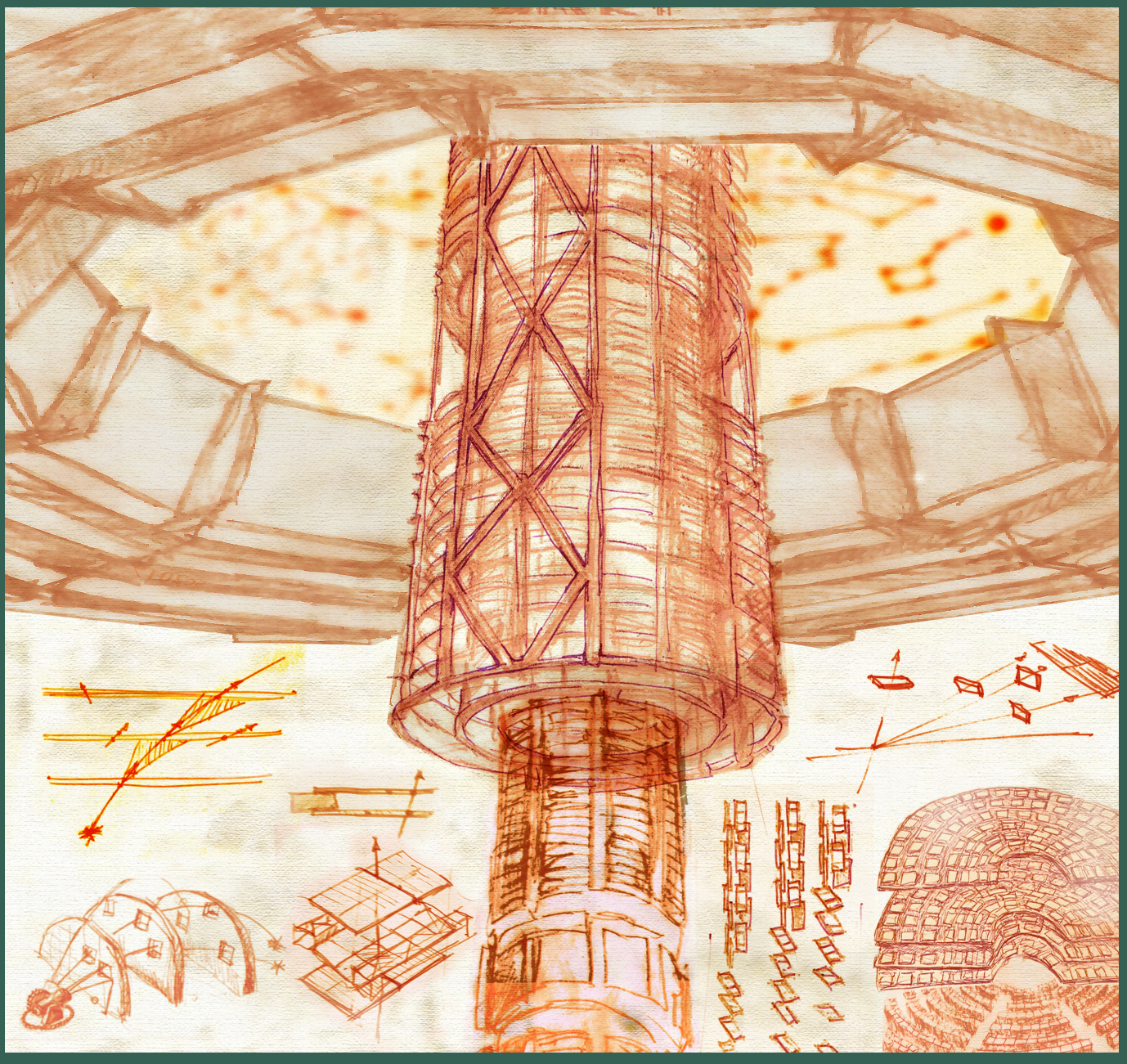
CERN European Organization for Nuclear Research
Organisation européenne pour la recherche nucléaire

CERN-LHCC-2017-009
CMS-TDR-17-001
1 July 2017

CMS



The Phase-2 Upgrade of the CMS Tracker

Technical Design Report

9.4.1 Associative memory plus FPGA approach

The AM+FPGA approach to L1 tracking has been described in Section 3.5.1.1. Proceeding from those general design considerations, a concrete AM+FPGA system configuration has been developed for implementation and demonstration in hardware. The demonstration system assumes 48 ($6\,\eta \times 8\,\varphi$) trigger towers and a time-multiplexing period of 20. Because DTC preprocessing functionalities are not yet fully specified, the design of the demonstration system seeks to avoid relying on specific DTC capabilities, and instead assumes only that these boards will at minimum simply pass information from the front-end detectors through to the L1 track finder system. Consequently, the AM+FPGA data delivery platform must de-bunch the 8 BX stub packets sent from the CICs (Section 3.2.2). The AM ASICs are assumed to provide ~ 150 k patterns each, which is consistent with the design goals of the two (28 nm planar and two-tier) CMS AM R&D efforts.

