Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Trigger vs streaming DAQs - Fast data links for readout: optical, electrical, wireless, track reconstruction on GPUs

Chin-Chia Kuo
2nd Jul. 2024

Outline

- Trigger system

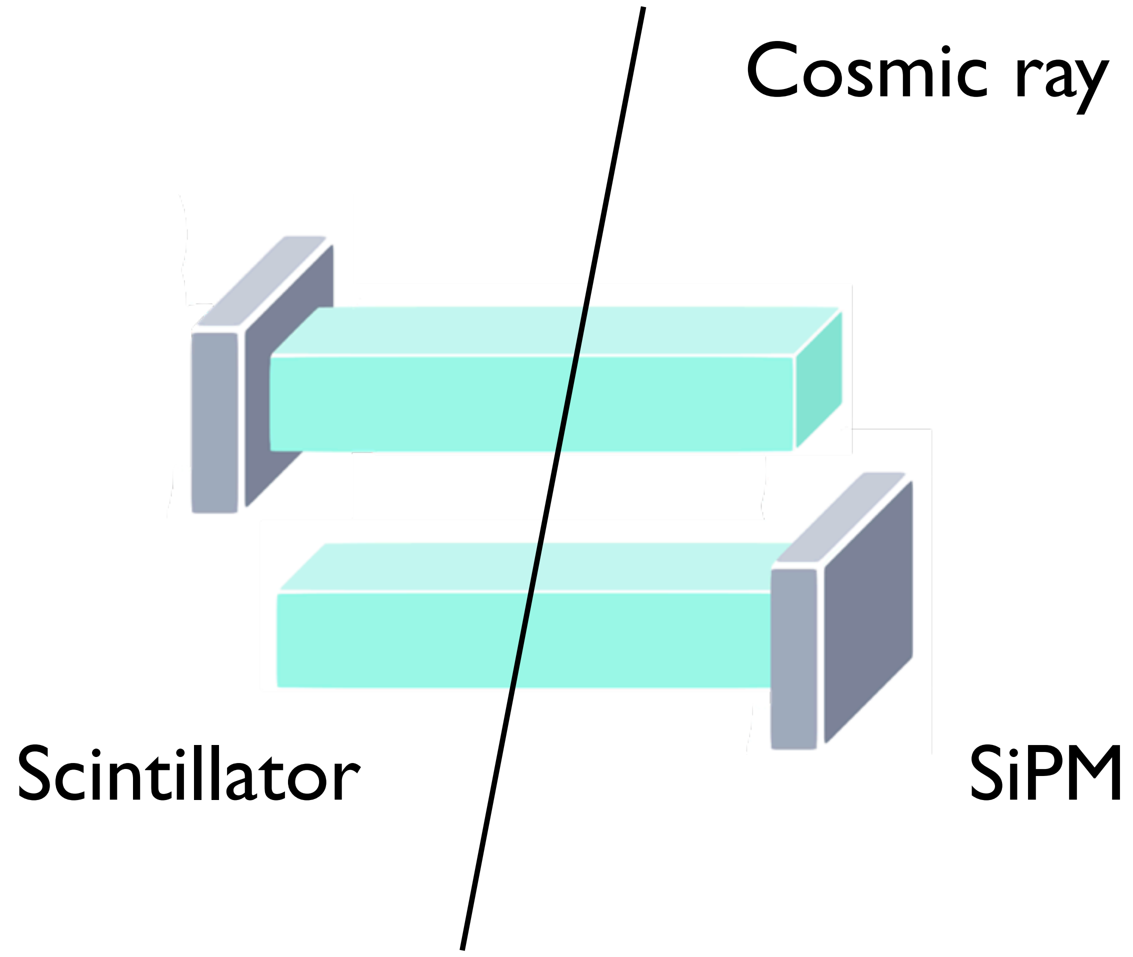
(i) Simple trigger system
(ii) Multilayer trigger system

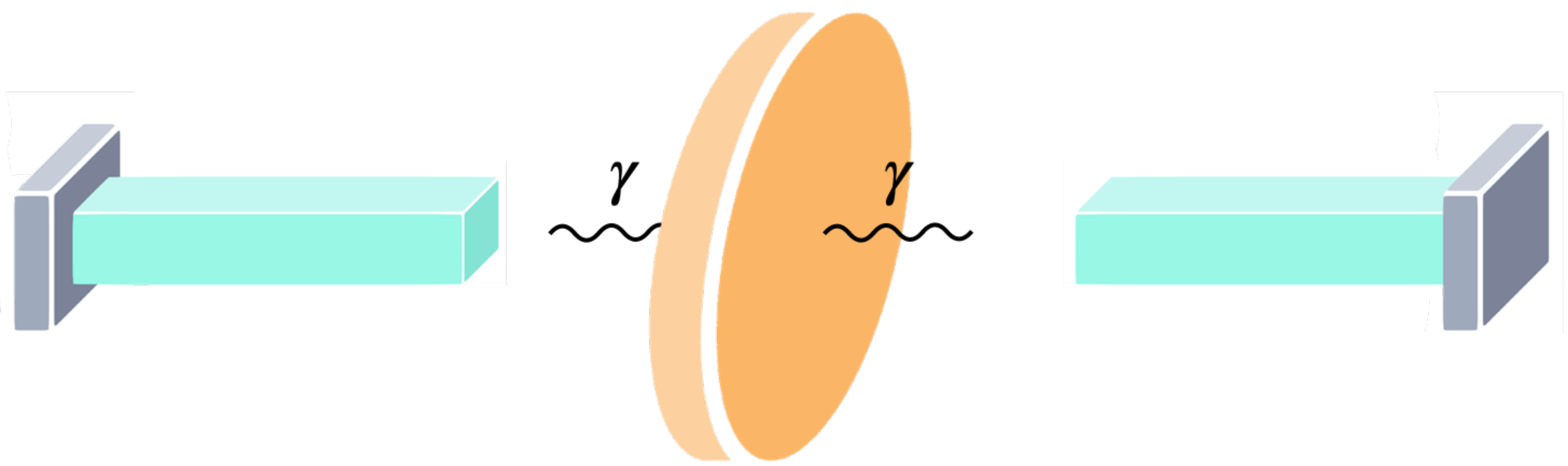
- Phase II upgrade of CMS trigger system

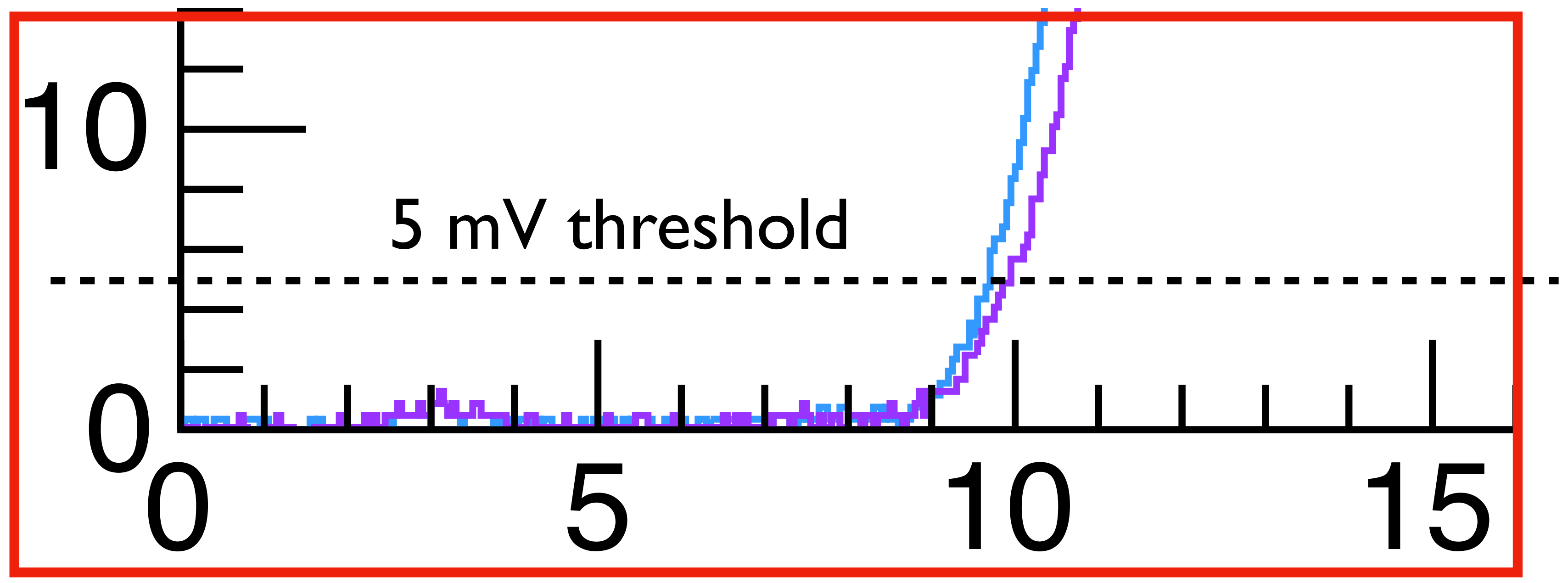
(iii) Tracking Triggers
(iv) GPU application
(v) Fast links

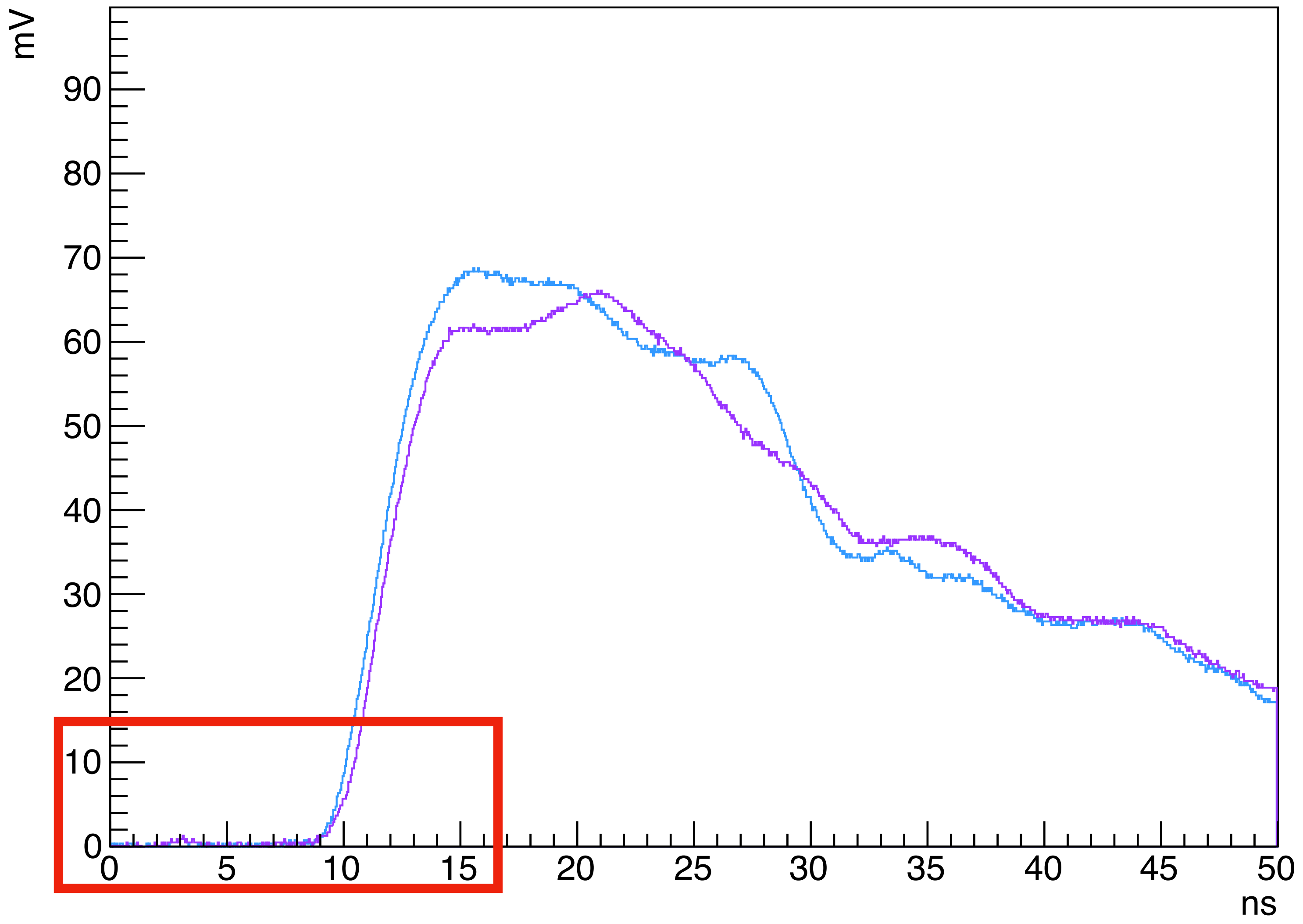
- Streaming DAQs

A simple system









Measure the signal time difference between two detectors

Reminder

- **Trigger basic requirements**

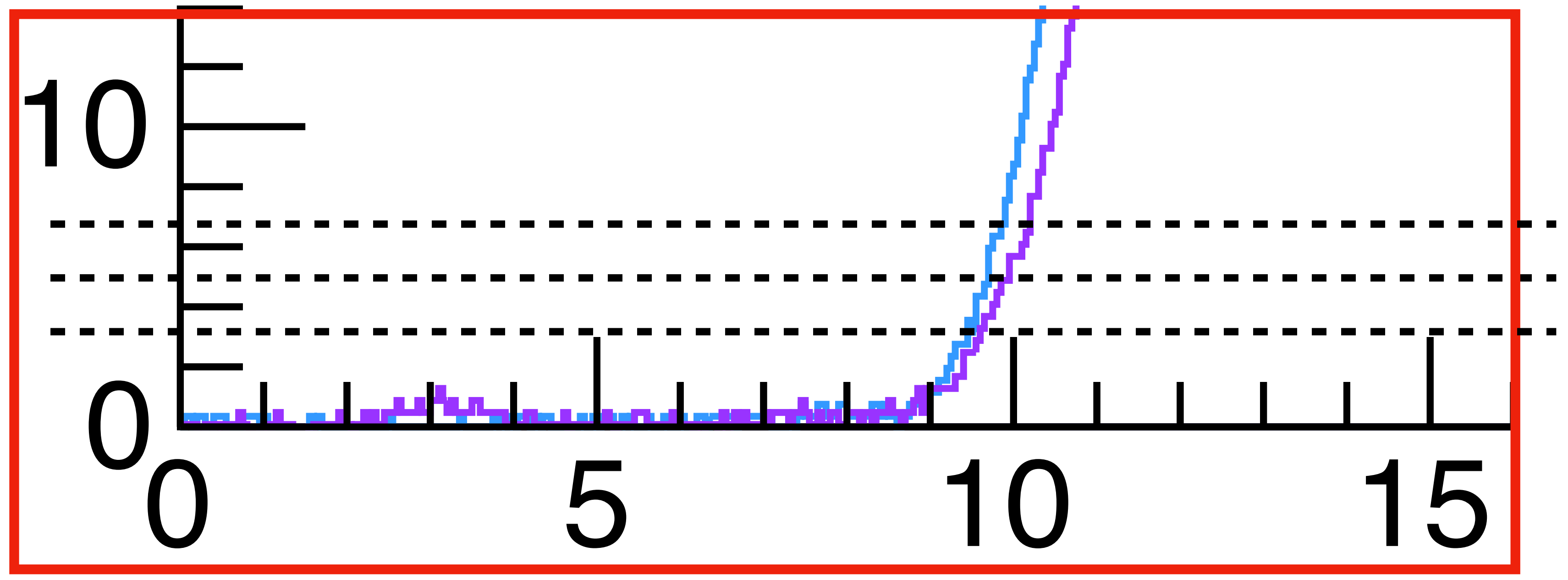
(i) Need high efficiency for selecting processes for physics analysis
(ii) Need large reduction of rate from unwanted noises
(iii) Robustness is essential
(iv) Highly flexible, to react to changing conditions
(v) System must be affordable

- **Pulse width**

(i) Limits the effective hit rate
(ii) Must be adapted to the desired trigger rate

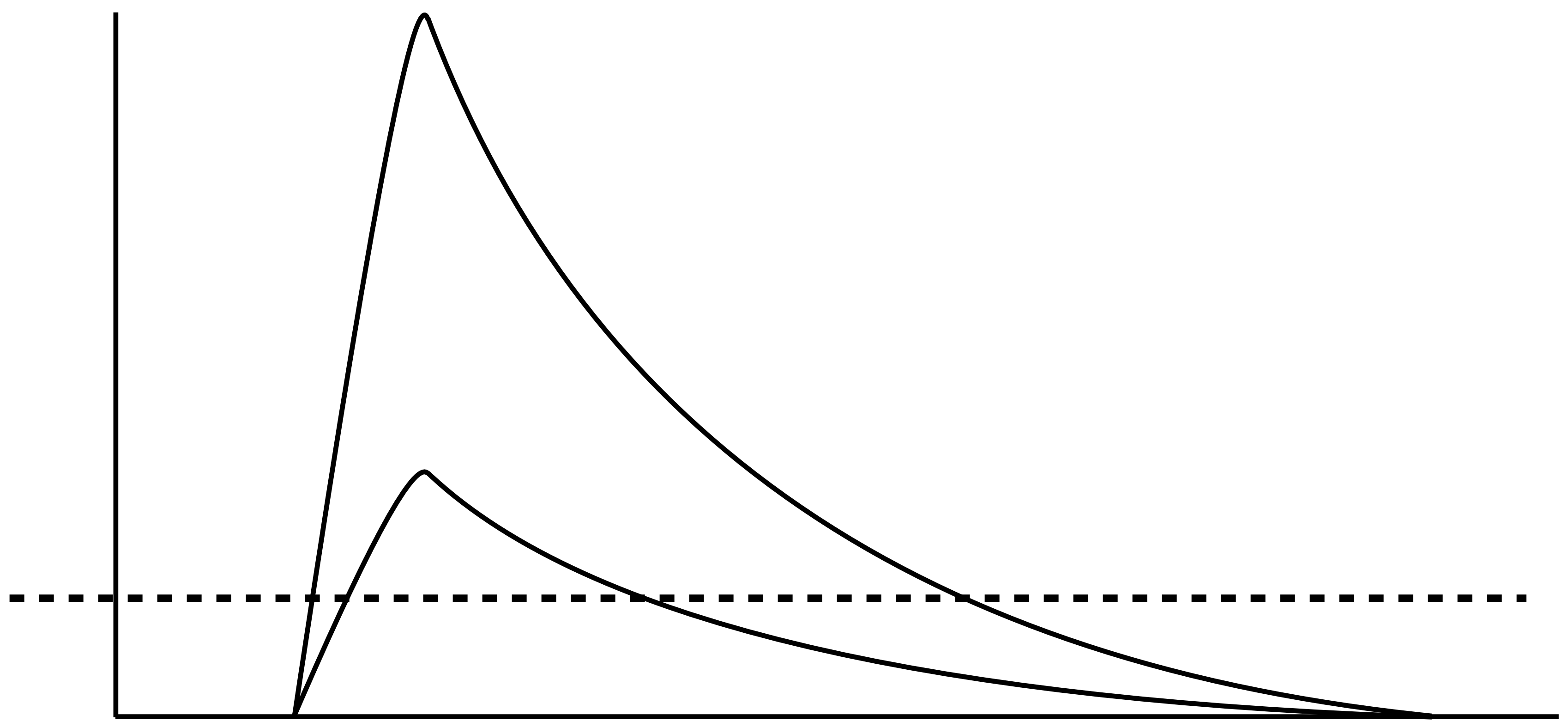
- **Time walk**

(i) The threshold-crossing time depends on the signal amplitude
(ii) Must be minimal in good trigger systems
(iii) Can be suppressed by triggering on total signal fraction

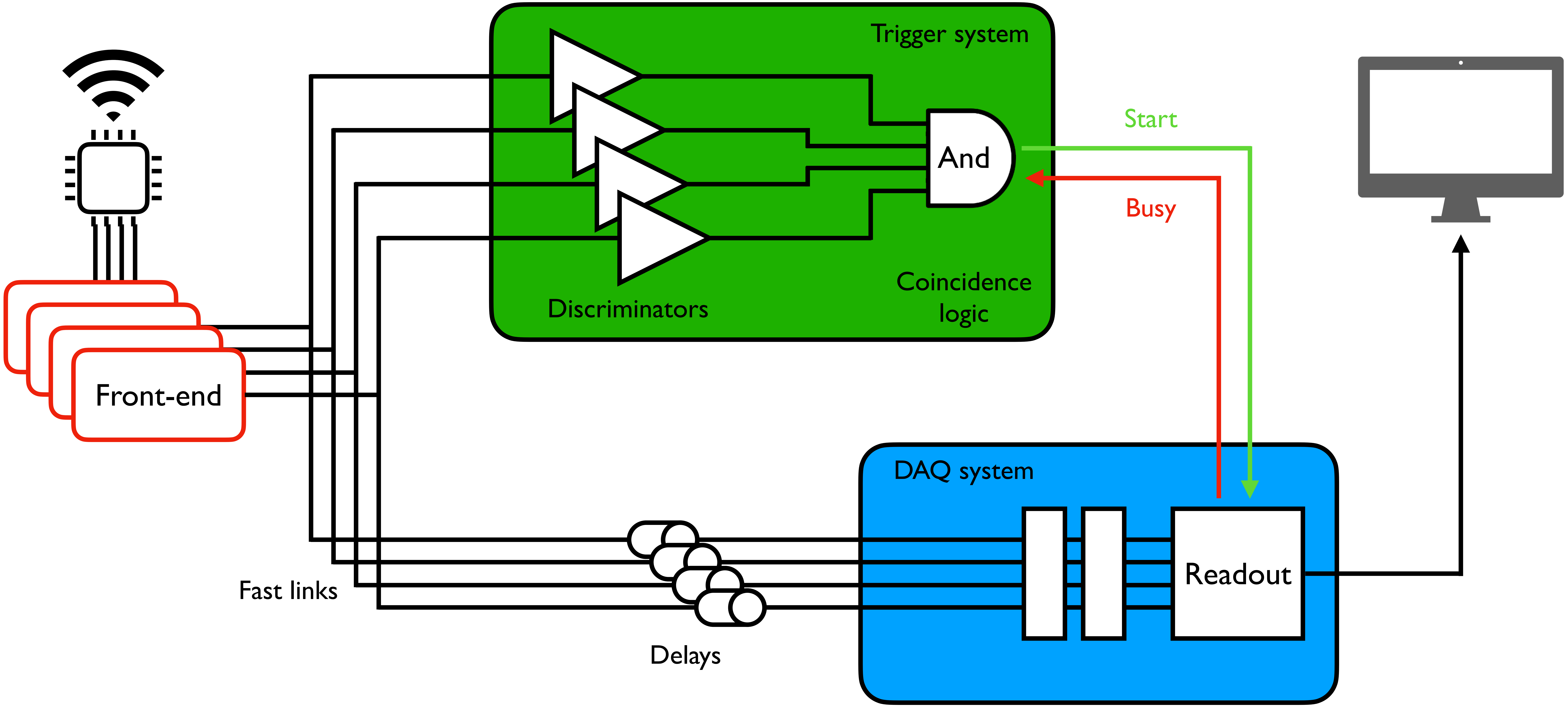


Find a trigger setting for good signal-to-noise ratio

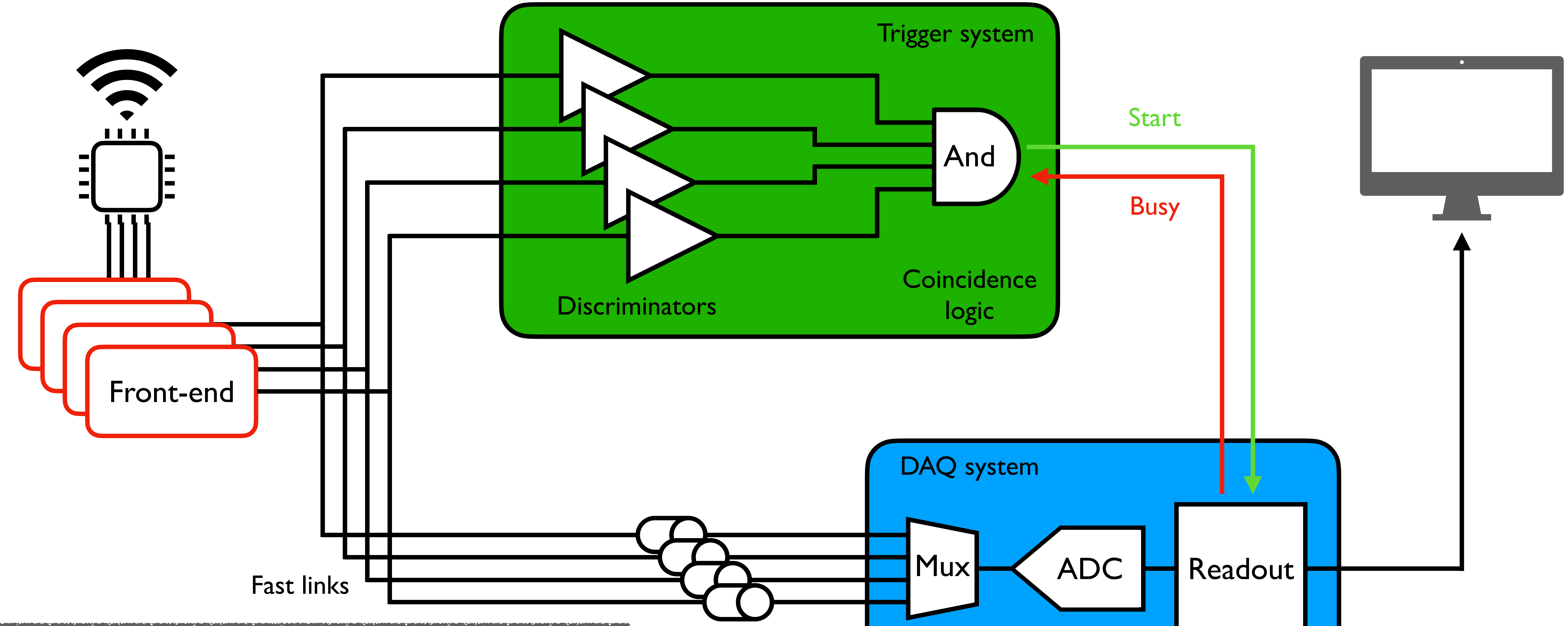
Fewer noises, lower efficiency



Simple trigger system



Simple trigger system



Multiplexer for reducing the number of wires or channels needed to transmit or process multiple signals

