

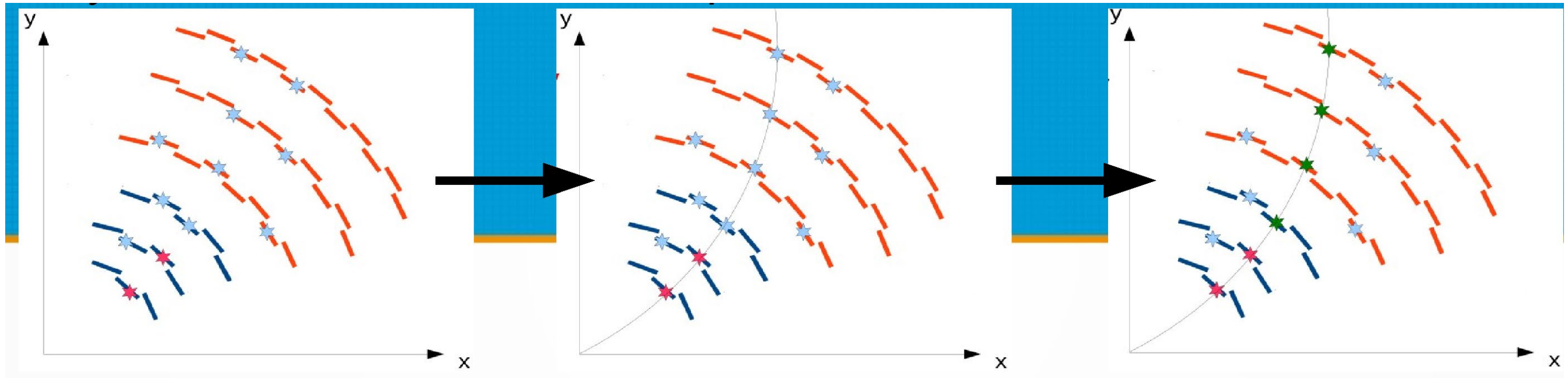
A 3D FPGA Track Segment Seeding Engine Based on the Tiny Triplet Finder

Jinyuan Wu
Fermilab
Sept. 2021

Introduction

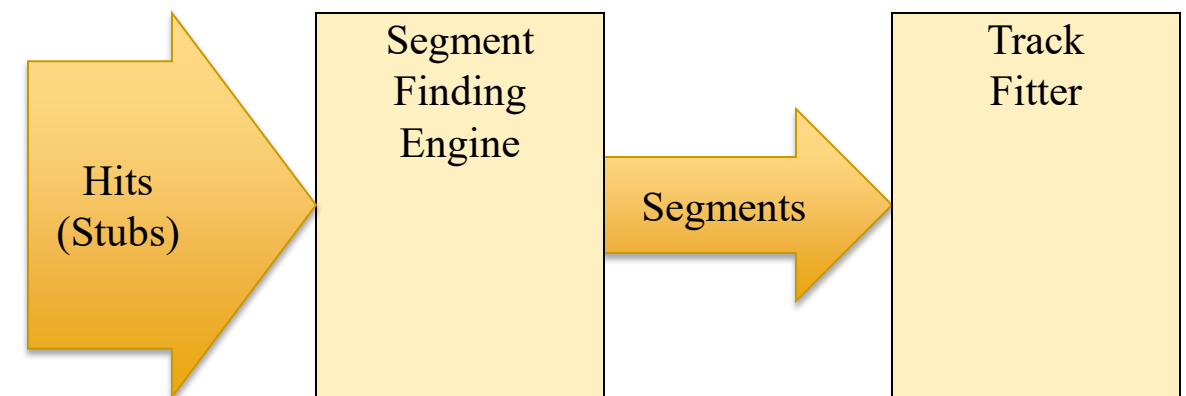
- An exercise of implementing a 3-hit 3-D track segment seeding engine based on the Tiny Triplet Finder is presented.
- It is intended to group hits into track segments and to feed into later Kalman filtering stage.
- The silicon resource consumption of the seeding engine is small, and it can be put into a low-cost FPGA.
- *In our work, we use CMS L1 tracking Hybrid Algorithm as the example because it is challenging. But this work is not in official CMS project*

Tracking in Multilayer Detectors

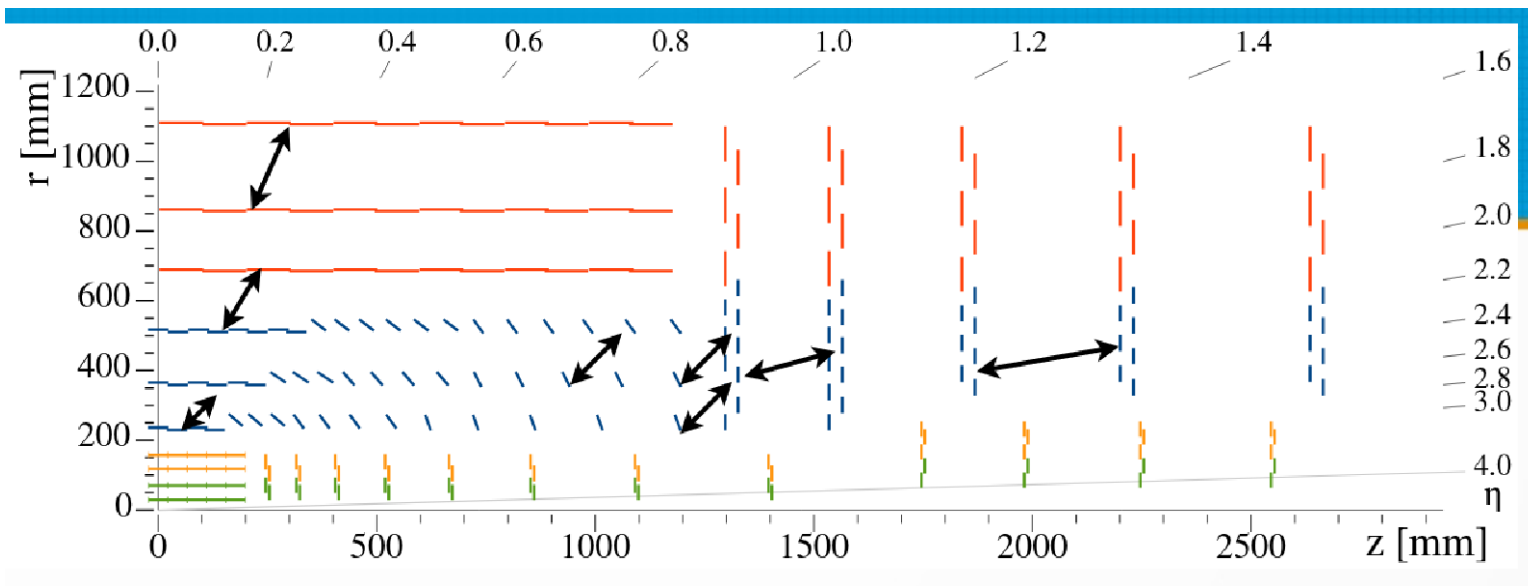


Andrew Hart et. al. 2020

- Well accepted tracking approaches take the following stages:
 - Hits or stubs are first grouped into track segments.
 - Track segments are fitted into tracks.

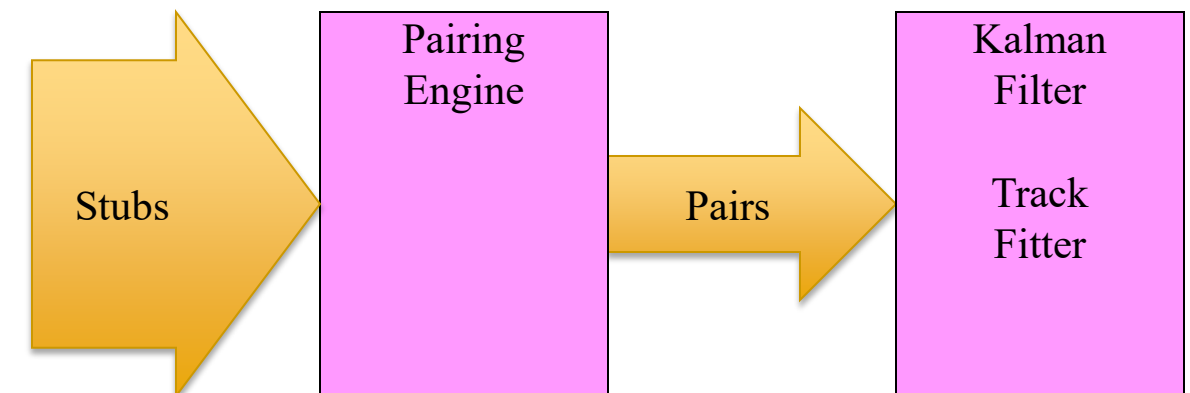
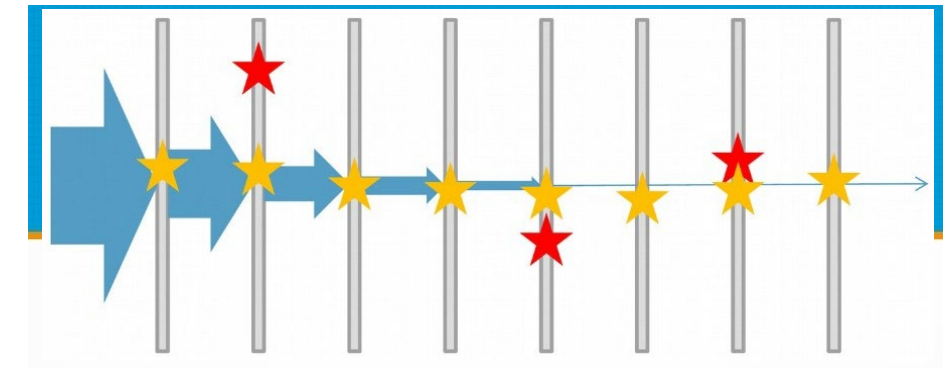


The CMS L1 Tracking Hybrid Algorithm



Andrew Hart et. al. 2020

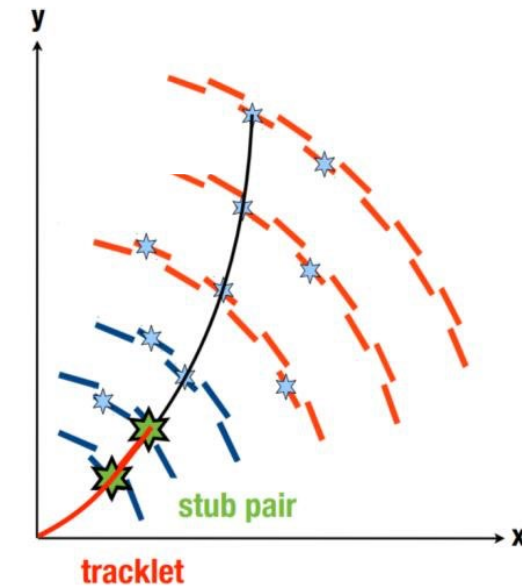
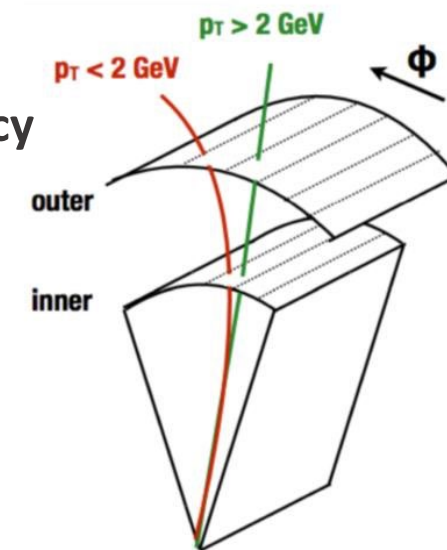
- Hits from different layers are first paired
- Track segments with paired hits are extended.
- Hits are fitted into tracks.



Pairing

Stub Organisation & Seeding

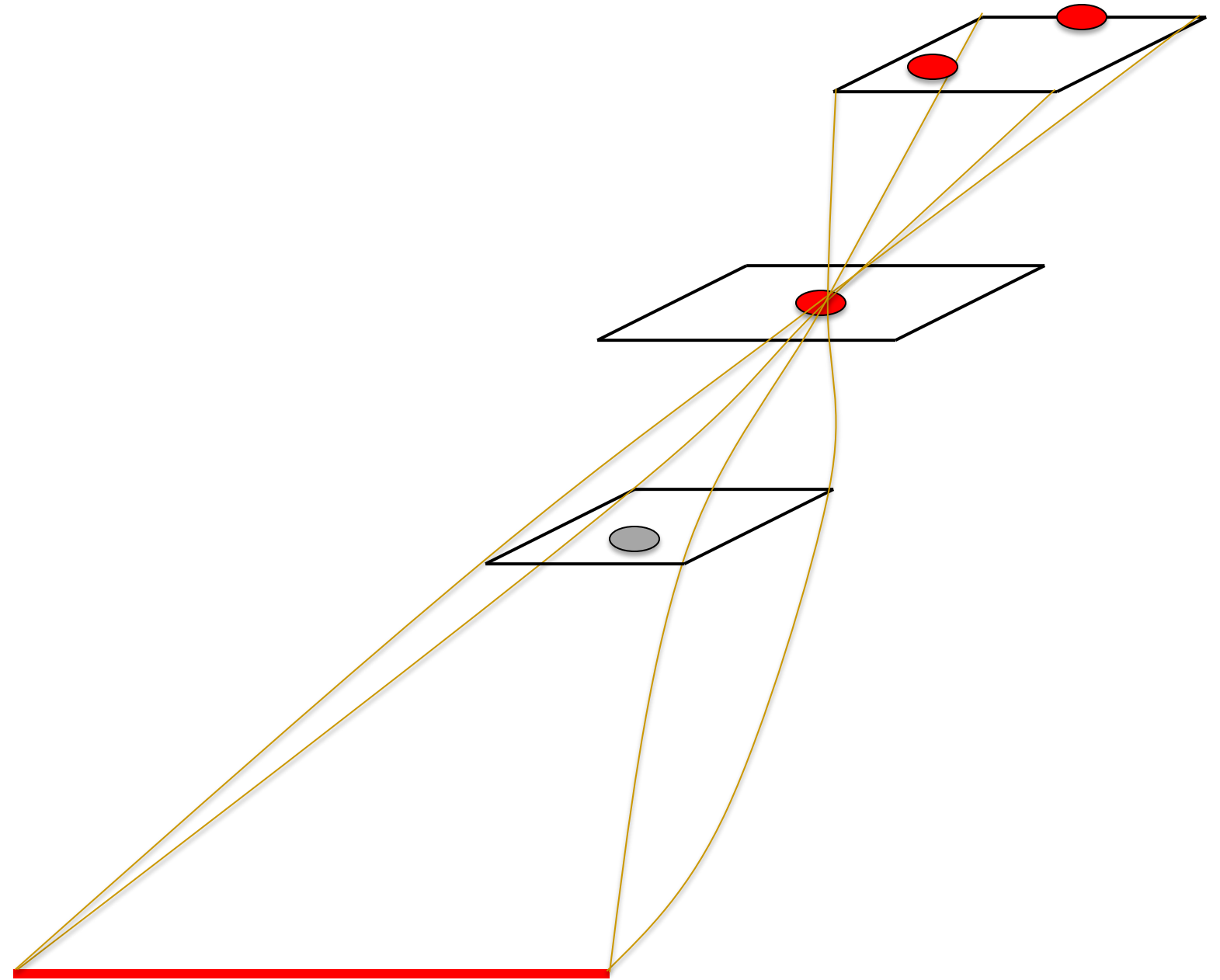
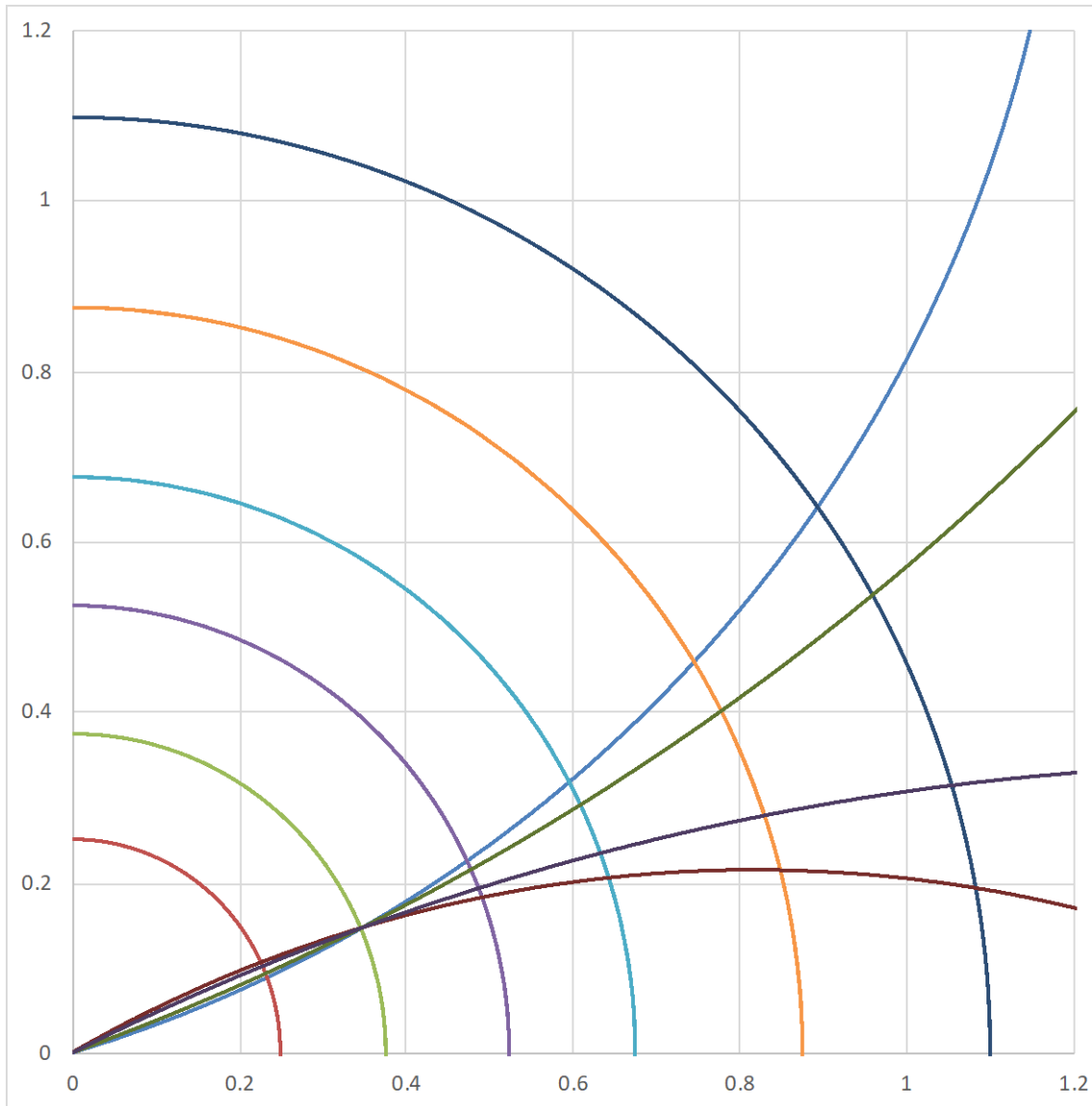
- Seeds (tracklets) are generated from **pairs of layers**
- Layer pairs selected to give full coverage in η , including **redundancy**
- Seeding step **massively parallelised** by internal geometrical partitioning & stub organisation (Virtual Modules - VMs)



Mark Pesaresi et. al. ACES 2020

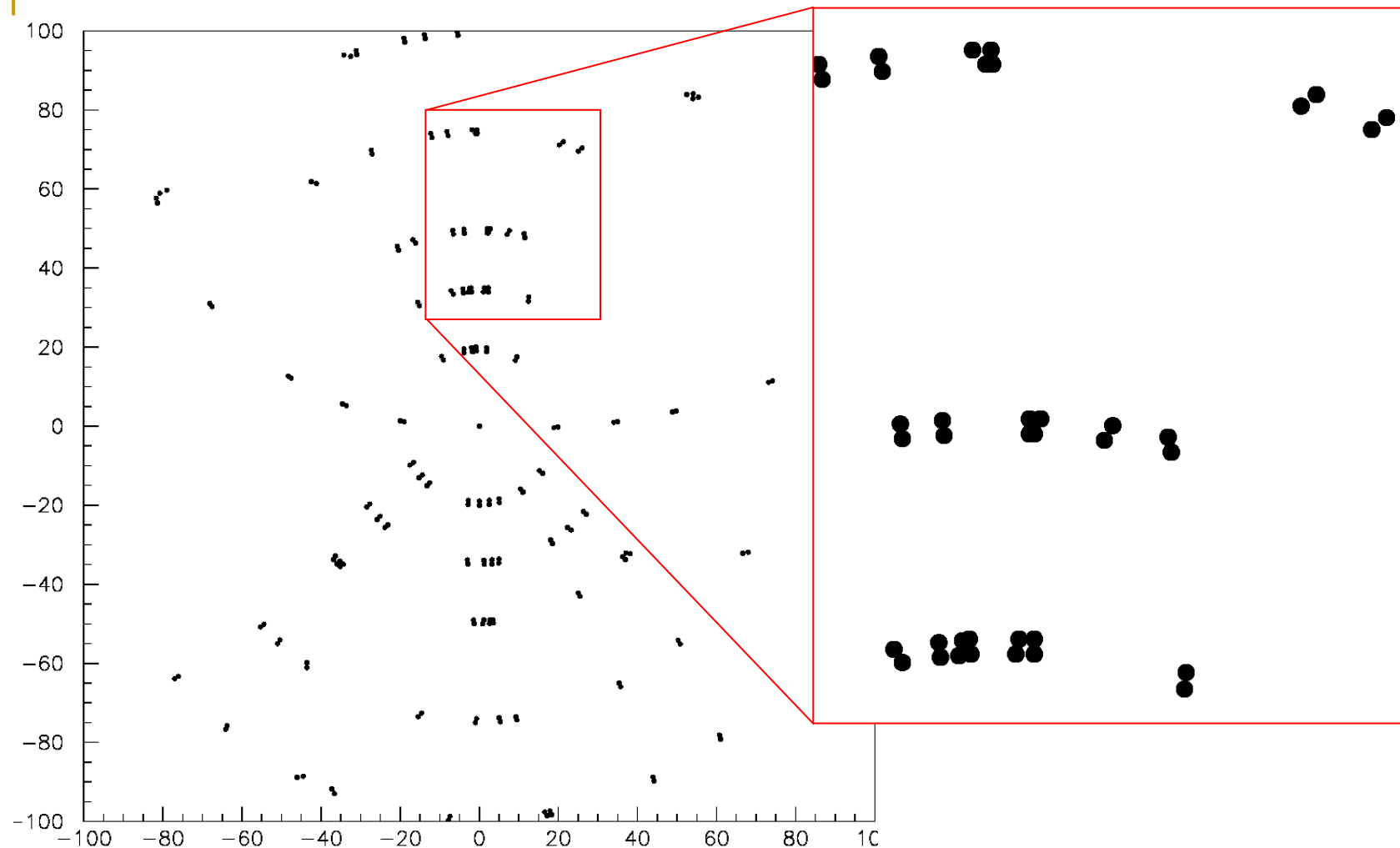
- The CMS L1 Tracking Hybrid Algorithm: find a **pair** => find third => fit and find others

Pairing Stubs



- For any stub in Layer 2, a range in ϕ and z must be searched in Layer 1 or/and Layer 3.
- Fake pairs may occur.

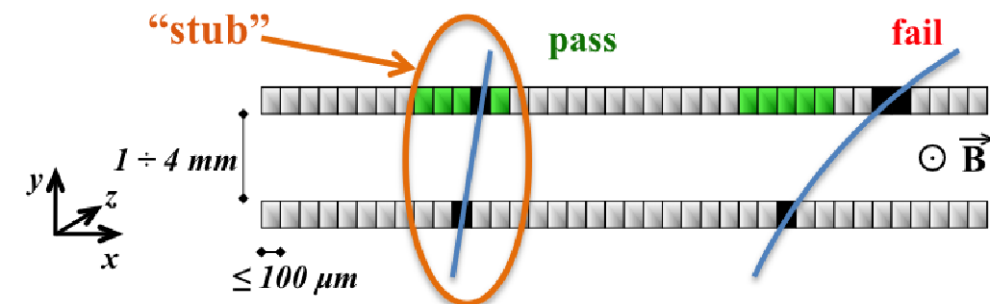
Pairing or Triplet Finding



- If the separation of silicon sensor layers were large (~ 10 mm), then stub **pairing** would be sufficient.
- Given the separation of ~ 4 mm in PS and 2S module, correlating **triplets** of stubs becomes necessary.

