TWEPP 2016 Karlsruhe

An FPGA based track finder at Level 1 for CMS at the High Luminosity LHC

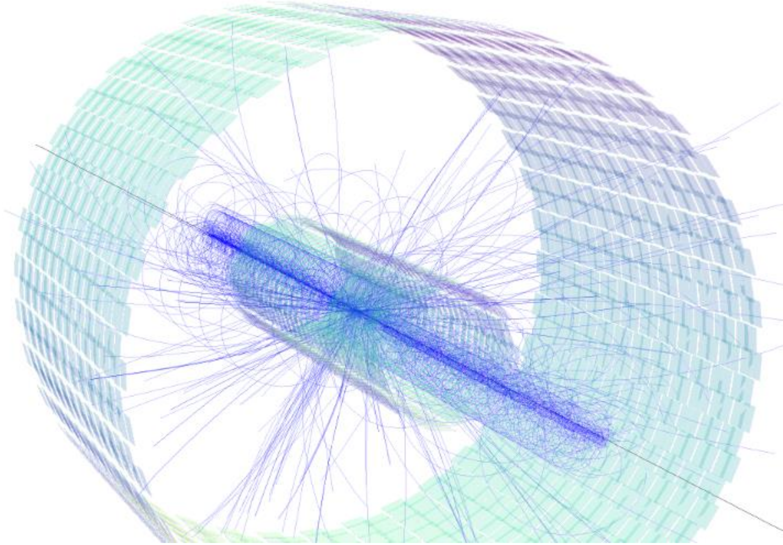
Mark Pesaresi (Imperial College London)

F.Ball, J.Brooke, E. Clement, D.Newbold, S. Paramesvaran, P.Hobson, A. Morton, I.Reid, P. Vichoudis, G.Hall, G.Iles, T.James, M.Pesaresi, A.Rose, A.Shtipliyski, S.Summers, A.Tapper, K.Uchida, L. Ardila, M. Balzer, M. Caselle, B. Oldenburg, T.Schuh, M. Weber, L.Calligaris, D.Cieri, K.Harder, K.Manalopoulos, C.Shepherd, I.Tomalin, T.Matsushita

upgrading the CMS tracker

tracker replacement essential at HL-LHC (post-2025)

- because of radiation damage and high pileup



additionally, Level 1 (hardware) trigger must be substantially upgraded

- **<#interactions> ~ 140 – 200** (currently ~20-40)
- 750 kHz max accept rate (currently 100kHz)

Calorimeter Trigger issues

isolation of e/γ/τ degraded by pileup
many more jets, which overlap

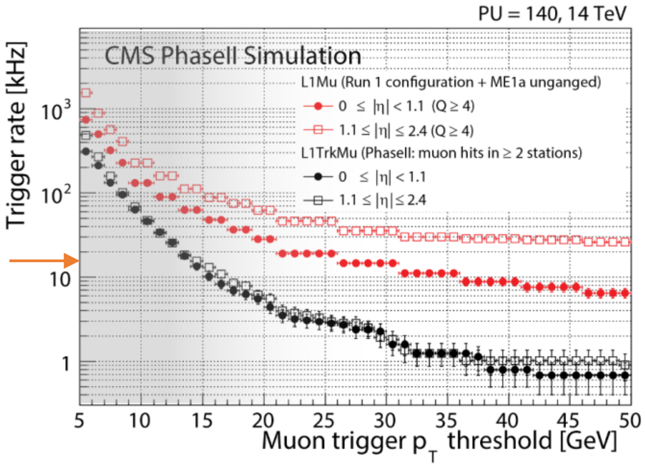
Muon Trigger issues

increased combinatorial fakes, enhanced by multiple scattering

to control much higher rate of L1 trigger only significant new source of data will come from tracker

- **CMS must access tracks at L1 to succeed at HL-LHC**

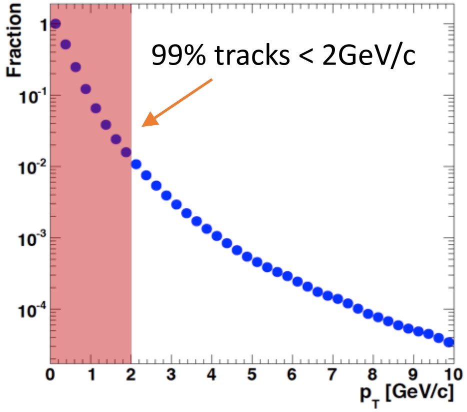
muon trigger example



L1 tracker primitives

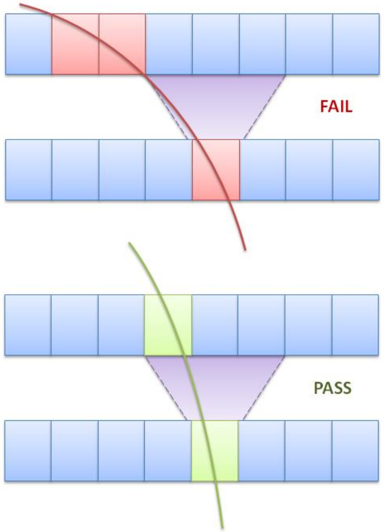
impossible to transfer all data off-detector for decision logic so **on-detector data reduction (or selective readout) essential**

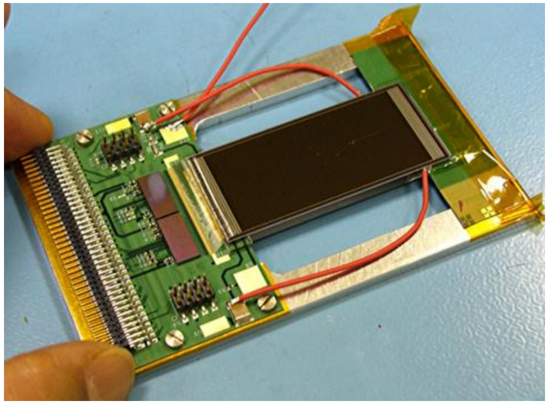
- tracks with transverse momentum < 2GeV/c not useful for triggering



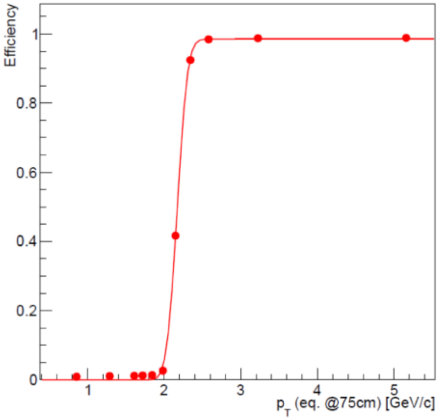
concept of stacked tracking

- modules made of closely spaced, O(mm) separated, sensors
- ASICs only forward hits detected on each sensor that lie within a pre-determined correlation window
- these **"stubs" indicate presence of a crossing high pT track**

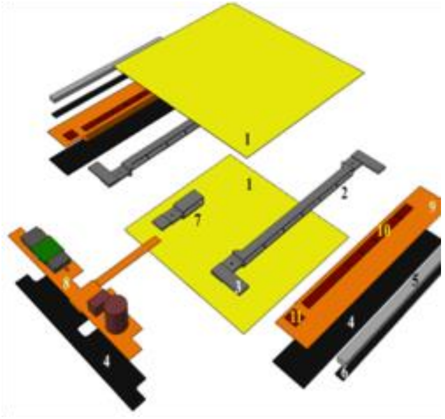




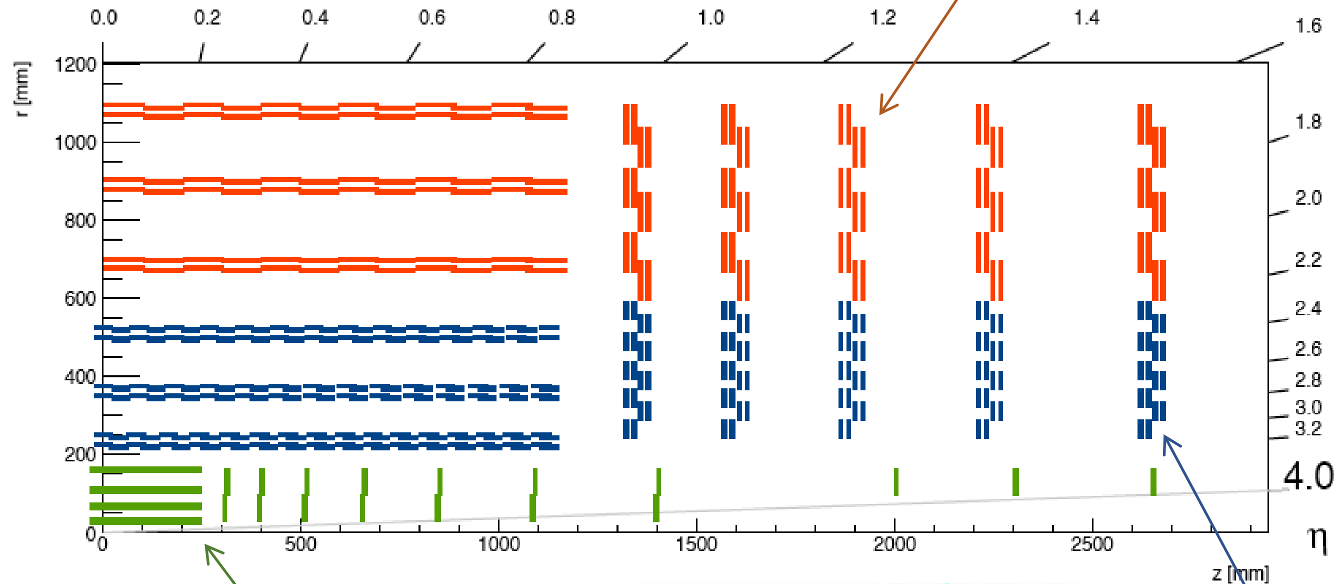
TWEPP2013 Characterization of the CBC2 readout ASIC for the CMS strip-tracker high-luminosity upgrade

