Directions in Realtime Tracking

Everything, Everywhere, All at Once

Kristian Hahn – Northwestern

CTD Realtime Tracking Workshop, Toulouse (FR), Oct. 2023

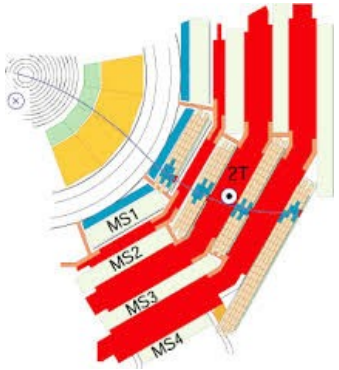
Preliminaries

For background, I've been engaged in:

- Development of L1 track trigger for the CMS HL-LHC upgrade
- Development of CMS Outer Tracker upgrade backend system
- A bias might be reflected (LHC, CMS, hardware) ...

Will focus on realtime tracking with inner detectors in mid-term HEP

- Real-time traditionally meant low-latency, ie: at the level of h/w trigger
  - With the advent of triggerless readout, real-time tracking applies more broadly
  - Now : "tracking that is able to keep up with the frontend event rate"
- Will not touch on realtime tracking with muon systems ...
- And will not discuss the FCC era directly ...
  - Although there will probably be some feed through



Some History

Earlier HEP experiments (eg : LEP, Tevatron) featured 2D tracking in their hardware triggers, Eg:

- CDF had full 3D tracking at Level-2 (XFT)
- ~20 us latency, ~40 kHz / ~180 Gbps input
- Obviously dealing with much higher data rates now ...

https://doi.org/10.1016/j.nima.2006.10.204

Realtime tracking has since evolved along two primary lines :

- More computation & full 3D tracking in Level-1 hardware
  - Examples : Belle-II, CMS high luminosity upgrade
- “Triggerless” readout : full rate readout to software trigger
  - And the application of heterogeneous computing in the trigger
  - Example : LHCb

This evolution will be apparent in all the talks today!

Along with a hybrid approach: L1 w/o tracker, accelerated s/w tracking (ATLAS)

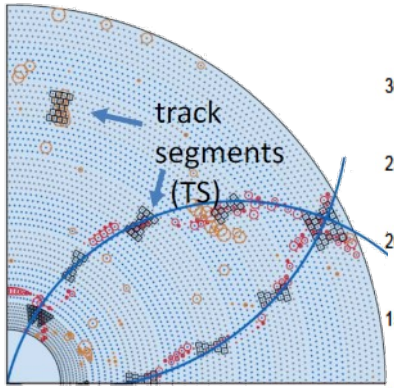
- An associative-memory Level-1(“0”) track trigger (ala CDF) was originally foreseen : FTK

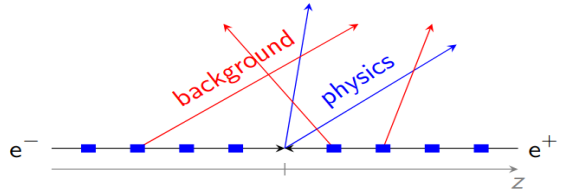
ATLAS-TDR-029-ADD-1

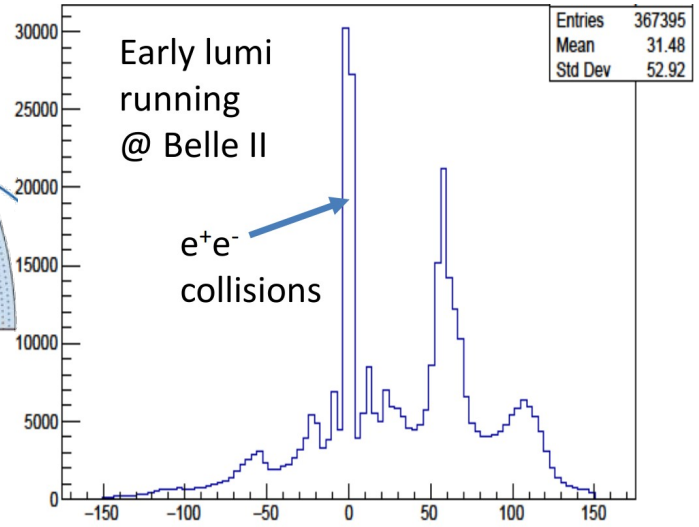
State of the Art

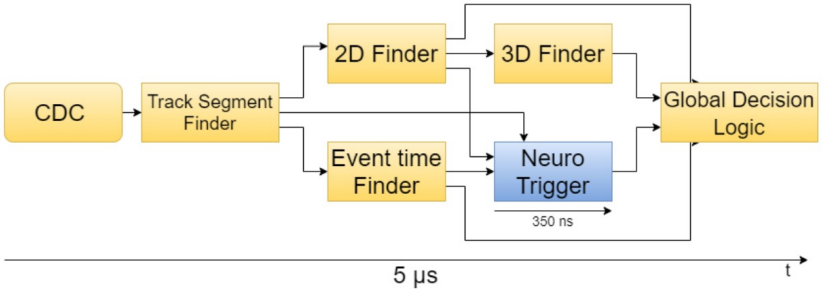
Full 3D hardware tracking at Level-1 : Belle-II

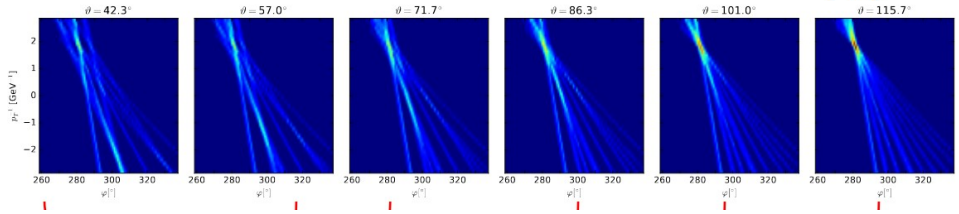
- Reject beam backgrounds via tracking with the Central Drift Chamber (14k sense wires)
  - Selecting on z position
- 1.7 Tbps rate out of the CDC
  - Data reduction via segment finding
- Average of 11 tracks / event
- FPGA-based system, 5 us latency budget
  - 2D/3D Hough Transforms, and Neural Net in firmware
- Hardware tracking operational since 2019, N.N. from 2021









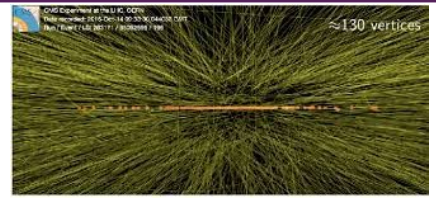


[Sebastian Skambraks et al 2020 J. Phys.: Conf. Ser. 1525 012102](#)

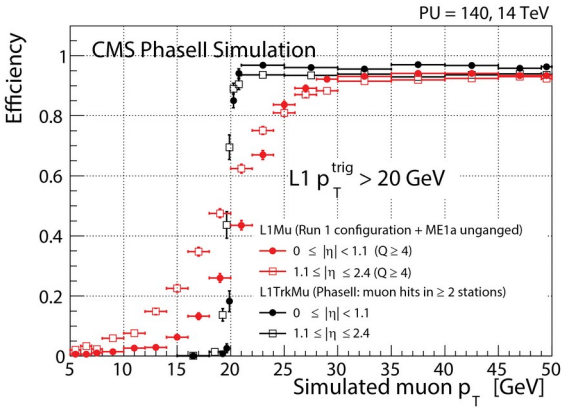
[Sebastian Skambraks et al 2018 J. Phys.: Conf. Ser. 1085 042026](#)

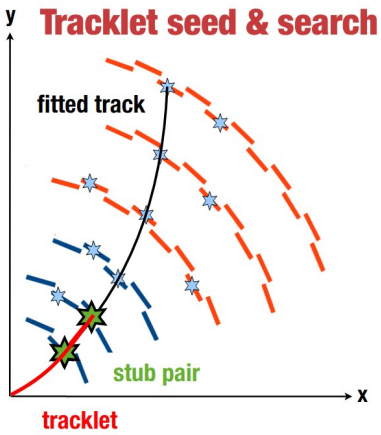
State of the Art

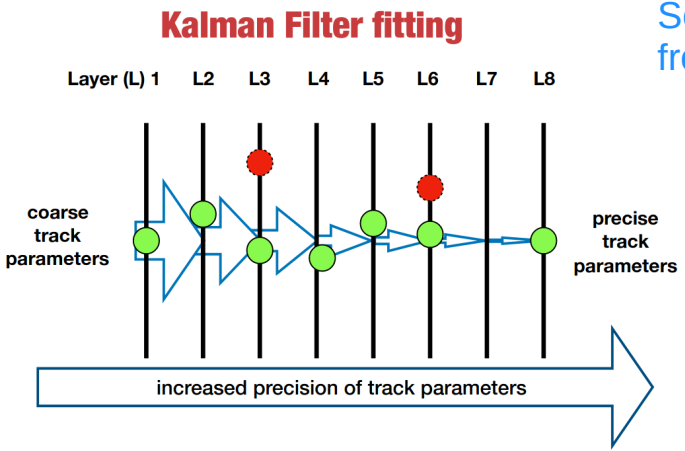
Full 3D hardware tracking at Level-1 : CMS upgrade