Dead time

- The key parameter in high speed trigger systems design

(i) The fraction of the acquisition time when no events can be recorded
(ii) Typically the order of a few percent
(iii) Reduces the overall system efficiency

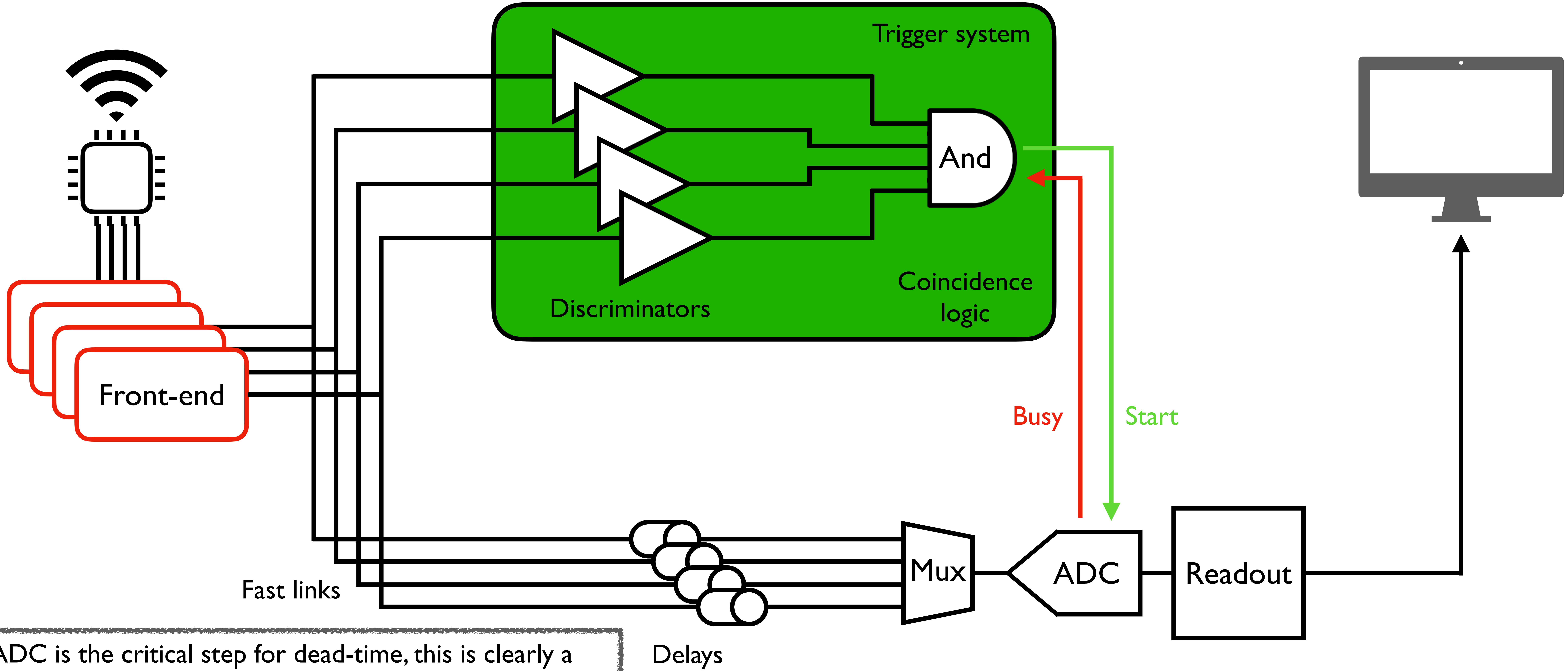
- Arises when a given processing step takes a finite amount of time

(i) Readout dead-time
(ii) Trigger dead-time
(iii) Operational dead-time

- Writing to disk or tape is much slow than accepting data into RAM

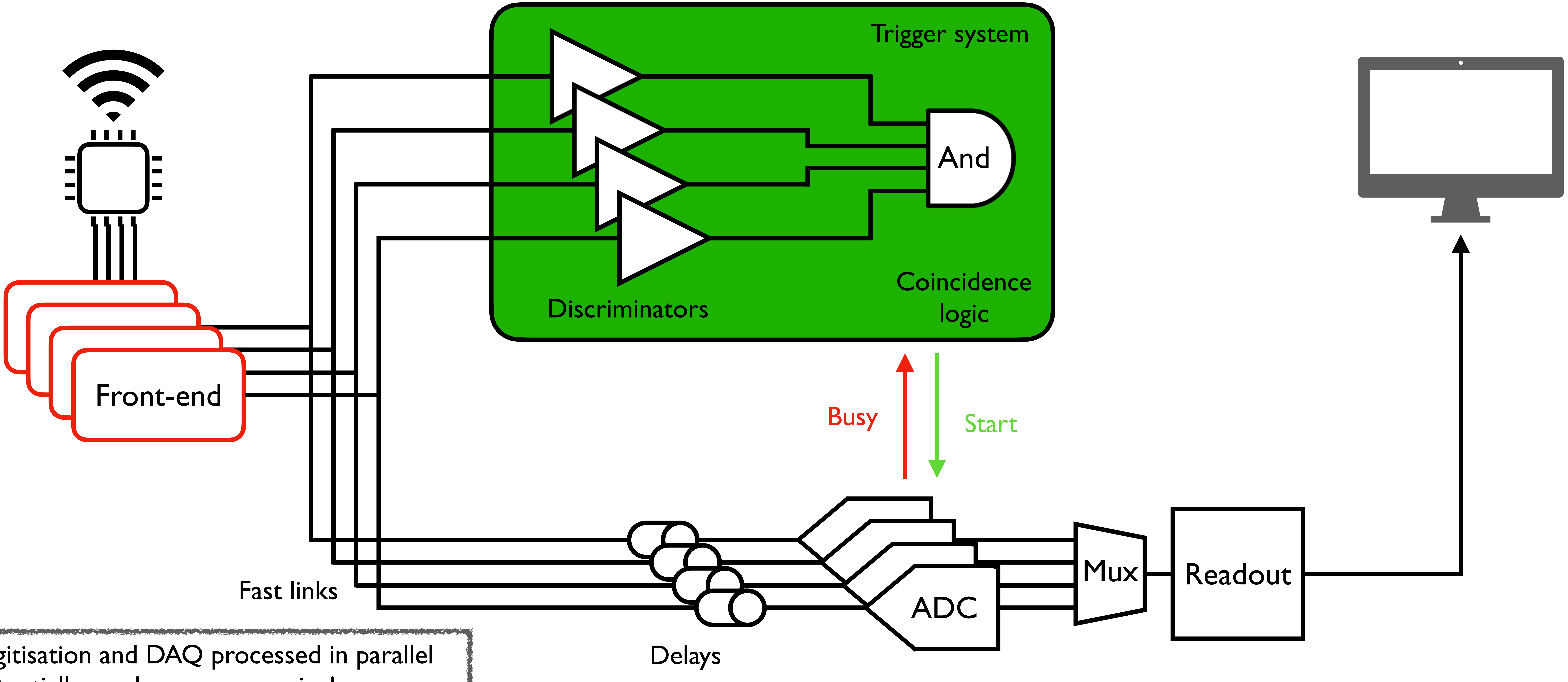
(i) Cannot accept any more events until writing finished after start writing to disk
(ii) Same principle applies to processing time (ex ADCs)

Simple trigger system



If ADC is the critical step for dead-time, this is clearly a really bad plan

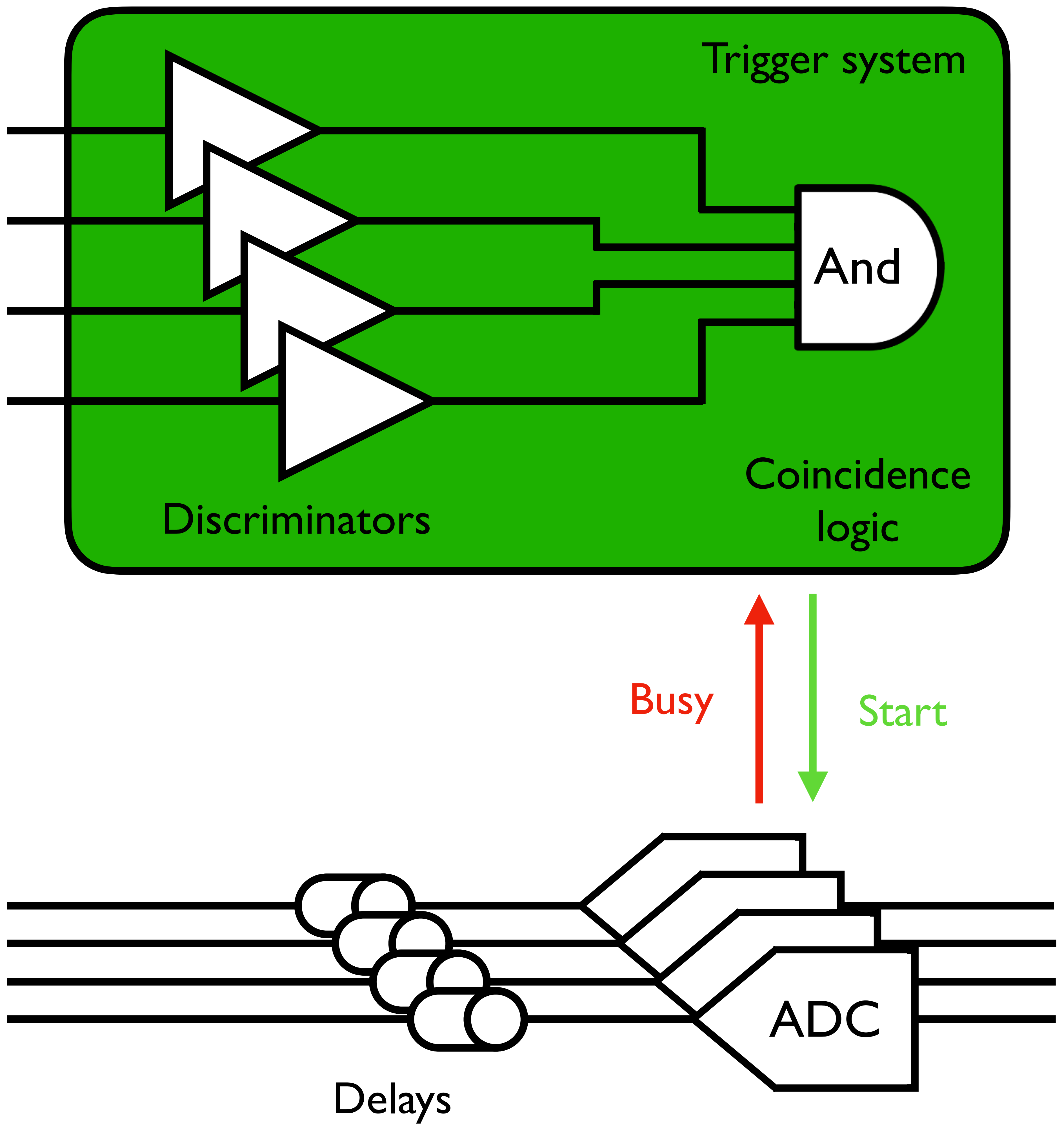
Simple trigger system



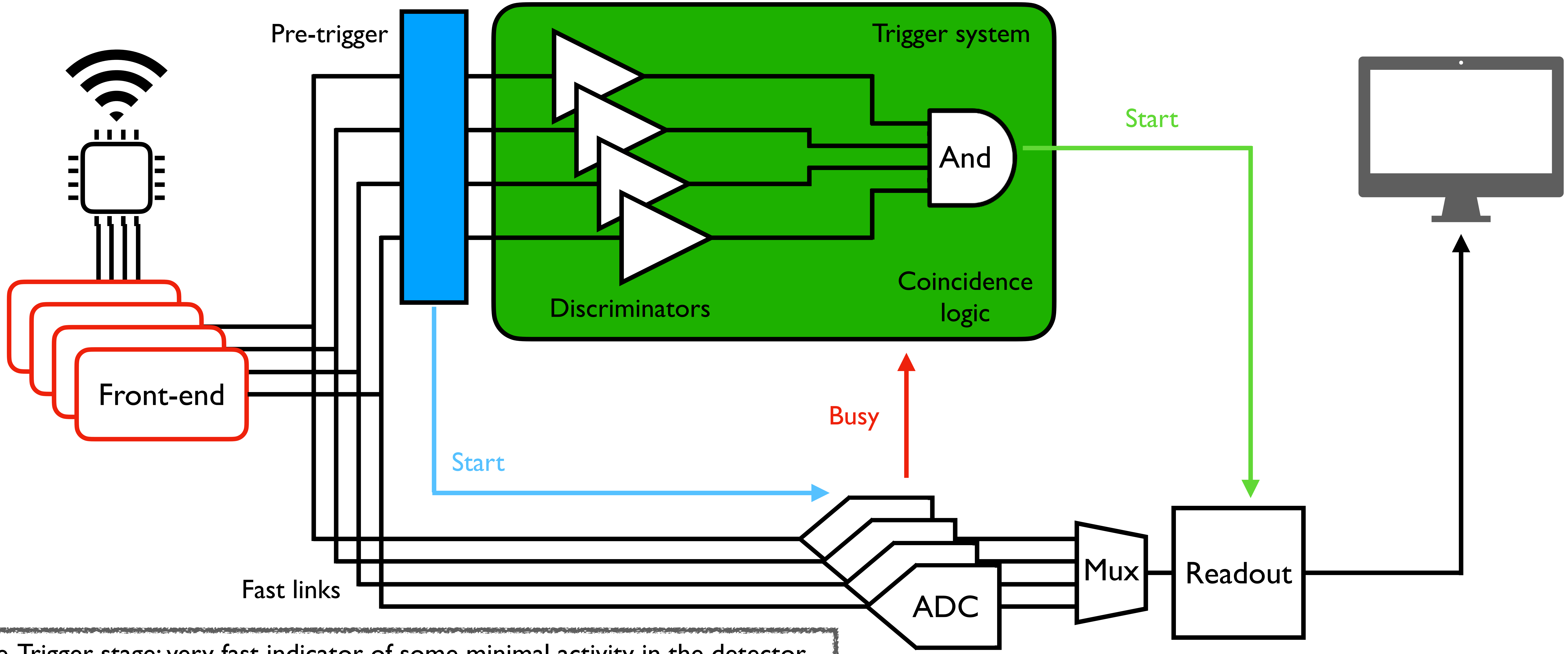
Digitisation and DAQ processed in parallel
Potentially much more expensive!

Latency

- Time to form the trigger decision and distribute to the digitizers
- Signals must be delayed until the trigger decision is available
- Latency is longer with a more complex selection

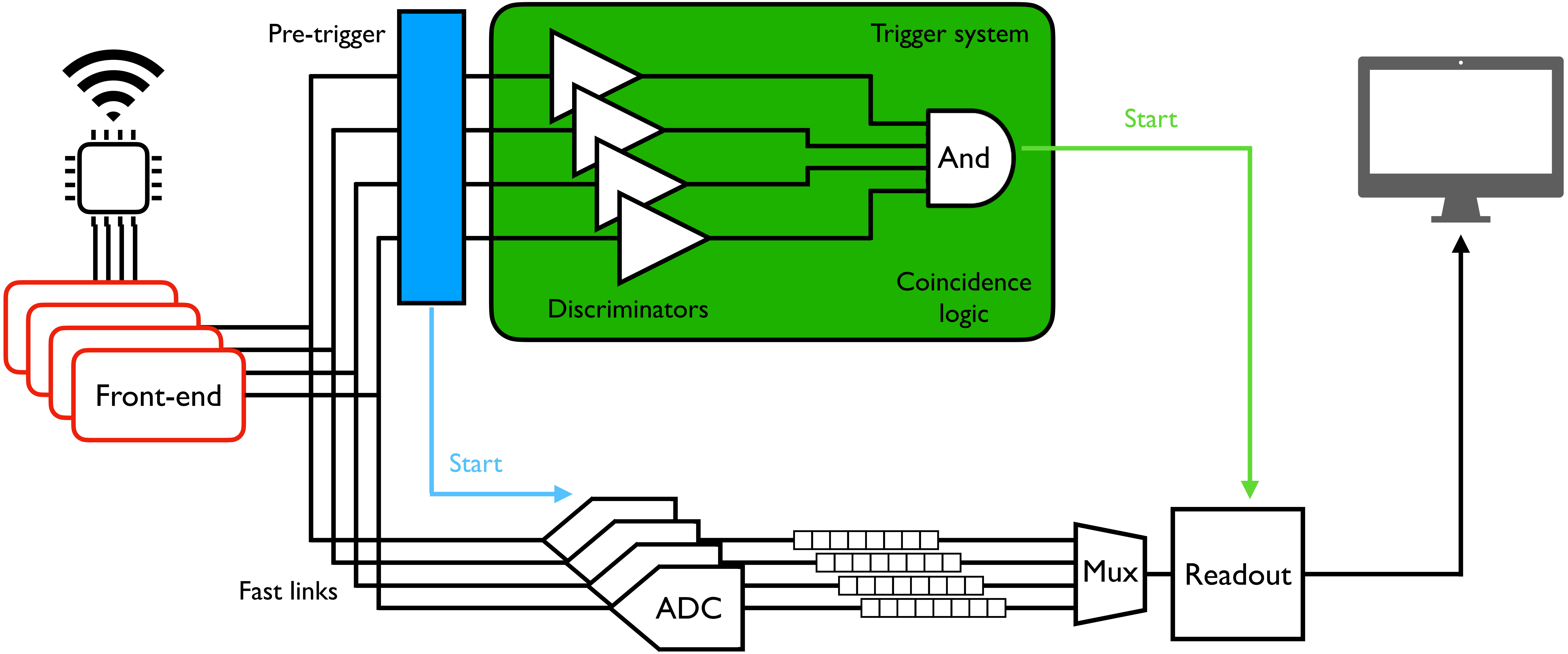


Simple trigger system



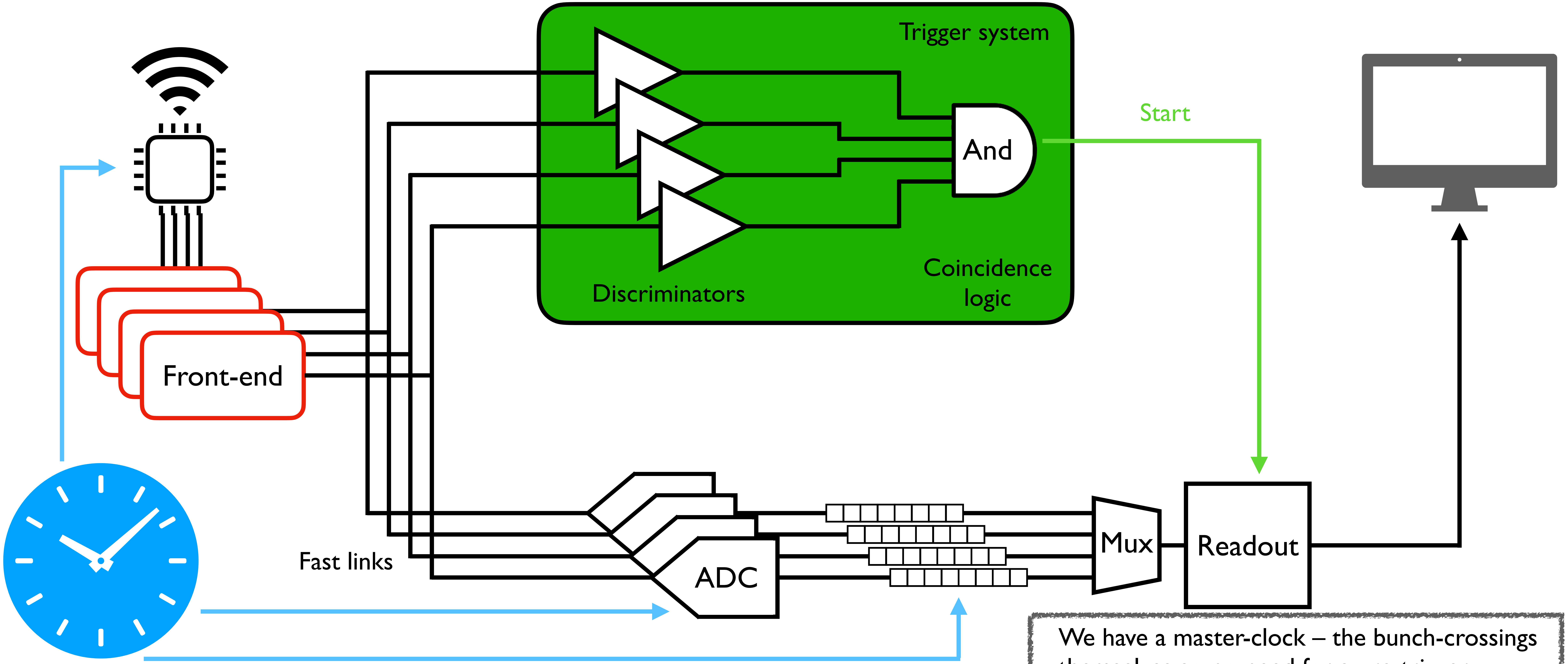
Pre-Trigger stage: very fast indicator of some minimal activity in the detector
Used to start the ADCs, with no delay

Simple trigger system



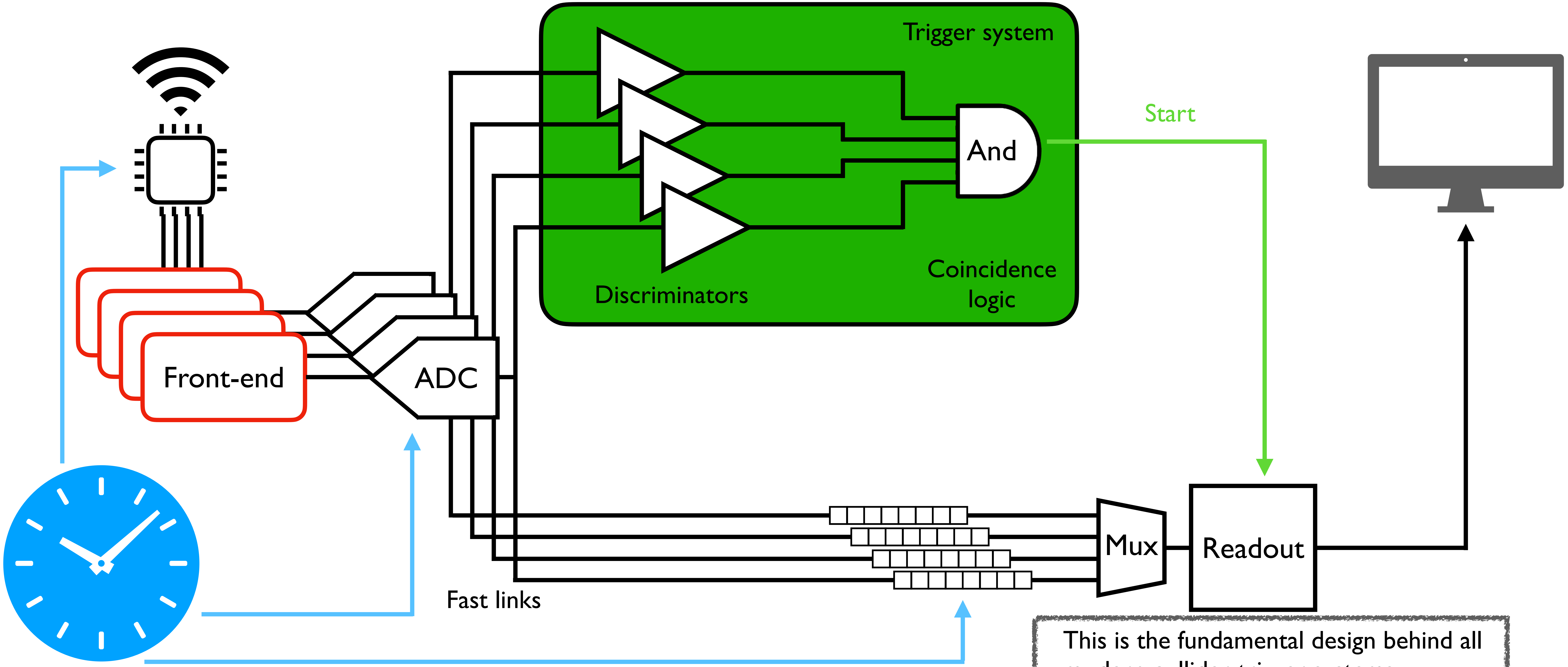
Can store the result of each digitization in RAM until trigger decision is made

