System Demonstration of a Tracking Trigger using an Associative Memory Approach

Zhen Hu on behalf of Fermilab/LPC
Tracking Trigger R&D group, in collaboration with

21st IEEE Real Time Conference
Colonial Williamsburg, Virginia - USA
Contents

• Motivation
• System overview
  – Multiplex in space and time
  – Associative Memory (AM) introduction
• System demonstration
  – Hardware introduction
  – Demonstrator at Fermilab
  – Performance
• Summary
• Over the past few years, Fermilab in close collaboration with CMS institutions has established a generic R&D program

  – Silicon-based Tracking Trigger based on Associative Memory approach
  – The initial R&D concept was presented at *Real Time 2012*
  – The initial hardware development was presented at *Real Time 2014*

• *Real Time 2018*: excellent prototype results
  – Allowed us to demonstrate at system level towards a working solution for the L1 tracking system in HL-LHC environment
  – The techniques developed are rather generic and could have other applications far beyond tracking trigger, both within and outside HEP
Physics Motivation: HL-LHC Trigger Upgrade

- High Luminosity LHC in 2026
  - 40 MHz bunch crossing
  - Up to 200 pileup (high occupancy)
  - Tons of data
- Current Level-1 Trigger will not work
  - Current maximum bandwidth
    - only 100 kHz (L1 output rate)
  - For HL-LHC, current trigger system would give
    - EG rate @25 GeV → 100 kHz
    - Overall Trigger Rate → >> 1000 kHz (unsustainable) to reach physics goals
  - Increasing trigger threshold → lose the opportunity of new physics with low threshold

- Upgrade trigger system
  - Must increase total bandwidth
  - Must increase trigger capabilities
  - Level-1 Tracking is a completely NEW handle
Track trigger advantage and challenge

- Silicon based tracking trigger is crucial for CMS Phase2 upgrade
  - Sharp turn-on efficiency curve
  - Background rate reduction → allows for low object threshold

- Huge challenges
  - How to handle readout of the entire tracker?
    - 260 M channel, 40 MHz, 100 Tbps data (after on-detector suppression using pT-modules), 2e4 hits
  - 4 μs latency:
    - Data distribution, track reconstruction, track fitting …
  - Silicon-based L1 tracking trigger has never been realized under these conditions
Solution: divide and conquer

- **Space parallel**
  - 6*8 trigger tower
  - 100 Tbps → ~2 Tbps per tower
Solution: divide and conquer

- **Space parallel**
  - 6*8 trigger tower
  - 100 Tbps → ~2 Tbps per tower
- **Time parallel**
  - 8x time multiplexing
  - 25 ns → 200 ns

Huge amount of cabling work without ATCA
Solution: Associative Memory

- Associative Memory
  - Based on CAM cells to match and majority logic to associate hits with a set of pre-determined hit patterns that represent allowed track trajectories (massively parallel, finishes right after all hits arrive)

Implementation of a High-Performance Pattern Recognition Associative Memory in an FPGA
Jun 12, 2018, 09:30
Jamieson Olsen (Fermilab)
Solution: Associative Memory

- **Associative Memory (AM)**
  - Based on **CAM cells to match** and **majority logic** to associate hits with a set of pre-determined hit patterns that represent allowed track trajectories (massively parallel, finishes right after all hits arrive)

- **Perfect tool to mediate the high pileup effect**
  - Avoiding the typical power law dependence of execution time on occupancy
The full mesh backplane interconnections effectively blur the distinction between FPGAs, and allow data sharing in both space and time.
We tested the data transfer performance for the full mesh back plan, Pulsar2b and RTM (10 Gbps).

- A custom ATCA full mesh enabled FPGA-based processor board
- Designed with the goal of creating a scalable architecture abundant in flexible, non-blocking, high bandwidth interconnections
We tested the data transfer performance between PRM and Pulsar2b (10 Gbps) and the interconnection between two FPGAs in PRM (16 Gbps).
Currently we emulate AM with FPGA
Full system demonstration at Fermilab

- Using the technology today to demonstrate tracking trigger feasibility
- For one Trigger Tower: two shelves fully loaded with Pulsar2b boards

Pattern Recognition Board (PRB) shelf
- 12 Pulsar2b with PRM Mezzanines
- Bandwidth between any pair of Pulsars is 20Gbps

Data Source Board (DSB) shelf
- 12 Pulsar2b to Emulates the detector output of ~400 modules

120 QSFP+ fibers
- Each with 4 bidirectional lanes each running at 10Gbps
- Capable of sourcing up to 4.8 Tbps data with full shelf
• All Pulsar2b boards in the system are synchronized to a common master clock optical link

1. Provided by a CERN TTCcx board that encodes a simulated LHC 40MHz bunch crossing clock and various control signals
2. This optical link is received by one Pulsar2b board in each shelf
3. From this Pulsar2b board the master clock and other control bits are distributed to other Pulsar2b boards in the shelf over dedicated clocks on the ATCA backplane
1. Simulated event data is first loaded into the DSB FPGAs
2. When triggered by bunch crossing signal, data is transmitted over the 120 QSFP+ optical links at full speed to the PRBs in the upper shelf
3. The 12 PRBs receive the incoming data, perform sophisticated time multiplexed data transfers over the full mesh ATCA backplane
4. Finally the event data sent to the FMC Pattern Recognition Mezzanine (PRM) cards
• Slave FPGA is used to emulate associative memory
  – Enables us to quickly evaluate and optimize the performance of the AM-FPGA interface and internal AM logic features
Demonstrator validation