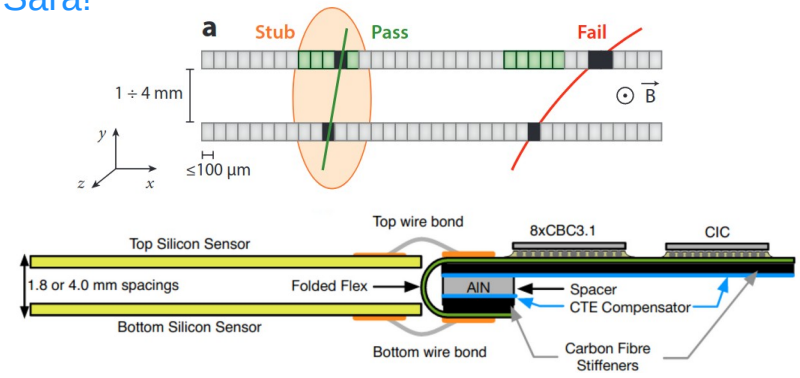
- Reject uninteresting “pileup” interactions
  - Avg. 200 per beam crossing (25 ns) @ HL-LHC
  - Improving object resolution, trigger efficiency, lowering rates
- Stubs : correlated signals in two closely separated silicon sensors. Send only stub data consistent with pT > 2 GeV
  - Reduces data by x10, but still 15k stubs/BX, O(50 Tbps)
  - Stubs passed to FPGA-based system, 4us latency budget
  - Road-seach (“tracklet”) track finding, Kalman filter







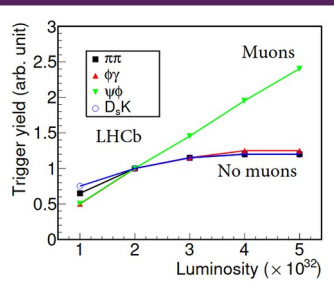
See talk from Sara!

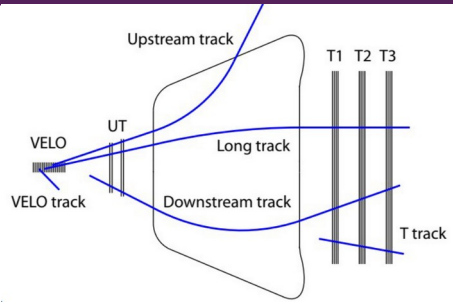


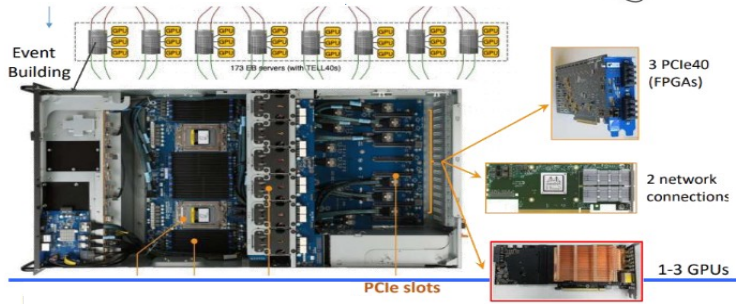
State of the Art

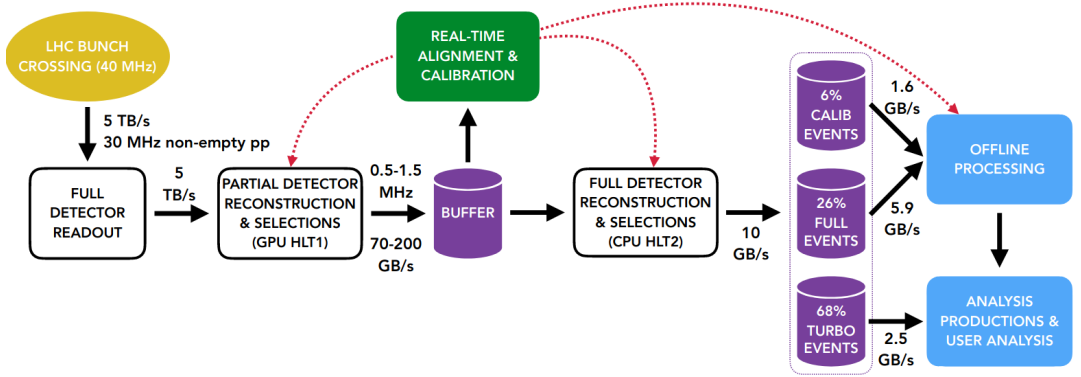
Triggerless software tracking : LHCb

- Level-0 hardware trigger removed for Run-3
  - Limits trigger yield for hadronic signals
  - 30 MHz rate (32 Tbps) into the HLT!
- Full tracking (incl. with pixels) performed on GPU
  - Co-located with FPGA/PCIe event building
  - Simple track selections at this stage (HLT1)
  - Full offline-quality reconstruction in CPU (HLT2)
- Object filtering applied for ~2/3 of the output data stream
  - Significantly reducing output rate to tape
  - Re-reconstruction not possible for these events









"Allen: A High-Level Trigger on GPUs for LHCb",
Computing and Software for Big Science (2020) 4:7

We describe a fully GPU-based implementation of the first level trigger for the upgrade of the LHCb detector, due to start data taking in 2021. We demonstrate that our implementation, named Allen, can process the 40 Tbit/s data rate of the upgraded LHCb detector and perform a wide variety of pattern recognition tasks. These include finding the trajectories of charged particles, finding proton–proton collision points, identifying particles as hadrons or muons, and finding the displaced decay vertices of long-lived particles. We further demonstrate that Allen can be implemented in around 500 scientific or consumer GPU cards, that it is not I/O bound, and can be operated at the full LHC collision rate of 30 MHz. Allen is the first complete high-throughput GPU trigger proposed for a HEP experiment.

Which Future Will Be Realized?

Outline

- **Everything** : Displaced/non-standard tracking, 4D tracking, extended track features
- **Everywhere** : expanding coverage, higher granularity
  - Also, wider application of RT tracking in HEP ...
- **All at once** : pixel in hardware tracking, broader movement toward the triggerless model, more intelligence in the front ends

Everything : Non-standard Tracking

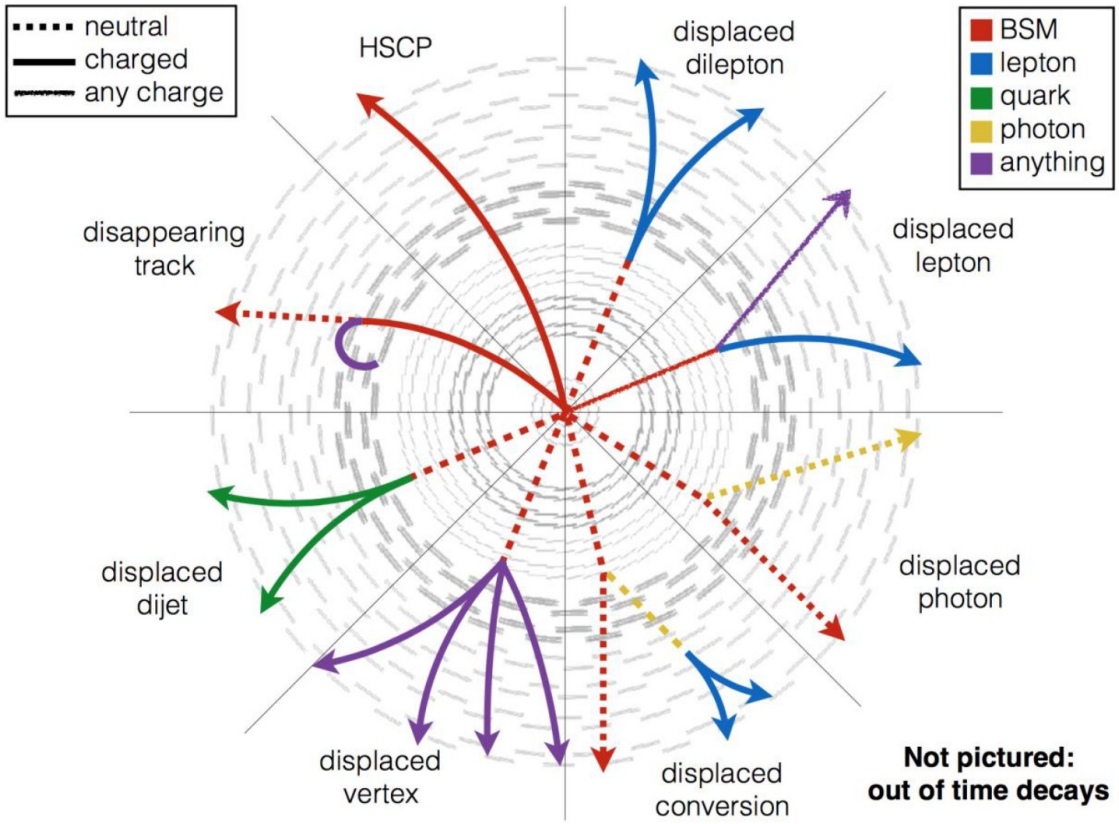
Increasing interest in searches for Long-Lived BSM particles

