Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 37 (2012) 1973 – 1982

Physics Procedia

TIPP 2011 – Technology and Instrumentation for Particle Physics 2011

A New Concept of Vertically Integrated Pattern Recognition Associative Memory

Ted Liu*, Jim Hoff*, Grzegorz Deptuch, Ray Yarema

Particle Physics Division, Fermilab, P.O. Box 500, Batavia, IL 60510 USA

Abstract

Hardware-based pattern recognition for fast triggering on particle tracks has been successfully used in high-energy physics experiments for some time. The CDF Silicon Vertex Trigger (SVT) at the Fermilab Tevatron is an excellent example. The method used there, developed in the 1990's, is based on algorithms that use a massively parallel associative memory architecture to identify patterns efficiently at high speed. However, due to much higher occupancy and event rates at the LHC, and the fact that the LHC detectors have a much larger number of channels in their tracking detectors, there is an enormous challenge in implementing fast pattern recognition for a track trigger, requiring about three orders of magnitude more associative memory patterns than what was used in the original CDF SVT. Scaling of current technologies is unlikely to satisfy the scientific needs of the future, and investments in transformational new technologies need to be made. In this paper, we will discuss a new concept of using the emerging 3D vertical integration technology to significantly advance the state-of-the-art for fast pattern recognition within and outside HEP. A generic R&D proposal [1] based on this new concept, with a few institutions involved, has recently been submitted to DOE with the goal to design and perform the ASIC engineering necessary to realize a prototype device. The progress of this R&D project will be reported in the future. Here we will only focus on the concept of this new approach.

Keywords: 3D Integration; Content Addressable Memories; CAM; Tracking Trigger; Associative Memory, Pattern Recognition

* *Email*: thliu@fnal.gov, jimhoff@fnal.gov

1875-3892 doi:10.1016/j.phpro.2012.02.521

1. Introduction

Future particle physics experiments looking for rare processes will have to address the demanding challenges of fast pattern recognition in triggering as detector hit density becomes significantly higher due to the high luminosity required to produce the rare processes. The Large Hadron Collider (LHC) at CERN has proposed a luminosity increase of a factor of ten over the original design as the goal for the upgrade, which will result in a corresponding increase in particle interactions and track densities in the detector. Most of these interactions contain events that are of no significance and should not be recorded. Since the quantity of data that can be stored for later analysis is limited, real-time event selection is imperative to retain the interesting events while rejecting the background and the capability to perform fast pattern recognition and track reconstruction of particle trajectories will be crucial.

The ultimate physics reach of the LHC experiments will crucially depend on the tracking trigger's ability to help discriminate between interesting rare events and the background. The CMS muon trigger, for example, will reach an unacceptably large rate at high luminosity due to the number of hits in the muon detectors. The first-level trigger can be reduced to an acceptable level if tracks are found in the inner detector and matched to the muon candidates. There are other important reasons for having tracking trigger capabilities at early stages of the trigger system. For example, the online identification of heavy fermions such as *b* quarks and tau leptons are important, since many interesting channels of new phenomena produce heavier elementary particles. Tracks coming from a secondary vertex not in the direction of the beam line identify a *b* quark. Tau jets can be separated from background using the number of tracks within a narrow "signal cone" and the number in a larger "isolation region".

Hardware-based pattern recognition for fast triggering on particle tracks has been successfully used in high-energy physics experiments for some time. The CDF Silicon Vertex Trigger (SVT) at the Fermilab Tevatron is an excellent example [2][3]. The method [4] used there, developed in the 1990's, is based on algorithms that use a massively parallel associative memory architecture to identify patterns efficiently at high speed. However, due to much higher occupancy and event rates at the LHC, and the fact that the LHC detectors have a much larger number of channels in their tracking detectors, there is an enormous challenge in implementing pattern recognition for a track trigger [6], requiring about three orders of magnitude more associative memory patterns than what was used in the original CDF SVT. Significant improvement in the architecture of associative memory structures is needed to run fast pattern recognition algorithms of this scale. Scaling of current technologies is unlikely to satisfy the scientific needs of future projects, so investments in transformational new technologies need to be made.