After fixing these general design parameters, the performance of the AM+FPGA approach can be explored in both simulation and hardware. Section 9.4.1.1 details the simulation and optimization studies of AM pattern recognition and of downstream FPGA operations that were performed to support the demonstration. Section 9.4.1.2 reviews the hardware implementation used in the demonstration. Demonstration results are presented in Section 9.4.1.3.

9.4.1.1 Simulation and system optimization

A software simulation of AM+FPGA pattern recognition has been developed to establish performance expectations and to guide the design of the overall L1 tracking system. The simulation covers the full tracker and includes all operations envisioned for the Pattern Recognition Mezzanine (PRM) cards. The mapping of front-end modules to trigger towers has been optimized so as to minimize the need for data sharing between towers and to reduce pattern bank sizes. With the trigger tower definition, the sharing of single module data is limited to at most four towers and ensures that a track with $p_{\mathrm{T}} > 2\,\mathrm{GeV}$ and $|\eta| < 2.5$ is contained in at least one tower.

In the AM+FPGA approach, pattern recognition is performed by the AM ASICs, which compare input stubs against pre-defined banks of stub patterns from valid tracks. Each trigger tower in the AM+FPGA system has its own pattern bank. The AM pattern recognition must provide high-quality track candidates and yet limit both pattern bank size (to reduce demands on the ASIC) and the number of matched roads and stub combinations in a given event (to reduce demands on downstream track fitting). The ideal pattern bank is small, efficient, and effective in reducing combinatorics. These characteristics define a figure of merit, sketched in Fig. 9.62, that is used to quantitatively assess the quality of a bank. The vertical axis of Fig. 9.62 indicates the number of patterns per trigger tower that must be stored. The horizontal axis corresponds to the overall number of stub combinations that must be handled by the downstream FPGA logic. The optimization parameters include the φ-width of the super-strip given in terms of a geometric scale factor ("sf") value, the z-segmentation ("nz") of the super-strip, and a factor by which similar patterns are merged ("mX") either for the full pattern bank or after truncating the bank to remove patterns that fire infrequently. The AM pattern bank optimization consists of finding sets of operable banks that are as close as possible to the origin of this plot. In the current optimisation, bank sizes range from 0.5 M to 1 M patterns per tower. A total of 24–48 M patterns would be needed for the full tracker if a $p_{\mathrm{T}} > 3$ GeV threshold is used. The number of patterns would need to increase by approximately 30–50% for a 2 GeV threshold.

Following AM pattern recognition, road-matched stubs can be used as input to track fitting

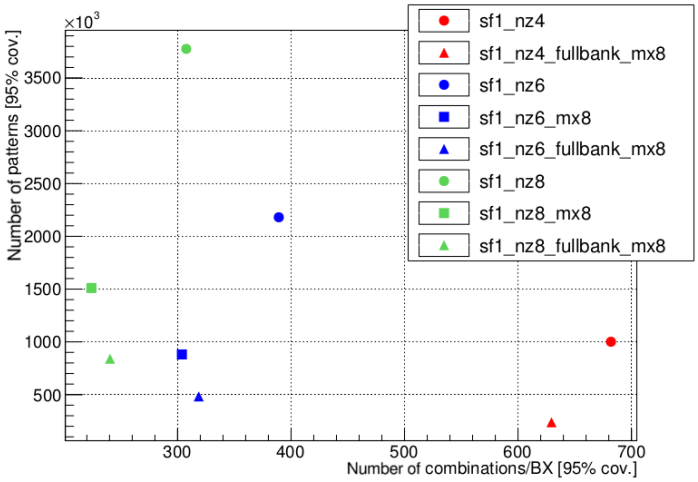


Figure 9.62: The figure of merit (the number of patterns versus number of combinations with 95% coverage) used to optimize pattern bank definitions in the AM+FPGA approach. Points near the origin of this plot represent the ideal scenario of small bank size and low stub combination multiplicity. The bank optimization parameters shown in the legend are defined in the text.

algorithms. Track fitting is performed with at most one stub per tracker layer. Due to the coarseness of the AM patterns, however, multiple stubs per layer can be matched to a given road. The Combination Builder (CB) is responsible for producing one or more sets of stubs with at most one stub per layer for each fired road. Two approaches to combination building have been explored for the demonstration. The simplest approach is to pass all possible combinations of stubs to downstream track fitting for each fired road. To preserve tracking efficiency in cases in which a stub is not registered by the detector, all sets of stubs in at least five out of six tracker layers are constructed, while ensuring that individual combinations are not repeated. This is the approach taken in the "Advanced Combination Builder" (ACB). An alternative approach involves the reduction to a single stub combination for each fired road. The "Track Candidate Builder" (TCB) achieves this by creating 3D seeds from pairs of inner-layer stubs within the AM roads. These seeds are projected to the outer layers and stubs within the road with the smallest residuals are associated with the trajectory. If multiple stub combinations remain after this step, the one with the widest road size and smallest residuals is selected. This procedure provides a unique stub combination for a given road. Both CB techniques have been extensively studied in simulation, and both have been deployed in the demonstration.