- Small couplings
- New symmetries
- Scale/phase-space suppression

Often requiring non-standard reconstruction algorithms

Not just for colliders

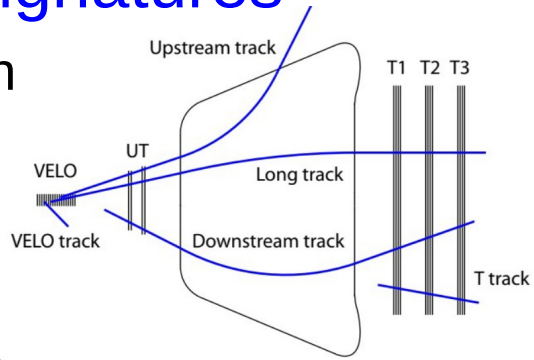
- Beam dump experiments : NA62, NA64, ShiP, LDMX, SeaQuest …
- Far detectors : FASER, Mathusla



Everything : Non-standard Tracking

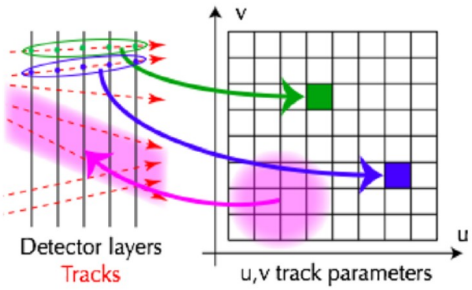
The triggerless model is well suited to exotic signatures

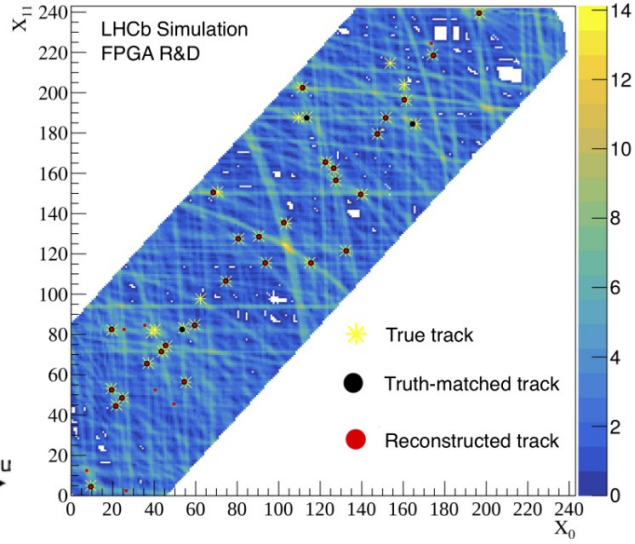
- HLT benefits from full detector resolution and information
- Variable latency → more time for complex tracking!



Yet computational load is high …

- LHCb HLT1 presently reconstructs “long tracks” only
  - NB: specspecialized Ks0 trigger, displaced VELO vertices
- Run-4 R&D : dedicated FPGA-based Downstream Tracker to provide SciFi(T)-tracks to HLT1
  - Artificial Retina : detector hits mapped to a defined collection of track parameters, weight & interpolate
  - < 1us latency demonstrated



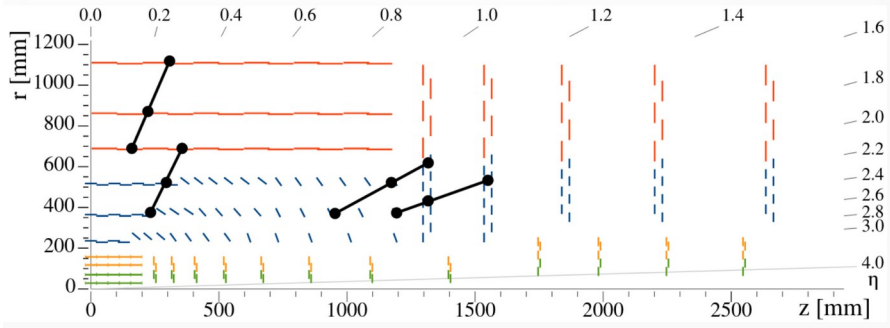


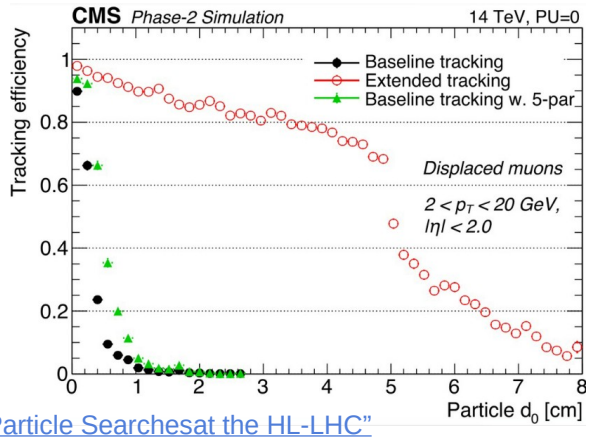
Riccardo Cenci et al (2020) J. Phys.: Conf. Ser. 1525 012101

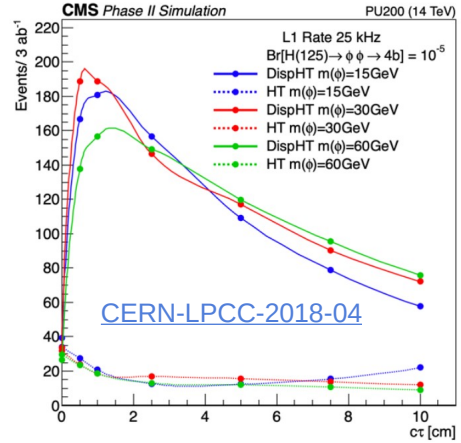
Everything : Non-standard Tracking

Extended Level-1 algorithms for displaced tracking, eg CMS:

- R&D progressing for an addition to the baseline prompt f/w algorithm
- Requires :
  - Triplet seeding to reduce combinatorics w/o the beamspot constraint
  - 5 parameter (vs 4) Kalman filter fit
- Decent efficiency up to $|d_0| < \sim 5$cm
- Track rate increases by ~40%
- Recovery of some BSM scearios that would otherwise be discarded
  - Can also help with electrons ...