In this paper, we are proposing a new concept of using 3D integrated circuit technology as a way to implement associative memory structures for fast pattern recognition applications. Adding a "third" dimension opens up the possibility for new architectures that could dramatically enhance pattern recognition capability. While our focus here is on the Energy Frontier (e.g. the LHC), the approach may have applications in experiments in the Intensity Frontier and the Cosmic Frontier as well as other scientific and medical projects. In fact, the technique that we are proposing is very generic and could have wide applications far beyond track trigger, both within and outside HEP.

2. The Associative Memory Approach: very fast track reconstruction

Typical track reconstruction in a tracking detector consists of two steps: pattern recognition followed by track fitting. Pattern recognition involves choosing, from all the hits present in the detector, those hits that were potentially caused by the same particle. This stage produces a set of "hits of interest", typically one to a few hits per detector layer depending on the hit resolution used. In the coarse resolution that is typically used at this stage, one bin can contain more than one actual hit. Track fitting involves extracting

track parameters from the coordinates of the "hits of interest". For cases for which time constraint is not so stringent, track reconstruction has been implemented using software computational techniques to identify patterns and perform track fitting, often using processors running in the upper levels of a data acquisition system to perform the task. However, such algorithms are usually time-consuming because the pattern recognition and track fitting steps are necessarily executed many times to find and fit all the tracks for each event. The software approach using processors are in general not suitable for fast tracking trigger applications due to the time constraints. As will be described below, with the associative memory approach, track reconstruction is made much faster by exploiting massive parallelism [4]. This is achieved in such a way that the pattern recognition is done by using associative memory devices, while the track fitting is done by using a simple liner approximation to the actual fitting of the analytical expression of the track trajectory to the hit locations in the detector.

The pattern recognition step is usually the most time consuming task, because one needs to test many combinations of hits to find those that potentially come from the same track and typically these tests are done sequentially (e.g. in software). The associative memory approach allows the testing of all hit combinations in parallel against a set of known patterns. To illustrate the concept of patterns in the associative memory approach, one can use an oversimplified case with a simple detector consisting of six

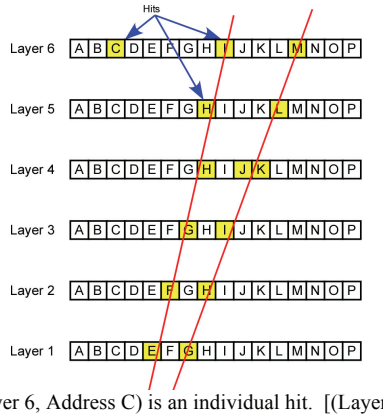

Fig. 1 - Tracks in a Tracking Detector: (Layer 6, Address C) is an individual hit. [(Layer 1, Address E), (Layer2, Address F), (Layer 3, Address G), (Layer 4, Address H), (Layer 5, Address H), (Layer 6, Address I)] are all hits from the same Track. A Road designed to discover such a Track would store [(Layer 1, Address E), (Layer2, Address F), (Layer 3, Address G), (Layer 4, Address H), (Layer 5, Address H), (Layer 6, Address I)] in the associative memory.

detector layers, as shown in Fig. 1. A charged track crossing the detector would produce a set of hits (pattern). A finite set of distinct patterns can be generated this way using valid tracks for a given experiment, and such sets are often called the pattern bank.

The Associative Memory (AM) architecture is based on Content Addressable Memory (CAM) cells [9][10] to efficiently identify track patterns (roads) at high speed using coarse-resolution "hits" recorded in the tracking detector. A block diagram [4] of the Associative Memory architecture is shown in Fig. 2 (for a case with four detector layers). Each pattern (shown as a cell) is composed of four hit coordinates each of which is stored in the CAM word for a given layer, and only four patterns (cell 0 to 3) are shown in Fig. 2. An incoming hit from a given detector layer is transmitted to the corresponding layer and the hit coordinate is compared against the stored words for all patterns in parallel for that layer. Any match to each incoming hit will be latched for that layer and for that pattern until reset (to rearm for next event). This process is repeated for all the incoming hits for each detector layer as the hits arrive one after the other. As soon as all hits from the same event are received, the hit matching stage is done and all latched

