Fermilab Technical Memo Number: FERMILAB-TM-2482-E-PPD

Proposal for the Development of 3D Vertically Integrated Pattern Recognition Associative Memory (VIPRAM)

Fermilab:
Gregory Deptuch, Jim Hoff, Simon Kwan, Ron Lipton, Ted Liu,
Erik Ramberg, Aida Todri, Ray Yarema

Argonne National Laboratory
Marcel Demarteau, Gary Drake, Harry Weerts

University of Chicago
Mel Shochet

University and INFN of Padova/Fermilab
Silvia Amerio

Tezzaron Semiconductor Corporation
Bob Patti

(comments to Ted Liu: thliu@fnal.gov)

Version 1.8, Oct. 3, 2010

Executive Summary

Future particle physics experiments looking for rare processes will have no choice but 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 process. We propose to develop a 3D Vertically Integrated Pattern Recognition Associative Memory (VIPRAM) chip for HEP applications, to advance the state-of-the-art for pattern recognition and track reconstruction for fast triggering.

3D technology is the integration of thinned and bonded silicon integrated circuits with vertical interconnects between IC layers using through silicon vias (TSVs). The technology has wide applications in industry, ranging from memories to pixel arrays to microprocessors and FPGAs. Performance can be improved significantly by reducing interconnect R/L/C for higher speed and density. In addition, it provides the freedom to divide functionality among tiers to create new designs that are simply not possible in 2D. As Moore's law is approaching severe limitations, it is expected that 3D technology will be the next scaling engine. Generally speaking, 3D technology becomes useful when a task can be partitioned into multiple sections that are physically and logically separable, and the interconnections among them are straightforward. Moreover, the use of 3D technology can have varied goals. For example, it can be used to increase transistor density – i.e. to increase the number of transistors per square micron. Such is a major goal of 3D DRAM design. Here, the DRAM task is first logically divided into a control/interface section and memory core. The control/interface section is physically separated onto its own tier, and the memory core is further divided into memory banks which are each vertically integrated onto their own tiers. A second, different example is the 3D integration of microprocessor systems. Here, different functions that have been traditionally separated can be brought together in a single monolithic structure and technological limitations can be eliminated. CPU and memory can be placed on separate tiers and the interconnect between them –i.e. the memory bus – can be reduced from on the order of tens of millimeters (a bus on a PC board) to a few tens of microns (the length of a through silicon via). Also, the memory bus itself can be expanded from a few bits to hundreds of bits wide, dramatically improving the memory access bandwidth.

With Pattern Recognition Associative Memory for HEP tracking trigger applications, the task is indeed logically and physically dividable and the interconnections among them are straightforward, making it a good candidate for 3D integration. Associative Memories, in HEP, are based on the concept of a Content Addressable Memory (CAM) – memory that is not accessed by providing the memory cell with an appropriate address, but rather a memory that is accessed by providing the memory cell with some related content. In the case of tracking triggers in HEP, that content is hit locations. However, in HEP tracking trigger applications, flagging an individual detector hit is not important. Rather, the path of a charged particle through many detector layers is what must be found. This effectively makes an HEP Associative Memory a "CAM of CAMs", meaning that individual hits must be flagged and accumulated first for a given event, then related sets of detector hits – commonly called "roads" – must ultimately be flagged. Therefore, in its essence, an HEP Associative Memory (AM) bank is a CAM array that is a collection of independent roads, i.e. independent sets of hit addresses from different detector layers which represent a path or

road that a charged particle might traverse through the detector. Like the 3D DRAM case, the AM can naturally be divided into a control/interface tier and a set of CAM tiers. What is unique for the 3D AM (VIPRAM) architecture is that when each CAM tier corresponds to a single detector layer, then the interconnections between the tiers become dramatically simplified. Logically, an AM road is an independent set of hit addresses from different detector layers, now physically in the VIPRAM architecture, a road is a simple independent vertical tube in a 3D monolithic circuit that is a collection of CAM cells each programmed to detect the hit on a particular detector layer for that particular road, and reports the match directly to the control tier. Routing in 3D can be very efficient, especially if functional elements are arranged such that the interconnections among tiers are mostly vertical. This is the case for VIPRAM architecture, not only are the interconnects among tiers vertical, they are uniform (in fact identical) across the tiers as well.

The goal is to improve the pattern density by about two orders of magnitude over the existing 180nm-based AMchip using 65nm technology. The R&D program will have two phases. In phase 1, we will use the novel 3D technology to improve the Associative Memory chip performance (density and speed) for fast pattern recognition. This will lead to the design and prototyping of a new 3D chip called VIPRAM. For Phase two, we plan to integrate the VIPRAM with track fitting stages (currently done using FPGA and RAMs) into a single chip, using either 3D stacking or a "system-on-package" approach to improve the bandwidth between the pattern recognition AM stage and the track fitting stage. The deliverable for Phase 2 is a single chip or chip package that has the potential to dramatically improve the overall track trigger performance. The chip should be flexible enough for applications far beyond tracking trigger, within and outside HEP.