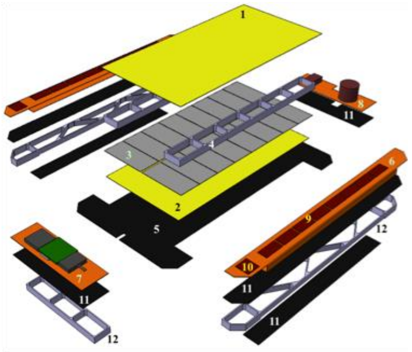


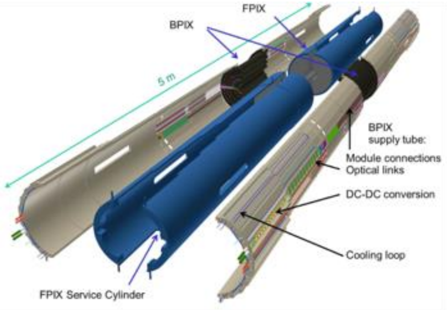
CMS tracker post-2025



Strip/Strip Modules
90 µm pitch/5 cm length







Inner Pixel
Covers up to η=4.0

Strip/Pixel Modules
100 µm pitch/2.5 cm length
100 µm x 1.5 mm "macropixels"

L1 track finder challenges

L1 trigger will require **quasi-full track reconstruction** for charged particles with transverse momentum > ~2 GeV/c

but full tracking at Level 1 is an incredible technical challenge

- data-rates: **O(100 Tbps)**
- occupancy & combinatorics: up to **20k stubs/event**
- latency: **~5 µs** (~12.5 µs for L1 overall)

how to find the tracks in ~5 µs with high efficiency and acceptable fake rates?

- **huge architectural & dataflow implications for hardware**

large fraction of hits are not track-related (fake combinations, conversions, secondaries,…)

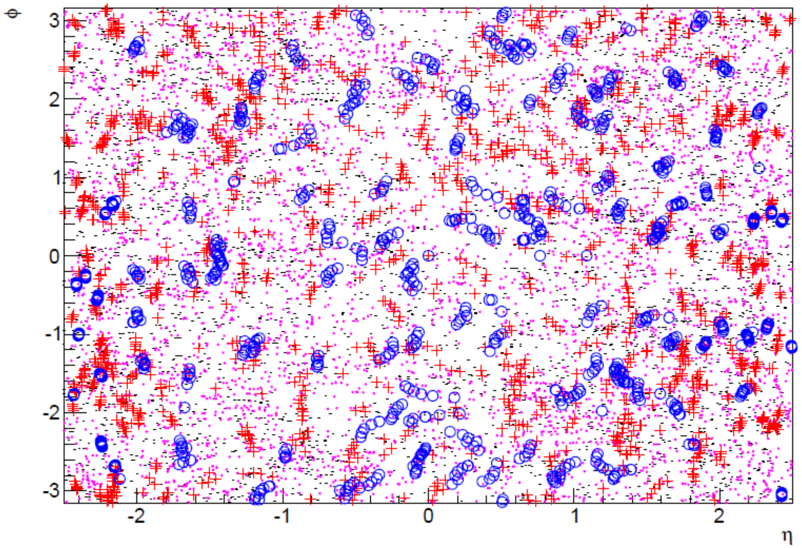
40% combinatorial/v. low energy electrons

40% charged particles, p_T < 1 GeV/c

13% charged particles, 1 < p_T < 2 GeV/c

7% charged particles, p_T > 2 GeV/c

140 interactions/event, no signal



L1 Track Finder R&D

CMS pursuing three complementary designs to confront the challenge

purely FPGA based:

1) Hough Transform Track Finder & Combined Filter/Fitter

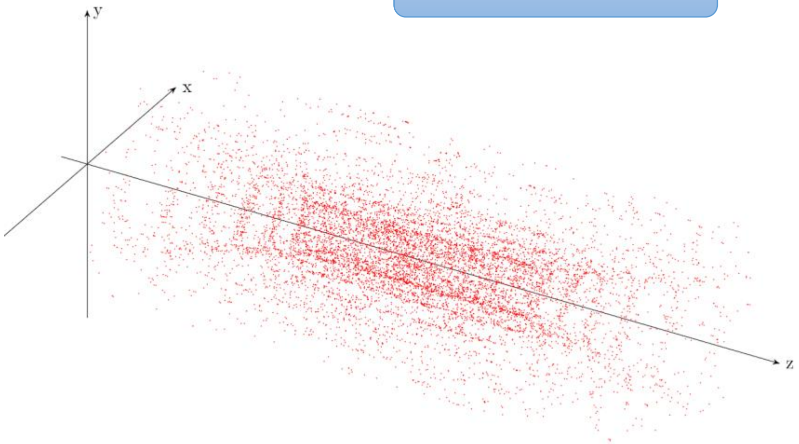
2) Combined Tracklet Builder & Linearized χ^2 Track Fit

ASIC assisted:

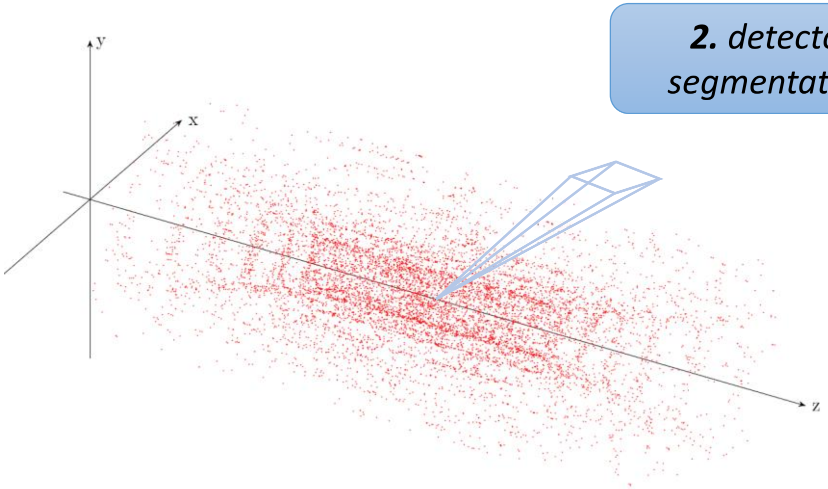
3) Associative Memory Pattern Recognition + FPGA PCA Track Fit

feasibility demonstrations by December 2016

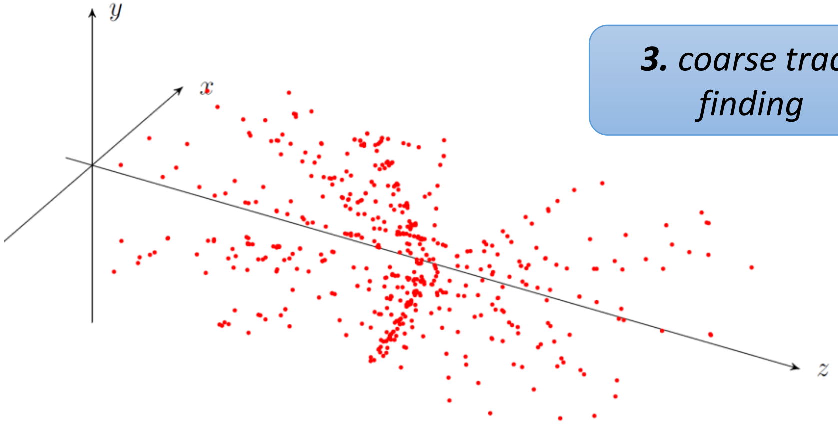
FPGA based track finding overview



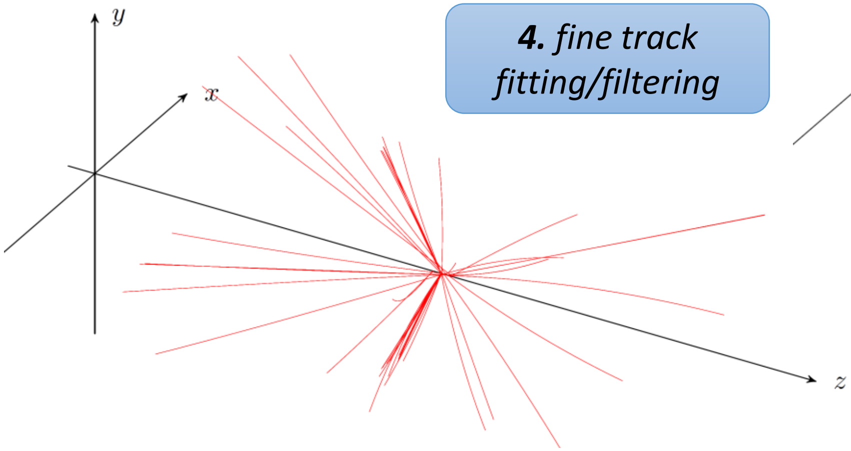
***1.** tracker stubs*



***2.** detector segmentation*



***3.** coarse track finding*



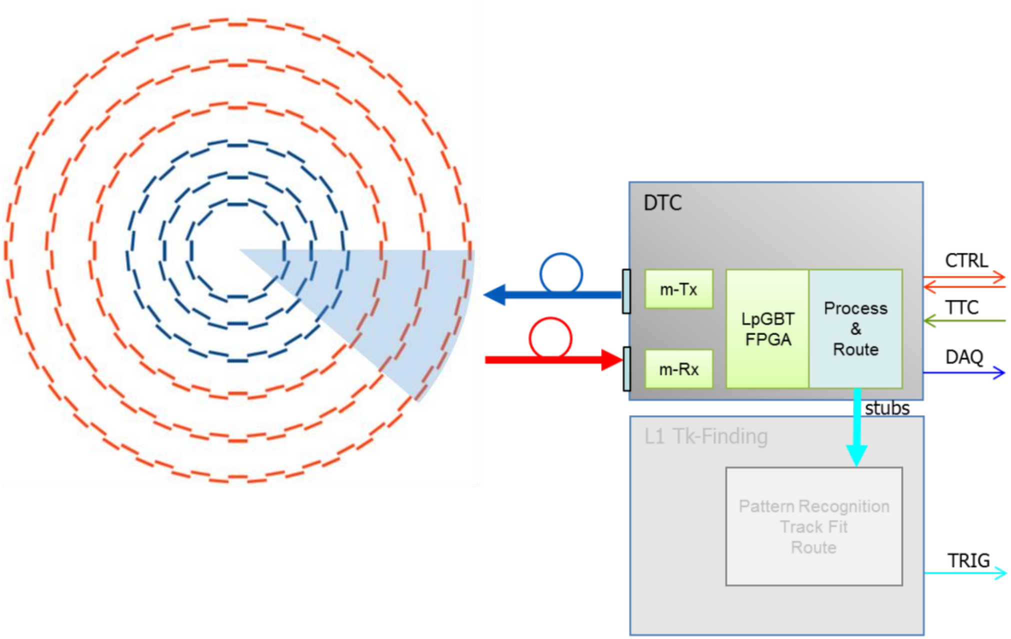
***4.** fine track fitting/filtering*

data organisation

detector to be interfaced to a first layer of off-detector hardware known as the **Data, Trigger and Control (DTC) system**