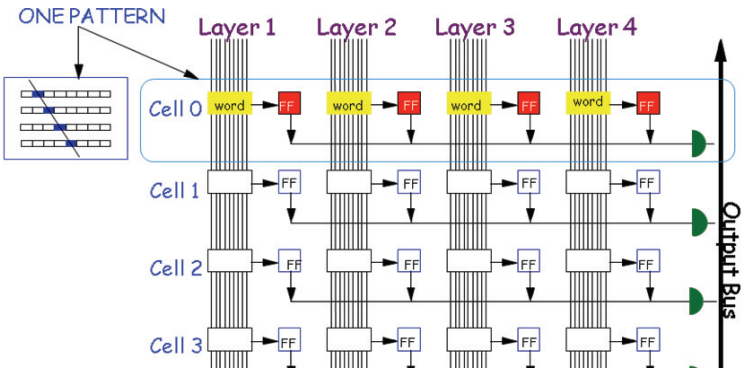

Fig. 2 - A Block Diagram of an Associative Memory Chip. The CAM Cells are shown as white (or yellow) boxes. The Majority/Glue Logic is shown to the right as green semicircles.

matches for each pattern (or cell) will be fed into a majority logic stage where a fired road will be found if the number of matched layers reaches a programmable threshold.

Associative Memory is sometimes called PRAM, Pattern Recognition Associative Memory. The AM method solves the combinatorial challenge inherent to the pattern recognition by exploiting massive parallelism of associative memories that can compare tracking detector hits to a set of pre-calculated patterns simultaneously. The found patterns or "fired roads" are then processed using fast FPGAs to perform track fitting with full detector resolution using all combinations of the "hits of interest" from the fired roads. Because each pattern or road is narrow enough, the usual helical fit can be replaced by a simple linear calculation. The track fitting stage for each matched pattern is much simplified and can be very fast [5].

3. Limitations of the 2D Approach

A critical figure of merit for an AM-based track reconstruction system is the number of predetermined track patterns or roads that can be stored in the Associative Memory bank. Generally speaking, wider roads using coarser resolution hits require less AM storage, but the number of AM roads satisfied by random hits and the number of fits at the track fitting stage downstream increases quickly due to the high detector occupancy. Also, the demand on the bandwidth would be higher because all the roads and hits have to be transferred from the AM stage to the track fitting stage. If the roads are very narrow, due to using finer resolution hits, the number of fake roads and fits are reduced, but the required total size of the AM would increase dramatically. Therefore, the road width must be optimized. The required AM pattern bank size will be different for different experiments and even different for the same experiment at different luminosities. For CDF SVT upgrade, the AMchip03 [7] was developed using 180 nm CMOS technology and standard cell based approach. The AMchip03 has 5K patterns for six detector layers and could work up to 50MHz. A new version, AMchip04, is currently being developed [8] using 65 nm for the Atlas FTK project [6].

Future associative memory designs will require many more stored patterns per unit area at less power per pattern and at greater speed. This is by no means a simple task, but the three elements – pattern density, power and speed – are related to one another through geometry. Obviously, with smaller feature sizes, more associative memory cells can be made in the same area. Furthermore, both power and speed are directly proportional to load capacitance which is itself related to feature size. Therefore, the logical approach to the requirements of future associative memory designs is to build them in smaller and smaller feature sizes. However, this approach eventually fails both economically and technically. Economically,

each new process node is averaging a factor of 2.5 times in production cost over the preceding technology node. Technically, the scaling of VLSI circuits reduces gate delay, but increases interconnect delay [11]. Load capacitance is a sum of the gate capacitance of any load gates, the diffusion capacitance of the driving gates, and the parasitic capacitance of the interconnect wires. As feature sizes get smaller and smaller, interconnect capacitance quickly begins to dominate, so load capacitance and therefore power and speed stop scaling with technology node [11]. Therefore, the ultimate solution is a conservative approach to feature size reduction side-by-side with an aggressive approach to interconnect reduction. Simply put, this is not possible in two dimension. It is a simple fact of geometry that for a given feature size and only two dimensions to work in, the design must be spread out.

