Trigger Upgrade Activities

DHEP Annual Meeting
May 4-6, 2022

Team Members: Kajari Mazumdar, M.R.Patil, Kushal Bhalerao , Mangesh Kolwalkar, Pramod Pathare, Dnyanesh Naik & other EHEP colleagues

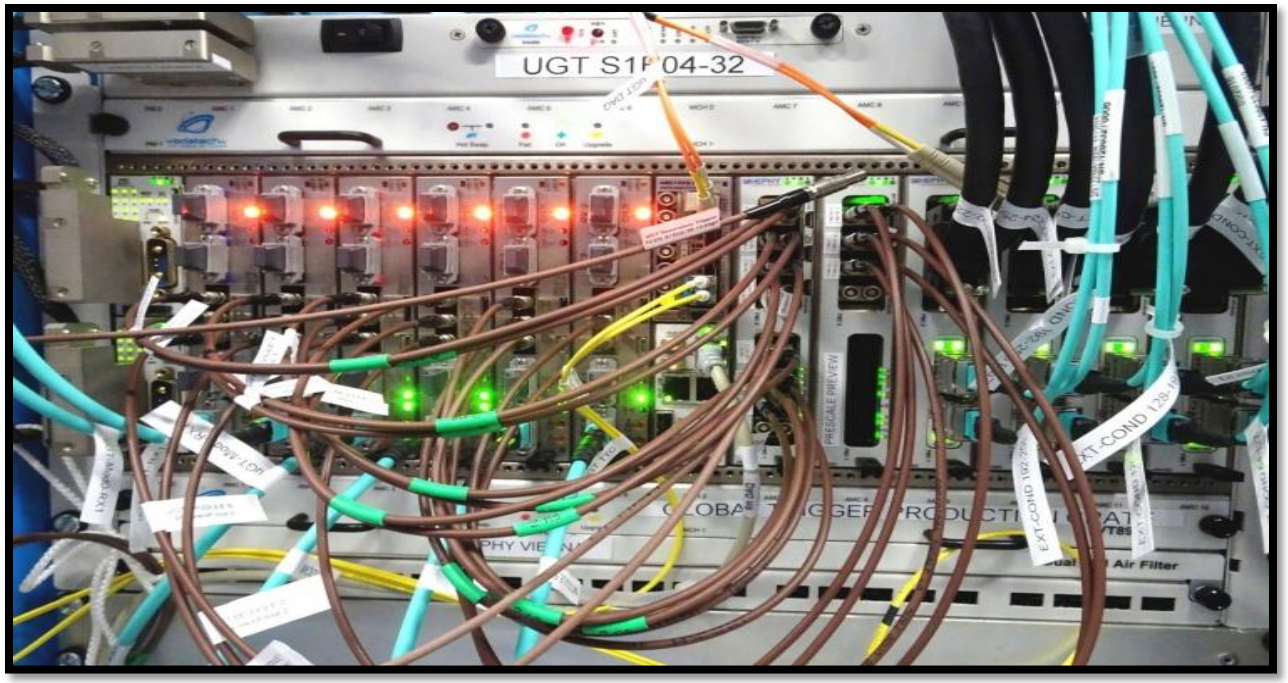
What is iperg

perf is a tool for network performance measurement and tuning. It is a cross-platform tool that can produce standardized performance measurements for any network.

Iperf has client and server functionality, and can create data streams to measure the throughput between the two ends in one or both directions.[2] Typical iperf output contains a time-stamped report of the amount of data transferred and the throughput measured.

The data streams can be either Transmission Control Protocol (TCP) or User Datagram Protocol (UDP):

Trigger Upgrade Motivation



CMS L1 Trigger System

- We are building a real time processing system that has to process an amount of data that is comparable to 5% of the total internet traffic and features high throughput real parallel processors. Our processors pull currents of the order of 200A at ~ 1V and dissipate heat of about 7W/cm2 . Data transmitted in each of our optical links corresponds to ~1 DVD / second .
- With these devices, we are able to bring the capabilities of the current HLT in L1 and implement algorithms that are more similar than ever to the offline reconstruction allowing us to reconstruct particles faster than microseconds. …Michalis Bachtis.

The CMS Two Level Trigger

A) Level-1 instrumented by custom hardware processor boards with dedicated FPGA and ATCA architecture:

The L1 currently receives information from calorimeter and muon systems generating an initial selection within 12.5μs(currently 4μs), with a maximum output rate of (currently 100 kHz) 750kHz