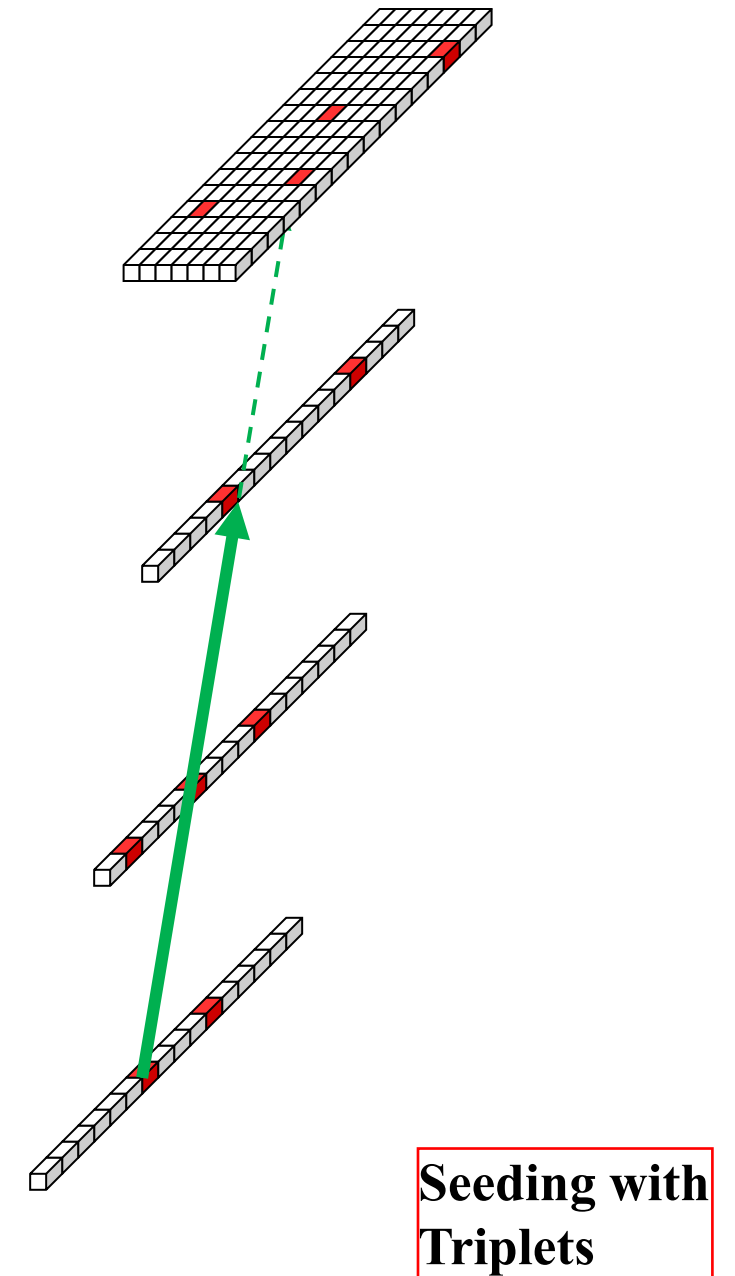
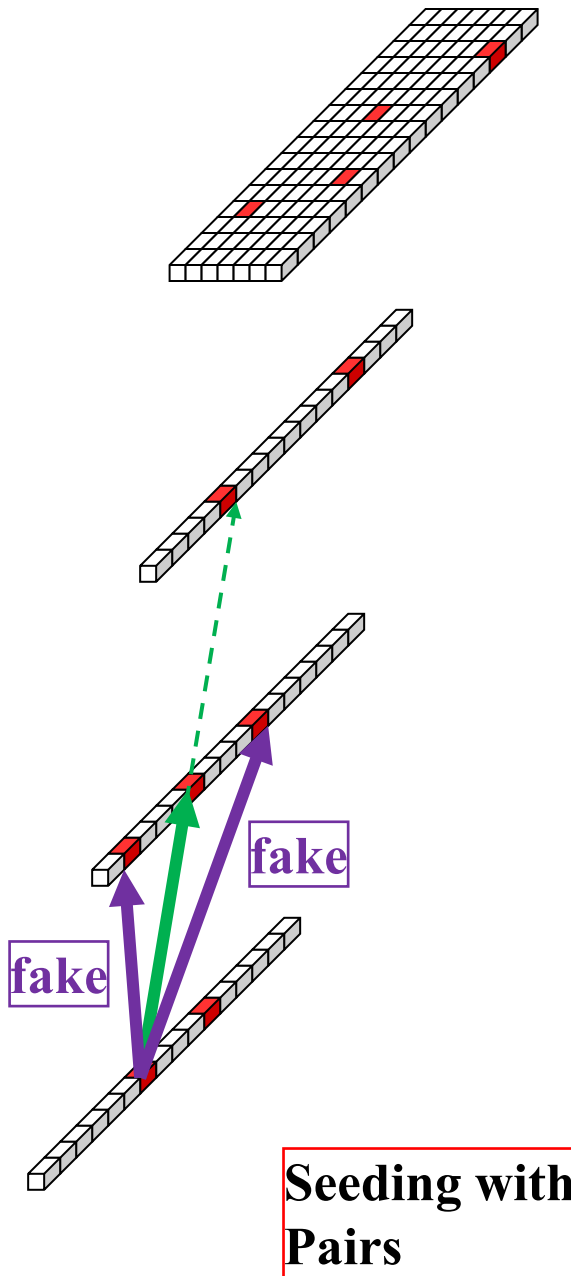
Wu et. al. 2004

https://indico.cern.ch/event/418639/contributions/1018451/attachments/868833/1216631/lowpt_lecc2004p.pdf



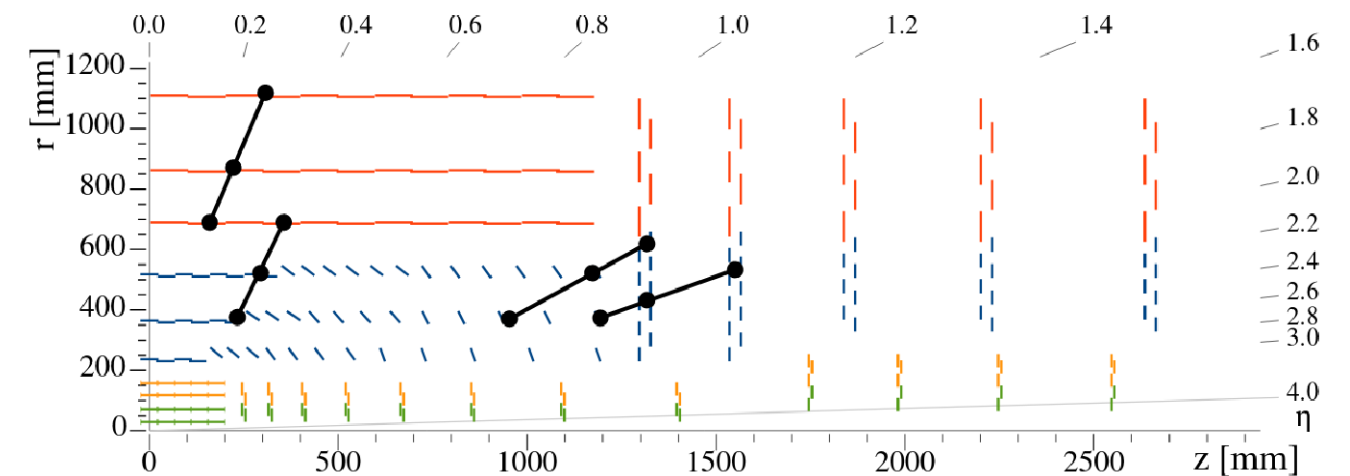
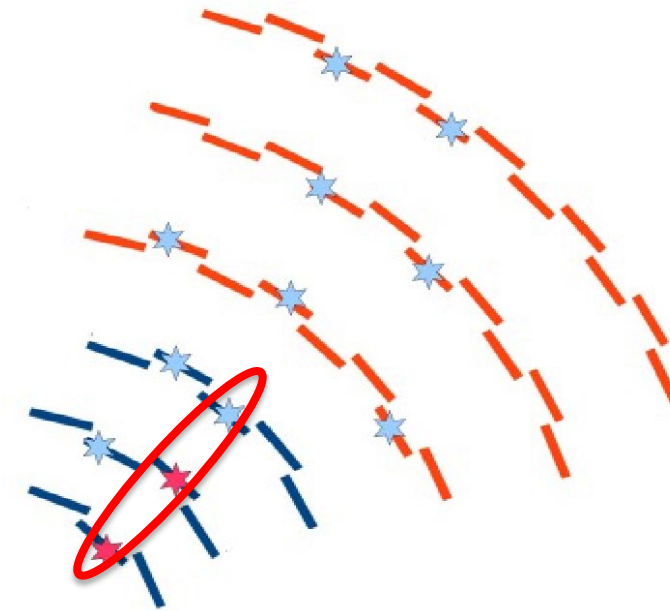
From Pairing to Triplet Seeding

- Simple pairing will create more fake hit pairs, which must be eliminated in later stages.
- When three or more hits (stubs) are picked up together in triplet seeding stage, less fake segments are expected.

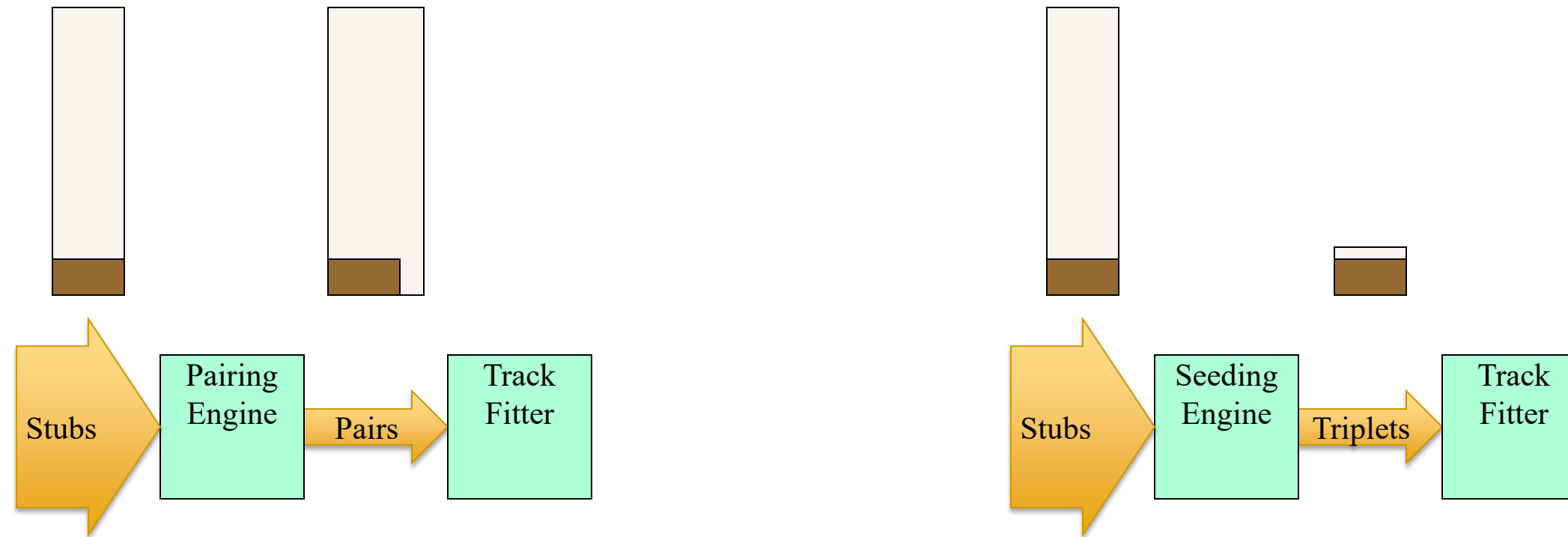


Triplet Finding

- If three hits can be correlated together in a single action, would that be useful in the Hybrid Algorithm?
- It seems that the triplet finding is more natural for:
 - Displaced track finding
 - Beam-spot constraint tracking



Triplet Seeding

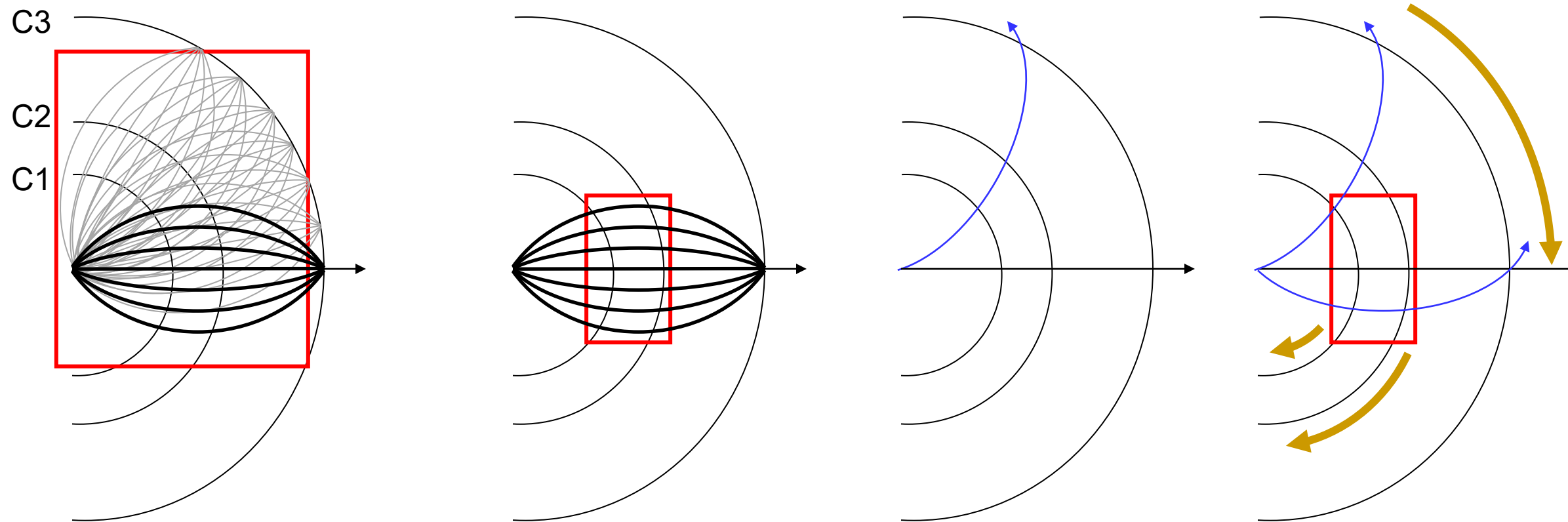


- Stub pairing won't eliminate stubs not associated with a high Pt track.
- Using three layers or more to most eliminate fake stubs in the seeding engine.

Triplet seeding is good.
The algorithm is straightforward.
The challenge is how to fit it into
FPGA.

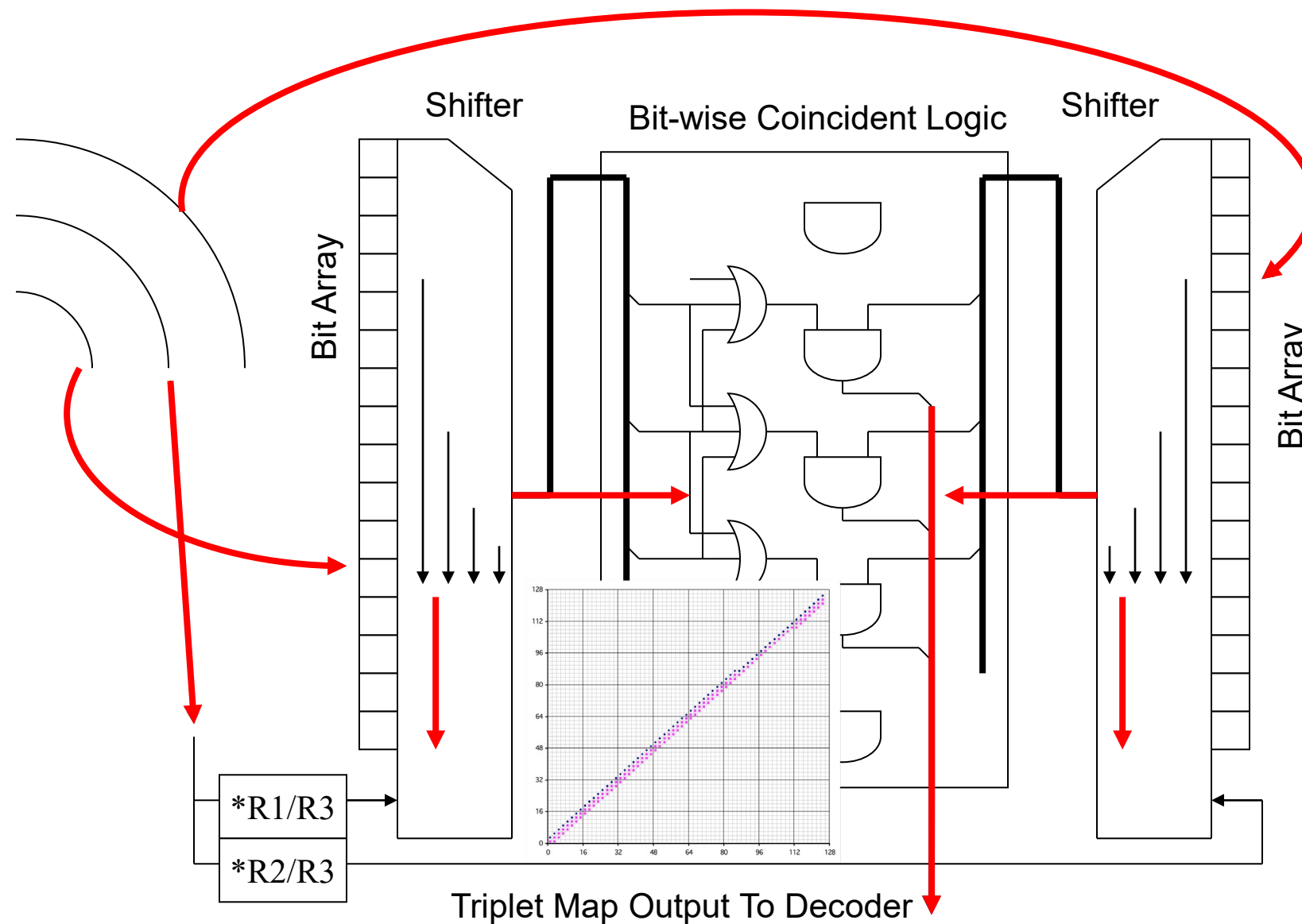
Tiny Triplet Finder:
-- A low resource usage scheme

The Tiny Triplet Finder for Cylindrical Detector Geometry



- The Tiny Triplet Finder is a scheme to reuse a single set of track road coincidence logic for whole detector.
- It uses shift (or rotate) to bring bit arrays to feed the track road coincidence logic.

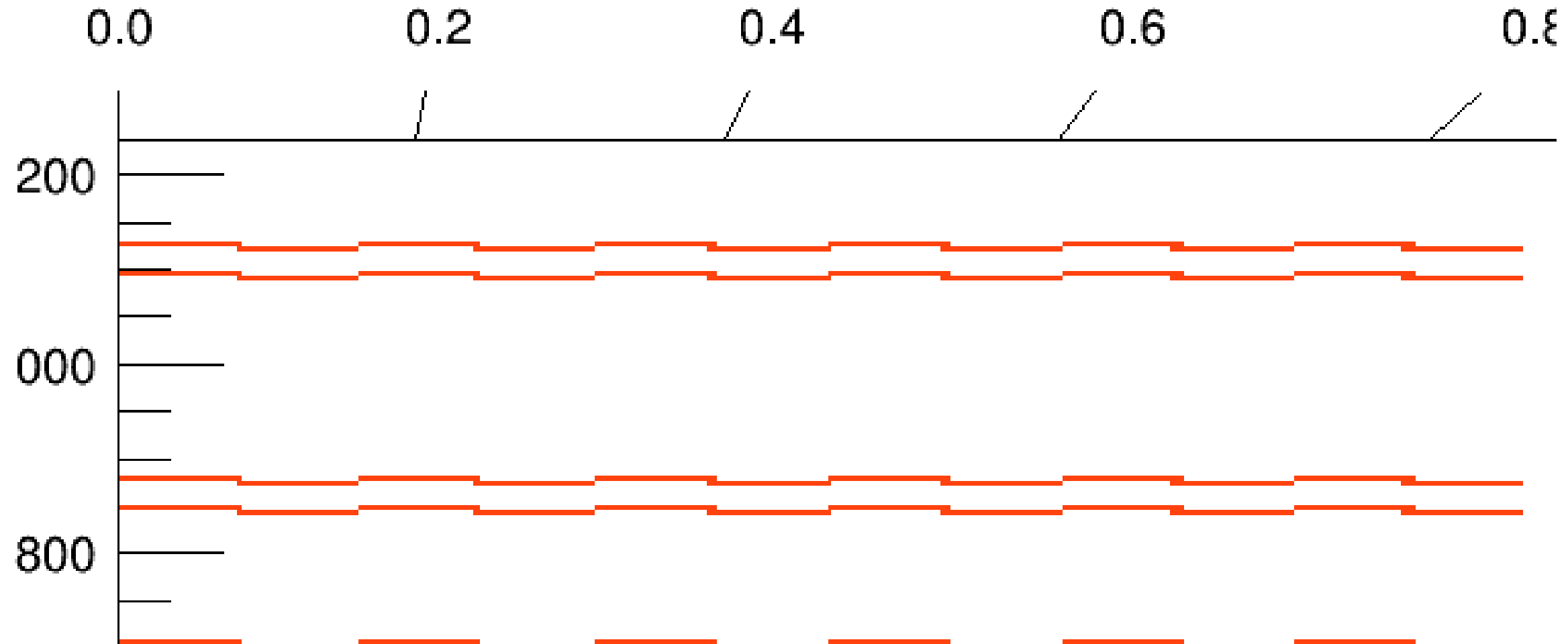
General Structure of the Tiny Triplet Finder



- A typical Tiny Triplet Finder includes:
 - Bit arrays for different layers of the detector.
 - Shifters
 - Bitwise coincidence logic.