3D technology is the integration of thinned and bonded silicon integrated circuits with vertical interconnects between IC layers using Through Silicon Vias (TSVs) [11][12]. The technology has wide applications in industry, ranging from memories to pixel arrays to microprocessors and FPGAs and it is a cornerstone of the International Technology Roadmap for Semiconductors [13]. First and foremost, integrating L layers one above another obviously gives a designer access a factor of L increase in area. At the same time, interconnect in three dimensions permits a significant reduction in wire resistance and capacitance and, consequently, interconnect delay [11]. As Moore's law is approaching severe limitations, it is expected that 3D technology will be the next scaling engine. Even better, it provides the freedom to divide functionality among tiers to create new designs that are simply not possible in 2D.

4. 3D Associative Memories

Associative memory chips can be said to belong to the same class of integrated circuits as SRAM and DRAM chips. They are large arrays of smaller cells that are reproduced many times and are ordered in a fashion that is periodic in two dimensions. Moreover, these smaller cells are mainly connected to peripheral cells and do not interact much with one another. Consequently, this makes any 3D memory or associative memory design different than, for example, 3D microprocessor design. Intuitively, an appropriate rendering of the repeated cell in 3 dimensions can be expected to yield significant benefits to the overall design and therefore deserves considerable attention.

Methods for dividing a design into 3-dimensions are, at this time, largely heuristic. One study [14] defined "critical length" as being a length of interconnect that resulted in an interconnect delay equivalent to the CMOS FO4 delay[b] for a given technology. The metric of a successful 3D design was defined to be the minimization of interconnects longer than the critical length. As an example, the authors of [14] calculated the critical length for 65nm CMOS to be approximately 110μm. Critical length drops significantly with feature size. Fig. 3 is a block diagram of a road pattern matching cell. It is a more detailed view of the highlighted block in Fig. 2 labeled "One Pattern". The road pattern matching cell is all the circuitry necessary to match individual address patterns from all of the layers, store the matches, and resolve and flag a road match. It is that cell in the associative memory design that is repeated many times and ordered in a fashion that is periodic. Within the road pattern match cell, the Stored Address Match Lines (from CAM cells to Majority Logic cell) are long with respect to the critical length and therefore should be shortened by the 3D implementation.

[b] The CMOS FO4 delay is the delay through an inverter that is being driven by an inverter one-fourth its size and which is driving an inverter that is four times its size. Staging inverters by factors of 4x is the accepted way of minimizing delay when driving a large capacitive load.