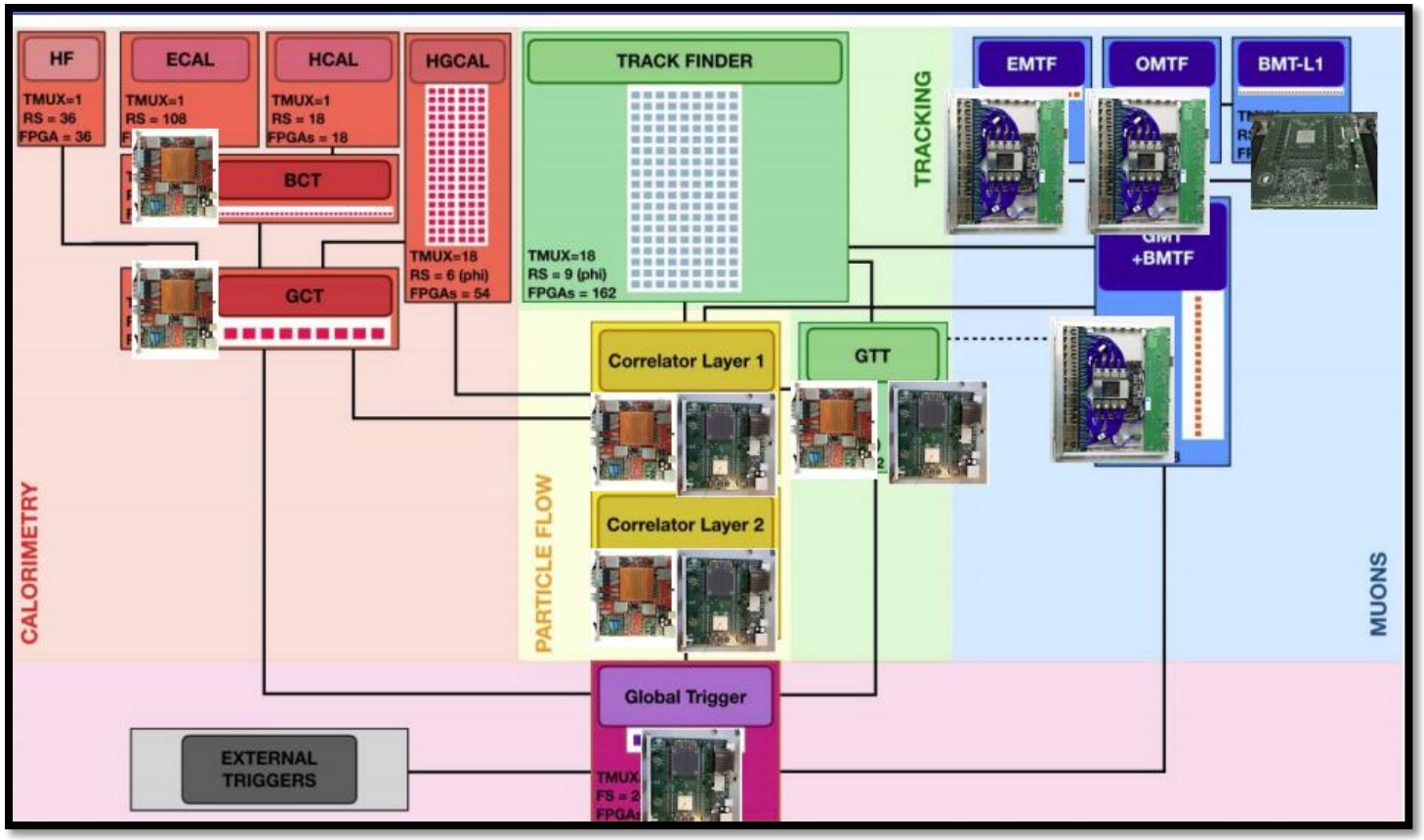
Upon a Level-1 Accept (L1A) received, the detector is fully readout and the selected event is reconstructed

B) High Level Trigger (HLT) software :

The HLT selection is based on this finer information reducing the output rate to about (currently 1 kHz) 7.5kHz. A first major upgrade of the L1 system has been conducted during the long-shutdown 1 (LS1 2013-2015).

A new architecture with improved performance was installed to maintain the high physics efficiency for the more challenging conditions experienced during Run-2 (2015-2018) and expected during Run-3 (2021-2023).