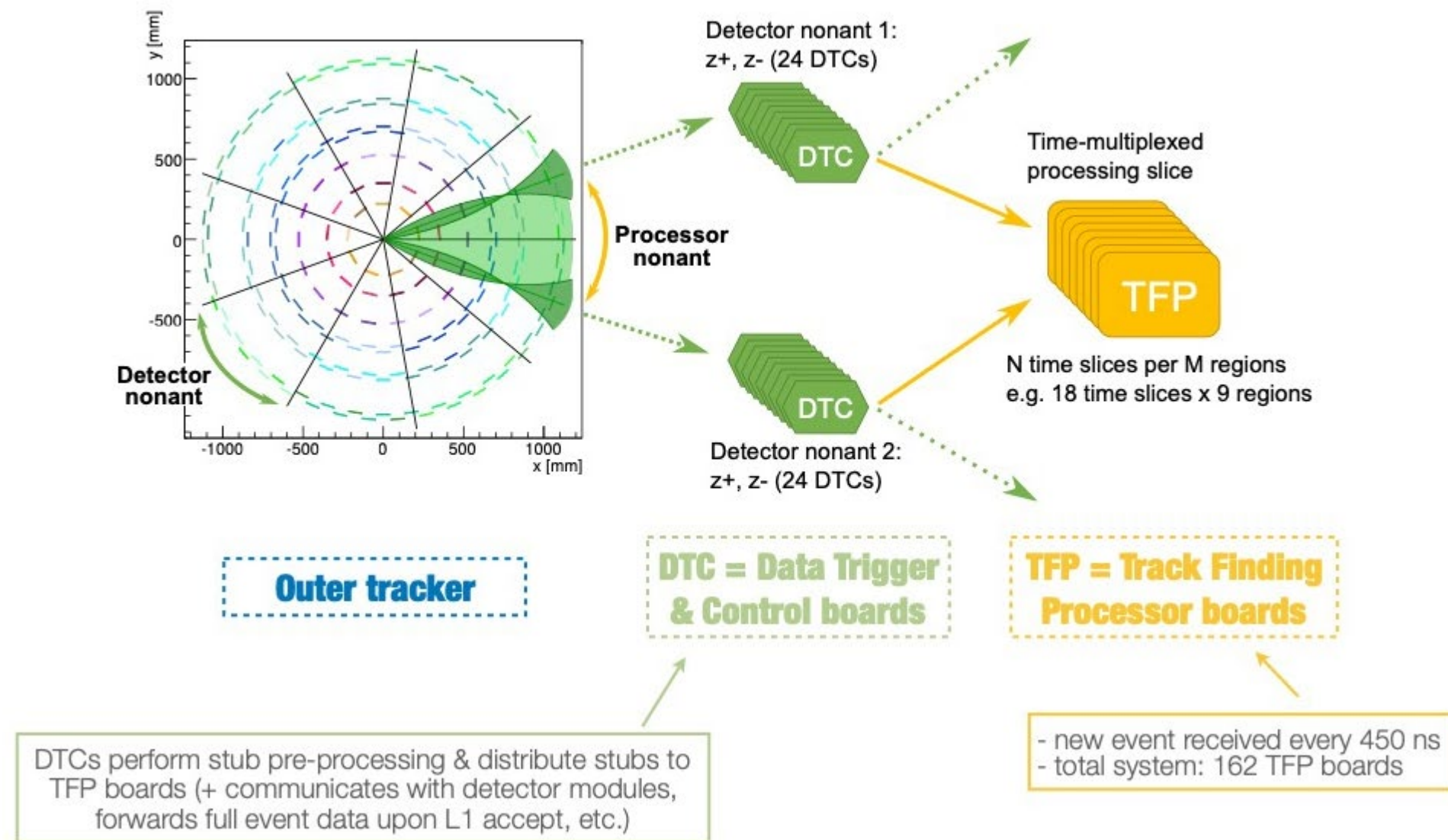
An Implementation Exercise

An Implementation Exercise Using CMS L1 Tracking as Example



- The detector in the exercise have the same radii and lengths as CMS Outer Tracker

System Division

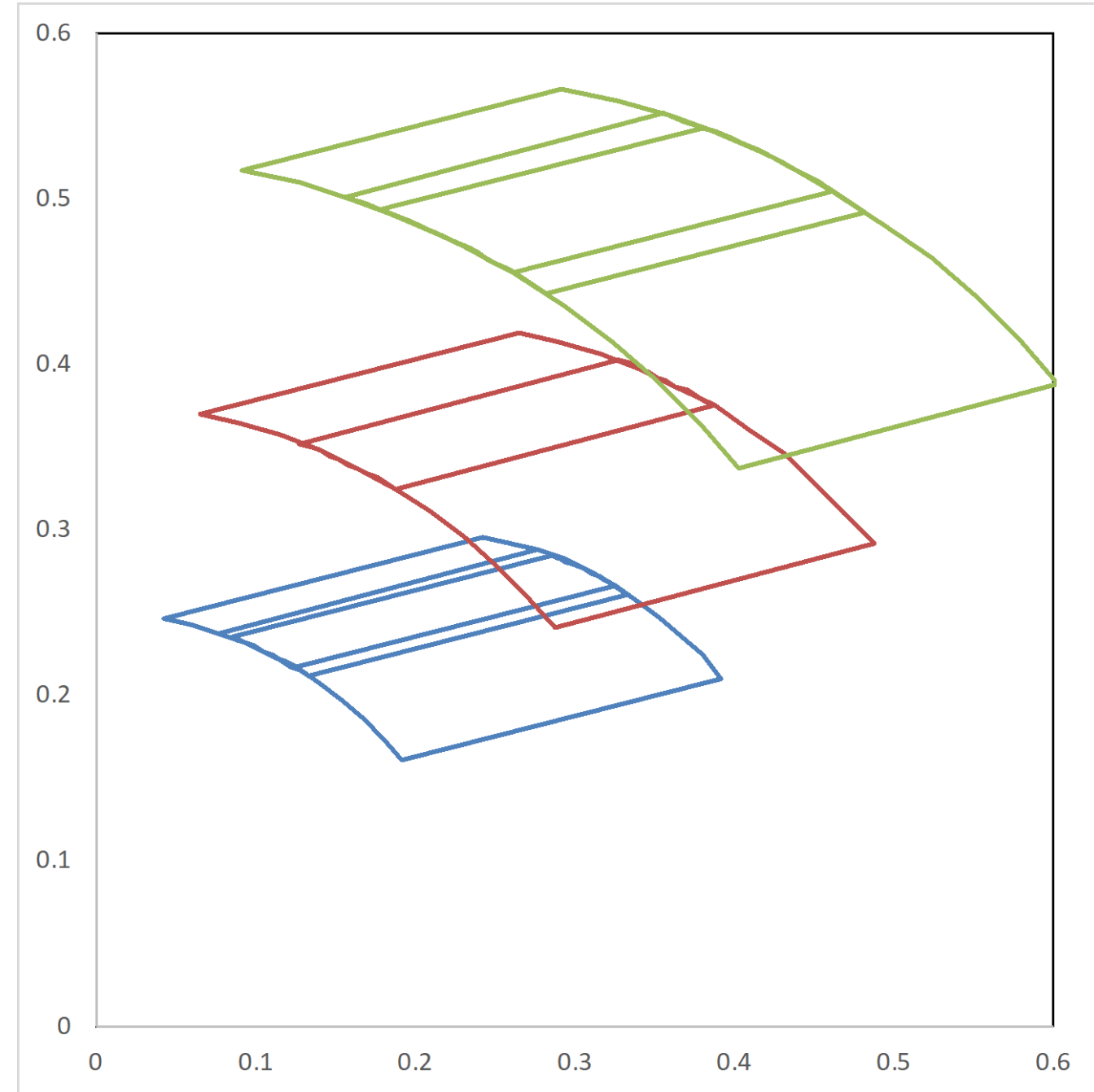


- **System Division:**
 - 18 time slices
 - 9 regions
- **Each TFP serves:**
 - 40 degrees.
 - $25 \text{ ns} \times 18 = 450 \text{ ns}$ or 225 clock cycles @500 MHz

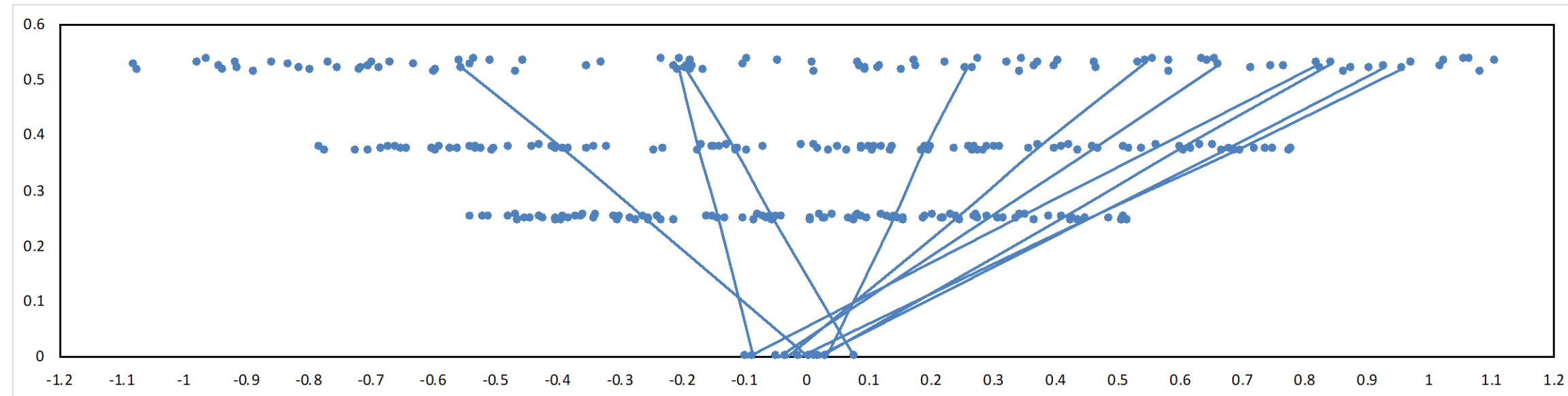
Louise Skinnari et. al. 2020

Serving Range of Seeding Engine

- In each TFP, 4 seeding engines are implemented, each serving 2.4 m (full length) x 10 degrees plus overlap.
- Each engine sees <100 hits/layer per event.
- The layer 2 is divided with exact 10-degree boundaries.
- Overlapping in:
 - Layer 1: ± 2.16 degrees, 14.3 degrees total
 - Layer 3: ± 2.62 degrees, 15.2 degrees total

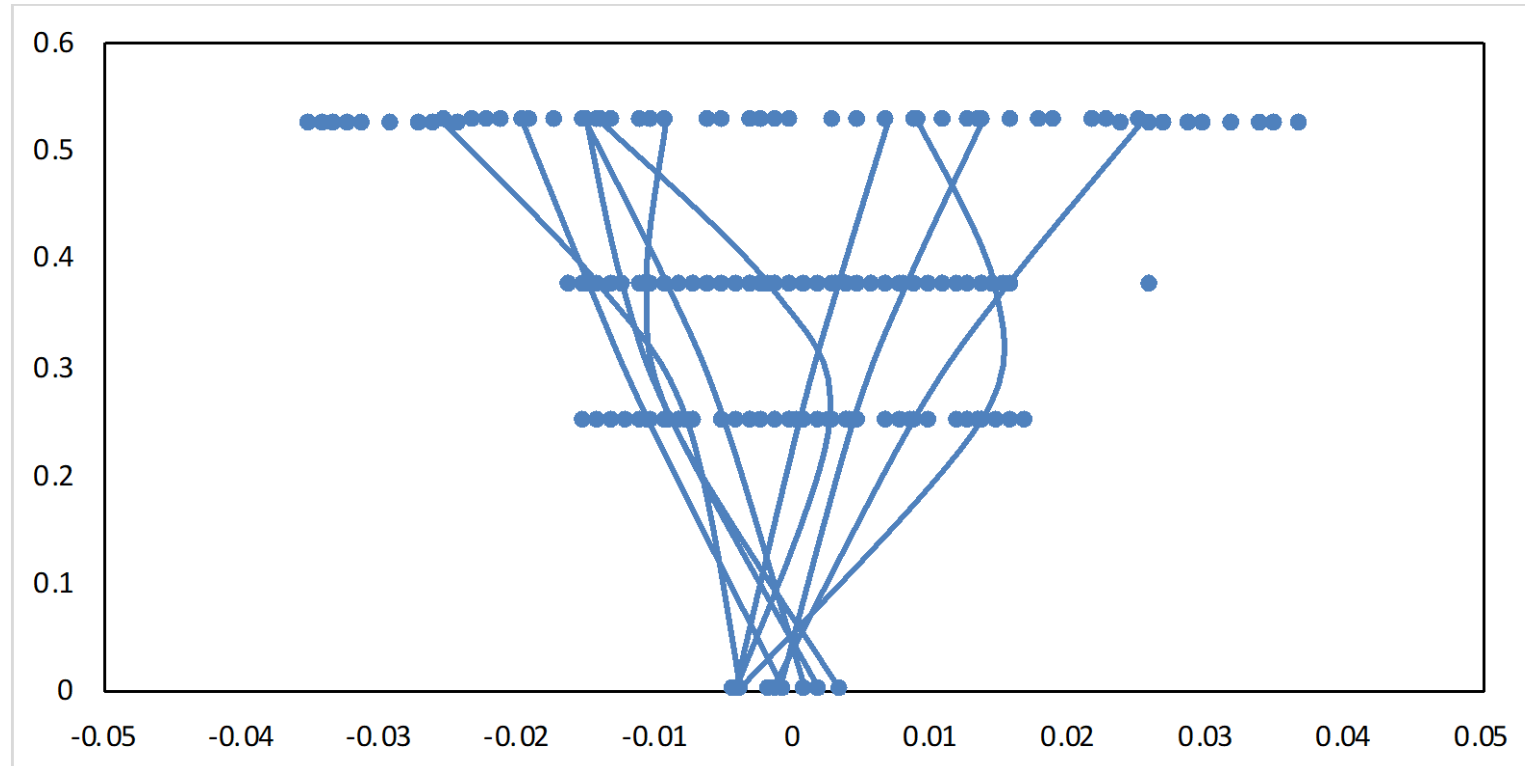
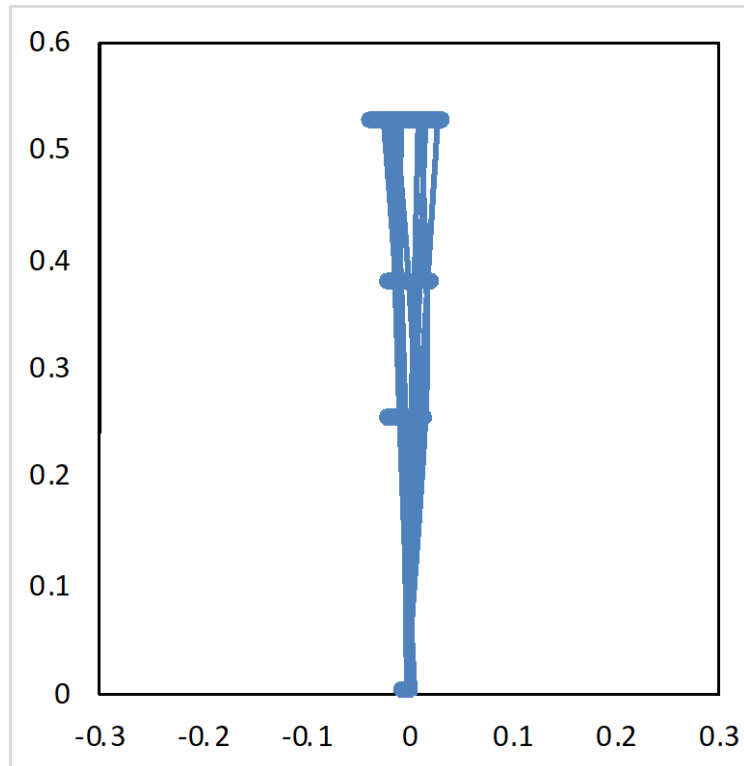


Generated Events: r-z View



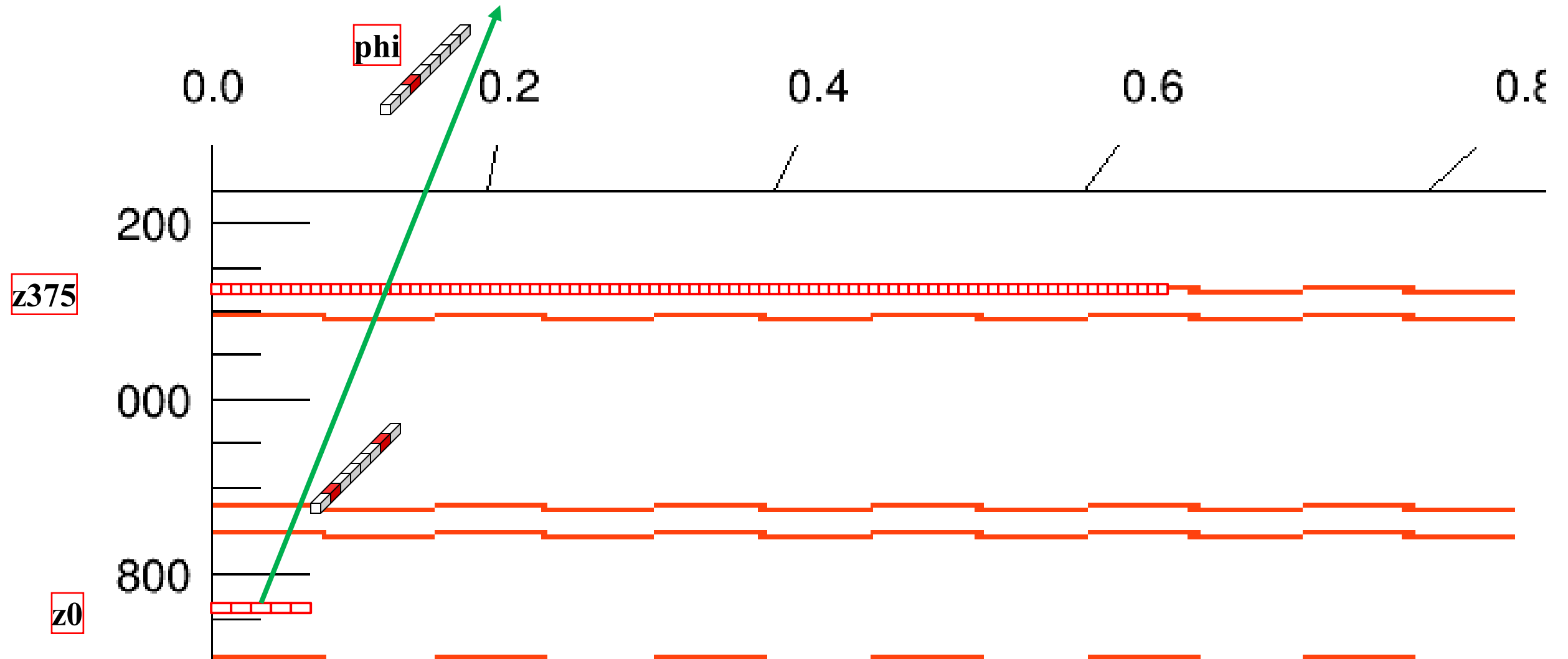
- In 240 cm (z) x 10 degrees (phi) range, up to 112 hits per layer are generated, out of which, 10 are from good tracks.
- The z_0 of the vertices spread in ± 10 cm along beam axis.

The Events: r-phi View



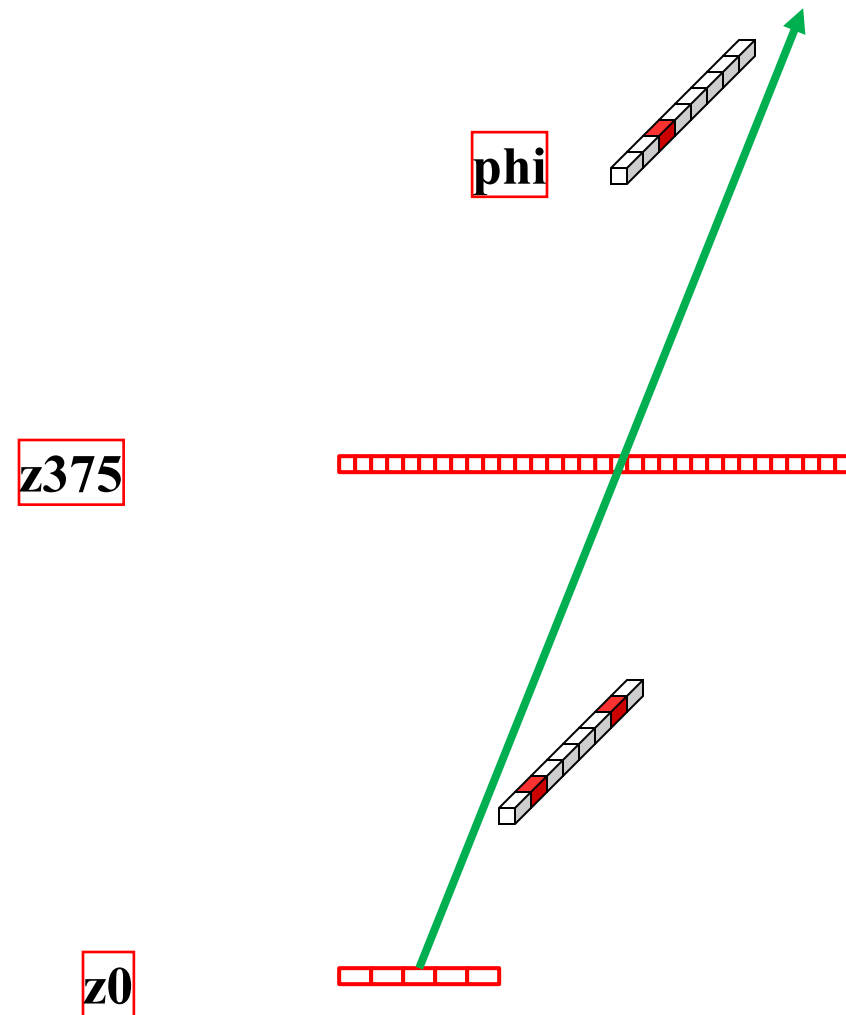
- Minimum PT: 2GeV/c.
- The hits on Layer 2 is limited to 10 degrees, but hits on Layer 1 and 3 spread up to 16 degrees.
- The impact parameters of the tracks: ± 2 mm.

The 3D Seeding Engine for SP Layers



- The parameters chosen in r-z plane are z_0 and z_{375} (z_{375} close to z of Layer 2).
- For any stub in Layer 2, the phi pattern in Layer 1 and Layer 2 with matching (z_0, z_{375}) are checked.
- The triplets matching not only in r-z, but also in r-phi are reported for further track fitting.

Bin Numbers

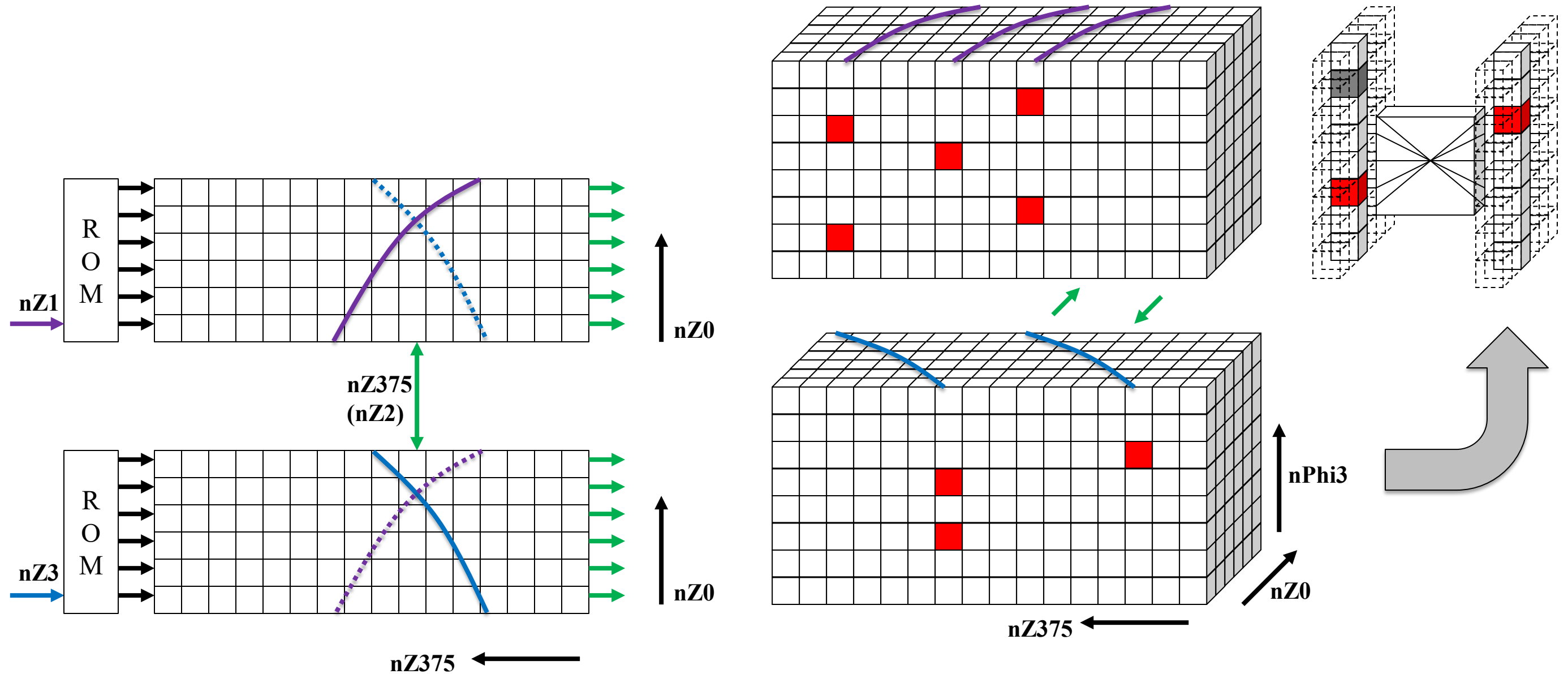


	Width	Number
z0	2 cm	10
z375	1 cm	240
phi	0.125 degrees (0.55 mm at 25 cm)	128 (/16 degrees)

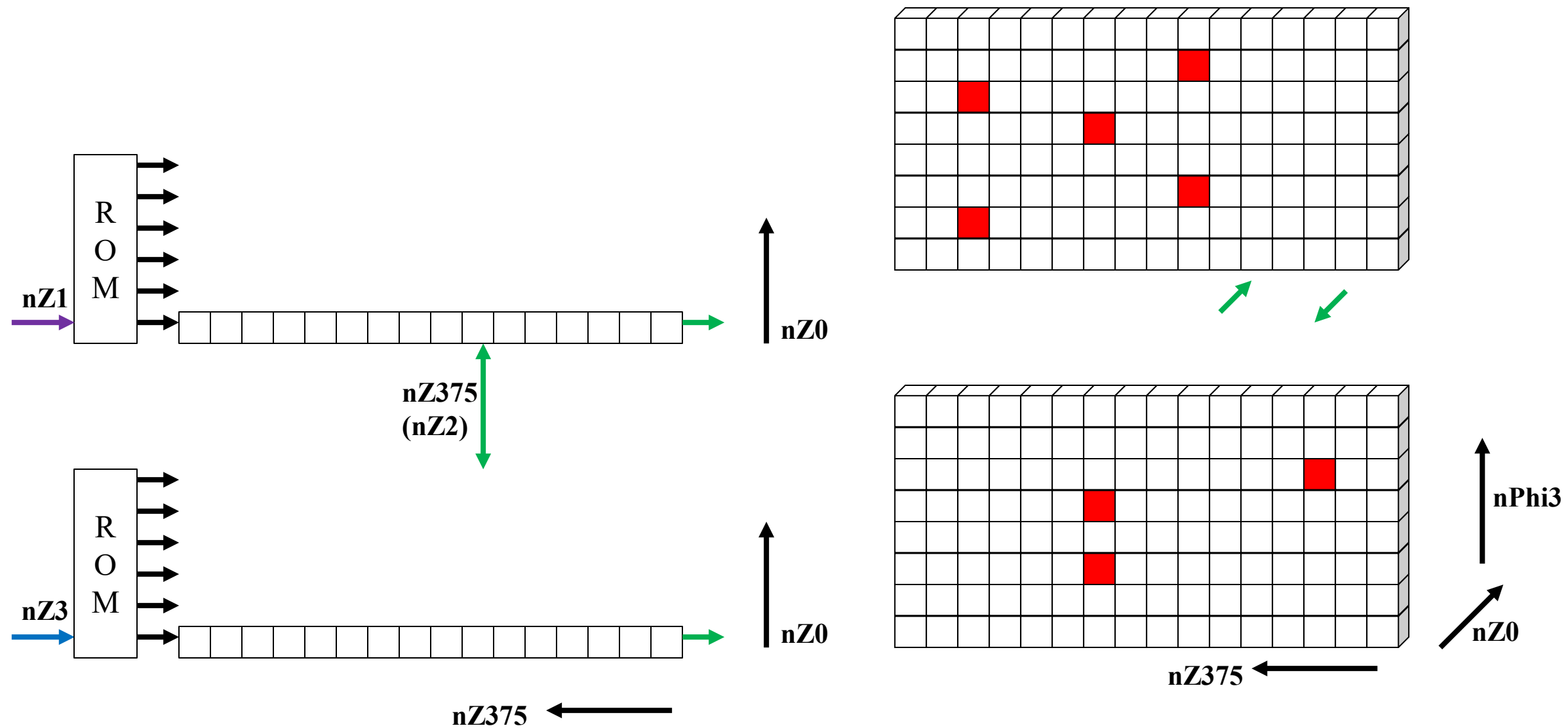
- Neither phi bins nor z bins can be too fine.
- Just checking r-phi or r-z would have too high fake rate.
- The 3-D track seeding (i.e., checking both r-phi and r-z) should provide sufficient fake stub rejection.