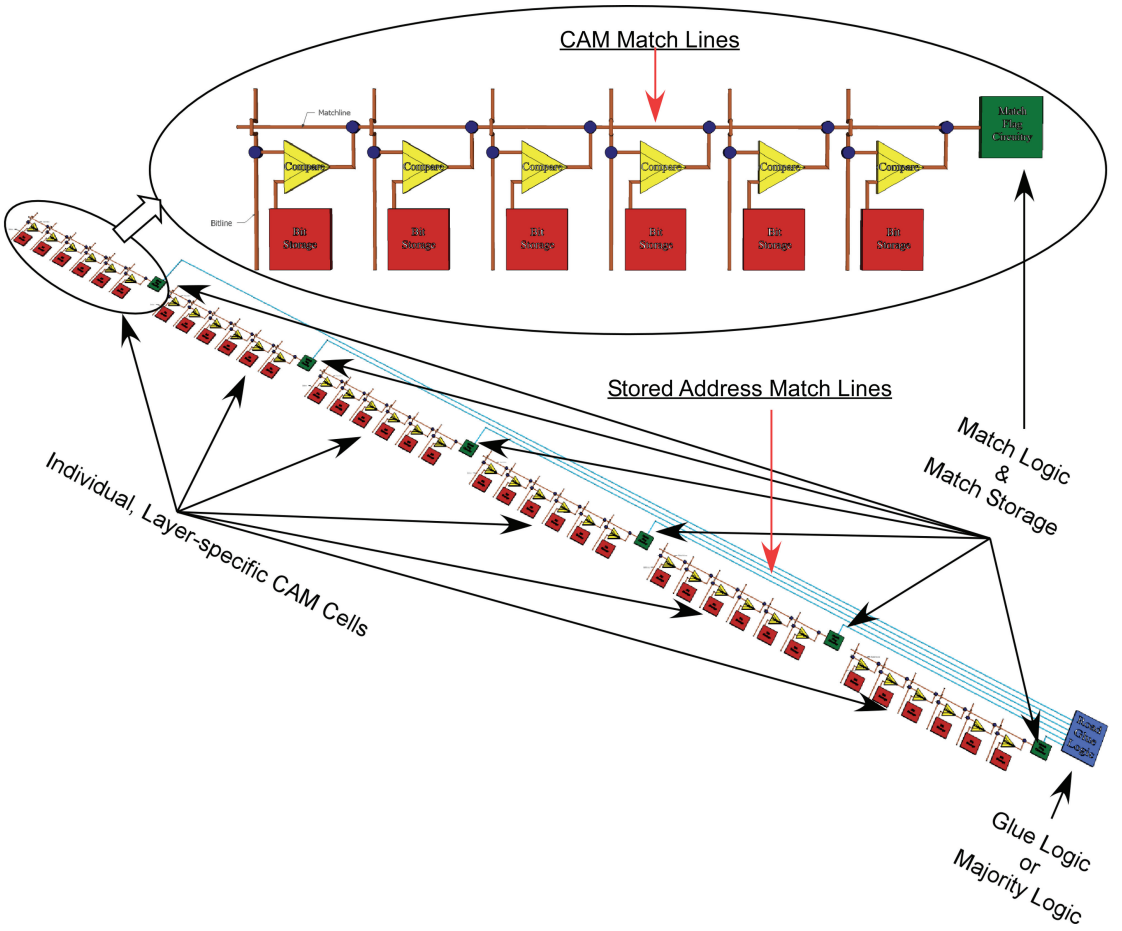

Fig. 3 – A Road Pattern Recognition Cell. A single CAM cell with memory is called out in the inset.

In Fig. 2, all data communication is rectilinear and orthogonal. Column-wise communication consists of the candidate hits on the left and also the "output bus" on the right. Candidate hits from one layer are never compared with stored patterns from another layer and the outputs are only taken from the Majority/Glue Logic cells. Row-wise communication is communication within the road pattern matching cell, and is limited in Fig. 3 to the Stored Address Match Lines. In other words, all column-wise communication is among identical cells whereas all row-wise communication is not. Finally, for a given application, there will be a number of columns equal to the number of detector layers plus one (with Majority Logic) whereas there could be thousands of rows. These two facts suggest that an efficient and cost effective 3D implementation of the associative memory design would be to move each column of Fig. 2 to a separate vertical tier. Each tier would then behave as a traditional CAM with the exception that address matches would be remembered. The inputs to each tier would be the candidate hits from one layer only and they would be routed to all cells on the tier. The vertical connections between tiers would be the Stored Address Match lines in Fig. 3. This would dramatically reduce their length from long with respect to the critical length to very short.

Fig. 4 shows the block diagram of Fig. 3 partitioned vertically side-by-side with a floor plan of a Vertically Integrated Pattern Recognition Associative Memory or VIPRAM chip. The "Individual, layer-

