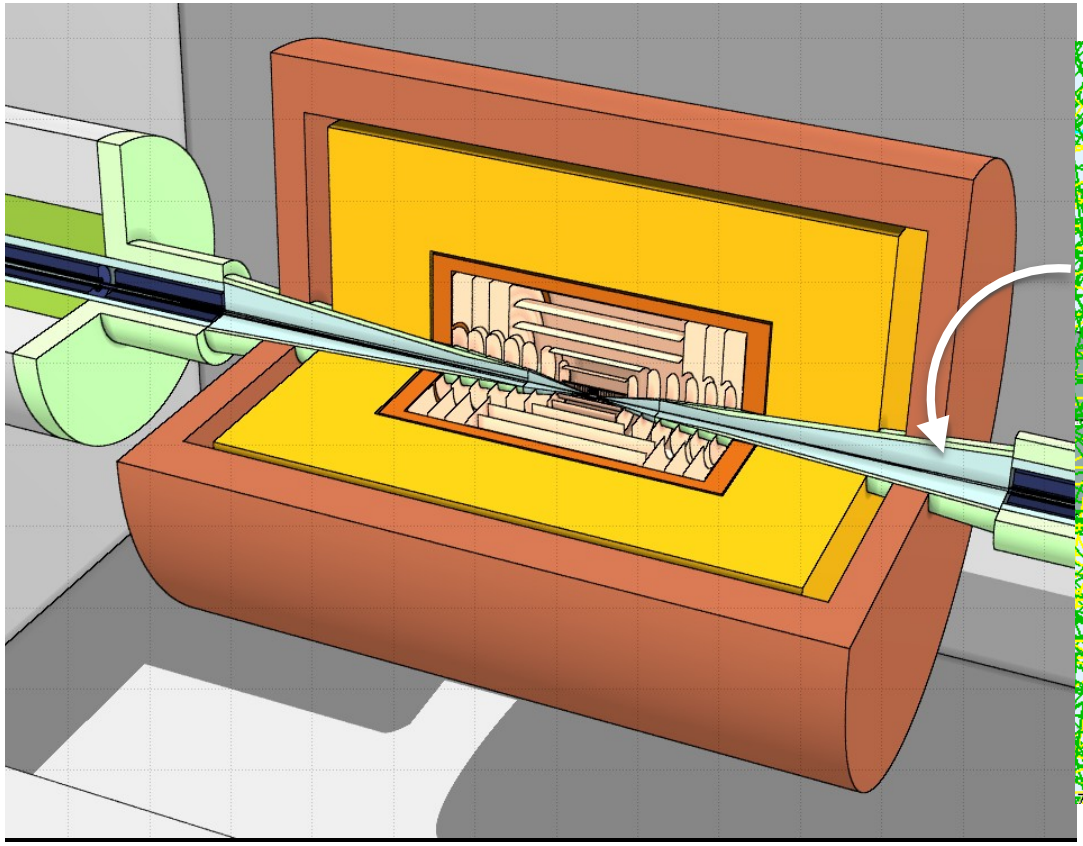
Longer term, hit rates at future colliders should far surpass the HL-LHC.

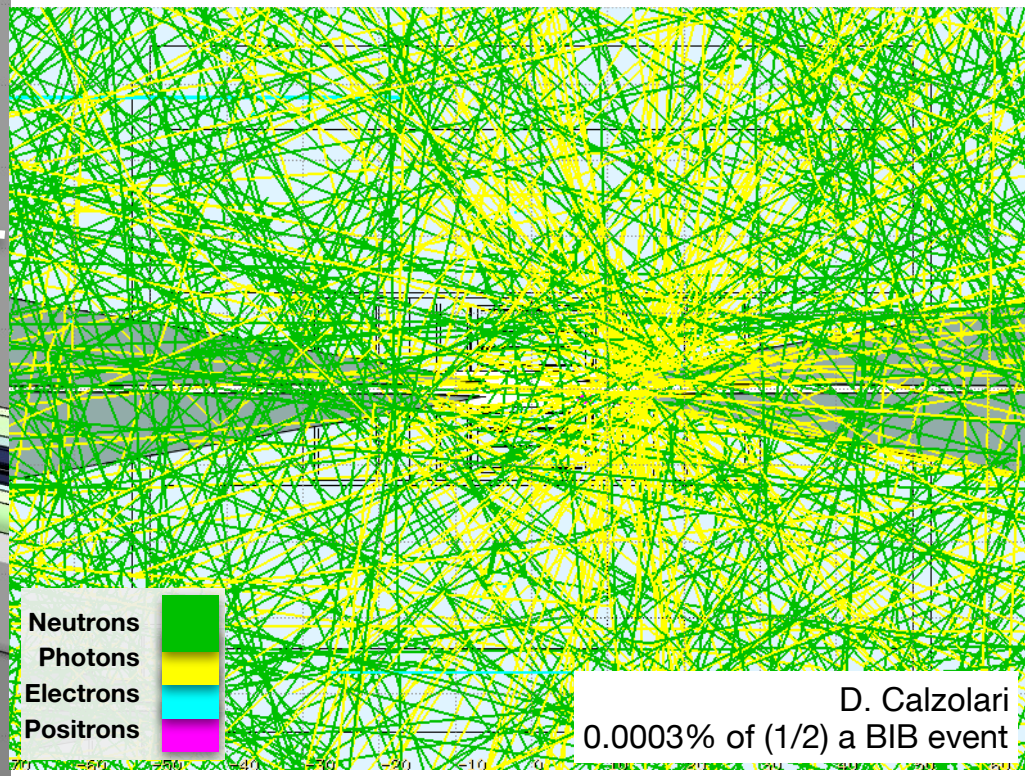
(Can extrapolate LHC to a 50-100 TeV hadron collider w/ ~1000 pileup)

Next-gen tracking detectors

Longer term, hit rates at future colliders should far surpass the HL-LHC.