DTC to configure and read out tracker modules – including trigger data at 40MHz

total system to comprise **256 boards**



32 DTCs control ~1900 modules (in 1/8 of tracker in ϕ, or **‘octants’**)

DTC to also implement low level stub manipulation e.g. global coordinate conversion, duplication, routing to next layer (L1 Track Finder)

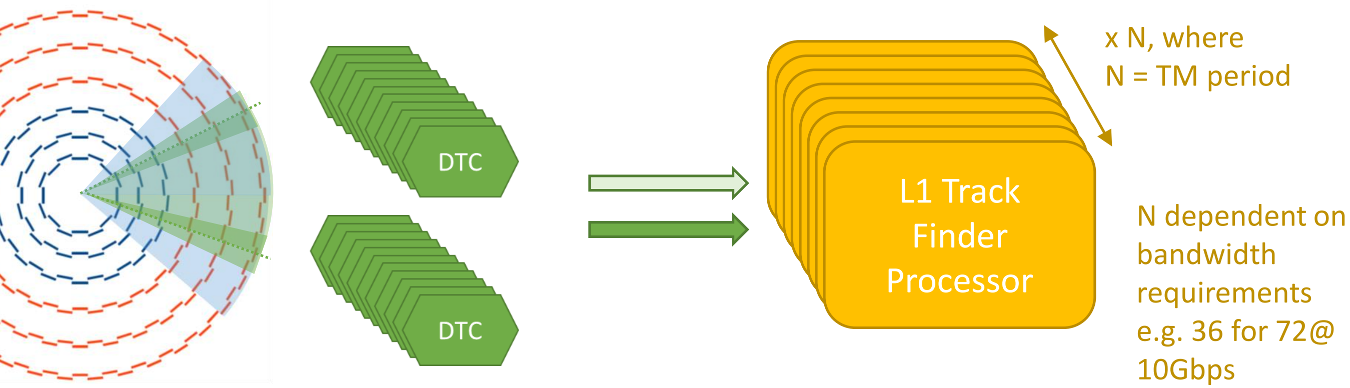
FE -> DTC output latency ~1μs

as part of routing to next layer, allow DTC to **'time-multiplex'** stubs

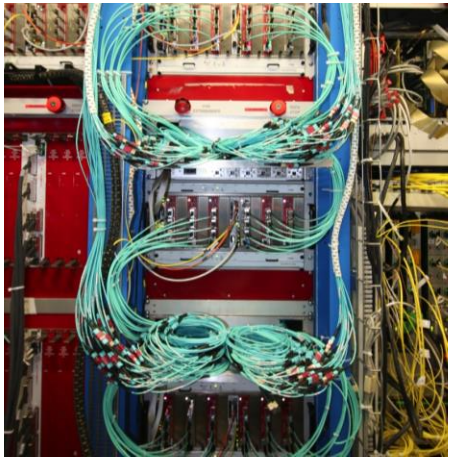
one L1 Track Finder Processor handles 1 in N events

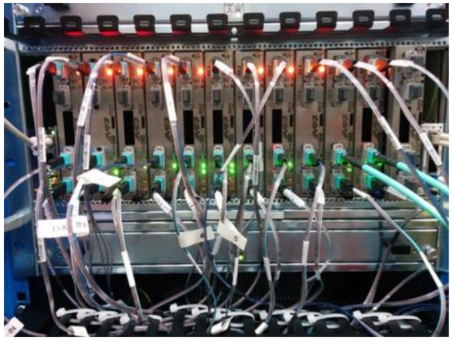
keep octant segmentation, with offset to allow DTC to handle duplication across detector cabled boundaries automatically

L1 Track Finder Processor handles all data in event from 2π/8 of tracker over TM period



N dependent on bandwidth requirements e.g. 36 for 72@ 10Gbps





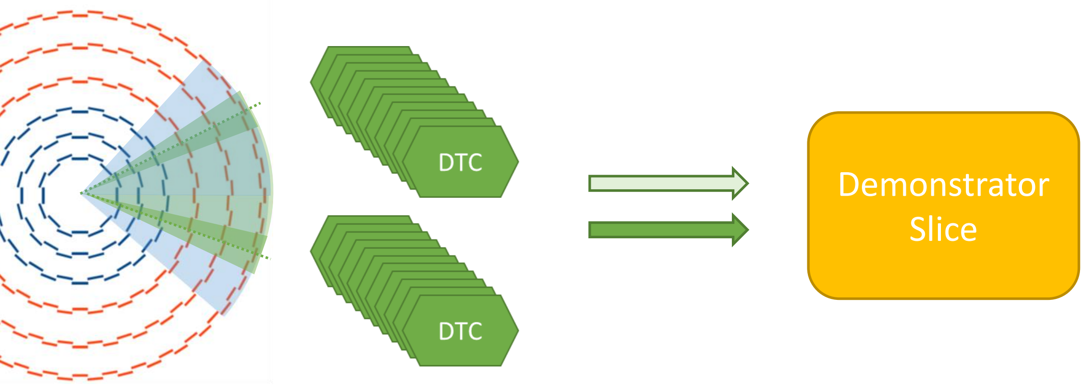
L1 calorimeter trigger (TM=9)

data organisation

no downstream communication/duplication between regions required

factorised system, N x 8 **independent and identical processors**

architecture well suited to slice demonstrations



receives 1/N events

data organisation in FPGA

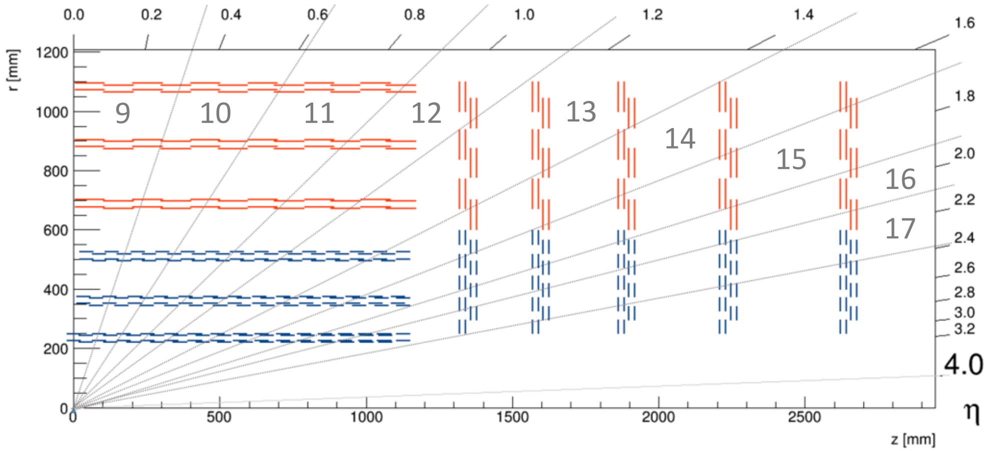
to simplify job of track finding, detector octant is further segmented in L1 processor

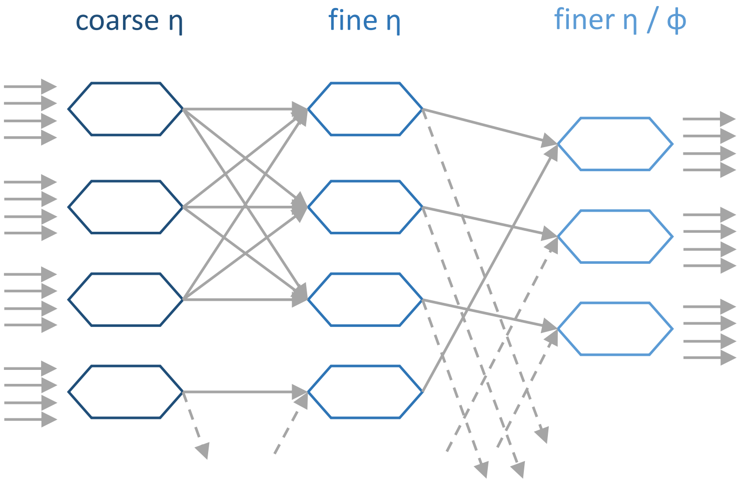
stubs are assigned to segments in η, ϕ according to their coordinates and local bend

