Preprint typeset in JINST style - HYPER VERSION

Design and Testing of the First 2D Prototype Vertically Integrated Pattern Recognition Associative Memory (Fermilab-PUB-14-478-E-PPD)

T. Liu[a]*, **G. Deptuch**[a], **J. Hoff**[a], **S. Jindariani**[a], **S. Joshi**[a], **J. Olsen**[a], **N. Tran**[a], **M. Trimpl**[a]

[a] *Fermi National Accelerator Laboratory, Particle Physics Division, Batavia, IL, 60565, USA*
E-mail: thliu@fnal.gov

ABSTRACT: An associative memory-based track finding approach has been proposed for a Level 1 tracking trigger to cope with increasing luminosities at the LHC. The associative memory uses a massively parallel architecture to tackle the intrinsically complex combinatorics of track finding algorithms, thus avoiding the typical power law dependence of execution time on occupancy and solving the pattern recognition in times roughly proportional to the number of hits. This is of crucial importance given the large occupancies typical of hadronic collisions. The design of an associative memory system capable of dealing with the complexity of HL-LHC collisions and with the short latency required by Level 1 triggering poses significant, as yet unsolved, technical challenges. For this reason, an aggressive R&D program has been launched at Fermilab to advance state of-the-art associative memory technology, the so called VIPRAM (Vertically Integrated Pattern Recognition Associative Memory) project. The VIPRAM leverages emerging 3D vertical integration technology to build faster and denser Associative Memory devices. The first step is to implement in conventional VLSI the associative memory building blocks that can be used in 3D stacking; in other words, the building blocks are laid out as if it is a 3D design. In this paper, we report on the first successful implementation of a 2D VIPRAM demonstrator chip (protoVIPRAM00). The results show that these building blocks are ready for 3D stacking.

KEYWORDS: Real Time Pattern Recognition; Tracking Trigger; Associative Memory; 3D IC; VIPRAM.

*Corresponding author.

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

Contents

1. Introduction

Hardware-based pattern recognition using associtive memory [1] for fast triggering on particle tracks has been successfully used in high-energy physics experiments. The CDF Silicon Vertex Trigger (SVT) at the Fermilab Tevatron is a good example. The method used there, developed in the 1990s, is based on algorithms that use an associative memory architecture to identify track patterns efficiently at high speed. The associative memory approach involves using content addressable memories (CAMs) to find matching detector hits and majority logic to associate hits from different detector layers to form track candidates. This massively parallel architecture is ideally suited to tackle the intrinsically complex combinatorics of track finding algorithms, avoiding the typical power law dependence of execution time on occupancy and solving the pattern recognition in times roughly proportional to the number of hits. This is of crucial importance given the large occupancies typical of hadronic collisions. 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 volume, there is an enormous challenge in implementing fast pattern recognition for a track trigger, requiring about orders of magnitude more associative memory patterns than were used in the original CDF SVT. Moreover, the rigorous technical requirements (such as low latency) of a silicon-based L1 tracking trigger push the limits of Pattern Recognition Associative Memories (PRAM) in pattern density, speed and power density. Approaches to this goal in simple 2D VLSI are limited. For this reason, a new concept to use emerging 3D technology has been proposed [2].

Design in 3D vertical integration is, in a sense, the logical partitioning of functionality into a third dimension. The PRAM structure is intrinsically adjustable in the 3rd dimension from the full