1. Introduction

Reminder:
Priorities of the DOE ➔ Focus on Transformational Science
Connect basic and applied sciences
Re-energize the national labs as centers of great science and innovation
Embrace a degree of risk-taking in research
Create an effective mechanism to integrate national laboratory, university, and industry activities

1.1. Future Challenges of Pattern Recognition

Many next generation science experiments will be characterized by the collection of large amounts of data, taken in rapid succession, from which the scientists will have to unravel the underlying physics processes. More often than not, large backgrounds will overwhelm the physics signal and real-time data analysis will be indispensible to immediately separate interesting events from background, select them for further analysis and reduce the data size to manageable proportions. Scaling of current technologies does not seem to meet the scientific goals of future projects and investments in transformational new technologies need to be made to enable new scientific projects.

Many areas in science can be identified that currently face these challenges. One area is the capability to perform fast pattern recognition and track reconstruction of particle trajectories in modern high-energy physics (HEP) hadron collider experiments. The Large Hadron Collider (LHC) at CERN has proposed a luminosity increase of a factor five to ten over the original design as the goal for the upgrade, which will result in a corresponding factor 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. The ultimate physics reach of the LHC will crucially depend on the tracking trigger capabilities of its experiments to handle these high luminosities to discriminate between the interesting events and the background. The overall goal is to identify particle tracks at the trigger level, a capability that is crucial for many important searches for new physics (use CMS L1 Muon as example here). There are other important reasons for having tracking trigger capabilities at early stage 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 could be separated from background using the number of tracks within a narrow "signal cone" and the number in a larger "isolation region".

Another example in the area of tracking at high-energy frontier is pattern recognition at a Muon Collider (MuC). The hit densities in a vertex and tracking detector at a MuC are dominated by backgrounds from the decays in-flight of the colliding muons and upstream muons entering the detector. The upstream muons will not originate from the interaction point, but rather travel along the beam axis. Fast, efficient pattern recognition could identify and possibly eliminate these tracks online (more later …).

Instrumentation at photon science facilities could also benefit from the development of technologies that allow for fast online pattern recognition of large sets of data. In Photon Correlation Spectroscopy (PCS), for example, the dynamics of a material is probed by analyzing the temporal correlations among photons scattered by the material. X-ray PCS (XPCS) offers the unprecedented opportunity to extend the range of length scales over which a material's low frequency dynamics can be probed down to inter-atomic spacing. With the advent of new coherent, brilliant X-ray sources, technologies that enable online correlation spectroscopy could be a major advantage (more later).

In this proposal, we describe the development of a new hardware-based technology that advances the state-of-the-art for pattern recognition and track reconstruction for fast triggering. The technology could have wide applications far beyond a track trigger, both within and outside HEP. 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 facilities.

1.2. Fast Pattern Recognition and Track Reconstruction

Traditionally, track triggers have been implemented using 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 relatively slow, since the computations require significant CPU processing time. It is desirable to push this type of trigger into earlier levels of a trigger system. The CDF Silicon Vertex Trigger (SVT) at the Fermilab Tevatron is a good example. The method used there [11], developed in the 1990's, uses algorithms implemented in fast logic. The technique has two-parts. The first part uses Associative Memory (AM) or Content Addressable Memory (CAM) architectures to efficiently identify track patterns (roads) at high speed using coarse-resolution "hits" recorded in the tracking detector. Then, the patterns are processed using fast FPGAs to perform track fitting with full detector resolution hits. A block diagram of the Associative Memory architecture is shown in Figure 1. The method solves the combinatorial challenge inherent to the tracking by exploiting massive parallelism of associative memories that can compare tracking detector hits to a set of pre-calculated patterns simultaneously. The assumption is that the resulting road is narrow enough so that a helical fit can be replaced by a simple linear calculation. The track fitting stage for each matched pattern is much simplified and fast by using tracking parameters with values for the center of the road, and applying corrections that are linear in the relative hit position in each layer.