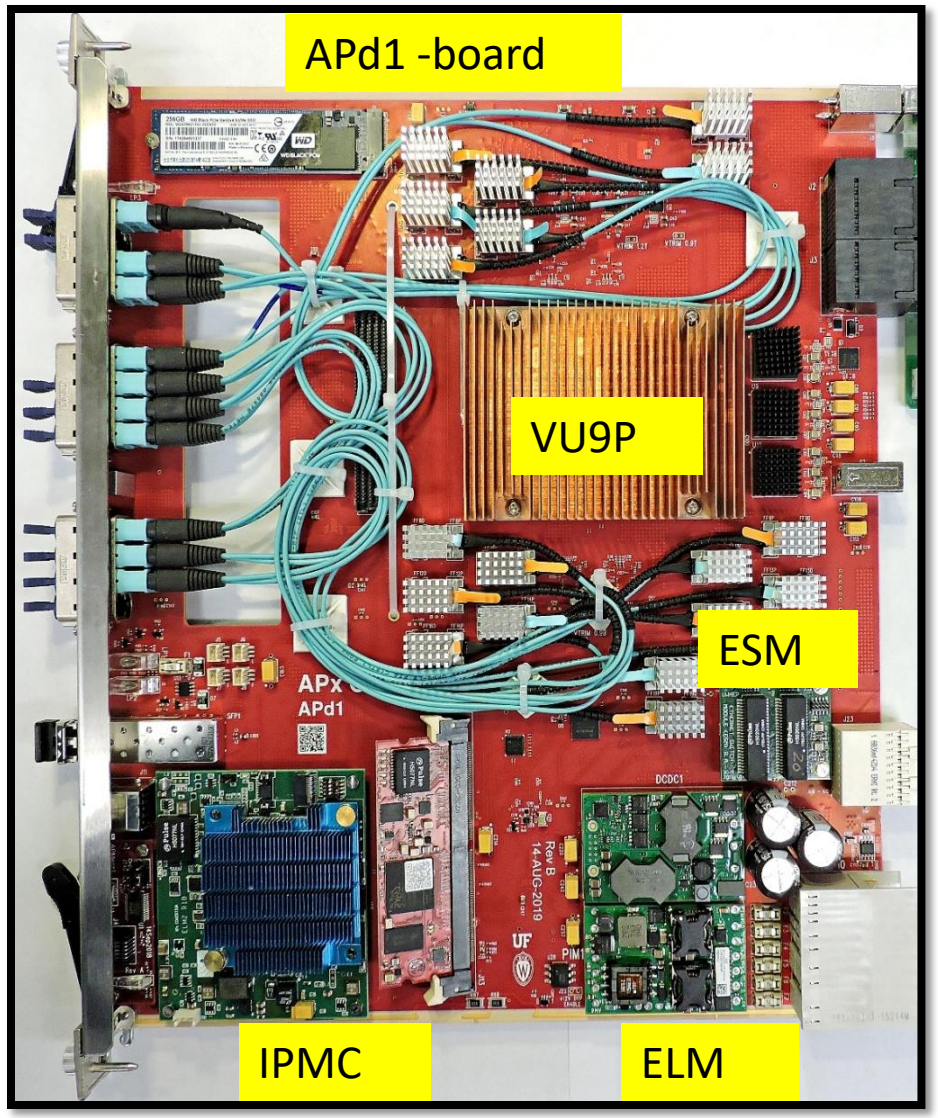
L1 Trigger Architecture



Calorimeters Triggers:

- Layer 1: Calibrates, merges and sorts ECAL/HCAL energy deposits into trigger towers (TTs)
- Layer 2: Reconstructs EG/jets/taus from TTs.
- **Muon triggers**: Three muon track finders (MTF) run in different |η| regions (Barrel, Overlap, Endcaps)
- **Global trigger (μGT):** Collects calo/muons objects, implement correlations, executes final algos for trigger decision

Hardware efforts : Peripheral Boards for APd1



APd1 –Advanced Processor demo board-The Main calorimeter Trigger Board powered by a VU9P (Virtex 9 Ultrascale) FPGA with 2.5M logic cells and 100 bidirectional links upto 28Gbps

--- Three daughter boards that are mezzannine on this Main board are ESM,IPMC and ELM board

ESM- Ethernet Switch Module (#50)

IPMC -Intelligent Platform Management controller (#10)

ELM -Embedded Linux Boards (#3) ;16 layer board that features a ZYNQ system on chip with dual core ARM Processor and FPGA logic

The above multilayer electronic boards were made in collaboration with Indian Industries (Micropack and Penisnula Electronics at Bengaluru) and the quality control of all of them done in house labs

Mezzanine Daughter Boards

- **IPMC (Intelligent Platform Management Controller)**

--Negotiates with crate for power, connectivity

--Controls power, monitors board conditions via sensors

-Monitoring Hardware –System Temperature and Power Supply

--Provides lower level configuration support (Recovery Control) I,e Booting ,Restarting and Shutting down the Server

-Logging-Out of range /failure states of the system

- **Embedded Linux Mezzanine (ELM)**

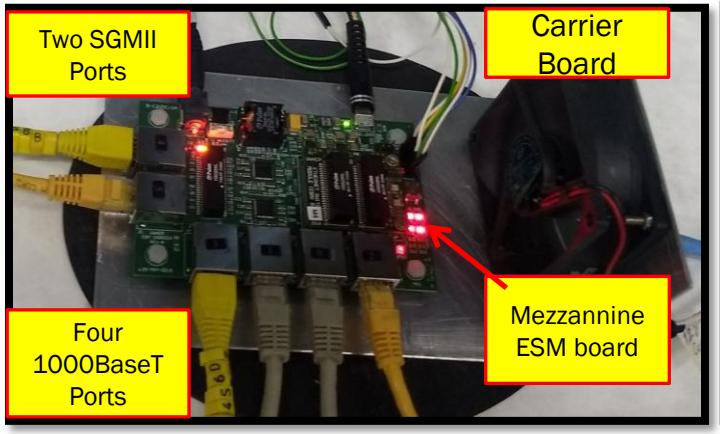
--Provides higher level configuration support (booting FPGAs, configuring memories, clocks, optics,..)

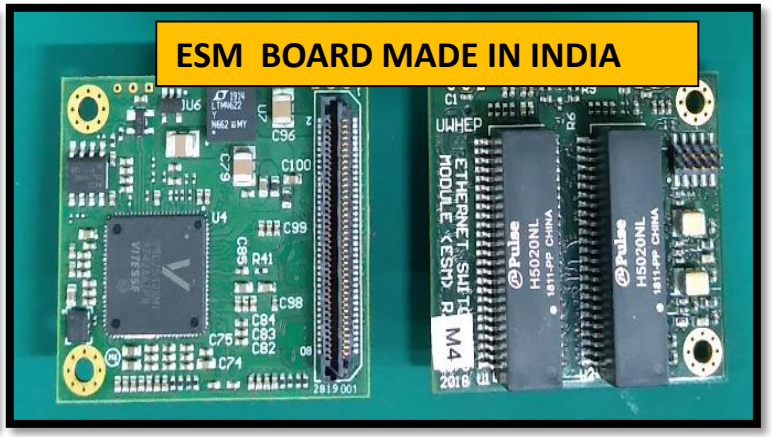
--Acts as primary network-connected TCP endpoint on the board during online operation

