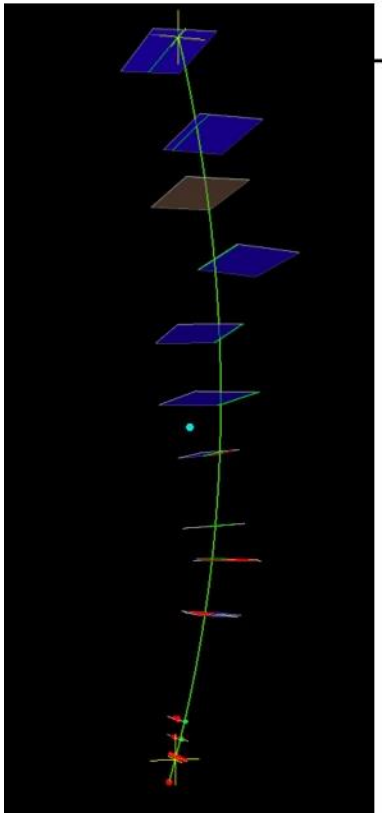
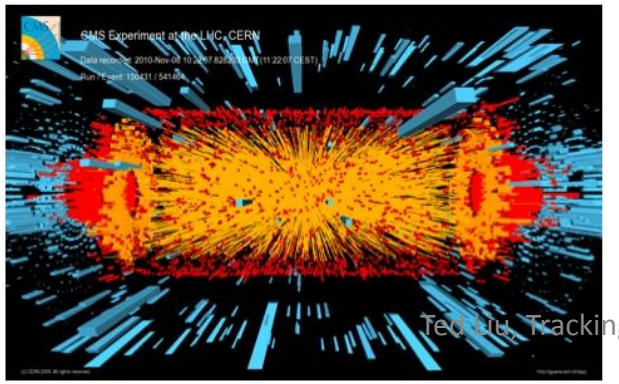
System Architecture and Vertical Slice Demonstration for CMS L1 Silicon-based Tracking Trigger

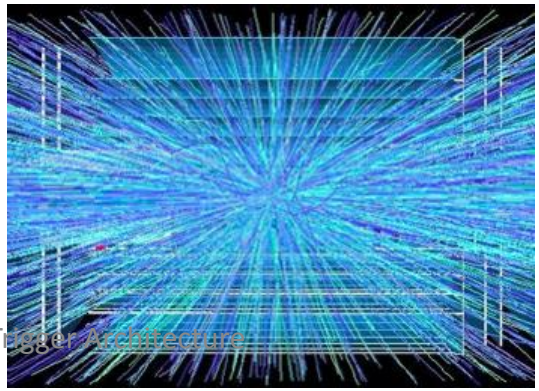


Ted Liu (FNAL)

May 14, 2014

Workshop on Intelligent Tracker (WIT), Upenn





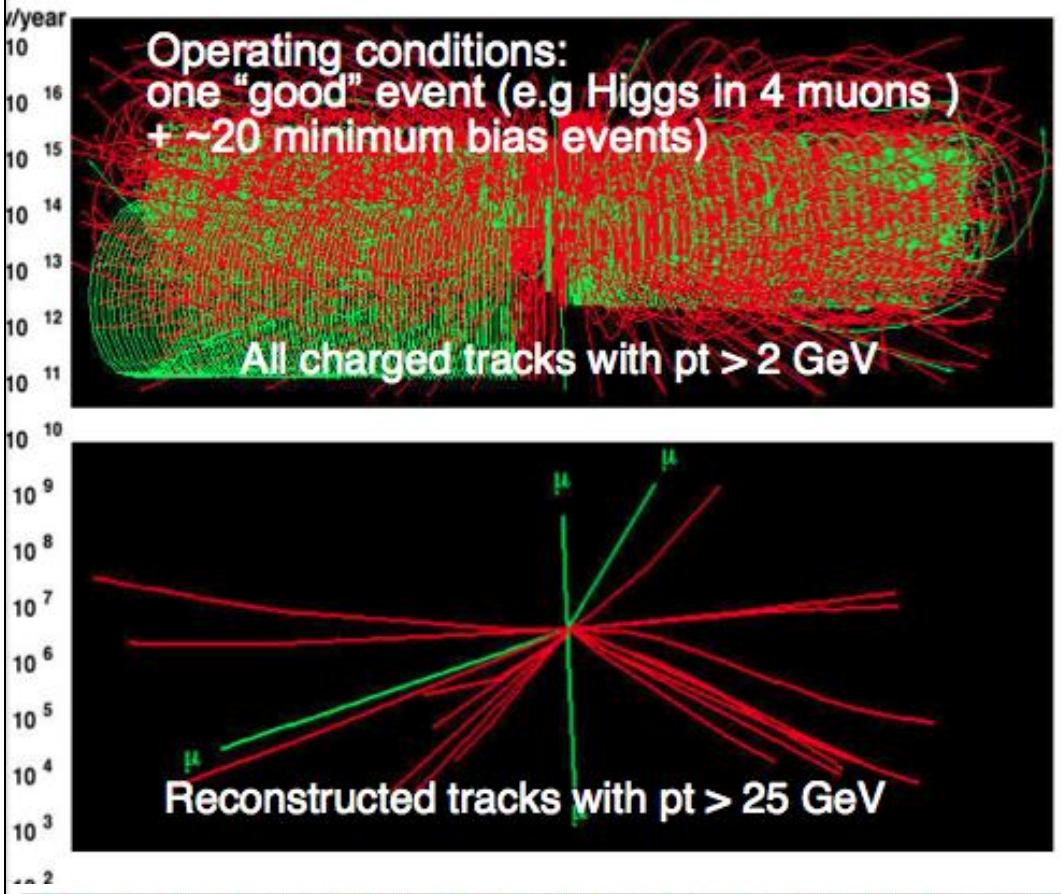
CMS L1 Tracking Trigger:

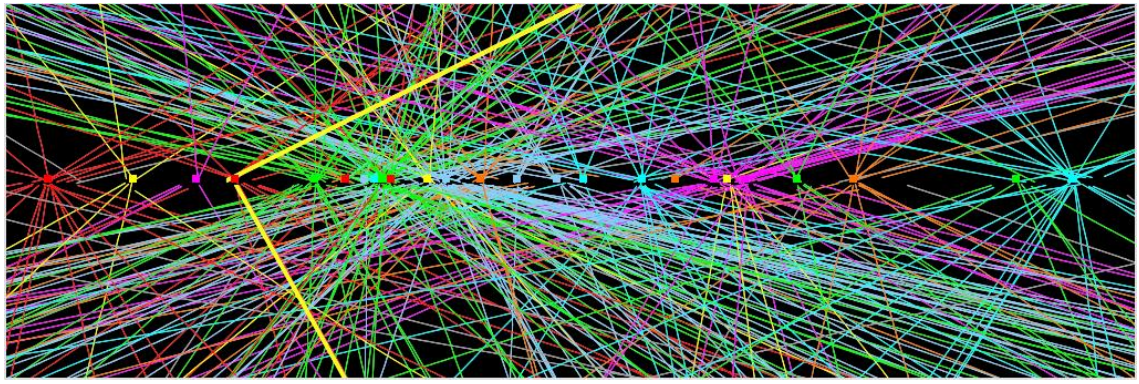
Will need to reconstruct charged particle trajectories "on-the-fly" for every beam crossing (25 ns, or 40 Million beam crossings per second), from an ocean of input data (bandwidth required to transfer up to ~ 50-100Tb/s)

This requires extremely fast high bandwidth data communication as well as massive pattern recognition power,
with lots known patterns to be compared against the multiple input data streams simultaneously with near zero latency (~ few μs)

This is challenging!

Pileup at HL-HC: > ~ 140 (only 20 shown here)



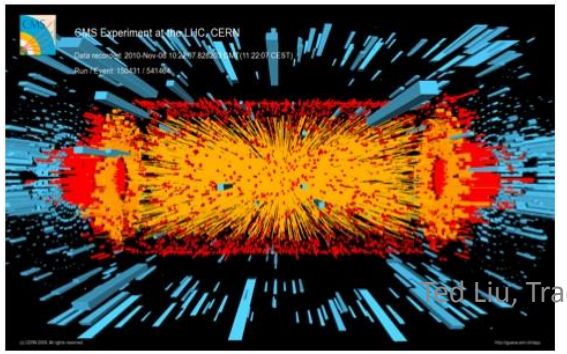


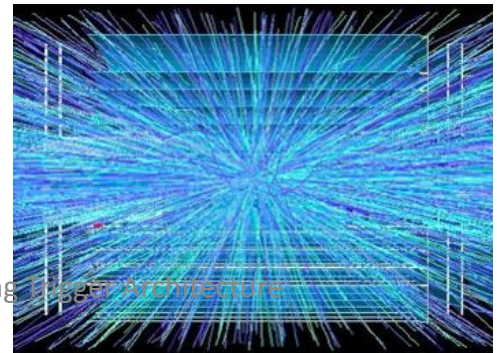
High Performance Computing

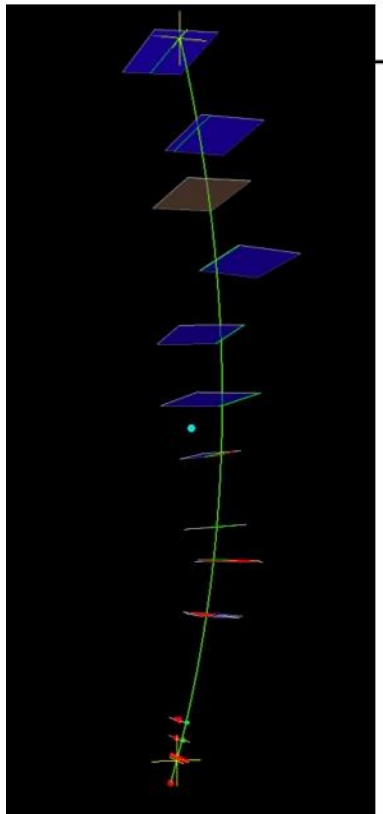
→from US "Report to the President and Congress" by President's Council of Advisors on Science and Technology, Dec. 2010 (page 65)

- Compute-intensive
  - massively parallel computation involving *very large number of processing elements*;
- Communication-intensive
  - *high-speed transfer of data* among processing elements;
- Data-intensive
  - *high-speed manipulation of* very large quantities of data

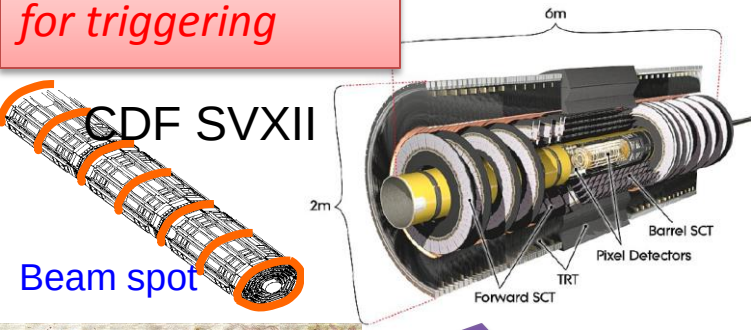
HL-LHC L1 Tracking Trigger is High Performance Computing
(Non-von Neumann approach)
but with very Low Latency and in Real Time
HL-LHC requires the most advanced Real Time processing technology

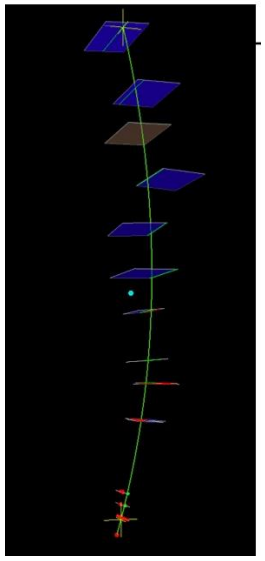






Detector design for triggering





Silicon Based Tracking Trigger at Hadron Colliders

Data transfer

Data formatting

Partition detector into trigger towers/sectors

Pick your favorite method:

Associative Memory (AM) Approach

Hough Transformation

tracklet-based

Adaptive Pattern Recognition

Biology Inspired …

your choice here…

Pattern Recognition

Finer pattern recognition

Track Fitting

FPGA vs GPU vs CPU

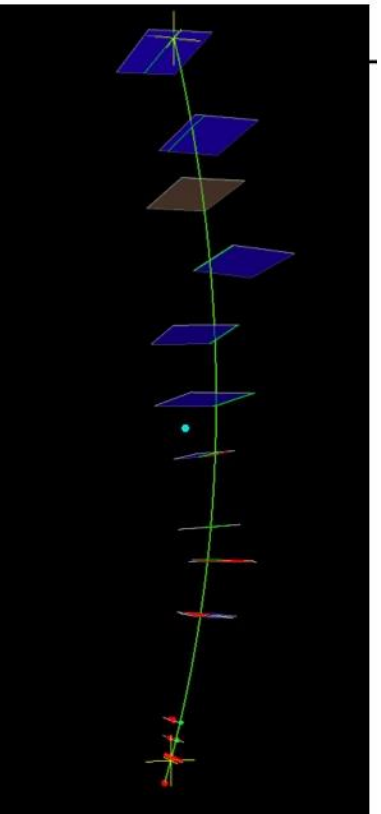
HLT

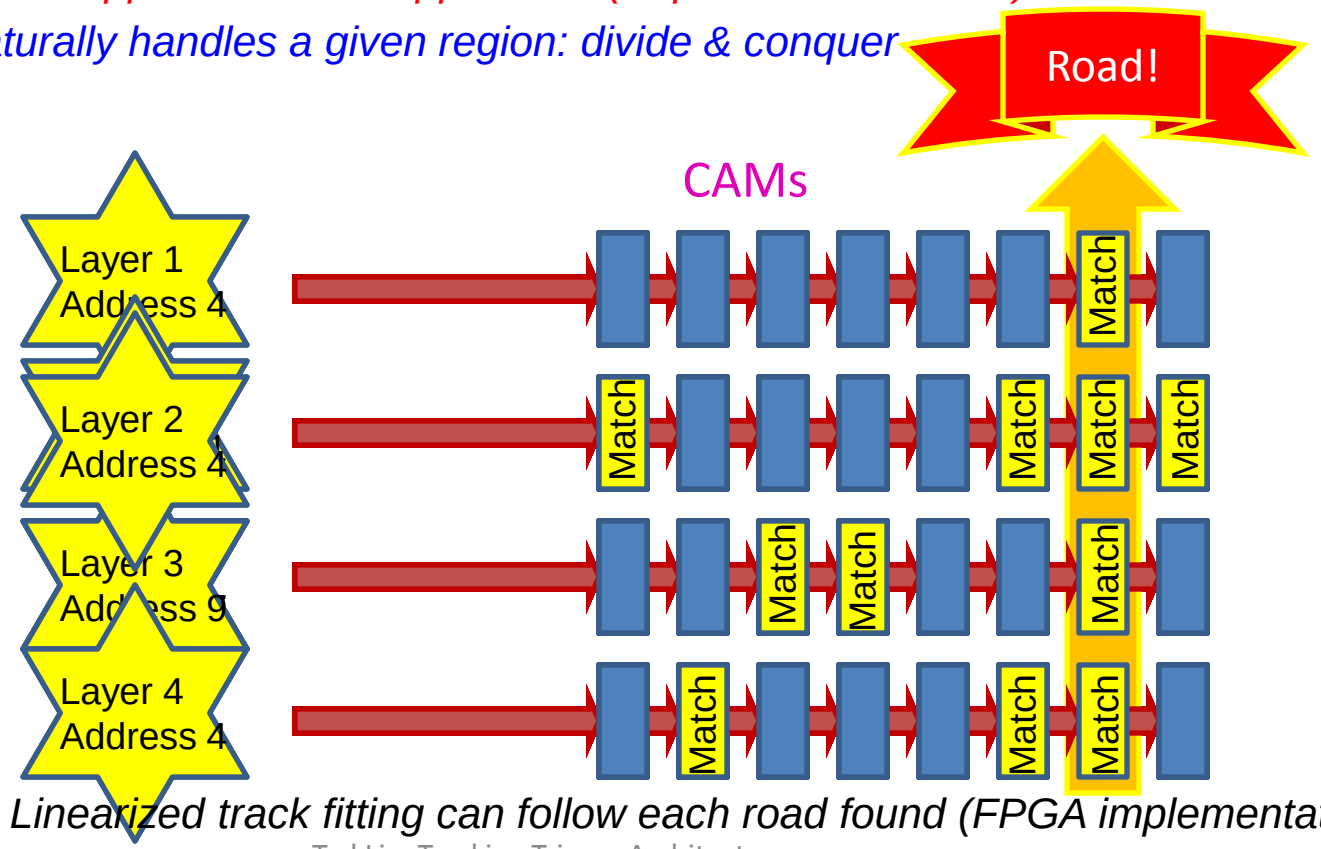
Will use Associative Memory approach as an example in this talk

The AM approach

- Pattern Recognition Associative Memory
  - Based on *CAM cells to match and majority logic to associate* hits in different detector layers to a set of pre-determined hit patterns (simple working unit, yet massively parallel)
  - *Pattern Recognition finishes right after all hits arrive (fast data delivery important)*
  - *Potentially good approach for L1 application (require custom ASIC)*