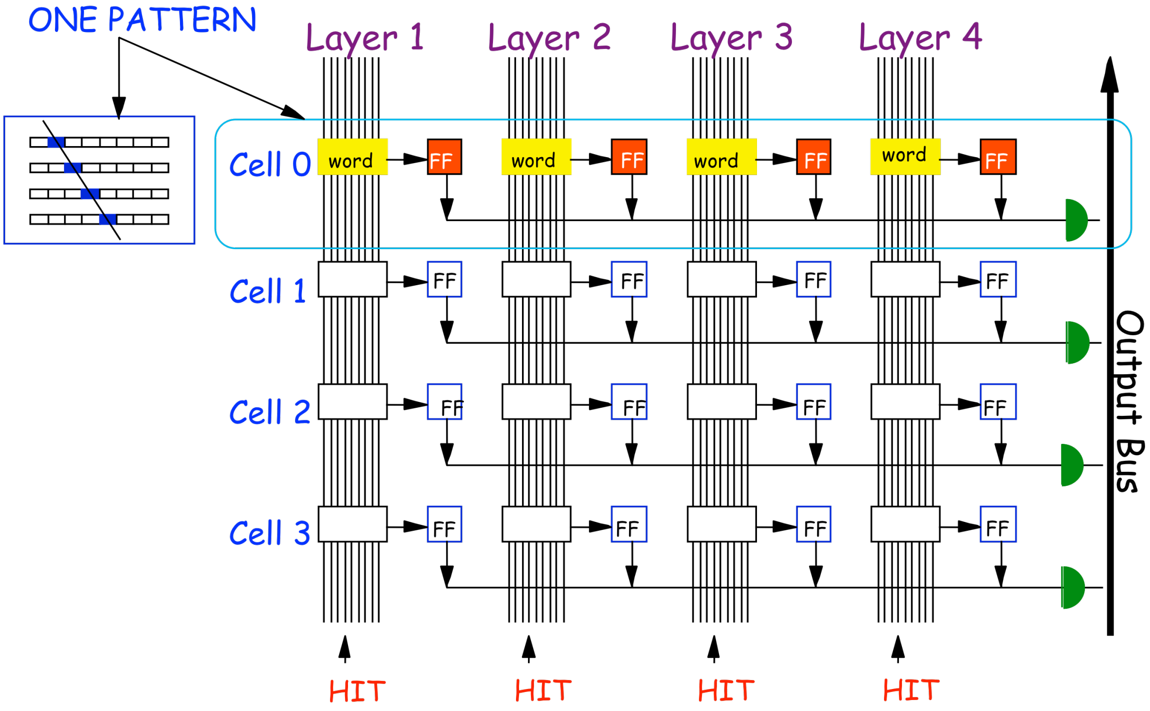


Figure 1. CDF SVT Associative Memory Architecture [11].

The SVT approach was highly successful, and CDF was the first hadron collider experiment in High Energy Physics to incorporate a fast secondary vertex track trigger. It finds all tracks emanating from each collision and precisely measures their properties within about 30 microseconds after the collision. This latency can be compared to the ~1 second required when track reconstruction is done inside a modern computer. The SVT has been essential to many of the physics results to come out of the CDF experiment and it has significantly improved the CDF physics reach. For example, it is the critical device that enabled CDF to measure the long awaited Bs mixing, a process that is important for understanding the matter-antimatter asymmetry in nature. It also allowed the observation of the decay of the Z boson, a carrier of the weak unclear force, into two energetic b-quark jets, a signature very similar to that of the Higgs boson.

In the era of the upgraded LHC (SLHC), it is desirable to implement this type of track finding capability into the early stages of the trigger because of the importance of identifying particle tracks at the trigger level for many important searches for new physics. However, due to the much higher occupancy and event rates at the SLHC, and the fact that the LHC detectors are much more massive with orders of magnitude more channels in their tracking volumes, it is a difficult challenge to perform pattern recognition and track fitting at the trigger level. In addition there is the obvious challenge of data transfer from the detector to the trigger system. While processing power and speed have increased steadily over time, certain future applications have levels of complexity that have exceeded the current hardware capabilities in fast track reconstruction. Significant improvements in both the pattern recognition (the associative memory) as well as the track fitting performance are needed.

1.3. Current Status

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, wider roads using coarser resolution hits require less AM storage, but the number of fake roads found by the AM and the number of fits at the track fitting stage downstream could increase quickly due to the high 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, by using finer resolution hits, the number of fake roads and fits will be reduced for a given AM bank size, but the required total size of the AM would increase dramatically. Therefore, the road width must be optimized. The required AM pattern bank size may be different for different experiments, or possibly different for the same experiment at different luminosities. As an example, consider the implementation of a hardware-based track trigger like that used in the CDF SVT in the context of what is needed for the LHC. For this comparison, we will use the Atlas FastTracK (FTK) project as an example, since the design requirements of the system are well known from extensive simulations, although this high-level extrapolation could apply to CMS as well. The original CDF SVT system, in operation during 2000 until 2005, had a total number of associative memory patterns of 384,000, while the proposed Atlas FTK system for the Level 2 trigger would require ~ 1 billion patterns in order to handle luminosities above 3×10^{34} $cm^{-2}s^{-1}$ [1]. This is three orders of magnitude more associative memory patterns. The Level 1 Track Trigger upgrade for both CMS and Atlas likely will require even more AM patterns running at higher speed.

A new AM chip (AMchip03) was developed in 2005 by the CDF Italian collaborators [7]. The AMchip03 was implemented in 180nm technology using the Standard Cell approach and the number of patterns per chip increased by a factor of 40 over the previous version used at CDF, from 128 to 5,000. This chip, which runs at 40MHz, was used to upgrade the SVT system and the total number of patterns has increased to more than 6 Million. According to extensive Monte Carlo simulation studies, it is possible that the AMchip03 could be used for the initial FTK system for low luminosity running. However, in order to meet the challenges for the LHC high luminosity running, another improvement factor of at least 50 in pattern density will be required for the Atlas FTK. The current technology using FPGAs and custom chips cannot be scaled in a simple manner to accommodate this. Significant improvement in associative memory performance (pattern density and speed) is needed to run fast pattern recognition algorithms of this scale.