Simple trigger system



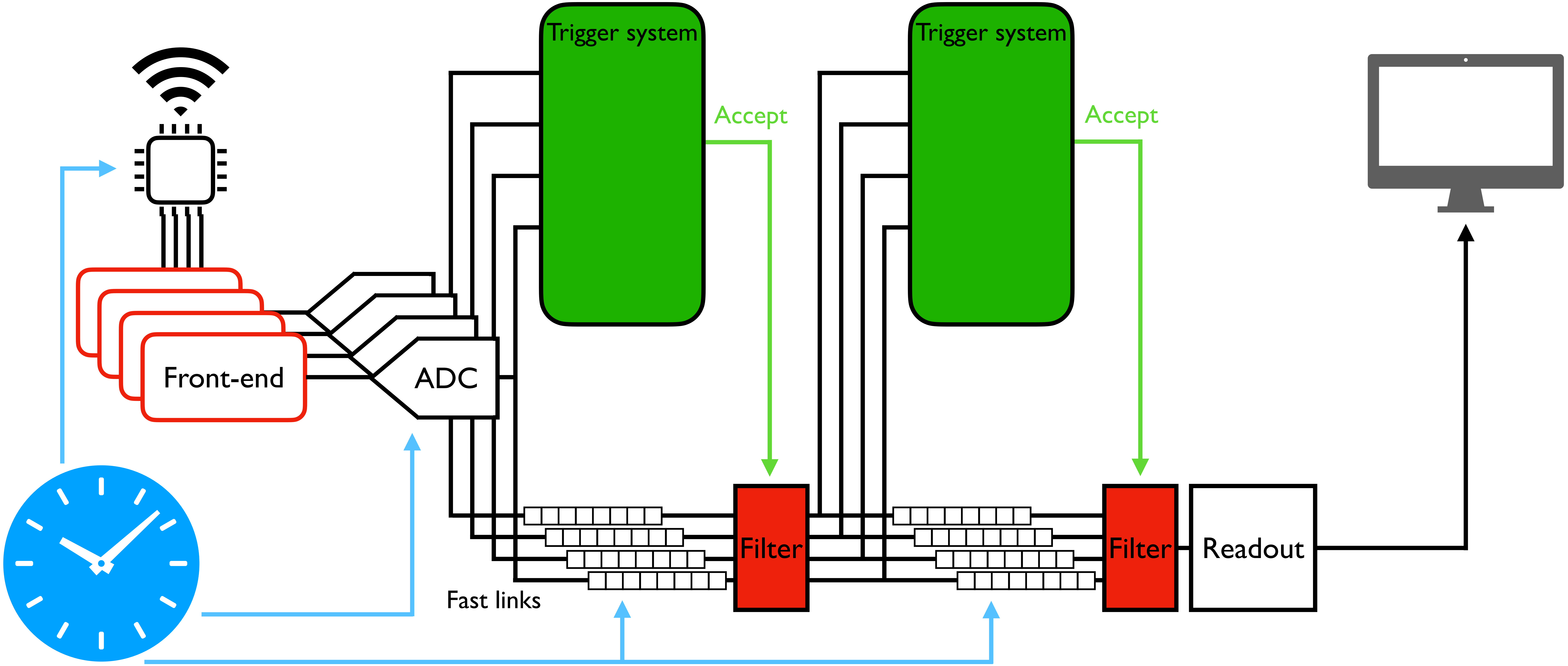
We have a master-clock – the bunch-crossings themselves so no need for a pre-trigger

Simple trigger system

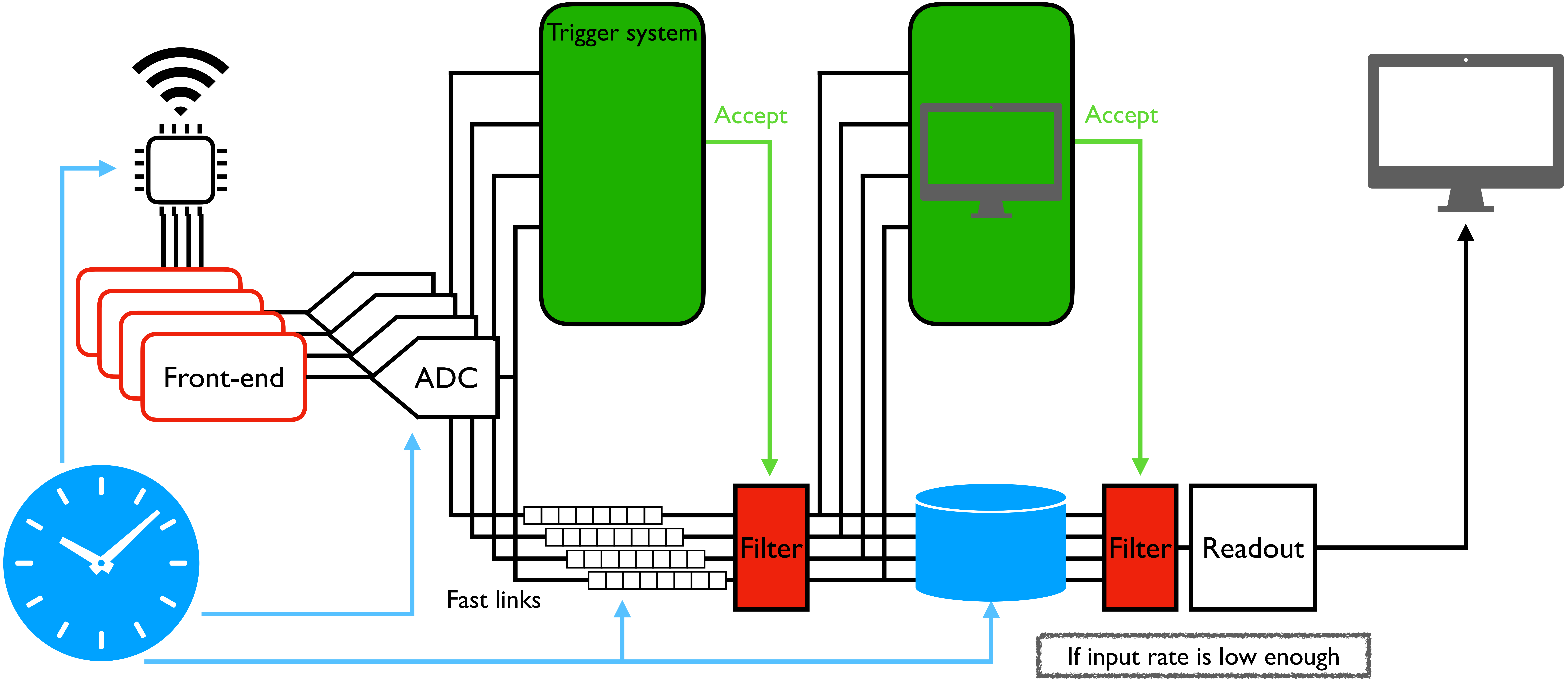


This is the fundamental design behind all modern collider trigger-systems

Multilayer trigger system



Multilayer trigger system



Multilayer trigger system in CMS

- Trigger System

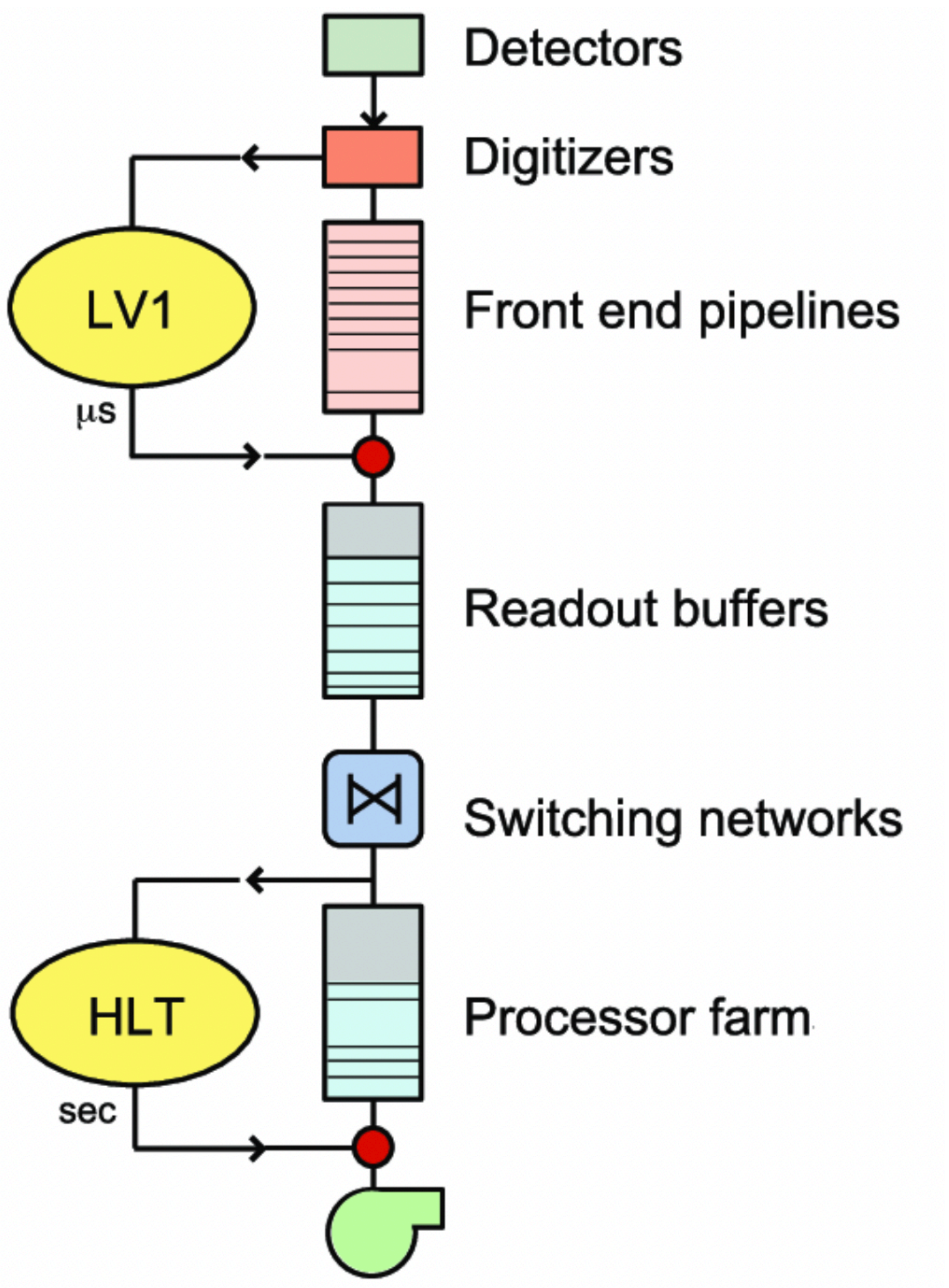
(i) Reduce input rate (40 MHz) to a data rate (~1 kHz) that can be stored, reconstructed and analyzed Offline maximizing the physics reach of the experiment

- Level 1 Trigger

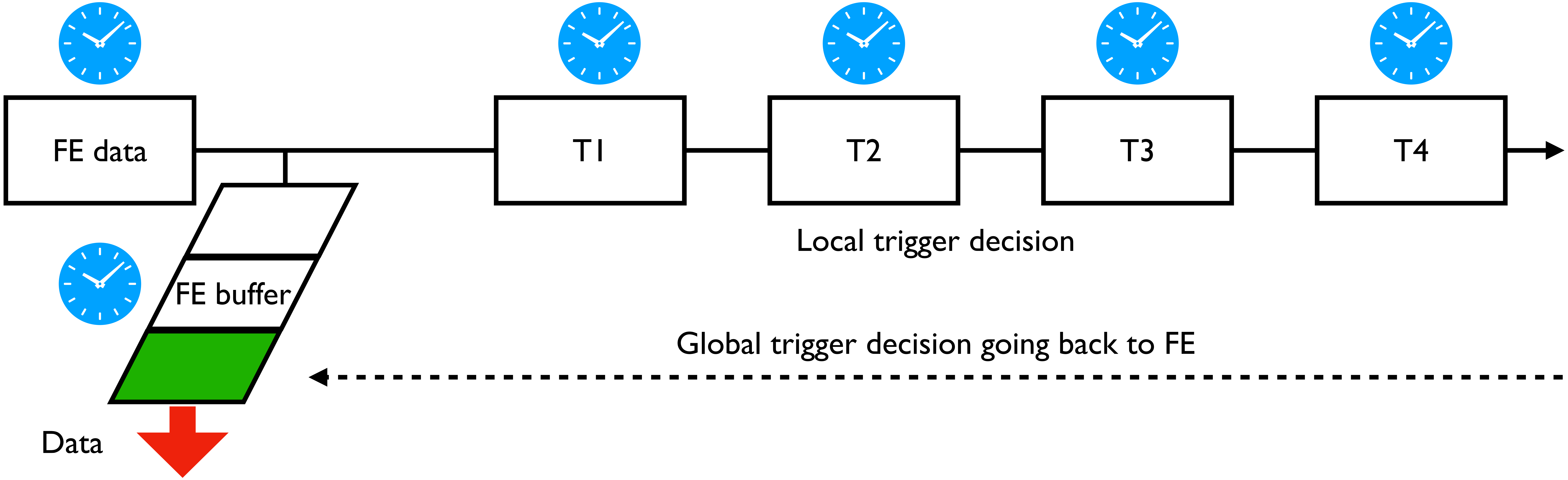
(i) Coarse readout of the Calorimeters and Muon detectors
(ii) Implemented in custom electronics, ASICs and FPGAs
(iii) Output rate limited to 100 kHz by the readout electronics

- High Level Trigger

(i) Readout of the whole detector with full granularity
(ii) Based on the CMS software, running on 22,000 CPU cores (before Run3)
(iii) Output rate limited to an average of ~1 kHz by the Offline resources



Synchronous trigger

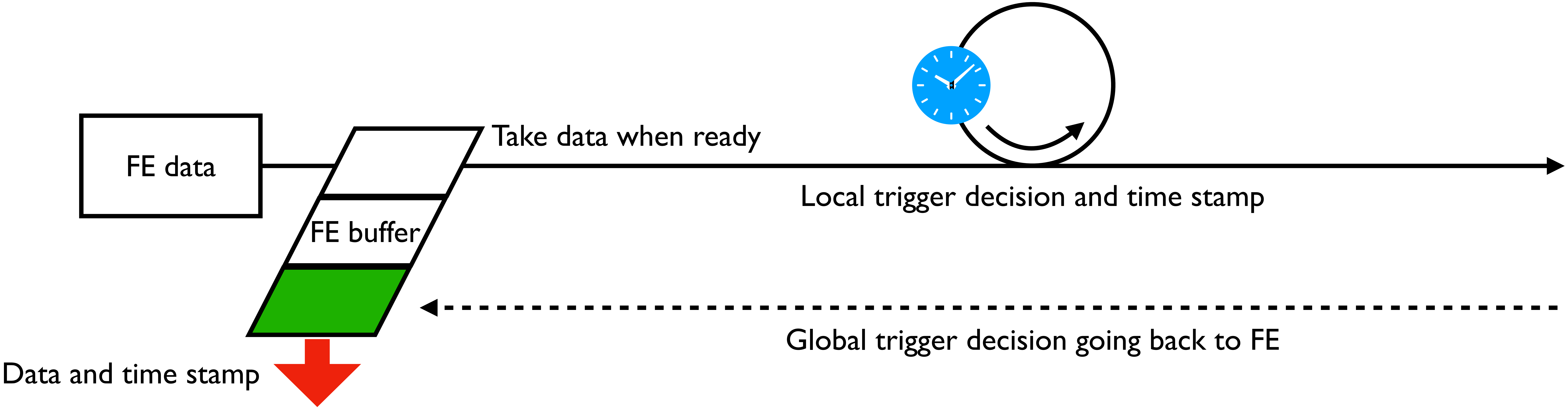


- Synchronous: operates phase-locked with master clock

(i) Data move in lockstep with the clock through the trigger chain
(ii) Fixed latency
(iii) The data, held in storage pipelines, are either sent forward or discarded
(iv) Used for L1 triggers in collider experiments, exploiting the accelerator bunch crossing clock

Pro's: dead-time free (just few clock cycles to protect buffers)
Con's: cost (high frequency stable electronics, sometimes needs to be custom made); maintain synchronicity throughout the entire system, complicated alignment procedures if the system is large (software, hardware, human…)

Asynchronous trigger



- Asynchronous: operations start when data ready or last processing is finished

(i) Used for larger time windows
(ii) Average latency (with large buffers to absorb fluctuations)
(iii) If buffer size ≠ dead-time → lost events
(iv) Used for HLT

Pro's: more resilient to data burst; running on conventional CPUs
Con's: needs a timing signal synchronised to the FE to latch the data, needs time-marker stored in the data, data transfer protocol is more complex)

HL-LHC

- Phase II upgrade of CMS trigger system

(i) Readout rate increased from 100 kHz to 750 kHz
(ii) Latency increased from 4 μs to 12.5 μs

- Tracking Triggers

(i) Hardware-based trigger at first stage (in Level 1 Trigger)
(ii) Reconstruct trajectories of charged particles with pT > 2 GeV for all pp interactions
(iii) Improved particle identification