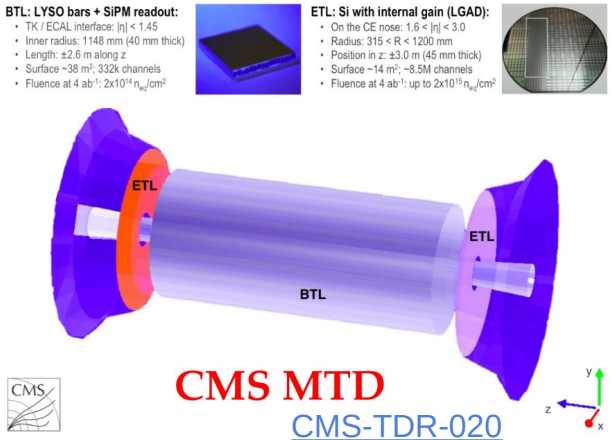
See also:

Y. Gershtein, “CMS Hardware Track Trigger: New Opportunities for Long-Lived Particle Searchesat the HL-LHC”

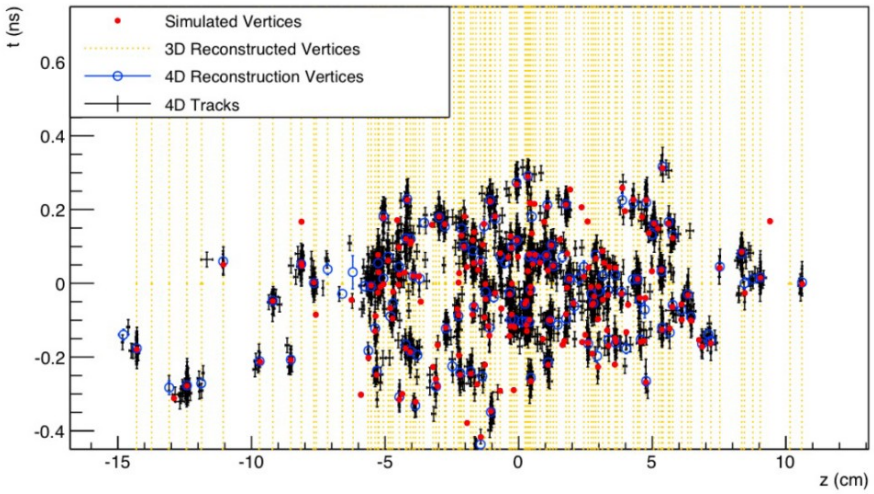
Everything : 4D Tracking

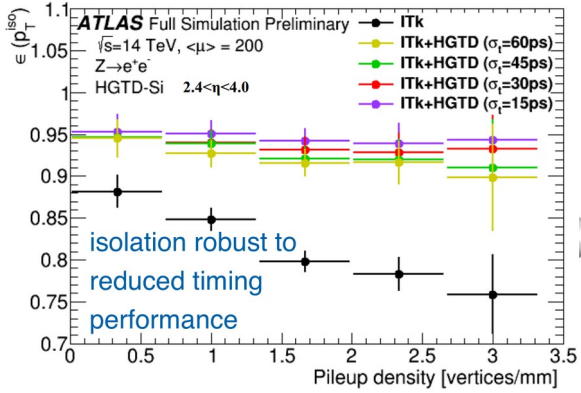
Combining position information with timing

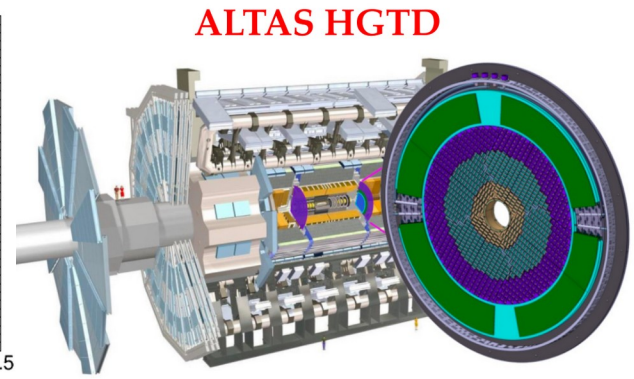
- ATLAS & CMS upgrades to feature timing detectors
- Timing information will be available to the s/w HLT
- 30 ps timing resolution in forward regions with LGADs
- Primarily intended for pile-up rejection
  - Improving lepton isolation, b-tagging, MET



CMS-TDR-020







ATLAS-TDR-031

Everything : 4D Tracking

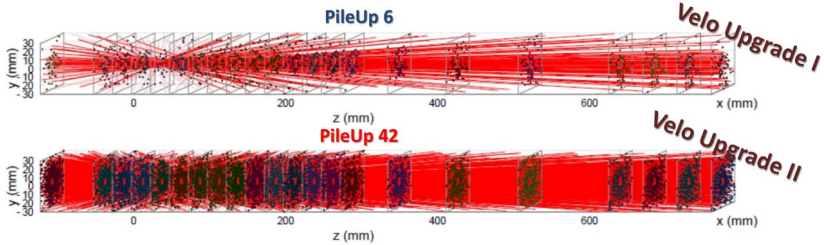
But single timing layers are not yet 4D tracking ...

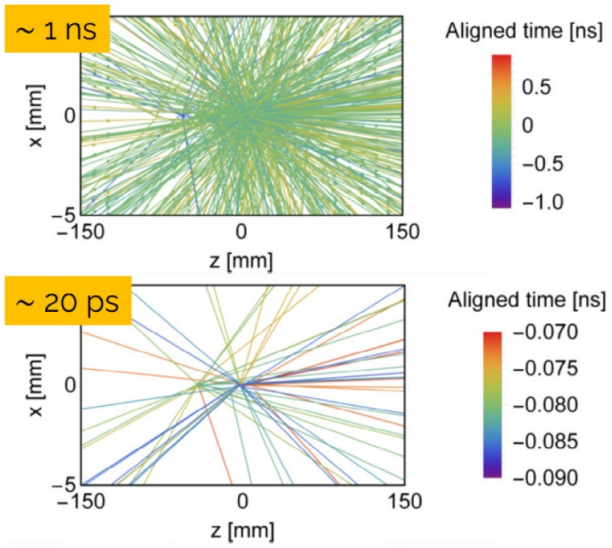
T. Parjero, “VELO Upgrade II : The LHCb 4D pixel detector”

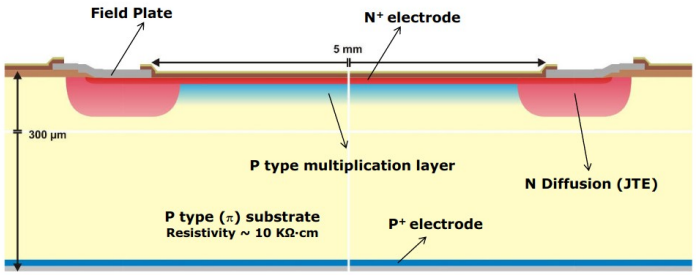
- Push to combine position and timing measurements in tracking detectors

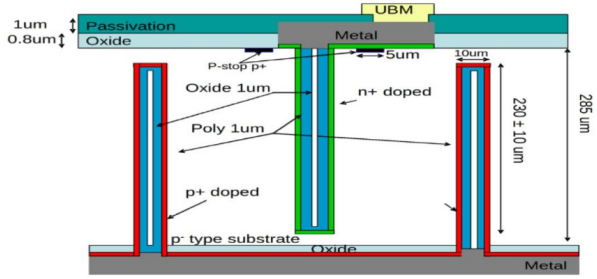
Example : LHCb VELO Upgrade 2

- 7.5-fold increase in luminosity in Run-5 (2030’s)
- Track reconstruction becomes very challenging
- Pileup vertices can be resolved with addition of timing information at the level of ~20 ps track / <50 ps hit
- Sensor (eg: LGAD, 3D pixel) and ASIC (28 nm) R&D is in progess









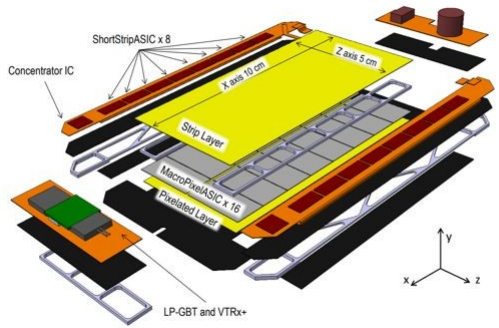
Outline

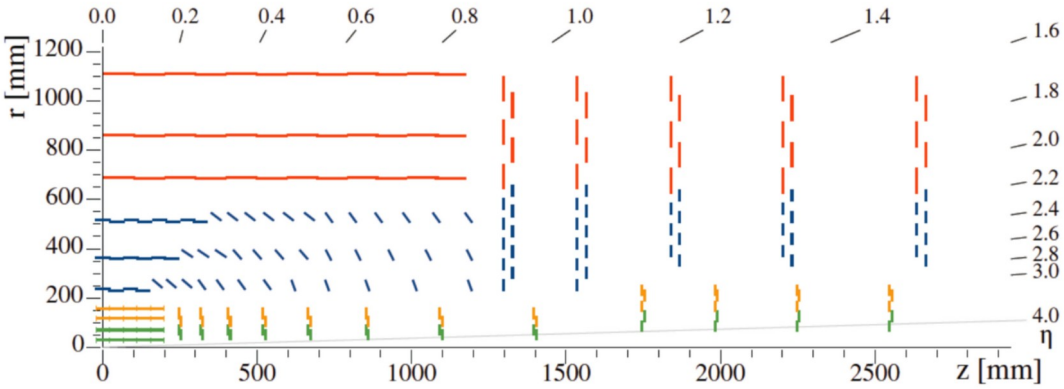
- **Everything** : Displaced/non-standard tracking, 4D tracking, extended track features
- **Everywhere** : expanding coverage, higher granularity
  - Also, wider application of RT tracking in HEP ...
- **All at once** : pixel in hardware tracking, broader movement toward the triggerless model, more intelligence in the front ends