flexibility to choose segmentation depending on track finder requirements

- e.g. 18 in η and 2 ϕ currently

pre-processing of stubs before subsequent track finding stage





small lightweight address based routing network
fast – tested to 450MHz on KU115

track finding with the Hough Transform

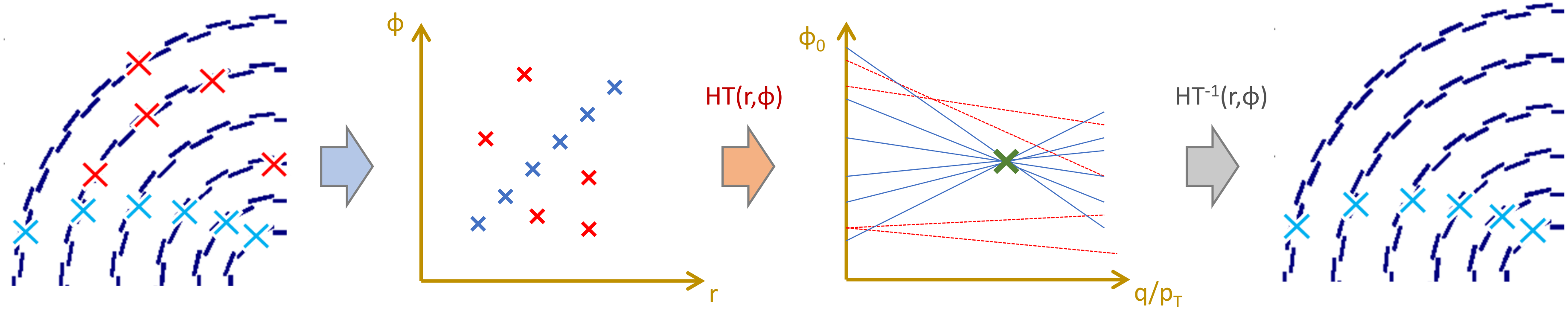
two step track finding approach based on coarse **2D Hough Transform**

well known technique used in image manipulation, including identifying tracks in bubble chamber photographs

the Hough Transform track finder is the workhorse of the design

- orders stubs into valid track candidates
- binning of stubs according to projections, determining coarse track parameters

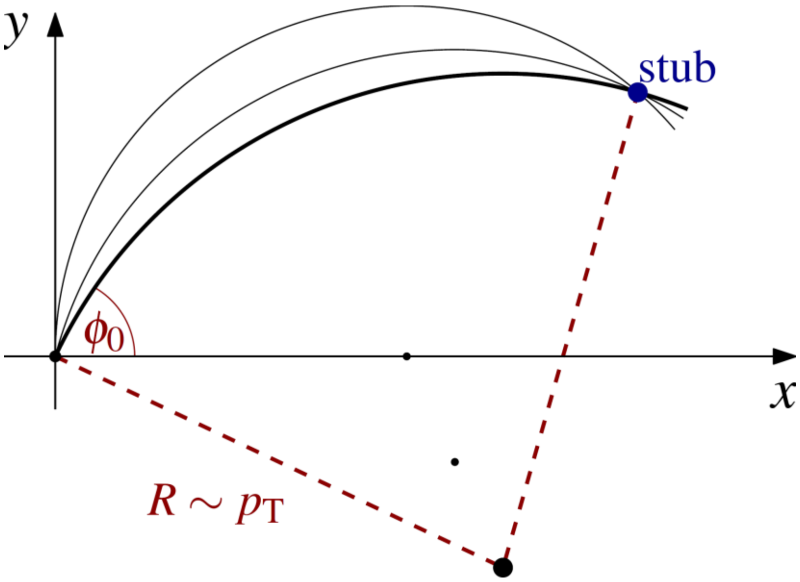
subsequent track fit uses full hit resolution and determines fine grained track parameters, on reduced data volume



r-ϕ Hough Transform

search for tracks in the transverse plane in r-ϕ

infinite number of circles with unique (R,φ_0) between origin and measured stub position at (r,φ)



but track parameters are correlated

$$\phi \simeq \frac{q}{p_T} r - \phi_0$$

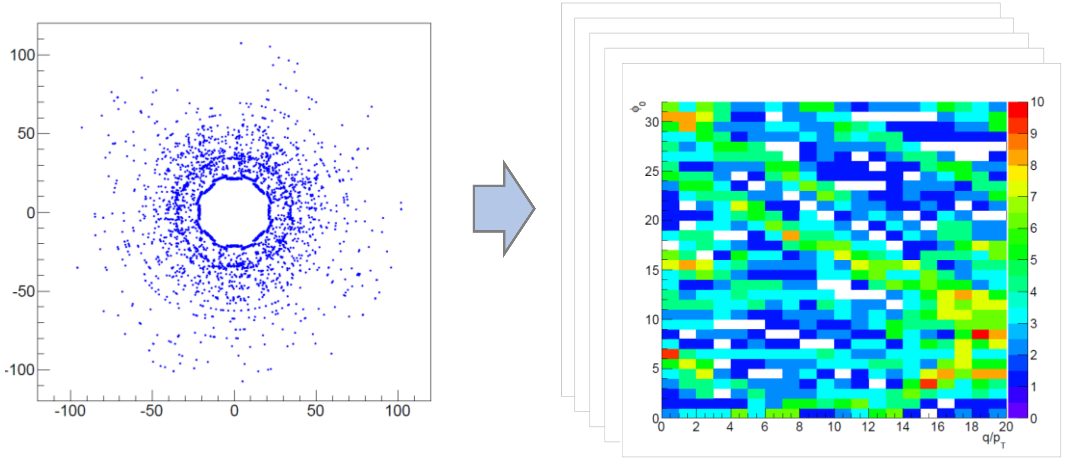
using small angle approximation

additionally initial coordinate transformation to r=58cm helps distribution of hits in parameter space

Hough Transform applied to Tracker

apply Hough Transform in 2D

bin stubs in multiple Hough arrays, segmented in η,φ



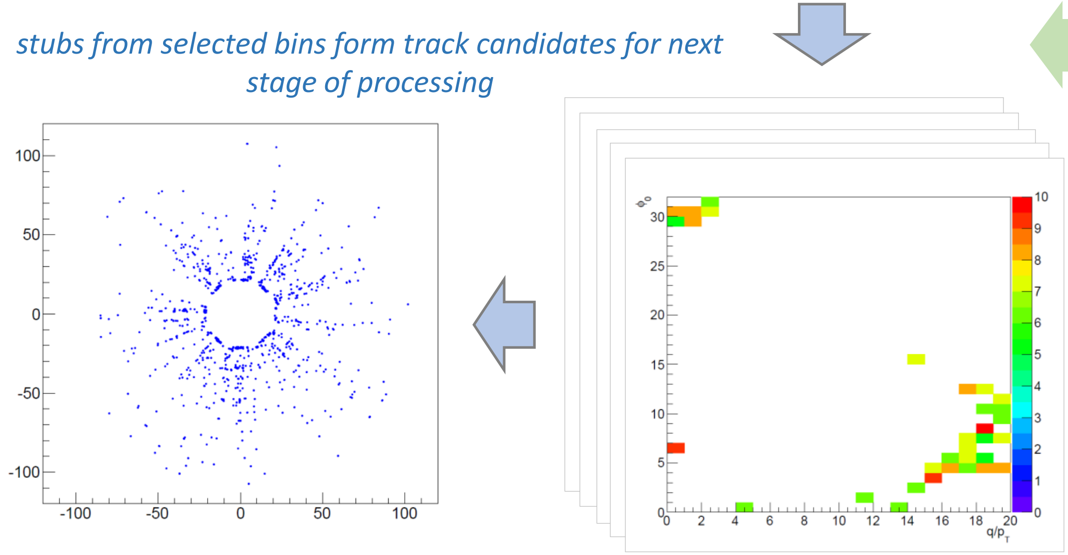
32x64 Hough array size required

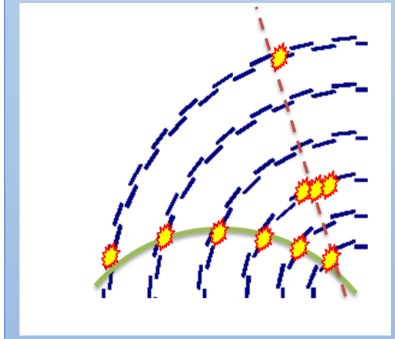
36 segments in ϕ, η per processor

min pT = 3 GeV/c

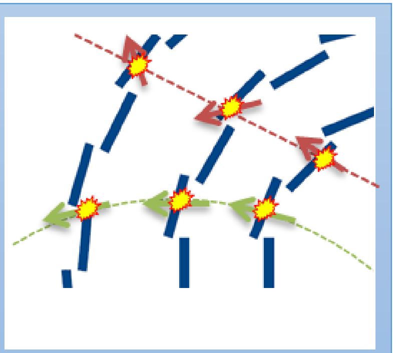
stubs from selected bins form track candidates for next stage of processing