A PR engine naturally handles a given region: divide & conquer





Linearized track fitting can follow each road found (FPGA implementation)

PRAM+TF/FPGA

- The PRAM stage:
  - Massive parallel processing to tackle the intrinsically complex combinatorics of track finding algorithms, avoiding the typical power law dependance of execution time on occupancy
  - and solving the pattern recognition in times roughly proportional to the number of hits, making the downstream task much easier
  - Usually requires custom ASIC

- The Track Fitting stage (FPGA) after AM:
  - Finer pattern recognition
  - Examples: linearized track fitting, Hough transform, Retina … etc
  - The more powerful the AM stage, the less demand on TF/FPGA
  - The more powerful the TF/FPGA, the less demand on AM

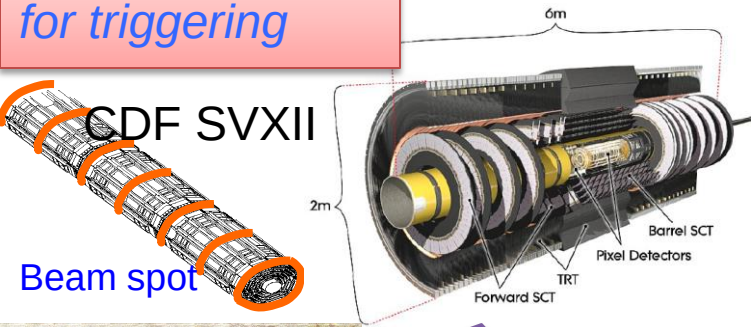
- *Some proposed algorithms do not have the AM stage*
  - Example: Track-let based track finding + linearized track fitting (see next talk by Jorge)

Comparison of the two approaches being currently explored at CMS

	AM + TF approach	Tracklet +TF approach
advantages	• Proven approach for silicon based track finding • AM pattern recognition algorithm: simple, fast and flexible • Can combine with different track fitting algorithms	• New approach for *hardware* silicon based track finding • Software simulation promising • *Can be implemented in FPGA in principle: no need for custom designed chips*
challenges	• *Requires custom ASIC: high performance AMchip* • Track Fitting *speed* in FPGA to be demonstrated for L1 • New architecture (see below)	• *It is new* • *Feasibility to be demonstrated in hardware (FPGA)* • *New architecture*
	Common:* *Fast* *Data delivery/sharing to Pattern Recognition Engines	

The rest of the talk …

Detector design for triggering



Silicon Based Tracking Trigger at Hadron Colliders

Partition detector into trigger towers/sectors

Pick your favorite method:

Associative Memory (AM) Approach

Hough Transformation

tracklet-based

Adaptive Pattern Recognition

Biology Inspired …

your choice here…

Data transfer

Data formatting

Pattern Recognition

Finer pattern recognition

Track Fitting

- *In this talk (focus on off-detector)*

(1) Pattern Recognition + Track Fitting options

-- AM + linearized track fitting (FPGA): traditional

-- AM + Hough transform (FPGA): new, being studied

-- Tracklet-base approach (FPGA): new, being studied

FPGA vs GPU vs CPU

(2) Data Formatting and System Architecture:

-- FPGA based Full-mesh enabled ATCA approach (Pulsar 2)

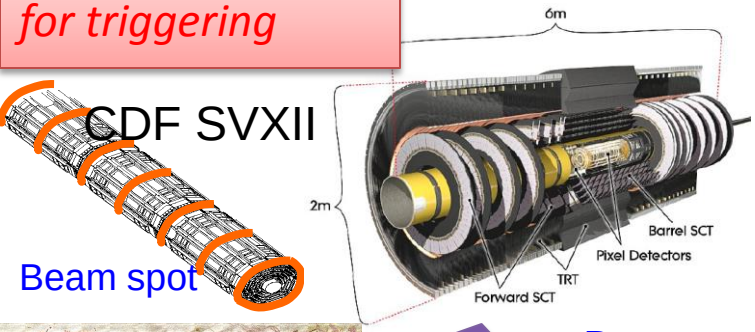
-- MicroTCA based (MP7): developed for CMS L1 CAL trigger upgrade (Geoff's talk)

HLT

(3) Summary (Vertical Slice Demonstration)

Detector design for triggering

Silicon Based Tracking Trigger at Hadron Colliders



Data transfer

Data formatting

Partition detector into trigger towers/sectors

Pick your favorite method:

Associative Memory (AM) Approach

Hough Transformation

tracklet-based

Adaptive Pattern Recognition

Biology Inspired …

your choice here…

Pattern Recognition

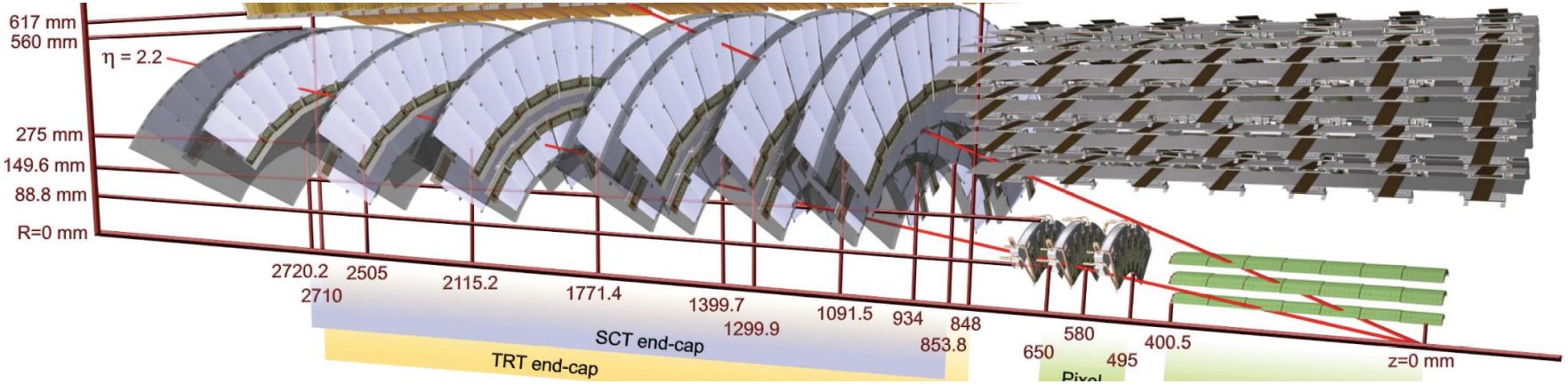
Finer pattern recognition

Track Fitting

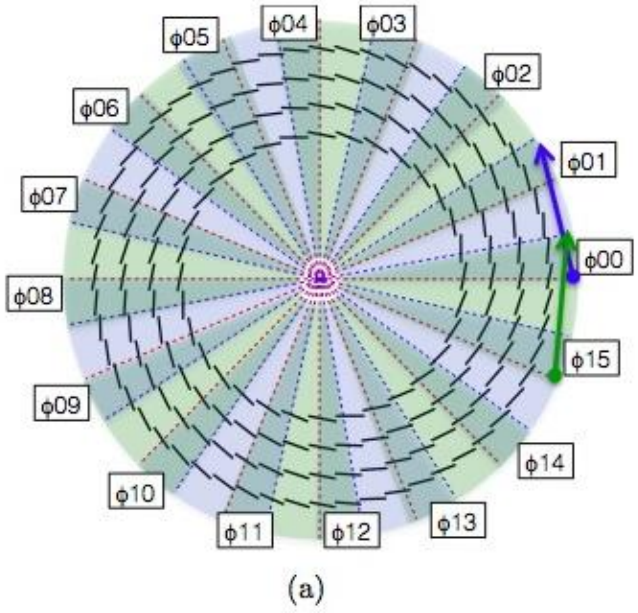
Need massive parallelism: to process in parallel
different regions of the detector for
the same beam crossing (regional multiplexing, L1&L2),
and different crossings (time multiplexing, L1).
Requires high bandwidth, low latency, and flexible real time communication

HLT

First take a look at: Data formatting challenges for Atlas FTK at L2

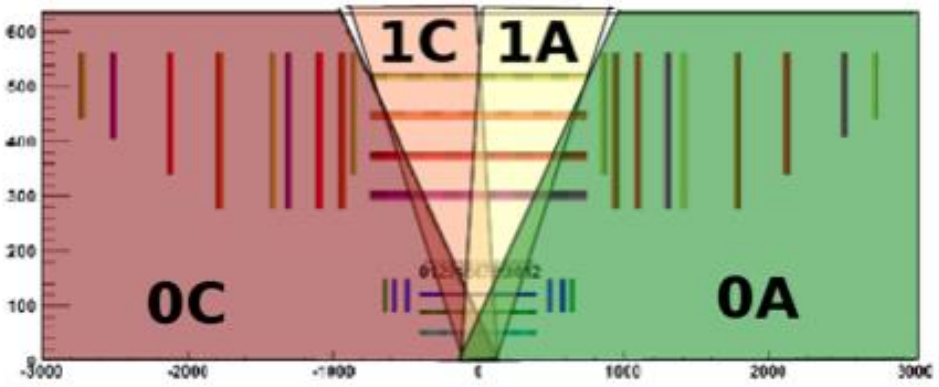


Input data from all silicon detector modules has to be formatted into 64 η-φ trigger towers after reformatting and sharing, ready for downstream pattern recognition

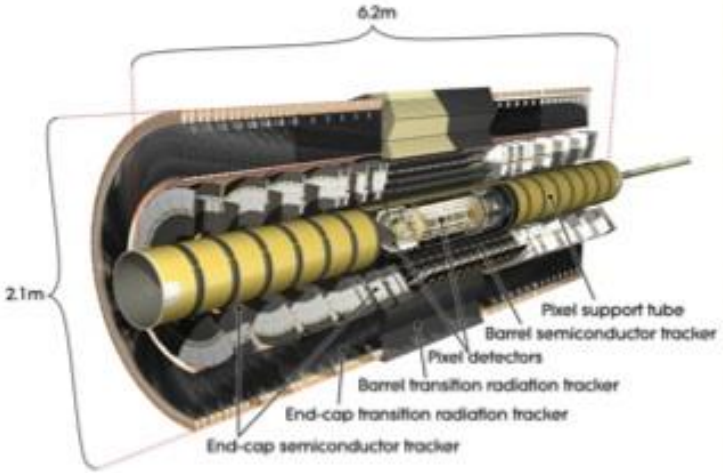


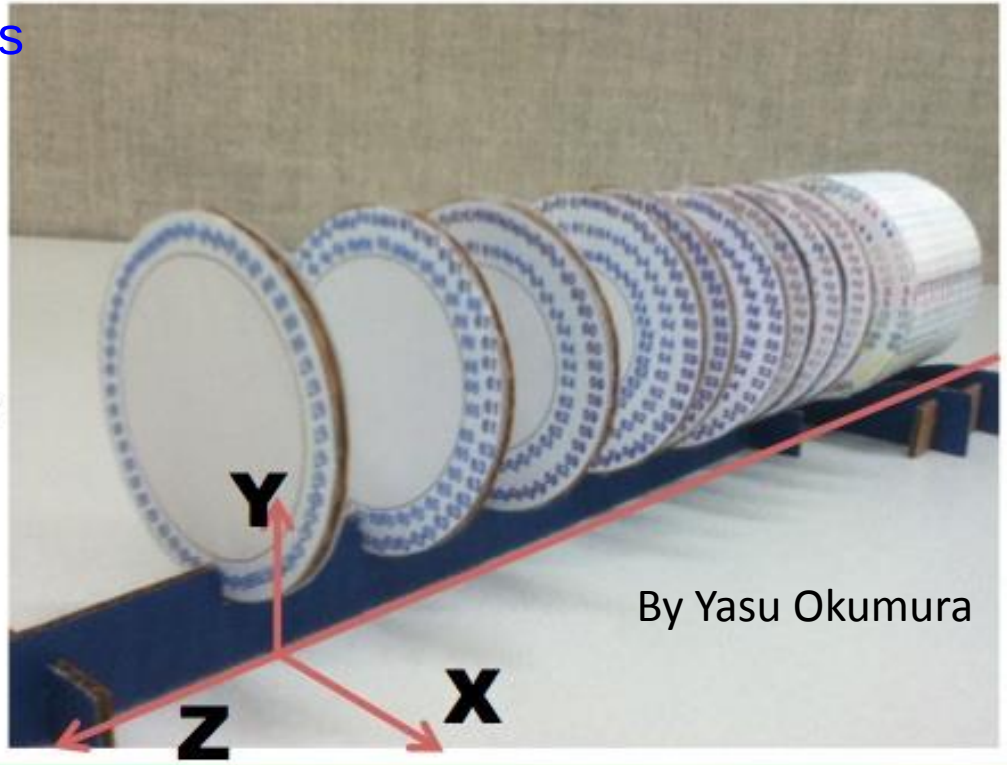
(a)

4 in η



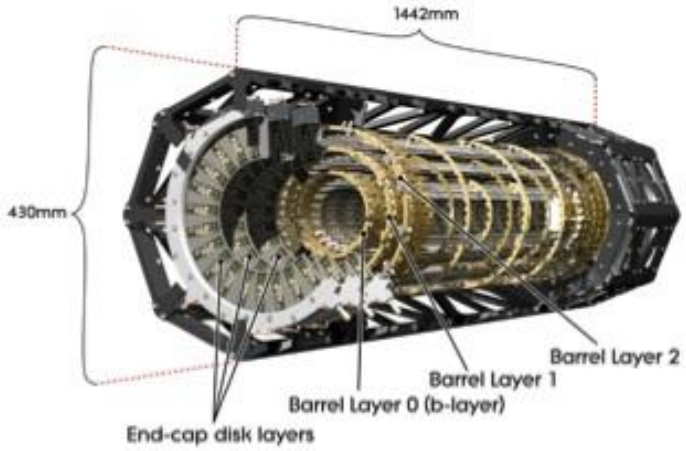
FTK Data Formatting Challenges

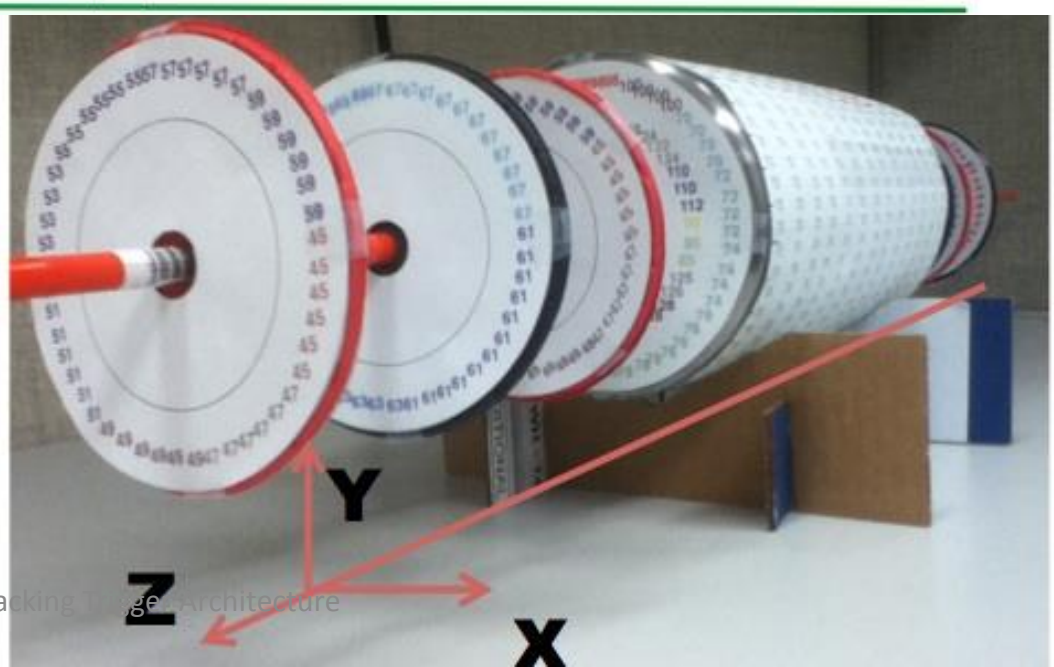


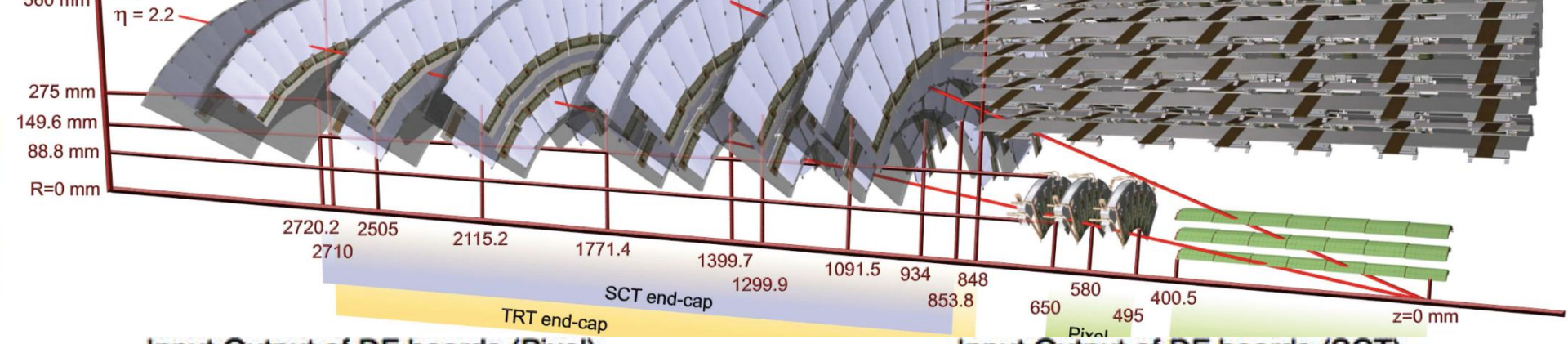


By Yasu Okumura

3D paper models at Fermilab with detailed ROD ID mapping, to help us understand the issues