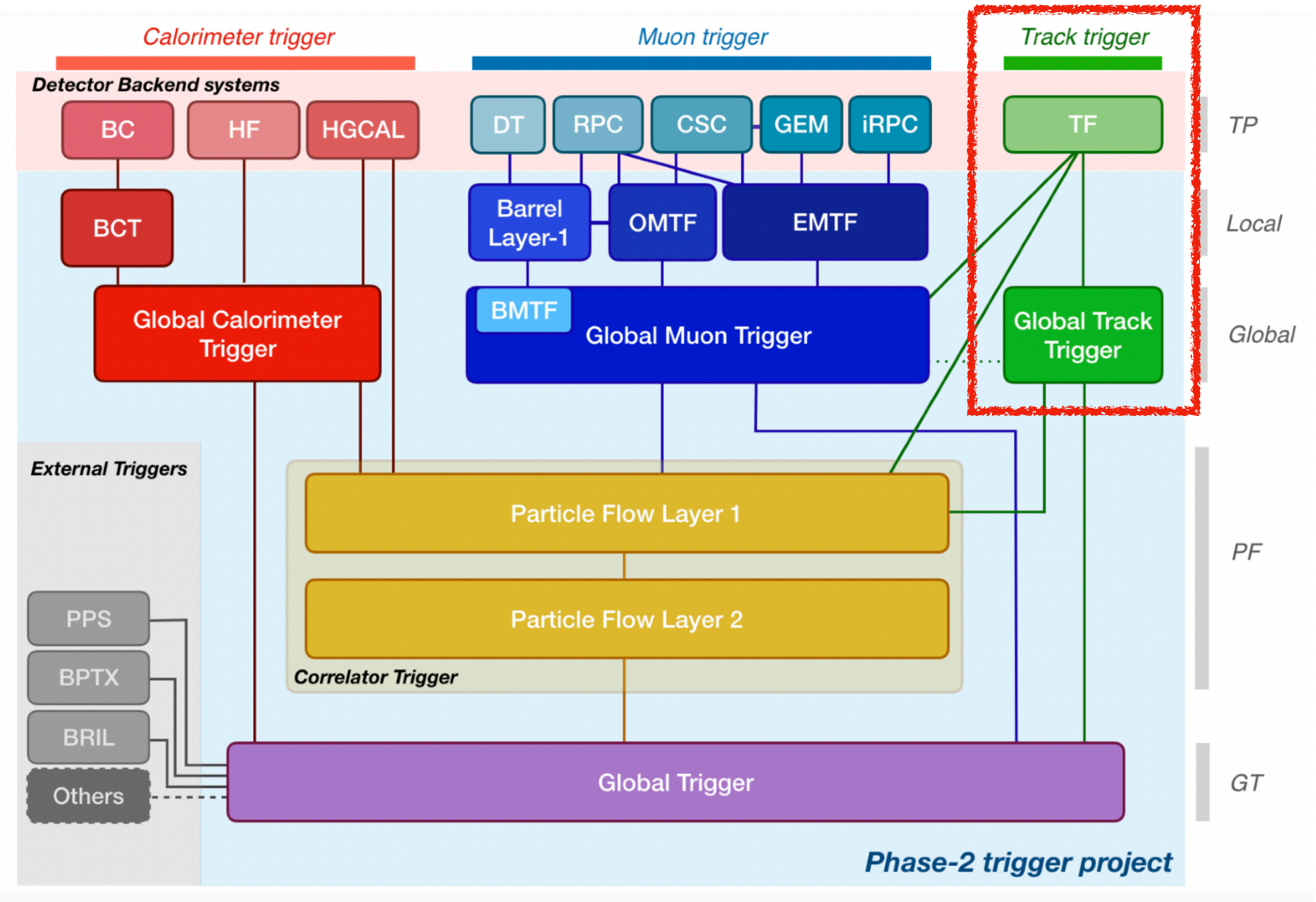
- GPU application

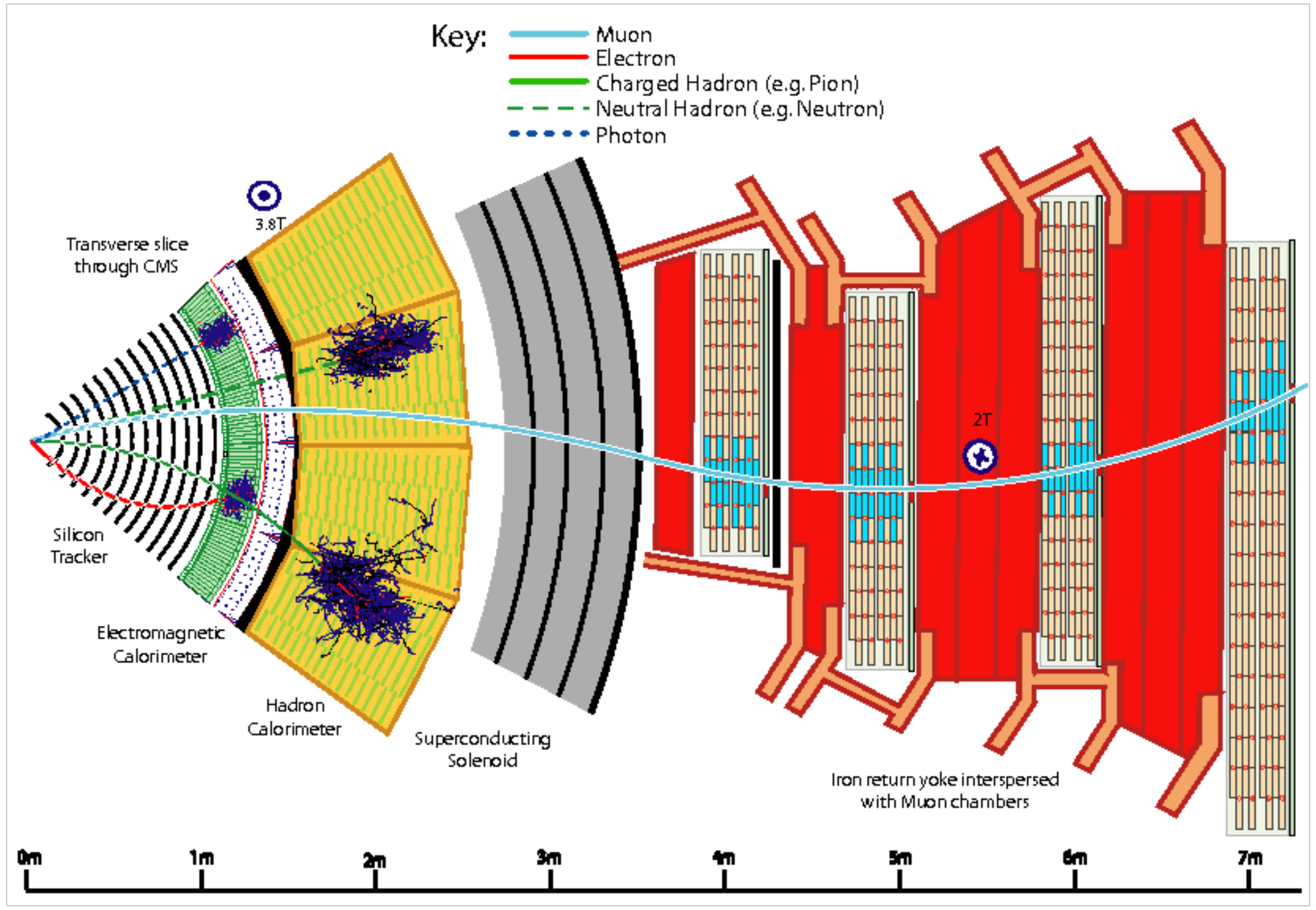
(i) Improve performance of HLT

- Fast links

(i) Up to 10 Gbps is required

Tracking Triggers



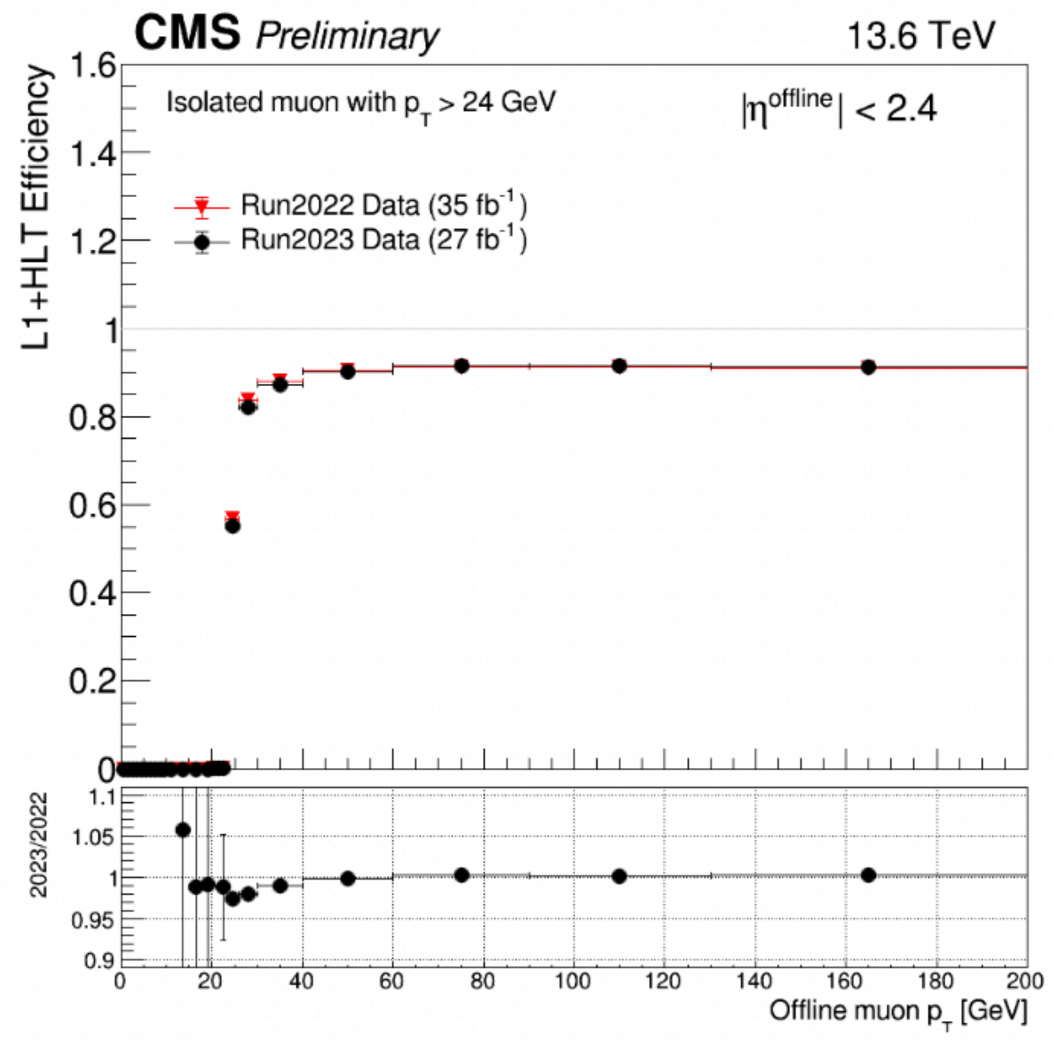


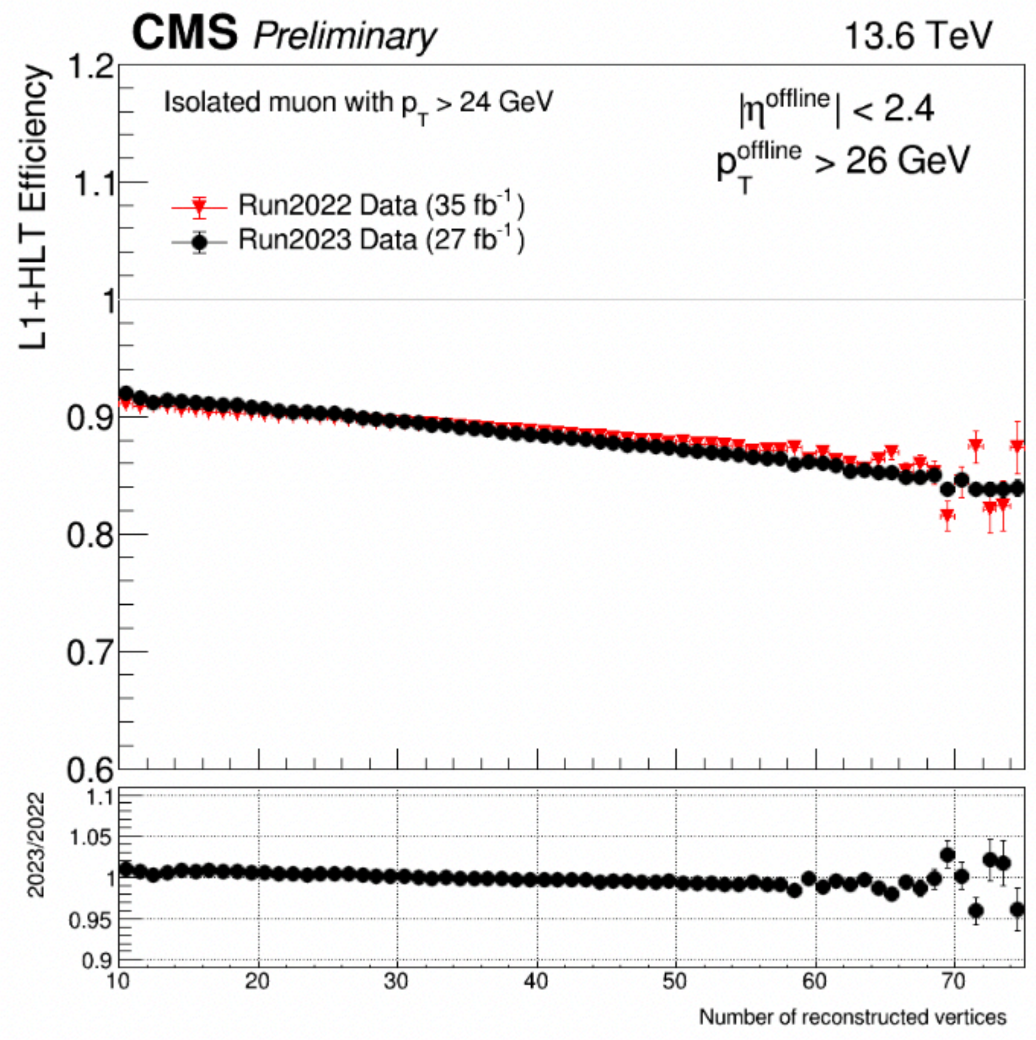
- CMS L1 trigger system for HL-LHC

(i) Barrel and endcap calorimeter trigger systems will process high-granularity information from the calorimeters
(ii) Barrel, endcap, and overlap muon track finding systems will provide triggering of muons up to |η| < 2.5
(iii) The Global Track Trigger will reconstruct primary event vertices and define track-only based trigger objects
(iv) A two-stage correlator system will match L1 tracks with information from the calorimeter and muon systems, and perform a L1-adopted version of full event reconstruction to identify physics objects such as e, γ, μ, hadronic τ, jets, and energy sums

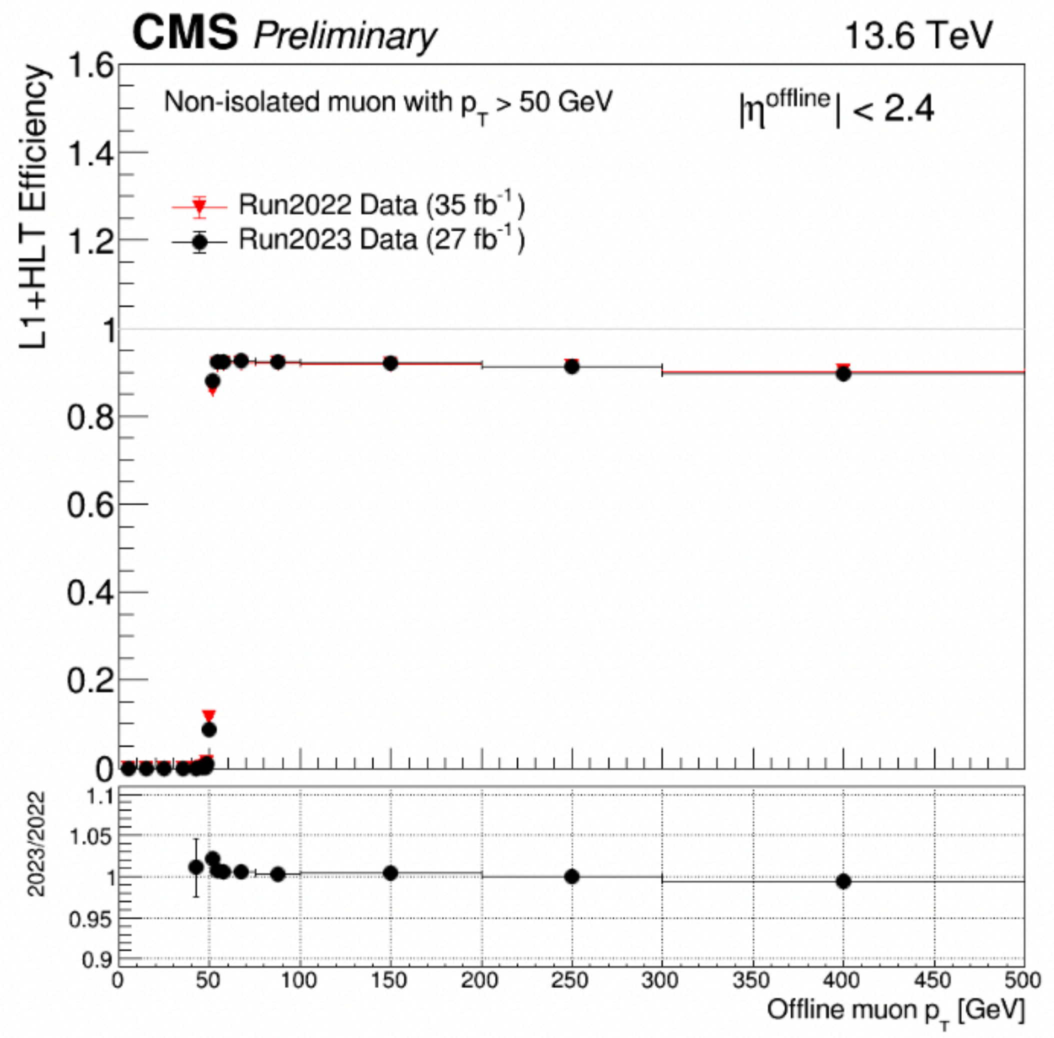
Muon Triggering in CMS

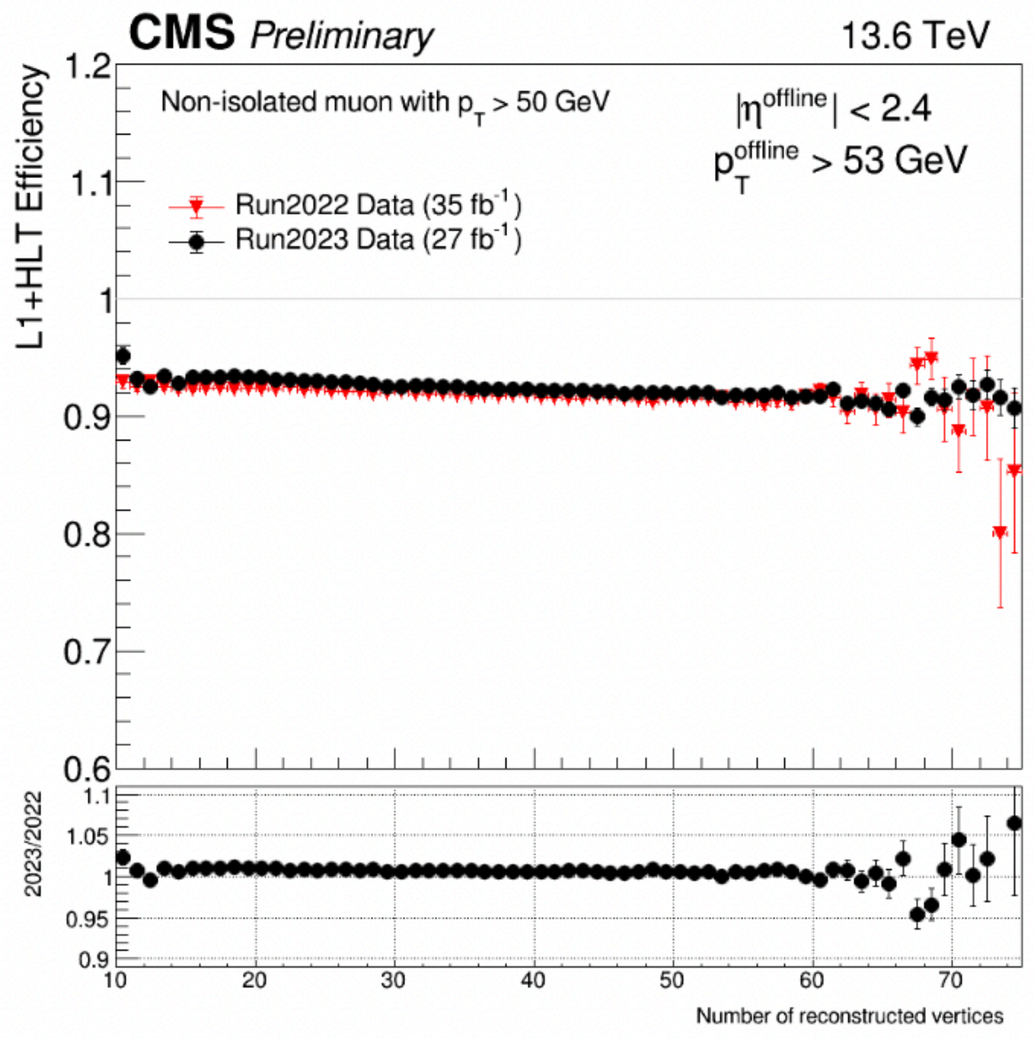
IsoMu24

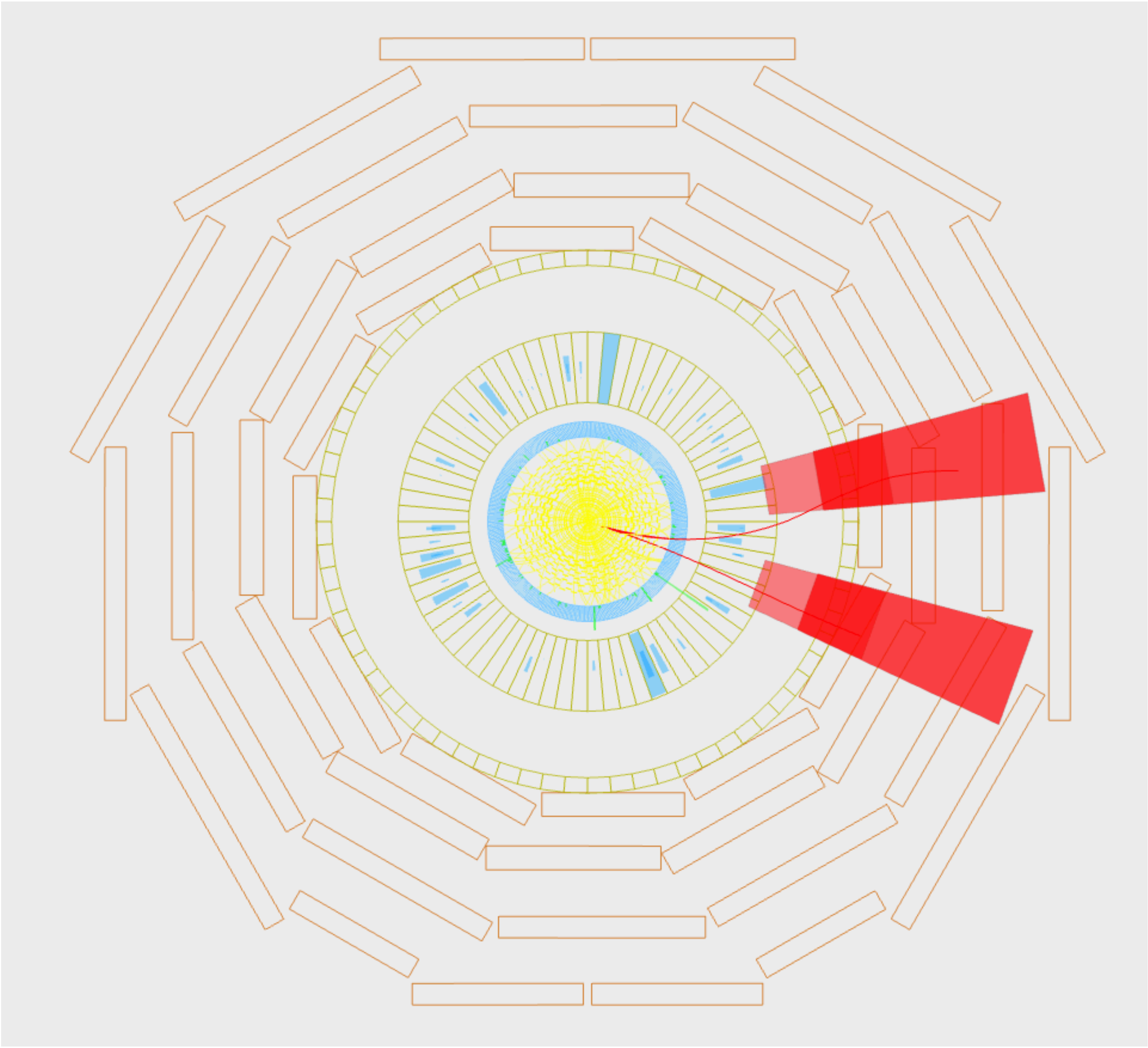


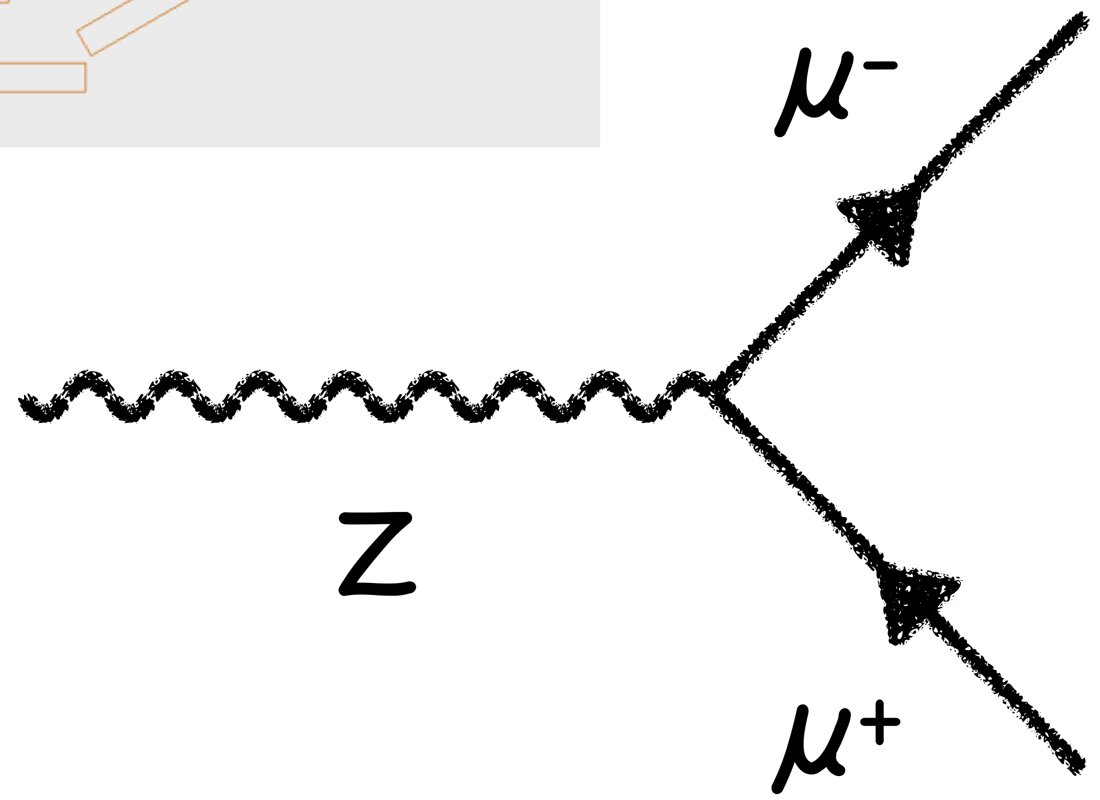


Mu50



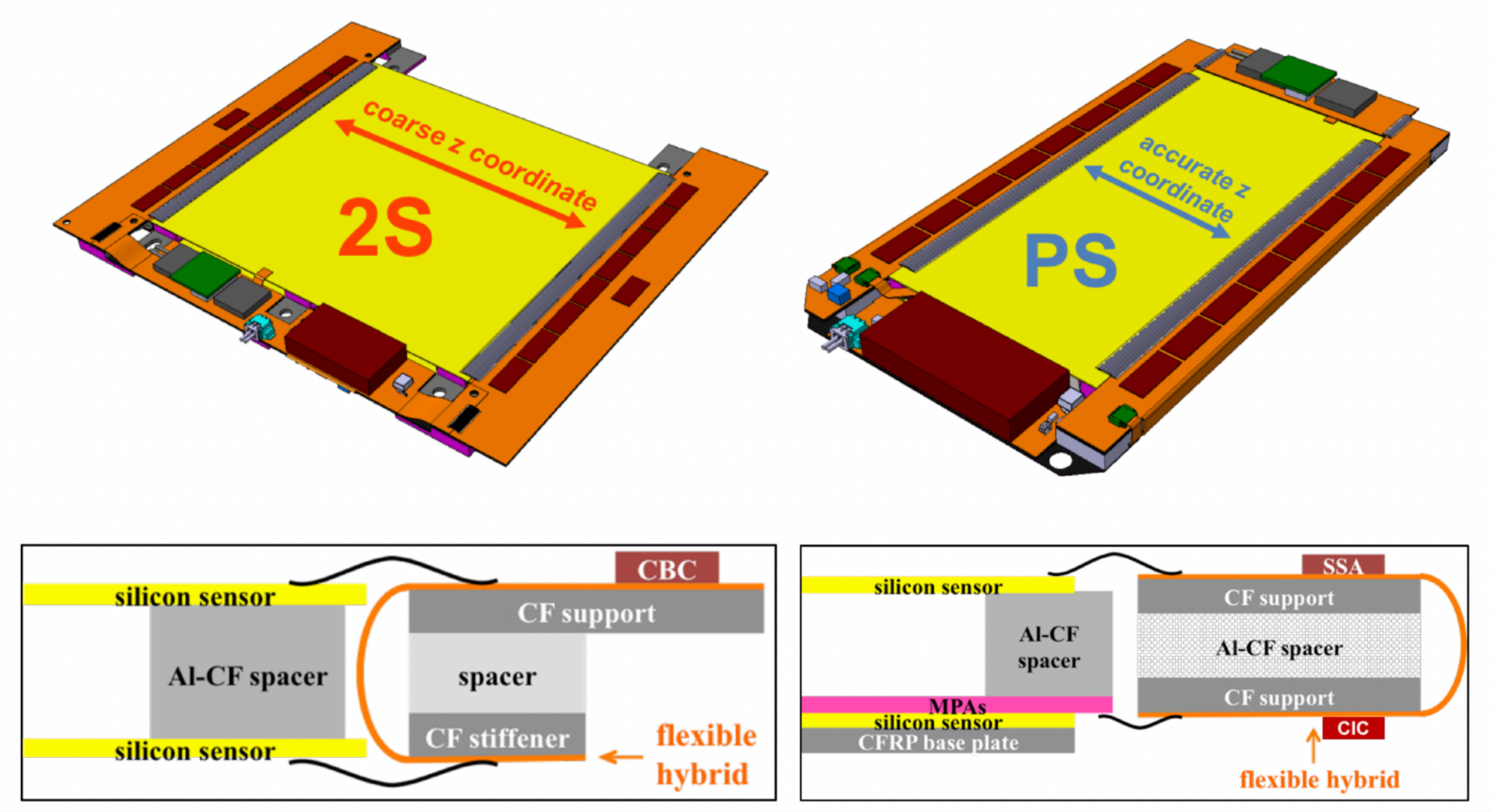






- Trigger turn-on curve
- Pileup dependency

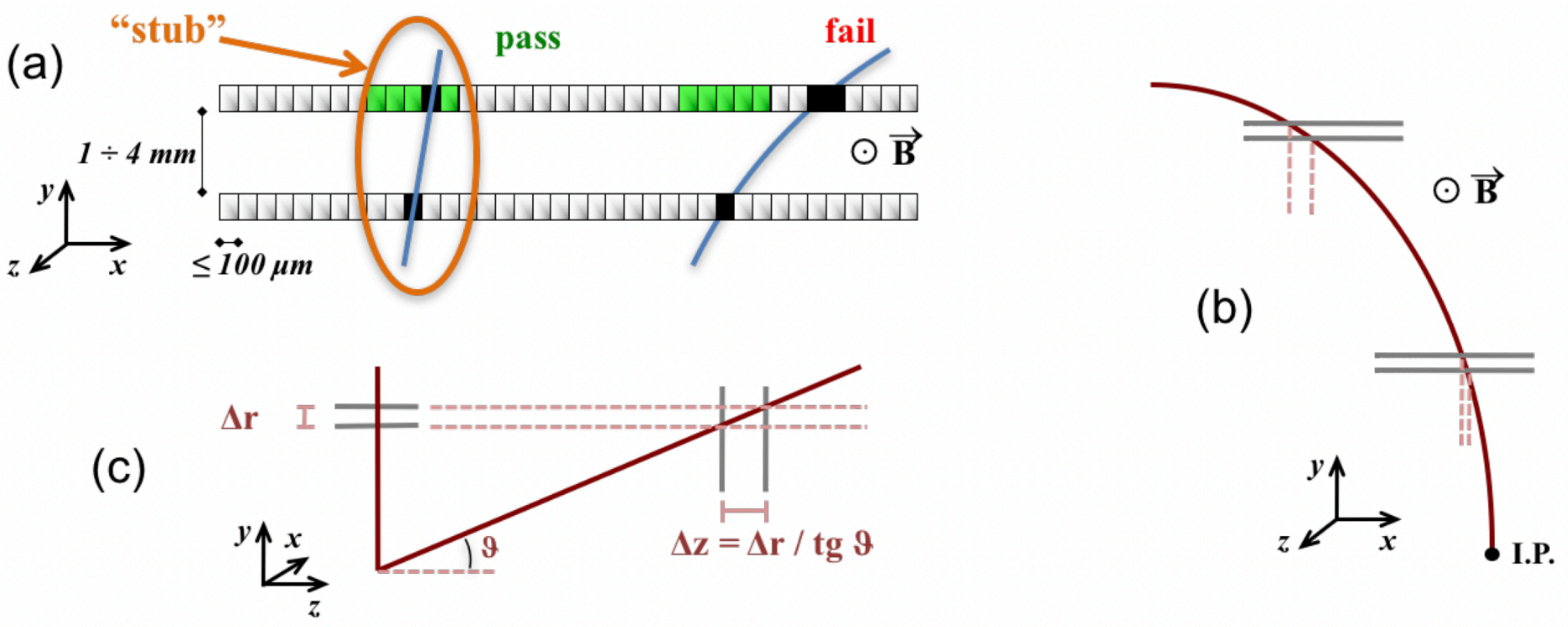
Tracking Triggers



- Outer tracker modules

(a) Pixel-Strip (PS) module: one sensor tier consists of macro-pixels (1.446 mm long by 100 μm) and the other tier of strips (2.4 cm long)