(Can extrapolate LHC to a 50-100 TeV hadron collider w/ ~1000 pileup)





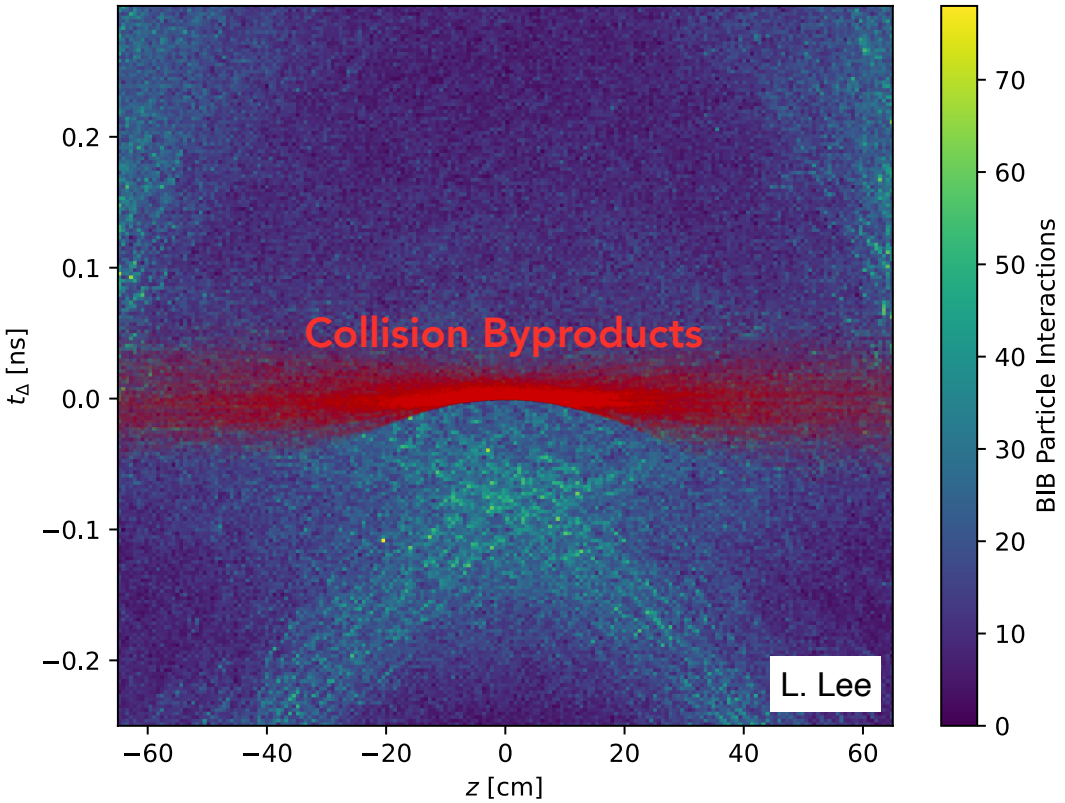
Muon collider is a bit of a different beast due to **Beam Induced Background**

Next-gen tracking detectors

Longer term, hit rates at future colliders should far surpass the HL-LHC.

(Can extrapolate LHC to a 50-100 TeV hadron collider w/ ~1000 pileup)

$$t_\Delta = t - t_{exp}(\beta = 1)$$



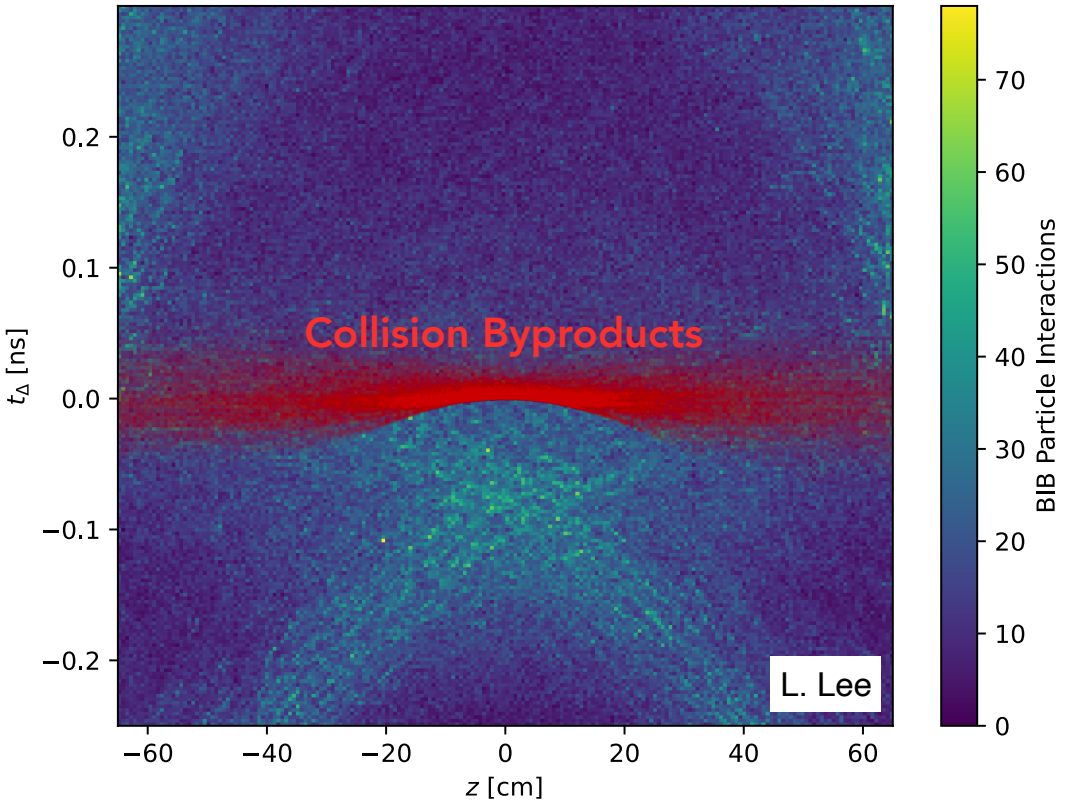
Sensors will require timing resolution at the scale of 10ps.