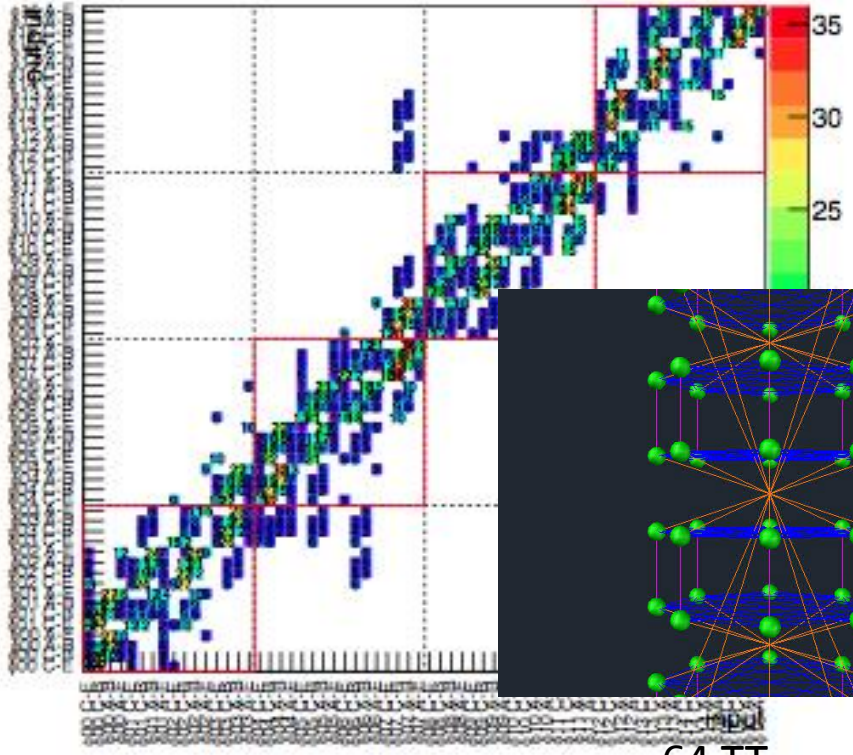






64 TT

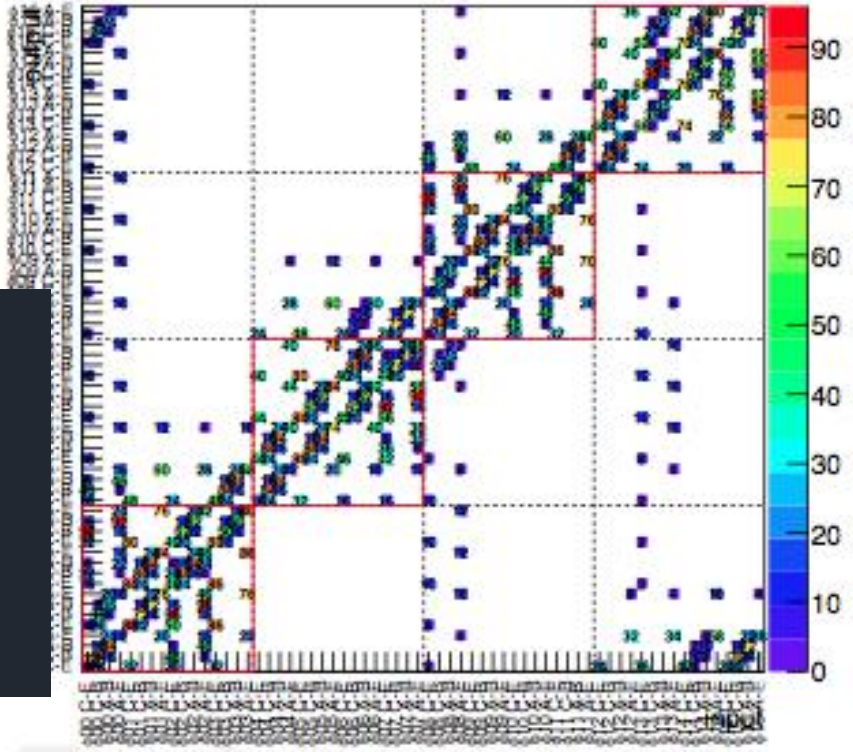
Input Output of DF boards (Pixel)



(a)

64 TT

Input Output of DF boards (SCT)



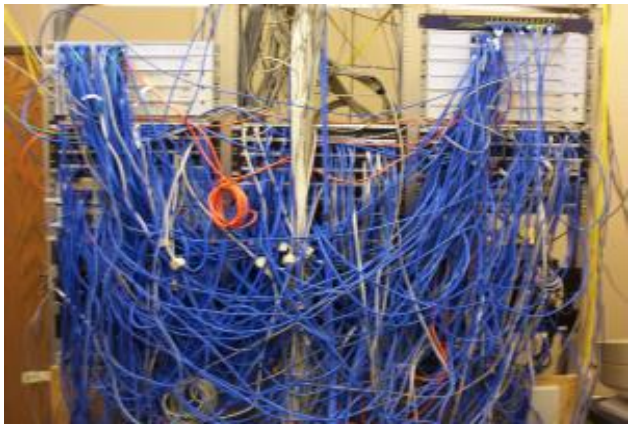
(b)

Detailed beam data analysis *using exact module/ROD cable mapping to trigger towers*

From “Data Formatter Design Specification”, Fermilab-TM-2553-E-PPD.

Data sharing/switching techniques

Traditional data sharing

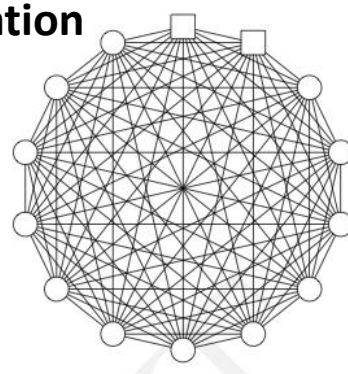


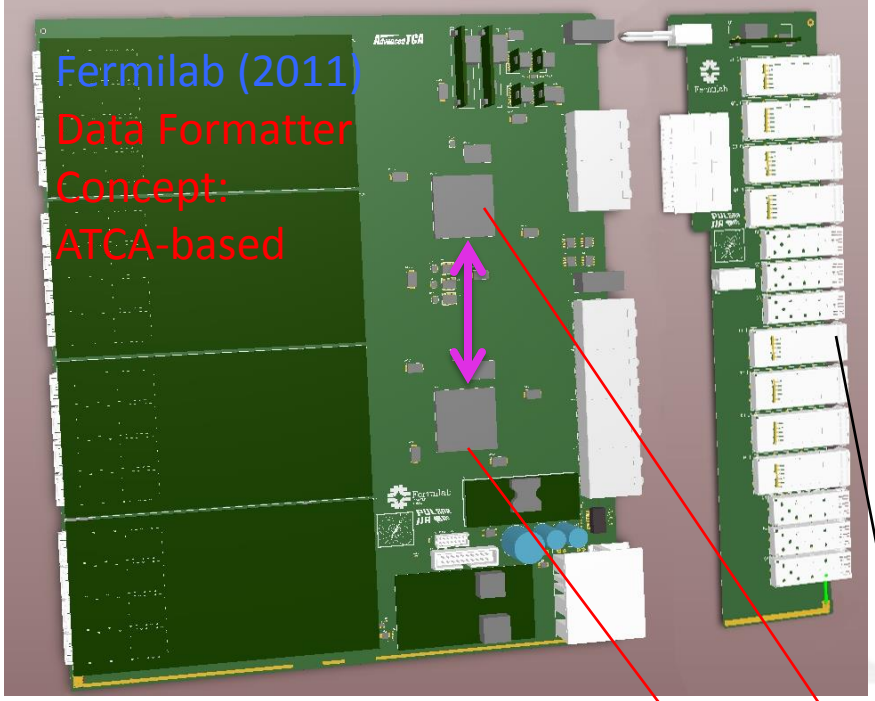
Data sharing with ATCA

high speed full-mesh backplane

~ 40 Gbps point to point

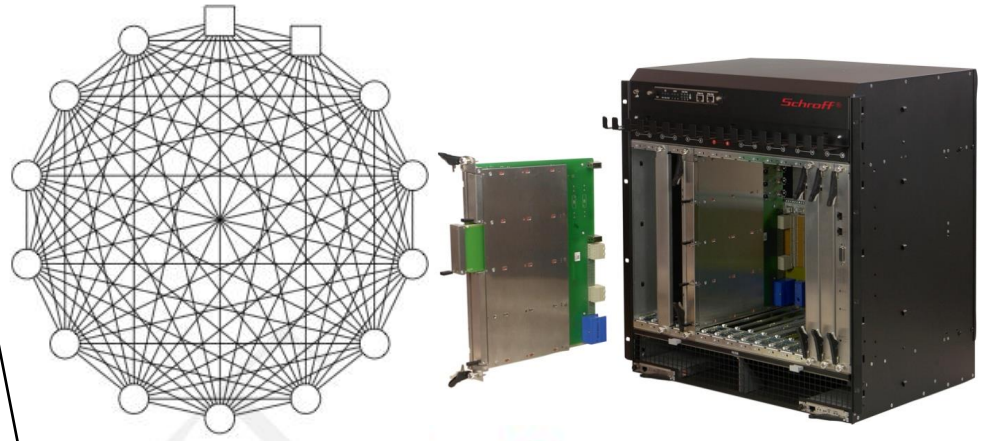
Advanced
Telecommunication
Computing
Architecture

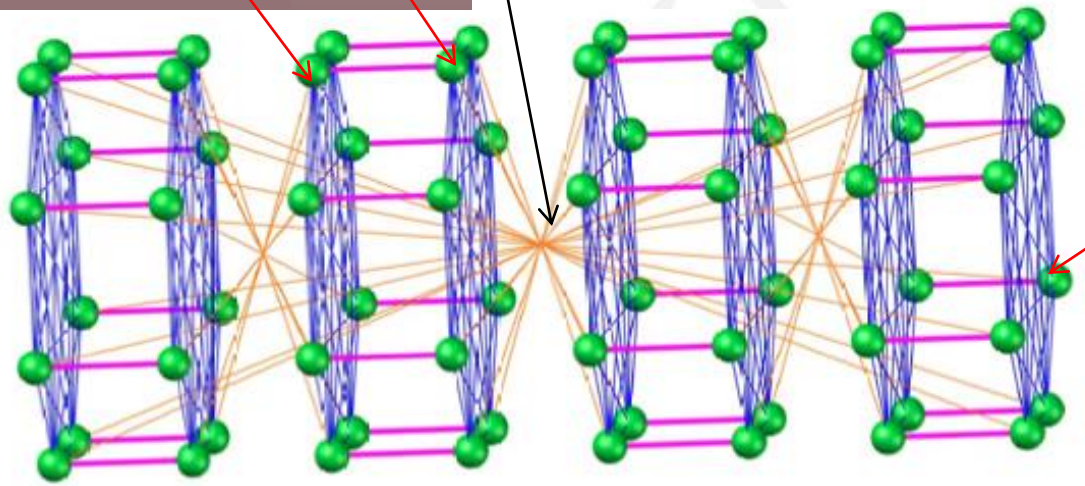


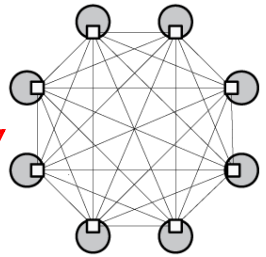


How to use ATCA backplane for data sharing

(1 FPGA per trigger tower, 2 trigger towers per board, 32 boards needed over 4 ATCA shelves)







ATCA Full-mesh backplane

Only data sharing links shown, not inputs/outputs

Figure 8: A 3D representation of FPGA interconnects in the Data Formatter system. 64 FPGAs (green) are connected through the ATCA backplane Fabric Interface (blue), local buses (purple) and inter-shelf links (orange). Each FPGA uses one inter-shelf link. This

Appendix P Unconstrained Data Volume Study

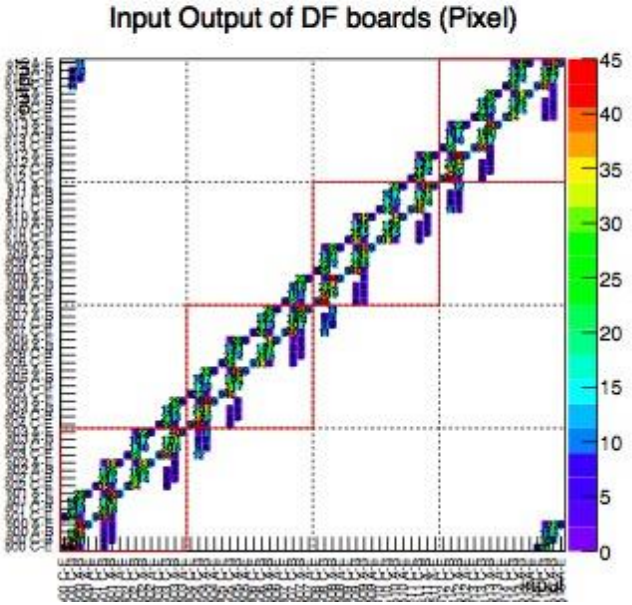
As previously mentioned the inner detector readout system was not originally designed for a track trigger. Modules were connected to RODs to minimize data rates and balance bandwidth. In this section we consider Data Formatter performance assuming an idealized module-ROD and ROD-DF mapping.

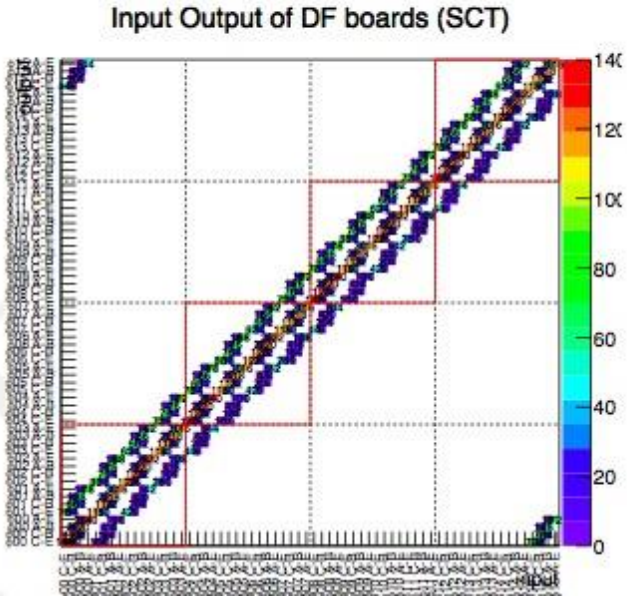
The band would be much cleaner without ROD constraints:

P.1 Data Sharing

The cabling was done to optimize for DAQ readout, not trigger

Refer to Figure 15 to compare these idealized results with the "real world" module-ROD cabling constraints.