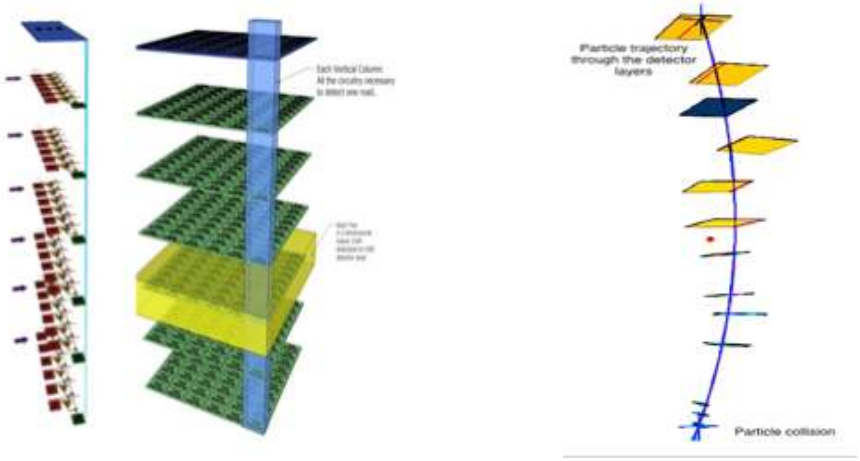

Figure 1. Left: 3D Block diagram of VIPRAM concept where each layer of the chip matches one layer of the detector. All together, one tube in 3D can detect one road. Right: Corresponding particle trajectory through detector layers.

pattern level down to the individual CAM level. The VIPRAM (Vertically Integrated PRAM) approach is to divide the PRAM structure among 3D VLSI tiers to reduce the area consumed by a single pattern, to reduce the parasitic capacitance of long runs, to increase the effective number of routing layers, and finally, to increase the readout speed significantly. The essence of VIPRAM is to divide this approach up into different tiers, maximizing pattern density while minimizing critical lengths and parasitics and therefore the power density.

As shown in Figure 1, the basic VIPRAM concept is a division of labor that places all control cells (i.e. the majority logic) in one "top" control tier and the CAM cells in individual tiers corresponding to each detector layer. The resulting pattern "tube", shown in Figure 1 and highlighted in blue, contains all circuitry necessary to select one candidate track passing through detector layers. Figure 1 (right side) also shows one example of a corresponding particle trajectory through detector layers. The pattern density directly depends on the cross-sectional area of the tube. The match lines from the CAM to the majority logical cell are long in the conventional 2D implementation. They are now implemented vertically and are therefore shorter in 3D because each tier will be thinned down to about 10 um during the 3D stacking process. As these lines are repeated throughout the chip, this can have a significant impact on performance. The vertical integration also provides flexibility in layout optimization of the building blocks, and therefore chip performance.
This paper presents the design, simulation studies and preliminary testing results of the first 2D prototype (`protoVIPRAM00`) of the Vertically Integrated Pattern Recognition Associative Memory (VIPRAM) concept.

2. `protoVIPRAM00` Design and Verification

2.1 Concept and Testing Methodology

Since 3D Vertical Integration is an emerging technology and the requirements of the L1 track trig-

ger have themselves been evolving, the first logical step is to test the two basic building blocks, the CAM cell and the majority logic cell, through a simple "2D" prototype run. This will provide verification of their functionality in preparation for the 3D stacking and low latency readout developments in the near future. In other words, the VIPRAM architecture allows us to test the 3D building blocks in a simple, low-cost 2D prototype.

The associative memory building blocks were laid out as if this was a 3D design. Space was reserved for as yet non-existent through silicon vias (TSV) and routing was performed to avoid these areas. To keep things simple, each PRAM pattern consists of only four identical CAM cells and a control cell resulting in the ability to recognize 4-layer pattern matches. The readout circuitry of the PRAM array is deliberately simplified to allow direct performance studies of the CAM and control cells. Any system interface including high-speed readout for Level 1 Tracking Trigger applications and architectural options made available by vertical integration are the subject of another presentation at the TWEPP 2014 workshop [3]. The testing setup has been developed in such a way to allow direct comparison between 2D and 3D implementations.