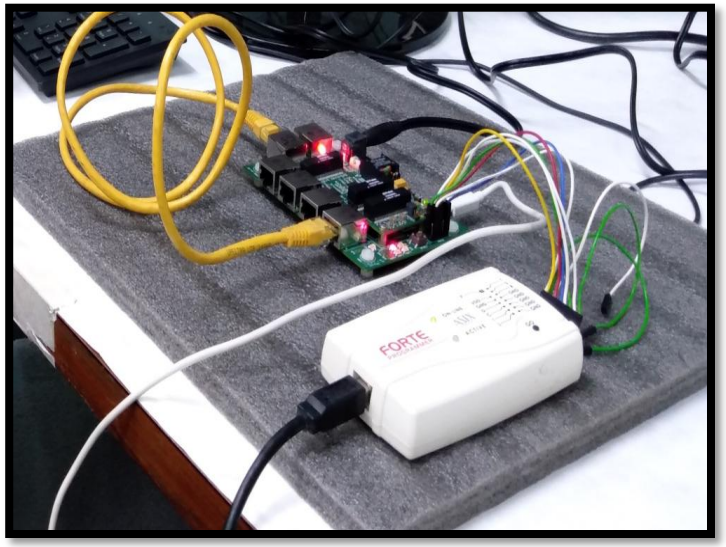
- **Ethernet Switch (ESM)**

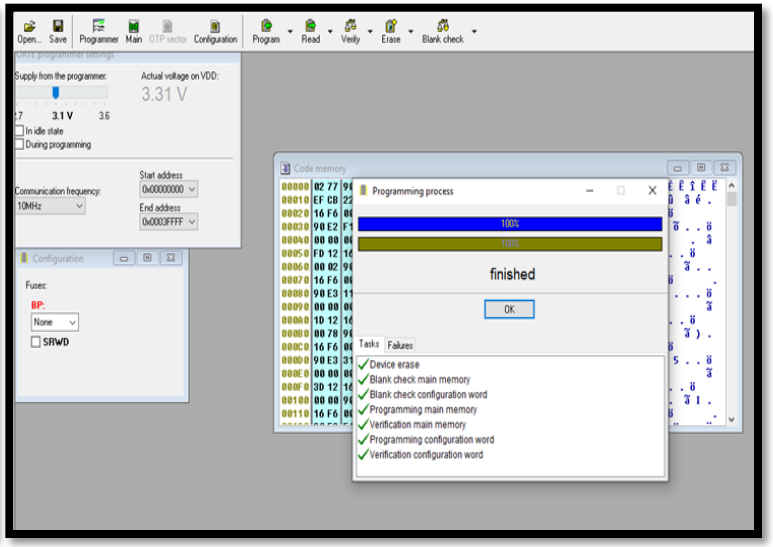
Connects the on-board endpoints such as the Linux & IPMC to the crate switch in an ATCA hub slot via back-plane connection (1000BASE-T is standard)

ESM (Ethernet Switch Module) test setup









Asix Forte Programmer used to update the firmware(the HEX code) to the onboard switch Processor via the Flash interface

ESM features

- It is intelligent than hub since it remembers the physical addresses of the devices connected to it.
- Micro semi gigabit layer-2 switch:
  - -Four integrated copper PHY ports
  - -Two 1G SGMII ports
  - +3.3V power source
- 256Kb flash on-board flash storage.
- 34x40 mm footprint

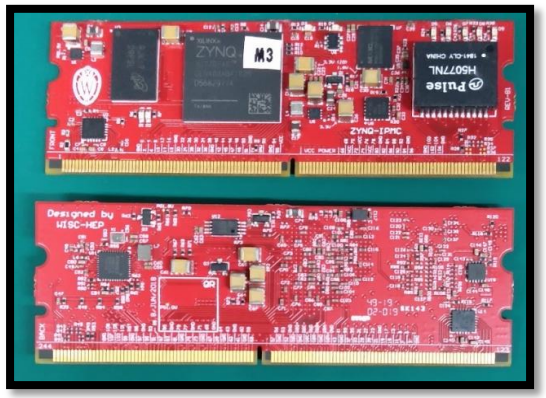
ESM test outputs

```
uPort iPort cPort MIIM Bus MIIM Addr   PHY/Serdes  CRC uPatch Link Status
-------------------------------------------------------------------------
    1     0     0        0      0x00 FERRET_7512  No uPatch   Up - 1GFDX FC(D)
    2     1     1        0      0x01 FERRET_7512  No uPatch   Up - 1GFDX FC(D)
    3     2     2        0      0x02 FERRET_7512  No uPatch   Up - 1GFDX FC(D)
    4     3     3        0      0x03 FERRET_7512  No uPatch   Up - 1GFDX FC(D)
    5     4     7        2      0xff -            Not PHY     Up - 1GFDX FC(D)
    6     5    10        2      0xff -            Not PHY     Up - 1GFDX FC(D)
>
```

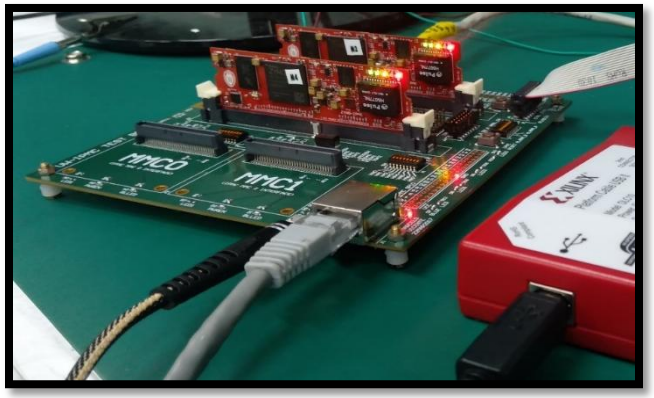
```
File  Edit  View  Search  Terminal  Help
hcal@hcal-desktop:~$ iperf3 -c 192.168.0.5 -p 5201 -P 1
Connecting to host 192.168.0.5, port 5201
[  4] local 192.168.0.3 port 45842 connected to 192.168.0.5 port 5201
[ ID] Interval           Transfer     Bandwidth       Retr  Cwnd
[  4]   0.00-1.00   sec   113 MBytes   950 Mbits/sec    0    362 KBytes
[  4]   1.00-2.00   sec   111 MBytes   935 Mbits/sec    0    380 KBytes
[  4]   2.00-3.00   sec   111 MBytes   935 Mbits/sec    0    399 KBytes
[  4]   3.00-4.00   sec   111 MBytes   935 Mbits/sec    0    399 KBytes
[  4]   4.00-5.00   sec   111 MBytes   935 Mbits/sec    0    399 KBytes
[  4]   5.00-6.00   sec   112 MBytes   936 Mbits/sec    0    419 KBytes
[  4]   6.00-7.00   sec   111 MBytes   934 Mbits/sec    0    419 KBytes
[  4]   7.00-8.00   sec   111 MBytes   934 Mbits/sec    0    419 KBytes
[  4]   8.00-9.00   sec   112 MBytes   936 Mbits/sec    0    438 KBytes
[  4]   9.00-10.00  sec   111 MBytes   935 Mbits/sec    0    438 KBytes
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bandwidth       Retr
[  4]   0.00-10.00  sec  1.09 GBytes   936 Mbits/sec    0             sender
[  4]   0.00-10.00  sec  1.09 GBytes   934 Mbits/sec                  receiver

iperf Done.
hcal@hcal-desktop:~$
```