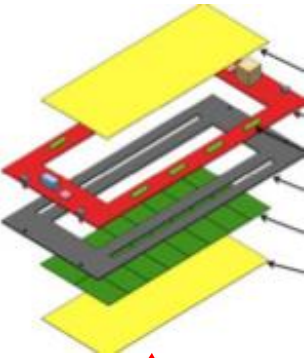
From "Data Formatter Design Specification", Fermilab-TM-2553-E-PPD, page 78.
Available at: http://www-ppd.fnal.gov/EEDOffice-w/Projects/ATCA/
(Pulsar IIa design spec)

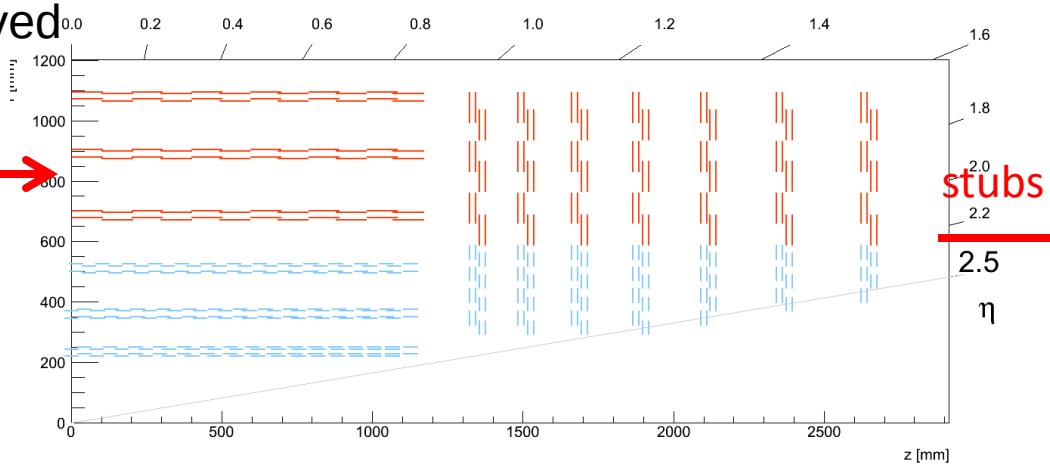
CMS: Module design vs Tracker design vs Trigger processing

Module Design | Tracker Design | Track Finding

Pt stub finding reduce the data volume by ~ 10-20,
making it possible to transfer the data out for off-detector L1 track finding

~15K modules involved

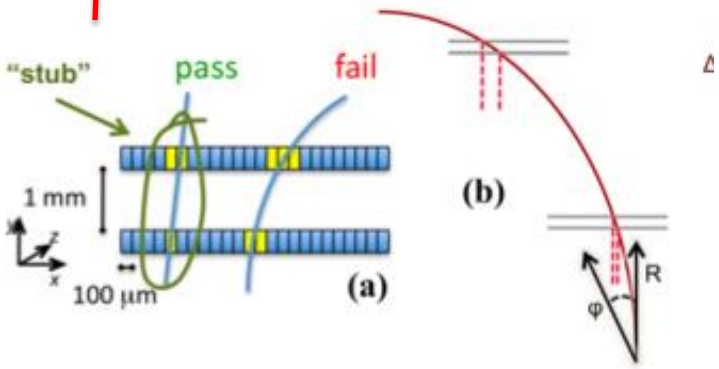




stubs

Back-end
Off-detector
Track Finding

CMS new tracker design



See Davide's talks earlier

Working assumptions:

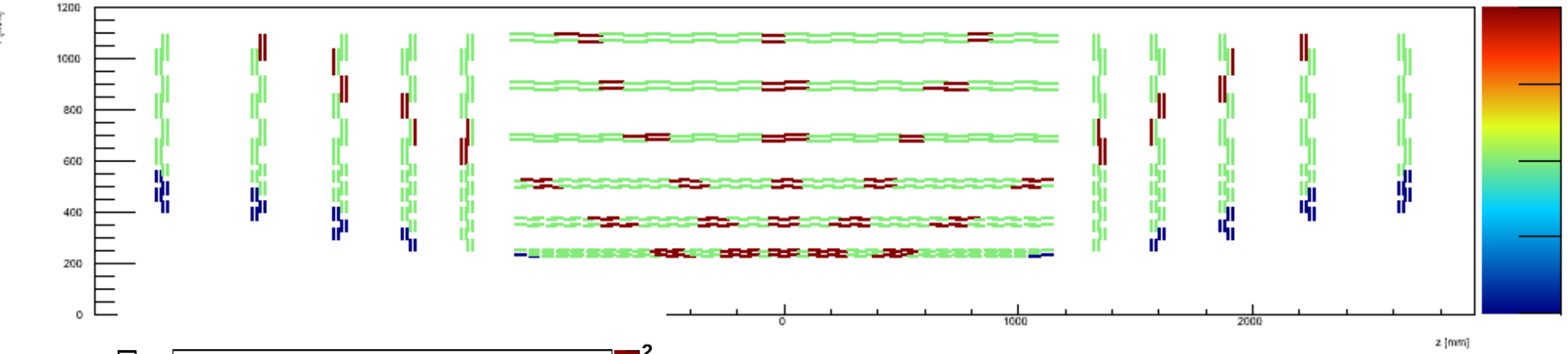
~15K fibers to trigger,
3.2Gbps payload each,
~ 50 Tbps input bandwidth

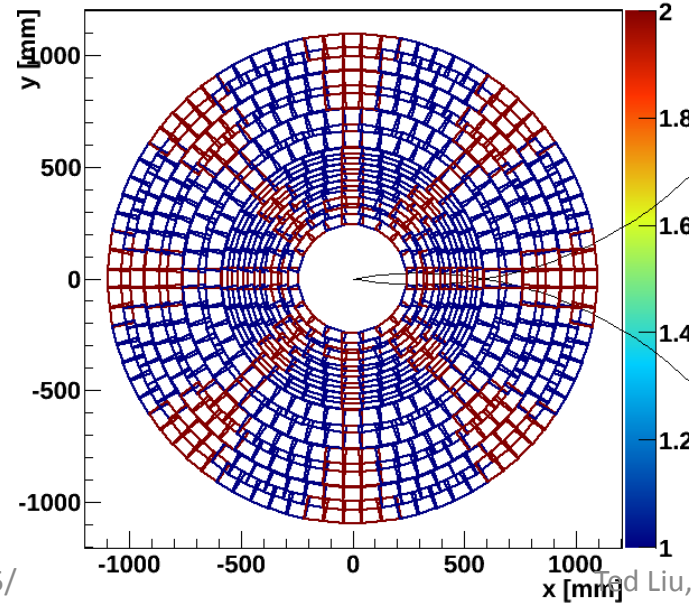
Block synchronous data transfer scheme …

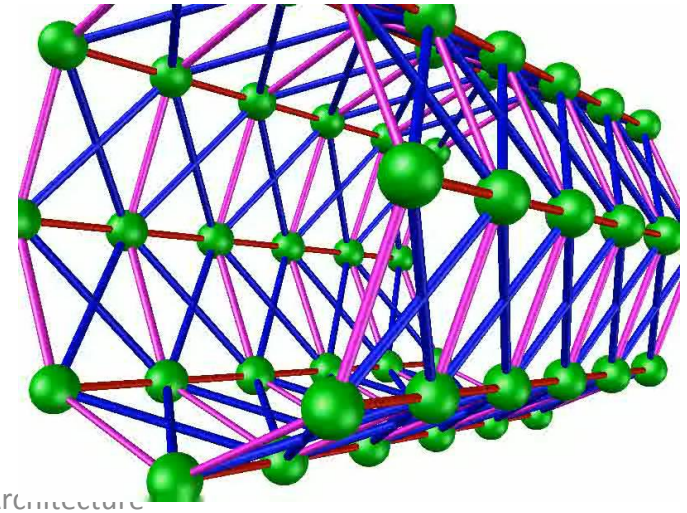
CMS Tracker Layout and Trigger Tower (6 in eta x 8 in phi)

- 15K modules

What we learned from FTK data formatting helps to understand CMS L1 case



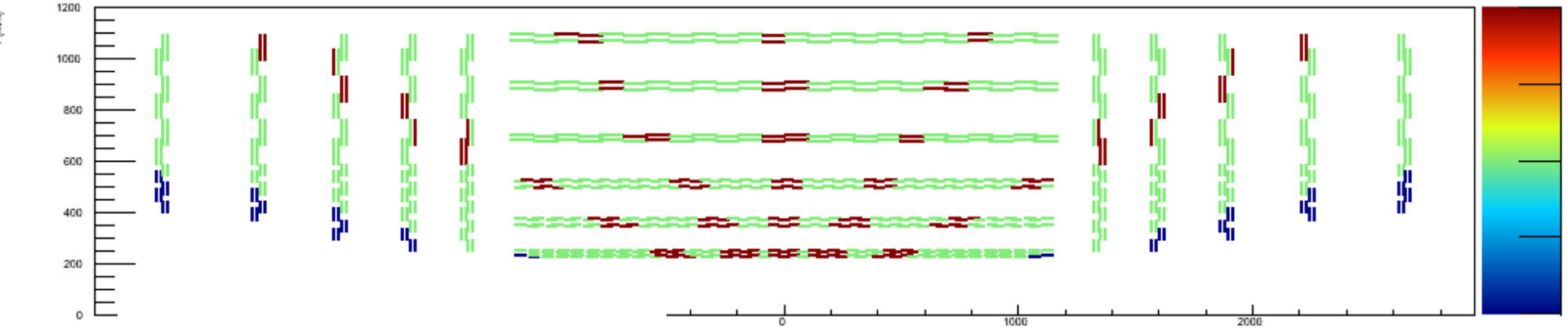


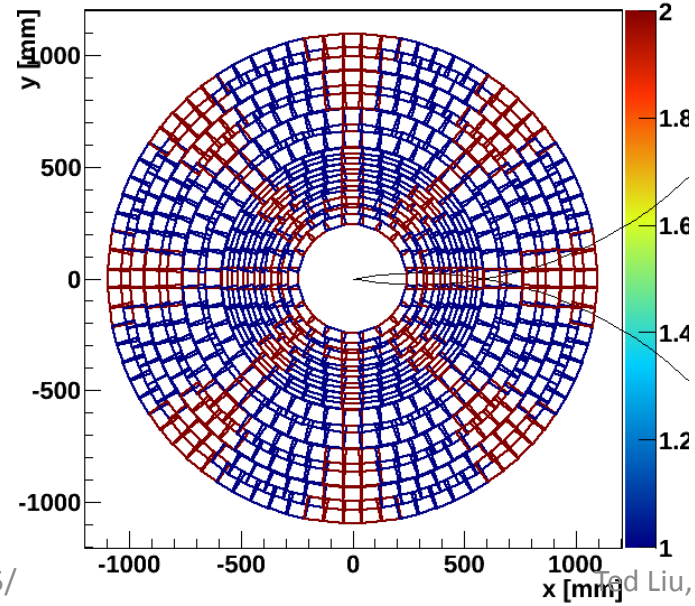


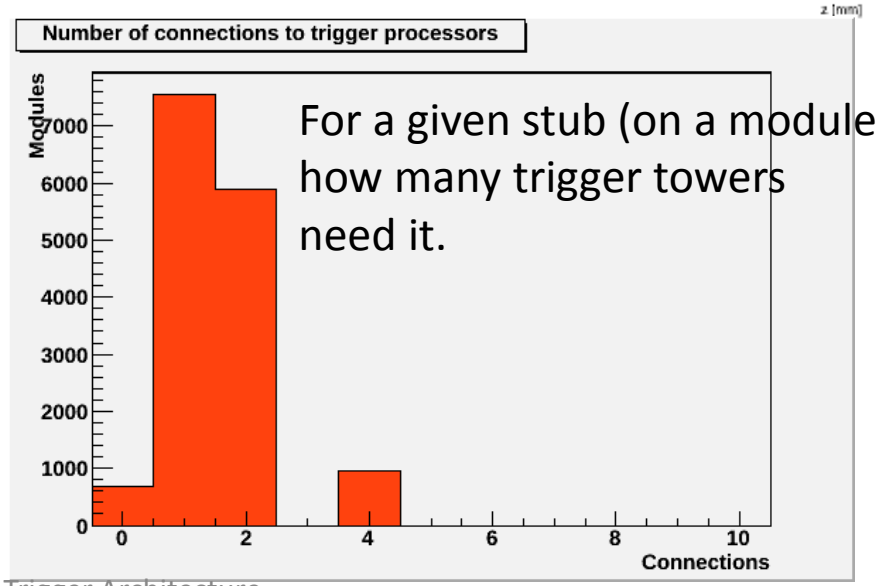
CMS Tracker Layout and Trigger Tower (6 in eta x 8 in phi)

- 15K modules/fibers

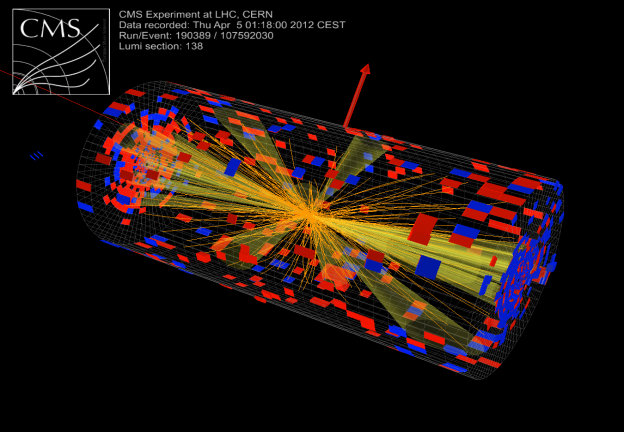
What we learned from FTK data formatting helps to understand CMS L1 case



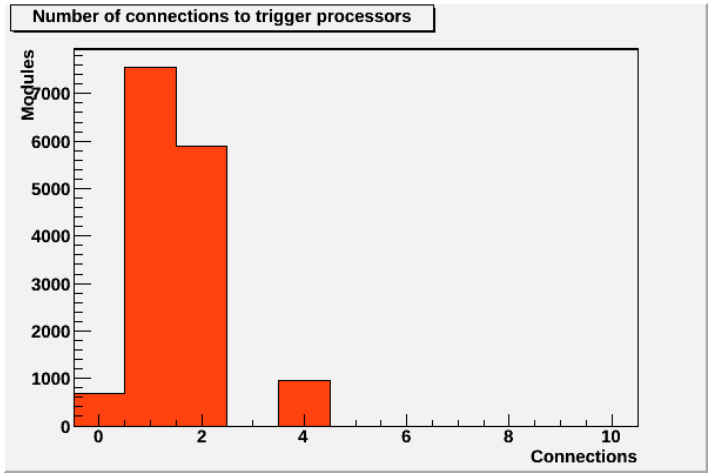


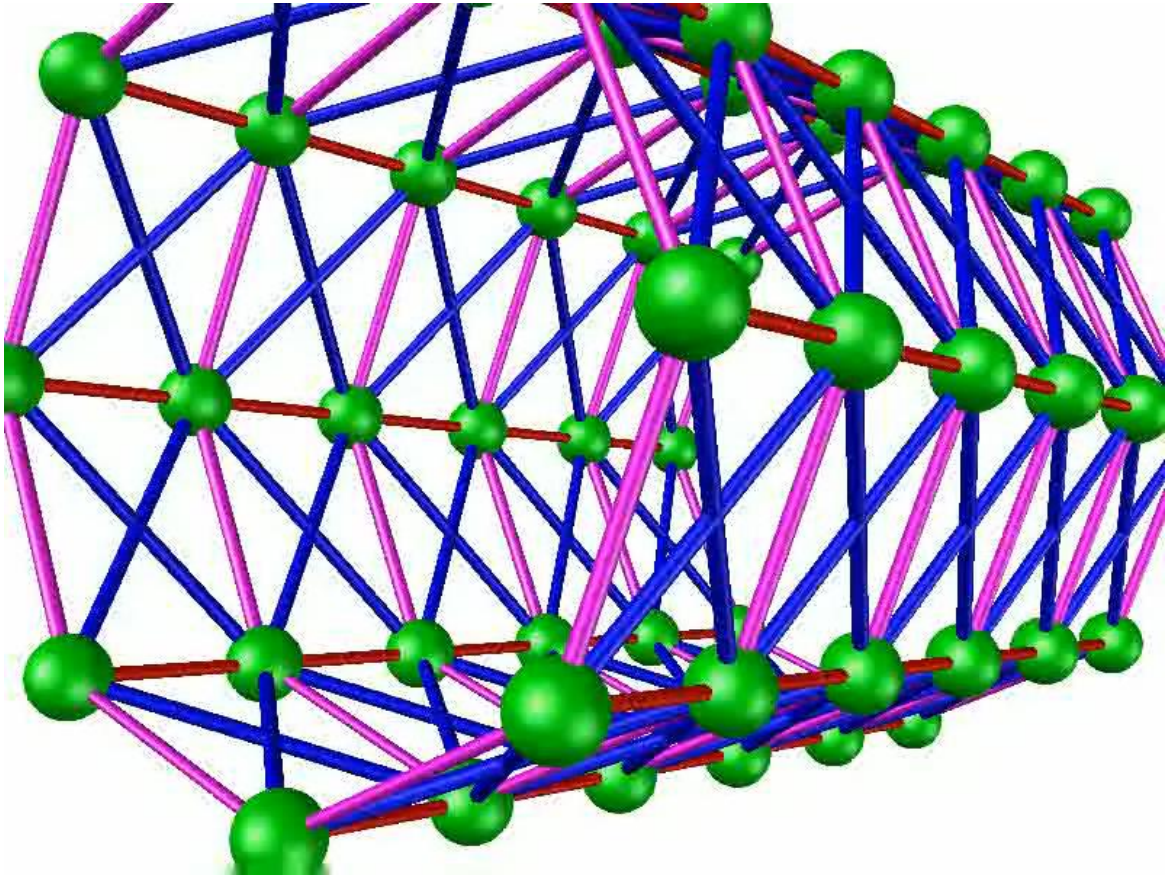


For a given stub (on a module), how many trigger towers need it.