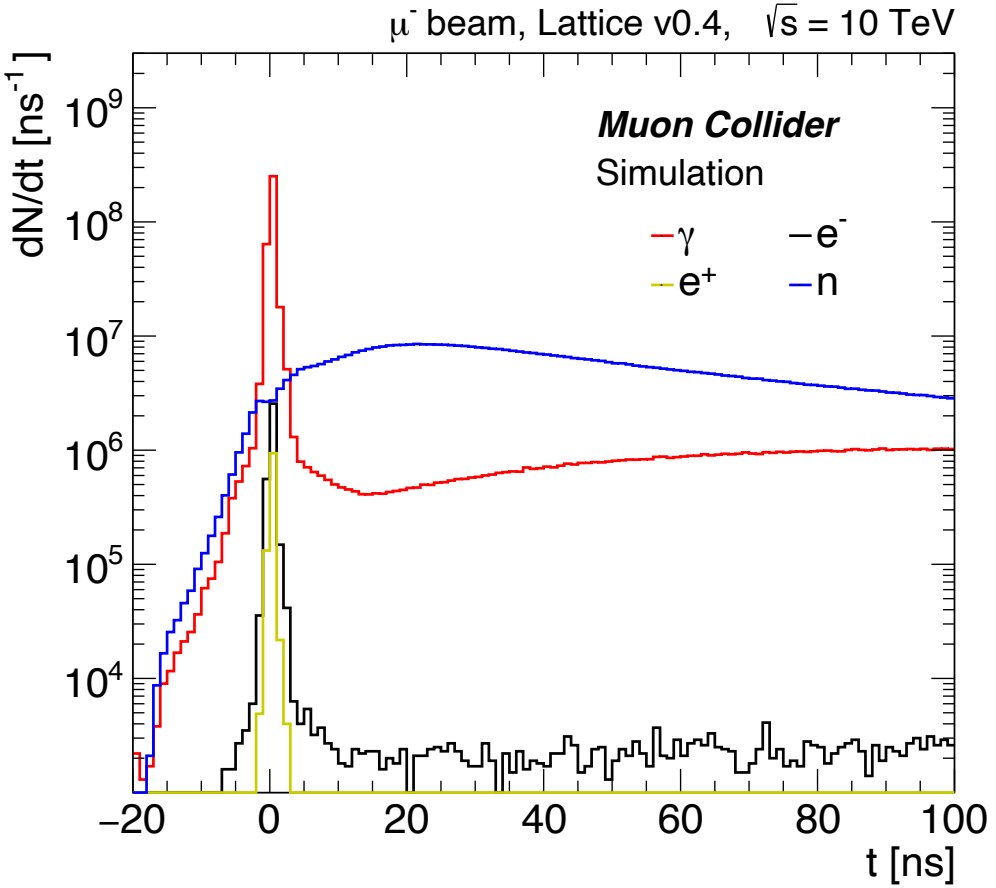
Next-gen tracking detectors

Longer term, hit rates at future colliders should far surpass the HL-LHC.

(Can extrapolate LHC to a 50-100 TeV hadron collider w/ ~1000 pileup)

$$t_\Delta = t - t_{exp}(\beta = 1)$$





Sensors will require timing resolution at the scale of 10ps.