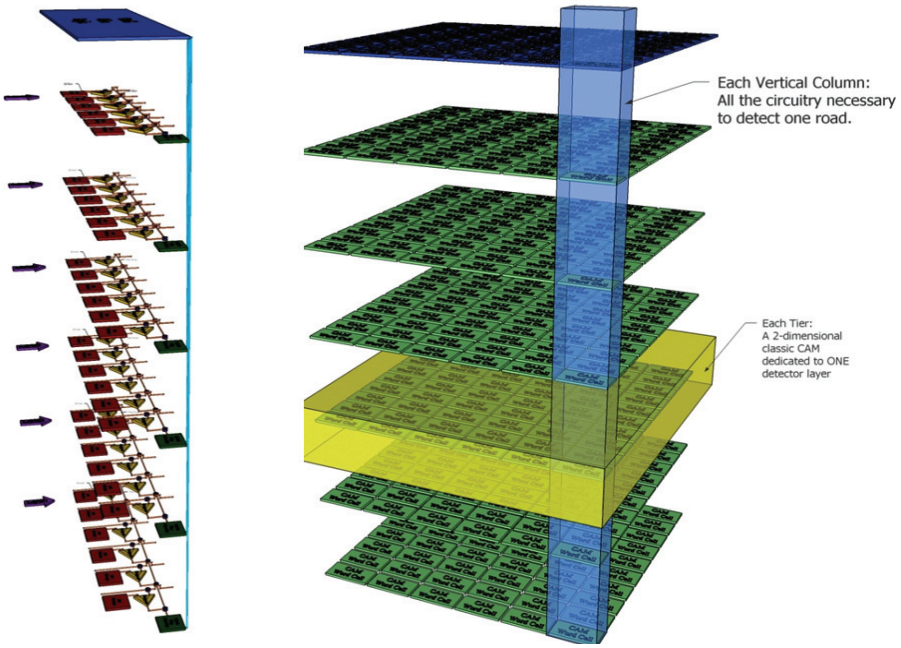

Fig. 4 - A 3D PRAM

specific CAM Cells" of Fig. 3 are now each on their own tier which are called CAM Tiers. In addition to the L CAM Tiers where L is the number of detector layers in the detector system, there is one extra tier. This, obviously, is the tier with the Majority/Glue Logic cells and each cell flags a road. This is referred to as the Control Tier because it also contains all the IO circuitry necessary to control the VIPRAM. The blue vertical tube in Fig. 4 highlights all of the circuitry for a single Road Pattern Recognition Cell. This approach means that in an area approximately equal to the area that once contained only one CAM word cell, a VIPRAM cell can process the L layers of a road pattern. This approach also means that the top tier of the VIPRAM is now resembles a 2-dimensional array of signals that indicate whether or not a road has been flagged. The location of the flag in the 2-dimensional array is indicative of which road has been flagged.

This approach directly shortens the longest of the lines in the road pattern matching cell by shortening the Stored Address Match lines. As these lines are repeated throughout the chip, this has a significant impact on performance. At the same time, this approach makes the layout of the CAM cells, Majority Logic cells, as well as the input and the output busses in Fig. 2 simpler, more uniform and more efficient.

5. Floor Plan and Diagonal Vias

VIPRAM design consists logically of two different types of tiers, the CAM Tiers which hold the CAM cells and the Control Tier which holds the Majority/Glue Logic cells, readout and the IO controls. Looking at Fig. 3, however, reveals that while all of the CAM Tiers might be logically identical, they cannot be physically identical without some additional effort. The Majority Logic on the Control Tier requires a unique input from each CAM Memory cell. Therefore, one unique vertical line is necessary in