The Track Seeding Engine

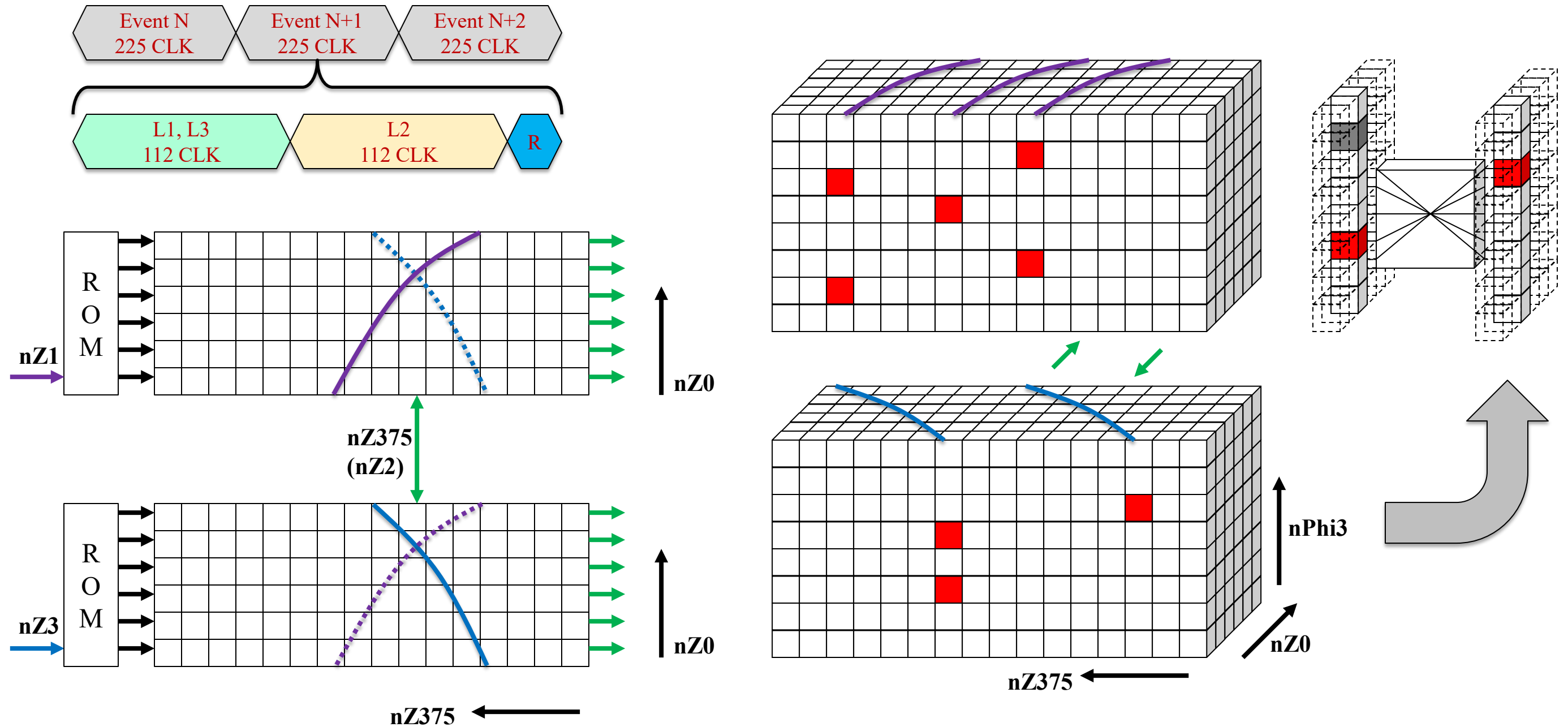
Structure of the Seeding Engine: Two RAM Banks



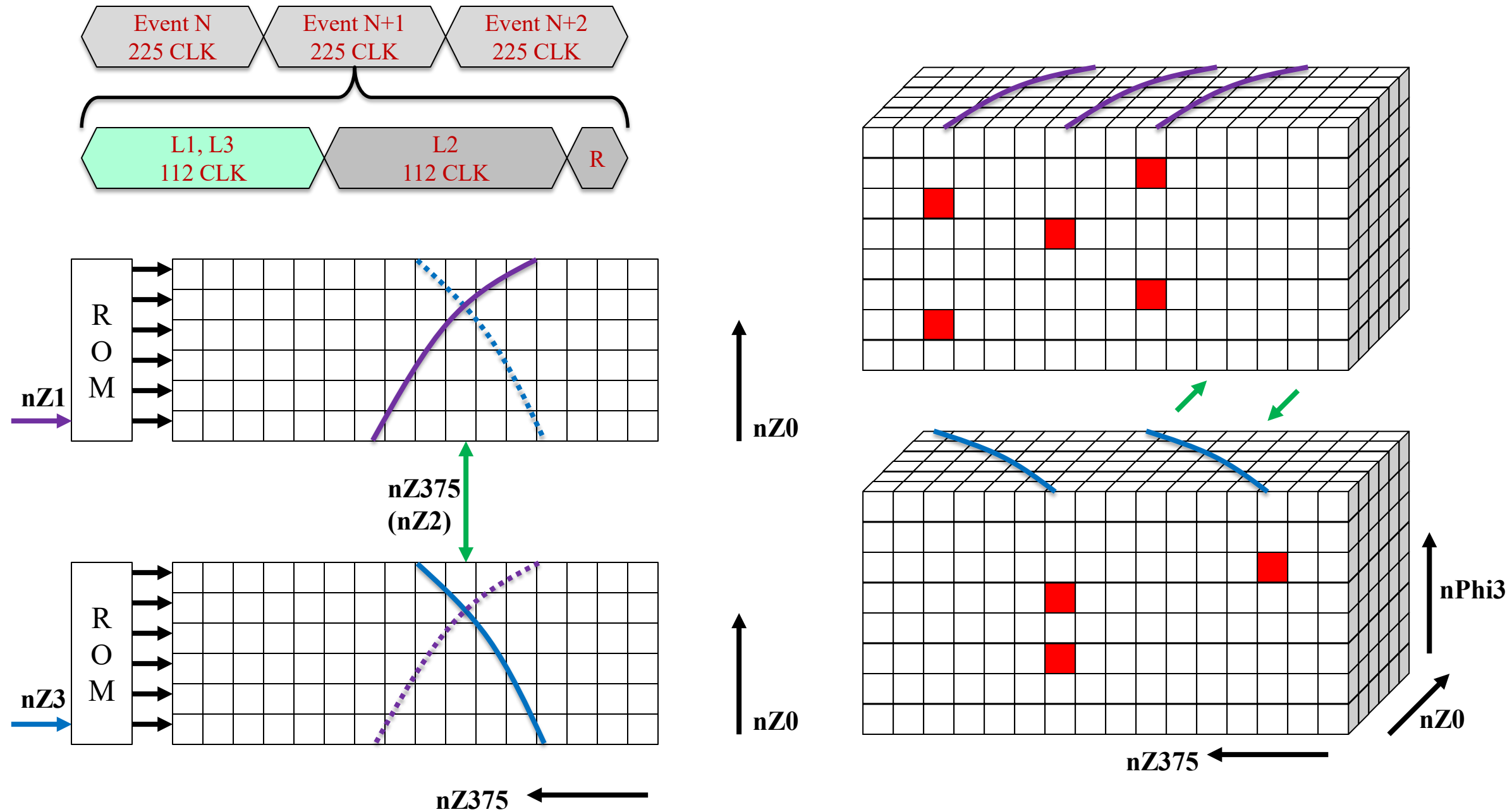
Structure of the Seeding Engine: Single Physical RAM View



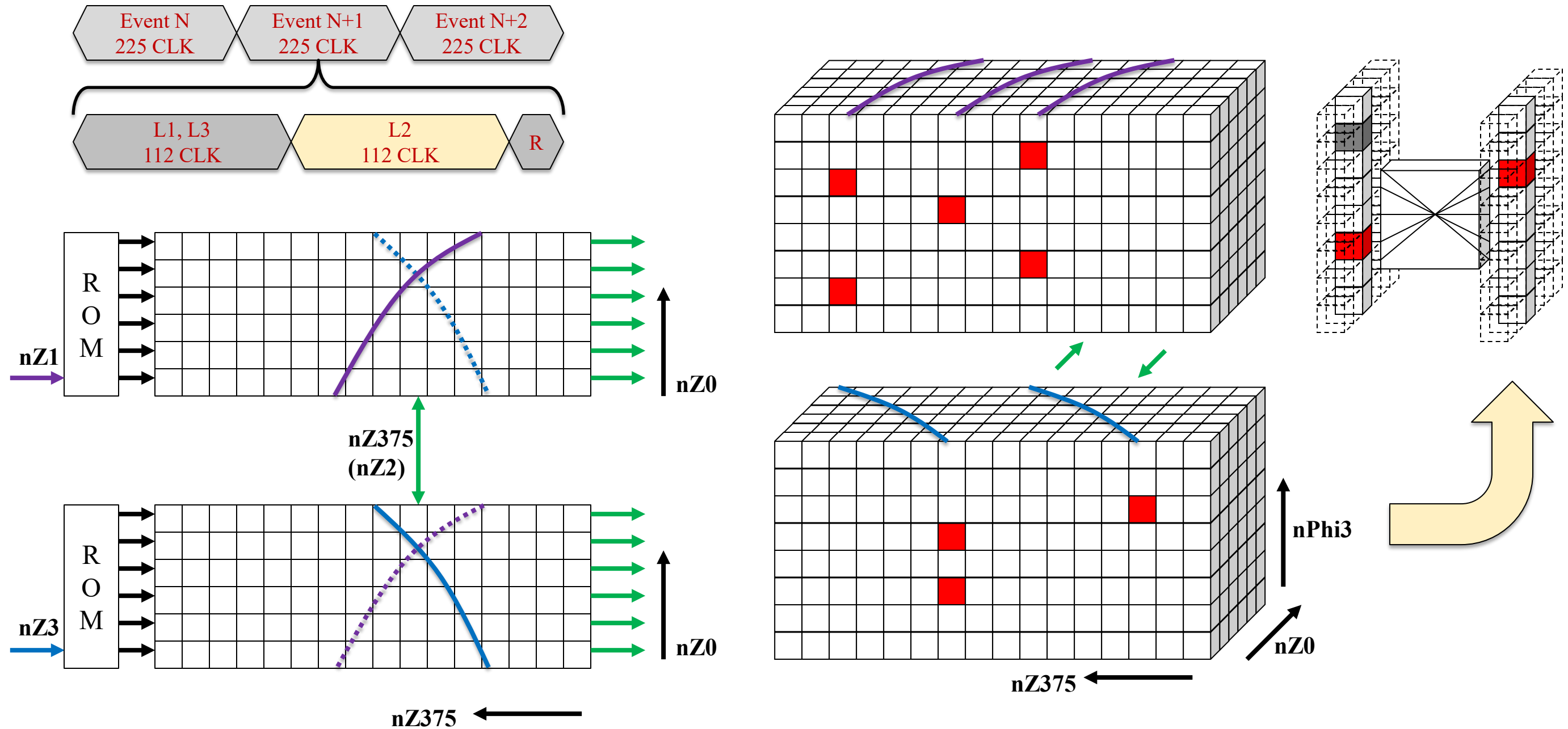
General Timing



Operation: (1) Filling Hits

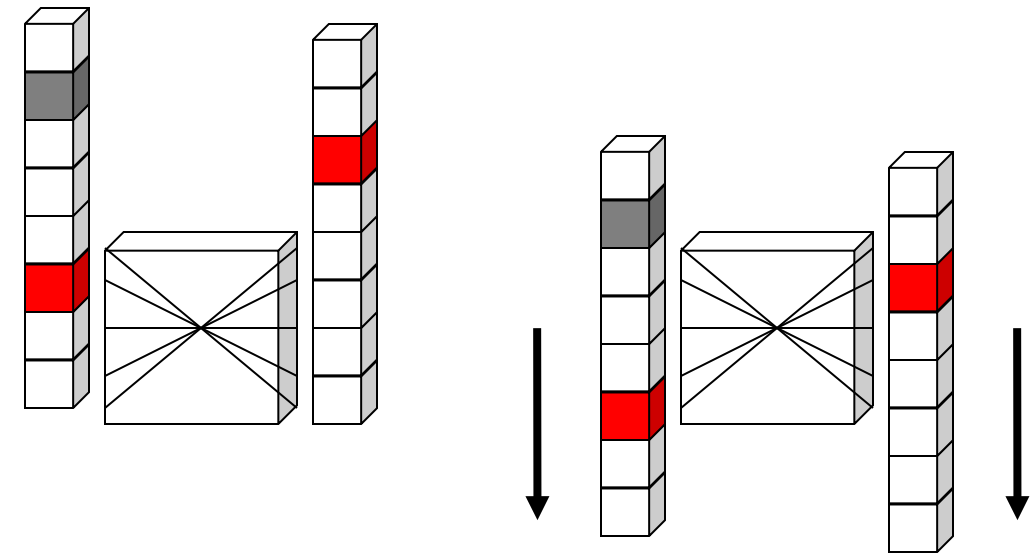
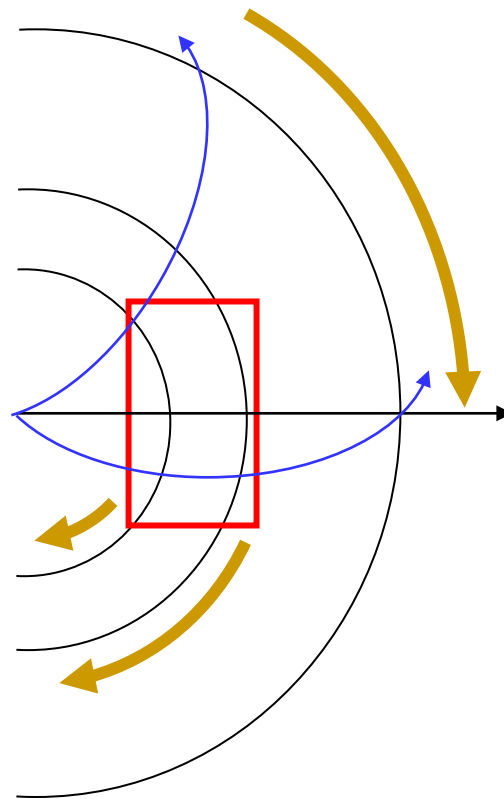


Operation: (2) Checking Coincidence

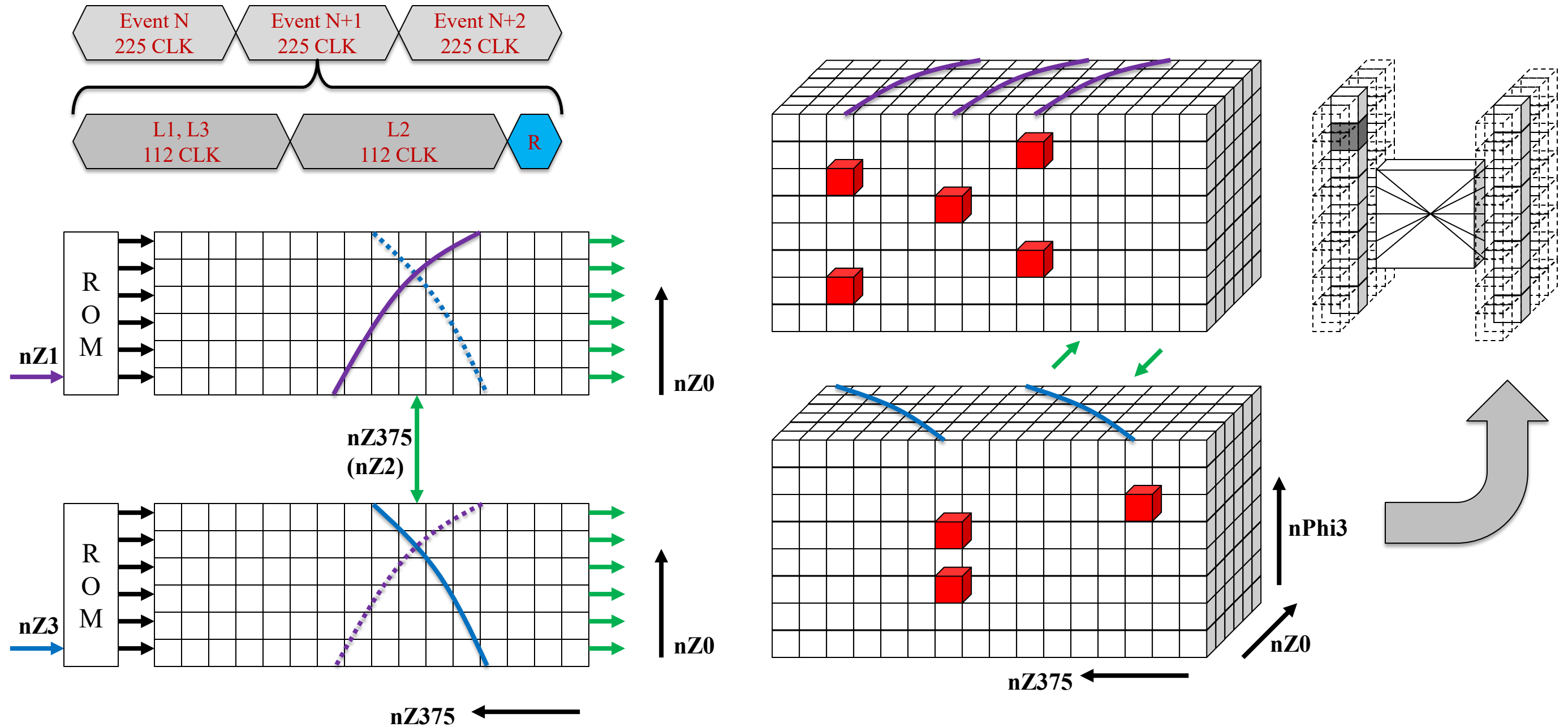


Pattern Shifting and checking

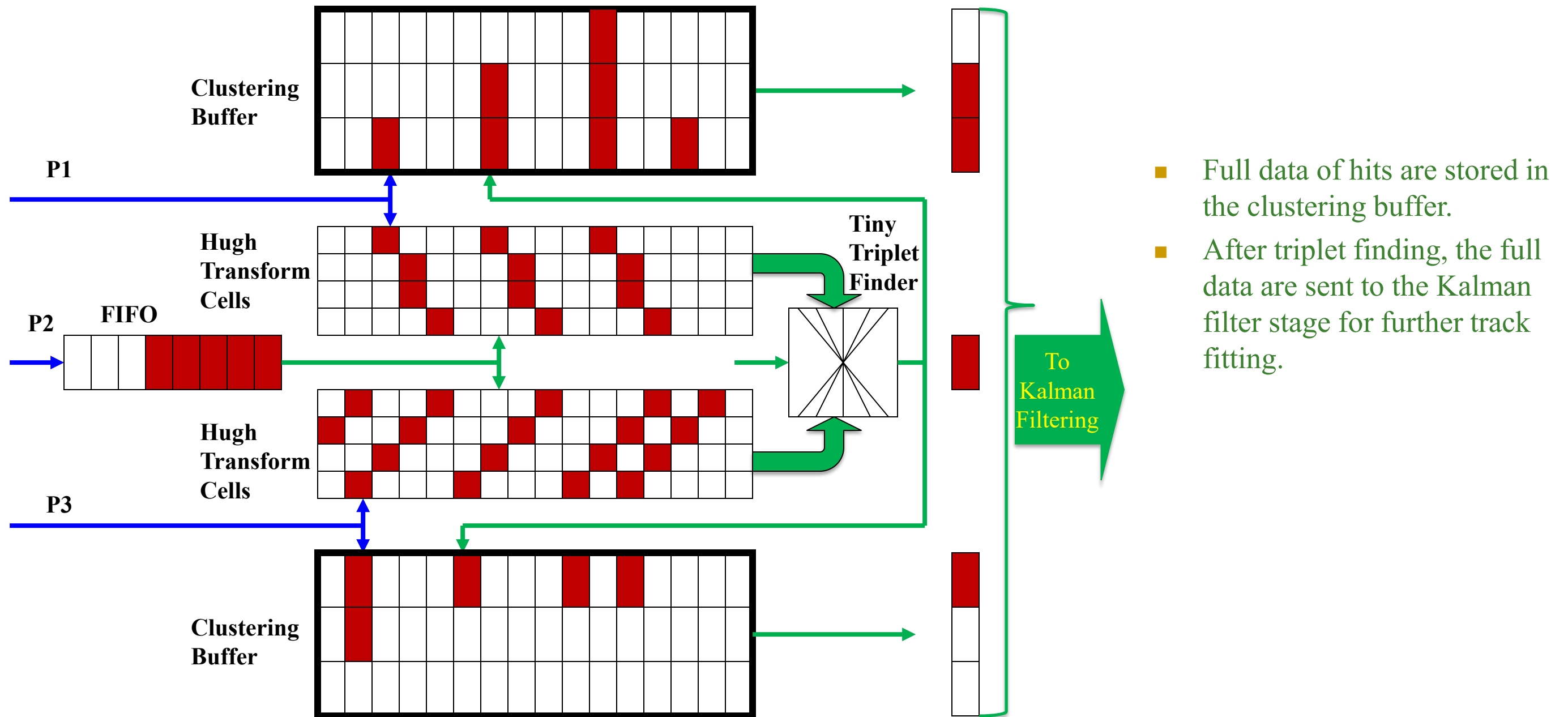
- The bit patterns with the same z_0 from Layer 1 and Layer 3 are shifted to align with the coincidence road array.
- The amount of shift is determined by ϕ_2 .
- Matching bits are output for further process.