The stub combinations produced by the CB are passed to a linearized track fit for precise track parameter determination. The fitting approach utilizes a principal component analysis (PCA) technique that obtains a vector of track parameters by multiplying a vector of stub coordinates by an appropriate translation matrix. The matrix multiplication operations involved in track fitting can be executed with low latency using Digital Signal Processing (DSP) resources in modern FPGAs [41, 42]. In a simple linearized track fit, track parameter resolution is typically degraded due to non-linearities in detector geometry. A novel approach has been developed to correct stub coordinates such that each tracker layer is effectively flattened into a perfectly cylindrical barrel. Figure 9.63 contrasts the original and transformed stub positions. The removal of detector non-linearities enables both excellent tracking performance and a significant reduction in the number of matrix constants needed for track fitting.

Track finding operations include a degree of redundancy so as to ensure high tracking effi-

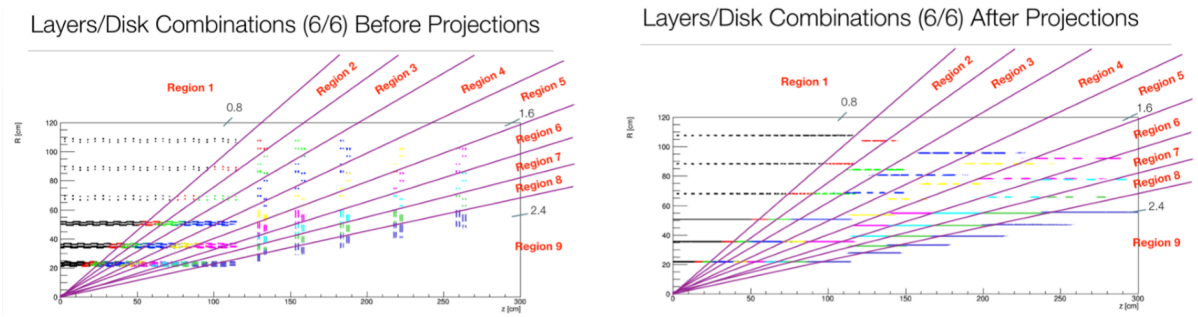


Figure 9.63: Stub positions in r-z view before (left) and after (right) the geometric corrections applied in track fitting. The corrections smooth the non-linearities in stub positions, yielding a uniformly cylindrical tracking geometry.

ciency. Consequently, a given track will typically be found multiple times. Pairs of tracks are considered to be duplicates if the tracks share a minimum number of stubs, where this minimum depends on the total number of stubs in the tracks and on stub type (from 2S or PS modules). For such pairs, the track with the larger reduced χ^2 is rejected. Duplicate removal (DR) is achieved by performing this pair-wise comparison/filtering for all tracks in the trigger tower.

The simulation of the full AM+FPGA pattern recognition and tracking chain, including both combination building solutions, is available within CMSSW. The framework includes a bitwise emulation of the AM06 chip [122]. Complete sets of banks for the $6\,\eta \times 8\,\varphi$ trigger tower scenarios are also available. In parallel, a flexible and standalone AM simulation framework has been developed to facilitate the optimization of the pattern banks. The frameworks provide an option to enable truncation effects (e.g. on the maximum number of roads or on the maximum number of stub combinations), as are experienced with the hardware implementation.

9.4.1.2 Hardware demonstration system

The hardware demonstration of the AM+FPGA approach implements the operations described in Section 3.5.1.1, including data sourcing, data delivery, pattern recognition, and track finding. Figure 9.64 shows the primary hardware configuration used in the demonstration. The data delivery platform in this system is implemented with Fermilab Pulsar2b ATCA boards [123, 124], which are equipped with Xilinx Virtex-7 690T FPGAs. The Pulsar boards are used both as Data Sources (DS) and as Pattern Recognition Boards (PRBs) in the demonstration. Ten DS boards are used to emulate the transmission of front-end stub data from DTCs corresponding to a single barrel tower. Each DS board sends its data over $40 \times 10\,\text{Gb/s}$ links to one of ten PRBs in a second ATCA shelf. The DS boards and PRBs are synchronized by means of CERN TTC mezzanine cards [125], which are supported on additional Pulsar2b boards in the ATCA shelves. The TTCci cards receive a 40 MHz LHC clock from a CERN TTCci board, which is distributed to all boards in the shelf over dedicated clock lines on the ATCA backplane.