- Each port of each board (i.e. 6 ports in all) tested with loop back cables for link UP status successfully
- Each Port tested for data Bandwidth and packet loss-No packet loss found and all Ports showed activity of nearly 1GBps

IPMC (Intelligent Platform Management controller)



IPMC BOARD MADE IN INDIA



Non-ATCA-based test fixture with some simple FPGA firmware to verify proper connectivity on all pins to validate the production lots for use in ATCA

The IPMC is responsible for power-up and environmental monitoring of the ATCA card, including voltage and temperature.

Commercial IPMC solutions are available, but are insufficient for the APx card series, which require faster response times to faults, wider input/output options, and additional monitoring features.

The custom IPMC is based on a Xilinx ZYNQ System-on-Chip (SoC) device in a 244-pin mini Dual In-line Memory Module (MiniDIMM) form factor

Running a Real Time Operating System (RTOS), the ARM Cortex-A series-based CPU is powerful enough to support TCP/IP connections for network-based I/O, firmware upgrades, and a Joint Test Action Group (JTAG) controller that can assist with main board debug.

Fast ADC channels integrated with the ZYNQ programmable FPGA logic can quickly detect fault conditions on the board and support rapid intervention to prevent damage in the event of over-temperature conditions or power supply faults, including a waveform capture capability around the time of the fault.

ELM2(Embedded Linux Board Ver2)

High level control is via a TCP/IP Ethernet endpoint in the form of a Xilinx ZYNQ based Embedded Linux Mezzanine (ELM)

The computing mezzanine is responsible for the high-level control functions of the card, specifically those associated with operating the one or more FPGAs on the main board.

Mezzanine form factor to facilitate use across APx ATCA board family

- ELM2: XCZU4/5CG-based
- PL-based IP library and custom peripherals support main board operation
- PL peripherals tailored to specific ATCA board requirements
- Key ELM functions in APx use:

-1GbE and 10GbE endpoint

- FPGA bitstream loading and register/memory IO (via AXI bridge)

-Support device configuration & monitoring (e.g. Refclk Synths, Fireflys, etc.)

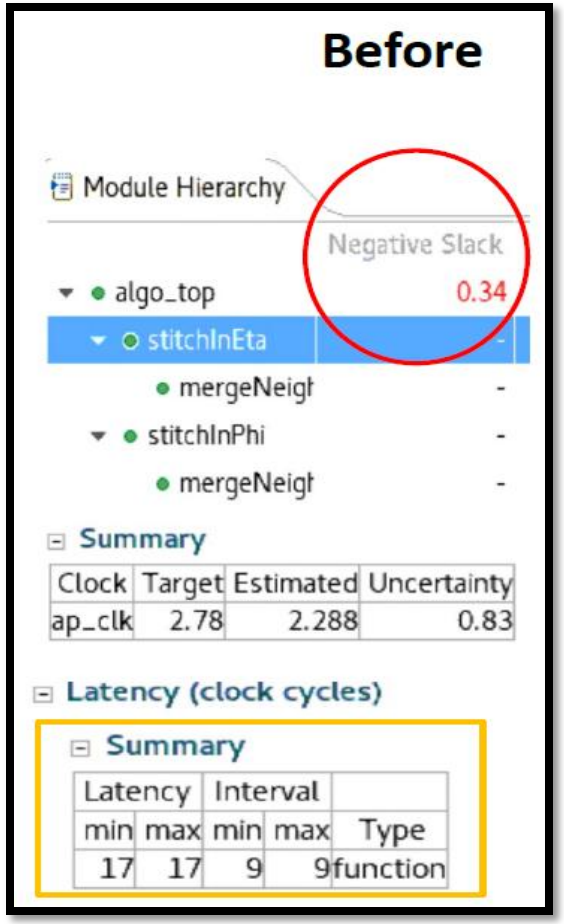
ELM BOARD MADE IN INDIA

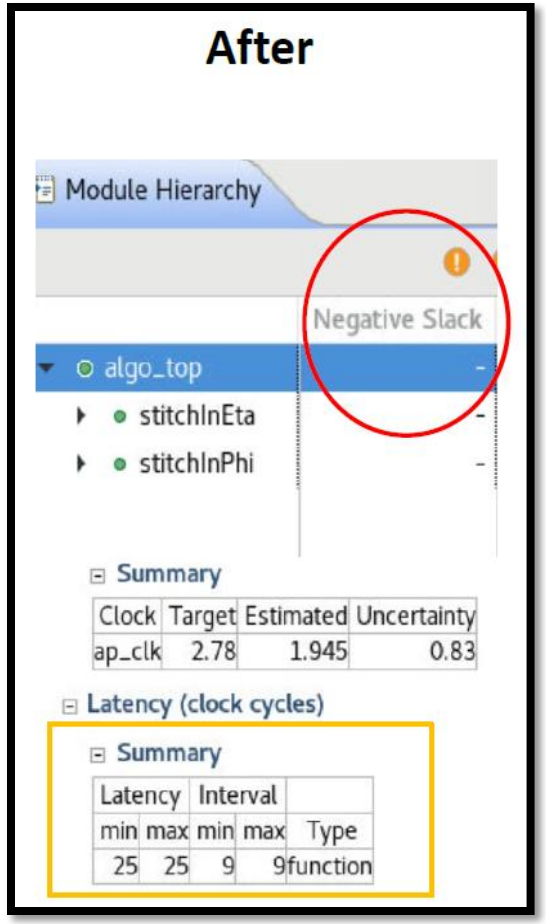
Trigger Algorithms and Firmware development