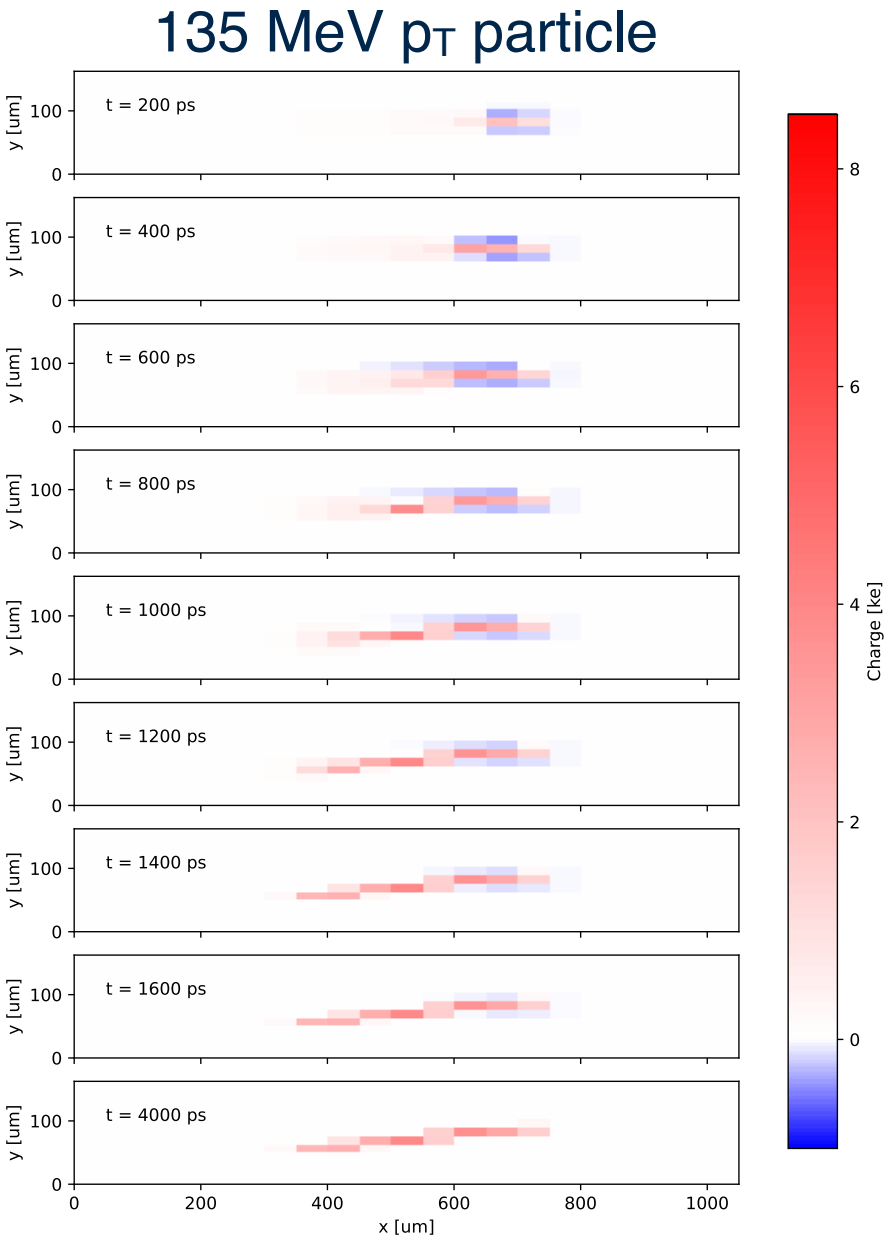
Next-gen tracking detectors

Implications for in-pixel intelligence?

A number of avenues to pursue:

1. ***Timing*** will play a key role, but how to cope w/ higher dimensionality?
2. Can we move from filtering to featurization (e.g. predict hit position + angle), to speed up trigger tracking?
3. More efficient algorithm designs (Spiking NNs, e.g.?)

Miniskar+, "Neuro-Spark: A Submicrosecond Spiking Neural Networks Architecture for In-Sensor Filtering," International Conference on Neuromorphic Systems (ICONS), 2024