The PRBs receive DS data via optical connections on their Rear Transition Modules (RTMs). The RTM links are implemented using the Aurora 64b/66b protocol, which provides a significantly more efficient means of high-speed data transfer, relative to conventional 8b/10b protocols. The PRB shelf is equipped with a Comtel 40G+ full-mesh backplane that enables point-to-point communication between each of the ten PRBs. Backplane links in the PRB shelf are also implemented using the Aurora 64b/66b protocol running at 10 Gb/s. The PRBs support two dual-width Pattern Recognition Mezzanine (PRM) cards that host the AMs and FPGAs for tracking operations. Each PRM is assigned a single bunch crossing for processing.

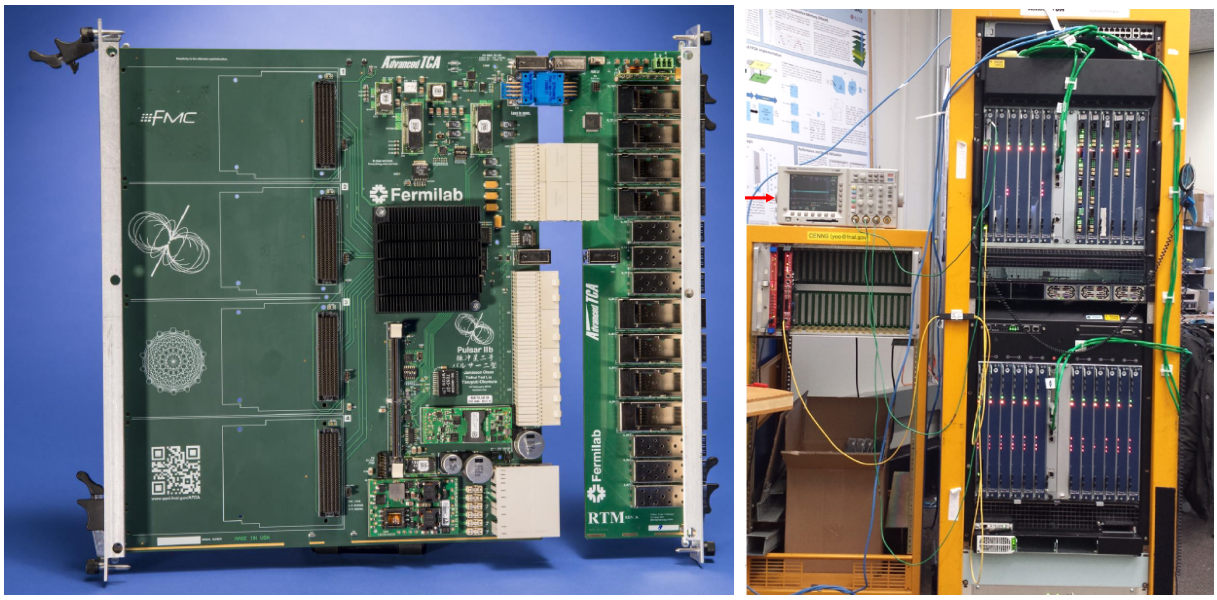


Figure 9.64: The Fermilab Pulsar2b ATCA board (left) is used in the AM+FPGA demonstration both as DTC-equivalent data sources (DSs) and as pattern recognition motherboards (PRBs). The primary demonstration system shown at the right comprises ten Pulsar DS boards (in the rightmost rack, bottom shelf) connected to ten PRBs (in the rightmost rack, upper shelf). Synchronization is achieved using a CERN TTCi (in the leftmost rack) and TTCi mezzanine cards on Pulsar boards in both ATCA shelves.

The overall time multiplexing factor in the AM+FPGA demonstration is therefore $f_{\mathrm{TMUX}} = n(\mathrm{PRB/tower}) \times n(\mathrm{PRM/PRB}) = 20$. Data delivery operations are fully pipelined, and each stage is designed to complete within the 200 ns window set by the 8 BX data transmissions from the DS. The time-multiplexing period of 20 nominally allows 500 ns for all operations on the PRM. This time window can be expanded by increasing the parallelism of track fitting operations on the PRM, as discussed below.

Each PRB must first determine whether the stub data it receives on its RTM belongs to one of the bunch crossings it is assigned. If this is so, the stubs must be formatted and routed to the designated PRM for processing. If the data instead corresponds to a different combination of PRB and PRM, the stubs must be formatted and routed to the appropriate PRB. Stub communication between the PRBs of a given tower proceeds via the backplane. Each PRB accepts and reformats the stub data it receives on the backplane, and then delivers this data, along with that received on the RTM, to the target PRM.