Everywhere : Instrumentation++

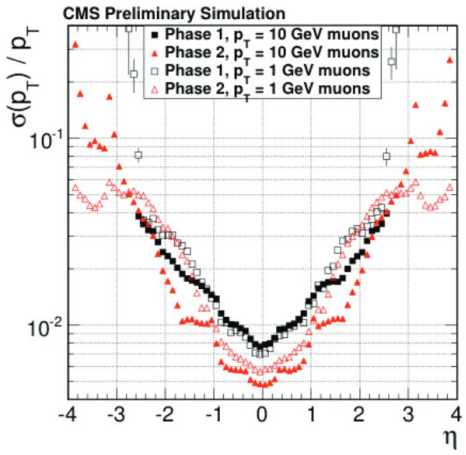
Experiments greatly expanding Si coverage, eg: CMS HL-LHC

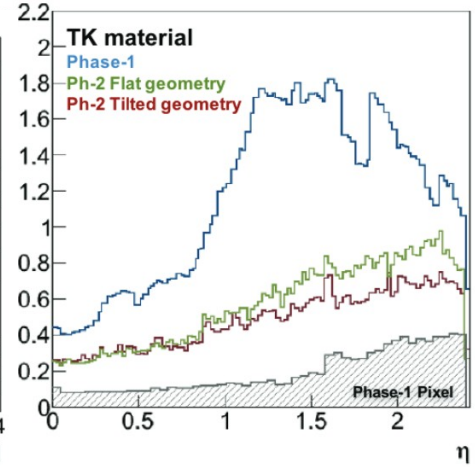
- Higher granularity : Outer Tracker 4.2 strips + 170M macro-pixel (was 9.3 M strips), Inner Tracker 2000M pixels (was 66M pixels)
  - Strip pitch 90-100 µm, length 5-2.5 cm (vs 80-200 µm, 10-20 cm)
  - MacroPixels 100 µm x 1.5 mm
  - Pixels 25 x 100 µm² (vs 100 x 150 µm²), 6 times smaller!
  - 3D pixels in layer 1
- Tracking to |n| < ~4 (was < 2.4), L1 rate to 100 kHz → 750 kHz
- And yet smaller material budget : better track parameter resolutions!





CERN-LHCC-2017-009 ; CMS-TDR-014

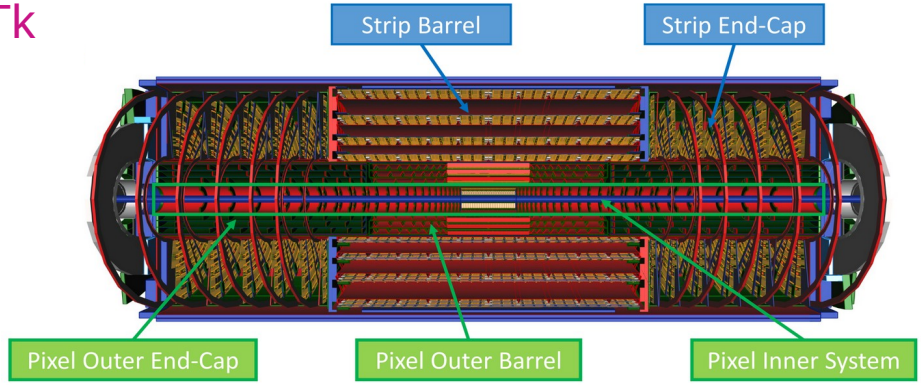


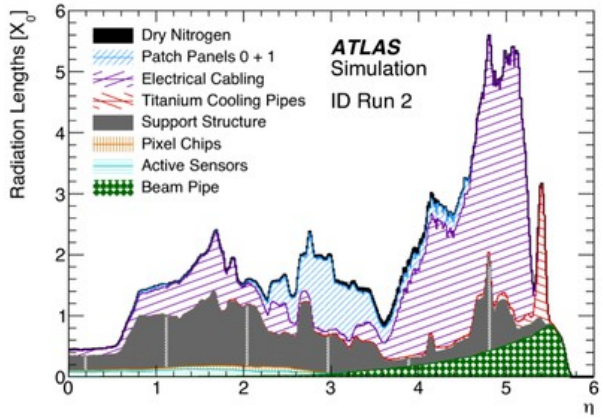


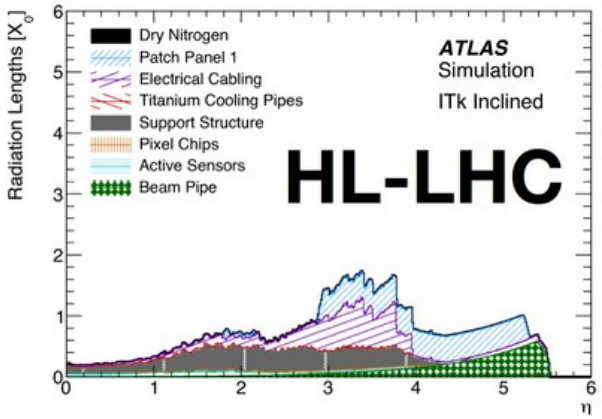
Everywhere : Instrumentation++

Experiments greatly expanding Si coverage, eg: ATLAS HL-LHC

- Moving toward an all-silicon tracking system, iTk
  - 4 layers of strips
    - 60M channels, factor of 10 increase
    - Smaller pitch, 70-80 μm
  - 5 layers of pixels
    - 5000M channels, factor of 60 increase
    - Smaller pixels : 25x100 μm^2 , 50x50 μm^2
    - 3D pixels in Layer 0
- Tracking to |n| < ~4 (was < 2.4)
- Trigger rate 100 kHz → 1 MHz
  - No Level-1 tracking
- Again, smaller material budget!







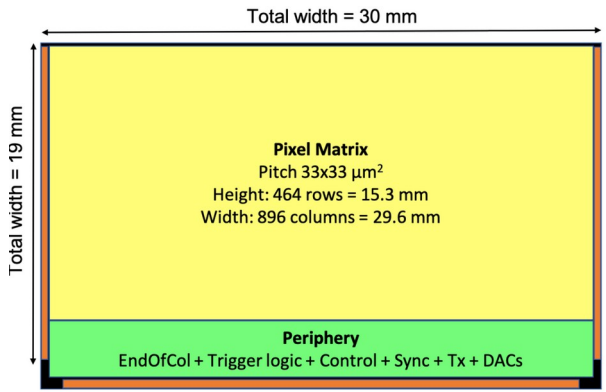
ATL-PHYS-PUB-2021-024

Everywhere : Instrumentation++

Belle-II VTX upgrade (‘26?), fully pixelated 5-layer vertex detector

- Super KEKB upgrade in 2027 : $2e^{35}$ cm^{-2} s^{-1} → $6e^{35}$ cm^{-2} s^{-1}
- OBELIX ASIC: DMAPS technology, thin sensors
- 33x33 μm^3 pixels, 15 μm resolution
- >800M channels? (vs. 7.7M pixel + 245k strips)

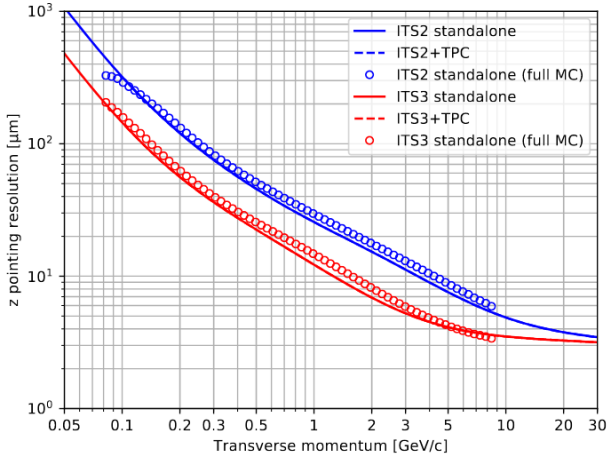
https://doi.org/10.1016/j.nima.2022.167616



ALICE ITS3 upgrade for LHC Run-4

- Replacing inner 3 layers of ITS2 pixel detector
- Ultra-thin, fully cylindrical, ~20x20 μm^2 pixels
- Very lightweight: 0.35% X_0 → 0.05% X_0
- Inner radius 2.3 → 1.8 cm
- Greatly improving momentum measurements at low pT