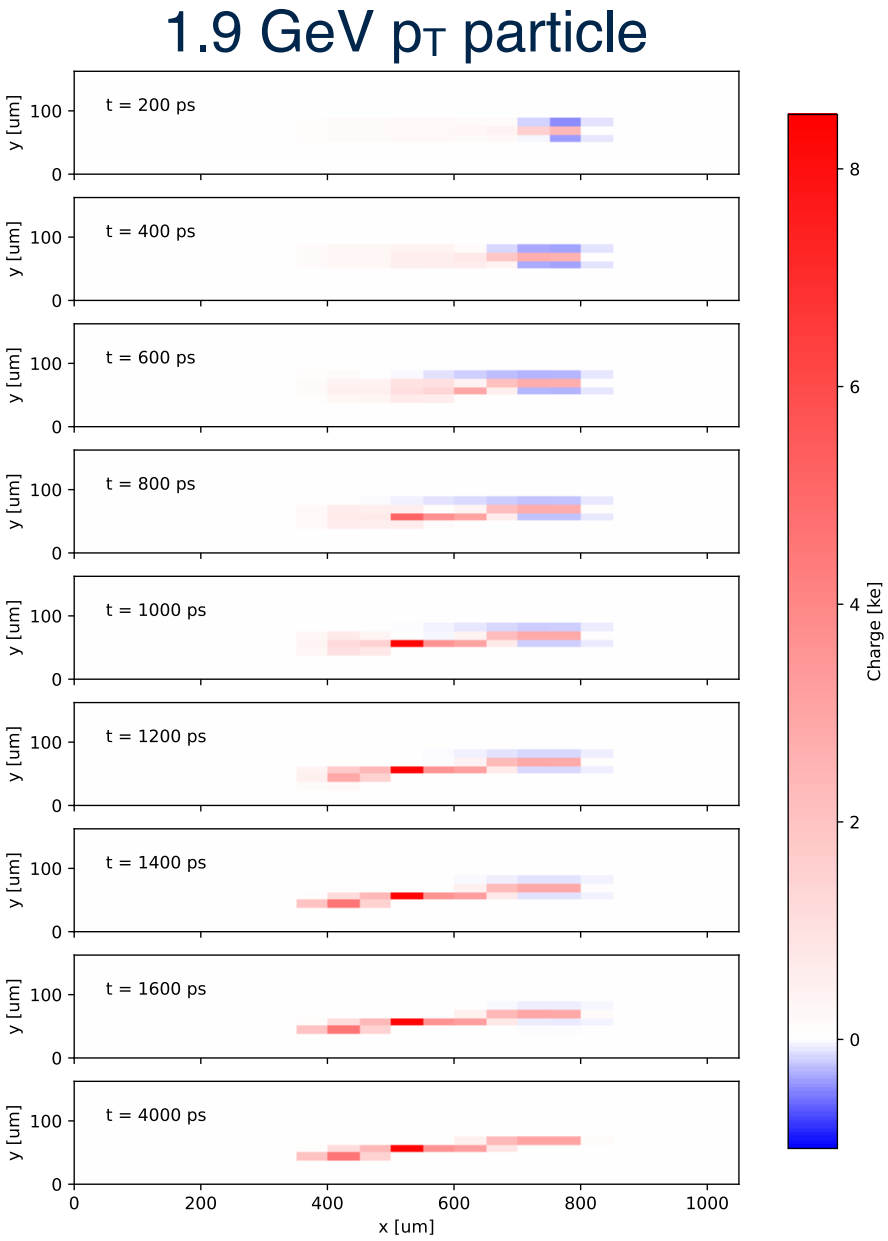
Next-gen tracking detectors

Implications for in-pixel intelligence?

A number of avenues to pursue:

1. ***Timing*** will play a key role, but how to cope w/ higher dimensionality?
2. Can we move from filtering to featurization (e.g. predict hit position + angle), to speed up trigger tracking?
3. More efficient algorithm designs (Spiking NNs, e.g.?)

Miniskar+, "Neuro-Spark: A Submicrosecond Spiking Neural Networks Architecture for In-Sensor Filtering," International Conference on Neuromorphic Systems (ICONS), 2024



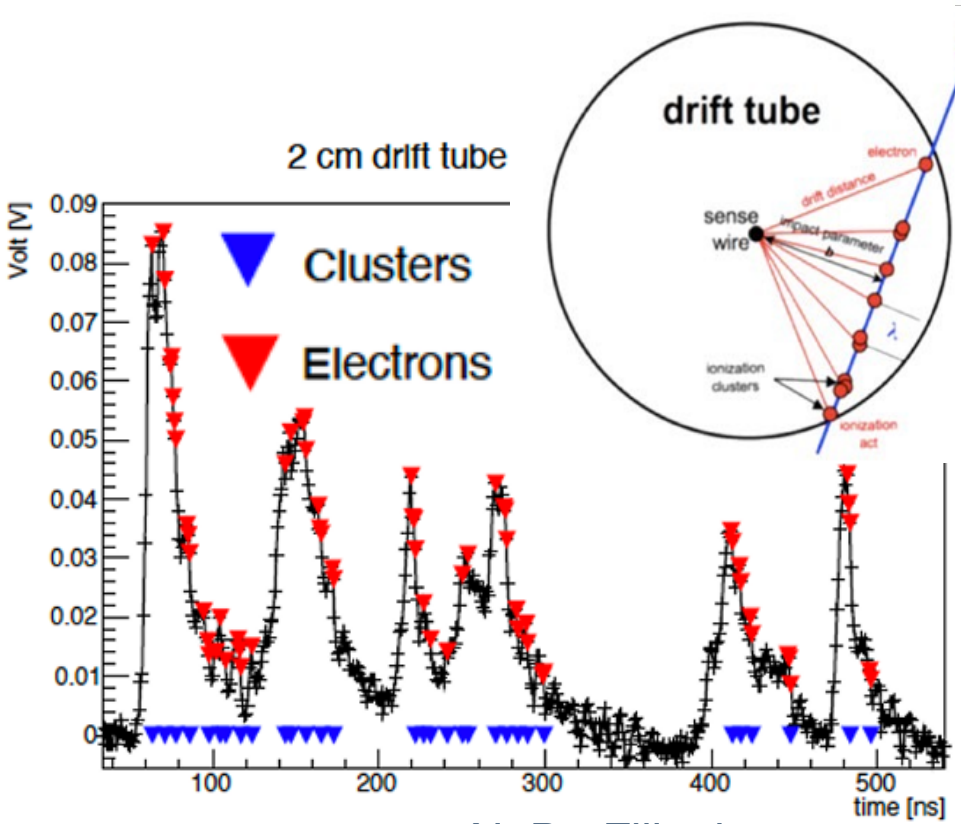
In-sensor featurization for FCC-ee?

Data rates at next-gen e^+e^- colliders are generally small (hadron, μCol).

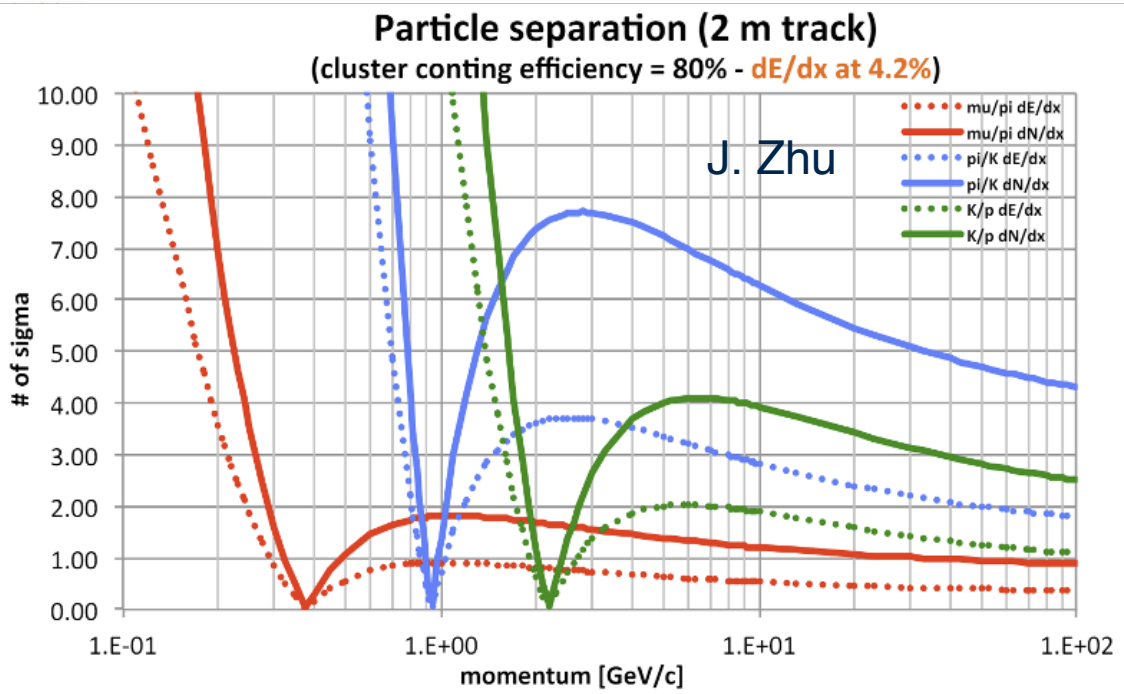
Total bandwidth @FCC-ee is ~3x LHCb (but 200kHz physics rate).