Two types of PRMs are used in the AM+FPGA hardware demonstration. The Fermilab Ultrascale PRM (Fig. 9.65, left) contains two Xilinx Ultrascale Kintex KU060 FPGAs and a socket that allows for the eventual integration of two-tier AM prototypes [126]. For the demonstration, a fully pipelined and cycle-accurate emulation of two-tier AM operations runs in one of the FPGAs, while the other implements the Data Organizer (DO), Combination Builder, and track fitter (TF) functionalities. Given the amount of block RAM (BRAM) available on the KU060, 1024 patterns can be stored in the emulation. This suffices for the demonstration of AM+FPGA latency, because even in a full-scale system only a small number of patterns will be matched in a given BX. Moreover, the time needed for pattern matching is independent of the size of the stored bank. Matched patterns for a particular BX are determined in advance from simulation using full pattern banks. These patterns are loaded in the emulation, ignoring those that do not match. This procedure allows a subset of events to be fully demonstrated with a single

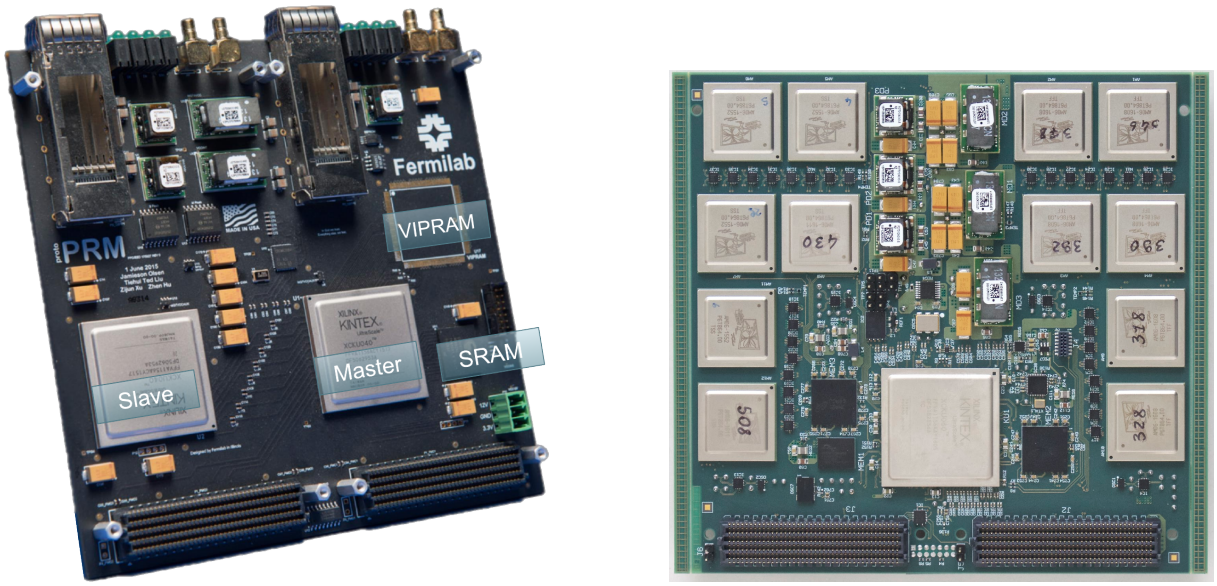


Figure 9.65: The Fermilab (left) and INFN PRM06 (right) pattern recognition mezzanine (PRM) cards.

configuration of the AM emulation.

All firmware running on the second KU060 is also fully pipelined. The design of the DO component utilizes newly supported read/write features in Xilinx block BRAMs to provide for low latency stub storage and retrieval. The track fitting pipelines are implemented using cascading DSP operations and small distributed RAMs for the matrix coefficients in order to minimize latency and resource usage. In the nominal demonstration scenario, the DO distributes roads (via 1 FIFO) to four pairs of CB and TF components that run in parallel. This scenario is referred to as $N_{\text{FIFO}} = 1$. The $N_{\text{FIFO}} = 1$ design runs at 240 MHz in the KU060 and the level of resource utilization for this scenario is comfortable (e.g. 20–30% FF (flip-flops), LUT (look-up table), and DSP). Alternative scenarios ($N_{\text{FIFO}} = 2, 3$) in which the DO distributes (via additional FIFOs) to two or three sets of four CB+TF pairs are also explored. The additional processing resources in these scenarios are used to provide an additional degree of time-multiplexing within the PRM, extending the time for tracking operations from 500 ns to 1000 ns and 1500 ns, respectively. Although the $N_{\text{FIFO}} = 1$ configuration is sufficient for processing $\text{t}\bar{\text{t}}$ + 200 PU events, the longer processing times in the $N_{\text{FIFO}} = 2, 3$ scenarios can be very useful for processing events with large numbers of stub combinations (e.g. in events containing high-p_{T} jets).