(b) Strip-Strip (2S) module where both sensor tiers are strips (5 cm long and 90 μm pitch)



- Stub formation

(a) Channels in green show the acceptance window to form a stub from the hit indicated in the inner sensor

(b) Same p_T corresponds to a larger distance between the two hits at larger radius for a given sensor spacing

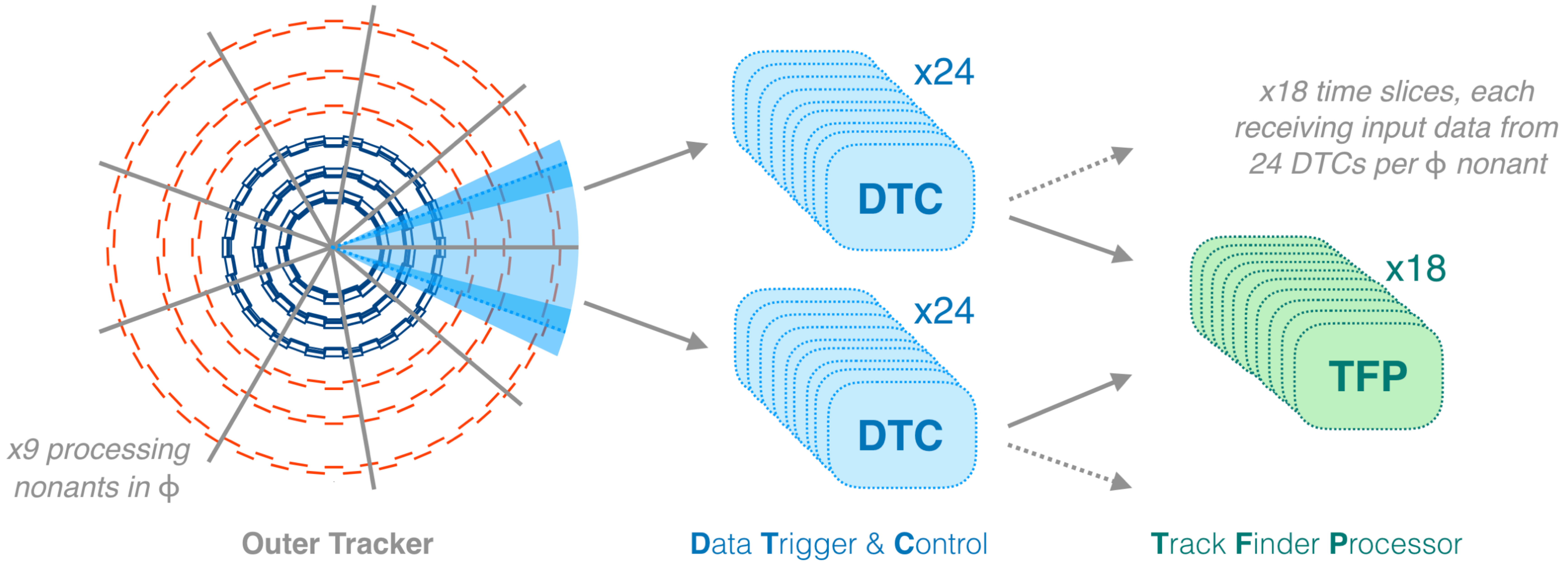
(c) For the endcap discs, a larger spacing between the sensors is needed to achieve the same discriminating power as in the barrel at the same radius

System architecture

- Two stages of data processing, with significant parallelization

(i) Split the detector into smaller geometrical region
(ii) Utilize time multiplexing to distribute the tasks of performing the pattern recognition, track fitting, and duplicate removal to many different hardware components
(iii) Process the roughly 15,000 stubs that arrive every 25 ns and find the tracks within about 4 μs

- All-FPGA processing system



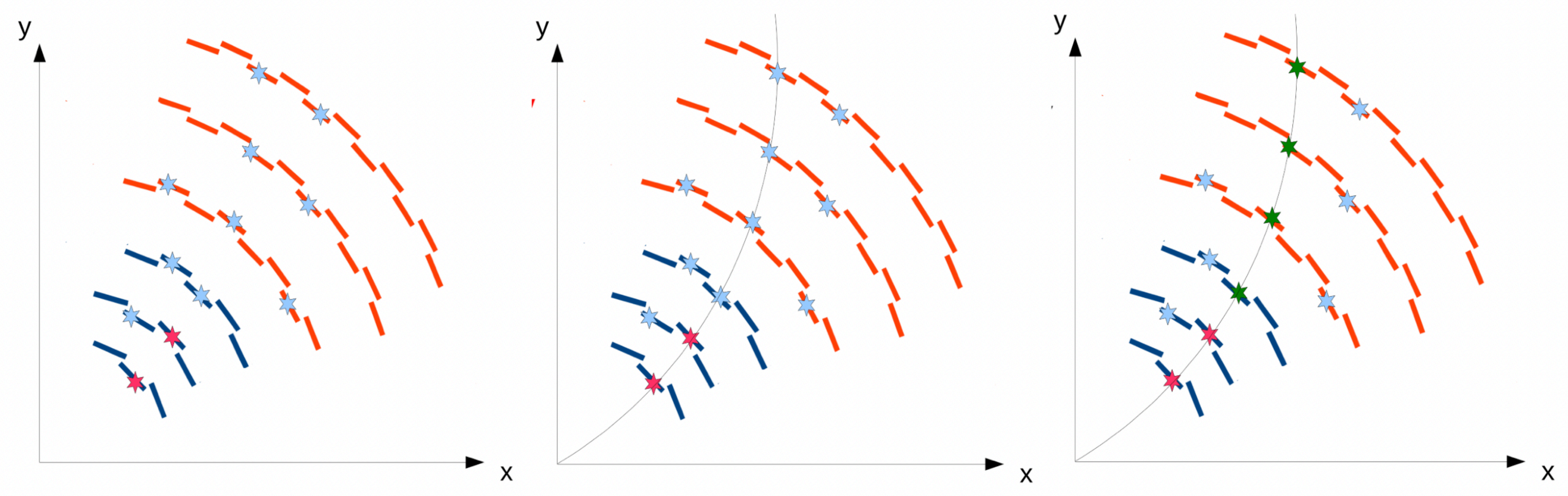
Hybrid algorithm

- Tracklet road search algorithm

(i) Tracklets, or the seeds, are formed from pairs of stubs in adjacent layers or disks
(ii) For each stub pair a trajectory is calculated, using the beam spot as a constraint in the transverse plane, and projections to other layers and disks computed
(iii) Using the projections to other layers and disks, matching stubs are found and used in the final track fit

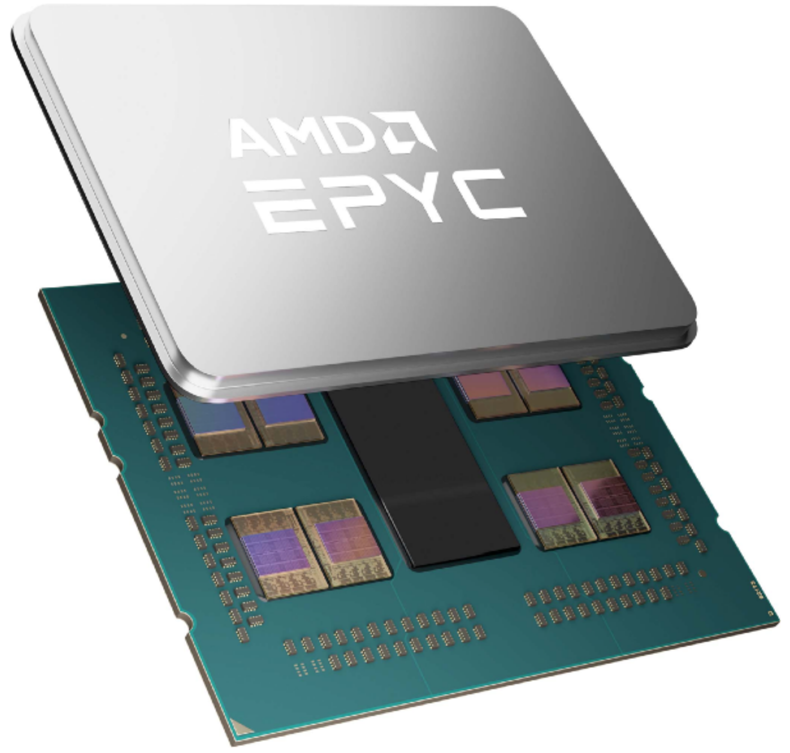
- Kalman filter track fit

(i) The Kalman state is updated iteratively, starting from stubs in the innermost layers
(ii) The seed trajectory is projected to the next layer or disk and using the position information from the next stub



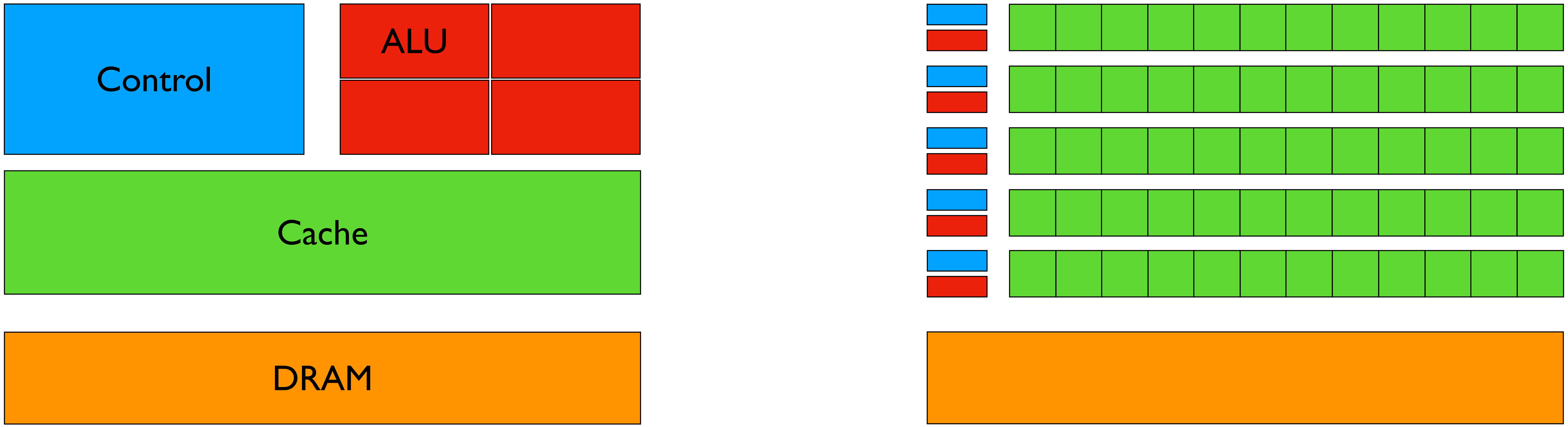
CPU vs GPU

- GPUs are used in HLT of CMS in Run3



Two AMD Milan 64-core CPUs and two NVIDIA Tesla T4 GPUs

CPU vs GPU



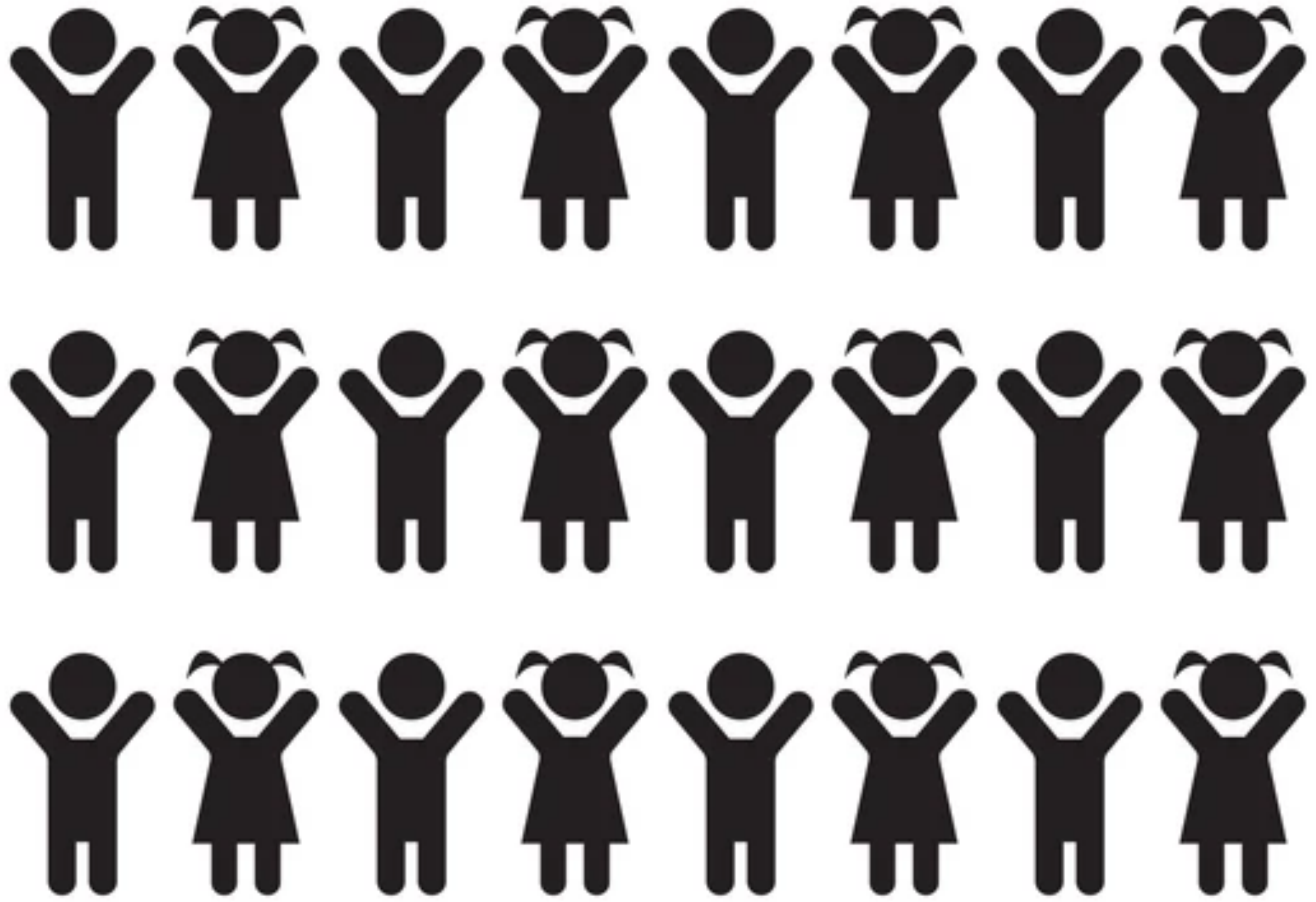
- CPU

(i) Powerful ALUs (Arithmetic Logic Units) and large Cache
(ii) Serial processing
(iii) Low bandwidth to memory (tens GB/s)
(iv) Suited to a wide variety of tasks

- GPU

(i) Many Streaming MultiProcessors consist of ALUs and control unit
(ii) Parallel processing
(iii) High bandwidth to memory (up to 1TB/s)
(iv) Designed for a specific purpose

CPU vs GPU



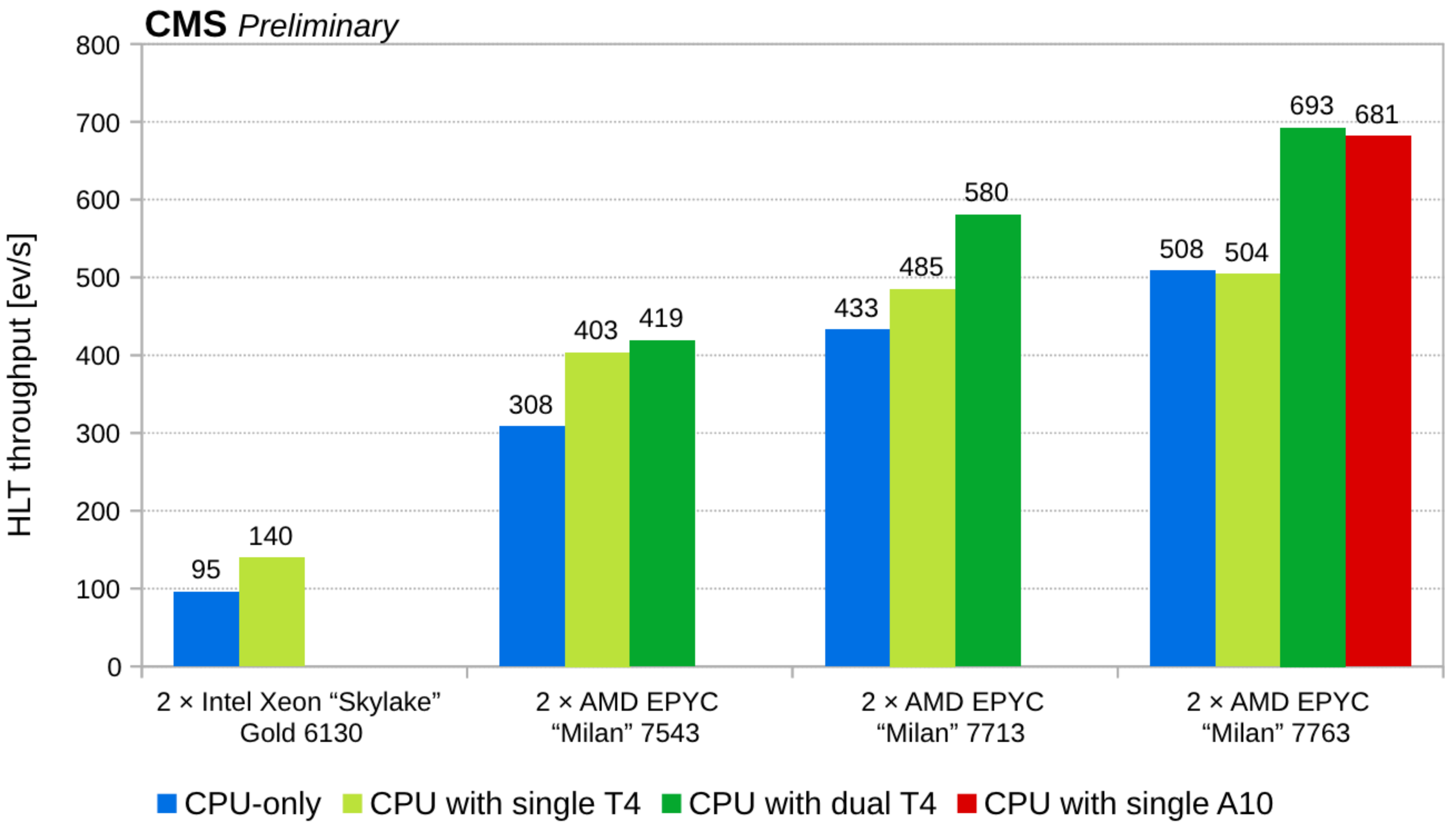
- CPU

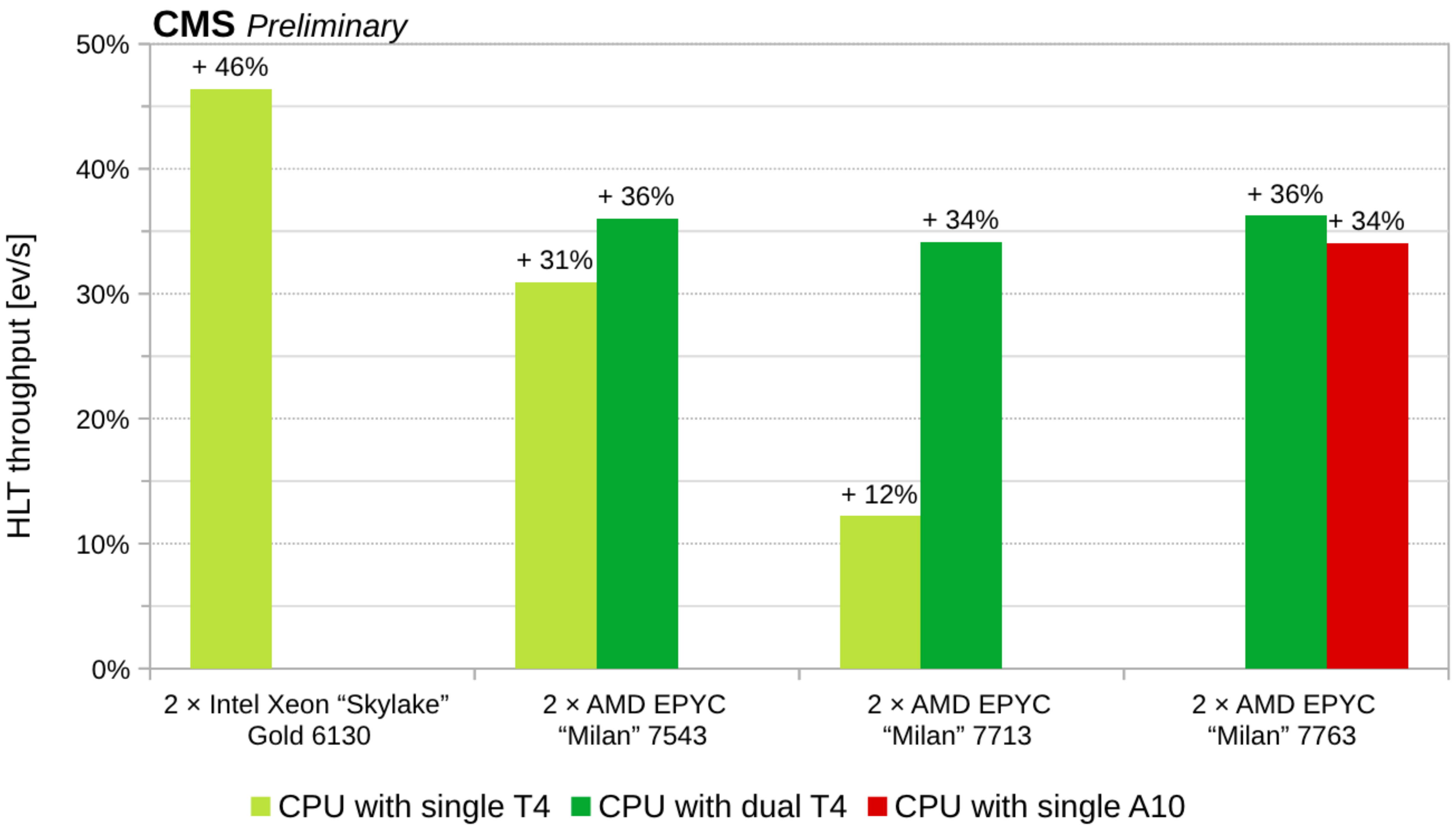
(i) Powerful ALUs (Arithmetic Logic Units) and large Cache
(ii) Serial processing
(iii) Low bandwidth to memory (tens GB/s)
(iv) Suited to a wide variety of tasks
(v) A professor can do any difficult math

- GPU

(i) Many Streaming MultiProcessors consist of ALUs and control unit
(ii) Parallel processing
(iii) High bandwidth to memory (up to 1TB/s)
(iv) Designed for a specific purpose
(v) Many kids can do lots of simple math