Increasing the AMchip pattern size is one approach to increase performance. As the AM bank density and size increases, the number of fired roads will increase at high luminosity due to higher occupancy. This has two consequences. First, more full resolution detector hits associated with roads have to be retrieved and transferred from the AM stage to the track fitting stage, which demands much higher bandwidth between the two. Secondly, for the track fitting, the fitting speed has to be high enough in order to keep up with the higher number of found patterns upstream. Motivated by existing system needs, CDF has recently upgraded the track fitting stage of the SVT system, the "GigaFitter Upgrade". It significantly improves the track fitting speed and performance by taking full advantage of some of the advanced features (imbedded DSPs) of modern FPGAs [3]. The speed performance of the Gigafitter is expected to approach one fit per nano-second, hence the name. This is the average time to fit the hits for a track candidate and extract a goodness of fit and track parameters, and it is several hundred times faster than in the original SVT. For the LHC trigger application at high luminosity using Atlas FTK as an example, even with almost 1 billion AM patterns, the total number of fits will still be on the order of a million for one proton-proton collision. Maintaining such a rate in a large system will be difficult even with the Gigafitter speed performance. In particular, there is a need to transfer large numbers of found roads and the associated full resolution hits from the AM stage into the track fitting stage. Since the track fitting stage typically has to be done on a separate module from the pattern matching stage, this would pose a significant design challenge at both the board and system level (for details, see FTK proposal [1]). It is highly desirable for the AM stage and track fitting stage to be implemented in such a way that the two are very close to each other, or preferably within the same chip. The track fitting stage can be viewed as the second stage of pattern recognition using full resolution information. The resulting integrated chip would be much more powerful for fast pattern recognition. In addition, both the board and system level design would be significantly simplified if this level of integration can be achieved.

1.4. Solution Based on a Novel Technology

In this proposal, we describe the use of 3-dimensional (3D) integrated circuit technology as a way to improve the fast pattern recognition and track reconstruction for HEP trigger applications. A 3D chip is generally referred to as a chip comprised of 2 or more layers of active semiconductor devices that have been thinned, bonded, and interconnected to form a "monolithic" circuit. These layers, or tiers, can be fabricated in different processes. Performance is improved by reducing interconnect resistance, inductance, and capacitance for higher speed. 3D integrated circuits can have increased circuit density due to multiple tiers and at the same time be very thin since the individual layers can be very thin. Moreover, the technology provides the freedom to divide functionality among tiers to create new designs that simply are not possible in 2D [8]. For those who are not familiar with the 3D technology, the relevant background materials can be found in Appendix 8.

As mentioned before, the AMchip pattern density can be improved by optimizing the design in 2D, using custom cell designs with smaller feature size technology. There is an R&D effort in Italy using 90nm technology to improve the standard cell based AMchip03 design [2]. Some of this work is described in Section 2 and this work is very important for the near term improvement of the performance. Due to the limitations in technology scaling, the gain in performance is rather limited and may not be sufficient for applications at higher luminosity for LHC in the future. By adding a "third" dimension to the signal processing chain, new possibilities for improvements and performance may be achieved.

There are different ways to take advantage of the 3D technology for improving AM/CAM performance. The simplest way would be to stack a few identical tiers of AMchips to increase the pattern density. The AMchip has been already designed in such a way to allow many chips to be daisy-chained to replicate the structure hierarchically for expansion to a larger array on a PCB board. This design feature makes the simple vertical stacking of identical AMchips natural. More advanced 3D stacking would partition different functionalities into different tiers to optimize the performance. For example, one could put common functionalities such as control, interface or even global matching into one tier, while keeping the rest of the tiers identical and mostly for AM/CAM implementations. It has some unique advantages that match the special AM/CAM needs of tracking trigger for HEP. With Associative Memory for Pattern Recognition for HEP tracking trigger applications, the task is logically and physically dividable and the interconnections among them are straightforward. The Associative Memory is based on the concept of Content-Addressable-Memory (CAM). In its essence, Associative Memory bank is a CAM array that is just a collection of independent toads, i.e. independent sets of hit addresses from different detector layers which represent a path or road that a charged particle might traverse through the detector. The AM can be naturally divided into a control/interface tier and a set of CAM tiers. What is unique for the 3D AM (VIPRAM) architecture is that if each CAM tier corresponds to a single detector tier, then the interconnections between tiers become dramatically simplified. Logically, an AM road is an independent set of hit addresses from different detector layers, now physically, a road is just a simple independent vertical tube that is a collection of CAM cells each programmed to detect the hit on a particular detector layer for that particular road, and reports the match directly to the control tier. Routing in 3D can be very efficient, especially if functional elements are arranged such that the interconnections among tiers are mostly vertical. This is the case for VIPRAM architecture, not only the interconnects among tiers are vertical, they are uniform (in fact identical) across the tier as well.

The true 3D design approach would be more flexible, with more room for optimization, albeit more difficult and complex. We have been performing preliminary design work to compare the advantages and disadvantages of different approaches, taking into account the actual 3D process required for each approach. We will describe these approaches in Section 3.