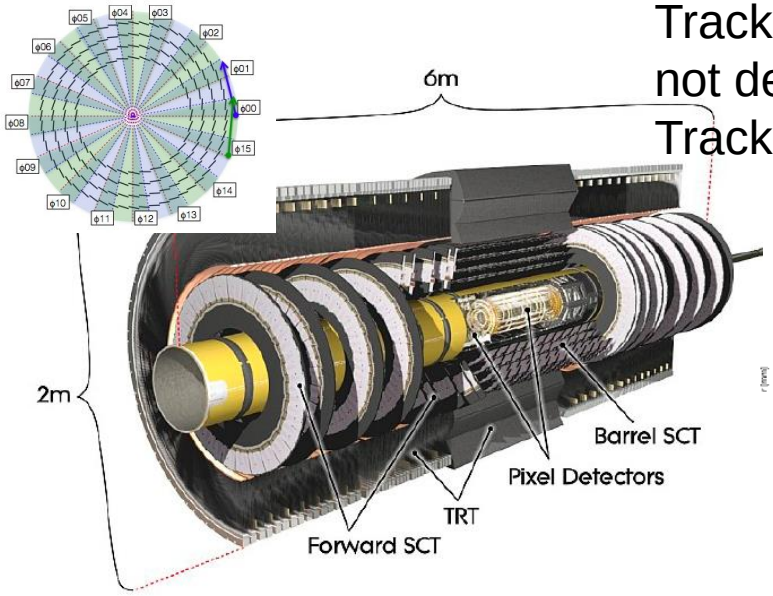
CMS L1 tracking trigger for Phase II: 6 (in eta) x8 (in phi) = 48 Trigger towers & their interconnections





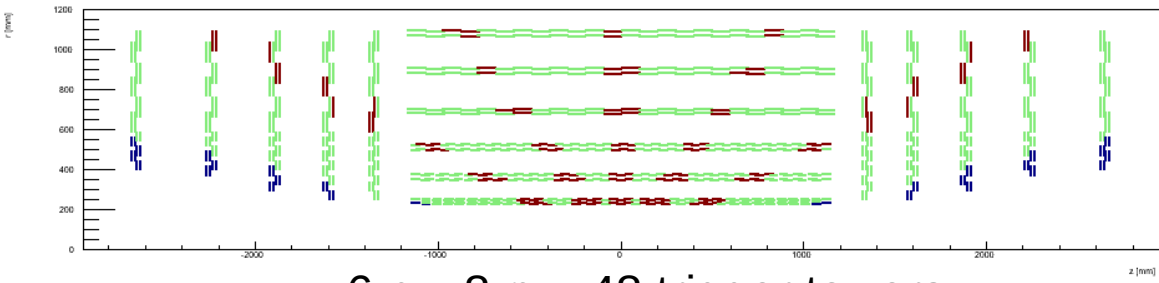
Data coming from a given trigger tower may need to be delivered to multiple trigger towers.
This happens,
when a stub comes from a detector element is close to the border Between trigger towers, due to the finite curvature of charged particles in the magnetic field and finite size of the beam luminous region along the beam axis.

Comparison: ATLAS L2 FTK and CMS L1 Track Trigger



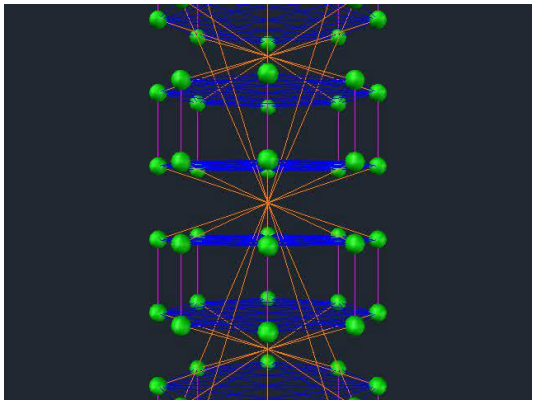
Tracker/cabling not designed for Tracking Trigger

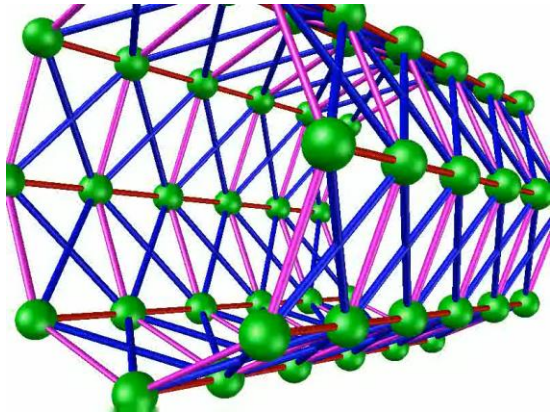
Tracker/Cabling designed for Tracking Trigger



4 η x 16 φ = 64 trigger towers

6 η x 8 φ = 48 trigger towers

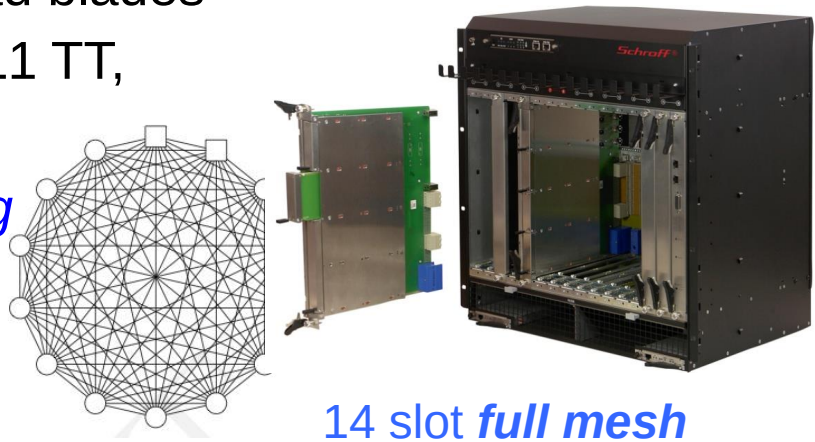




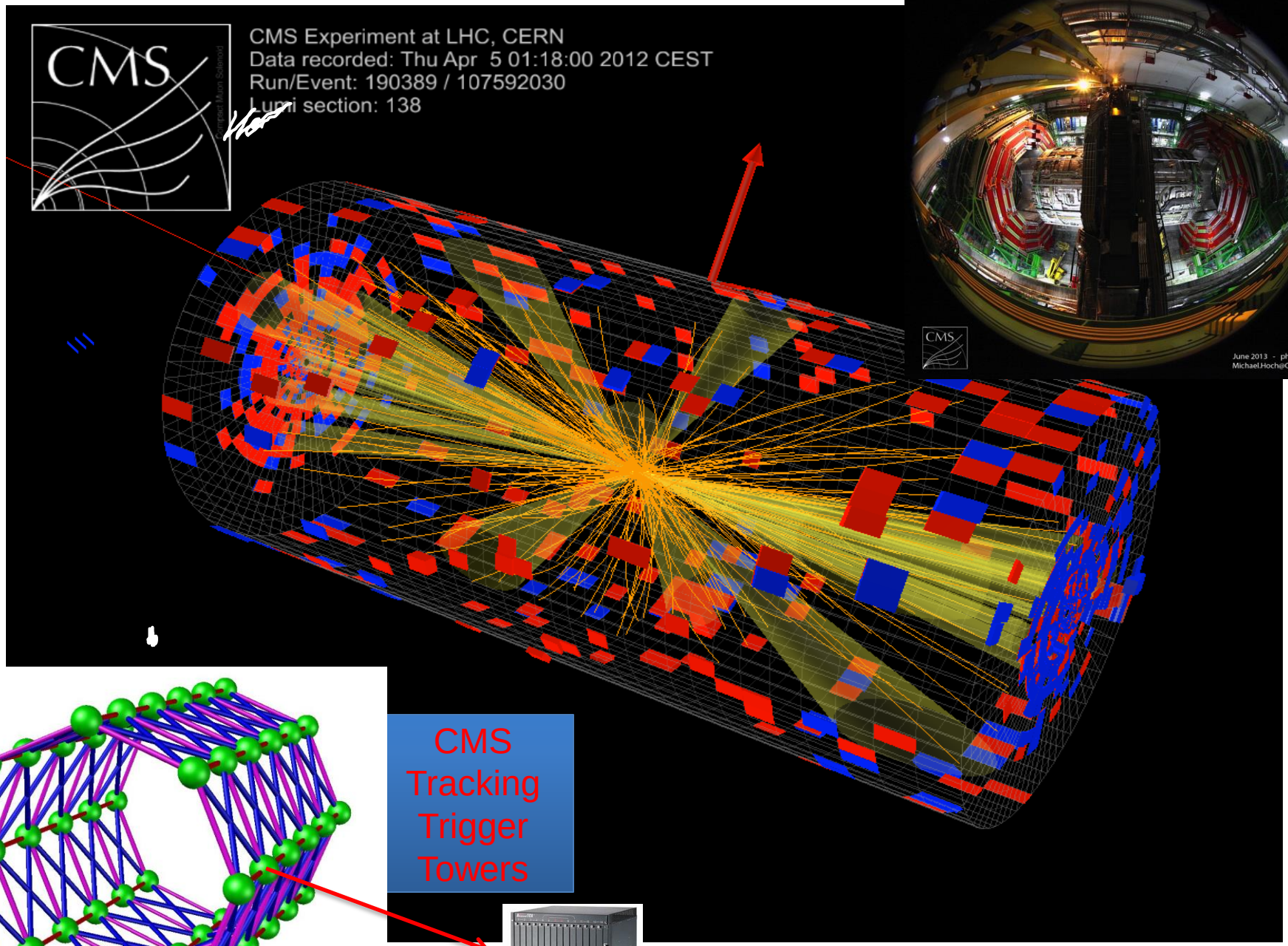
This leads to new Architecture….

General considerations for the tower processor Platform for silicon based tracking trigger system

- The tower processor platform must support large numbers of fiber transceivers, used for receiving input links and data sharing
- A flexible, high bandwidth backplane is desirable to quickly transfer data between boards
- The boards should be large enough to support pattern recognition engines and fiber connections, in a comfortable way
- ***A Full Mesh***, 14 slot ATCA shelf is a natural fit as the platform with 12 slots available for processor or payload blades
- This applies to both Atlas FTK and CMS L1 TT, but *architecturally they are very different:*
  Atlas FTK: full-mesh used for data sharing
  CMS L1 TT: full-mesh mostly used for time-multiplexing



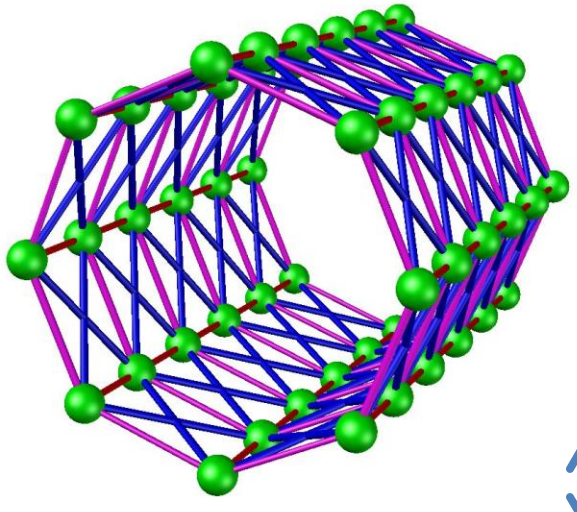
14 slot ***full mesh ATCA backplane***:

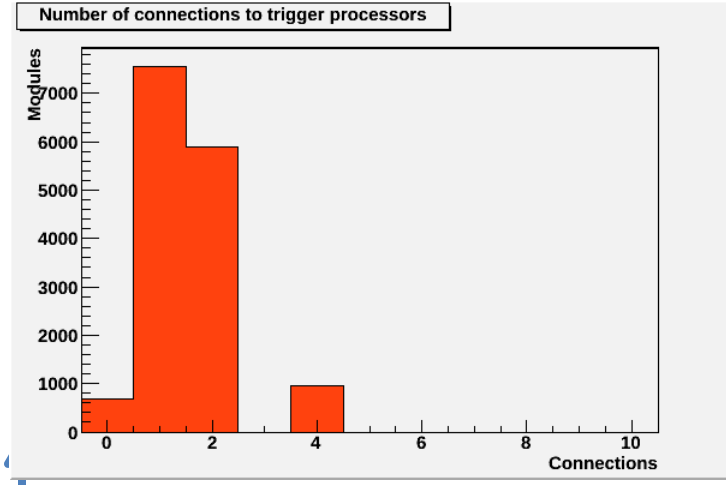


CMS Tracking Trigger Towers

ATCA

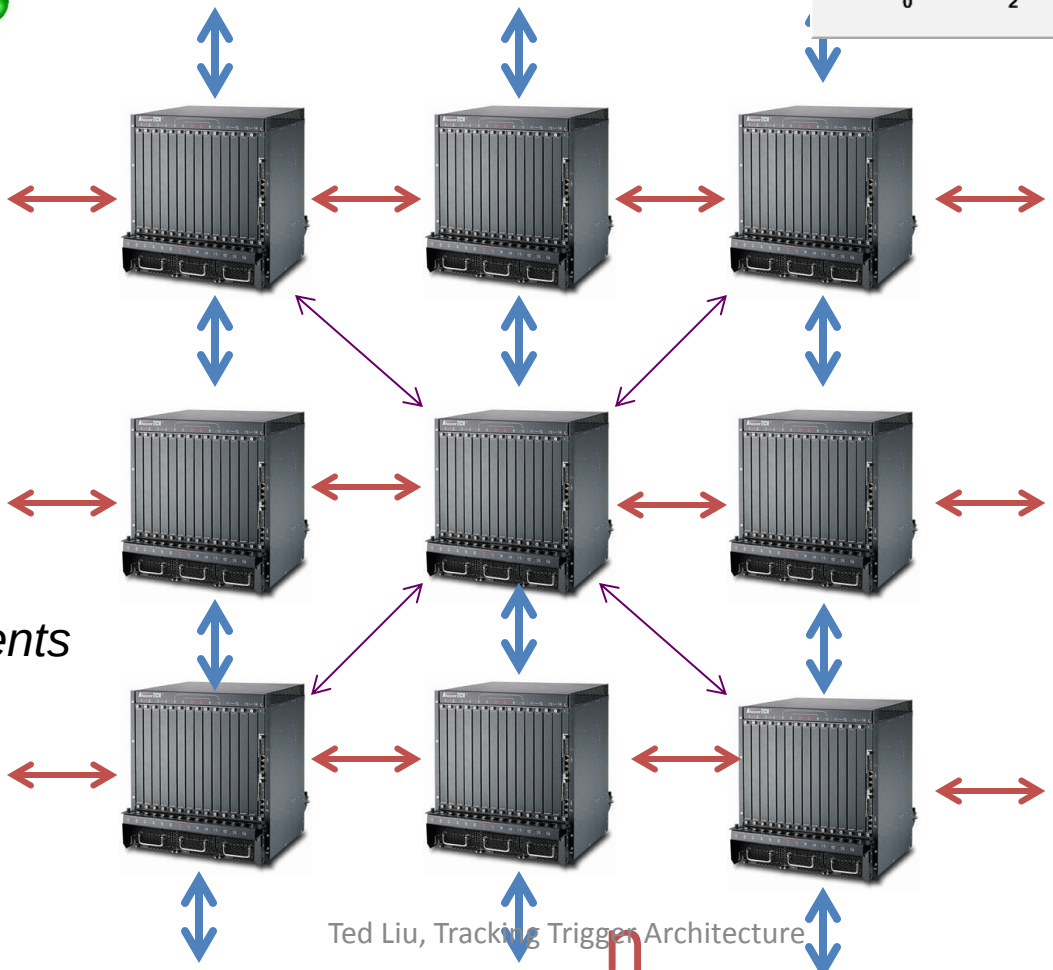
For simplicity, let's assume one crate is assigned to one trigger tower

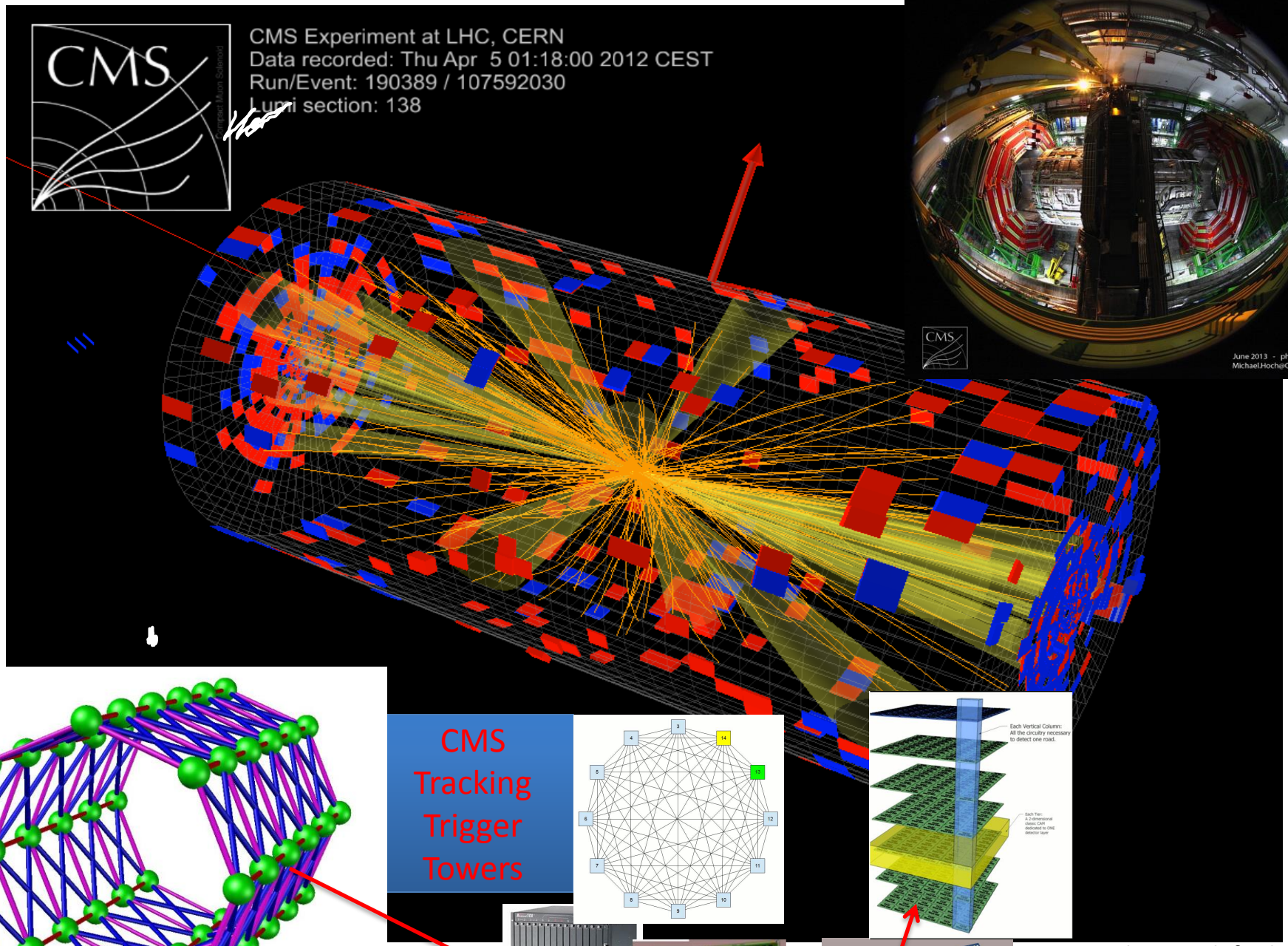




Simple
Trigger Tower
Interconnections.