CPU vs GPU



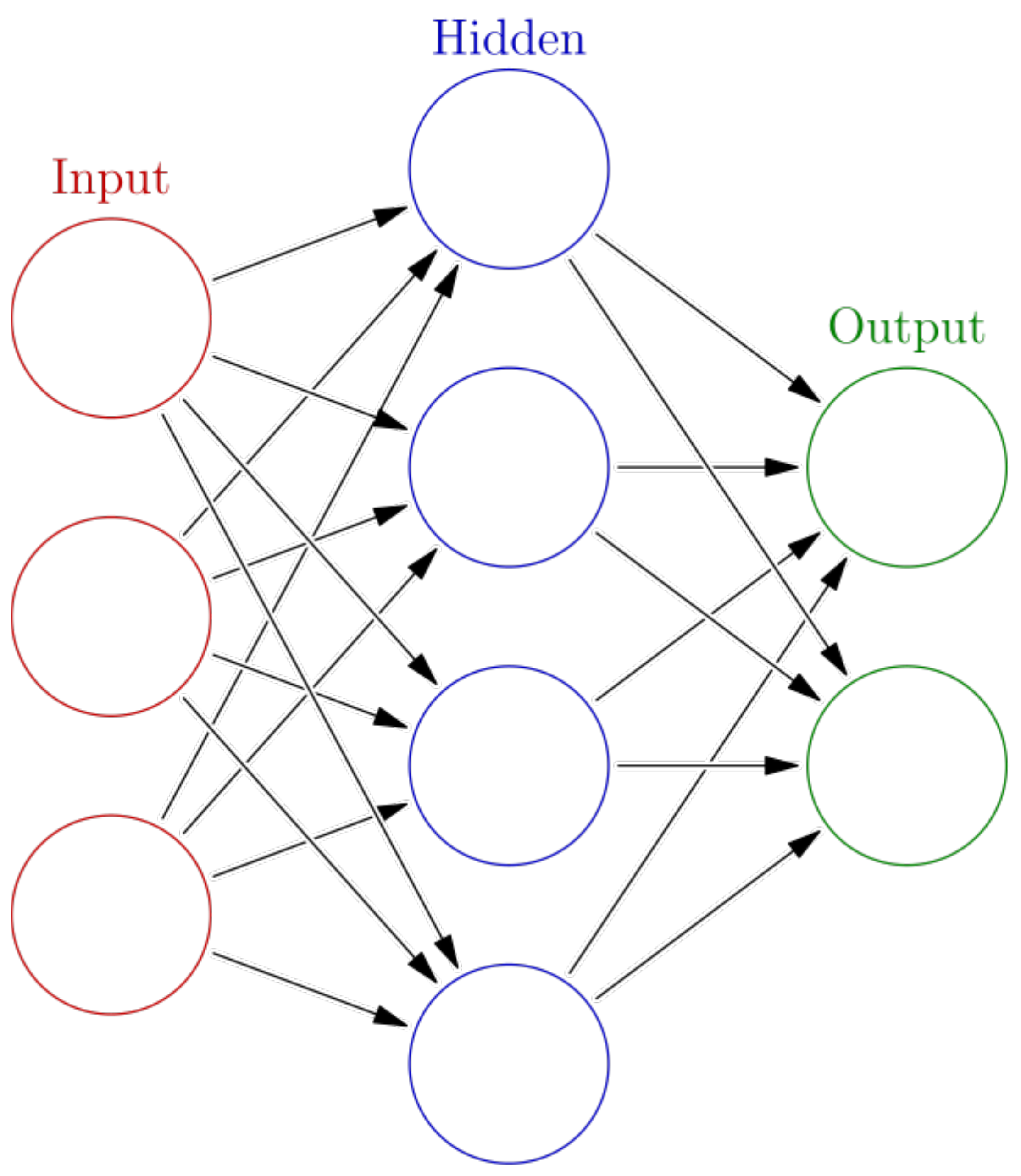


- Adding a second T4 to the dual 7543 gives only a small improvement, suggesting that these CPUs are evenly matched to a single T4

- As the performance of the CPUs increases this GPU becomes the limiting factor in the system performance, requiring a second GPU or a single more powerful one. When enough processing power is available from the GPUs, the performance improvement is stable around 35% for all AMD CPUs: the application is limited by the performance of the CPUs and the fraction of the algorithms that can be offloaded

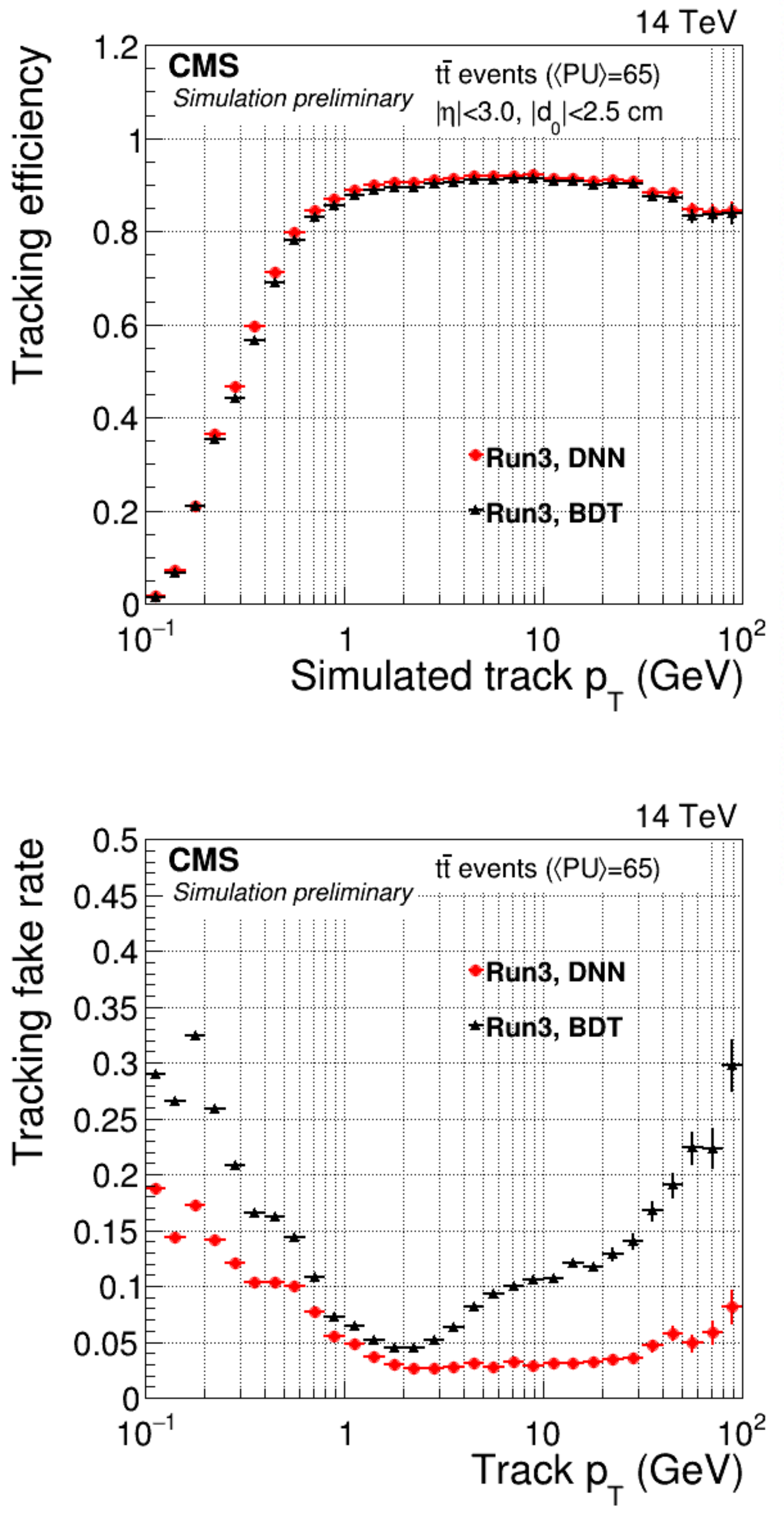
What else can GPU do for HEP?

Neural Net Work

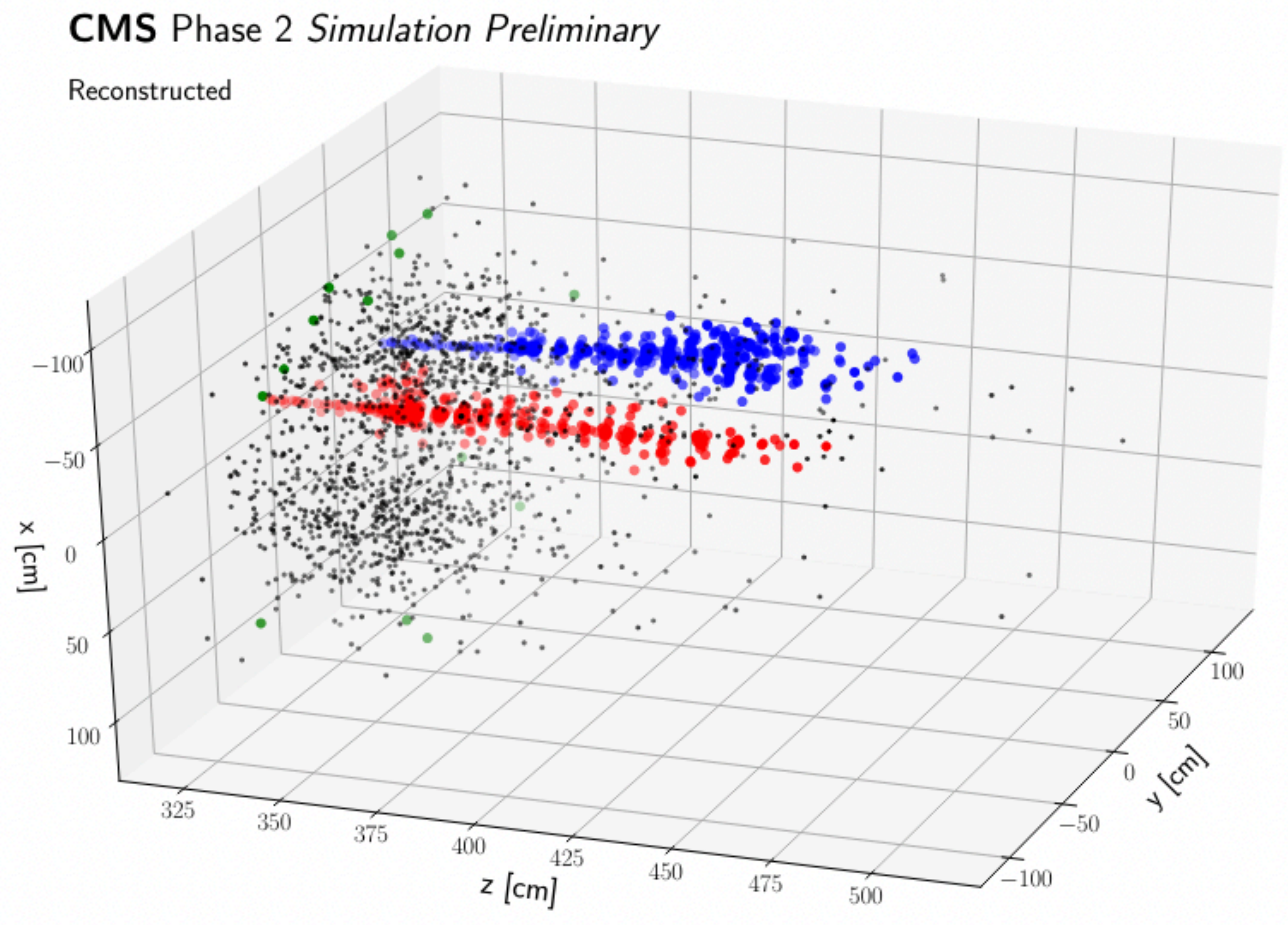


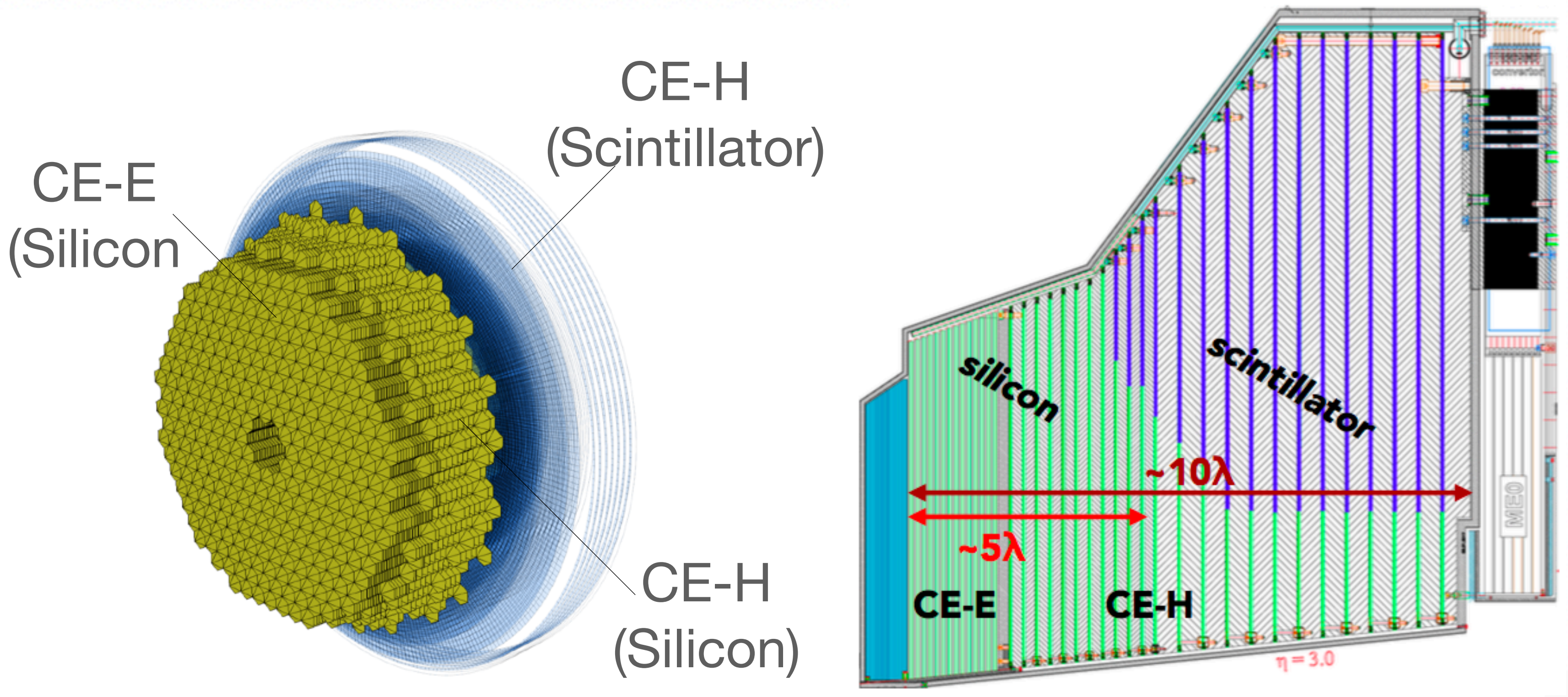
- Using interconnected nodes or neurons in a layered structure
- Can be used for data training
- Training speed can be improved by GPU