We plan to carry out the R&D work in two stages. For the first stage we propose to use 3D technology to improve the AM chip performance (density and speed) for fast pattern recognition applications. Initial conceptual design work on a new 3D chip called Vertically Integrated Pattern Recognition Associative Memory, or VIPRAM, has already begun. For Phase II we plan to integrate both the AM stage (VIPRAM) and track fitting stage (FPGA + DRAMs + SRAMs) into a single chip, using either 3D stacking technique or a “system-on-package” approach. The final chip or package has the potential to dramatically enhance the overall track trigger performance for both Level 1 and Level 2 trigger applications by not only significantly improving the AM pattern density and speed, but also by increasing the data transfer bandwidth between the AM stage and the track fitting stage (as both are within the same chip or package). Because the track fitting stage is implemented in an FPGA, the final integrated chip will be flexible enough that it could be used for applications far beyond fast tracking trigger for HEP.

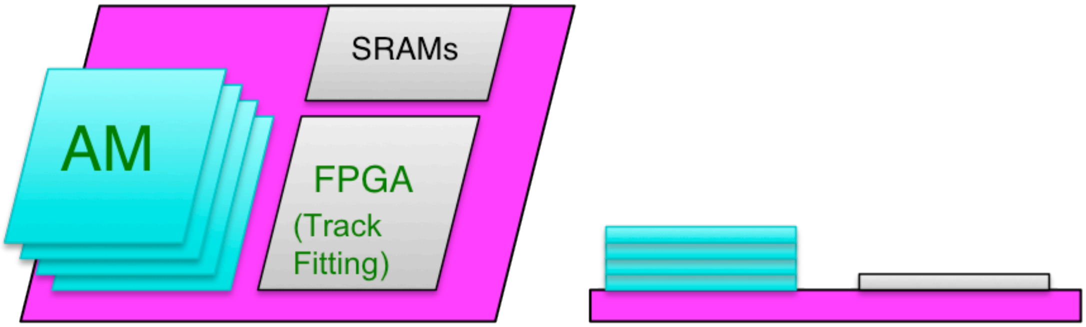


Figure 2. Possible VIPRAM design integrated with the FPGA-based track fitting.

In this proposal, we seek to first develop the design of 3D stacking of the VIPRAM, and perform the ASIC engineering necessary to realize a prototype device of the 3D stacked VIPRAM as our Phase I program. This can be done initially in 130 nm as prototyping (with initial goal to increase the pattern density by a factor of 40 x AMchip03) , and then likely move to 65 nm for the final chip (with > 100 x AMchip03). In Phase II, we plan to integrate the VIPRAM design with the FPGA-based track fitting stage (VIPRAM +FPGA+DRAMs+SRAMs) into a single chip, either using a system-on-package approach, or possibly, as the technology matures, to use vertical integration into one chip. The goal of Phase I is to solve the pattern density limitation in 2D design by vertical integration, while the goal of Phase II 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 ultimateexily what must 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 can 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 hit buffer to store the full resolution input hits into a database organized for rapid retrieval, as well as lookup tables for large sets of constants for track fitting purpose. With the addition of FPGA and RAMs, one could also implement the Tree Searching Processor (TSP) [4] using the same chip or package.

One of the main challenges of the 3D stacking and integration approach will be power and thermal issues [5]. There has been a lot work in reducing the power consumption in the new AM design in 2D for the AMchop04 R&D, and the 3D design will benefit directly from that effort. In addition, since we plan to follow Tezzaron's 3D DRAM stacking approach, we will learn a great deal from the Tezzaron extensive experience in addressing power and thermal issues of the 3D stacking. The AM/CAM structure is simple and repetitive but has relatively high power consumption, making it ideal for developing power and thermal analysis tools (for other 3D ASIC design applications). Because the tiers will be purely digital and the CAM tiers identical, this project is ideal in studying many of the challenging 3D stacking and packaging issues faced by other 3D R&D projects such as integration of sensors and digital electronics tiers. For this reason, it could be interesting to develop a power and thermal modeling of the design and perform power and thermal analysis. Some of the details are described in Appendix 9.

Besides the fast tracking trigger applications for the LHC energy frontier experiments, the proposed R&D might be useful for other experiments in the future as well. Possible examples include intensity frontier experiments such as μ2e, and cosmic frontier experiments such as ground-based telescope arrays where fast triggering on the correlation of images from multiple telescopes into the sky is needed. The 3D stacking and packaging technique to be developed from this proposal for both Phase I and Phase II will be useful for other 3D developments for particle physics researches. In addition, the power and thermal analysis methodology that could be developed (as a possible by-product) will be useful for future ASIC design as well.

1.5. The collaboration

The proposed R&D would be carried out as a collaborative effort between Fermilab, Argonne, UC, and INFN Italy, as well as other interested institutions. Some of the physicists in this collaboration have been involved in the design, building, commissioning, operation and upgrade of the CDF SVT system, including the recent AMchip upgrade and Gigafitter upgrade, as well as the current design work of the FTK system. The extensive experience in associative memory and track fitting approach within the collaboration will be important for carrying out this R&D project. In addition, this proposal will leverage unique areas of engineering expertise at Fermilab.