2.2 Design

For demonstration purposes, it is useful to compare a new design to existing designs with the same core functionality such as INFN Amchip04 [4]. Therefore, the prototype was designed as a 15-bit CAM with a 4-bit selective precharge and 3 ternary CAM cells. The selective precharge is made with four NAND cells while the remaining eight bits are NOR cells. The `matchLine` is the single signal that connects the different bits in the CAM cell and its parasitic impedance impacts the

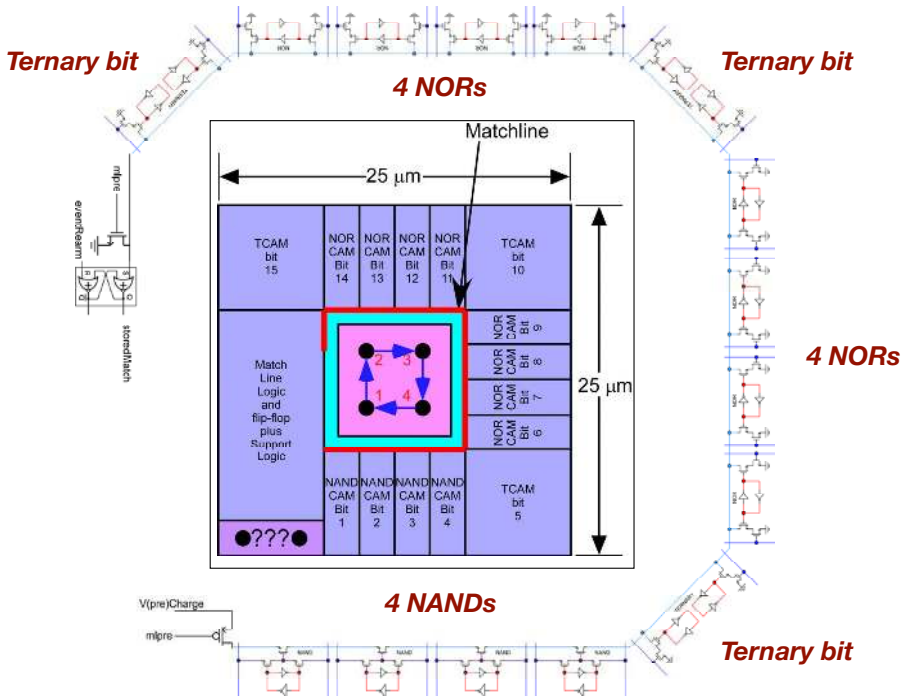

Figure 2. CAM cell design block diagram

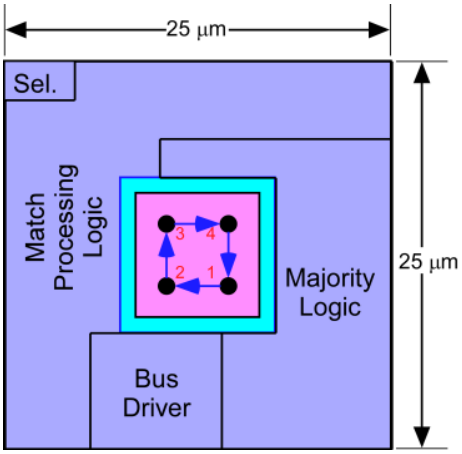

Figure 3. Majority Logic cell diagram

chip performance. A schematic of the CAM layout is shown in Figure 2. The CAM design is pre-discharged low before the input data arrives. Because the CAM is pre-discharged, the precise timing of data arrival is not critical. The `matchLine` is charged to a high through the V_{charge} node in case of a match or discharged to ground in case of a mismatch. Minimizing the `matchLine` increases the maximum clock frequency and minimizes power density for a given clock frequency. The layout optimization permitted by vertical integration allows a redesign of the `matchLine` in the CAM cell itself, and the `matchLine` is shortened considerably by wrapping it in a square as shown by the red line in Figure 2.

The majority logic cell (see Figure 3) is designed to match the CAM Cell in footprint. The Majority Logic uses Pass Transistor Logic to produce a 3-bit code indicative of the match stage: 111 (All Layer Match), 011 (One Missing Layer), 001 (Two Missing Layers), and 000 (Three or more Missing Layers). The Match Processing Logic compares the output of the Majority Logic with the user supplied threshold and, if met, flags a Road Match.