In addition to the Fermilab PRM, the INFN PRM06 mezzanine (Fig. 9.65, right) was also studied in the demonstration. This mezzanine card hosts 12 AM06 associative memory chips, a Xilinx Kintex Ultrascale KU060 FPGA, and two RLDRAM3 1.125 Gbit external memories. As with the Fermilab PRM, the PRM06 connects with the Pulsar2b via two FMC connectors, both of which carry data signals and power supply voltages. Six bi-directional high-speed serial links (three links per each FMC connector) are used to send and retrieve data to/from the PRM, and were tested up to 12.5 Gb/s. An additional 34 LVDS lines are used for slow control. Each AM06 chip is pre-loaded with 128 k patterns (for a total of 1.5 M patterns per PRM), which is sufficient to cover an entire trigger tower. Each pattern is constructed from eight independent 16-bits words, one per tracker layer.

Stubs are distributed from the FPGA to the AM06 chips using a cascade of fan-outs. It is also possible to send different set of stubs to two groups of six AM06 chips independently. Matched roads are read out from each AM06 chip in parallel, in order to minimize latency. The serial

Table 9.9: Cumulative latency of the processing stages in the AM+FPGA demonstration for $t\bar{t}$ + 200 PU events. Results are provided in terms of both the beginning (t_{first}) and the end (t_{last}) of the output from the various processing stages.

Processing stage	t_{first} [μs]	t_{last} [μs]
Data sourcing	0	X
Data delivery	1.2	1.7
Pattern recognition	1.85	2.3
Tracking $N_{\text{FIFO}} = 1$	2.03	2.53
$N_{\text{FIFO}} = 2$		3.03
$N_{\text{FIFO}} = 3$		3.53

links between the FPGA and the AM06 chips were tested at their nominal operating speed of 2 Gb/s. A working firmware suite has been developed, which implements local-to-global coordinate conversion, a Data Organizer, 12 instances of the Track Candidate Builder, and one PCA track fitter. The demonstration firmware operates at 200 MHz and occupies roughly 50% of the resources of the KU060 FPGA. The AM06 was not designed for low-latency L1 applications, thus detailed timing studies involving the PRM06 were not performed. Instead, the PRM06 offers an opportunity for present-day tests with real AM ASICs, and for system development that includes the I/O and control of multiple chips.

The bit-level emulations of all PRM operations have been made to match the hardware outputs of the Fermilab and INFN PRMs. All truncation effects are captured by the emulations and, where present, were introduced in the simulation via equivalent cuts on the stub, combination, and track multiplicities. Having achieved perfect agreement between hardware, emulation, and truncated simulation in the barrel for a representative subset of events, simulation can then be used to reliably extrapolate tracking performance to the full tracker.

9.4.1.3 Demonstration results

The latency of AM+FPGA tracking has been characterised with a two-shelf demonstration system. This system includes all of the data sourcing, data delivery, and pattern recognition functionalities inherent to the approach. The system also includes firmware implementations of nearly all of the FPGA tracking logic (DO, CB, TF); only the duplicate removal step, which is estimated to contribute about 100 ns of latency, has not yet been implemented. Total latency using the two-shelf system is determined from the start of stub transmission from the DS shelf to the time of the first (t_{first}) or last (t_{last}) stub output from track processing on the Ultrascale PRM. Intermediate latency measurements are performed at each of the relevant internal processing stages. For those measurements, t_{first} and t_{last} refer to the beginning and the end of the output of the respective stages.

Table 9.9 summarizes the latency results obtained for a sample of $t\bar{t}$ + 200 PU events. Latency accumulates when descending rows in the table. The results incorporate all link latencies necessary for full system operation, including that for communication between the PRB and PRM, which was not implemented in the demonstration system. The 64b/66b links were separately characterised and found to contribute $\sim$ 150 ns of latency (stemming primarily from serializer/deserializer operations) per communication stage. This value was added to the measured latency for data delivery and is reflected in the table. Importantly, the data sourcing and delivery latencies incorporate all DTC operations relevant for L1 tracking, including that due to data transfer to the PRBs and that resulting from the de-bunching and sorting of the 8 BX front-end data structures.