Tracker dominates the total data rate (~% X_0: drift chamber or straws).

Pulse structure is critical for **particle ID**!



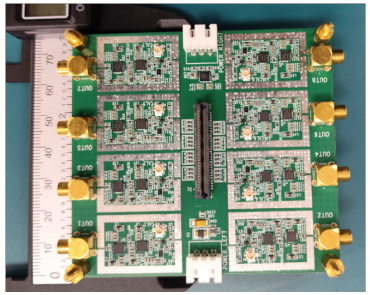
N. De Filippis

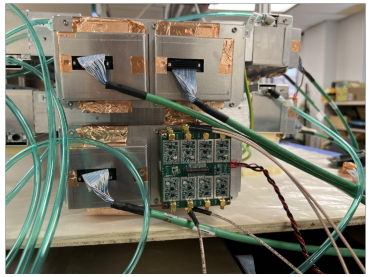


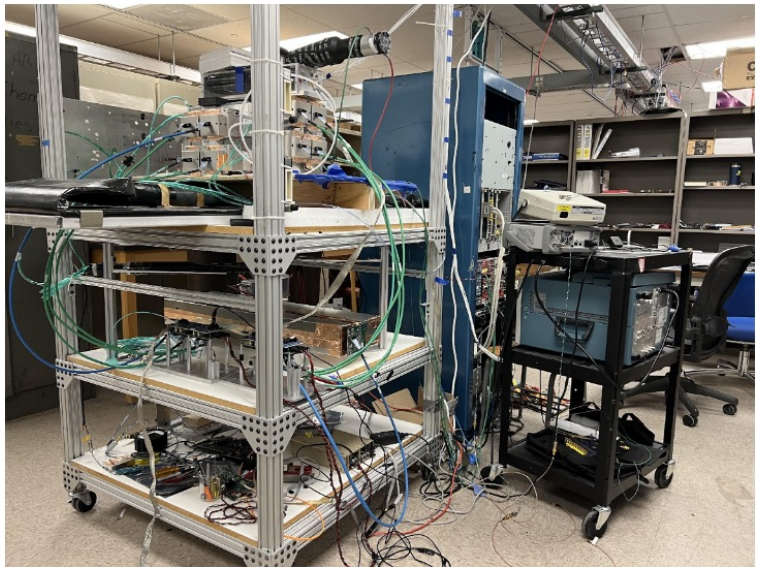
Cluster counting ~2x better than dE/dx (but never used in experiment)!

In-sensor featurization for FCC-ee?

Efforts are underway for a straw tracker demonstration at Michigan.