Each pattern has 4 CAM cells and one control cell (which contains the majority logic/flag logic) and performs complete pattern recognition for one 4 layer pattern. The layout of all the cells

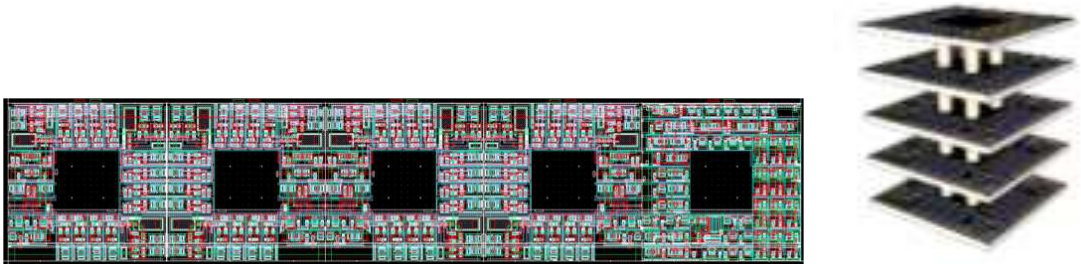

Figure 4. A PRAM pattern layout in 25um X 125 um. Each pattern has 4 CAM cells and one control cell (which contains the majority logic/flag logic) and performs complete pattern recognition for one 4 layer pattern. To keep the design fully compatible with 3D implementation (right), space is left in the middle for TSVs.

for a pattern requires an area 25 um x125 um. An area 10 um x125 um was added on the side to allow for routing within the pattern as well as for horizontal routing. The prototype chip has 4096 patterns distributed in 128 rows and 32 columns. The chip operates in two modes: in load mode patterns are stored in the CAMs one at a time, and in run mode incoming data is compared to the stored patterns and the `roadFlag` outputs are generated based on matches found and threshold conditions asserted. Each CAM in a pattern performs a layer-specific pattern recognition operation on incoming data in a single clock cycle. At the end of an event, the eventRearm signal clears the matches. In each pattern the outputs from the 4 CAMs go to the control cell where threshold condition can be applied to determine if a single `roadFlag` is asserted. A total of 32 `roadFlag` outputs corresponding to all the columns of a particular, selected row on the chip are sent to the output pads. The readout implementation is kept very simple to allow easy testing of the pattern matching performance.

The `protoVIPRAM00` was designed and fabricated in a 130nm Low Power CMOS process that has been used previously in HEP 3D designs. The size of the chip is 5.46mm x 5.46mm. The layout was implemented such that, in future 3D design, the basic building blocks can be directly reused and placed on different 3D tiers.

2.3 Design Simulation and Verification

The design has been thoroughly simulated at all levels with timing, signal dispersion and power consumption. Corner analysis and Monte Carlo tests were done on all parts of the core and the periphery. In order to study the CAM cell performance as a function of V_{charge}, mixed signal simulations were performed using Verilog data generator . The AMS (Analog Mixed Signal) simulator was used in `Cadence ADE XL` to simulate the designs with connection rules for 1.5V logic.

A number of critical delays in the chip were simulated. This includes 50% to 50% propagation delay (PD) from the release of `matchLine` pre-charge to the rising edge of stored match signal, rise time of the `matchLine` from 10% to 90% of V_{charge}, and the delay between the last CAM match in a pattern and the corresponding `roadFlag` output of control logic. The former two delays are shown in Figure 5 and their mean values were found to be in the range of 3 to 4 ns. The third delay was found to be insignificant.

3. `protoVIPRAM00` testing setup and results

The `protoVIPRAM00` chips are wire bonded to a standard 144 pin Thin Quad Flat Pack (TQFP). The socket is mounted on a test mezzanine card (see Figure 7, right) which is controlled by a Kintex-7 FPGA. Test vector data are stored in the FPGA internal RAM blocks and clocked into the `protoVIPRAM00` with frequencies up to 200 MHz. Likewise, the `protoVIPRAM00` outputs are sampled and stored in FPGA RAM for later readout via an optical Gigabit Ethernet connection, which is based on simple UDP packet transfers (the IPBus-based protocol [5]). The generated input patterns and the sampled outputs from the chip are verified against functional simulations done on the full chip design using Cadence NC-Sim, while the software needed to initialize, control, and transfer data to and from the FPGA was written in `Python`. Additional lower-level validation of the testing results are checked using `ChipScopePro` to verify detailed bitwise performance.