Trigger Algorithm Optimization

Trigger Algorithms for the Regional Calorimeter Trigger(RCT)

- Takes crystal level input from ECAL and tower level input from HCAL
- 36 RCT cards are needed to cover entire barrel calorimeter of CMS experiment
- Algorithm is being developed to cluster ECAL energy and add corresponding HCAL energies
- Regional Hadronic/Electromagnetic energy ratios and isolations are evaluated and sent upstream to GCT
- Algorithm is being developed for Xilinx's Ultra scale plus VU9P FPGA based cards running at 16Gbps for inputs and 25Gbps outputs at a clock frequency of 240 MHz
- Algorithms optimized for the latency and resource utilization done with thorough understanding of HLS (High Level Synthesis and architecture of ultra scale FPGAs)





Firmware Development :ELM Test Suite

ELM Board Test Suite :

- The firmware for testing all the peripherals on the ELM board was developed using Vivado Design Suite and customized Linux built using petalinux
- The peripherals like EEPROM, GEM (gigabit Ethernet Module) , I2C Clock Synthesizers and reference clocks used for frequency measurements were tested using this suite
- DDR was however tested in stand alone way and is however automated from Zynq configuration till application launch

For details refer the repository links:

https://github.com/bkushal26/ELM_Bring_up_Petalinux
https://github.com/bkushal26/ELM_BringUP_PSDDR

ELM test Base Board : Designed the ELM test base board for testing it in stand alone way before integrating it on the Apd1.

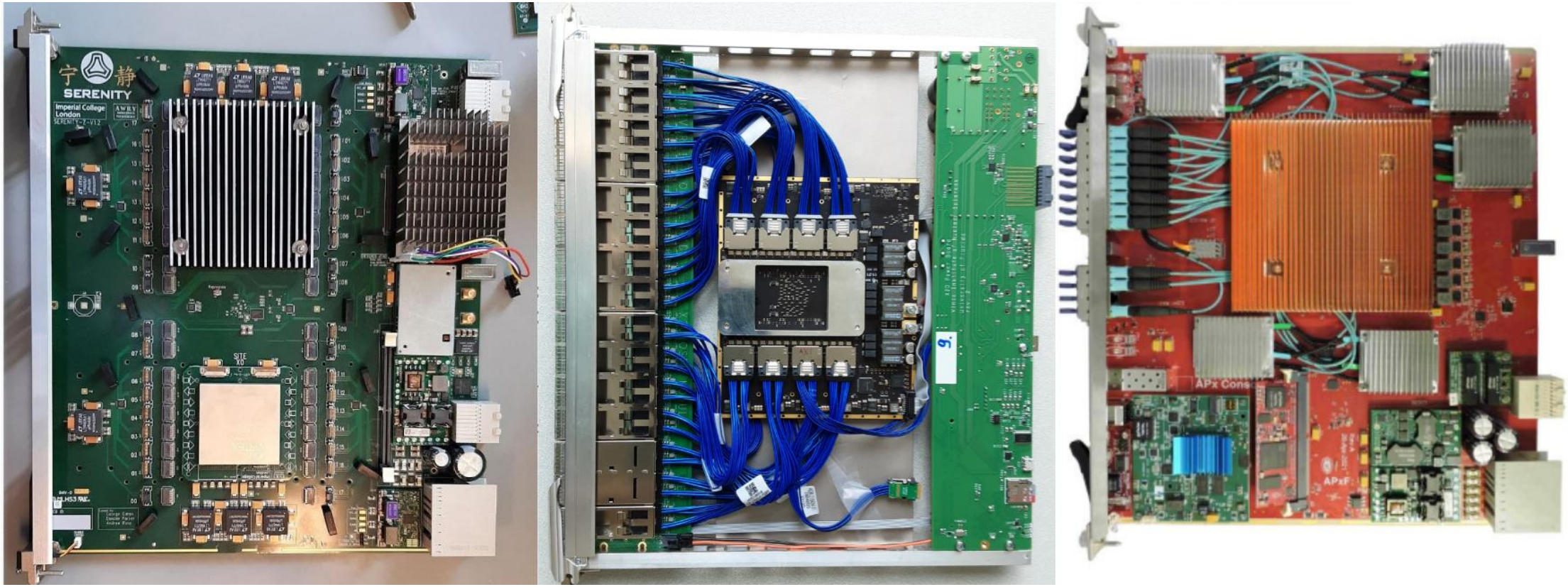
- The PCB is out for fabrication and once ready we will get it assembled and test the ELM boards here

Summary

- Working with the high end Ultra scale FPGA with SOC and Arm Processors embedded is a very enriching experience
- The Daughter boards developed in Indian Industries for the Calorimeter trigger board are showing excellent performance and we are ready to go into production of these boards
- Lot of scope for the Students to get involved in the trigger algorithms optimization and firmware development work
- A mini DAQ (Data Acquisition) system is also developed using DRS4 for Cosmic Muon detection study using Plastic Scintillators and PMTs (Photo Multiplier Tubes) used for the readout
- Use of Raspberry pi, Zynq SOCs(system on Chips ,GBT (Gigabit Transceivers, Serdes (Serializer Deserializer chips) used in the development and design of the ,in house test stand .to test the Trigger daughter boards (ESM,IPMC and ELM2)

Thank You !