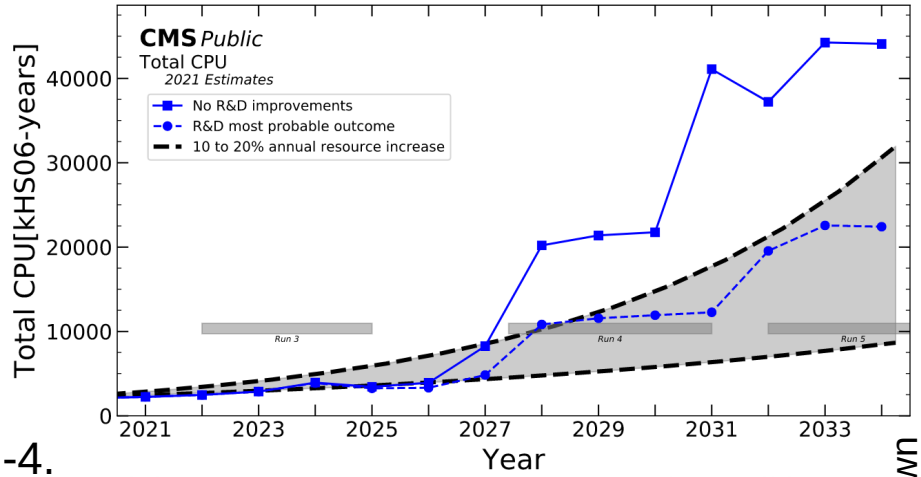
https://doi.org/10.1016/j.nima.2022.167315



Everywhere : Instrumentation++

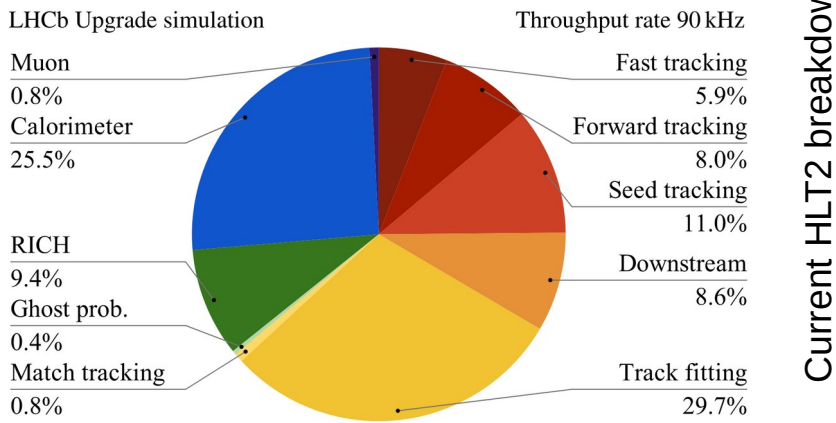
More instrumentation means more work for tracking

- Tracking dominates HLT and offline processing loads
- CMS/ATLAS : GPU acceleration deployed in Run-3 to fit within computing budget



Situation even more challenging for the triggerless model. LHCb:

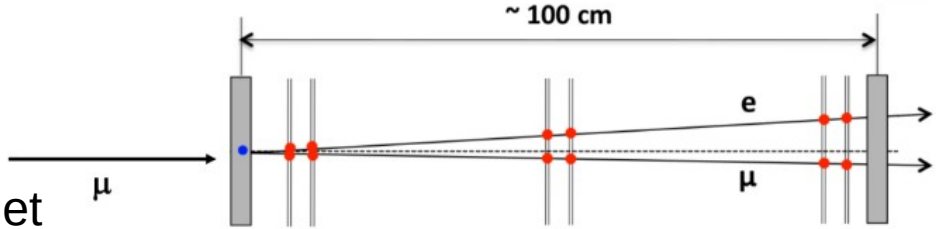
- LHCb envisions HLT2 migration to GPUs in Run-4. In combination with other accelerators?
  - FPGA, IPU, ASICs …
  - Capitalize on momentum in AI
- Several distinct tracking detectors
  - VELO2, UT2, Mightytracker, Magnet Station
  - Detector data formats can impact performance gains from accelerators ...



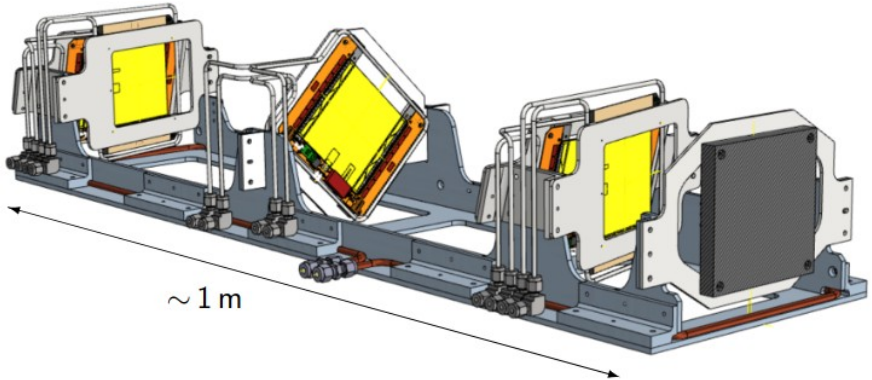
Everywhere : Realtime Adoption

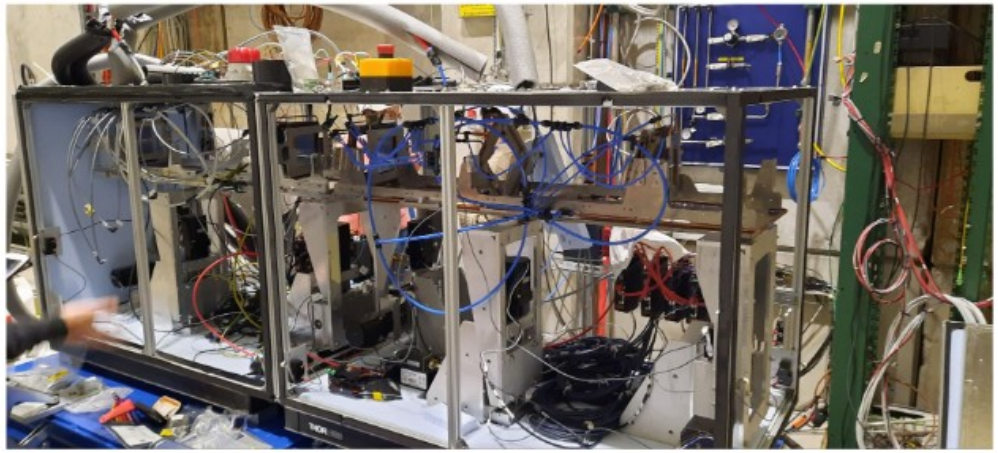
Small experiments also looking to leverage real-time tracking

Example : MUonE



- Proposes to measure hadronic contribution to the running of α_μ . Relevant for g-2
- 150-160 GeV muons scattering on a low-Z target
- M2 beamline at CERN, max intensity ~5e^7 μ/s
- Using double sided strip sensors from CMS upgrade, reading out the 40 MHz stub data
- 450 TB collected since ‘22, streamed to EOS
  - Pilot runs, no online selection yet ...

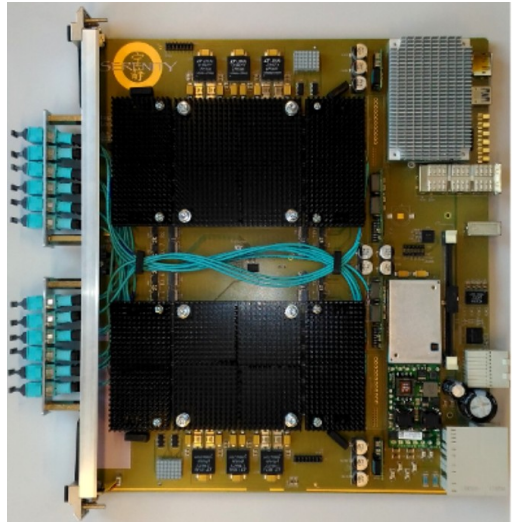


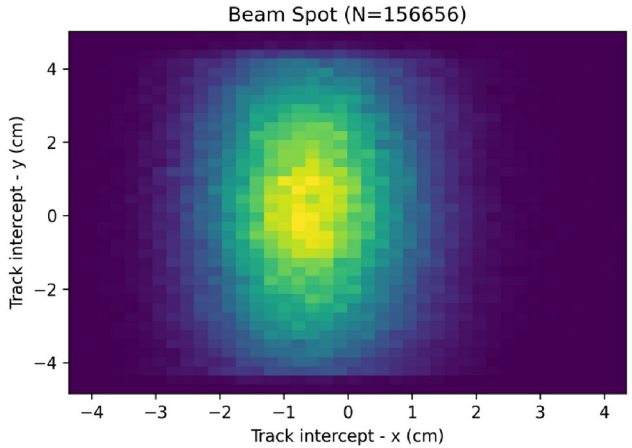


Everywhere : Realtime Adoption

Demonstration of realtime tracking in MuonE

- During a joint beam test with CMS in Aug/Sept'23
- Using prototype CMS 'Serenity' ATCA board
  - 2xKU15p Xilinx FPGA
- Least squares algorithm written in Xilinx High Level Synthesis
  - HLS transforms C++ code to RTL
- Present implementation:
  - No multiple scattering, independent x/y fits
  - 3.75 µs latency @ 320 MHz, not yet pipelined
  - 3% KU15P resources
- Preliminary results are encouraging
  - Online/offline track comparison underway
  - Work progressing on online vertex fitting