Operation: (3) Refresh Cycle

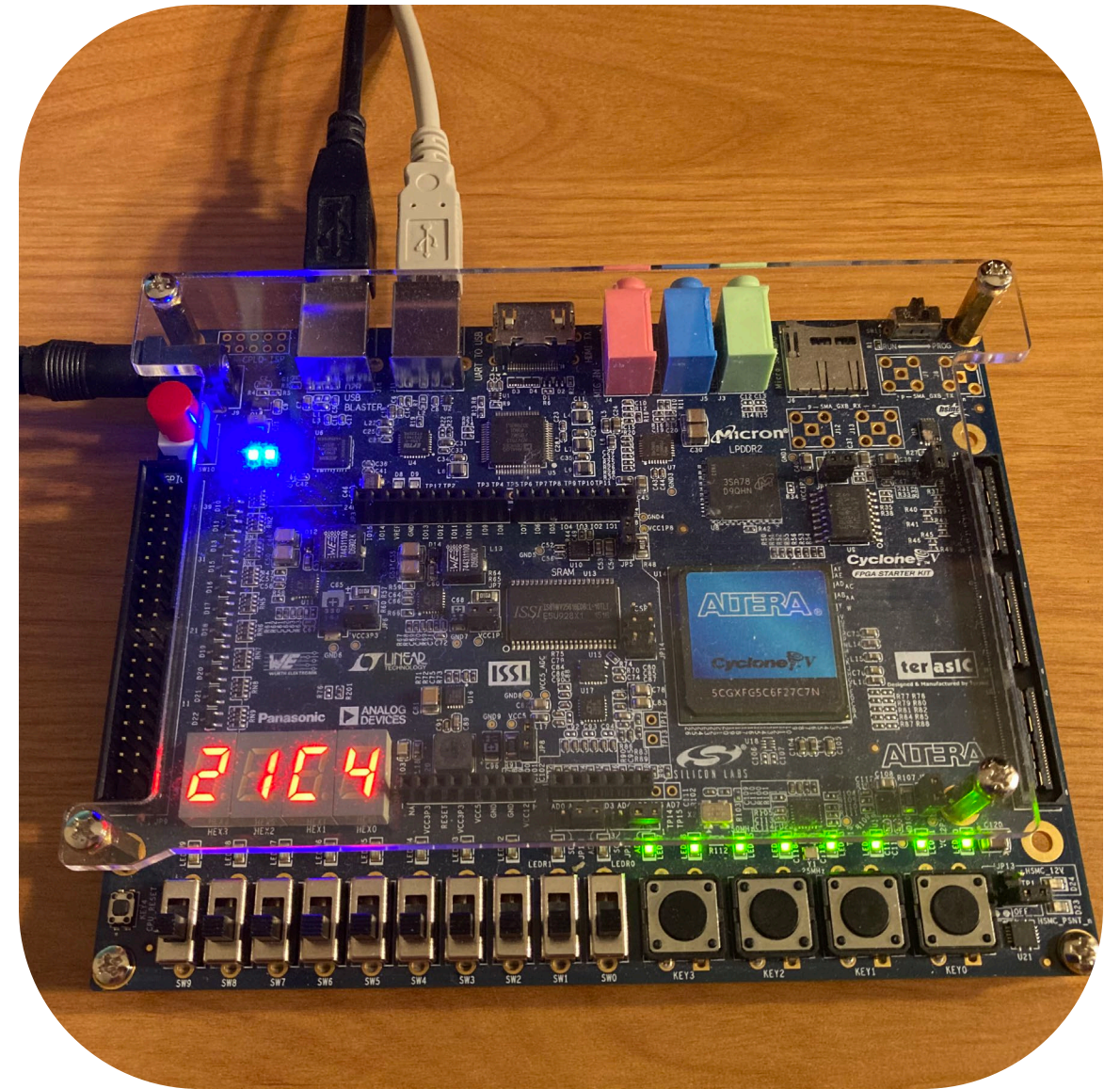
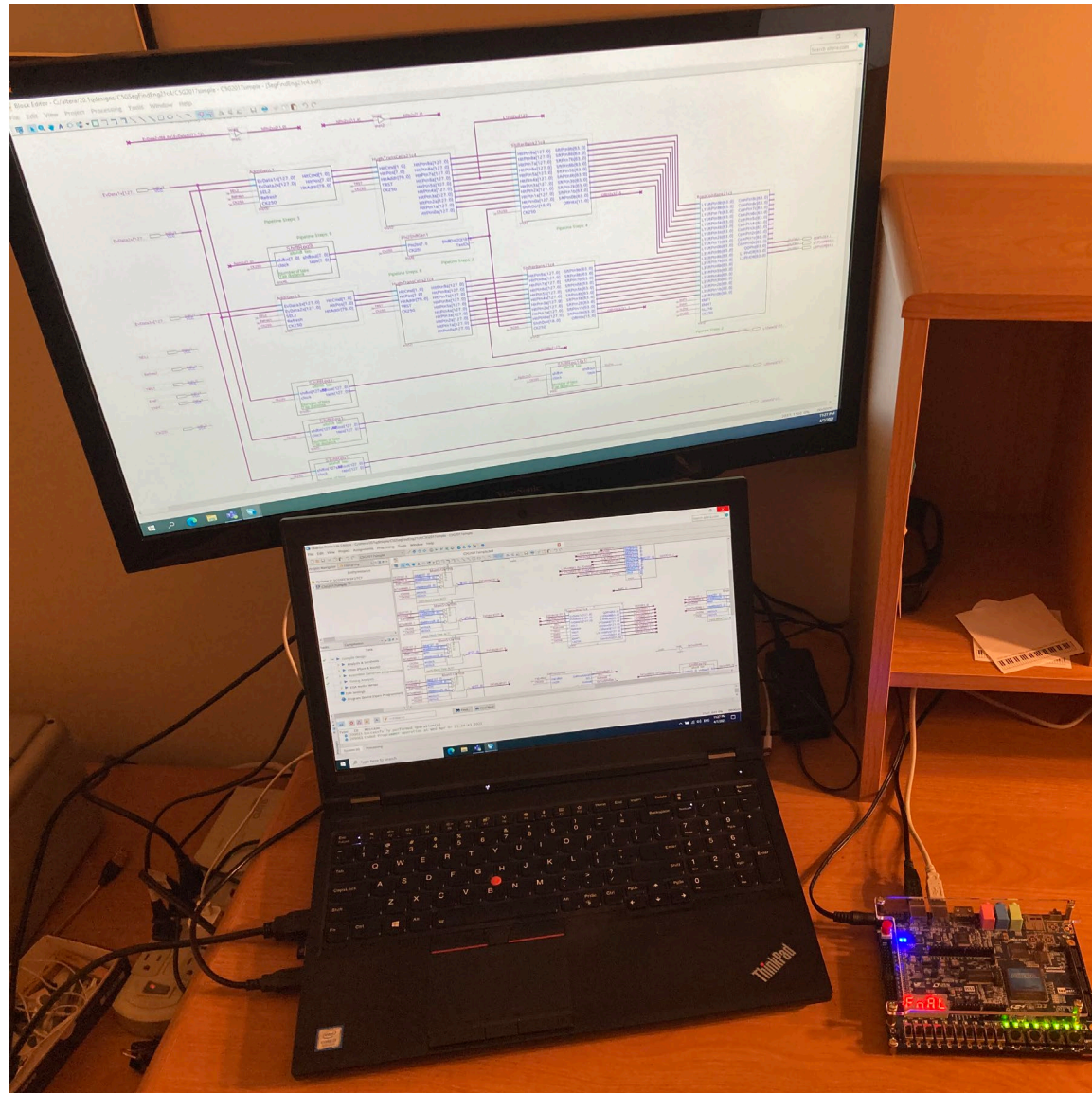


The Seeding Engine & Clustering Buffer



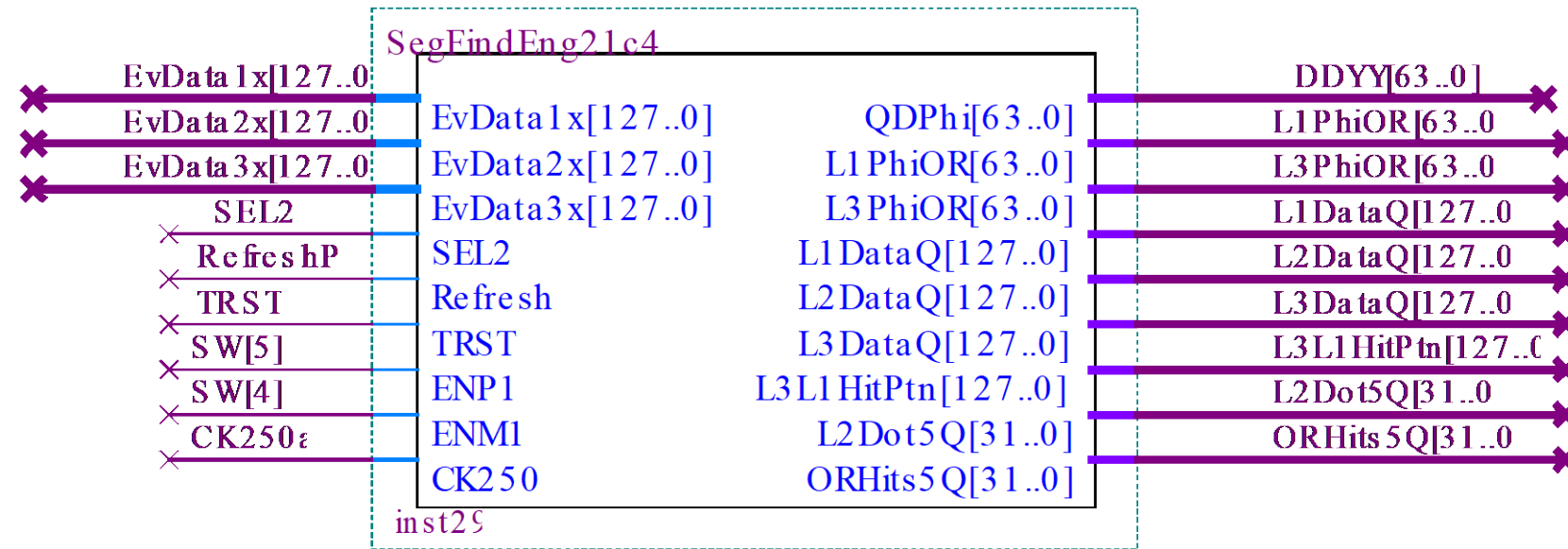
Implementation in Cyclone 5 FPGA

Home Test Stand and the Evaluation Module



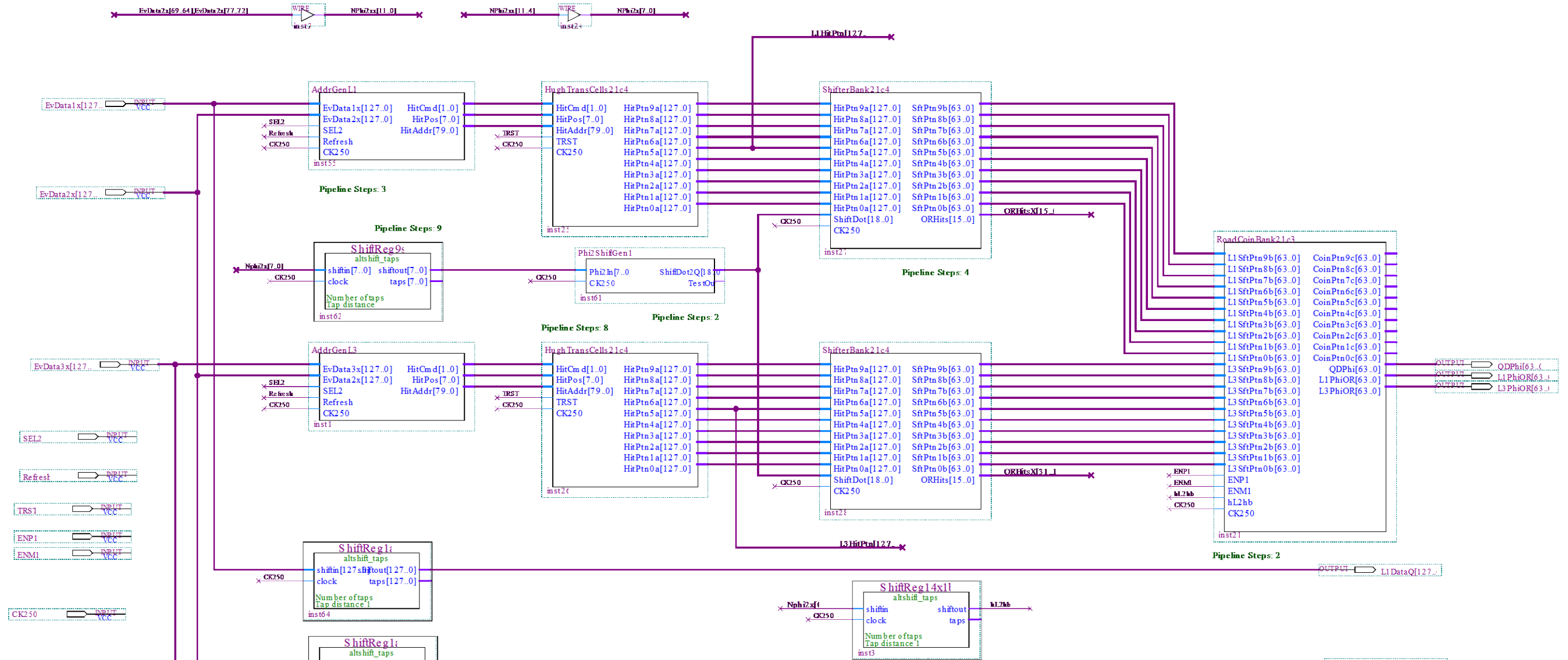
- The seeding engine core firmware is implemented in Altera C5G evaluation module (179 USD ea.).
- The FPGA is Cyclone V GX 5CGXFC5C6F27C7N (171 USD ea.).

The Seeding Engine Block I/O



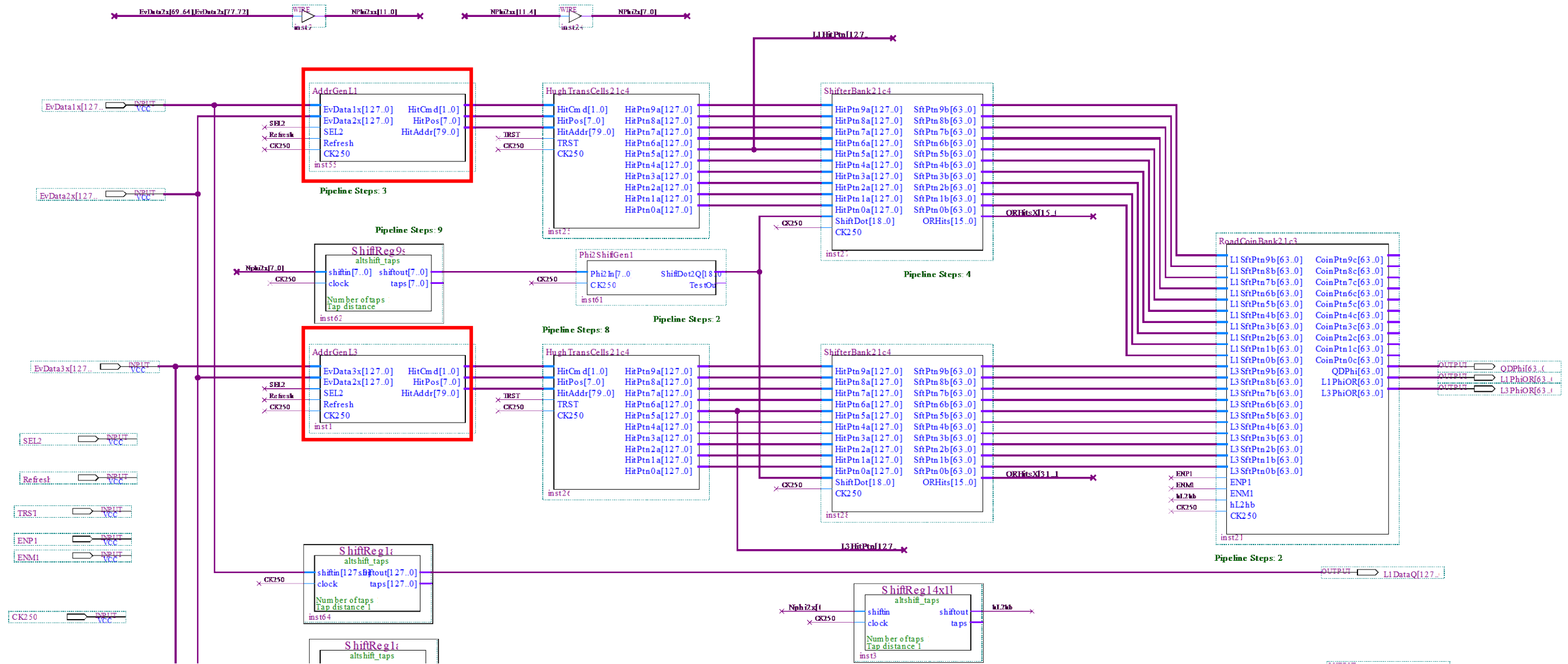
- The hit data from Layer 1, 2 and 3 are fed into the engine, one hit per clock cycle.
- The primary coincidence outputs are in port DDYY[63..0] which represents up to 64 possible 1/PT values of the track segment.

Structure of the seeding engine

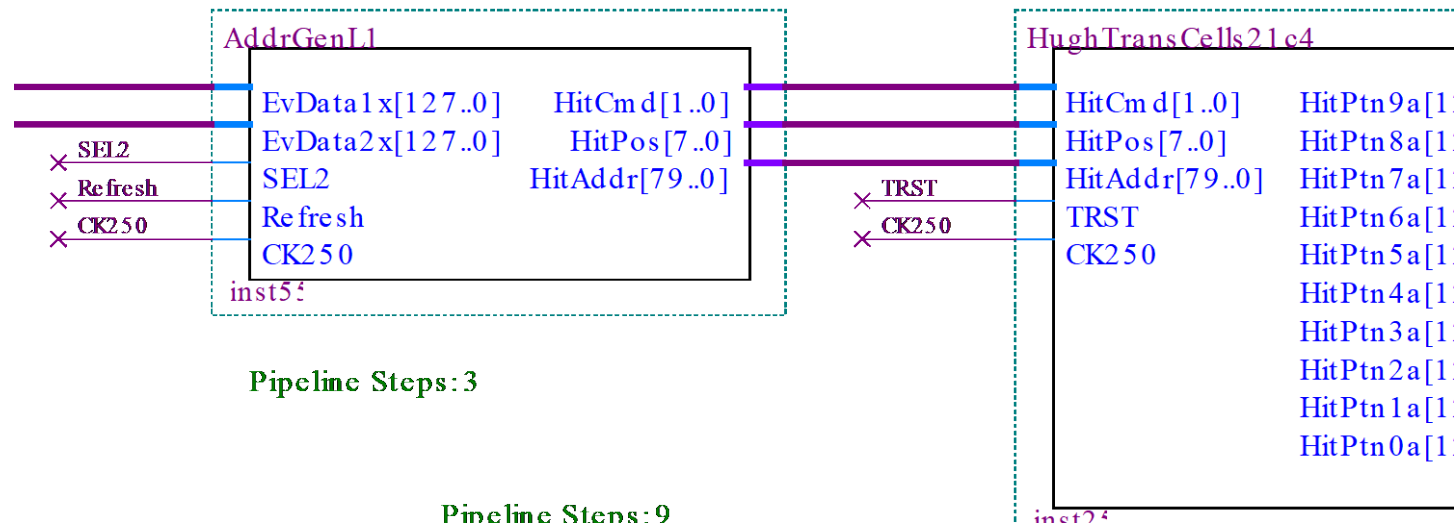


- The entire engine is organized as a pipeline so that it can run at high speed (250 MHz in Cyclone V).
- Blocks: Address Generation; Hugh Trans. Cells; Shifter Bank; Road Coincidence

The Seeding Engine



Address Conversion for Hugh Transform Cells



- For Layer 1 or Layer 3 Hugh Transform Cells have 10 RAM blocks corresponding 10 bins in z_0 .
- The z -coordinate z_1 or z_3 is converted to z_{375} for each z_0 by checking the lookup table ROM.
- The ROM contents depend on detector geometry and are generated using simulation data.

```
MemInit1.txt
--Format: Quartus .mif

WIDTH=80;
DEPTH=256;

ADDRESS_RADIX=HEX;
DATA_RADIX=HEX;

CONTENT BEGIN
00 : 00000000000000000000;
01 : 00000000000000000000;
02 : 00000000000000000000;
03 : 00000000000000000000;
04 : 00000000000000000000;
05 : 00000000000000000000;
06 : 00000000000000000000;
07 : 00000000000000000000;
08 : 00000000000000000000;
09 : 00000000000000000000;
0a : 00000000000000000000;
0b : 00000000000000000000;
0c : 00000000000000000000;
0d : 00000000000000000000;
0e : 00000000000000000000;
0f : 00000000000000000000;
10 : 00000000000000000000;
11 : 00000000000000000000;
12 : 00000000000000000000;
13 : 00000000000000000000;
14 : 00000000000000000000;
15 : 00000000000000000000;
16 : 00000000000000000000;
17 : 00000000000000000000;
18 : 00000000000000000000;
19 : 00000000000000000000;
1a : 00000000000000000000;
1b : 00000000000000000000;
1c : 00000000000000000000;
1d : 00000000000000000000;
1e : 00000000000000000000;
1f : 00000000000000000000;
20 : 00000000000000000000;
21 : 00000000000000000000;
22 : 00000000000000000000;
23 : 00000000000000000000;
24 : 00000000000000000000;
25 : 00000000000000000000;
26 : 00000000000000000000;
27 : 00000000000000000000;
28 : 01000000000000000000;
29 : 02010000000000000000;
2a : 04030201000000000000;
2b : 05040302010000000000;
2c : 07060504030201000000;
2d : 08070605040302010000;
2e : 0a090807060504030201;
2f : 0b0a0908070605040302;
30 : 0d0c0b0a090807060504;
31 : 0e0d0c0b0a0908070605;
32 : 100f0e0d0c0b0a090807;
33 : 11100f0e0d0c0b0a0908;
```