apply track criteria to accept



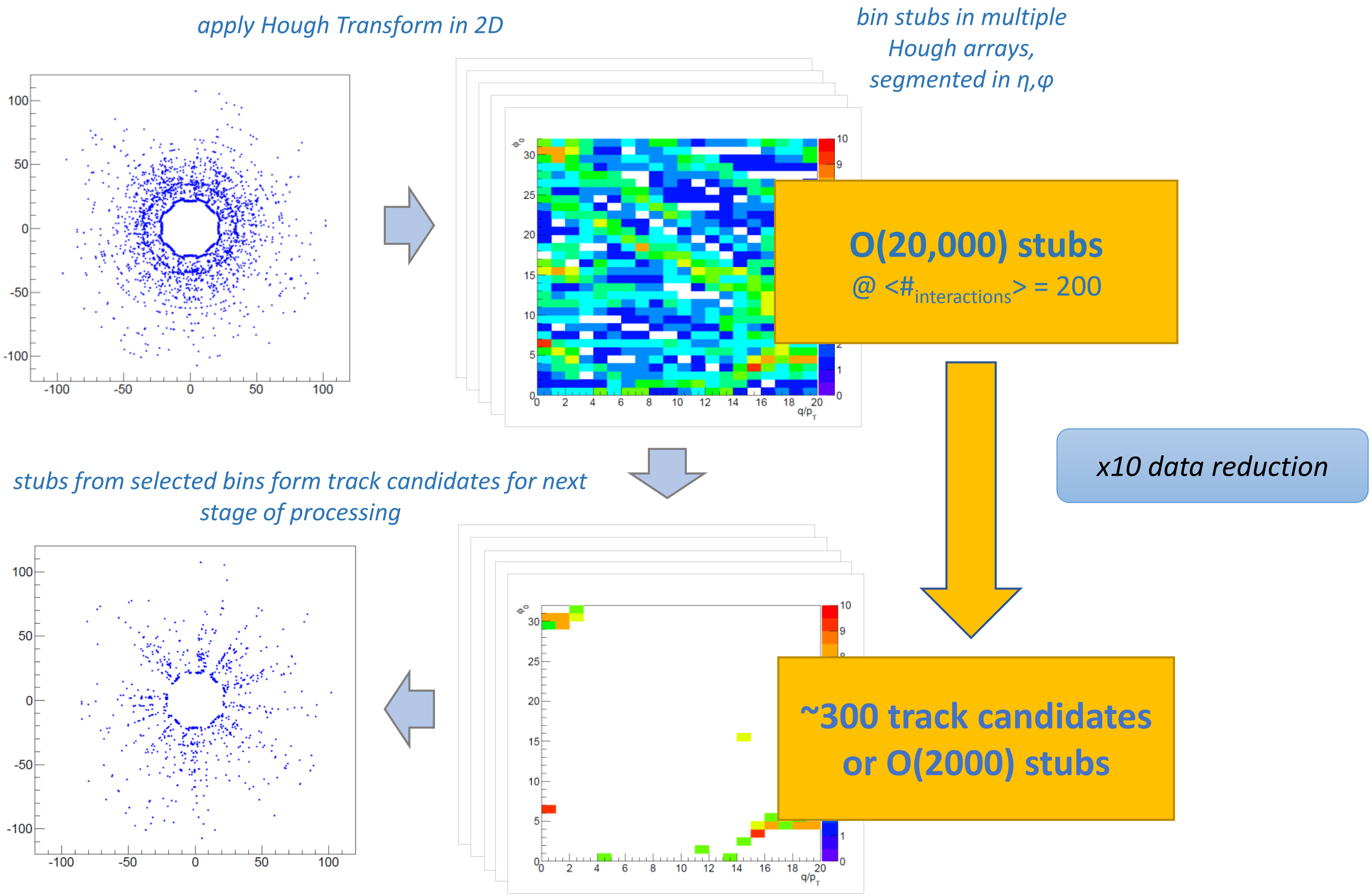


bins with more than 5 stubs at unique radii



bins with stubs that have compatible local bend

Hough Transform applied to Tracker



apply Hough Transform in 2D

bin stubs in multiple Hough arrays, segmented in η,φ

O(20,000) stubs

@ <$\#_{\text{interactions}}$> = 200

x10 data reduction

~300 track candidates or O(2000) stubs

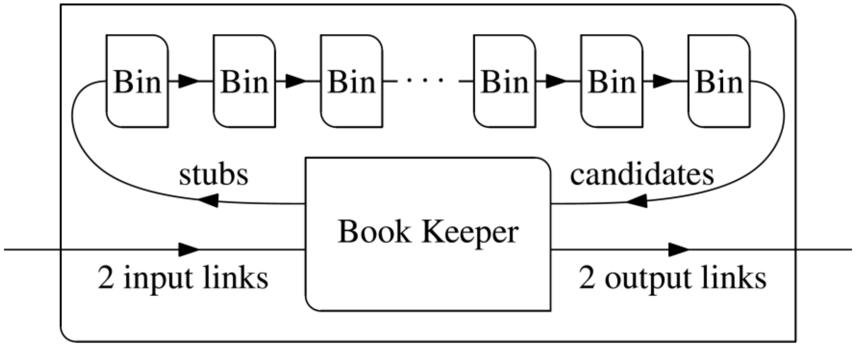
stubs from selected bins form track candidates for next stage of processing

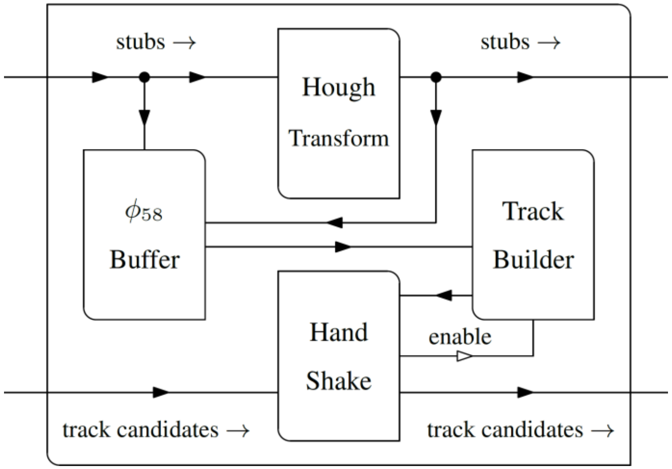
r-ϕ Hough Transform in FPGA

one array per segment

array is implemented as a pipeline, processing one stub per 240MHz clock cycle

first step is to **fill the array**, second step is to **read out the track candidates**





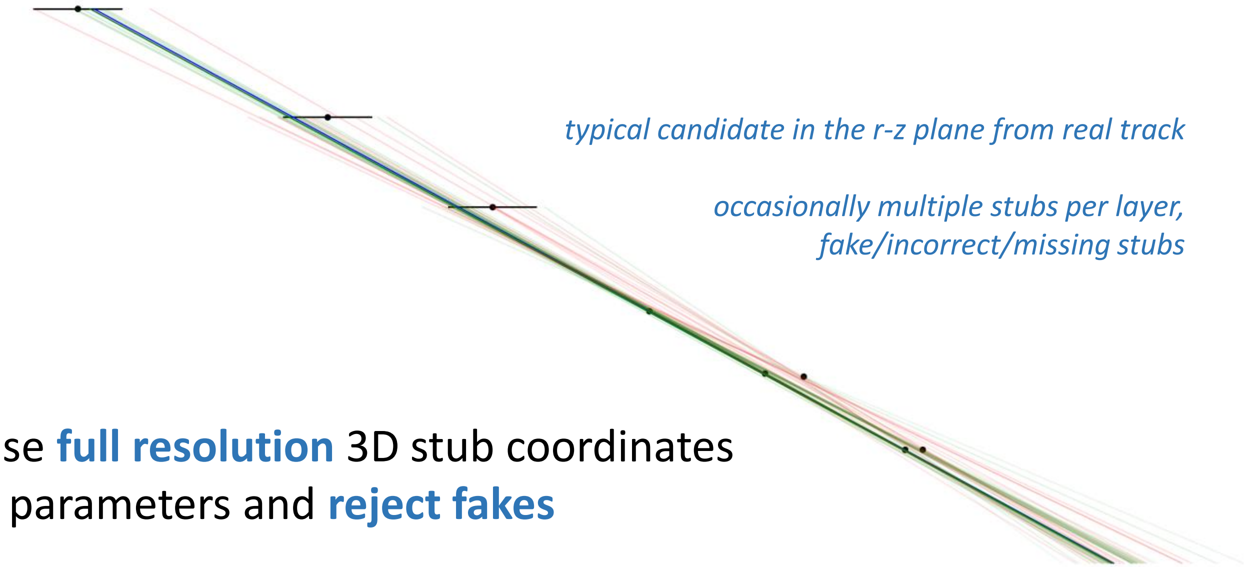
1 bin corresponds to a column in the Hough array

1 paged block RAM per bin implements the 64 rows in the array column

track fitting and filtering

rates out of HT **vastly reduced** and stubs grouped into candidates

but candidate quality is generally poor



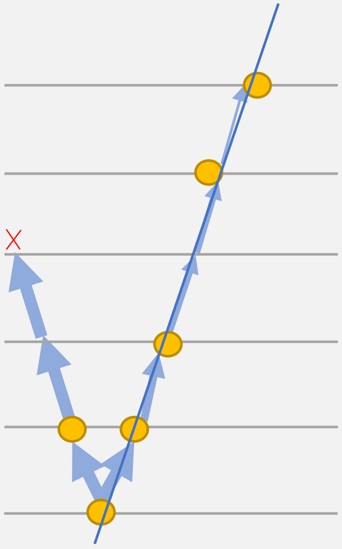
typical candidate in the r-z plane from real track

occasionally multiple stubs per layer, fake/incorrect/missing stubs

combined fitters/filters to use **full resolution** 3D stub coordinates to reconstruct precise track parameters and **reject fakes**

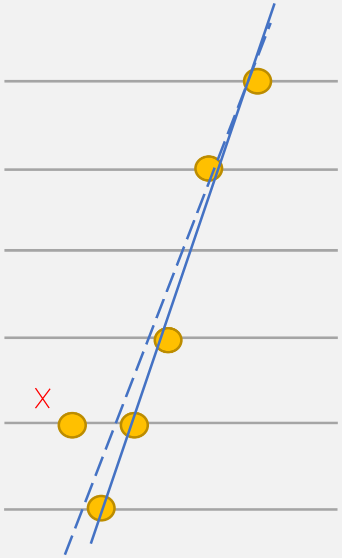
Kalman Filter

- default offline fit
- uses full information
- incorrect trajectories rejected
- allows for scattering, 5 or 4 parameter fits
- mathematically heavy