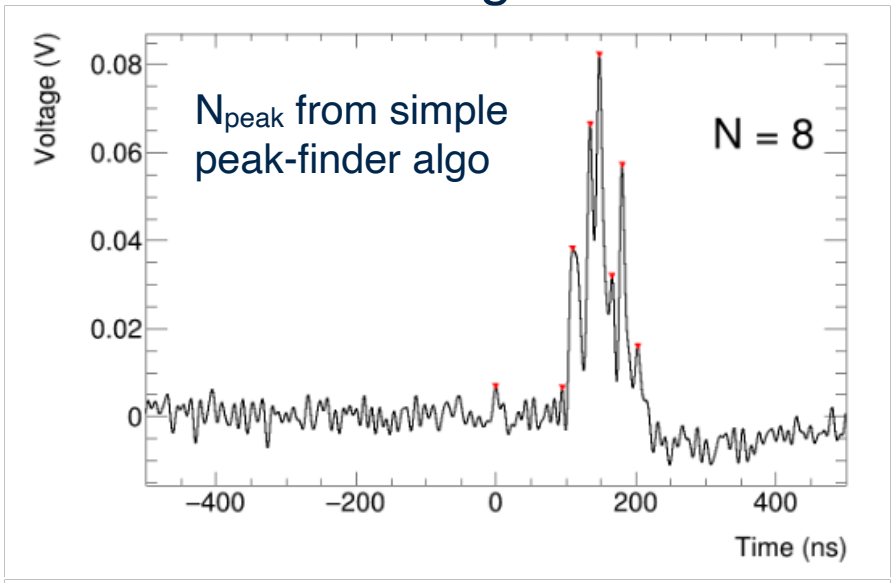


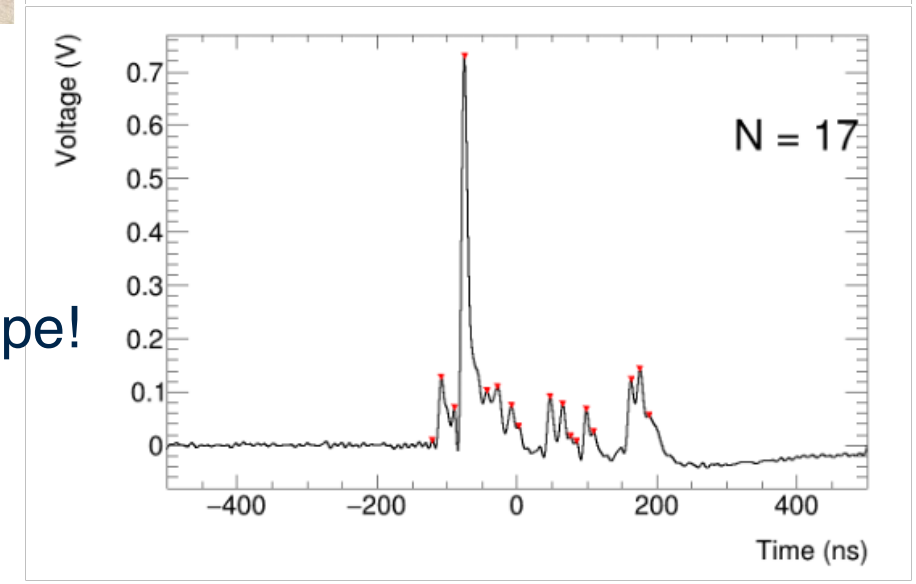




3+1 ATLAS sMDT chambers to measure ‘dx’ of cosmics, and test new readout electronics (low-noise, high gain).

Achieve σ~100μm hit resolution in 90:10 He:Isobutane.



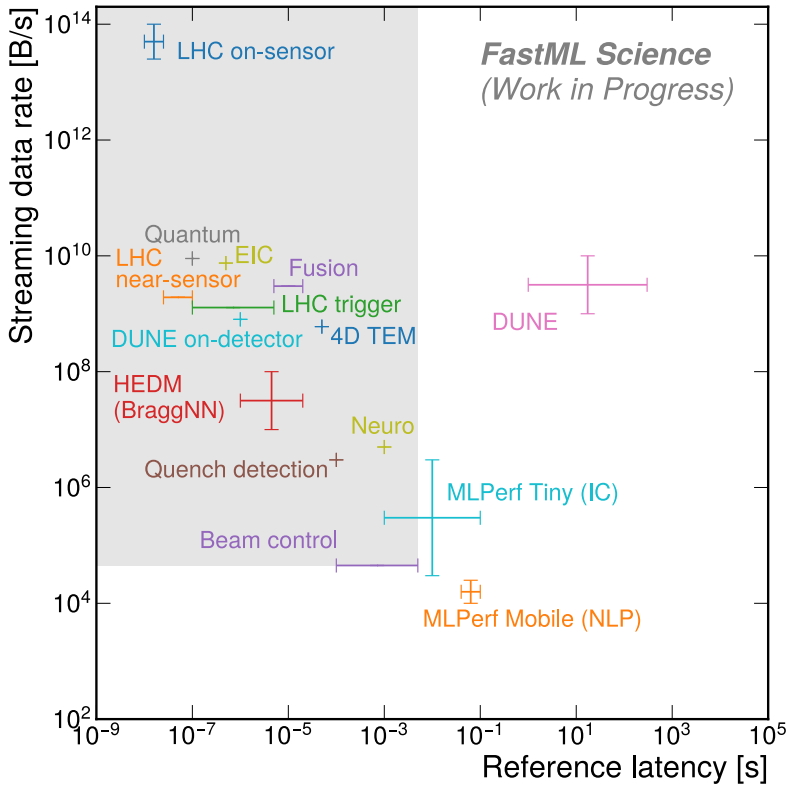


Individual clusters evident on the oscilloscope!

ML-based clustering should give a nice improvement in performance!

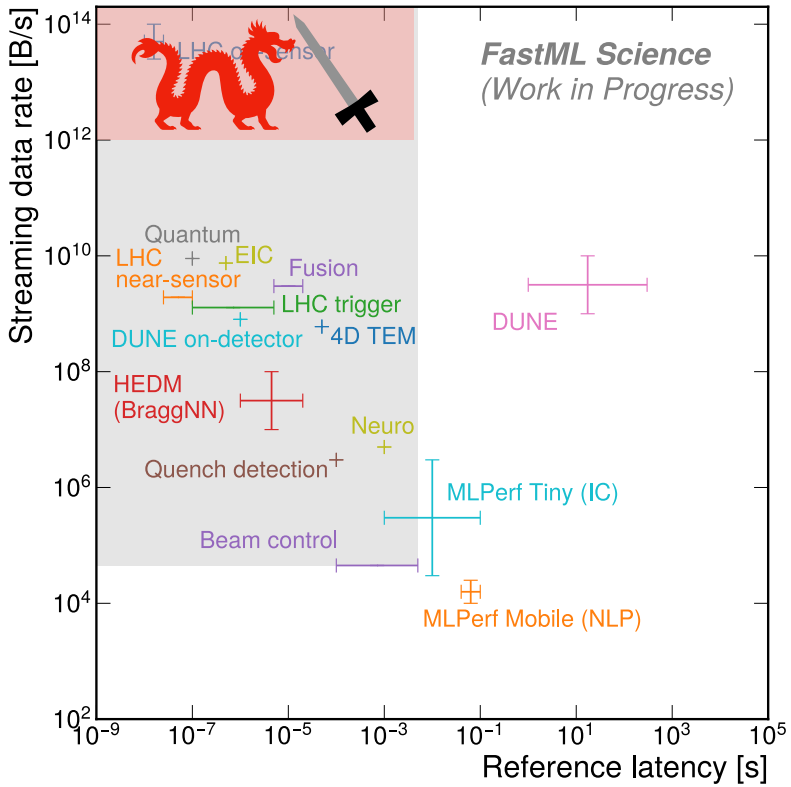
Conclusions and outlook

Despite the inherent challenges, intelligent processing within custom ASICs will be critical to unlock the full potential of next-gen experiments.