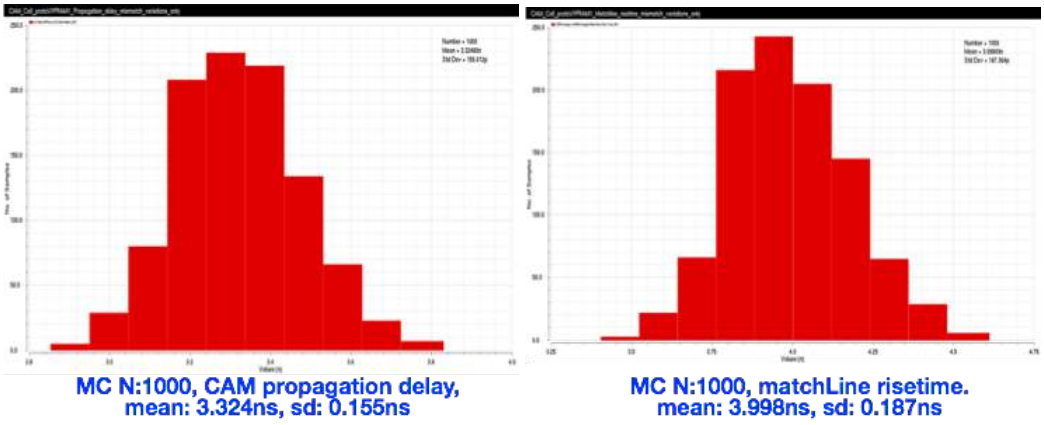

Figure 5. Left: 50% to 50% PD from the release of `matchLine` pre-charge to the rising edge of stored match signal. Right: Rise time of the `matchLine` from 10% to 90% of V_{charge}.

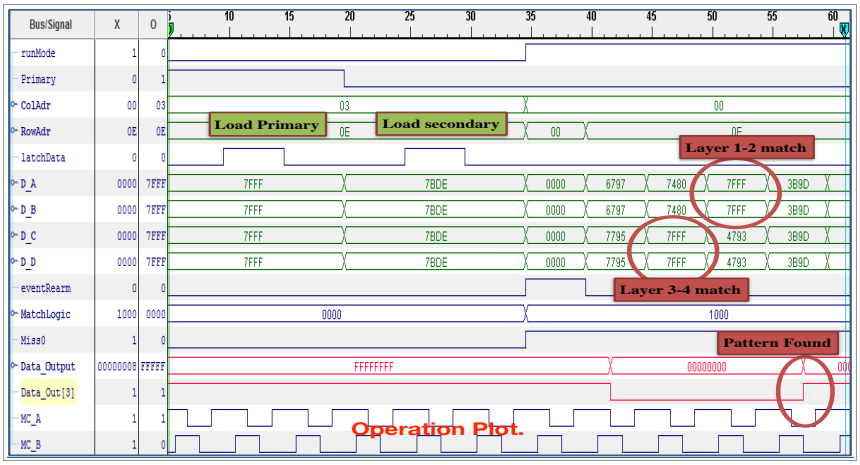

Figure 6. Timing diagram of chip functionality; a pattern is loaded into the chip and then matched during run mode

Figure 7. Left: `protoVIPRAM00`, 130nm Global Foundries Low-Power CMOS, 5.5mm X 5.5mm. Right: FMC Test Mezzanine card features a Xilinx Kintex XC7K160T FPGA, 4 SFP+ transceivers, 128MB DDR3, and a 144 pin socket used for testing custom ASIC chips, such as `protoVIPRAM00`.

The testing results obtained show that the chip operates with full functionality. We illustrate basic validation of the performance in Figure 6 with `ChipScopePro` where a pattern is loaded into the

chip and then is found during run mode. Each location of the chip was tested exhaustively with varying data and threshold conditions.