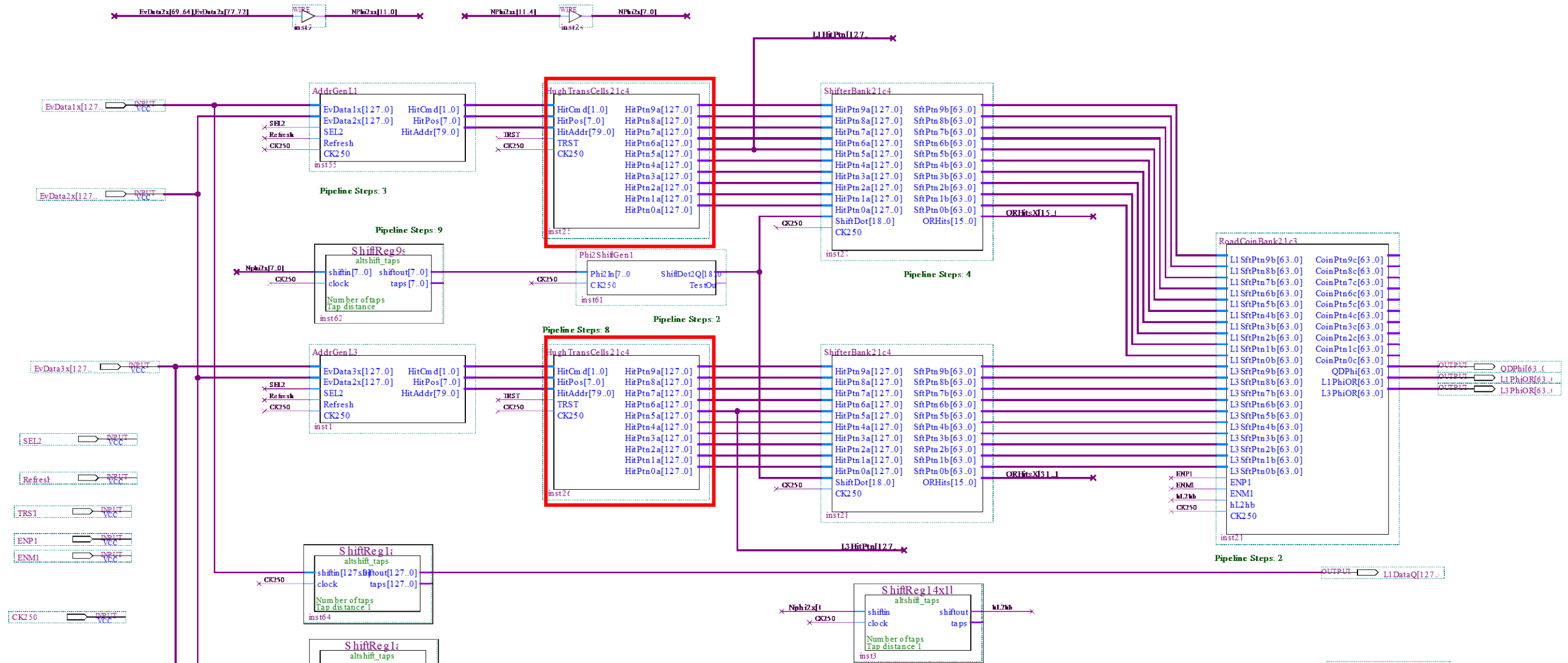
```
MemInit3.txt
--Format: Quartus .mif

WIDTH=80;
DEPTH=256;

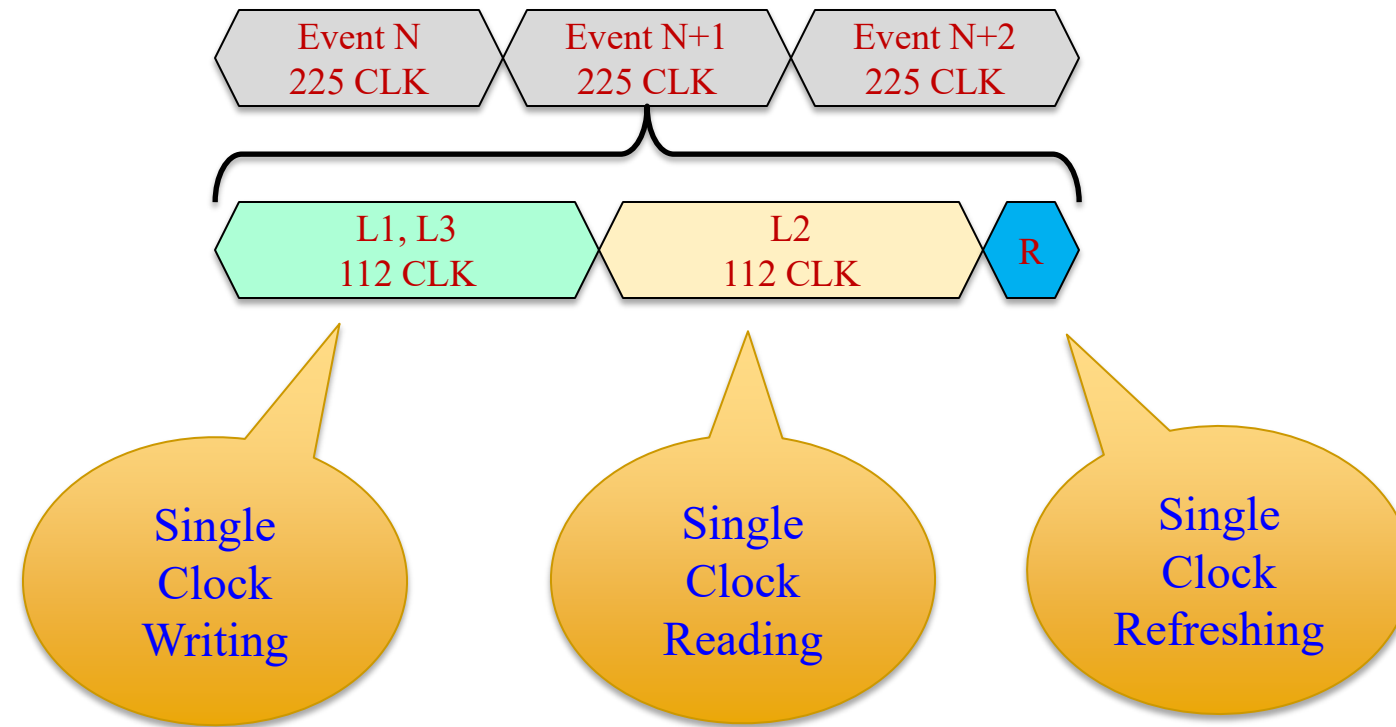
ADDRESS_RADIX=HEX;
DATA_RADIX=HEX;

CONTENT BEGIN
00 : 2222324242525262627;
01 : 23232424252526272728;
02 : 23242425262627272828;
03 : 24252526262727282929;
04 : 2525262627282829292a;
05 : 25262727282829292a2b;
06 : 262727282829292a2b2b;
07 : 27272829292a2b2b2c;
08 : 282829292a2b2b2c2c2d;
09 : 2829292a2b2b2c2c2d2d;
0a : 292a2a2b2b2c2c2d2e2e;
0b : 2a2a2b2b2c2c2d2d2e2f;
0c : 2a2b2c2c2d2d2e2f30;
0d : 2b2c2c2d2d2e2f3030;
0e : 2c2c2d2e2f303031;
0f : 2d2d2e2f30313132;
10 : 2d2e2f3031313232;
11 : 2e2f303131323333;
12 : 2f30313132333334;
13 : 2f3031313233333435;
14 : 303131323333343535;
15 : 313132333334353536;
16 : 323233333435363637;
17 : 323333343536363738;
18 : 333434353637383839;
19 : 34343536373838393a;
1a : 343536373838393a3b;
1b : 3536373838393a3b3c;
1c : 36373838393a3b3c3d;
1d : 37373838393a3b3c3d3e;
1e : 373838393a3b3c3d3e3f;
1f : 3839393a3b3c3d3e3f40;
20 : 39393a3b3c3d3e3f4041;
21 : 393a3b3c3d3e3f404142;
22 : 3a3b3c3d3e3f40414243;
23 : 3b3c3d3e3f4041424344;
24 : 3c3d3e3f404142434445;
25 : 3c3d3e3f40414243444546;
26 : 3d3e3f4041424344454647;
27 : 3e3f404142434445464748;
28 : 3e3f40414243444546474849;
29 : 3f40414243444546474849;
2a : 4040414243444546474849;
2b : 41414243444546474849;
2c : 414243444546474849;
2d : 4243444546474849;
2e : 43444546474849;
2f : 43444546474849;
30 : 444546474849;
31 : 454546474849;
```

The Seeding Engine



Register-Like Block RAM for Hit Patterns



Register-Like Block RAM: Implementation, Testing in FPGA and Applications for High Energy Physics Trigger Systems

Jinyuan Wu

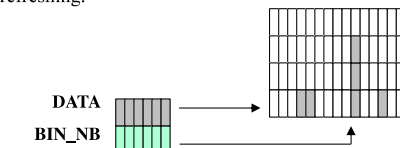
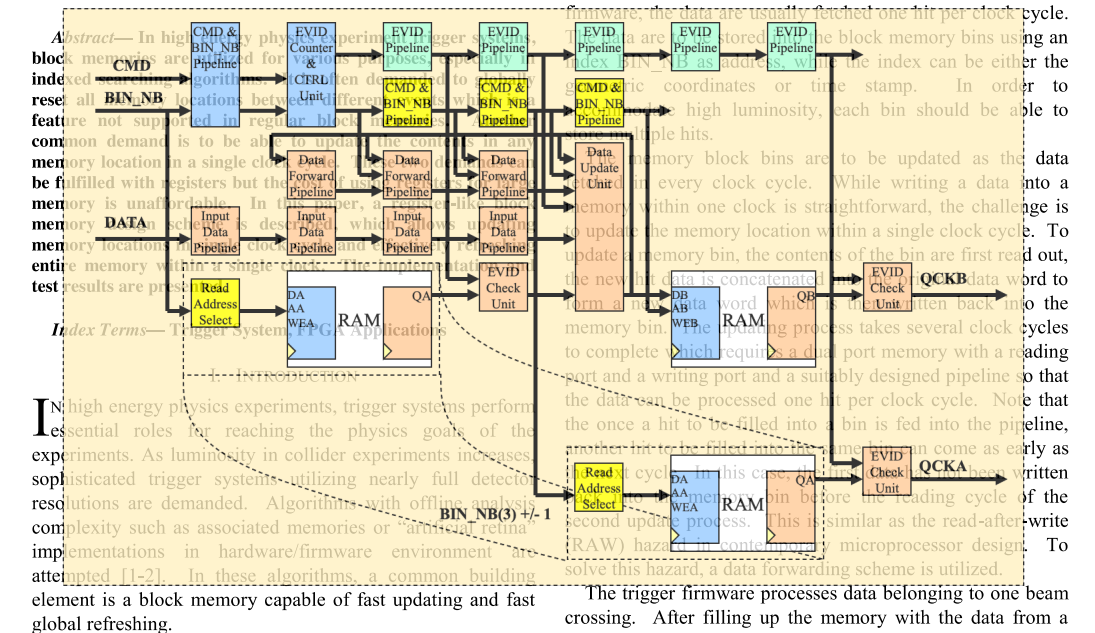


Fig. 1. An application of block memory bins used for in high energy physics experiment trigger system.

A typical application of the block memories organized in memory bins is shown in Fig. 1. The detector data representing hit coordinates and timing information are fed into the trigger system and in the processing stages of the

Manuscript received May 20, 2016. This work was supported in part by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy.

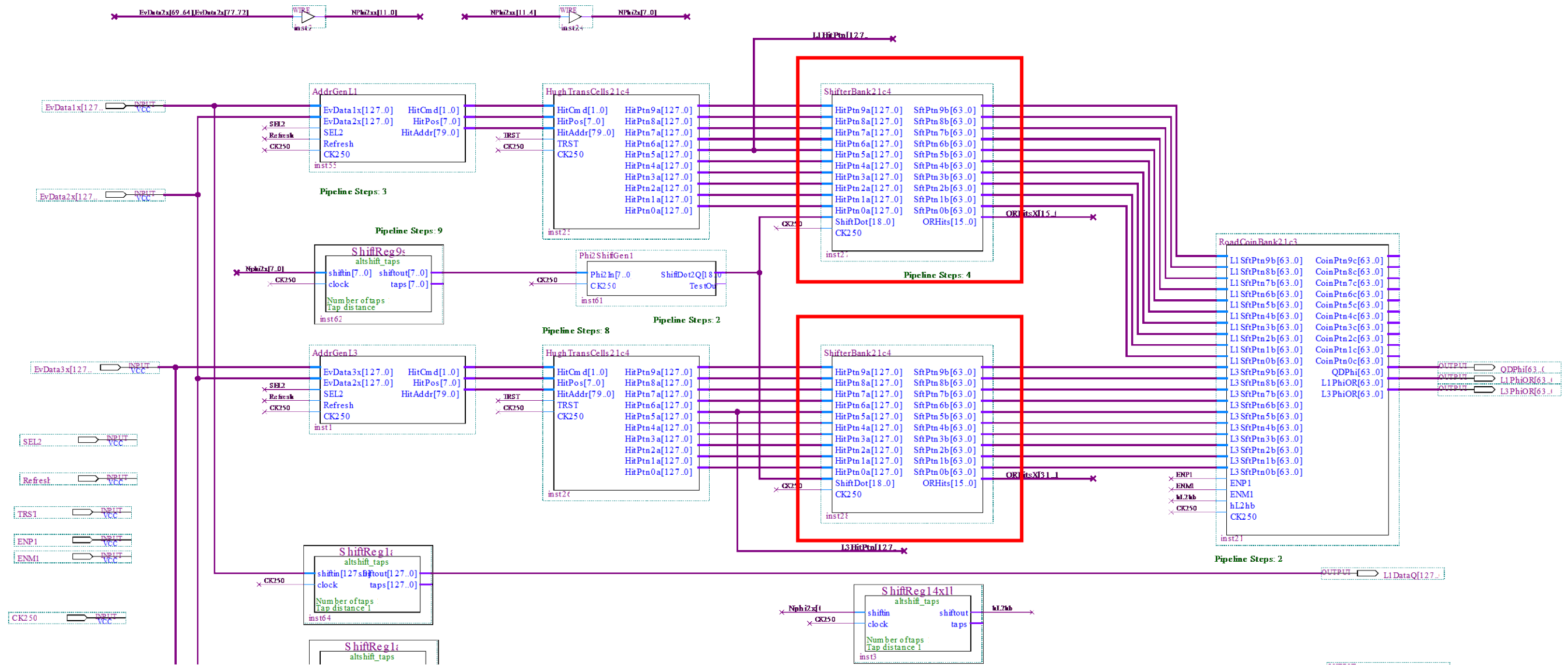
The author is with Fermi National Accelerator Laboratory, Batavia, IL 60510 USA (phone: 630-840-8911; fax: 630-840-2950; e-mail: jywu168@fnal.gov).

II. GLOBAL REFRESH SCHEME FOR BLOCK MEMORY

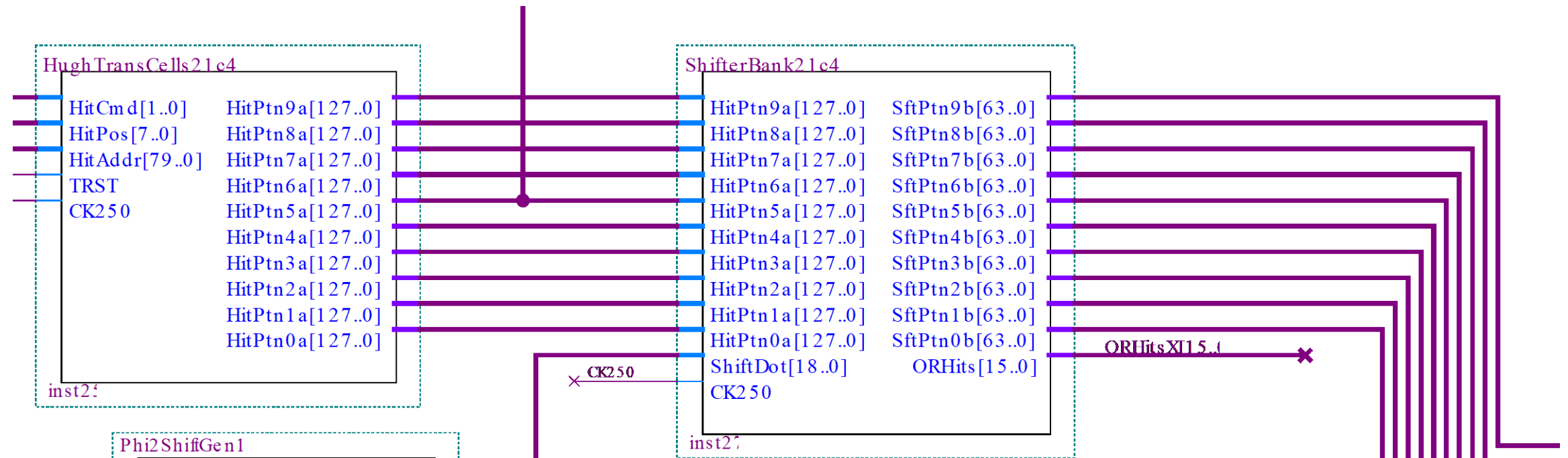
Intrinsically, the block memories can only be accessed one

- The hit patterns are stored in the register-like BRAM.
- Each hit is processed in a single clock cycle.
- Single clock cycle refreshing allows fast preparation for next event.

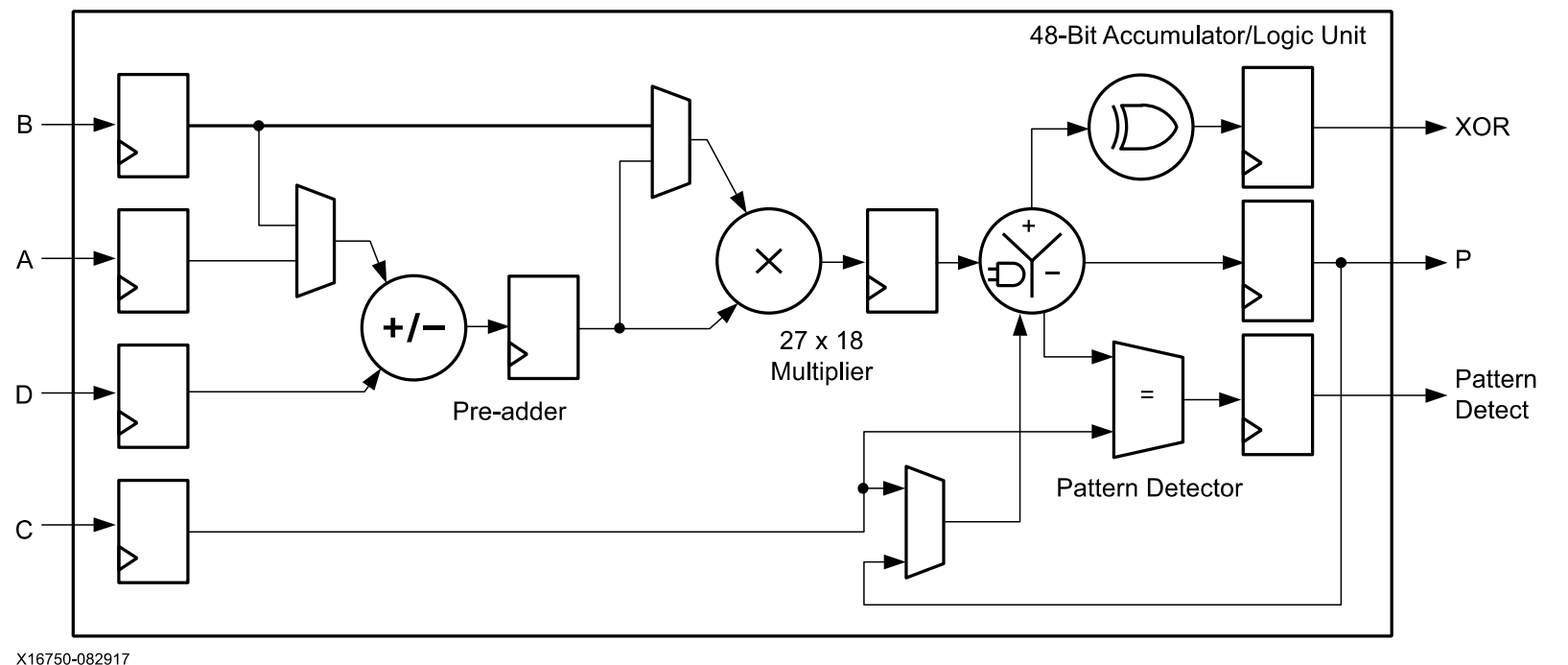
The Seeding Engine



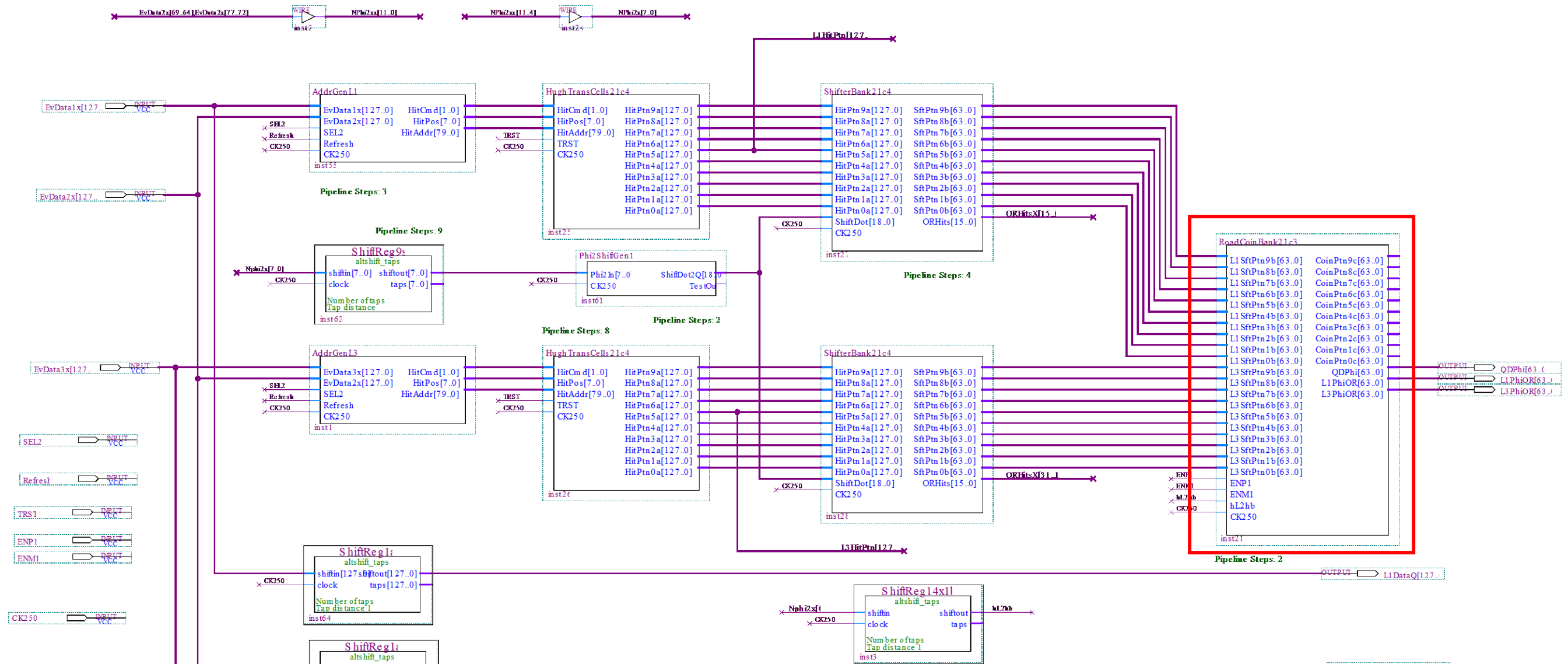
Shifters



- The multipliers in FPGA DSPs are used to implement shifters.