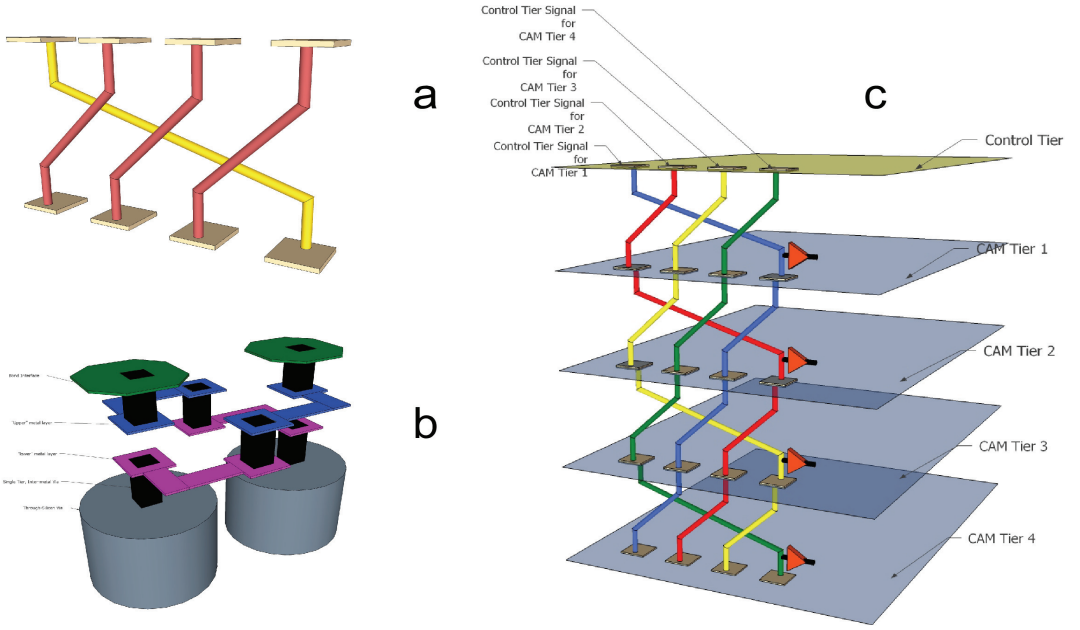

Fig. 5 - Diagonal Vias: (a) shows a cartoon of the function of 4 diagonal vias; (b) shows a simple VLSI implementation of two diagonal vias from the Through-Silicon Vias on the bottom (in gray) up to one layer of metal (purple) to a second layer of metal (blue) and finally up to the bond interface (green) where it would connect to the next tier. (c) shows a cartoon of four 4-via diagonal vias working together to connect four different signals from a Control tier uniquely to four different CAM Tiers.

each Road Pattern Match Cell for each CAM Tier. This is a problem because though logically identical, each tier must be physically separable. It is economically very desirable to have only one CAM Tier mask set and since there are L CAM tiers and L vertical lines to the Majority Logic but only one CAM Tier mask set, somehow each otherwise identical CAM tier must be told which unique vertical line to use. This is called automatic tier-self ID and it turns out that this problem has been solved in industry long ago using the so-called diagonal via structure [15].

The diagonal via structure allows inter-tier communication with automatic tier-self ID, without the need for any extra transistors. Fig. 5 (a) shows the concept: the leftmost signals are shifted one place to the right and the rightmost signal is shifted to the leftmost position. Fig. 5 (b) shows how this might be implemented in VLSI. Finally, Fig. 5 (c) shows how this is done for signals driven from Control Tier to each of four CAM Tiers. In this case, the Control Tier is sending layer/tier specific data to each tier (such as an input data bus from each detector layer). This same structure works with drivers on each CAM tier and with each CAM tier sending layer/tier specific data to the Control Tier (such as Stored Address Match Lines from each CAM cell). In a structure with one Control Tier and four CAM tiers, the Control tier sees four vias, one for each CAM tier. All CAM tiers have exactly the same mask layout.

6. Conclusions and Future Work

The associative memory approach to track finding and the PRAM devices that implement it are well suited to modern 3D integration. The algorithm is easily divisible into logical partitions that are physically separable from one another due to the simplicity and consistency of the interconnects between these logical partitions. Moreover, integrating them vertically yields an immediate pattern density

improvement to the associative memory approach. Diagonal Vias permit simple automatic tier self-ID which allows the VIPRAM design to be accomplished with only two mask sets regardless of the number of detector layers in the final design. The implementation presented here is specifically geared towards the tracking trigger pattern recognition application for High Energy Physics. However, the VIPRAM 3D structure is inherently open and flexible, and would facilitate design reuse, making possible the design of more general-purpose fast pattern recognition devices with potential applications far beyond the original Associative Memory used for particle physics experiments.

The VIPRAM effort is very much an ongoing project. A generic R&D proposal [1] based on this new concept, with a few institutions involved, has been submitted to DOE with the goal to design and perform the ASIC engineering necessary to realize a prototype device. The first step should be a 2D prototype of the new sub-cells. The first 3D implementation should probably be a single Control Tier with a single CAM tier to test the vertical interconnections. This should be followed by a single Control Tier with two or three CAM tiers. This prototype will test all the necessary processing steps involved in a complete VIPRAM. As 3D technology evolves, the spacing of Through Silicon Vias and other structures unique to 3D integration will also evolve. For the moment, it makes sense to remain at a reasonable technology node such as 130nm rather than pursue a more aggressive node. We expect up to 200K patterns per cm square with 130nm and 4 um TSV spacing [12]. This would allow for relatively inexpensive prototyping. When all of the processing steps for a final VIPRAM are prototyped, then the selection of a final VLSI technology node will be clearer.