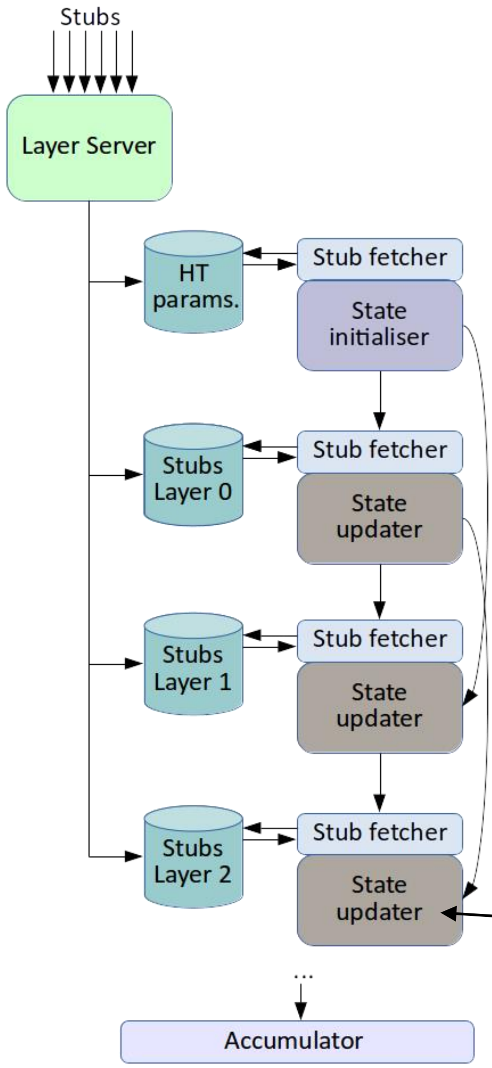


Linear Regression

- tracks are essentially straight lines
- residuals are minimised
- worst residual stubs are rejected iteratively
- simple & potentially fast
- non-linear effects not modelled



Kalman filter in FPGA



4 parameter fit to begin with, (q/pT, ϕ_0, z_0, η)

propagate track parameters inside-out, **using HT parameters and segment to seed state**

use stubs to update track parameters

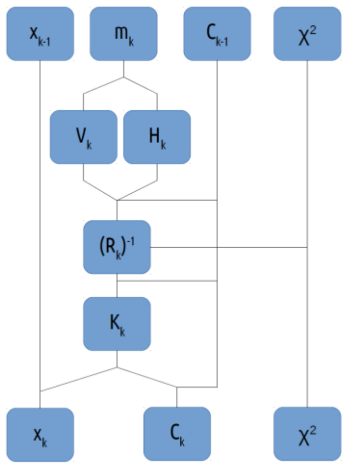
hits incompatible with the measurement & propagation errors are **rejected at each step (layer)**

multiple stubs per layer split into new candidates

implemented with High Level Language (MaxJ) – see talk from S. Summers

$$x_k^{k-1} = \mathbf{F}_{k-1} x_{k-1}$$
$$\mathbf{C}_k^{k-1} = \mathbf{F}_{k-1}\mathbf{C}_{k-1}\mathbf{F}_{k-1}^T + \mathbf{Q}_{k-1}$$
$$r_k^{k-1} = m_k - \mathbf{H}_k x_k^{k-1}$$
$$\mathbf{R}_k^{k-1} = \mathbf{V}_k + \mathbf{H}_k \mathbf{C}_k^{k-1}\mathbf{H}_k^T$$
$$\mathbf{K}_k = \mathbf{C}_k^{k-1}\mathbf{H}_k^T\left(\mathbf{R}_k^{k-1}\right)^{-1}$$
$$x_k = x_k^{k-1} + \mathbf{K}_k r_k^{k-1}$$
$$\mathbf{C}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{C}_k^{k-1}$$
$$\chi_+^2 = r_k^{k-1T}\left(\mathbf{R}_k^{k-1}\right)^{-1} r_k^{k-1}$$
$$\chi_k^2 = \chi_{k-1}^2 + \chi_+^2$$

- k-1, k: previous layer, current layer
- x: track helix parameters
- **C**: their covariance matrix
- m: a measurement (stub)
- **F**: forecast matrix
- **H**: measurement matrix
- **K**: the Kalman gain



Linear Regression in FPGA

also a 4 parameter fit

assumes that:

i) tracks are **straight lines** in each projection

ii) initial candidates are of **reasonable quality**

at each step, track is progressively cleaned by fitting then **removing hit with worst residual**

filter criteria vary depending on track quality at each step

r-ϕ plane

$$m = \frac{n\overline{r\phi} - \overline{r}\,\overline{\phi}}{\text{denom}}$$

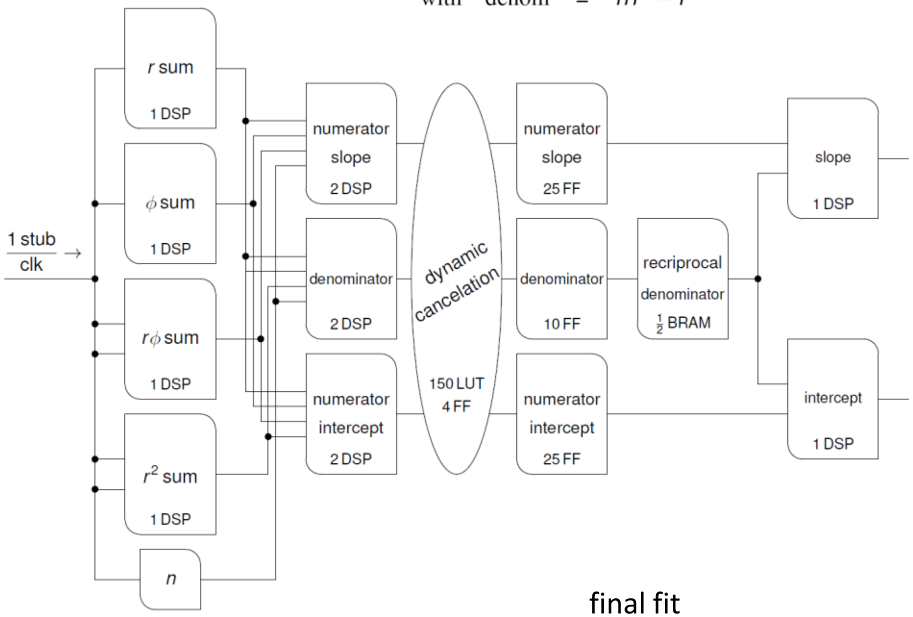
$$c = \frac{\overline{r^2}\,\overline{\phi} - \overline{r}\,\overline{r\phi}}{\text{denom}}$$

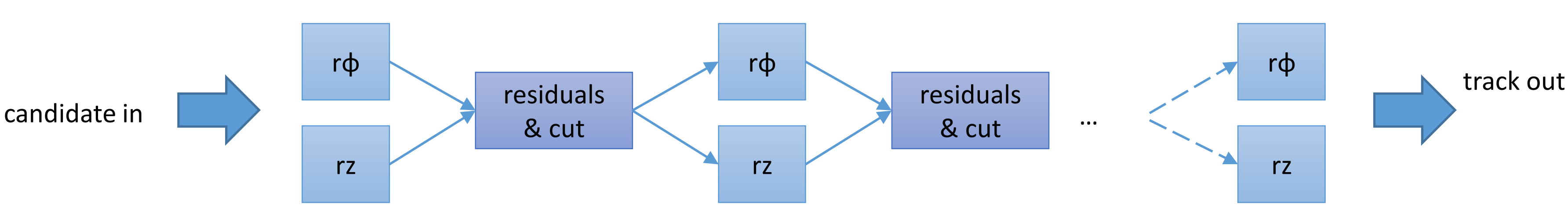
r-z plane

$$m = \frac{n\overline{rz} - \overline{r}\,\overline{z}}{\text{denom}}$$

$$c = \frac{\overline{r^2}\,\overline{z} - \overline{r}\,\overline{rz}}{\text{denom}}$$

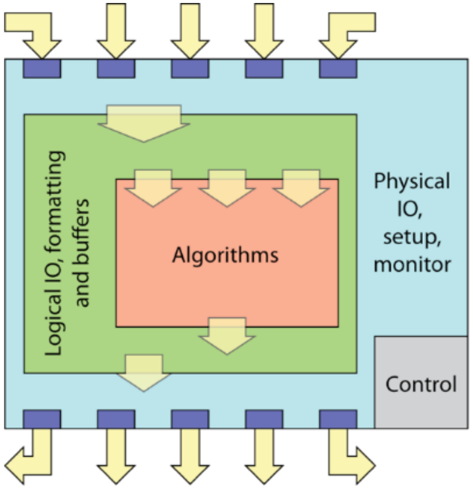
$$\text{with} \quad \text{denom} = n\overline{r^2} - \overline{r}^2$$





based on the **Imperial MP7**

- Virtex 7 690
- 72 optical I/O up to 12.5Gbps
- MTCA, total optical b/w 0.9Tbps


Logical IO, formatting and buffers
Algorithms
Physical IO, setup, monitor
Control

segregated infrastructure and algorithm payload firmware regions

links augmented with buffer RAMs

integrated build system & firmware management

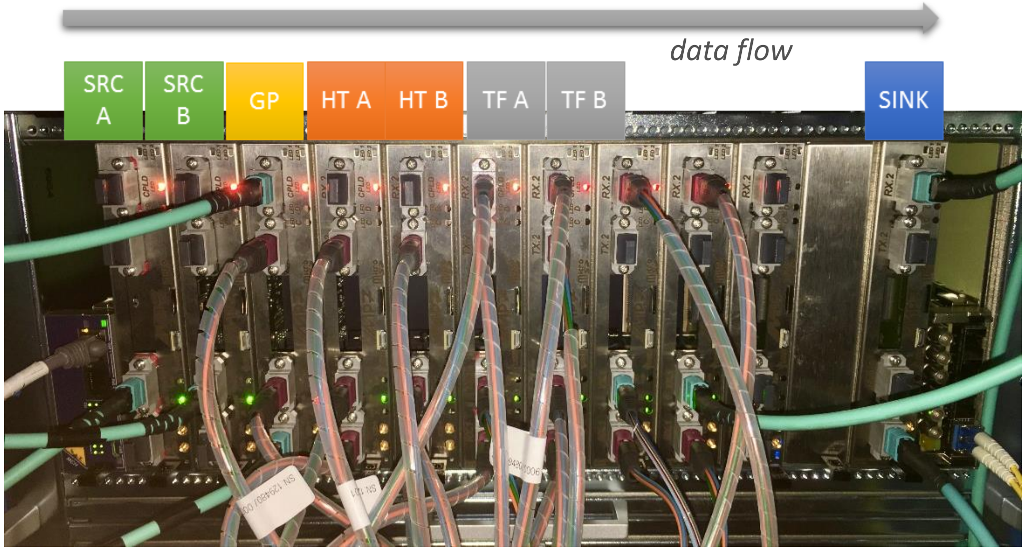
infrastructure supported as part of CMS L1 trigger