The Seeding Engine



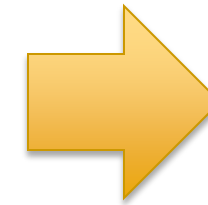
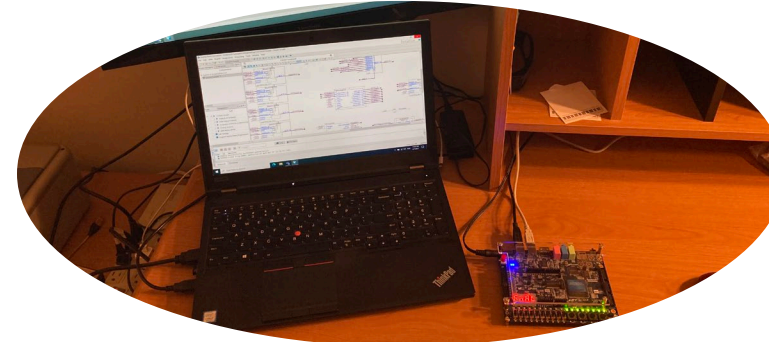
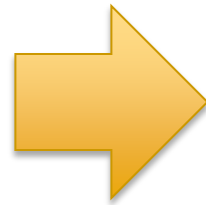
inst21

- 42

Test Results

Raw Input Data and Output Data

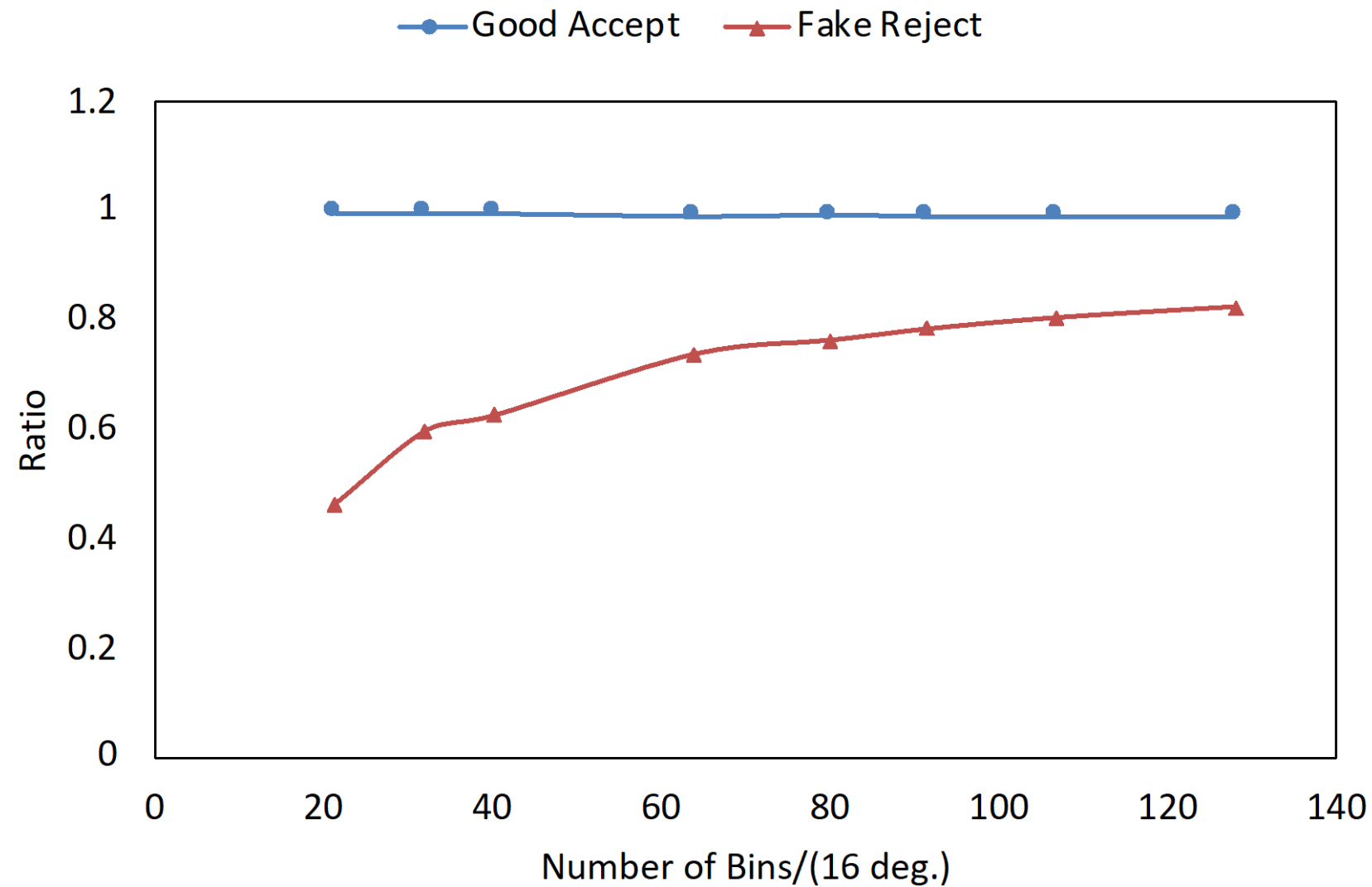
```
HitD... Edited
00L1PXFmkP`PV]
00L1PXFN8m^4V^
00L1PXF0[nkLV_
00L1PP@@@@@WP
00L1PP@@@@@WQ
00L1PP@@@@@WR
00L1PP@@@@@WS
00L1PP@@@@@WT
00L1PP@@@@@WU
00L1PP@@@@@WV
00L1PP@@@@@WW
00L1PP@@@@@WX
00L1PP@@@@@WY
00L1PP@@@@@WZ
00L1PP@@@@@W[
00L1PP@@@@@W\
00L1PP@@@@@W]
00L1PP@@@@@W^
00L1PP@@@@@W_
00L2PY@L@^VPP
00L2PY@AT0gJPQ
00L2PY@B[UMXPR
00L2PY@CaL[UPS
00L2PY@D\l1MPT
00L2PY@EaD\hPU
00L2PY@F0ZM9PV
00L2PY@GYH3iPW
00L2PY@HU8jHPX
00L2PY@IN:V:PY
00L2PX@JjjU`PZ
00L2PX@K0W0KP[
```



- Hits (stubs) are sent into the seeding engine.
- Each line is a hit in a detector layer.
- Typical data: 100 hits/layer with 10 good tracks with 90 random hits.
- The seeding engine outputs a 64-bit pattern, and the non-zero bits represent a potential track segment.

```
temp1.txt Edited
00000000000000000000000000000000
000000200000000040405950324C38
0000000000000400041405950324C38
000001000000000042405950324C38
0000000000001000043405950324C38
000000004000000044405950324C38
000000000000800045405950324C38
000020000000800046405950324C38
000000000000800047405950324C38
000000020000000048405950324C38
000400000000000049405950324C38
00000000000000004A405850324C38
00000000010000004B405850324C38
00000000000000004C405850324C38
00000000000000004D405850324C38
00000000000000004E405850324C38
00000000080000004F405850324C38
000000000000000040415850324C38
000000000000000041415850324C38
000020000000000042415850324C38
000000000200000043415850324C38
000000000000000044415850324C38
000000000000000045415850324C38
000000000000000046415850324C38
000000000000000047415850324C38
000000000000000048415850324C38
```

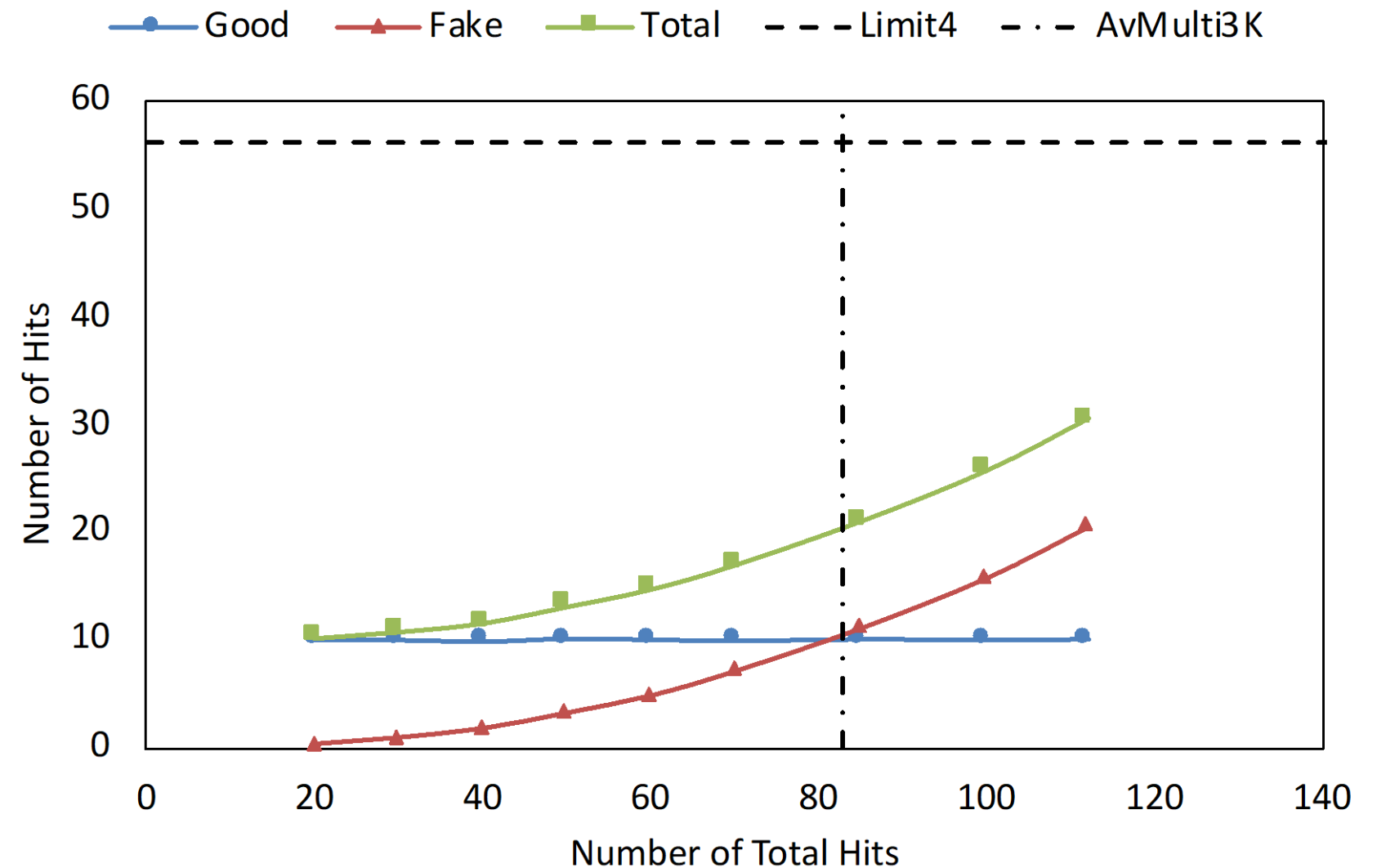
Good Track Acceptance and Fake Rejection



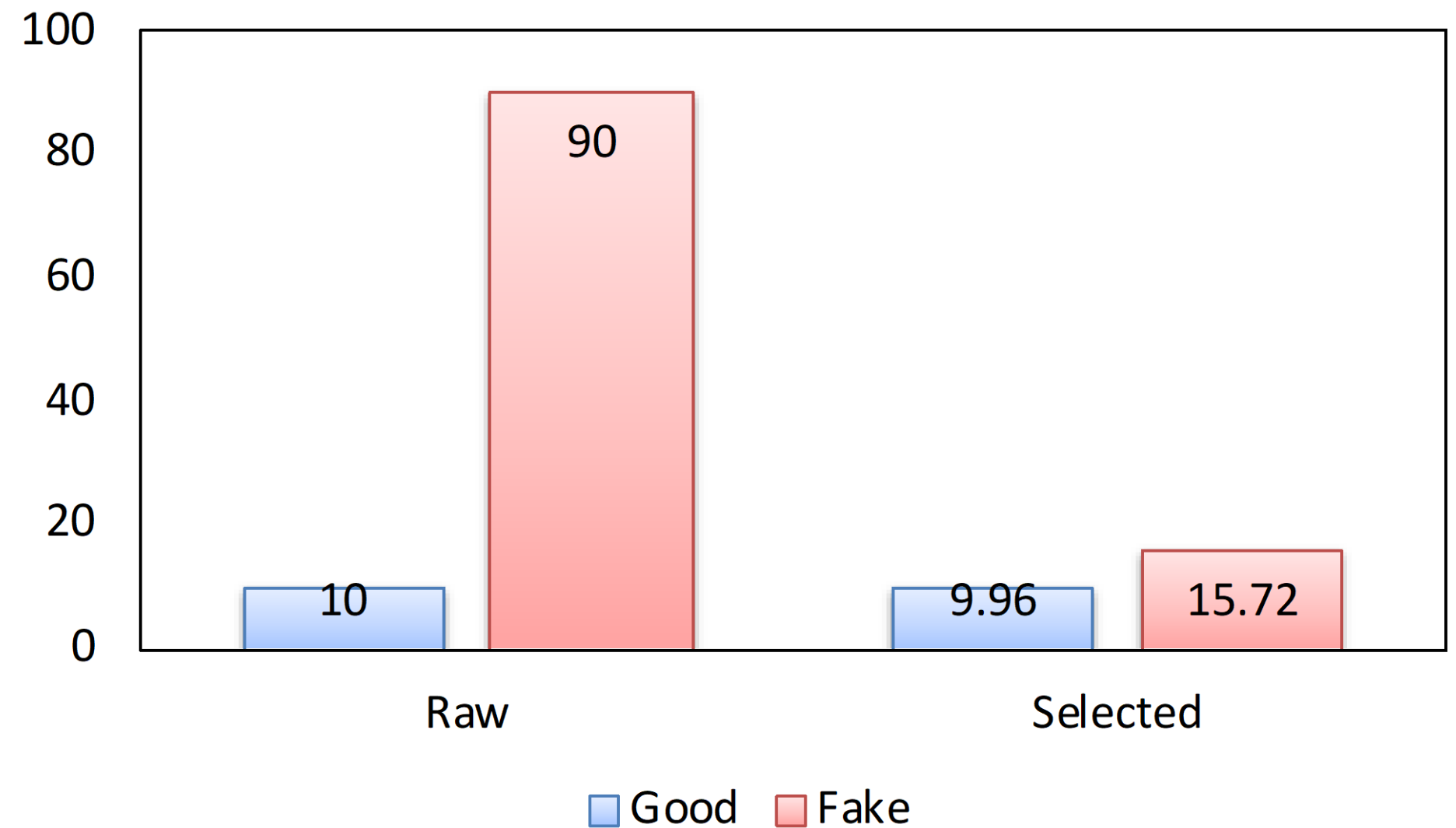
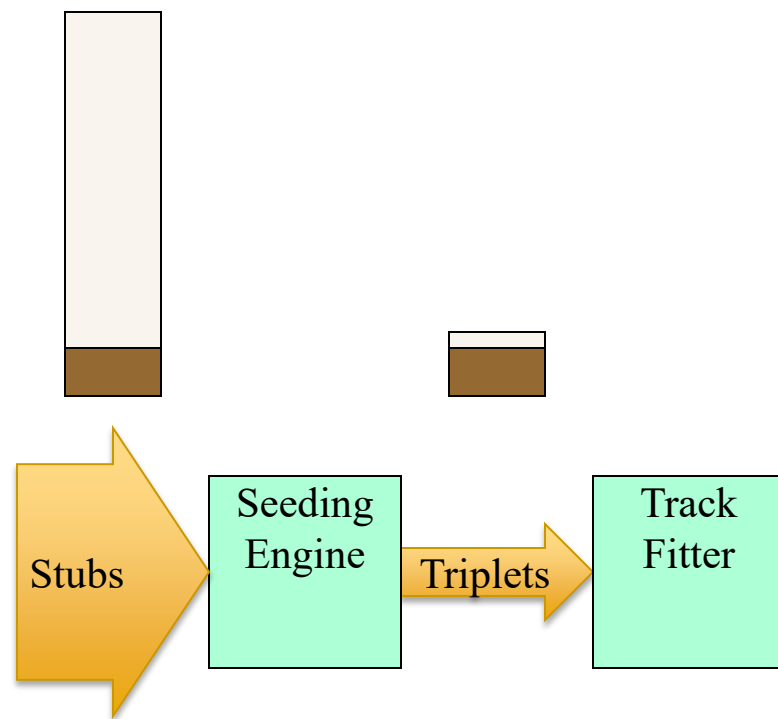
- Good track acceptance is better than 99%.
- Increasing number of bins in phi causes better fake rejection.

Fake Tracks due to High Occupancy

- When number of total hits per event (BX) is low, fake rate is low.
- Event beyond occupancy of 3000 stubs/layer for the whole detector (the AvMulti3K line), total number of coincidence is still manageable.
- The dash line Limit4 indicate total coincidence allowed in fitting stage if each coincidence is processed with 4 clock cycles.



Performance of the Seeding Engine

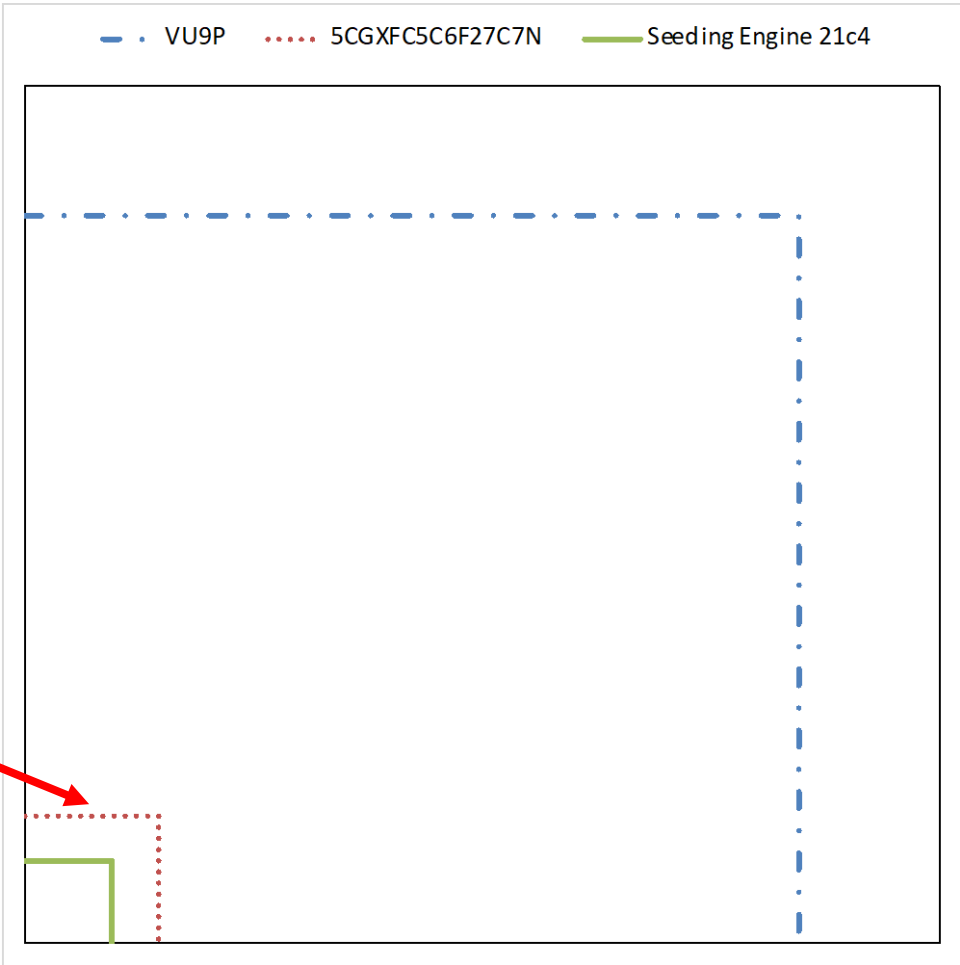
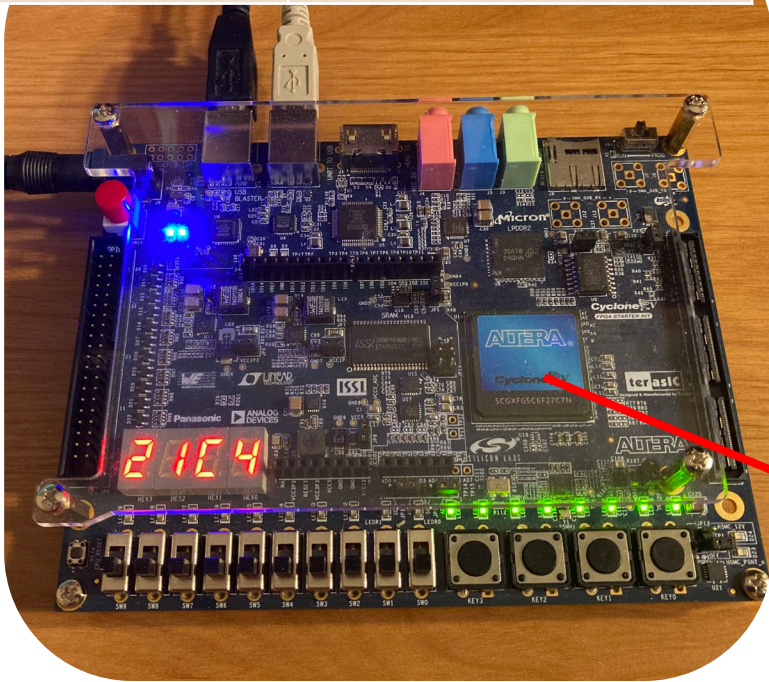


- Most of fake hits are rejected while good tracks are preserved.
- Data volume sent to the fitting stage is much smaller.

Resource Usage:

	Used in a Seeding Engine core	Available in Cyclone 5 5CGXFC5C6F27C7N	Available in Ultra-scale Plus VU9P
Logic Cells	12446 (ALM) Eq. 32.9K (LC)	29080 (ALM)	2586K (LC)
Block RAM	150 (M10K) 743K (bits)	446 (M10K) 4567K (bits)	75.9M (bits)
DSP	100 (16x16 Multipliers)	150 (DSP 27x27) Eq. 300 (16x16)	6840 (DSP 18x27)

- The seeding engine core fits the Cyclone 5 device with ~45% resource usage.
- Four cores are needed in VU9P and it looks possible.



Summary

- **Algorithms:** well-known in the field.
 - The fake rate of track segment seeding with triplets is lower.
- **Implementations:** exist in the field.
 - The challenge is how to fit into a reasonable sized FPGA.
- **Methods** of solving the implementation challenge: this work.
 - Using Tiny Triplet Finder so that only a small number of coincidence roads (product terms) are implemented.
 - Hit patterns are implemented with “register-like RAM” which allows single clock update, read and refresh operations.
 - Lookup tables and coincidence roads are auto initiated using simulation data to accommodate detector geometry or alignment variations.

An aerial photograph of a racetrack complex. The track is a dark, winding oval shape, surrounded by green grass and dense trees. In the background, there are several large, light-colored industrial or warehouse buildings. To the right of the track, there is a large, white, curved structure that looks like a grandstand or a viewing platform. The overall scene is a mix of natural greenery and man-made structures.

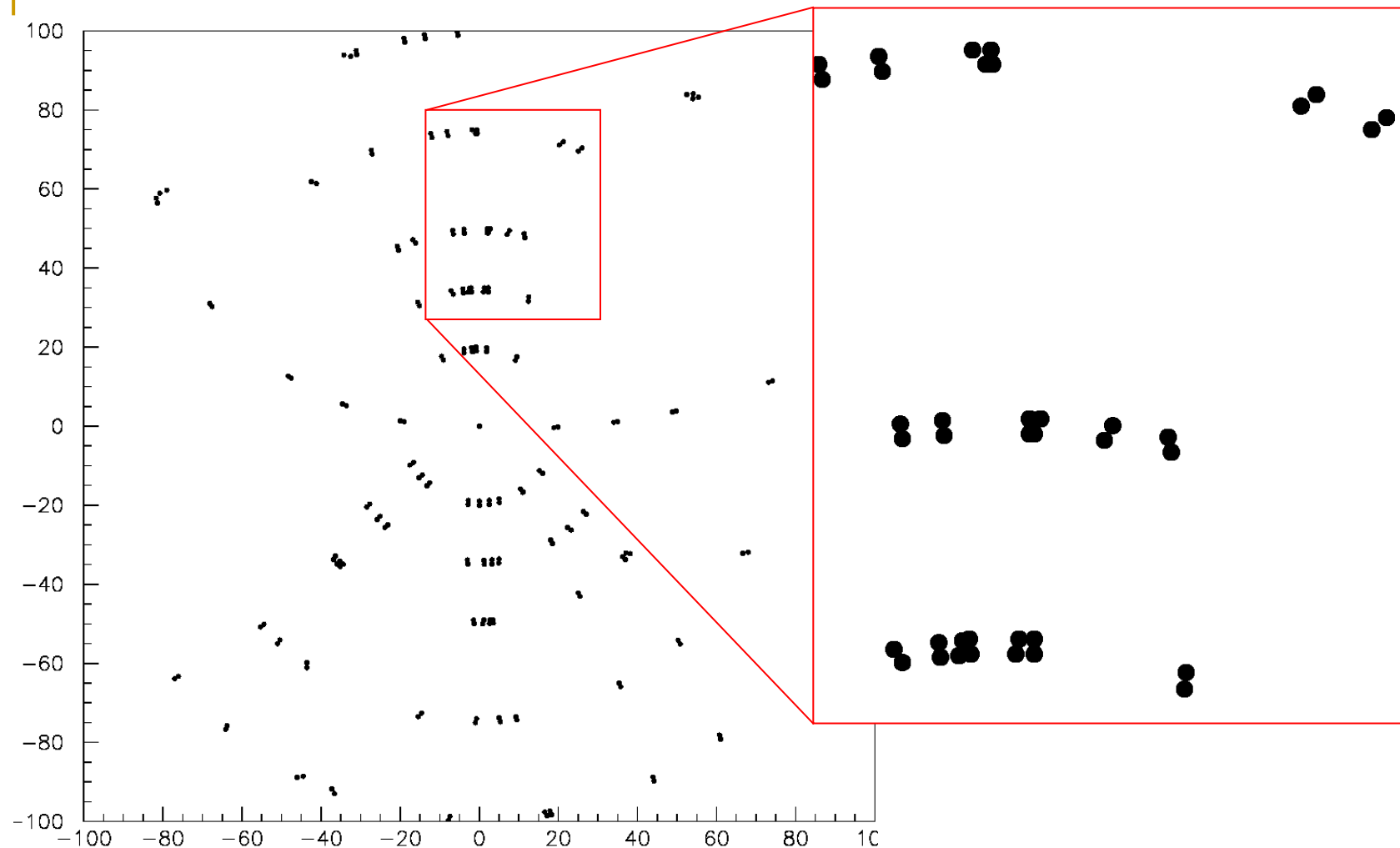
The End

Thanks

Introduction

- For high-luminosity operation, 3D track segment seeding becomes necessary:
 - Usable bin sizes in z or phi direction can't be too fine due to:
 - Multiple scattering
 - Non-zero impact parameter
 - The 2D segment seeding would yield too many fake segments.
 - Constraints in both r-z and r-phi planes help to reduce fake segments.
- Using the Tiny Triplet Finder to implement 3D track segment seeding engines
 - The Tiny Triplet Finder is a low-resource usage scheme for track segment finding in FPGA.
 - It becomes affordable to implement 3D seeding engines since the resource usage is low.
- A true 3D track segment seeding engine has been implemented and tested.
 - Detector partition, data timing etc. are based on the seeding-fitting approach in recent CMS presentations: [Mark Pesaresi, ACES 2020] and [Luigi Calligaris RT2020].
 - Stubs are generated on idealized Outer Tracker 3 PS layers.
 - Operates as intended.