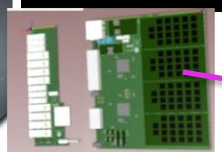
Each box represents a trigger tower.

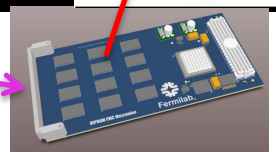




CMS Tracking Trigger Towers

ATCA





AM or other track finding approaches implemented on mezzanine (PR engine)

Commericial HUB boards

One simple configuration as an example

4 Pattern Recognition Boards

ATCA Full-Mesh Backplane

8 Data Input Boards

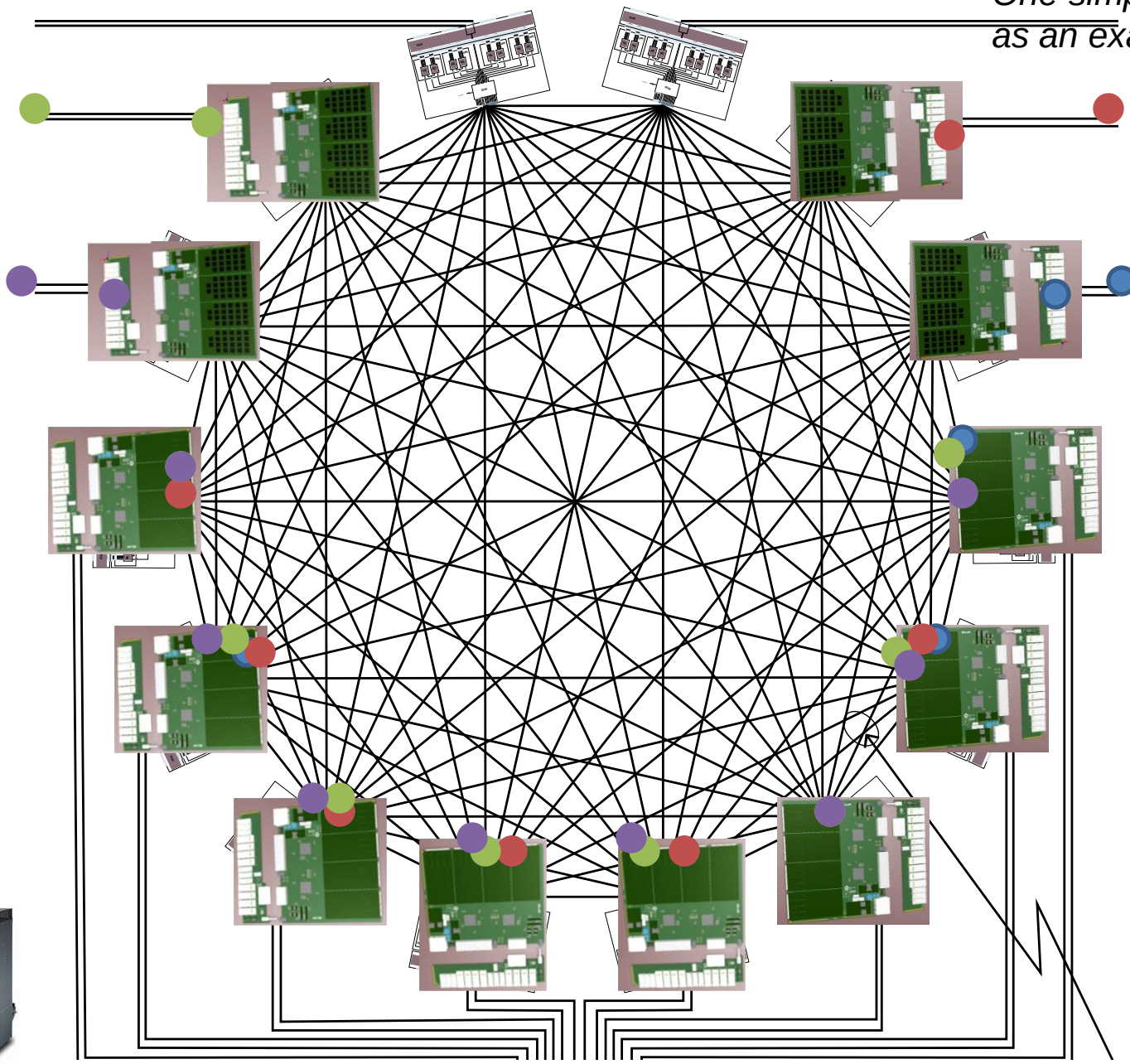
What's inside one trigger tower/crate

Fibers from upstream

40Gbps

Commericial HUB boards

One simple configuration as an example



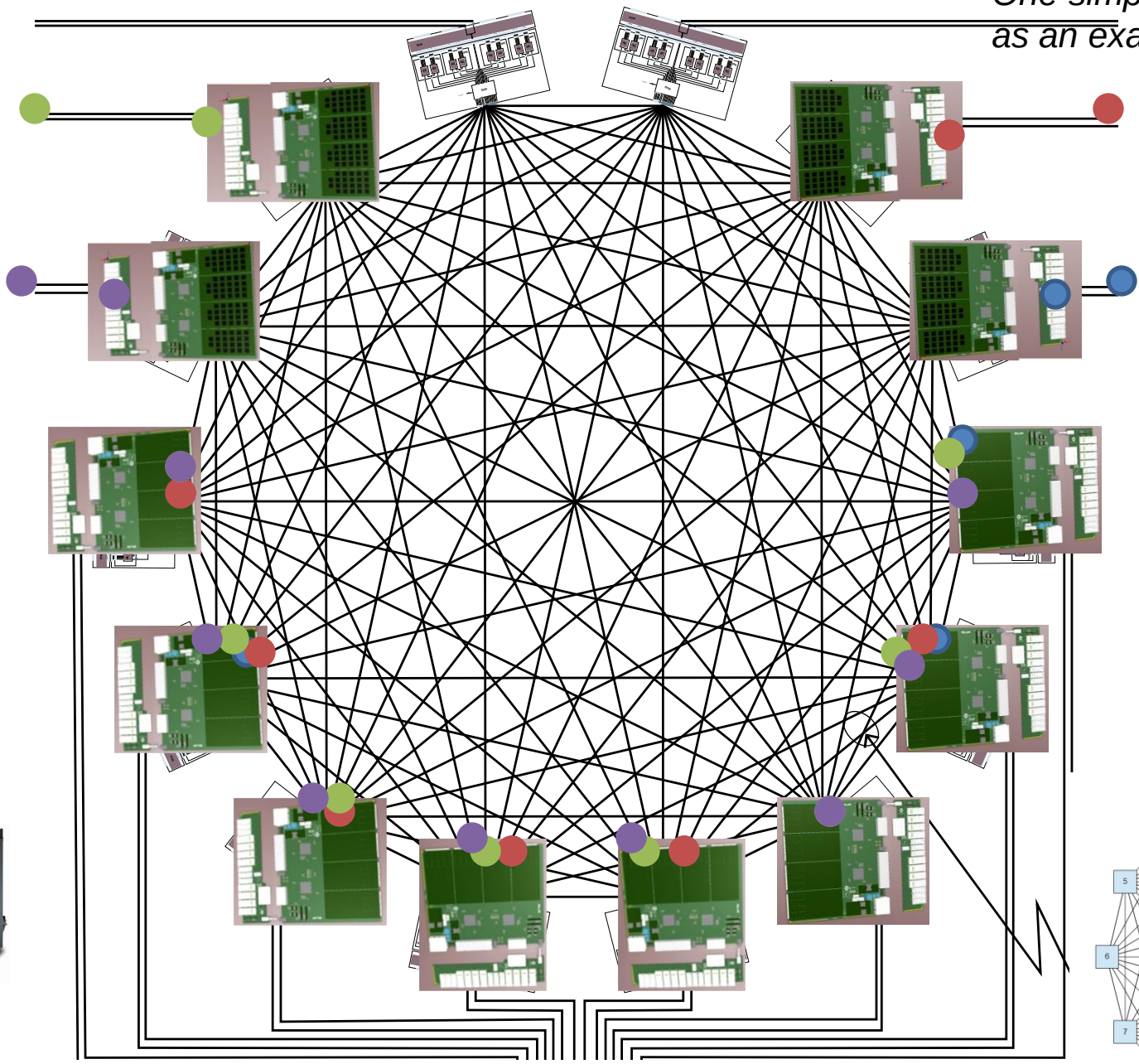
4 Pattern Recognition Boards (PRB)