The 3D integrated circuit technology is actively being pursued by industry, since it enables heterogeneous integration of IC technologies, dense packing of transistors, and close integration of sensors and electronics. Partnering with an experienced industrial partner is considered key to the success of this project. Our partner in this R&D project is the company Tezzaron Seminconductor, located in Naperville, Illinois. Tezzaron is one of the world-leaders in developing the 3D technology and specializes in cutting-edge memory products, 3D wafer stacking and TSV processes. As will be described later, Tezzaron's revolutionary FaStack technology, which integrates several layers of DRAM with a powerful controller layer, will be used for the VIPRAM R&D work.

Fermilab was the first high-energy physics laboratory to recognize the potential of 3D integrated circuits for particle physics and has started a focused R&D program to explore this technology, and is currently recognized as the world leader in exploring this technology for high-energy physics applications. In addition, Fermilab has already been developing a 3D chip (VICTR) to demonstrate the application of 3D technology to the formation of track-trigger primitive for CMS Level One tracking trigger upgrade. The proposed 3D fast pattern recognition and track fitting R&D would leverage this other work in 3D technology that has already begun. Moreover, Fermilab has built a successful relation with Tezzaron over the course of the last few years. We plan to further develop our collaboration with them as part of this work.

2. R&D on Associated Memory Chips in 2D

2.1. Description of the Existing AMchip03

In 2004, colleagues at the Universities of Pisa and Ferrara developed a new CAM chip, the AMchip03, for the CDF SVT upgrade as well as future HEP needs [7]. The AMchip03 was implemented in 0.18um CMOS technology and a strictly standard-cell based VLSI design approach to minimize the design effort. The size of the die is approximately 9.8 x 9.8 mm^2. The 0.18um CMOS process (with 1 poly and 6 metal layers), available from the silicon foundry UMC, was chosen.

Figure 3 shows a block diagram of AMchip03. The pattern bank uses approximately 80% of the silicon resources in the device and contains 5120 patterns, corresponding to a total of approximately 500,000 content-addressable memory bits. For each pattern, approximately 30% of silicon resources is used for majority logic while 70% is actually used for patterns for the 6 layers of the SVT. In addition to the default 6-layer mode, the chip can be configured to perform pattern recognition for a detector of up to 12 layers, combining pairs of 6-word patterns into a 12-word pattern. In this case the 18th bit of each data bus is used to distinguish hits coming from two layers, multiplexed on a single bus. With the 12-layer configuration, the number of available patterns is 2560.

The minimum number of layers that have to be hit for a road to be declared matched is a free parameter and is set in a control register as a single programmable threshold common to all roads in the chip. Threshold comparison is introduced to account for silicon detector inefficiency. Priority is given to the matched patterns involving higher number of matched layers. The CAM technology, which allows this search to be performed within just a few clock cycles, is obviously the key component of the AMchip03. The AMchip03 has been designed in such a way that multiple chips can be cascaded in a daisy chain to build a larger pattern bank, and the daisy chains can in turn be multiplexed to form even larger banks.

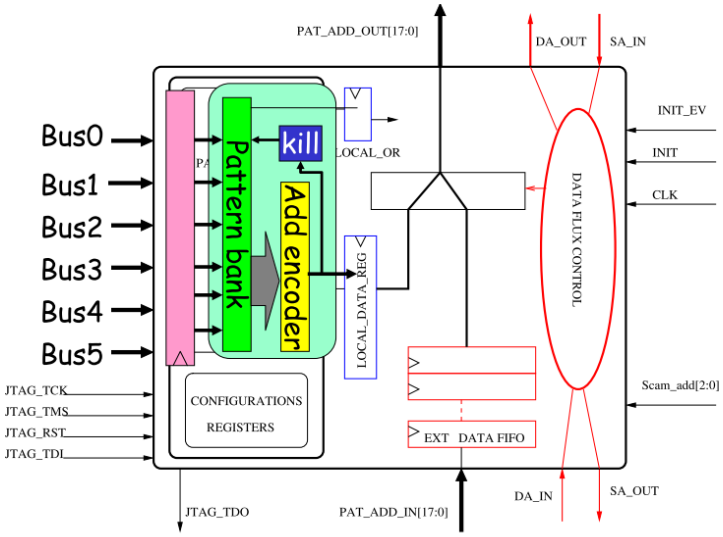


Figure 3. Block diagram of AMchip03.

It should be emphasized that compared to commercially available CAMs, such as Network Search Engine, the AMchip03 has the unique ability to search for correlations among input words received at different clock cycles. This is essential for tracking trigger applications since the input words are the detector hits arriving from different layers at different times. They arrive at the chip, serialized on six buses, without any specific timing correlation. Each pattern has to store each fired layer until the pattern is matched or the event is fully processed and thus all patterns can be reset. Even in the case of Level One trigger application, where things run in sync, this feature will be still useful.