L1 trigger Boards



Processing boards currently under development for the Phase-2 Level-1 Trigger upgrade project. These prototypes feature large Xilinx FPGAs hosting the trigger algorithms and more than 100 Input/Output high-speed optical links (28 Gb/s) for receiving/transmitting the data. From left to right: Serenity, X2O and APx boards.
Credits: Michalis Bachtis

2.7 Summary of trigger input bandwidth and latency

Table 2.11 summarizes the trigger primitives inputs from the Phase-2 sub-detectors. The time-multiplexing period for each system is provided. The number of output links, link speed and expected latency are also specified. The number of output links is given without including the (available) number of links to the 40 MHz scouting system as this is not expected to be a limiting factor. Table 2.12 provides the required bandwidth from each sub-detectors. A total input bandwidth of 62 593 Gb/s is expected to be digested by the Phase-2 L1 trigger logic downstream.

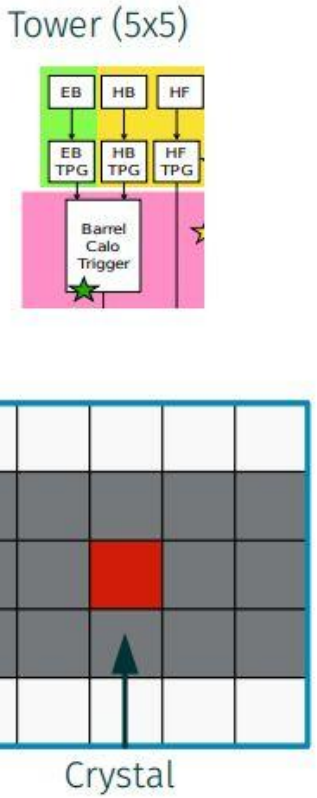
A summary diagram displaying the links between the trigger primitives, the trigger objects, the Level-1 algorithms used in the menu and the physics channels, is shown in Fig. 2.24

Table 2.11: Trigger primitive (TP) inputs expected from Phase-2 sub-detectors. This table summarizes the main features of the trigger backend system of each sub-detector sending TP to the Phase-2 L1 trigger. Reported here: the TMUX period, the number of output links, the link speed and the latency (defined as the time after the bunch crossing at which the first data are received by the trigger system). The numbers in parenthesis account for the links required to transmit data from overlapping sector regions. Note that for some sub-detectors (DT, RPC. iRPC and ME0) the latency would need to be estimated for the Phase-2 detector.

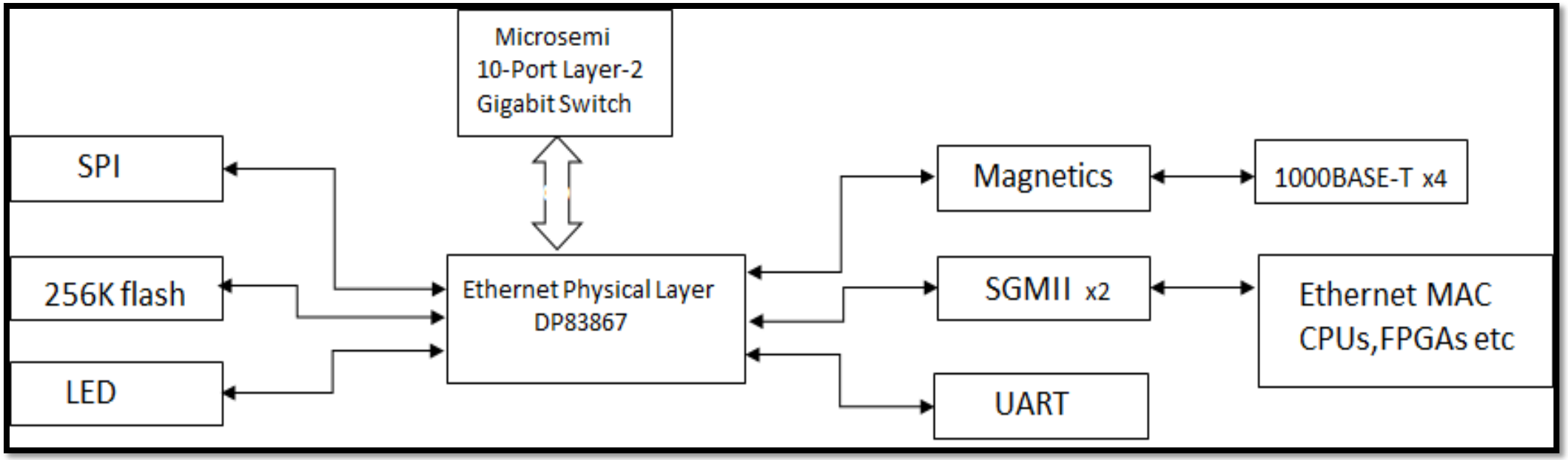
Detector	TMUX period	Output links	Link speed (Gb/s)	Latency (μs)
Track Finder	18	$9\times2\times18 = 324$	25	5
ECAL	1	3060	16	1.5
HCAL	1	144	6.4	1.5
HF	1	36	6.4	1.5
HGCAL	18	$2\times3\times4\times18= 432$	16	5
DT+RPC to BMTF	18	$60\times18 = 1080$	25	
DT+RPC to OMTF	1	72 (+18)	25	
RPC(endcap) to OMTF	1	42 (+6)	25	
RPC(endcap) to EMTF	1	48 (+12)	25	
CSC to OMTF	1	360 (+30)	3.2	1.75
CSC to EMTF	1	480 (+108)	3.2	1.75
iRPC	1	24 (+12)	16	
GEM (GE1/1)	1	96 (+12)	9.6	1.0
GEM (GE2/1)	1	36 (+12)	25	1.0
ME0	1	24 (+12)	25	