8 Data Input Boards

Fibers from upstream

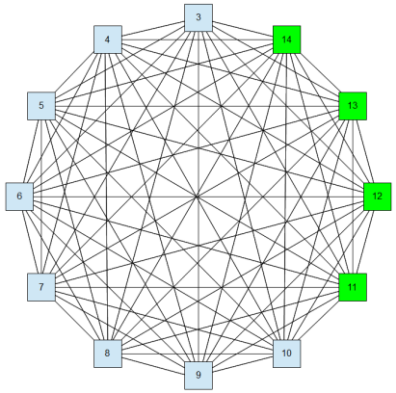
Commericial HUB boards

One simple configuration as an example



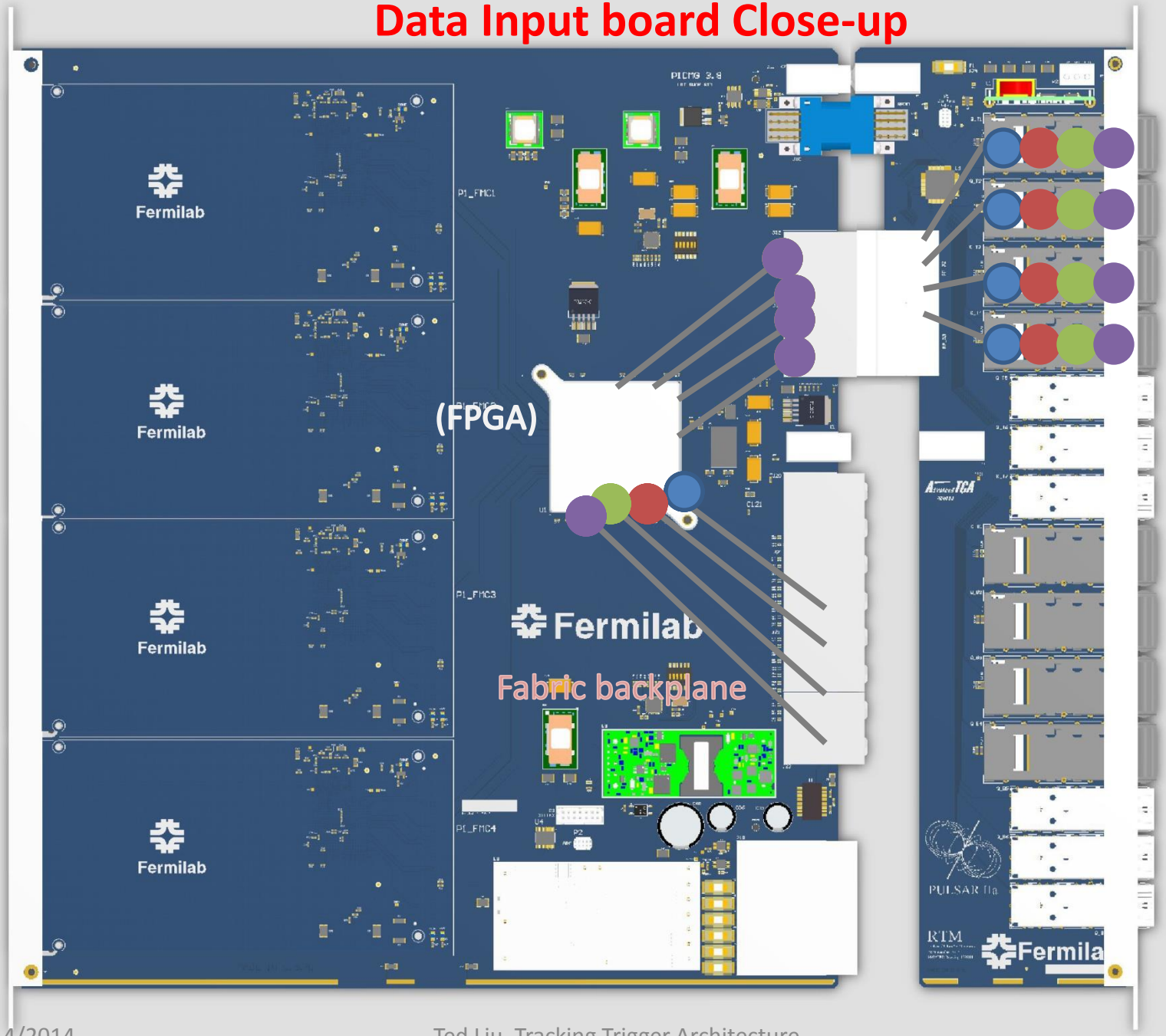
4 Pattern Recognition Boards

8 Data Input boards

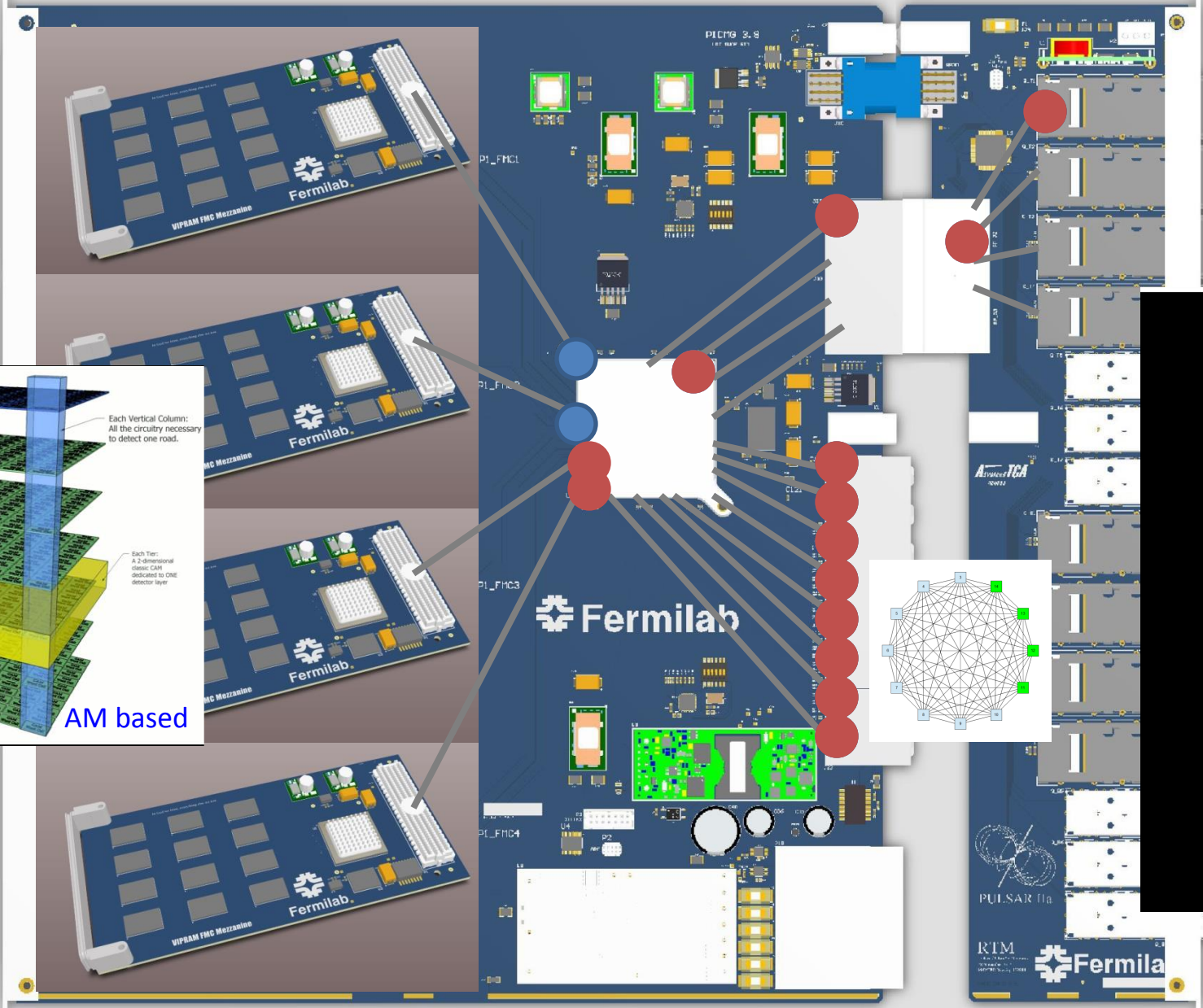


Fibers from upstream

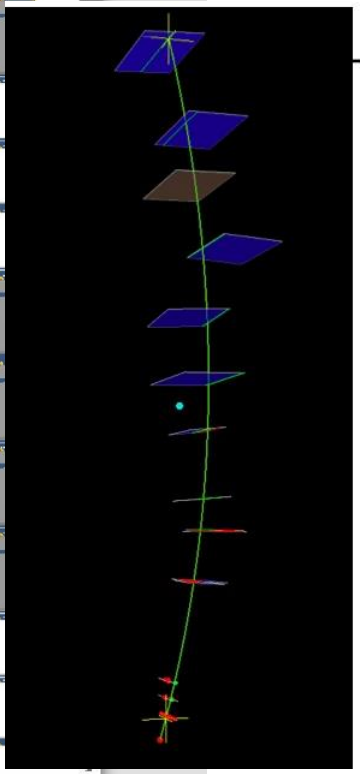
Data Input board Close-up



Pattern Recognition Board (PRB) data flow



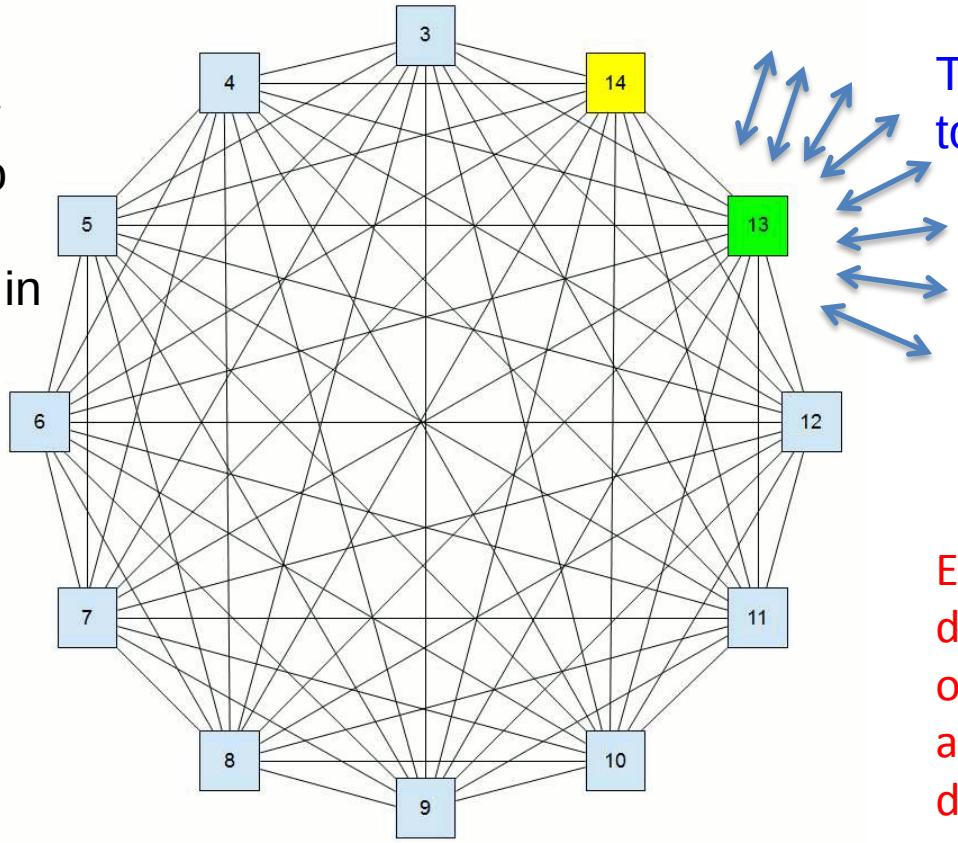
AM based



Neighbor data sharing to/from another Trgger Tower

More advanced configuration

Ten Processors and the Gateway send the event to the target Processor Blade in a round robin scheme.



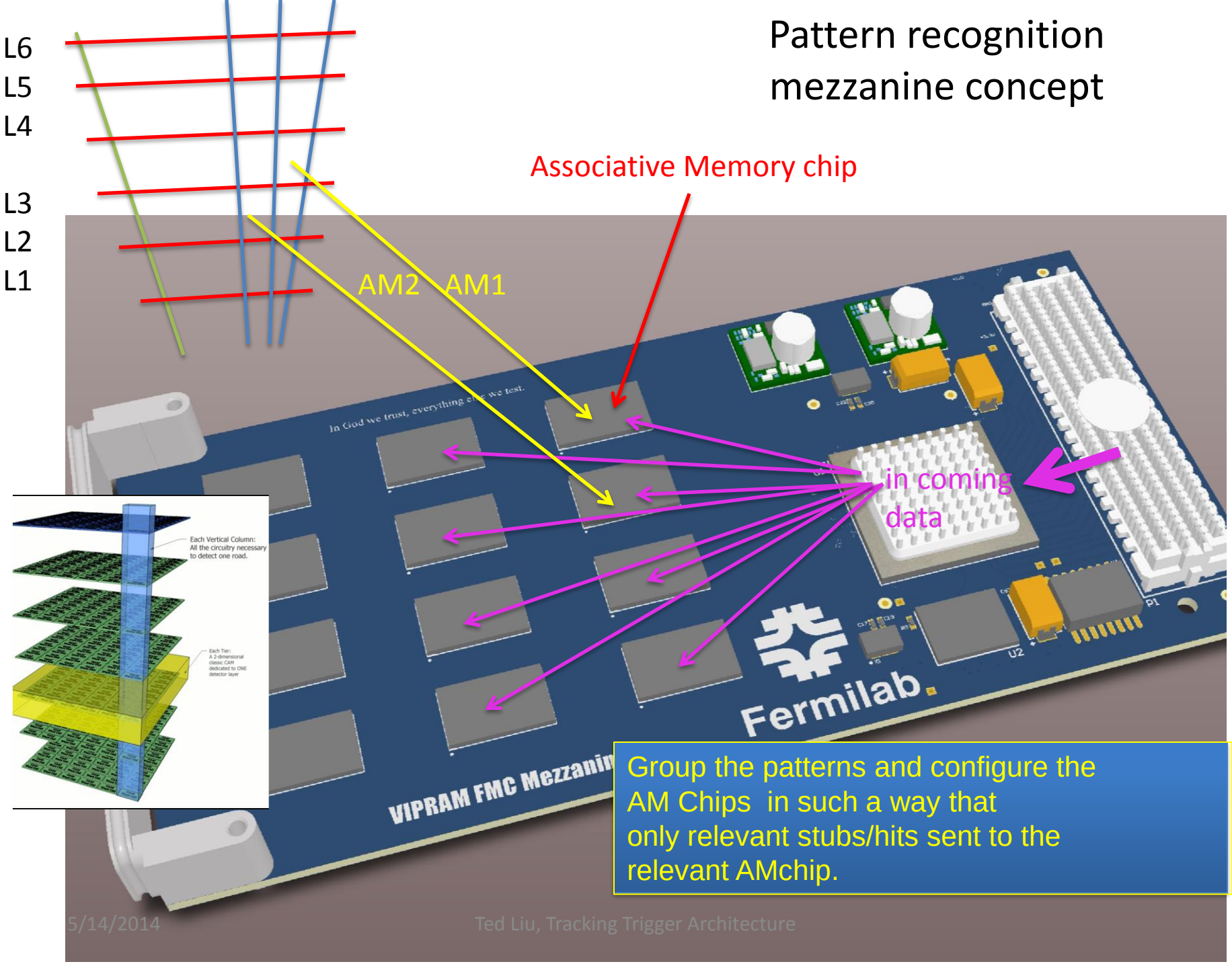
To/from eight neighbor towers

Each processor receives data directly from upstream on RTM,
and then take turn to handle different event for each trigger tower

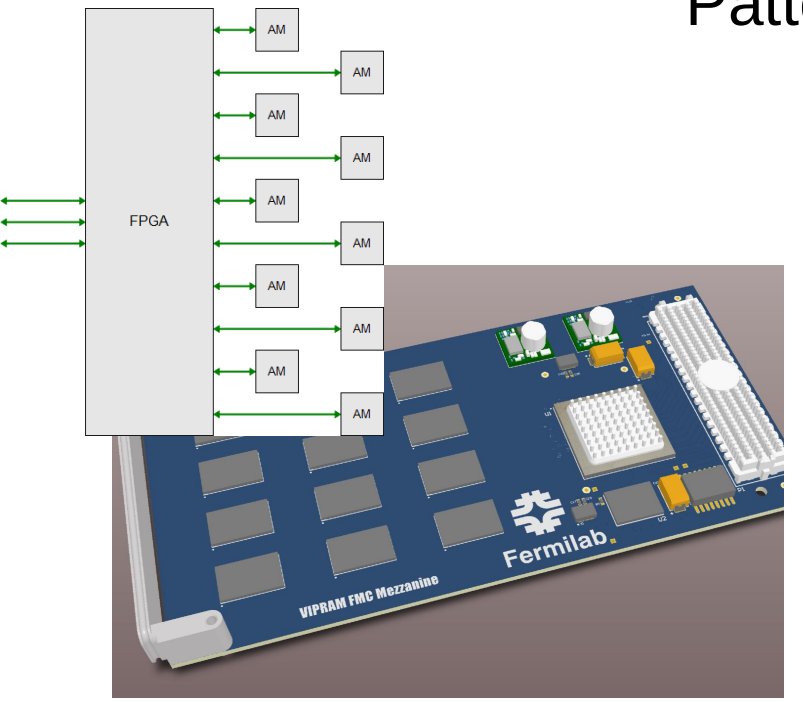
The full mesh based architecture is highly flexible.
Many performance and bandwidth bottlenecks can be solved/avoided/relaxed simply by better configurations.

This also makes an early technical demonstration feasible using today's technology.
The flexible architecture is a good platform for a vertical slice demonstration and beyond.

Pattern recognition mezzanine concept



Pattern Recognition Mezzanine (PRM)



Relaxed Performance Requirements (in the case of 10 PRBs with ~40 PRMs):

- 40MHz input handled by 40 PRM mezzanines in round robin, each handles ~1MHz event input rate
- Event Processing >= 1MHz (out of 40MHz)
- Input BW >= 16Gbps

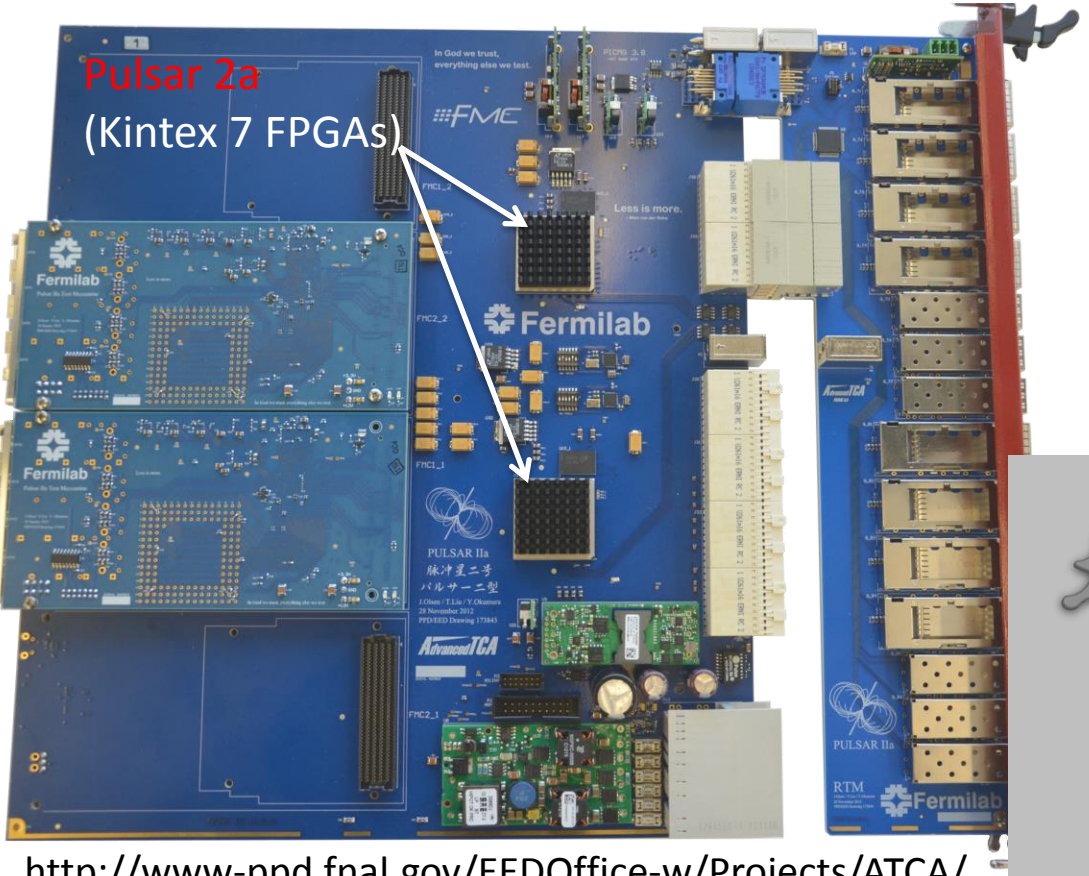
- In the case of AM approach:
  - ~10 AM chips / PRM
  - ~200k patterns / AMchip
  - ~ 2M patterns / tower
  - (2M x 48 towers ~ 100M patterns)

System size shrinks with better AMchip performance:

If 2X more AM pattern density,
or 2X higher AM speed,
→ 2 x less system size
(48 crates → 24 crates)

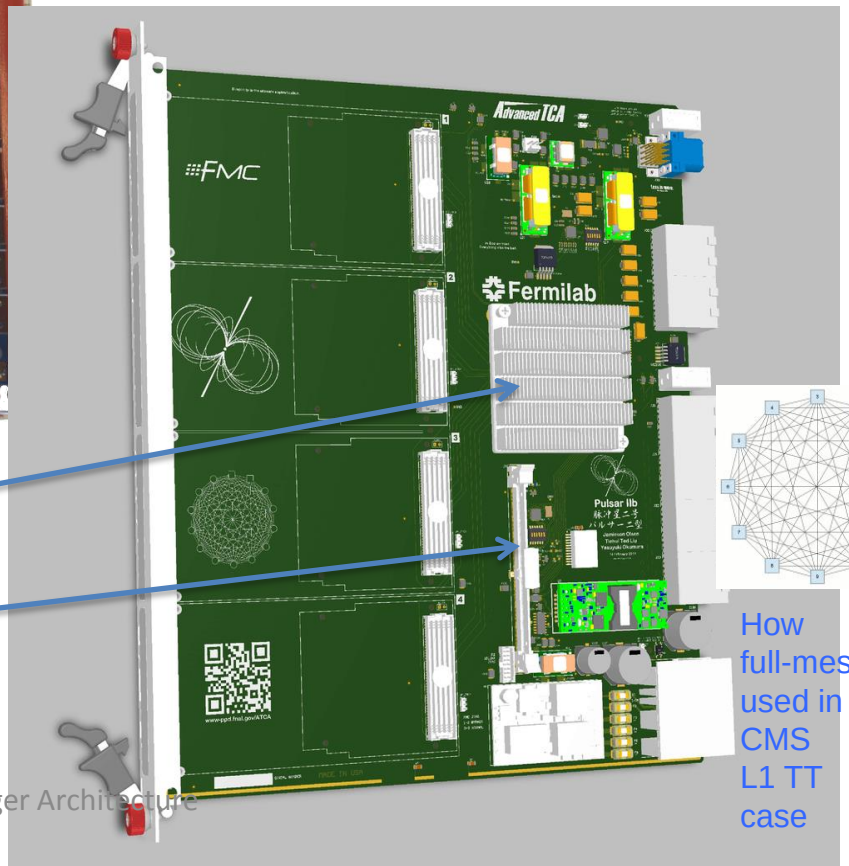
The relaxed performance requirement would make early technical demonstration easier for different track finding approaches.