2.2. AMchip04: Ongoing R&D Effort

As mentioned in the introduction, there is an on-going effort at Universities of Pisa, Ferrara and Frascati to develop a new AM chip, AMChiP04, using 90nm CMOS technology [2]. Reducing the feature size from 180nm to 90nm technology allows for higher pattern density integration. The CAM cell for the patterns is full custom design and this also allows further reduction of the size of a pattern. In fact, it has been shown that the size of a 15 bits CAM cell has been optimized to about 67.2 um x 2.8 um in area, with the full custom design in 90nm (see Matteo's talk at the First Mini-workshop on AM R&D, April 15th 2010; need a better reference than this). To reduce the power consumption, a pre-match power saving method is applied. (add a brief description of the pre-match power saving method here). The rest of the design is identical to the design of AMchip03. The initial plan of this ongoing R&D project is to produce a 90nm mini-ASIC to serve as a prototype for full custom 90nm technology while exploring power saving technique. The size of the mini-ASIC is 6.48 mm^2, about 7% of the AMchip03 area. In the future, it will be possible to explore the 65 nm technology, depending on the cost. The ultimate goal of this R&D is to improve the AMchip04 pattern density in 2D by a factor of 10 to 20 over that of the AMchip03. The design will be for 8 detector layers instead of 12, since the FTK system is being designed to operate on 8 layers at a time for a given pattern recognition step. Note that this part of the R&D has already been funded by INFN Italy, and the custom design of the CAM cell as well as the power saving techniques developed will be directly useful for the VIPRAM design.

Power consumption poses a challenge on the amount of patterns that can be stored in a chip, and this motivates investigation and the development of design techniques to further reduce the power consumption in the AM chip. Power dissipation in CAM is dominated by the dynamic power which is consumed by the match-line (ML) and search-line (SL) toggling during each clock cycle for search and match operations. The search-lines are switching to represent the new words to be compared and as a result match-lines are continuously switching based on the miss/match results. The dynamic power increases proportionally with the memory size. In the current AMchip03, the power dissipation is estimated as 1.8W for 5120 patterns stored in the memory, with 0.35mW per pattern [7]. AMchip04 design aims to reduce power consumption in a significant way using the pre-match power saving technique.

3. VIPRAM: Vertically Integrated Pattern Recognition Associative Memory

3.1. General Design Considerations

To increase the pattern density of the basic AM chip, we propose to investigate AM implementation using the 3D technology. Associated memories and content addressable memories are good candidates for 3D technology due to their regularity and uniform architecture. As described earlier, there are different ways to take advantage of the 3D technology to enhance the AM/CAM performance. The simplest way would be to simply stack identical chips, connected with through silicon vias (TSV). This can be done without dramatically altering the footprint or the architecture of the 2D design. This simple approach would allow one to make use of the existing 2D design, as well as the ongoing 2D design improvements, as a way to reach higher performance through the use of multiple tiers. In fact, the existing physical architecture of the AMchip is already designed to allow the chaining of multiple chips for expansion to a larger AM array on a printed circuit board, which makes it straight-forward for extrapolation to 3D stacking. This is the Identical Tier Architecture approach with all tiers sharing the same design and are identical. Alternatively, one can separate the AM/CAM cells from the rest of the control and interface logic, and use identical tiers for AM/CAM tasks while keeping the control and interface and other common functionalities in a single (different) tier. This is the True 3D Architecture with a control_tier and multiple identical CAM tiers. This is similar to the approach that Tezzaron has taken to implement their 3D DRAM stack. As will be described later, it turns out that this approach has some unique advantages that match well with the tracking trigger pattern recognition needs for HEP. The advantages and disadvantages with the two approaches will be described later.

3.2. Paths toward a 3D AM: VIPRAM

3.2.1. Identical Tier 3D Architecture

The Identical Tier Architecture uses as much of the existing AMchip03 as possible and incorporates 3D enhancements to convert the 2-dimensional AMchip03 into a 3D VIPRAM stack. Note that Identical Tier Architecture does not imply that each tier is identical to an AMchip03. Rather, Identical Tier Architecture is so named because every tier in the architecture is identical in every way to every other tier. Only one mask set is necessary for this architecture, and this is a cost advantage to this scheme. Some modifications to the existing AMchip03 design will be required. For example, the ability to automatically determine a particular tier’s location in the stack, called Self ID or Tier ID, is required by this architecture due to the fact that all tiers are physically identical.

Note that this is different compared to the case where many Amchip03 chips are chained together to form a larger array on a PCB, where the glue logic block has dedicated address lines to address each AMchip in the chain.

A small change is needed to adapt the AMchip03 design to 3D to allow each tier to be independently recognized and addressed. . Therefore, the first step in the Identical Tier Architecture is to create a chip that contains an AMchip03 wrapped in an IO interface that knows about the existence of other 3D tiers and converts any necessary signals to "tier specific" 2D signals as shown in Figure 4.

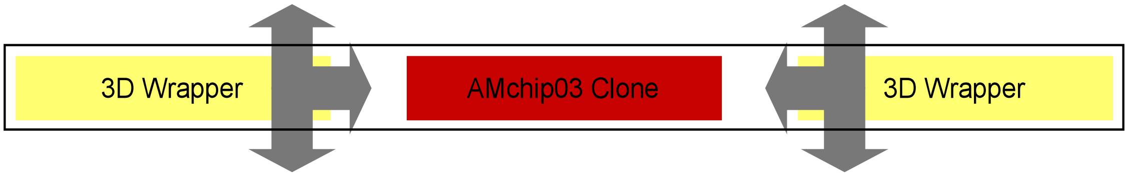


Figure 4. 3D integration of AMchip03 wrapped with IO interface.