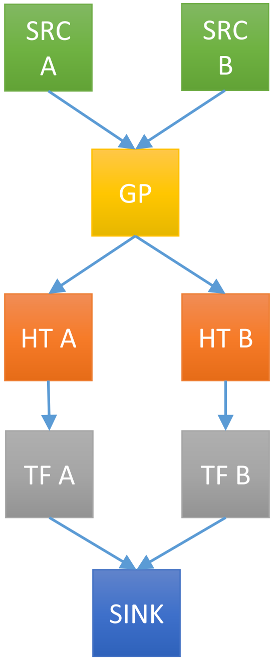
- **maximise reuse of existing technology and lower barrier to entry for algorithm development & testing** demonstrator is divided into **logical elements**, each on separate MP7 boards

- 1/36 event time-multiplexing period

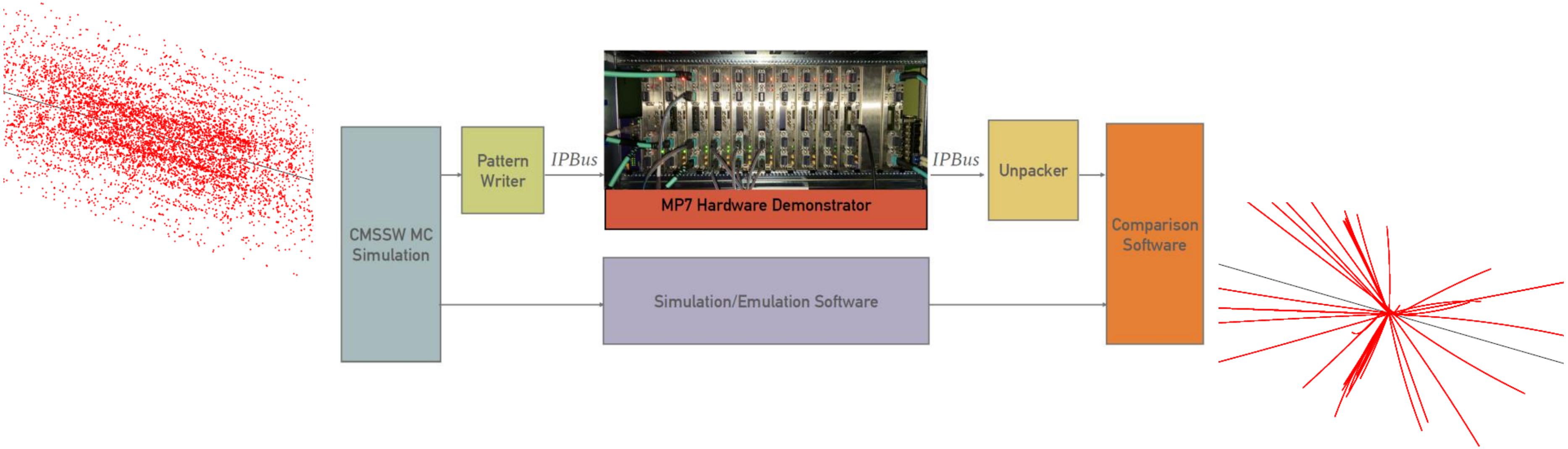
simplifies division of labour and testing (algorithms tested **individually** or **in chain**)

present-day FPGA resources not a limit to the scale/performance of system





demonstrator system



CMSSW MC Simulation

Pattern Writer

IPBus

MP7 Hardware Demonstrator

IPBus

Unpacker

Comparison Software

Simulation/Emulation Software

stub data (from simulation) **loaded into RAM buffers** in source boards and **played through demonstrator**

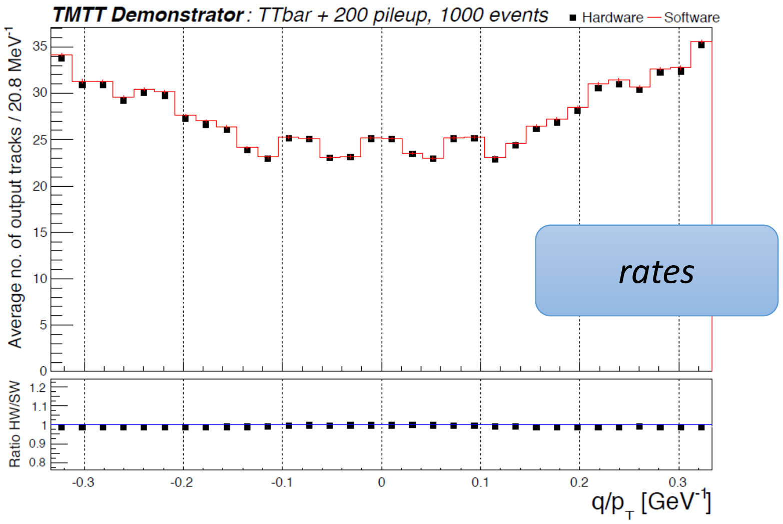
1/8 of tracker at a time, 30 events per run

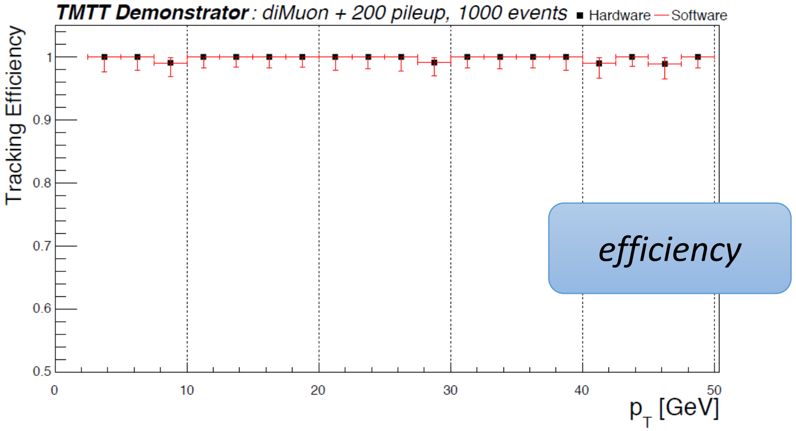
track candidates or tracks for each event are extracted from the sink

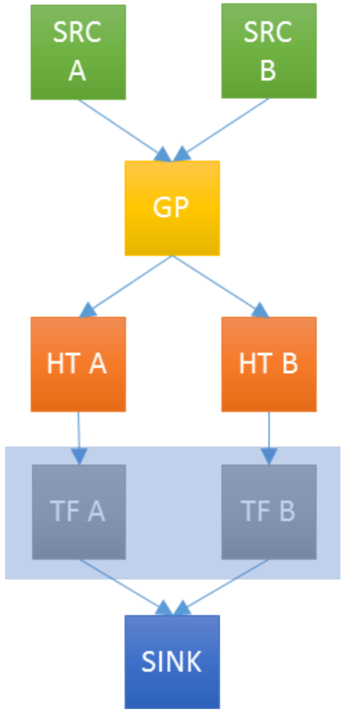
hardware output compared with C++ simulation/emulation software

track finding on full tracker is demonstrated, latency can be measured

demonstrator results to date







fitting stage not yet included

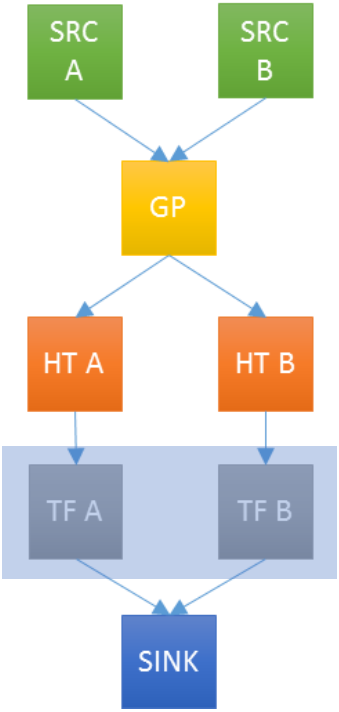
rates match expectation for worst case scenarios at output of HT