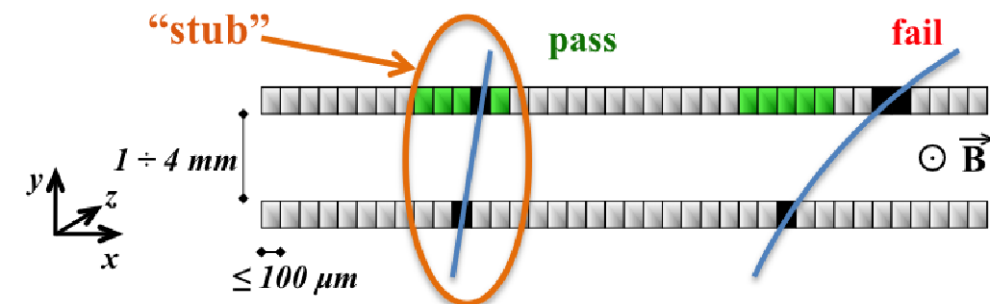
Pairing or Triplet Finding



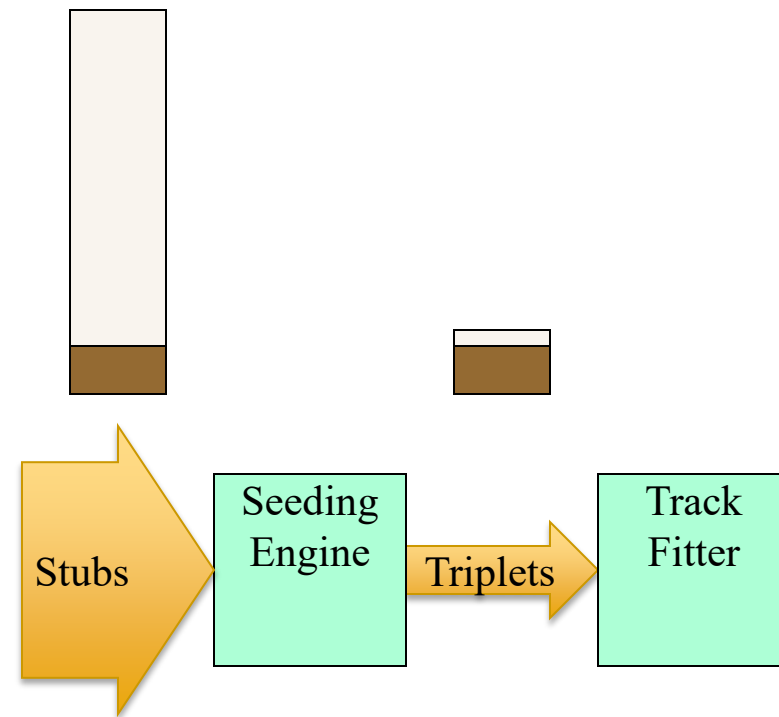
- If the separation of silicon sensor layers were large (~ 10 mm), then stub **pairing** would be sufficient.
- Given the separation of ~ 4 mm in PS and 2S module, correlating **triplets** of stubs becomes necessary.

Wu et. al. 2004

https://indico.cern.ch/event/418639/contributions/1018451/attachments/868833/1216631/lowpt_lecc2004p.pdf



Why Seeding



L1 TRACKING CHALLENGE

Immense data rates at ultimate luminosity

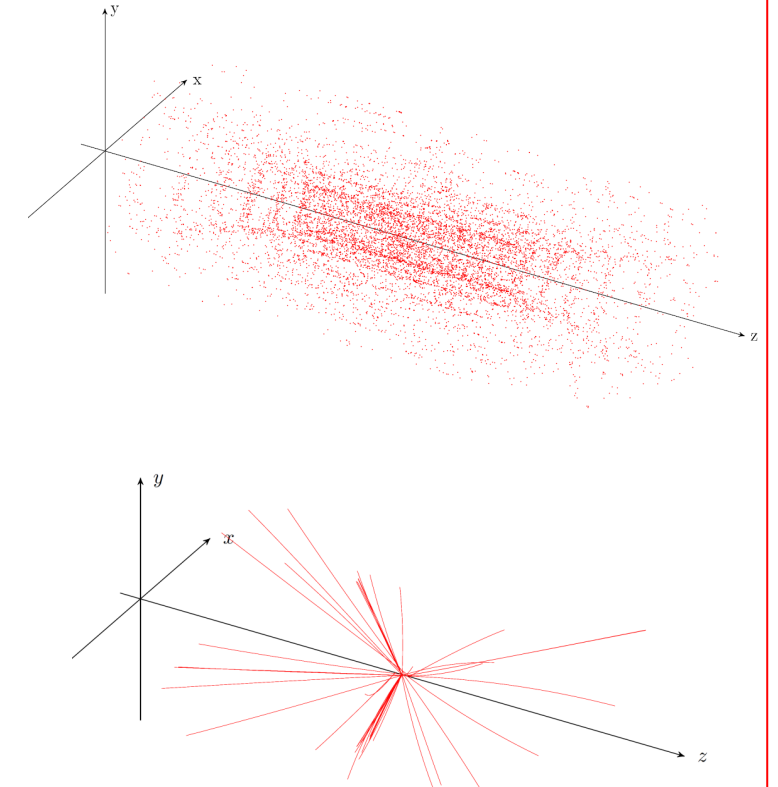
- Average **~30 Tb/s** from detector

Significant combinatorial tracking problem

- **15k – 20k stubs** every 25 ns from Tracker (200PU)
- Stubs are input for tracking

Limited time for track finding

- L1 latency limited to 12.5 μ s
- Must provide tracks within **5 μ s** after collision
- **~300 tracks** with $p_T > 2$ GeV (200PU)

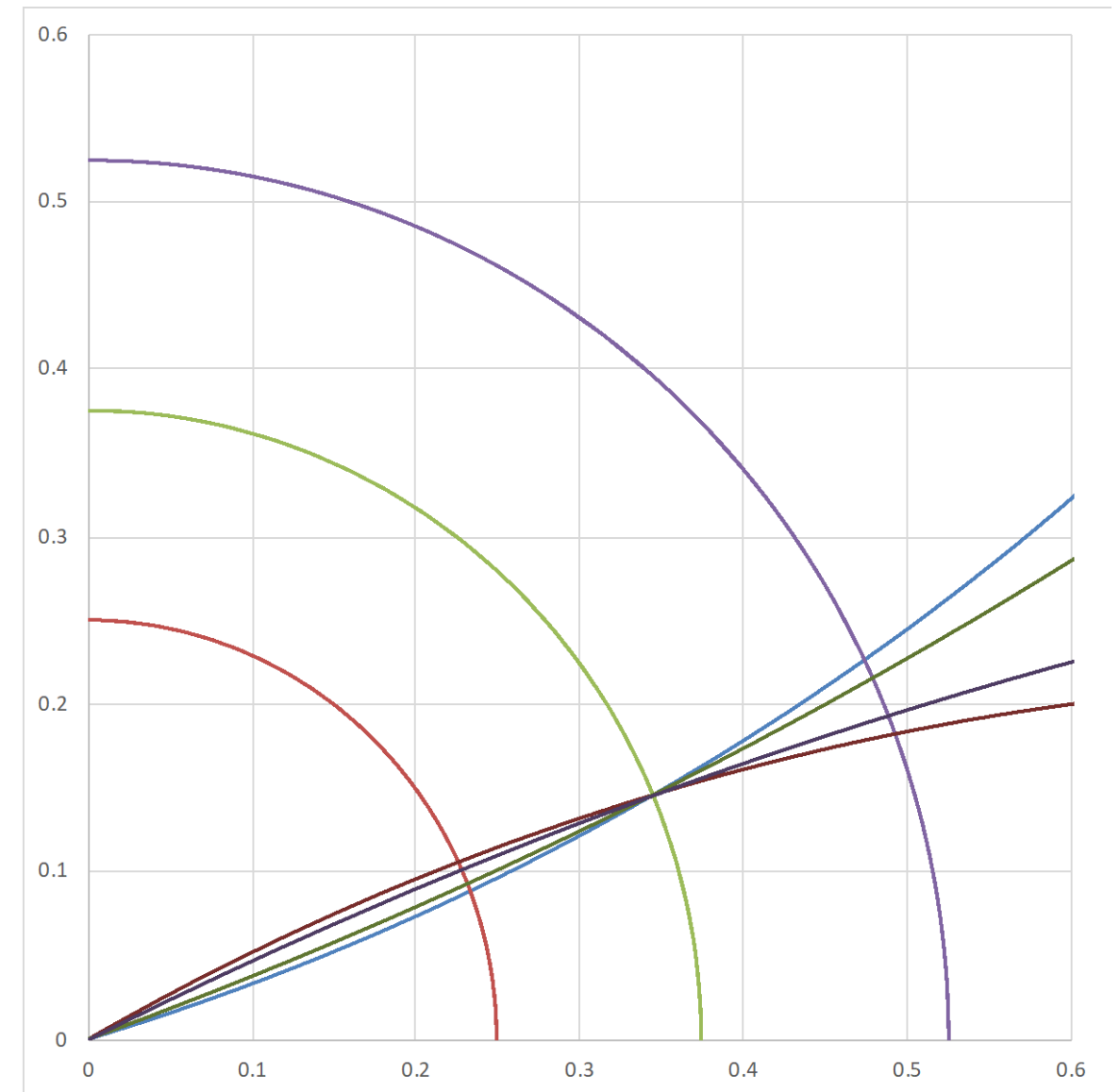


Mark Pesaresi et. al. ACES 2020

- At 200 pile-up, the tracker sees 15k-20k stubs $\Rightarrow \sim 3000$ stubs/layer.
- There are ~ 300 high Pt tracks ($Pt > 2$ GeV).
- About 90% stubs are not associated with any high Pt track. \Rightarrow more than 90% stub pairs are fake.
- Requesting additional constraints from 3 or more stubs (triplet) will pre-exclude fake pairs.

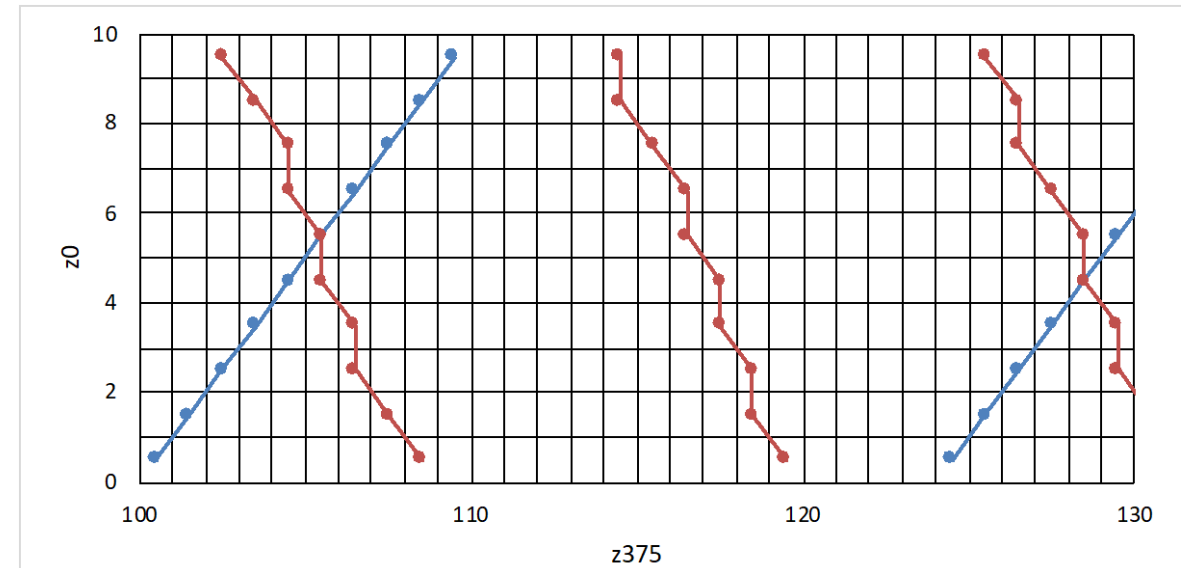
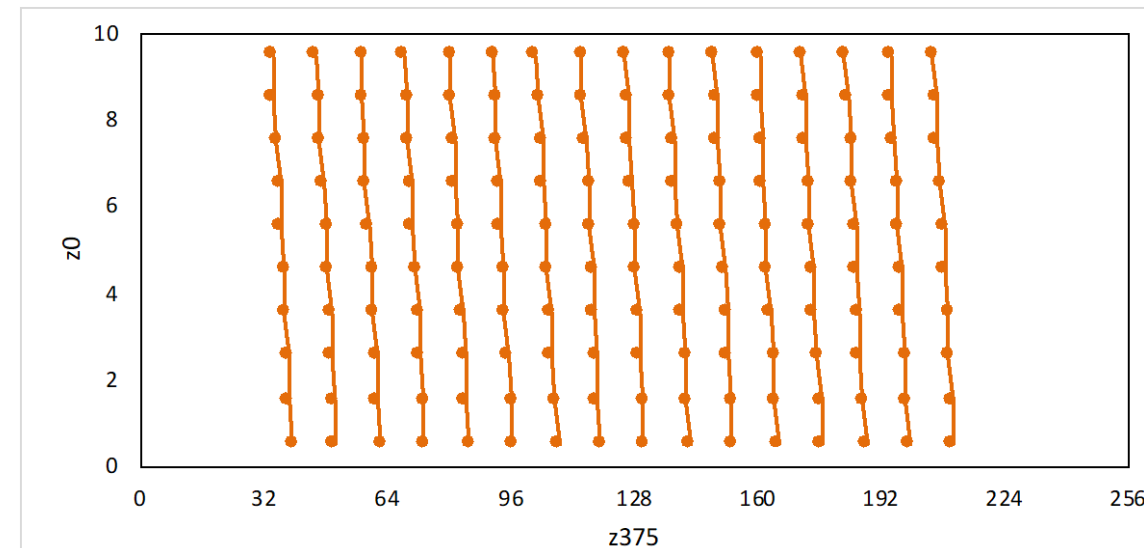
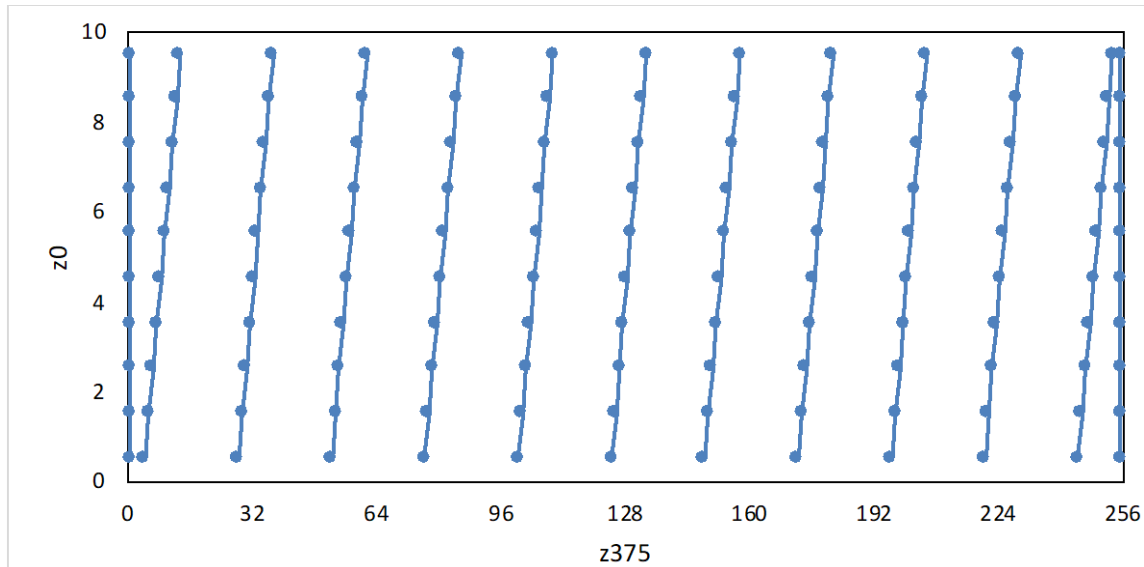
The Searching Range: Layer 2 to 1 and 3

PT(GeV/c)	2.000		
B(T)	4.000		
R0(m)	1.667		
Z0(m)	0.200		
Detector Length(m)	2.400		
Number of stubs per layer	3000.000		
	Layer 1	Layer 2	Layer 3
R(m)	0.250	0.375	0.525
Phi(rad)	0.075	0.113	0.158
Searching range in phi (rad)	0.075		0.091
(deg.)	4.316		5.205
(mm)	18.834		47.691
Search range in Z (mm)	66.667		80.000
Total Layer Area (m^2)	3.770		7.917
Number of Hits in searching area	0.999		1.446



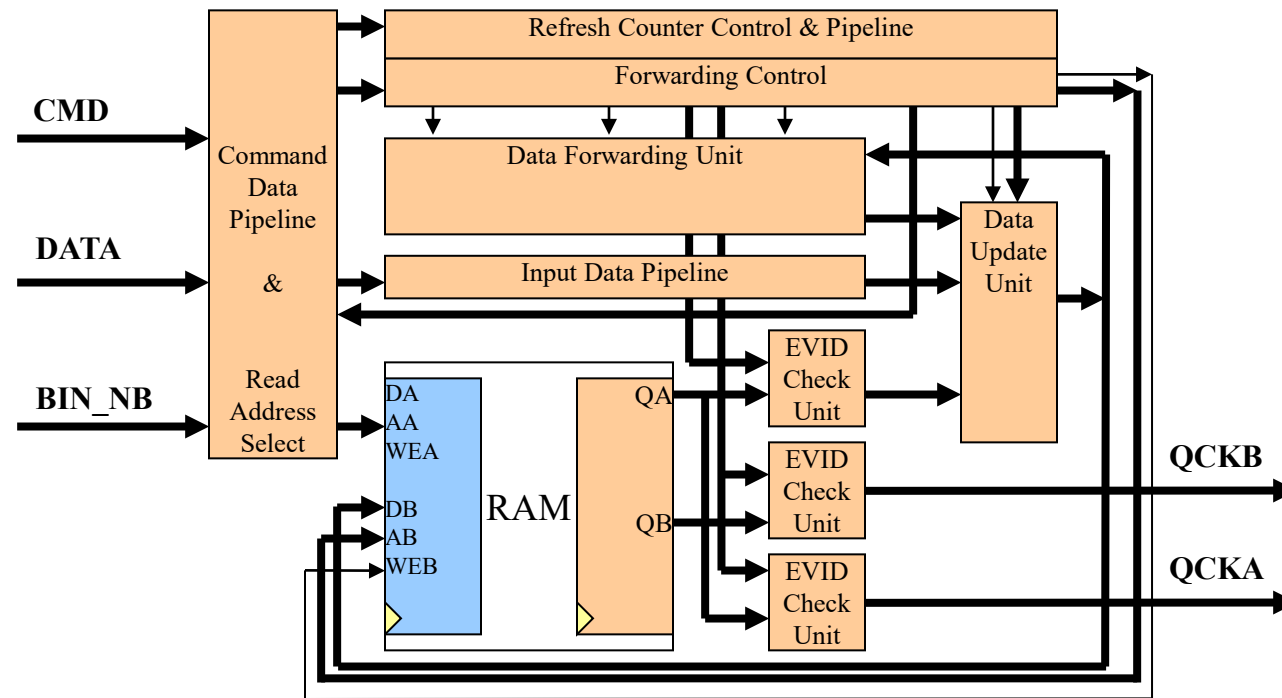
- For almost all 3000 stubs in Layer 2, on average a hit can be found in Layer 1 or Layer 3.
- About 90% of these pairs are not part of true tracks.
- When 3000 stubs/layer are sent to the stub pairing stage, there will be about 3000 pairs output to the tracking stage without significant reduction.

Hits in the Hugh Transform Cells



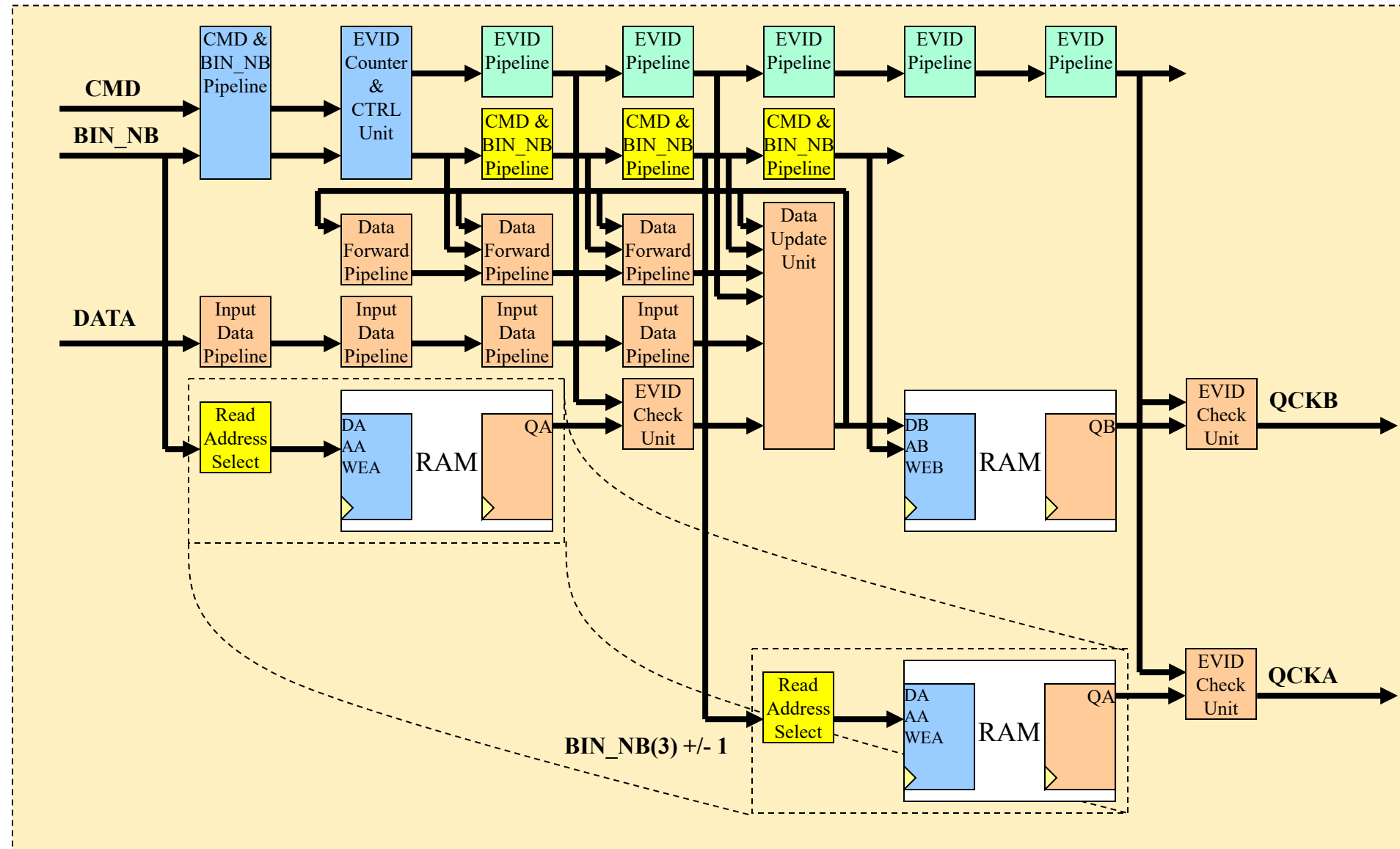
- A hit is written into 10 z_0 RAM blocks.
- The hits in Layer 1 and Layer 3 have different angles.
- The z coordinate of Layer 2, z_{375} is used to address all RAM block simultaneously.
- If an intersection of Layer 1 and Layer 3 hits are met, it might be a potential track segment.

Register-Like Block RAM Block Diagram



■ The

Detailed Block Diagram



- Boundary coverage is also supported during reading operations.