Beamspot from realtime tracks

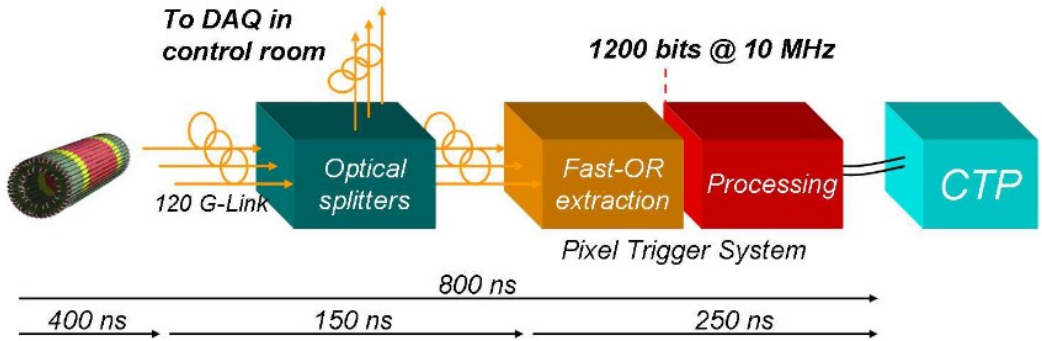
Outline

- **Everything** : Displaced/non-standard tracking, 4D tracking, extended track features
- **Everywhere** : expanding coverage, higher granularity
  - Also, wider application of RT tracking in HEP ...
- **All at once** : pixel in hardware tracking, more intelligence in the front ends, broader movement toward the triggerless model,

All at Once : Level-1 Pixel

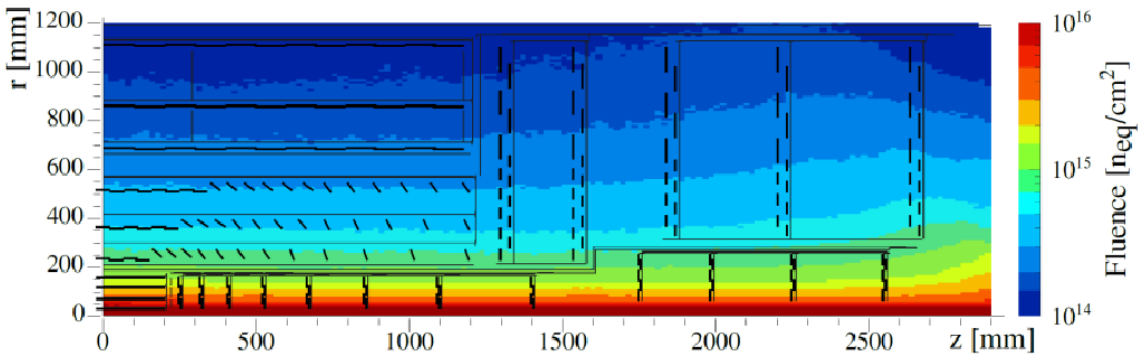
Pixel information ideally available in the hardware trigger

- ALICE had a Level-0 pixel trigger …
  - SPD: 2 layers pixel detectors
  - OR among all pixels on the readout chip
  - 1200 signals, each covering 13x14 mm^2
  - Relatively low input rate, 77 Gbps



G Aglieri Rinella et al 2007 JINST 2 P01007

- Pixel triggering not presently used
  - ALICE migrated to full 50 kHZ Pb-Pb readout
  - Not implemented presently in Belle, nor in the CMS HL-LHC upgrade, for example
- Challenges
  - Enormous data rates
    - Eg: > 3.2 Ghz/cm^2 hit rate in the innermost CMS layer
  - Additional services
  - Integration time vs L1 latency ...

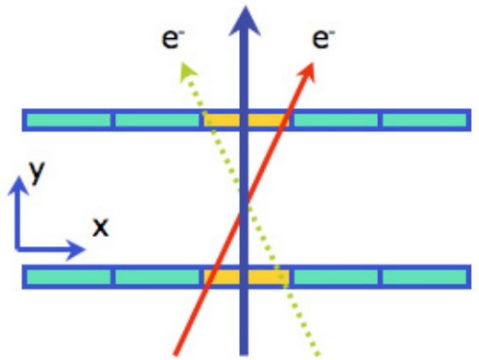


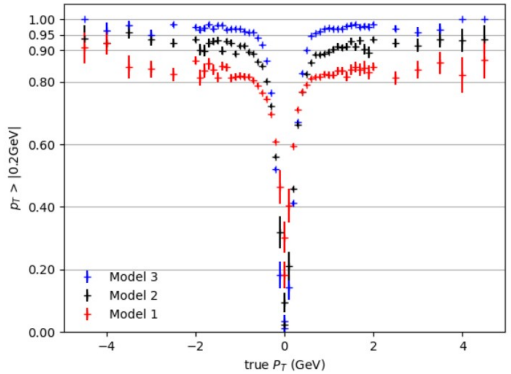
All at Once : Level-1 Pixel

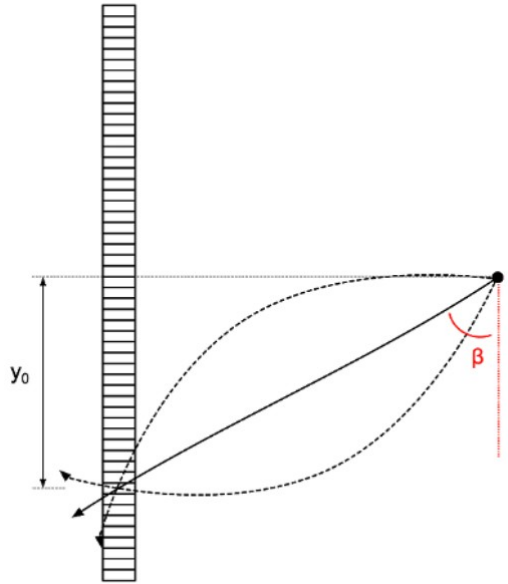
Frontend data reduction

- This would at least reduce the challenges of data rate
- A “stacked pixel” detector was investigated as precursor to the CMS Outer Tracker upgrade
- Further CMS studies conducted on a “Level-1.5” pixel trigger
  - ROI seeded by the calorimeter [arXiv:2211.15276v1]
- Investigations continue, eg: “SmartPixels”
  - Filter out low pT particles, noise, background using cluster features
  - With a data-trained Neural Net implemented on-chip, 28 nm
  - Potentially achieveing 50-75% bandwidth reduction

[arXiv:2310.02474v1]