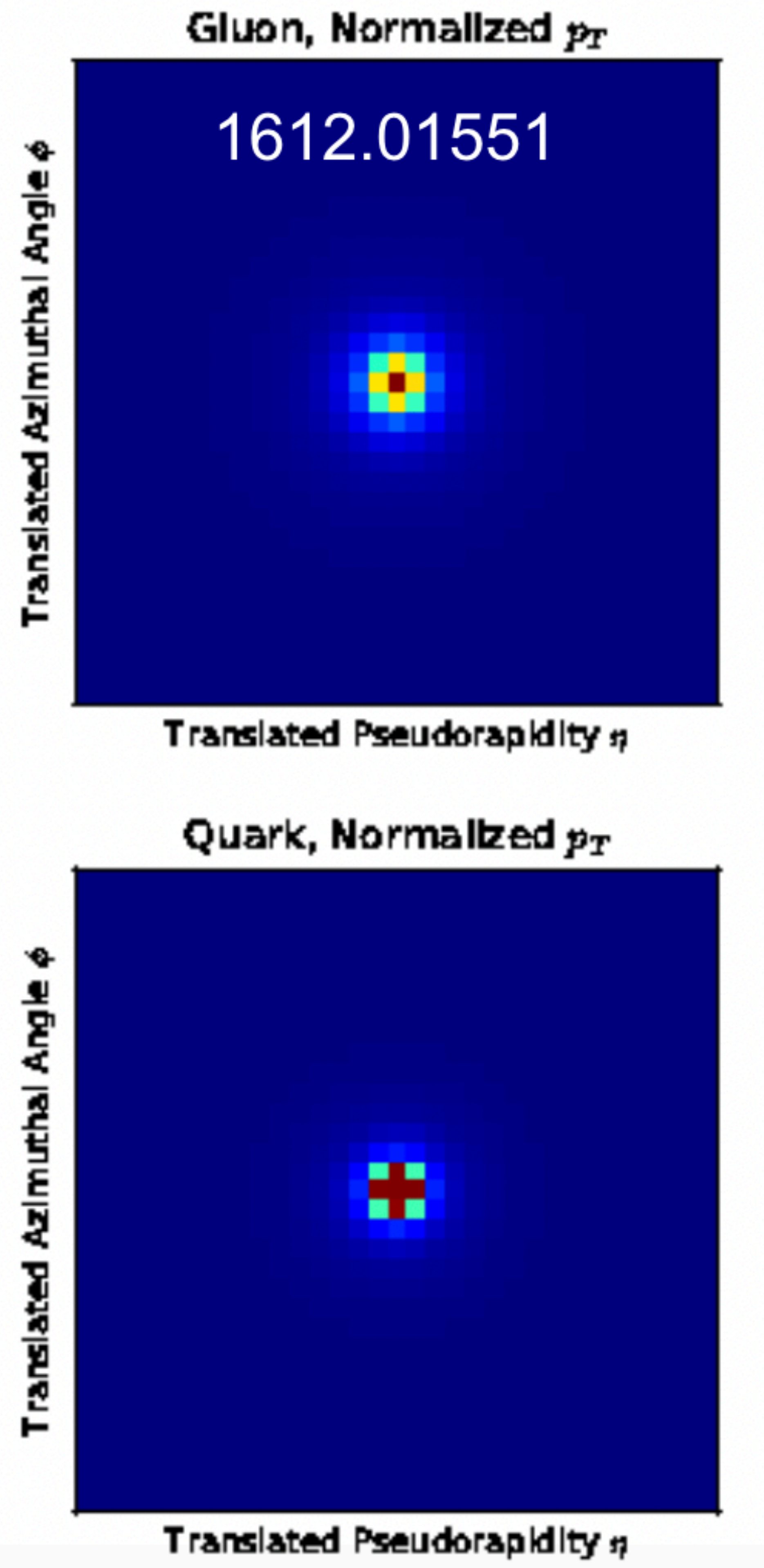
Deep learning



Particle Reconstruction and Identification

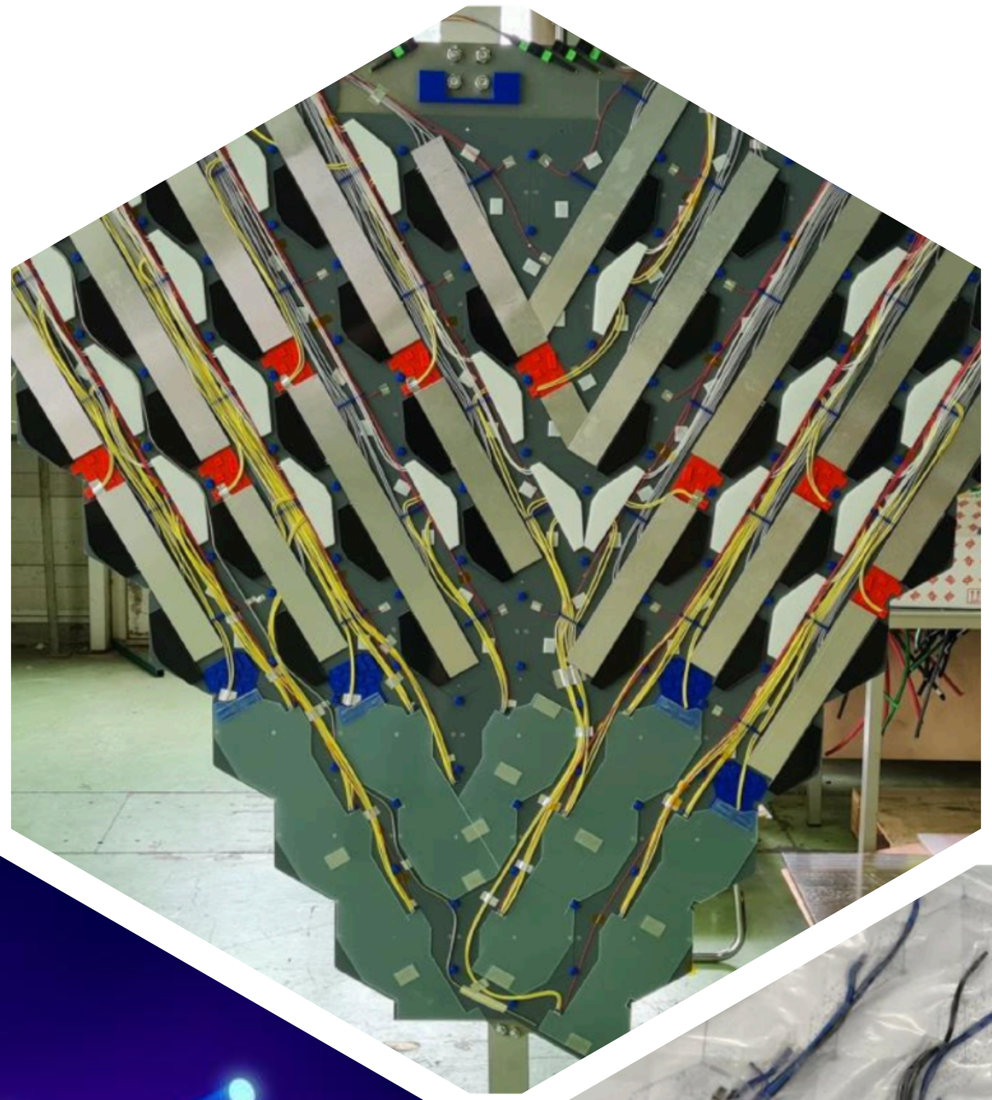


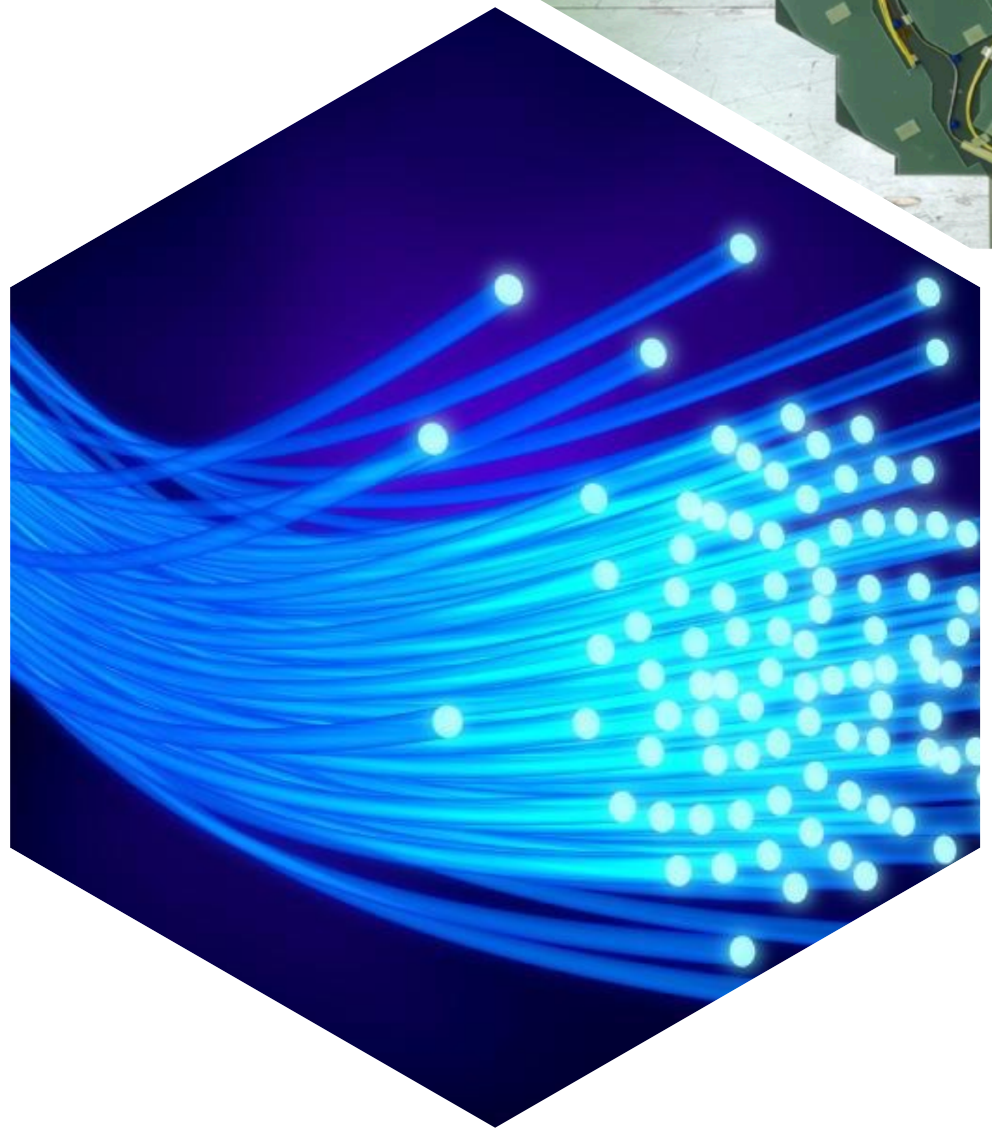


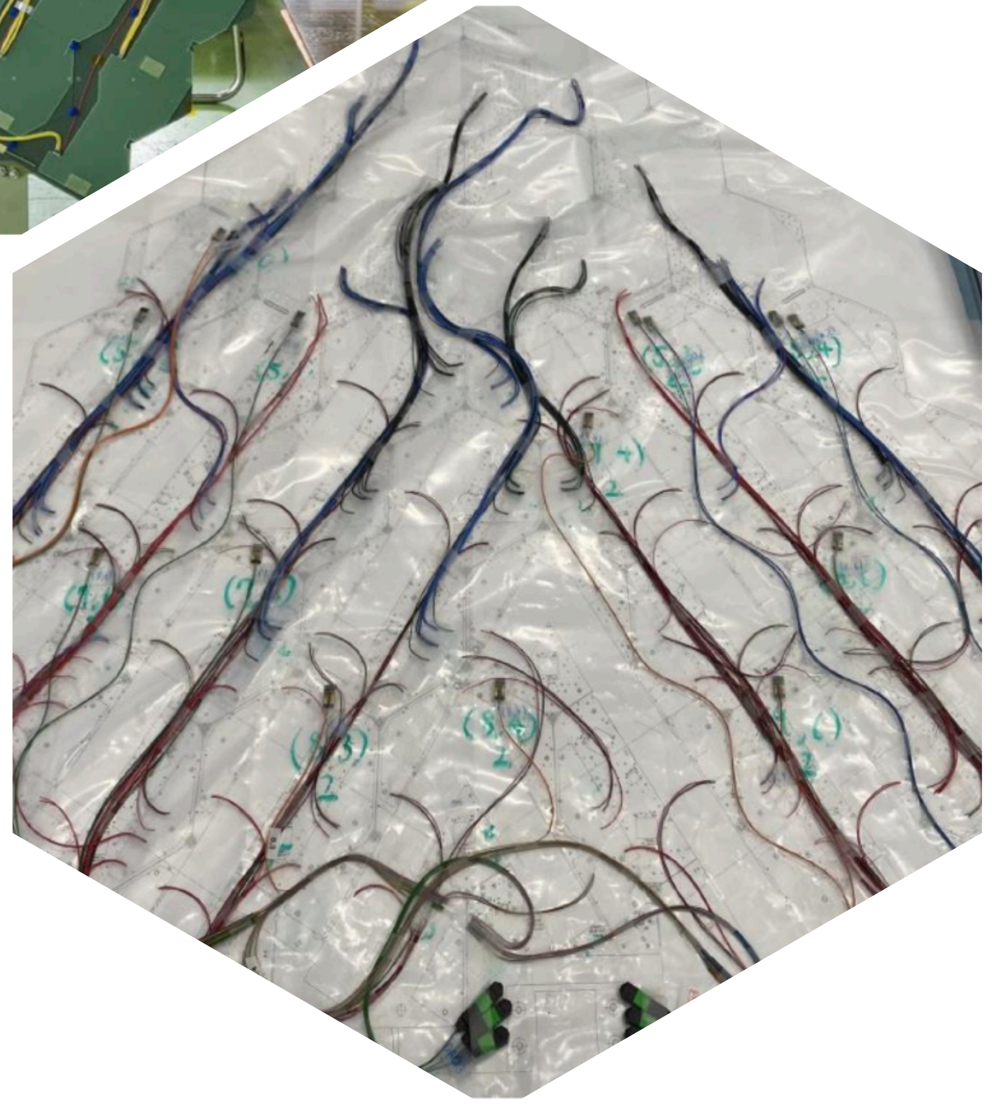


Fast data links

- Optical transmission
- Electrical transmission
- Wireless transmission







Comparisons between optical and electrical data transmission

I. Bandwidth

✔ Fiber optic cables have a much higher bandwidth than copper wires (up to 100 Gbps or more) while copper wires are limited to a bandwidth of around 30 Gbps

II. Distance

✔ Fiber optic cables can transmit data over longer distances without signal loss or degradation while copper wires are more limited in their distance capabilities, and they typically start to lose signal quality over distances of a few kilometers.

III. Immunity to interference

✔ Fiber optic cables are immune to electromagnetic interference, which can be a significant problem for copper wires

IV. Mechanism

✔ Fiber optic cables have lower size, mass, and power consumption

V. Durability

✗ Copper wires are more durable than fiber optic cables, and they are less likely to be damaged by physical stress or environmental factors. However, fiber optic cables are becoming more durable as technology improves

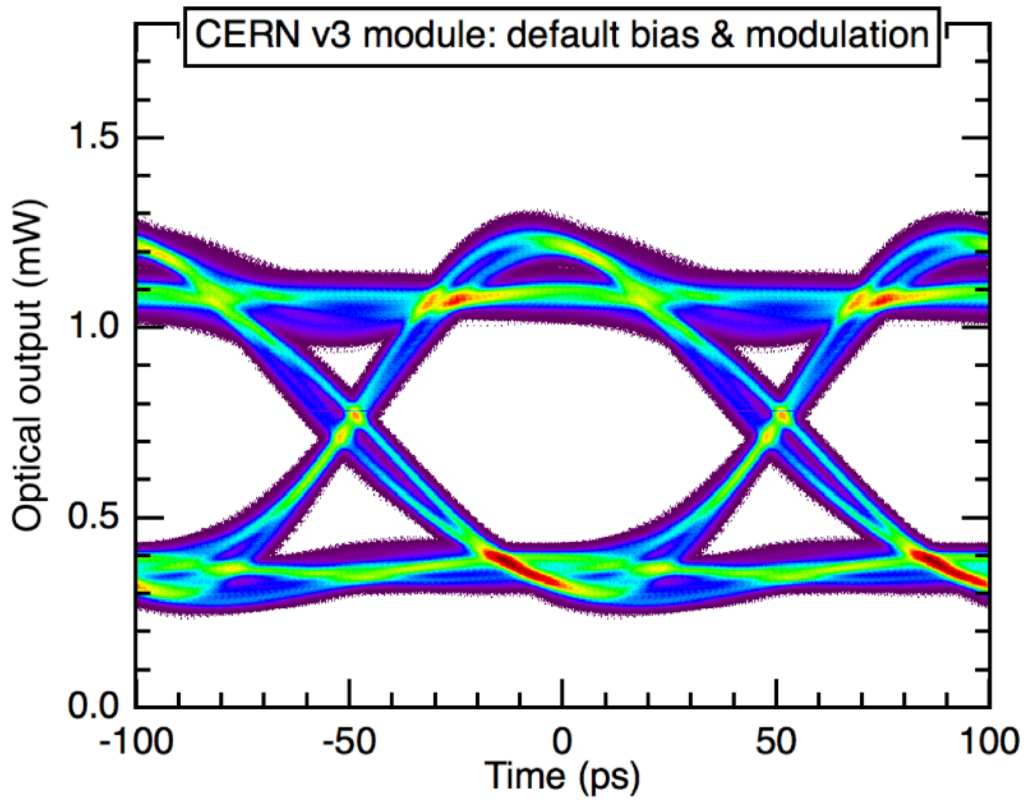
VI. Cost

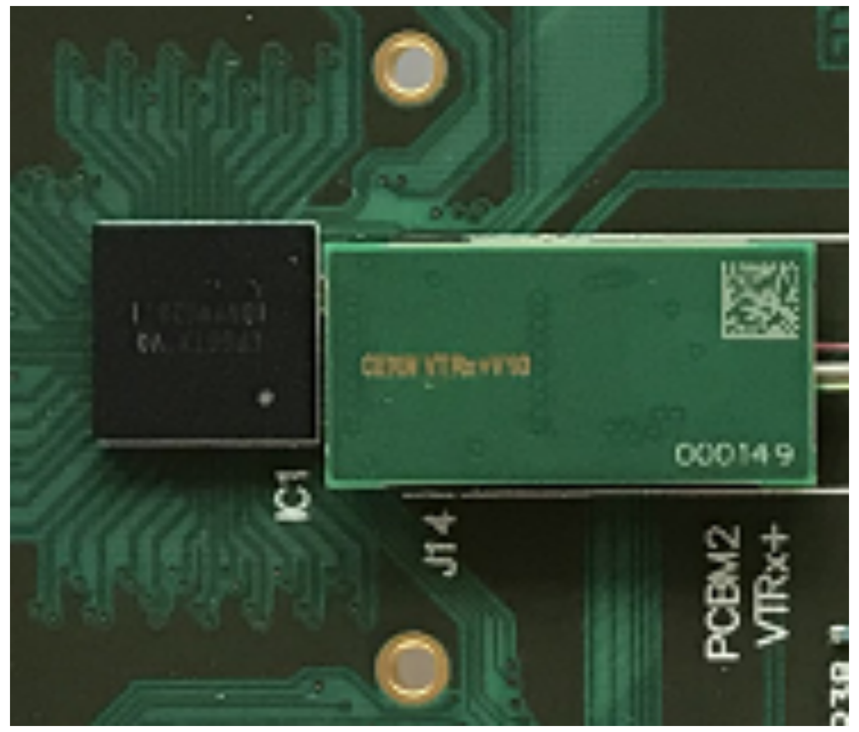
✗ Fiber optic cables are generally more expensive than copper wires, which can be a significant factor in some applications

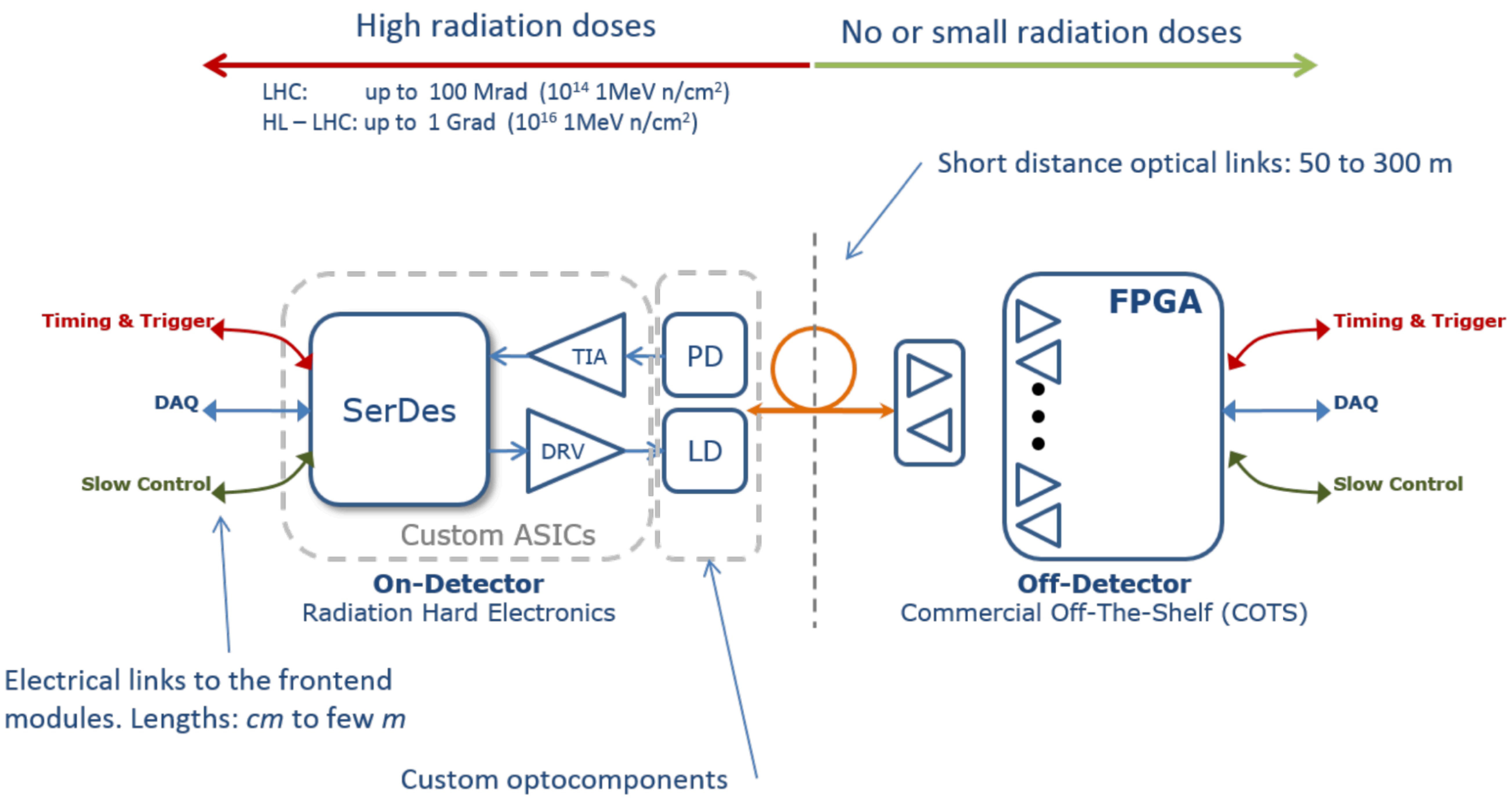
Fast optical links in CMS

- Low Power Giga Bit Transceiver (lpGBT)
  (i) Providing higher bandwidth and adequate radiation tolerance

- Versatile Link PLUS (VL+) optoelectronic components
  (i) replaces the original Versatile Link, matching the design goals of the lpGBT







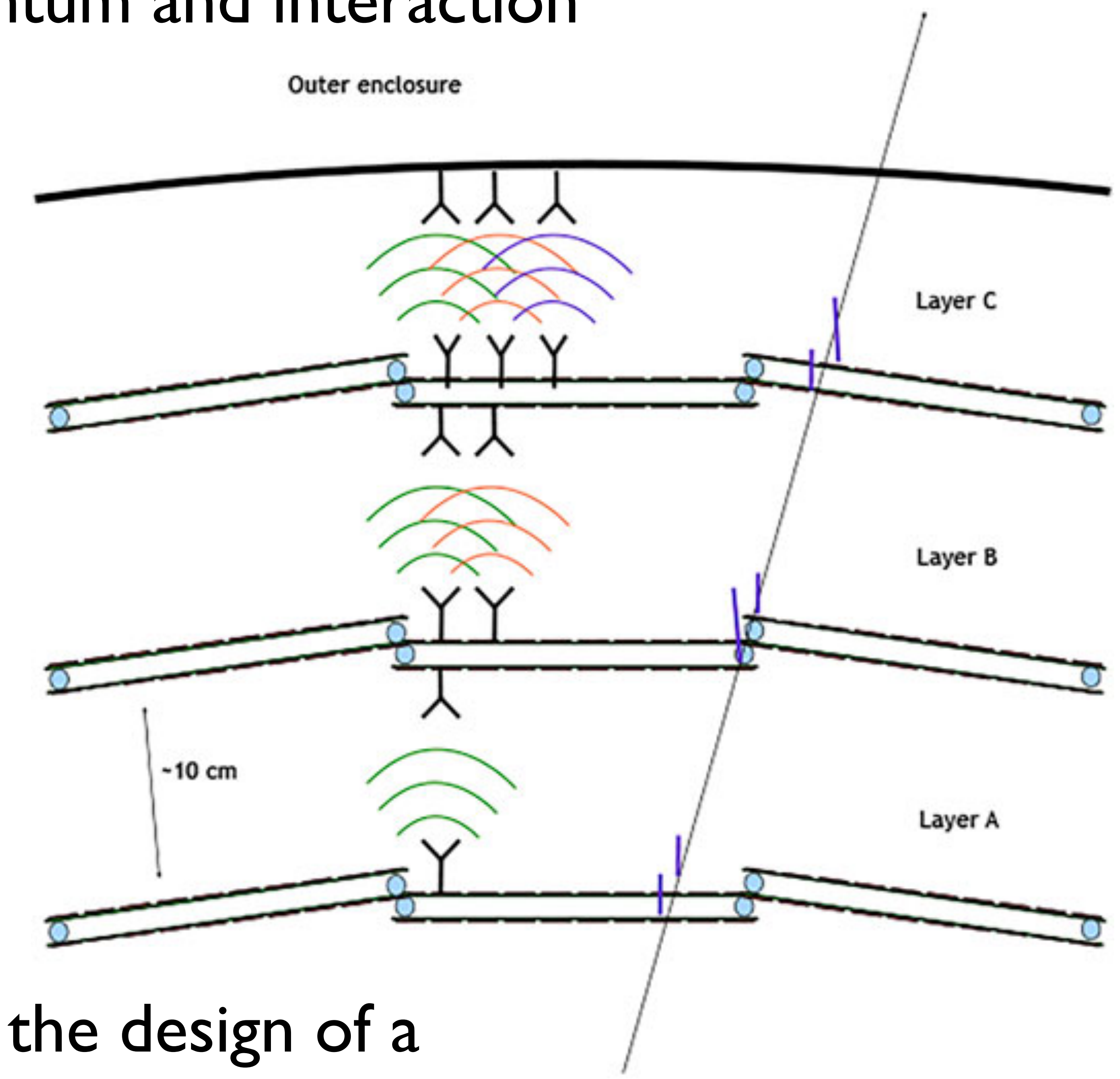
Era	Component	Line speed	Capacity
Phase-1	GBT	4.8 Gb/s	DAQ: 2.56 Gb/s TTC: 640 Mb/s Slow-control: 160 Mb/s
Phase-2	lpGBT	Down: 2.56 Gb/s	User: 1.28 Gb/s ASIC control: 80 Mb/s External control: 80 Mb/s
		Up: 5.12/10.24 Gb/s	DAQ: 3.84(4.48) Gb/s with FEC12 (FEC5) ASIC control: 80 Mb/s External control: 80 Mb/s
			DAQ: 7.68(8.96) Gb/s with FEC12 (FEC5) ASIC control: 80 Mb/s External control: 80 Mb/s

Era	Component	Max. link speed	Details
Phase-1	Versatile Link (VL)	5 Gb/s	Single-mode or multi-mode Package: Tx+Rx or 2Tx
Phase-2	Versatile Link Plus (VL+)	Rx: 2.5 Gb/s Tx: 10 Gb/s	Multi-mode (850 nm) Package: 4Tx+1Rx

Fast Wireless links

- Advantages

(i) Not constrained to the mechanical modularity of the tracker. The technology offers already today data rate comparable with present optical links and no connectors are needed
(ii) Minimizing the amount of material in the region of the tracking detectors will reduce multiple scattering and nuclear interactions that degrade the precision on the measurement of track momentum and interaction vertices, and in addition will reduce the number of fake hits arising from secondaries



- Consideration

(i) Data rates and frequency band (up to 10 Gbps at 60 GHz)
(ii) Transmission and reflection
(iii) Reduction of crosstalk
(iv) Link density
(v) Noise pickup
(vi) Radiation hardness

- Future Developments

(i) The integration of wireless technologies in tracking detector would certainly require the design of a dedicated radiation hardened mmw transceiver IP, to be integrated either as a companion chip or within silicon trackers

Advice for trigger designing

- Keep it as simple as possible

(i) Easy to commission, debug and understand

- Be as inclusive as possible

(i) One trigger for several similar analyses
(ii) Your trigger should be able to discover the unexpected as well as the signal you intended it for

- Make sure your trigger is robust

(i) Be prepared for any strange condition as triggers run tens of millions of times a second
(ii) Make sure your trigger is immune to detector problems as detectors don't work perfectly ever
(iii) Be prepared for different beam conditions

- Build in redundancy

(i) Make sure your signal can be selected by more than one trigger
(ii) Helps to understand biases and measure efficiencies
(iii) Also for safety, if rates are too high or there is some problem you still get your events

Advice for trigger designing

Triggers are not new but they are constantly evolving as the accelerators and detectors do

- Consider together with the design of the whole system
- Understand the requirements for your signals
- Follow the latest technology which could be helpful
- New studies for special requirements

Streaming DAQs

- Why not read out all the events?

(i) It is impossible to acquire all the data from every channel in a detector without a trigger
(ii) Even if it could be acquired it would be impossible to store
(iii) Even if it could be stored it would be an impractically large dataset to ever be able to process

- In a Streaming readout

(i) Data is read continuously from all channels
(ii) Validation checks at source reject noise and suppress empty channels
(iii) Rather than a trigger or event number the data is organized by time
(iv) The data then flows unimpeded in parallel streams to storage or a local compute resource
(v) Data flow is controlled at source
(vi) Data is processed as it is taken to reduce data volume

Advantages of streaming DAQs

- No trigger anymore

(i) Potential unintended bias due to the trigger is eliminated
(ii) We could run groups of experiments in parallel
(iii) The system is simplified – no expensive custom electronics
(iv) Readout speed is independent of detector response time

- Requires robust and accurate time stamp generation and distribution

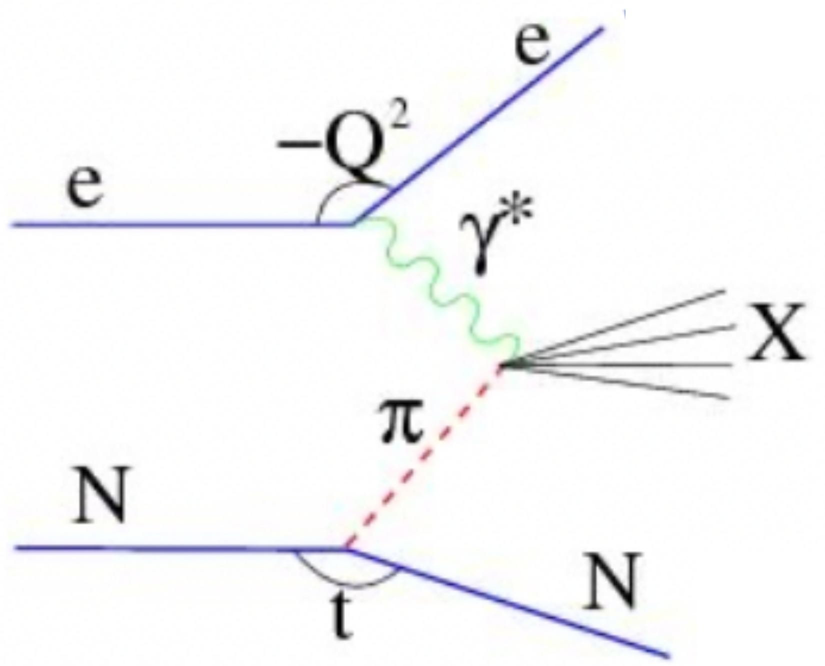
(i) Is still a simpler task than an online trigger

- Parallel timestamped streams mean

(i) System is robust against minor hardware or firmware glitches
(ii) Can use novel analysis techniques such as AI/ML looking for patterns rather than numerically processing events

- Experiments with streaming DAQs

(i) TDIS (Tagged Deep Inelastic Scattering)
(ii) SoLID (Solenoidal Large Intensity Device)
(iii) EIC (Electron-Ion Collider)



Backup

Reference

(i) Trigger and DAQ Part1 and Part2 by Andrew Rose

(ii) Tracking Triggers for the HL-LHC by Anders Ryd and Louise Skinnari

(iii) Performance of Muon Reconstruction in the CMS High Level Trigger using pp Collision data at $\sqrt{s}$ = 13.6 TeV in 2023

(iv) FIRST COLLISIONS RECONSTRUCTED WITH GPUS AT CMS by Andrea Bocci

(v) The challenge of Heterogeneous Computing: the CMS case for 2020s by Felice Pantaleo

(vi) CMS High Level Trigger performance comparison on CPUs and GPUs

(vii) ML applications in CMS by Markus Stoye

(viii) Paving the Way to Reconstruct the 5D Information of the CMS HGCAL Detector at the HL-LHC by Jingyu Zhang

(ix) CMS HGCAL by Jia-Hao Li

(x) The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger Technical Design Report