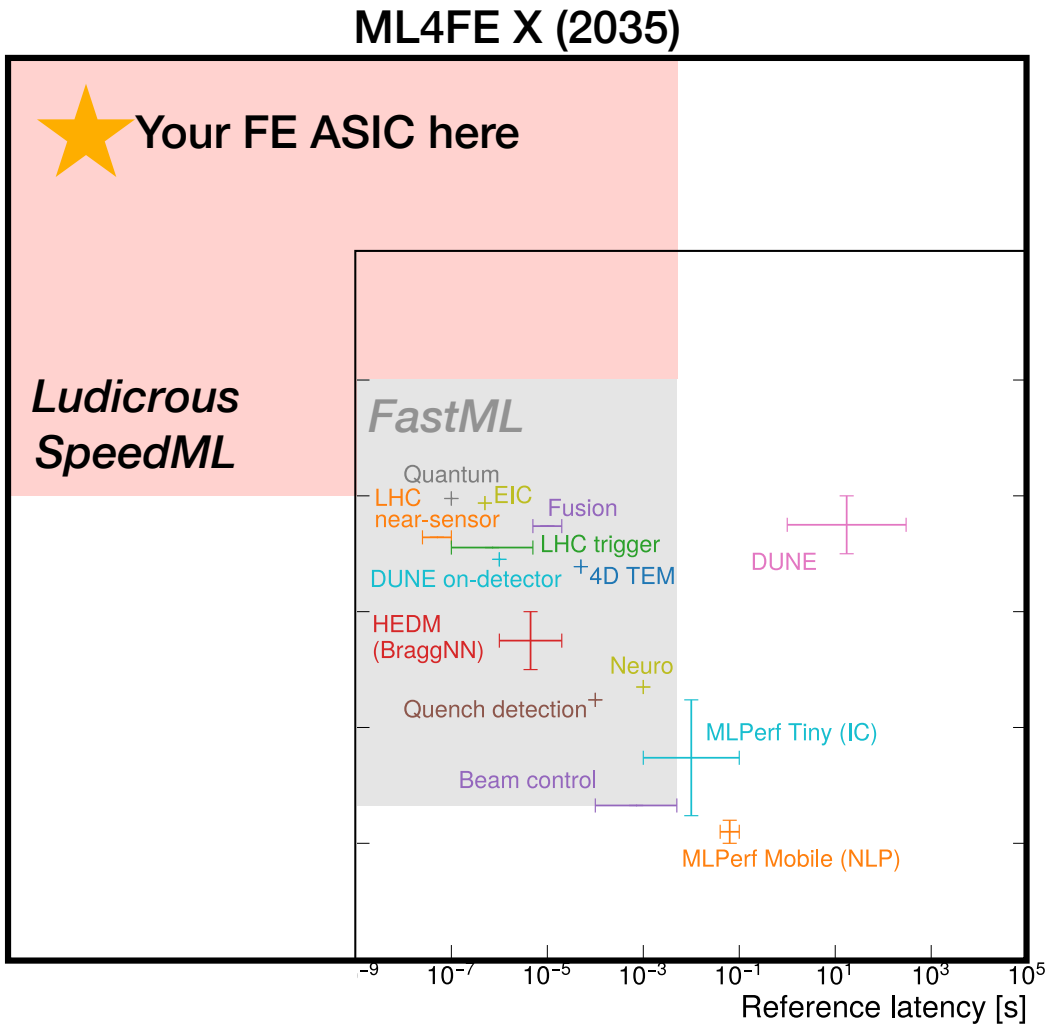
Conclusions and outlook

Despite the inherent challenges, intelligent processing within custom ASICs will be critical to unlock the full potential of next-gen experiments.



Conclusions and outlook

Despite the inherent challenges, intelligent processing within custom ASICs will be critical to unlock the full potential of next-gen experiments.



Conclusions and outlook

Despite the inherent challenges, intelligent processing within custom ASICs will be critical to unlock the full potential of next-gen experiments.

Past 5 years have seen exciting new efforts, but much is still theoretical.

- ~10k ECONs will “go live” in CMS in a few years… how will they work?
  - What lessons can we learn for the next-generation of detectors?

Conclusions and outlook

Despite the inherent challenges, intelligent processing within custom ASICs will be critical to unlock the full potential of next-gen experiments.

Past 5 years have seen exciting new efforts, but much is still theoretical.

- ~10k ECONs will "go live" in CMS in a few years… how will they work?
  - What lessons can we learn for the next-generation of detectors?

In the coming years, I will be eager to see:

- What possibilities new sensor tech enables (e.g. fast timing).
- As well as how "competitor" tech (non-ASIC) evolves.
  - Staying tuned for (e)FPGA, … and perhaps most important: links!

Thanks for your attention!