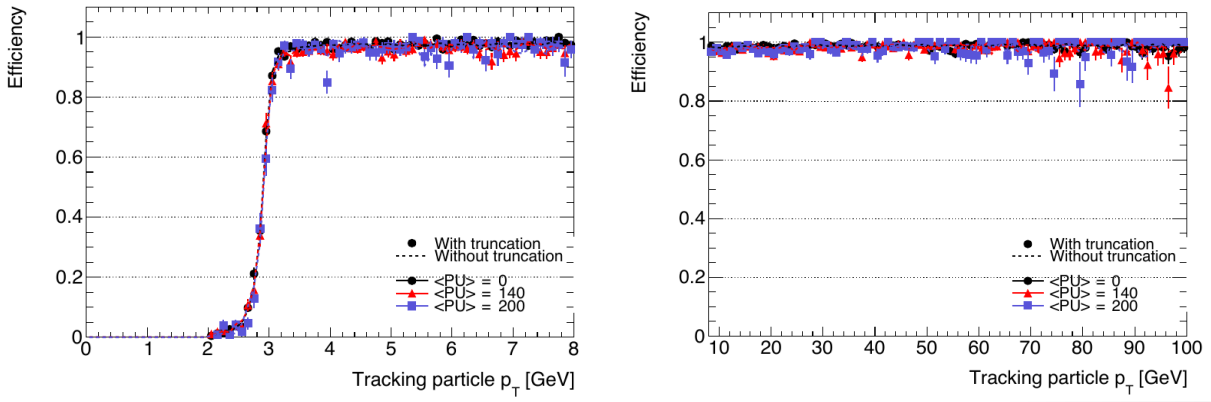


Figure 9.66: Tracking efficiency as a function of p_T as measured in the AM+FPGA hardware demonstration for muons with (left) $p_T < 8\,\text{GeV}$ and (right) $p_T > 8\,\text{GeV}$ in $t\bar{t}$ + pileup events.

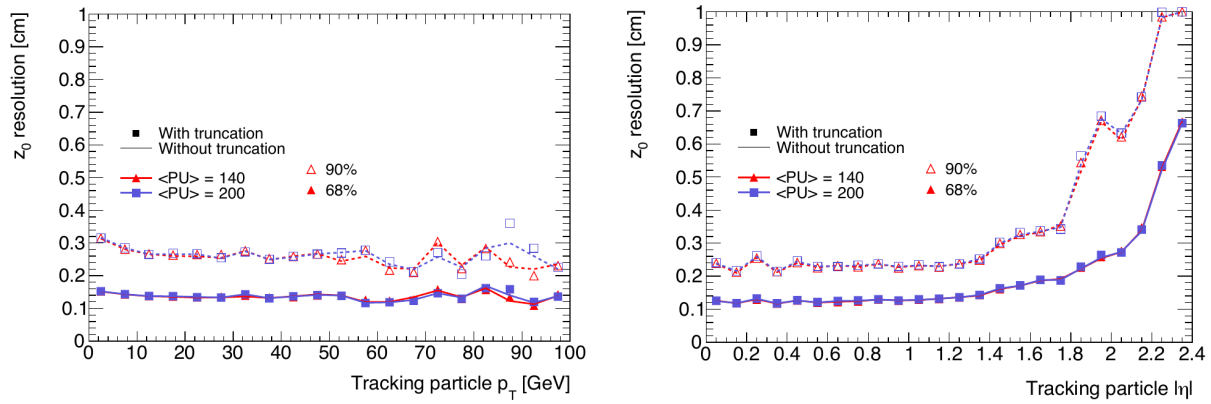


Figure 9.67: The z_0 track parameter resolutions for muons in $t\bar{t}$ + pileup events as a function of the muon p_T (left) and pseudorapidity (right) from the AM+FPGA demonstration.

The overall latency for the nominal $N_{\text{FIFO}} = 1$ scenario is approximately 2.5 μs, which is comfortably lower than the 4 μs target. No data truncation effects are observed with the $t\bar{t}$ + 200 PU dataset in any of the N_{FIFO} scenarios explored. As was discussed in Section 9.4.1.2, the latencies of the $N_{\text{FIFO}} = 2$ and $N_{\text{FIFO}} = 3$ scenarios are larger than that of the nominal scenario by design; the additional track processing capabilities in these alternative scenarios provide for additional time-multiplexing within the PRM, which increases latency but which could be desirable for processing higher occupancy events, as discussed below.

The tracking efficiency obtained with the full-tracker AM+FPGA emulation is shown in Fig. 9.66. These results correspond to muons within the $t\bar{t}$ + 0, 140, and 200 PU samples, split by muon p_T. Figure 9.66 shows a sharp efficiency turn-on at 3 GeV, the track p_T threshold used for the demonstration. Beyond this, tracking efficiency plateaus at 95–100%, with little degradation observed due to increasing pileup.