high track finding efficiency down to 3 GeV/c

demonstrator results to date

	LUTs	FFs	BRAM36	DSP48
GP	145k	273k	318	1488
HT	260k	288k	1584	126

resources



SRC -> GP	146 ns
GP -> HT	292 ns
HT -> SINK	1221 ns
TM period	900 ns

latency

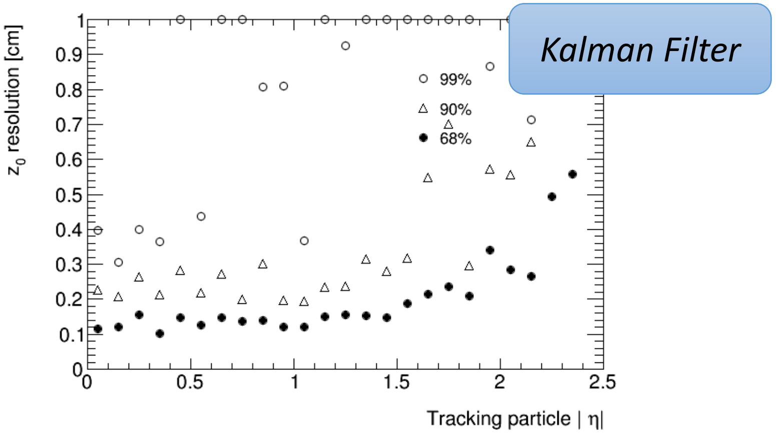
fitting stage not yet included

remaining latency budget 1.44µs

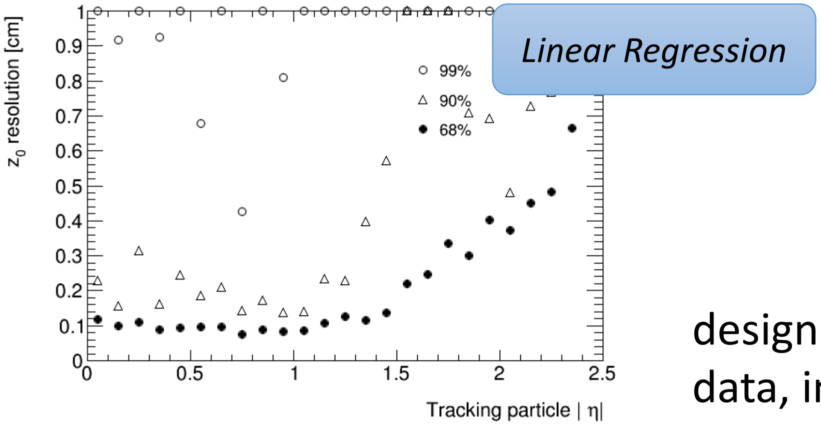
demonstrator prospects & outlook

ready for integration of track fitters

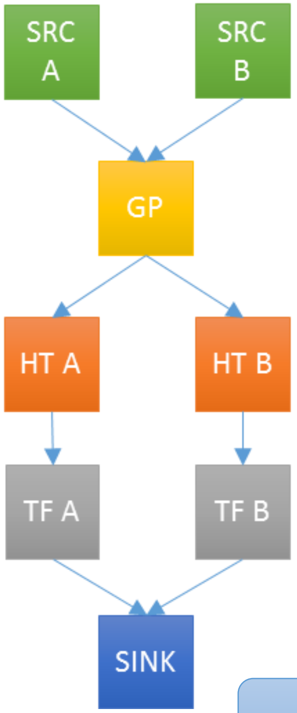
tracking performance is **close to offline**, Linear Regression fitter is performance competitive



Kalman Filter



Linear Regression



resources – first look

	LUTs	FFs	BRAM36	DSP48
Kalman	160k	270k	393	1836

design now requires dataflow validation with realistic data, including latency measurement

summary and conclusions

at HL-LHC new Level 1 (hardware) trigger must ensure tracks can be reconstructed within 5μs

Hough Transform based demonstrator using MP7 hardware in operation, reconstructing track candidates within ~1.5μs, reducing input rate by order of magnitude

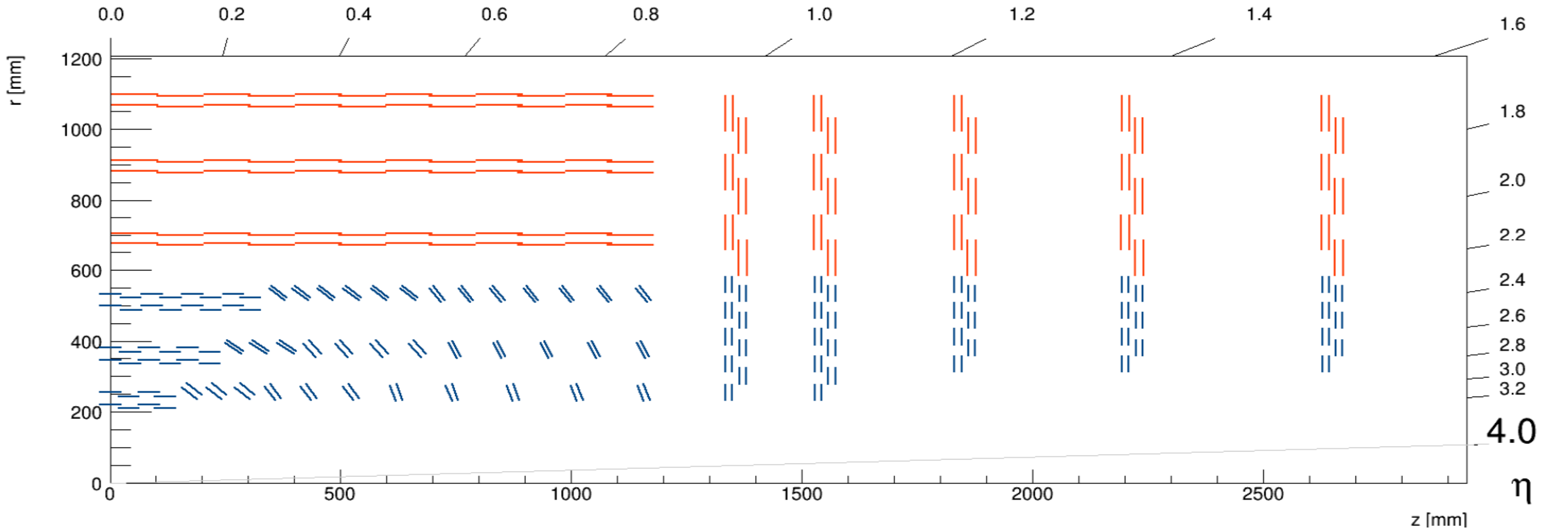
- excellent performance, efficiency and matching with simulation so far

track fitting stages currently under test and results to be extracted in next months

further performance studies to be carried out in parallel, e.g. robustness to dead modules, coping with higher rates or lower pT thresholds, alternative beam profiles, impact of tilted barrel geometry

backup

tilted barrel geometry

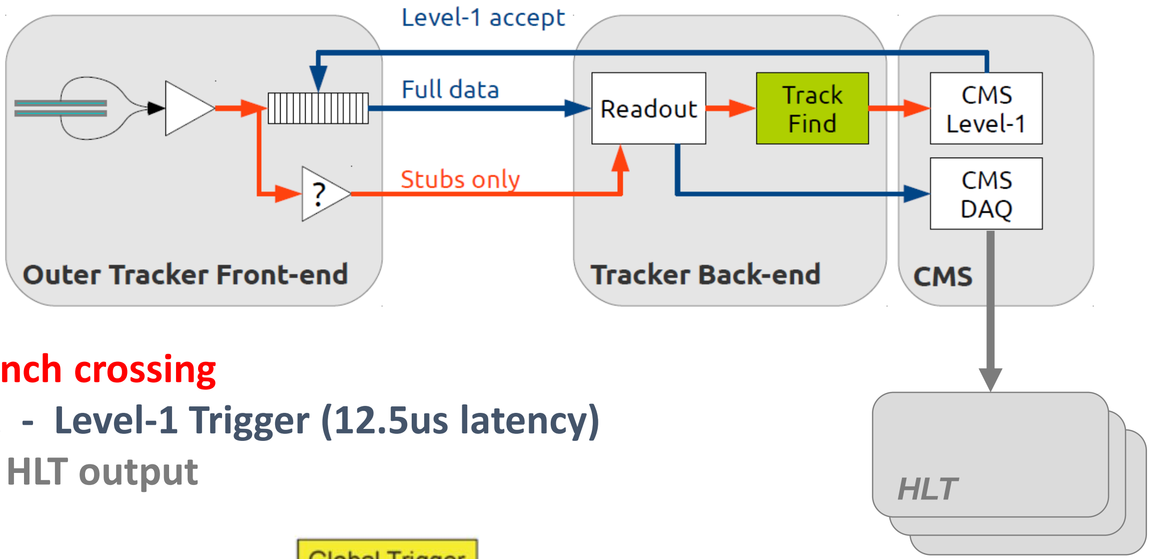


less material

fewer modules

improves trigger efficiency at high eta, and reduces #hits in inner barrel layers

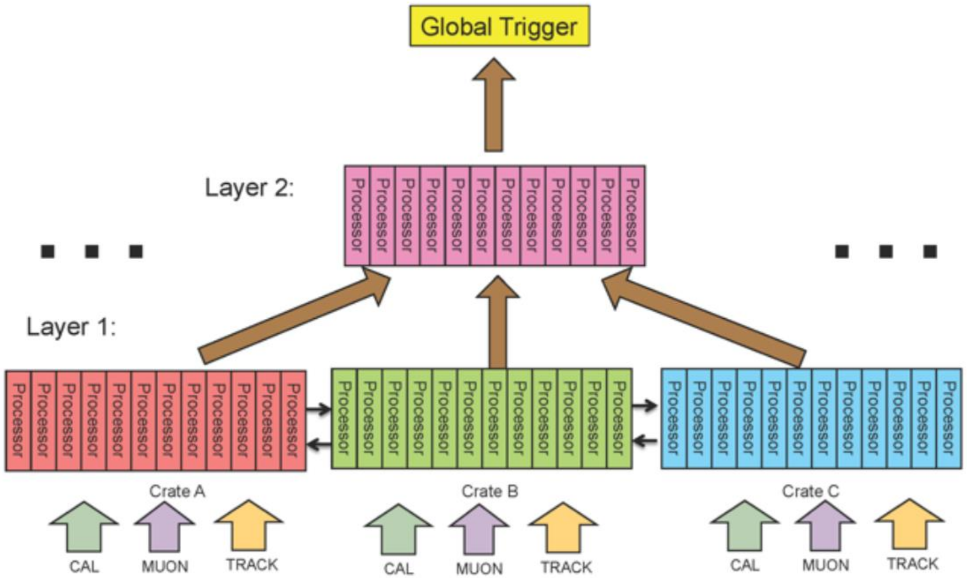
L1 and Tracker Readout



@ 40 MHz - Bunch crossing

@ 0.5-0.75 MHz - Level-1 Trigger (12.5us latency)

@ 100-500 Hz - HLT output


Global Trigger
Layer 2:
Processor
Layer 1:
Processor
Crate A
Crate B
Crate C
CAL
MUON
TRACK