Full simulation & emulation software

- Hardware and emulator perfectly matched
  - Output from each stage validated bit-by-bit
- With the full chain demonstrator, we have measured
  - Latency, FGPA resource usage, efficiency, resolution
Latency well within specs

Data delivered to PRM starts/ends: @1.20 – 1.70 μs
Pattern Recognition output starts/ends: @1.84 – 2.34 μs
Track Fitting output starts/ends: @2.04 – 3.04 μs
An example event in Vivado

4xTF, 7191.666 ns to 7716.666 ns = 525 ns, 126 clk @240MHz
This is with a selected ttbar + PU200 event with high tail of combinations
FPGA Resource Utilization (KU060)

Data Organizer only

- Very light weighted design
- BlockRAM mainly used for DO
  - TF does not increase BlockRAM usage, leaving enough room for TF
- TF: modest increase in registers and DSP blocks
  - Plenty of room for parallel copies of the fitter

Data Organizer and 8 Tracker Fitters

Graph: Table
Post-Synthesis Post-Implementation
High efficiency up to PU250

- System is robust against higher luminosity or increase in stub occupancy
  - We demonstrate that the system reconstructs all tracks for events with PU250 within 2.5 μs (no truncation needed)
  - Only for very high pT jet, truncation needed to meet the pipeline window

Tracking efficiency in ttbar+PU jets
Resolution

- Excellent performance for L1 trigger application
Summary

• Demonstrated with a vertical slice
  – Achieved excellent performance in terms of tracking efficiency and momentum resolution
  – Very low total latency (2.5 μs): data dispatch to the trigger towers, pattern recognition, track fitting

• The success of the demonstration system
  – An existence proof of fast data delivery, fast pattern recognition and track fitting implemented using the full mesh ATCA and associative memory approach
  – Technology advancements could lead to reduced size of the system in the future
Backup
Group photo taken at recent weekly meeting at FNAL/LPC

Close collaboration:
FNAL, Northwestern, U. Florida, Texas A&M, Brazil(SPRACE/UERJ) and China (Peking).
With FULL support of LPC
Excellent hardware performance

- **ATCA shelf**
  - 10 blades for parallel processing
  - Full mesh backplane is a natural solution for time multiplexing
    - All of the 56 bidirectional links among 8 Pulsar2b boards were tested at 10Gbps

- **Rear Transition Module**
  - 10 QSFP bidirectional links
    - 10 Gbps per link achieved

- **PRM performance**
  - Communication between Pulsar2b and PRM FPGAs
    - 10 Gbps achieved
  - Two latest generation of Xilinx FPGAs
    - Interconnection achieved 16.3 Gbps
AM Pattern and Bank

- A pattern is a low resolution track
  - Made of 1 superstrip (SS) per layer
    - A SS is a group of adjacent strips

More powerful AM => less demand on the FPGA
More powerful FPGA => less demand on the AM
AM in FPGA: very closely follows the AM ASIC (chip) design

- Match two silicon tiers in ASIC with two modules in FPGA firmware
  - CAM Tier -> a 2D array of Pattern Modules
  - I/O Tier -> fired roads serialization and output
- Pipelined operation
  - CAM tier: processes pattern matching with stubs for current event N
  - I/O tier: outputs road addresses for event N-1 at the same time

CAM tier logic is optimized for 7-Series/UltraScale FPGA architecture
## Latency well within specs

### Data Sourcing
- Module data arrives at RTM
  - Latency: 125 ns
  - Start: 0 ns
  - End: 200 ns
- Stubs after formatting
  - Latency: 325, 525 ns
  - Start: 200 ns
- Stubs transfer on Full Mesh Backplane
  - Latency: 625, 825 ns
  - Start: 642 ns

### Data Formatting /Delivery
- Stubs after layer sort
  - Latency: 200 ns
- PRB stubs after layer sort
  - Latency: 950, 1450 ns
- PRB PRBF-2 Formatting, FIFO
  - Latency: 6 clk
- MGT TX
  - Latency: 250 ns

### Pattern Recognition
- PRB stubs after layer sort
  - Latency: 950, 1450 ns
- PRB PRBF-2 Formatting, FIFO
  - Latency: 6 clk
- MGT TX
  - Latency: 250 ns

### Track Fitting
- PRB stubs after layer sort
  - Latency: 950, 1450 ns
- PRB PRBF-2 Formatting, FIFO
  - Latency: 6 clk
- MGT TX
  - Latency: 250 ns

### Data delivered to PRM starts/ends: @1.20 – 1.70 µs
### Pattern Recognition output starts/ends: @1.84 – 2.34 µs
### Track Fitting output starts/ends: @2.04 – 3.04 µs
Demo with FPGA

- Full pattern bank in ASIC (for a trigger tower)
  - Pattern 1
  - Pattern 2
  - Pattern 3
  - Pattern 4
  - Pattern 5
  - Pattern 6
  - Pattern 7
  - Pattern 1M

- Partial pattern bank demo in FPGA
  - Pattern 6
  - Pattern 7
  - ... ...
  - ... ...
  - ... ...
  - ... ...

- ~100 patterns fired per event

- Test ~10-20 events at a time
  - more than 1-2 bunch train worth, the rest is repeating
  - only need ~1K patterns loaded at a time