In the future, we plan to integrate the VIPRAM design with the FPGA-based track fitting stage [5] into a single chip, possibly using the interposer approach recently used for Xilinx Virtex 7 FPGA. The 3D VIPRAM design is to solve the pattern density and performance limitation in 2D design by vertical integration, while the goal of the single chip integration is to solve the problem of very large data flow between the AM stage and the track fitting stage by integrating the two stages into one chip. This second part is ultimately what should be done to address the fast pattern recognition and track fitting challenges/issues for the LHC at very high luminosity. Note that since modern FPGAs can be used, the data input bandwidth will be significantly improved as well. In addition, large memories can be integrated into the same package this way. The large memory array could be used as a hit buffer to store the full resolution input hits in a database organized for rapid retrieval, as well as lookup tables for large sets of constants for track fitting purpose.

Acknowledgement

The authors would like to thank Bob Patti of Tezzaron Semiconductor for numerous beneficial discussions.

References

[1] Development of 3D Vertically Integrated Pattern Recognition Associative Memory (VIPRAM), FERMILAB-TM-2493-CMS-E-PPD-TD

[2] J. Adelman et al., “The Silicon Vertex Trigger upgrade at CDF”, Nucl. Instr. And Meth. In Physics Research A, vol. 572, Issue 1, pp 361-364, March, 2007.

[3] J. Adelman et al., “Real time secondary vertexting at CDF”, Nucl. Instr. And Meth. in Physics Research A, vol. 569, pp 111-114, 2006.

[4] M. Dell’Orso and L. Ristori, “VLSI Structures for Track Finding,” Proceedings in Nuclear Instruments and Methods, vol. A278, pp. 436-440, 1989.

[5] A. Annovi et al., "The GigaFitter: A next generation track fitter to enhance online tracking performances at CDF," Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE , vol., no., pp.1143-1146, 2009.

[6] “FTK: A Hardware Track Finder for the ATLAS Trigger,” Technical Proposal, 2010.

[7] A. Annovi, et al., “A VLSI Processor for Fast Track Finding Based on Content Addressable Memories,” IEEE Transactions on Nuclear Science, vol. 53, no. 4, pp. 1-6, 2006.

[8] S. Amerio, et al., “Associative Memory Design for the FastTrack Processor (FTK) at ATLAS”, Proceedings of IEEE Nuclear Science Symposium, October, 2011.

[9] T. Kohonen, “Content-Addressable Memories,” 2nd edition, New York, Springer-Verlag, 1987.

[10] K. Pagiamtzis and A. Sheikholeslami, “Content-Addressable Memory (CAM) Circuits and Archtectures: A Tutorial and Survey”, IEEE Journal of Solid-State Circuits, Vol. 41, No. 3, pp. 712-727, March, 2006.

[11] K. Banerjee, S. Souri, P. Kapur, and K. Saraswat, “3-D ICs: A Novel Chip Design for Improving Deep-Sub micrometer Interconnect Performance and Systems-on-Chip Integration,” Proceedings of the IEEE, pp. 602-633, 2001.

[12] Tezzaron Semiconductor, http://www.tezzaron.com/.

[13] *International Technology Roadmap for Semiconductors (ITRS)* 2009, http://www.itrs.net/.

[14]Fujita, S., Keiko, A., Nomura, K., Yasuda, S., Tanamoto, T., “Perspecitives and issues in 3D-IC from designers’ point of view”, IEEE International Symposium on Circuits and Systems, ISCAS 2009, May, 2009, p. 73

[15] Patti, Robert, Connection Arrangement for Enabling the Use of Identical Chips in 3-dimensional Stacks of Chips Requiring Address Specific to Each Chip, U.S. Patent 6,271,587, filed September 15, 1999 and issued August 7, 2001