Pulsar 2a prototypes work well: "plug & play" (summer 2013)



http://www-ppd.fnal.gov/EEDOffice-w/Projects/ATCA/

Pulsar 2b design

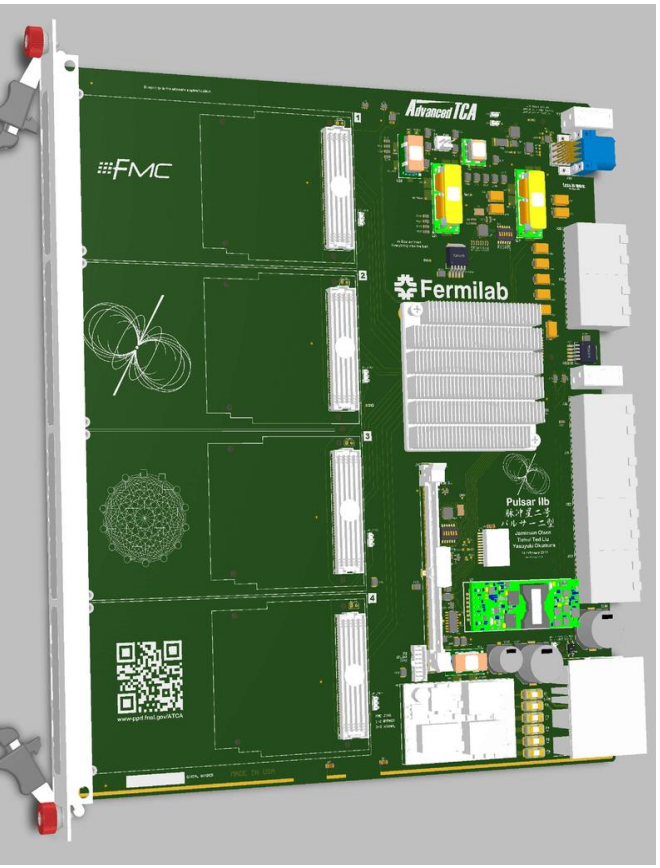


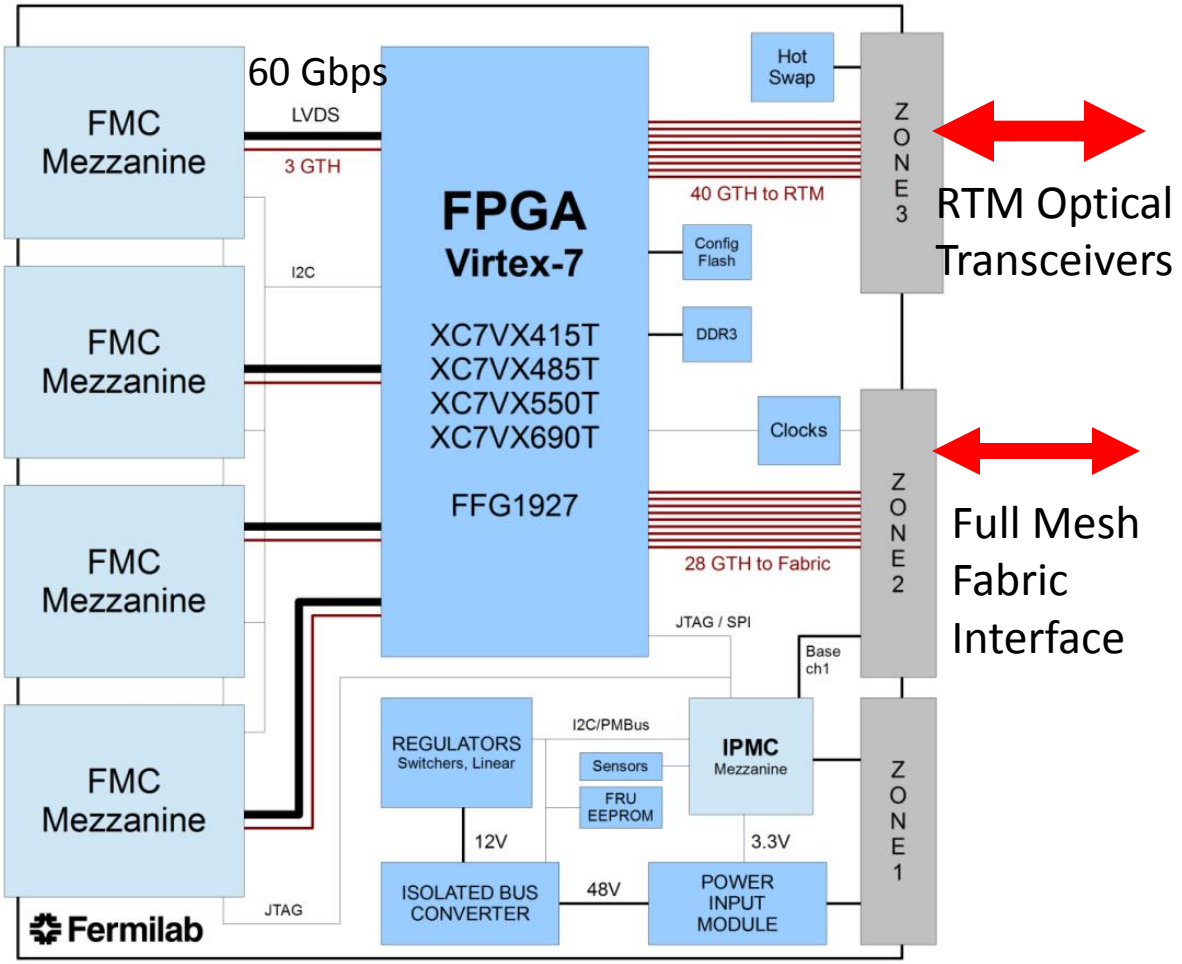
Pulsar 2b:

- ***Vertex 7 FPGA*** (XC7VX690T)
- ***80 GTH lines***
- Compatible with LAPP IPMC module
- FMC TTC compatible, backplane clock dist.
- CMS IPBus user interface
- General purpose design
- I/O ~ 1 Tbps

How full-mesh used in CMS L1 TT case

Pulsar 2b Block Diagram



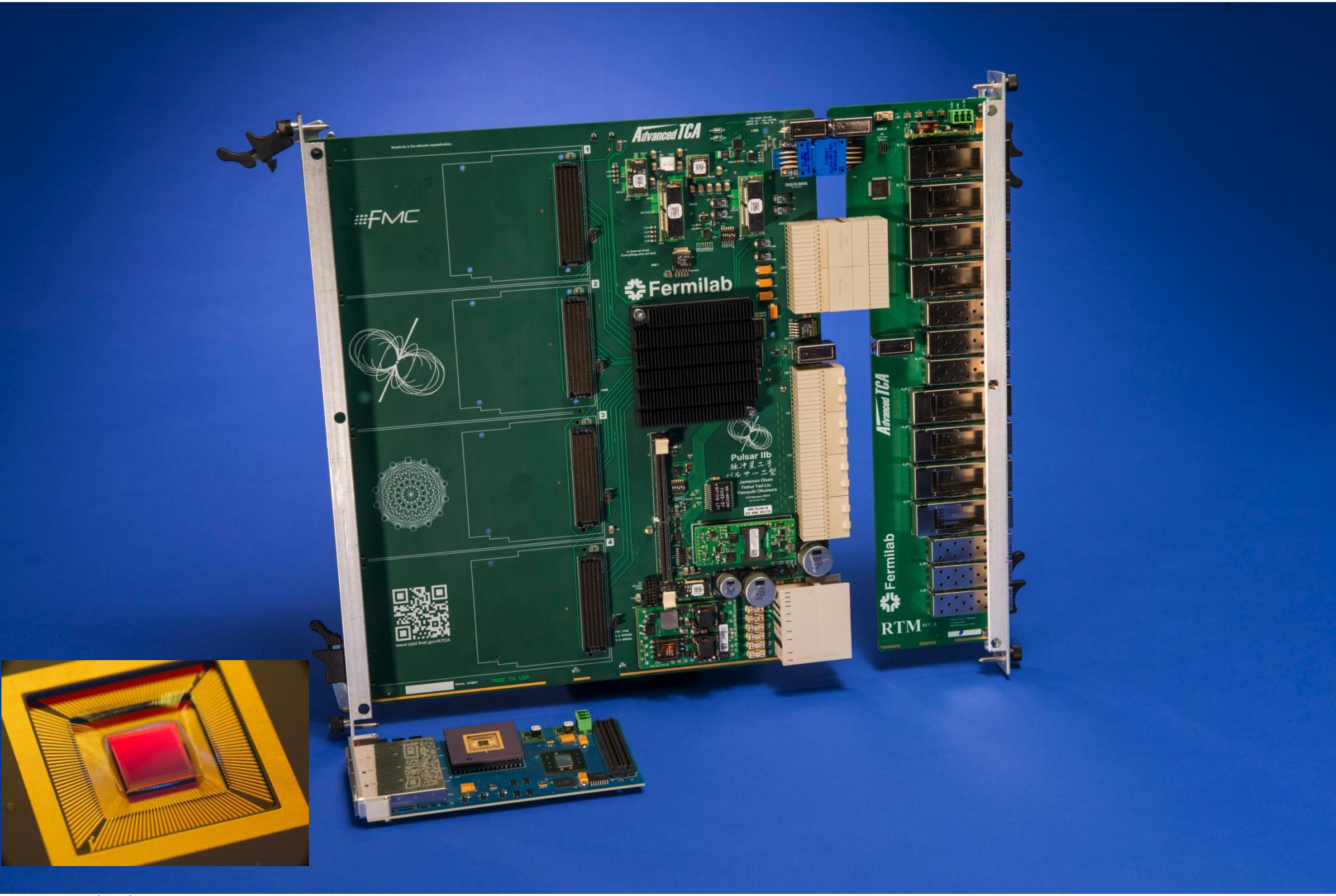


RTM Optical Transceivers

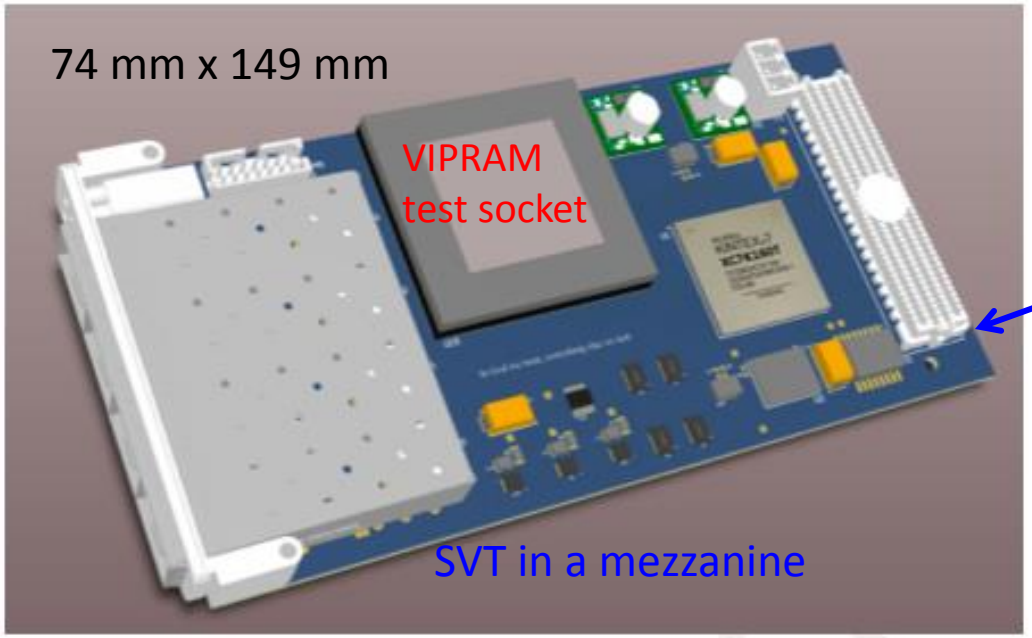
Full Mesh Fabric Interface

I/O: ~ 1 Tbps

http://www-visualmedia.fnal.gov/VMS_Site/gallery/stillphotos/2014/0000/14-0089-24D.hr.jpg



Pulsar 2b arrived 3 weeks ago



Test mezzanine card

design

actual card

protoVIPRAM chip

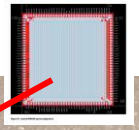
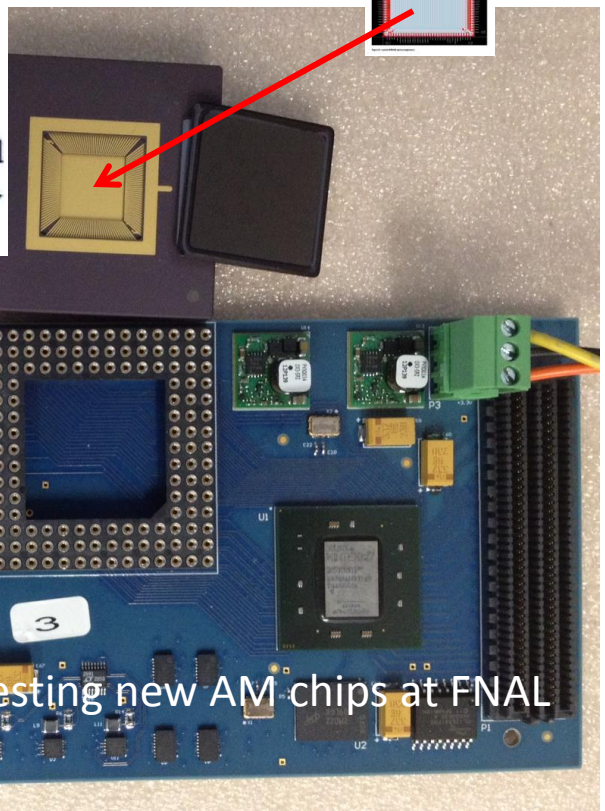
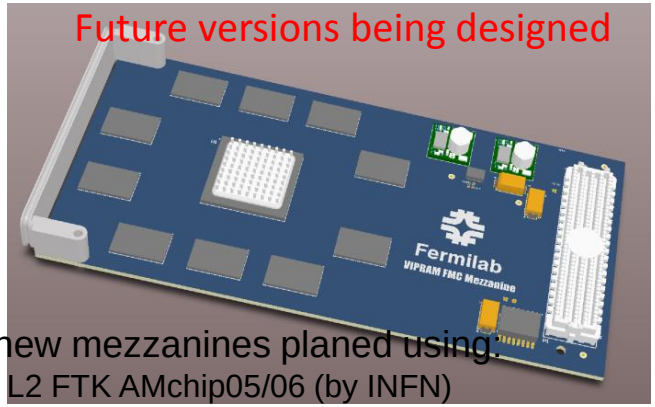


Figure 13: A test Mezzanine Card. This design features four SFP+ pluggable serial transceivers, a small Kintex FPGA, configuration flash memory, DDR3 memory, power supplies, local oscillators, a test socket and FMC connector.



Future versions being designed



Two new mezzanines planed using:

INFN L2 FTK AMchip05/06 (by INFN)

FNAL CMS AMchip dedicated for L1

Being used for testing new AM chips at FNAL

Some related abstracts for TWEPP 2014

- Pulsar 2b design and performance
- Pulsar 2b mezzanine design for AMchip05/6 (by INFN)
- Pulsar 2b application for FTK Data Formatter
- ProtoVIPRAM1: design and testing results
- Next version of protoVIPRAM for CMS L1 demonstration
- Power and Thermal analysis results for ProtoVIPRAM (SMU EE)

details will be presented at TWEPP 2014 this Sept.

Vertical Slice System Demonstration over next few years

Can and will be Implemented in stages:
mezzanine, board, crate and multi crate level

With the goals:

- Performance study (latency, efficiency etc)
- Identify issues/bottlenecks
- Guide future R&D, find solutions
- *A common platform to explore new ideas/algorithms/approaches*
- An important step towards TDR and beyond
- *A major undertaking* !

CMS people involved:
Lyon/INFN/Cornell/Northwestern/
Florida/Purdue/KIT/UK/CERN/FNAL …