3.1 Efficiency with "Pseudo-realistic" HL-LHC Test Conditions

To benchmark performance, simulation data based on the upgraded CMS tracker geometry was used. The input data contained hits associated with tracks from hard scattering overlaid with on average 140 secondary interactions, as expected in High Luminosity LHC running. Multiplicity of hits in each layer was varied event-by-event according to random Poisson distribution with means derived from simulation. The chips were tested using hits from the inner and outer four layers of the central part of the tracker. Efficiency of pattern finding has been measured at different operating frequencies ranging from 50 MHz to 143 MHz.
It is observed that the chip operates with 100% efficiency (i.e. correctly finding all the patterns) up to a frequency of approximately 100MHz with decreased efficiency above 100MHz. This is consistent with the original design goal as well as being consistent with simulation. Tests are repeated on all 12 protoVIPRAM chips available with similar performance.

3.2 Power Consumption

Programmable voltage regulators on the mezzanine card support current readback on the `protoVIPRAM00` power rails. Line Drivers consume most of the power, typical power consumption at 100 MHz for pseudo-realistic patterns is about 200 mW. Since this chip only contains 4K patterns, it is not straightforward to measure the power consumption directly related to the pattern matching operation.

The chips were also tested beyond ordinary operating conditions to include situations where all of the patterns matched simultaneously in order to study the power consumption at the extreme condition. As expected, efficiency drops with the number of simultaneous matches or near-matches, but still fully working with 100% efficiency at 50 MHz with all patterns matching simultaneously all the time. Such "torture" testing has been useful in studying the robustness of the building blocks and load capability of the power rail structure. Initial power and thermal analysis of the chip is described in a separate paper [6]. More detailed study is on-going and the results will be reported in the future.

4. Summary and Outlook

The first `protoVIPRAM00` chip was designed and fabricated in a 130nm Low Power CMOS process. The layout was deliberately implemented so that the basic associative memory building blocks can be directly re-used for 3D stacking. The design has been thoroughly simulated at all levels and the prototype has been successfully tested both for functionality and performance using a custom test setup, with special test patterns as well as input hit patterns in "pseudo-realistic" HL-LHC conditions. The testing results show that the basic associative memory building blocks, the CAM and the control cell that comprise `protoVIPRAM00` are ready for 3D vertical integration for proof-of-principle demonstration of the VIPRAM concept. The 3D VIPRAM design is work in progress, with minimal modifications to the existing 2D design and only those changes that are consequences of 3D are being permited. This would allow an apples-to-apples comparison of the

2D and 3D designs, using the same test setup. In addition, the `protoVIPRAM00` design can be used as a solid starting point for more improvements in the next version, a 2-tier design optimized for CMS L1 tracking trigger application.

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 [7] allows mutli-layer VIPRAM design to be accomplished with only two mask sets. As 3D technology evolves, the spacing of TSVs and other structures unique to 3D integration will also evolve. For the moment, it makes sense to remain at technology node such as 130nm to allow for relatively inexpensive prototyping. When all of the processing steps for VIPRAM are prototyped, then the selection of a final VLSI technology node will be made.

Acknowledgments

This work has been supported by DOE CDRD (Collider Detector Research and Development) program.

References

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

[2] T. Liu, J. Hoff, G. Deptuch, R. Yarema, "A New Concept of Vertically Integrated Pattern Recognition Associative Memory", published in the Proceedings of the TIPP 2011 Conference: http://www.sciencedirect.com/science/article/pii/S1875389212019165.

[3] J. Hoff, T. Liu and J. Olsen, VIPRAM architecture: from 2D to 3D, TWEPP 2014 talk.

[4] A. Annovi, et al, Associative memory design for the fast track processor (FTK) at ATLAS, Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2011 IEEE Page(s): 141-146

[5] R. Frazier, G. Iles, D. Newbold, A. Rose, "Software and firmware for controlling CMS trigger and readout hardware via gigabit Ethernet", published in the Proceedings of the TIPP 2011 Conference: http://www.sciencedirect.com/science/article/pii/S1875389212019098.

[6] W. Xia, et al, Thermal Analysis of the ProtoVIPRAM00 Chip, Poster at TWEPP 2014, and submitted to JINST.

[7] 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.