Algorithm Inputs

- Barrel ECAL, 5x5 crystals → single crystal
  - 16b x 61200 = ~39 Tb/s over 3060 x 16 Gb/s links
  - Improved position resolution and object id
- Barrel HCAL trigger primitives carry depth info from 7 segments
- Barrel calo trigger computes clusters finds standalone trigger objects
  - L1, 36 boards, finds proto-clusters in region
  - L2, 3 boards, stitches proto-clusters in region boundaries, computes final clusters and standalone objects
- 2.3 Tb/s output bandwidth



ESM + Carrier Board architecture



Why ATCA (Advanced Telecommunication Architecture) in HEP

The focus on high availability and reliability, the large bandwidth offered by the shelf backplane, and the large availability of electrical power and thermal dissipation make ATCA systems very attractive for use in high energy physics experiments, where extreme detector read-out rate, low latency, high availability and limited physical space occupation are requirements for the successful installation and operation of the back-end systems of the detectors.

Optimization of resources

- Use of custom data types, like RTL buses, supports arbitrary data lengths in contrast to C/C++ native data types.
- HLS provides arbitrary precision data types --> smaller bit widths --> faster execution -->more free logic space in FPGA fabric.
- Use of optimization directives, like, *Array Partitioning*-
- Breaks the arrays into individual elements/ smaller arrays --> smaller chunks of memories.
- Potentially improves throughput of the design

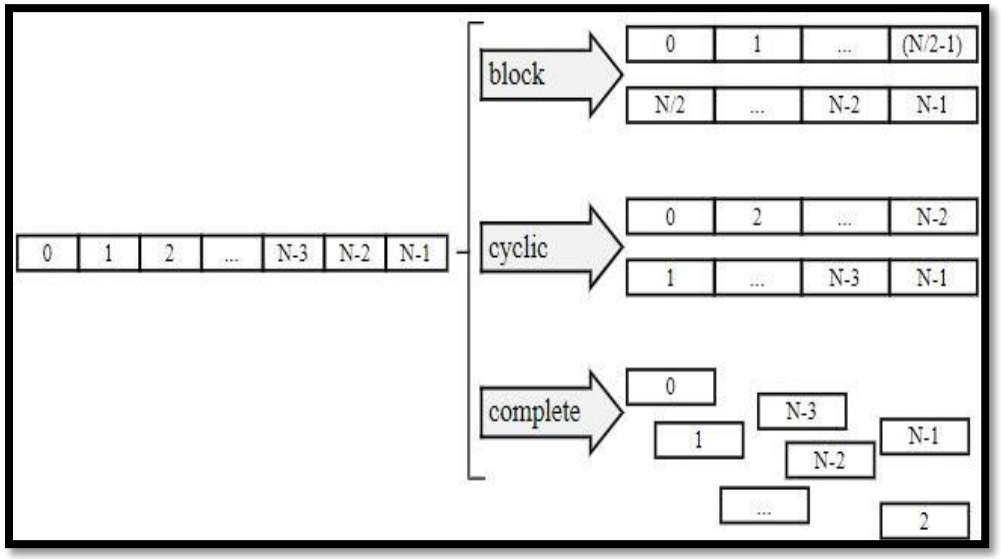


Image Source: UG902 by Xilinx

L1 Trigger Key feature in Phase 2 Upgrade

The L1T system architecture has been designed to process efficiently the 63 Tb/s input bandwidth (compared to 2 Tb/s in Phase-1) relying on state-of-the-art FPGAs (with 8 times more resources than today) and high-speed optical links reaching up to 28 Gb/s (compared to 10 Gb/s in Phase-1).

Directly inspired by the current system, the data processing is carried out by more than 250 generic-processing cards based on Advanced Telecommunications Computing Architecture (ATCA) technology.

Optimization

- Certain constraints within the design may limit the performance, like latency.
- The optimum solution obtained by tuning these parameters.
- (i) *Pipelining*- the next instruction can be launched into execution before current instruction is complete. An example of optimization using pipelining-

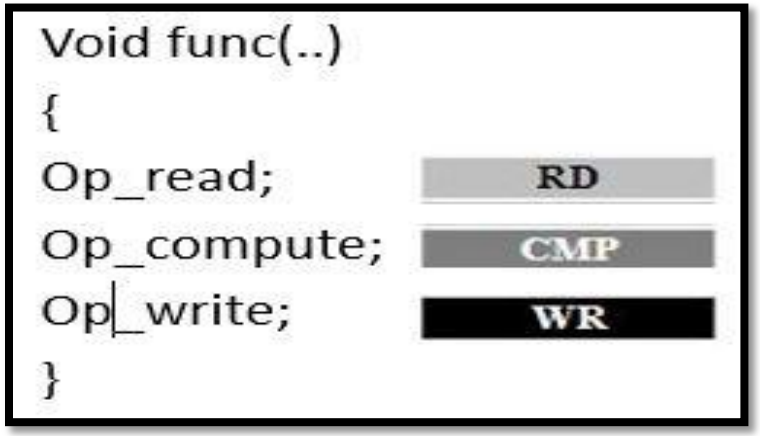


Image Source: UG902 by Xilinx