Reference

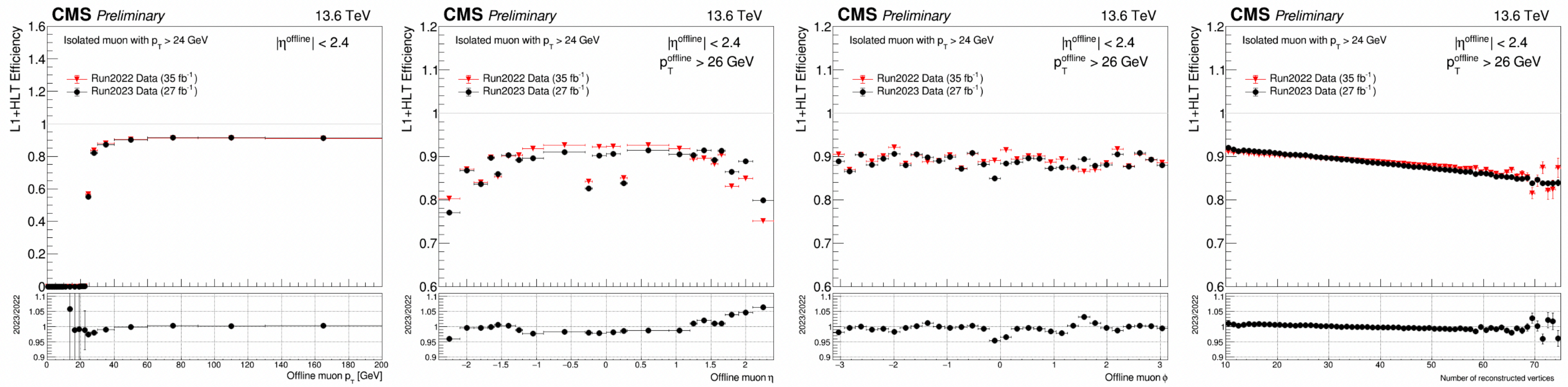
(xi) lpGBT Documentation

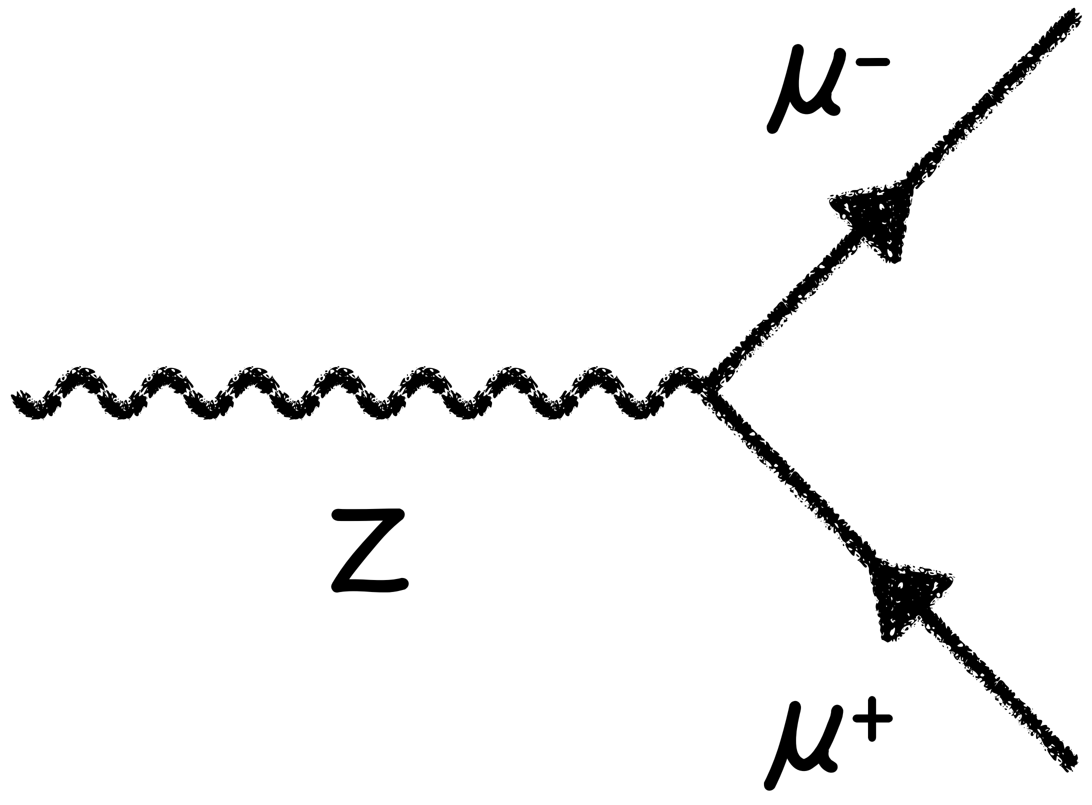
(xii) Multi Gigabit Wireless Data Transfer in Detectors at Future Colliders by R. Brenner, C. Dehos and E. Locci

(xiii) DAQ, Streaming, calibration and triggering by Graham Heyes

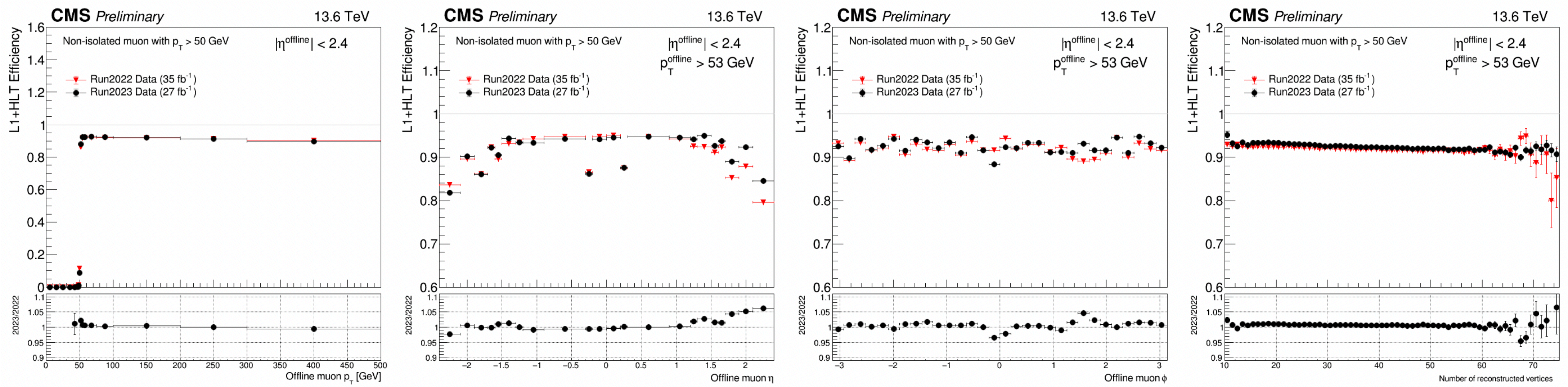
Muon Reconstruction in CMS

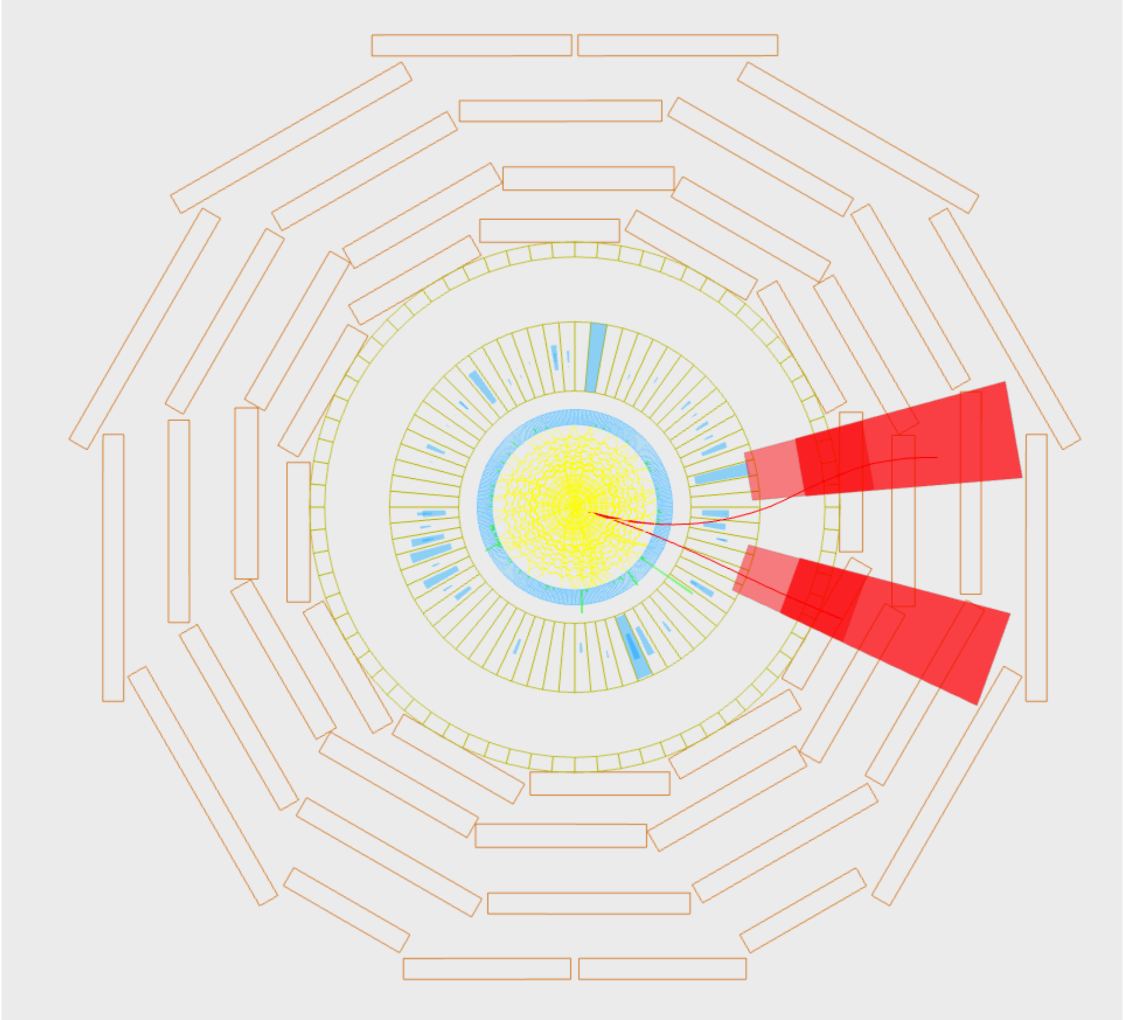
IsoMu24





Mu50





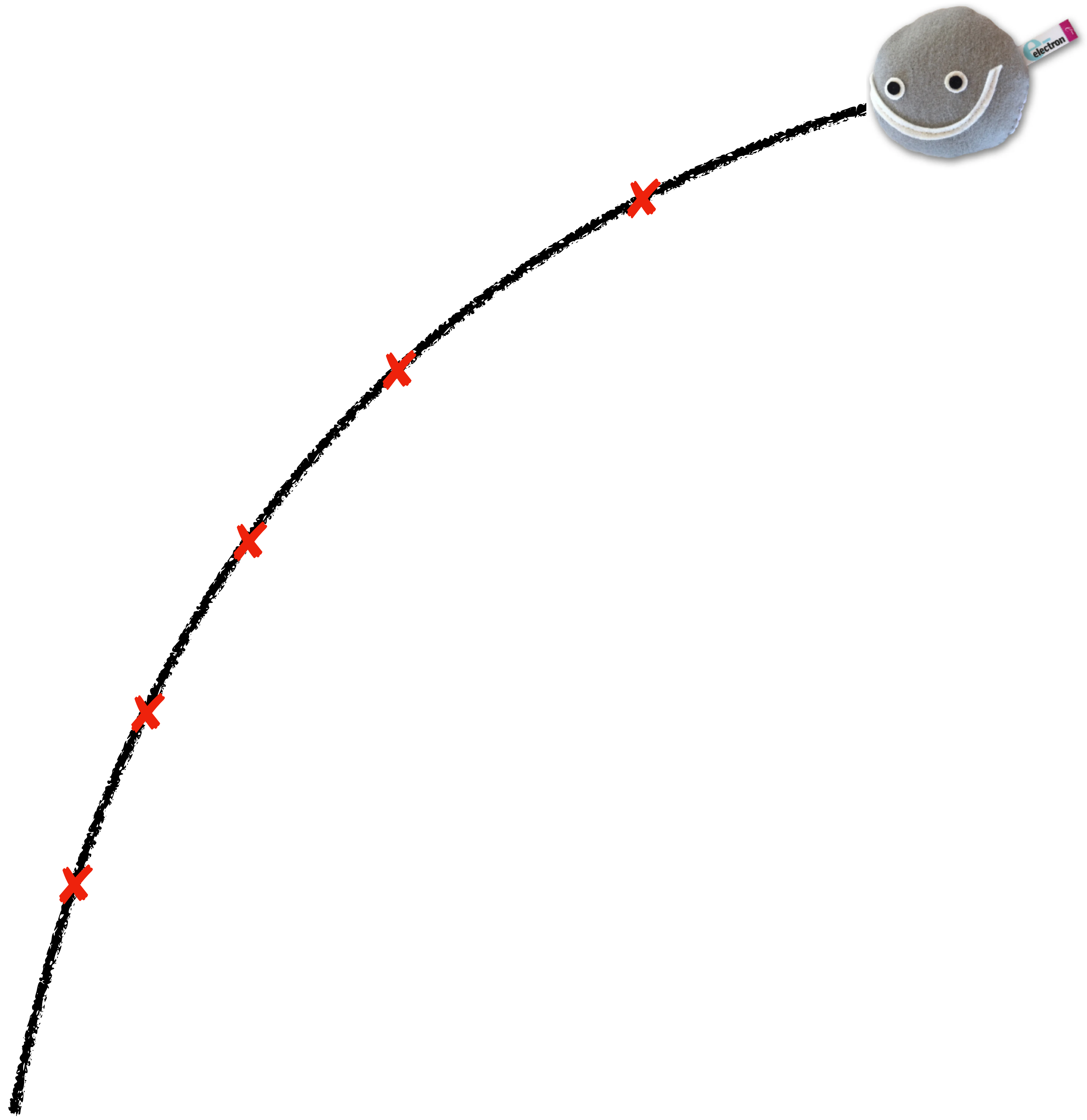
Tracking approaches

- Common point and requirement

(i) Split the detector into smaller geometrical region
(ii) Utilize time multiplexing to distribute the tasks of performing the pattern recognition, track fitting, and duplicate removal to many different hardware components
(iii) Process the roughly 15,000 stubs that arrive every 25 ns and find the tracks within about 4 μs

- Three approaches studied for CMS

(i) Tracklet Approach
(ii) Hough Transform + Kalman Filter Approach
(iii) Associative Memory + FPGA Approach



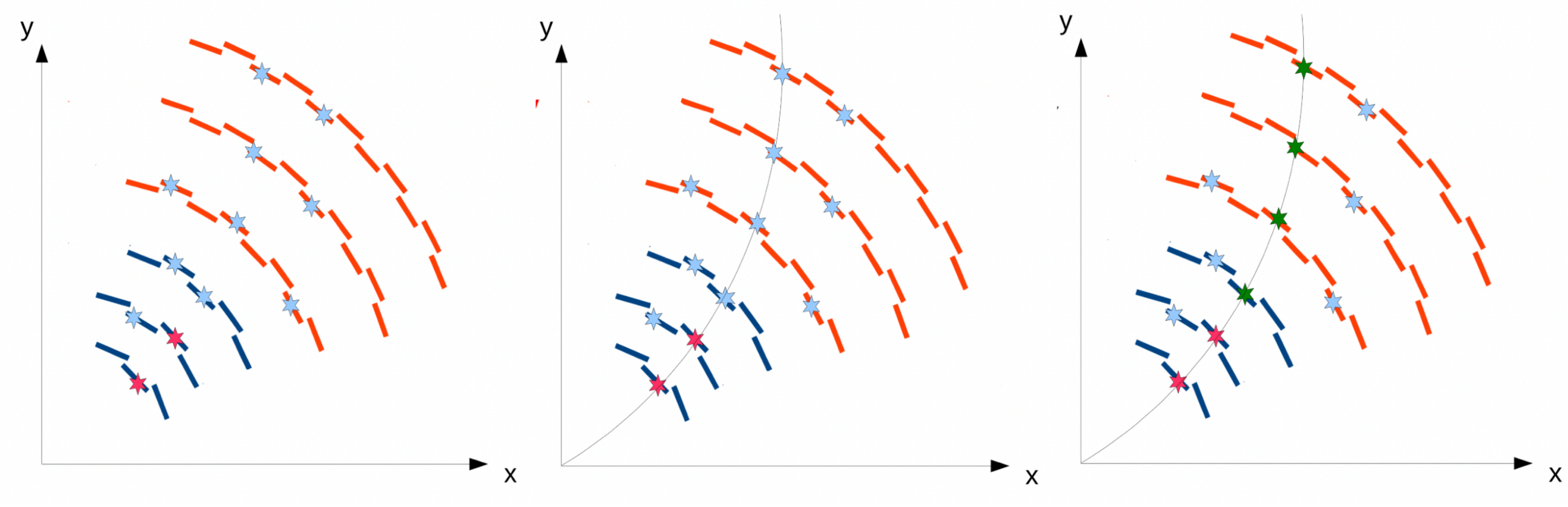
Tracklet Approach

- Tracklet road search algorithm

(i) Tracklets, or the seeds, are formed from pairs of stubs in adjacent layers or disks
(ii) For each stub pair a trajectory is calculated, using the beam spot as a constraint in the transverse plane, and projections to other layers and disks computed
(iii) Using the projections to other layers and disks, matching stubs are found and used in the final track fit

- Linearized χ^2 track fit

(i) provides a refinement to the track parameters



Hough Transform + Kalman Filter Approach

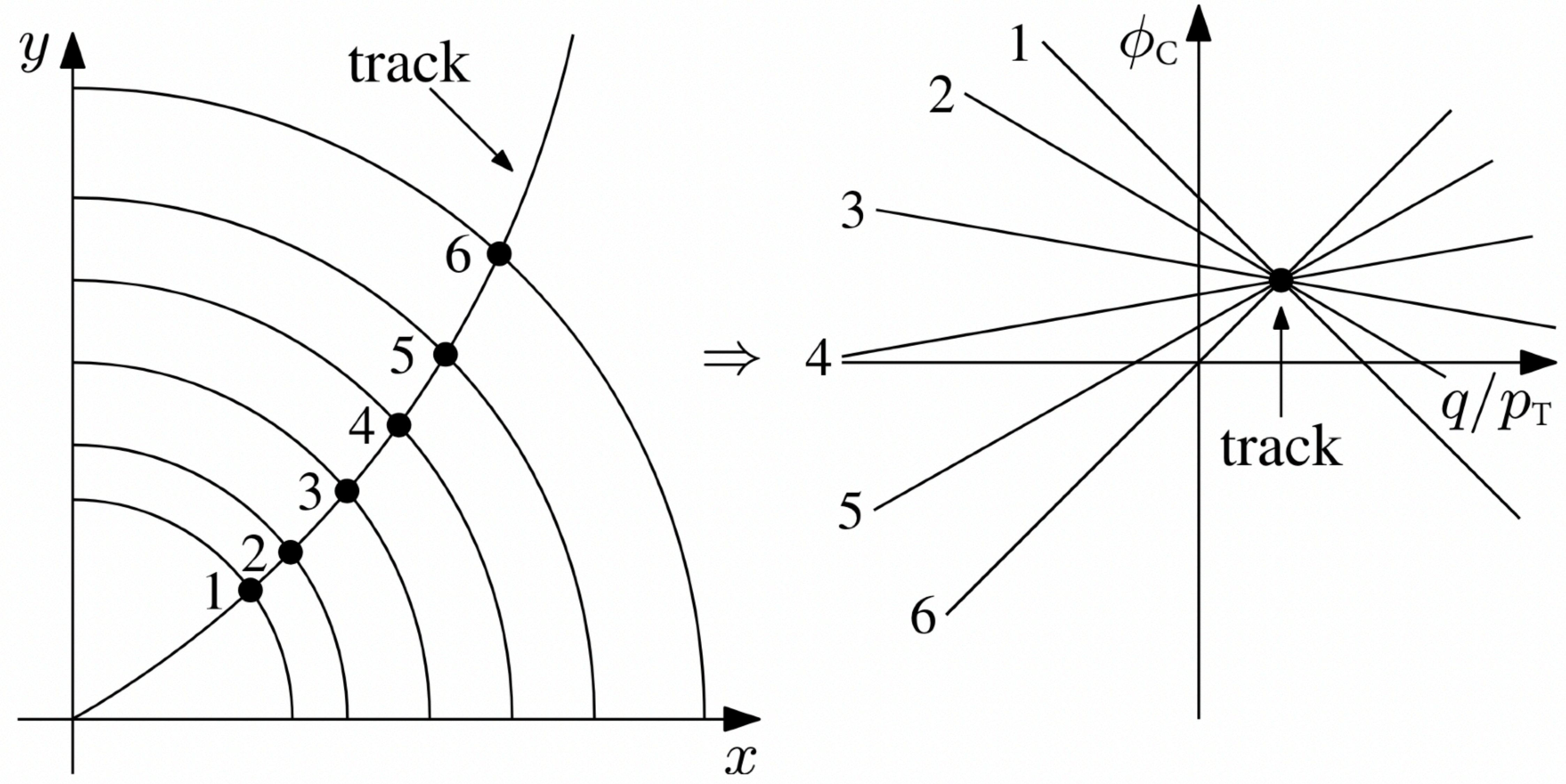
- Hough transform pattern recognition

(i) Each stub at a given r_c and ϕ_c describes a straight line in the ϕ_c vs q/p_T plane
(ii) Stubs belong to the same particle trajectory these lines will all go through the same $(q/p_T, \phi_c)$ point
(iii) Tracks can therefore be identified by looking for points in the $(q/p_T, \phi_c)$ plane where multiple lines overlap

- Kalman filter track fit

(i) The Kalman state is updated iteratively, starting from stubs in the innermost layers
(ii) The seed trajectory is projected to the next layer or disk and using the position information from the next stub

$$\phi_c = \phi - \frac{qB}{2p_T} \cdot r_c$$



Associative Memory + FPGA Approach

- Associative memory pattern recognition

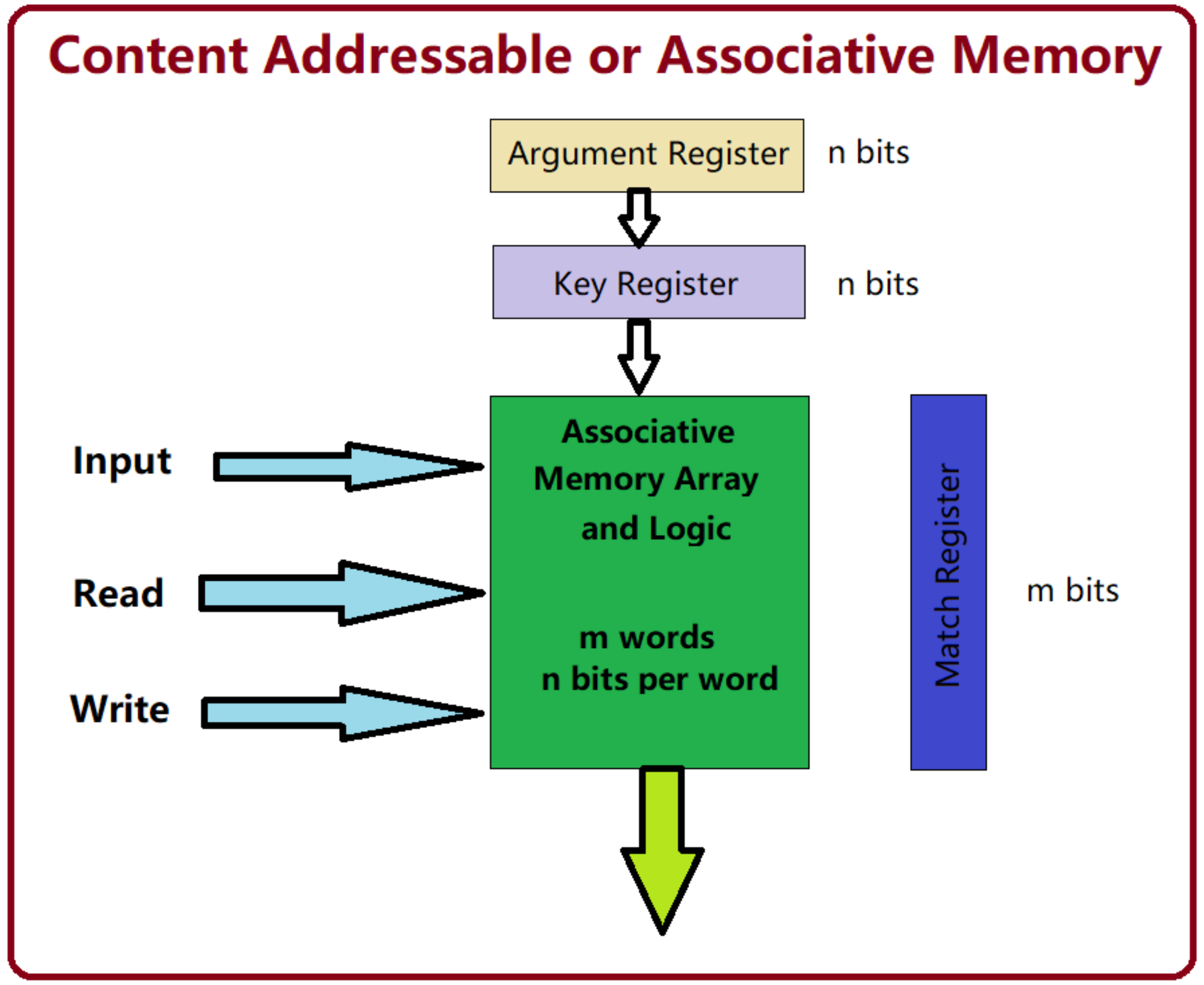
(i) Uses a content addressable memory (CAM) ASIC
(ii) AM chips are preloaded with hit patterns representing possible valid trajectories for detector's geometry
(iii) AM chips allow simultaneous matching to a large number of patterns, providing a fast response once all hits have been loaded
(iv) The patterns are coarse hit positions, super strips, on the modules

- Linearized χ^2 track fit

(i) provides a refinement to the track parameters

- Content addressable memory

(i) Also known as associative memory or associative storage and compares input search data against a table of stored data, and returns the address of matching data



Hybrid implementation

- All-FPGA based approaches are decided to reduce the risks associated with the development of the AM ASIC

- Tracklet algorithm + Kalman filter are chosen