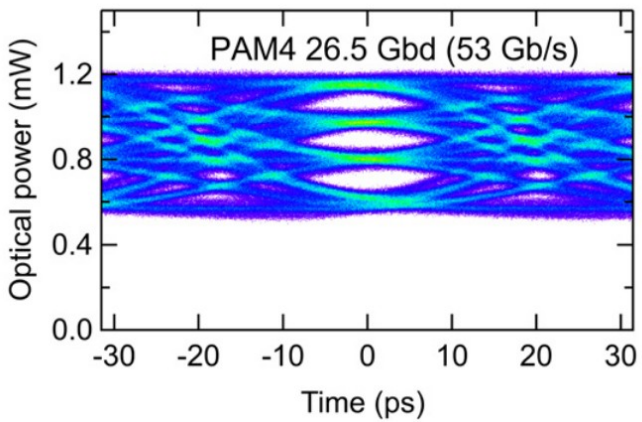
All at Once : Level-1 Pixel

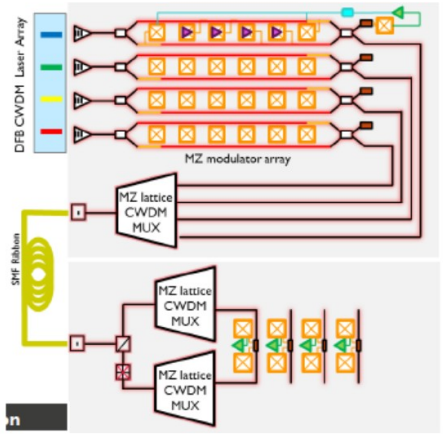
Improving Bandwidth

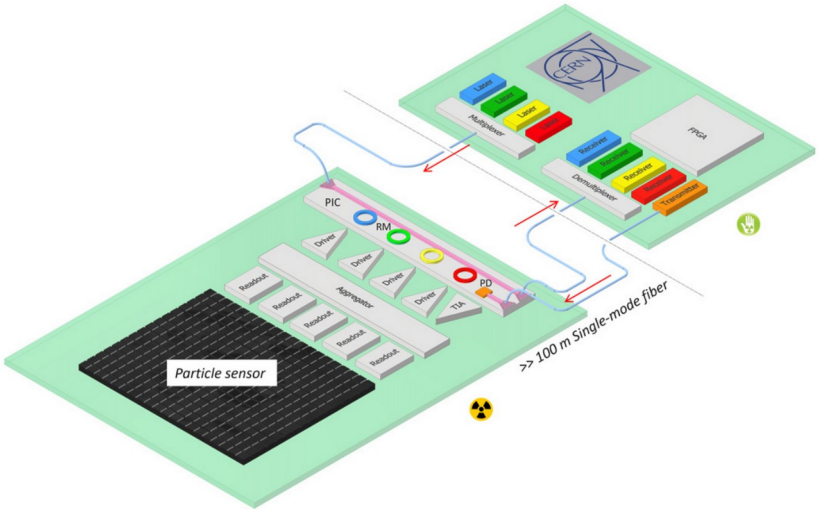
- Even with data reduction, a Level-1 pixel trigger would likely need more bandwidth …

Focused efforts in radiation-hard silicon photonics

- Eg : ring modulators (small, low power), MZM (temperature insensitive)
- CMOS process compatible, several multiplexing schemes
- 25-50 Gbps links already demonstrated in HEP







DRD7 - Future Links

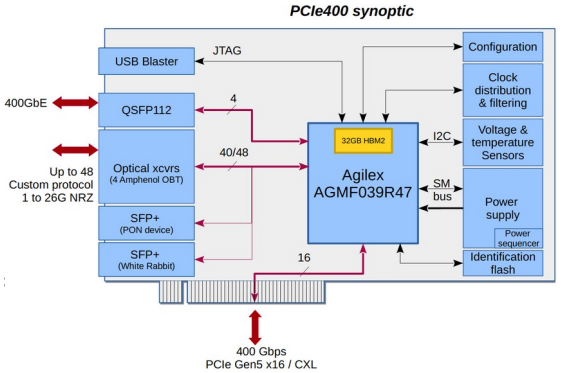
C. Scarcella et al 2023 JINST 18 C03002

All at Once : Software Trigger

Improving Bandwidth

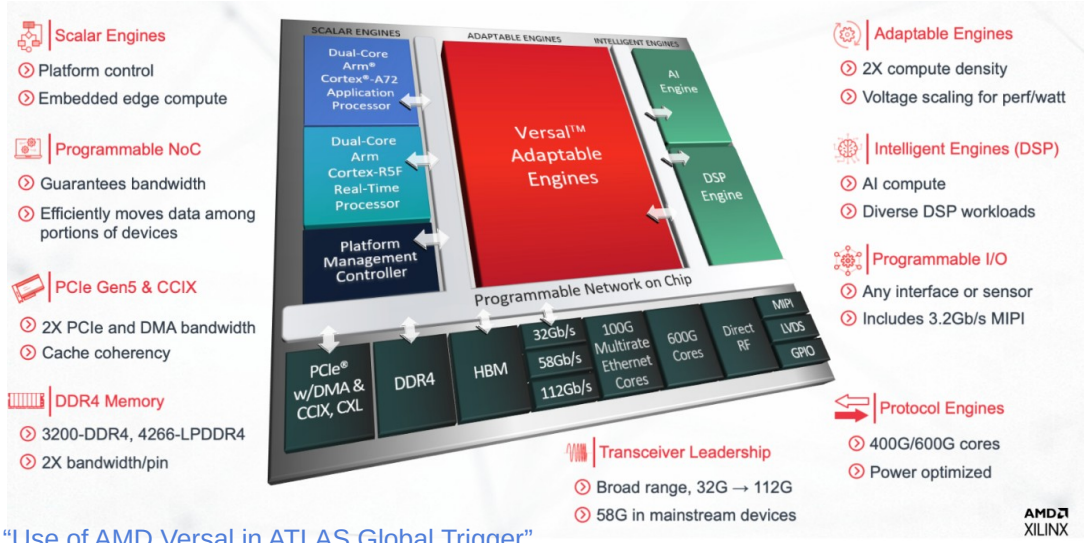
J. Langouët , “Future DAQ Boards : PCIe400”

- Triggerless model will also need much more bandwidth
- LHCb R&D on PCIe*400*, possibly w/ photonics in VELO2 FE



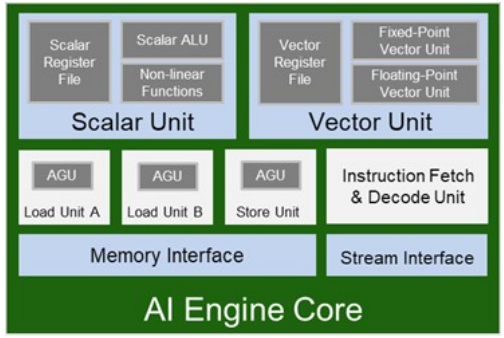
Tighter integration of heterogeneous computing

- See ATLAS use of Xilinx Versal in the HL-LHC L0 Global Trigger
- Relevant for s/w: various capabilities on-chip



D. Sankey, “Use of AMD Versal in ATLAS Global Trigger”

- Scalar Engines: traditional ARM + hard realtime cores
- Adaptable Engines : FPGA fabric, including DSPs
- AI Engine : RISC & vector SIMD processors0



Other Ideas

HIP bits, dE/dX in the h/w : exotic physics (HSCP)

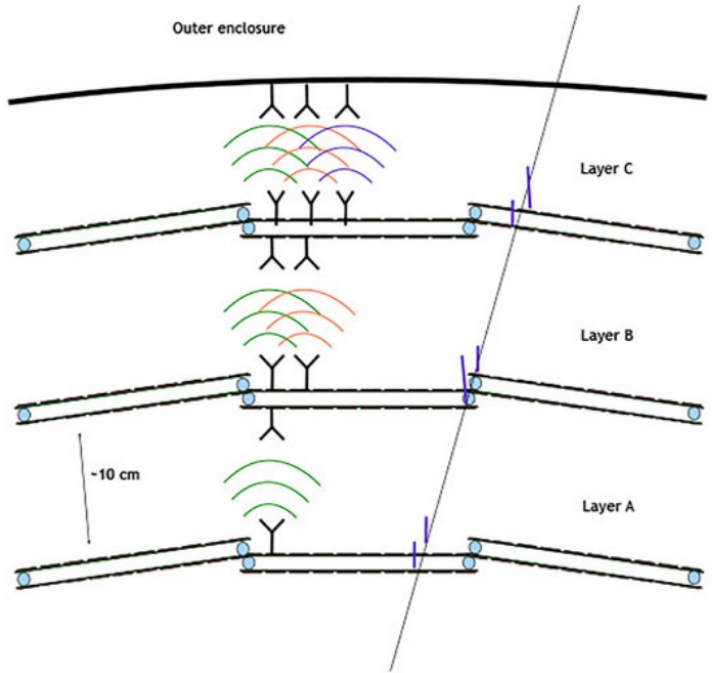
Low power ARM processors for HLT applications …

- Data center efficiency a growing concern

L1 Scouting with tracking data ...

A favorite : wireless inter-module communication

...



Summary

Not able to cover all of the interesting R&D projects

Nor do any of them justice ...

- Today's talks will paint a more interesting and complete picture!

Hopefully it's already clear that the landscape of track triggering R&D is *not static* ...

Summary

Not able to cover all of the interesting R&D projects

Nor do any of them justice ...

- Today's talks will paint a more interesting and complete picture!

And that its future is unfolding in

realtime