Figure 9.67 shows the z_0 resolutions obtained for fit tracks from the same $t\bar{t}$ samples. Resolutions of $\sim$ 1 mm are obtained for muons with $|\eta| < 2$. In addition, all fit tracks obtained from the $t\bar{t}$ samples have been used as input to an offline vertex finding algorithm [13] that approximates the procedure that is anticipated to run in the downstream L1 trigger. Using this procedure, the z_0 resolution of reconstructed vertices is observed to be $\sim$ 0.5 mm.

The L1 tracking performance observed with the $t\bar{t}$ + PU samples is excellent, which demonstrates that the approach is fully capable of operating in busy physics environments within

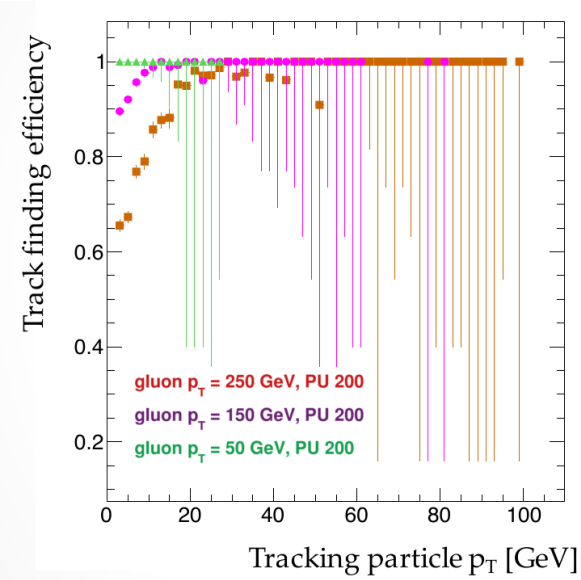


Figure 9.68: Tracking efficiency as measured in the AM+FPGA hardware demonstration for high-p_T jet + PU samples. The gluon p_T is 50 GeV (green), 150 GeV (violet), and 250 GeV (red).

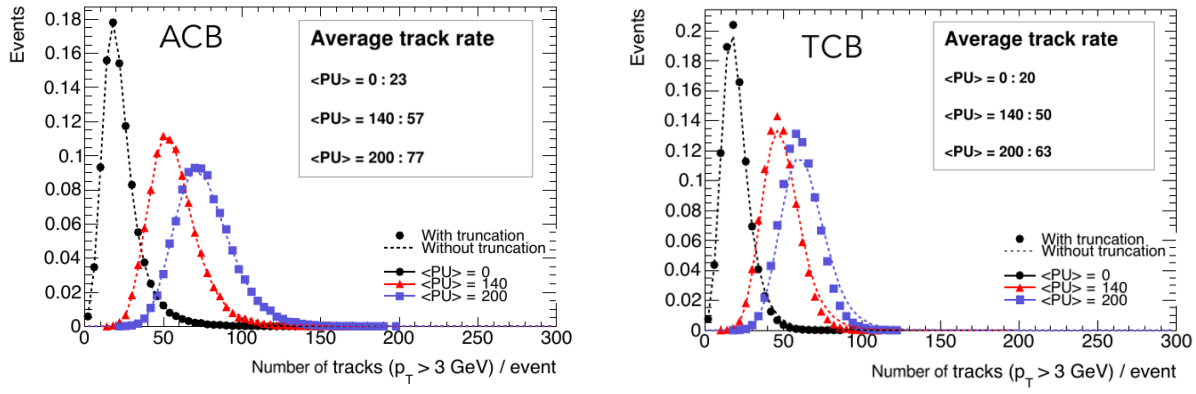


Figure 9.69: Track rates for $t\bar{t}$ events as measured in the AM+FPGA hardware demonstration with the ACB (left) and TCB (right) combination builders.

the target latency. Additional studies were performed in order to identify and understand the conditions under which the performance of the proposed system begins to degrade. The studies utilized samples of single gluon events with 200 PU events and a gluon p_T ranging up to 250 GeV. Figure 9.68 shows the tracking efficiency obtained from these samples for the $N_{FIFO} = 2$ configuration. In all cases, high-p_T tracks are found with $\sim$ 100% efficiency and with good resolution, however the efficiency and resolution of low-p_T tracks within high-p_T (> 150 GeV) jets clearly degrade. This degradation stems from truncation in the pattern recognition stage, and is concentrated at low-p_T because the AMs have been configured to prioritise the readout of high-p_T roads. Future system optimization, such as increasing parallelism within the PRM (e.g. via $N_{FIFO} = 3, 4$ scenarios), are expected to mitigate these effects.

The hardware track rates obtained from the $t\bar{t}$ + PU samples are shown in Fig. 9.69. These results correspond to system operation with the two stub combination builders introduced before, the Advanced Combination Builder (ACB) and the Track Candidate Builder (TCB). In both cases, the average number of tracks with $p_T > 3$ GeV output per event is less than 80 for a pileup of 200, a rate that should be easily accommodated by the downstream L1 trigger.