The advantage of this approach is that the AMchip03 at the core of each tier is essentially unchanged from the existing design. The AMchip03 was designed to operate in a daisy chain. All chips on a chain received candidate hits in parallel, but road addresses output by one chip would be accepted as inputs by the next chip on the daisy chain. This next chip would append its road addresses onto the road addresses output by the previous chip and then pass the set on down the daisy chain. This capability can be mapped into 3-dimensions once the self-ID (or Tier-ID) mechanism is in place. There are different ways to implement the Tier-ID. One possible approach involves a new 3D pad structure and is described in Appendix 8.

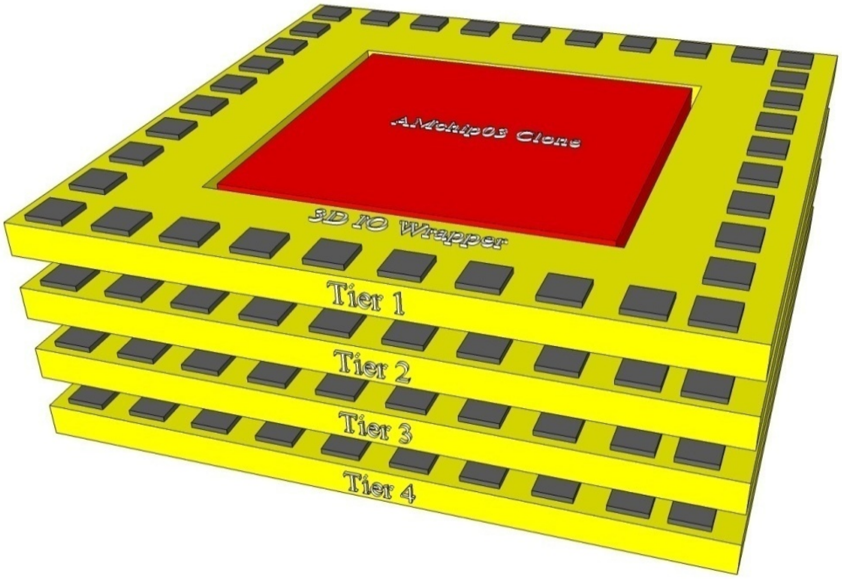


Figure 5. Identical tier integration of the AMchip03 device.

Although the AMchip03 at the core of each tier is unchanged from the original, 2D AMchip03, it does not imply that no design effort is necessary. The 3D IO Wrapper must be designed and fabricated around the AMchip03 to ensure that all tiers act as a single chip as shown in Figure 5. Even for prototyping purpose, it is not possible to simply take an existing AMchip03 and place it inside a rectangular doughnut-shaped 3D IO Wrapper. There are several ways to proceed with a 3D version of the AMchip03. One way is to have the AMchip03 redesigned in a 3D process like Tezzaron/Chartered, and then the 3D IO Wrapper could be designed around it. This method has no obstacles to its 3D fabrication. However, it does require the redesign of the AMchip03. Another approach is to stick with the 180nm CMOS UMC process which could be used for 3D development even though UMC does not have a 3D process. This method requires no redesign of the AMchip03, but it does require UMC to be willing to participate in a "Via Middle" process in which after a certain number of fabrication steps, the wafers are shipped to a "Via Middle company" (e.g. Tezzaron) where the first steps of the Through Silicon Via process are started. Then the wafers are shipped back to UMC where the 2D processing is completed. Finally, UMC ships the completed wafers to the Via Middle Company where the 3D processing is completed. Not all companies are willing to participate in a Via Middle process.

This simple stacking of identical tiers approach has the advantages of reusing (most if not all) the existing design and testability. In addition, by having identical tiers with the same footprint, only one mask set will be needed. Moreover, the required 3D interconnection density is very low, mostly on the pads, so that the demands on the new 3D technology is very moderate. The overall gain in performance on the other hand, could still be potentially large and increase linearly with the number of tiers.. As mentioned earlier, power and thermal issues are a concern. If power and thermal issues prevent the stacking of 4 or more tiers (to run at high enough speed), then it will be necessary to consider alternative designs, such as True 3D. This is a real concern with the identical tier design approach and power consumption reduction in 2D is crucial. As will be described later, the True 3D approach with Control + CAM tier is potentially a much better approach in this regard.

3.2.2. General Issues with the Identical Tier Architecture

The Identical Tier Architecture may not be the long-term solution unless new information comes to light over the course of the research. The main issues are the following:

1. **Power/Thermal** – Power and therefore heat dissipation are already issues with CAM architectures in general and with the AMchip03 architecture in particular. The majority of the power consumption in the AMchip0X designs is in the input/output sections of the chip because of the speed of operation and in the CAM match lines because of the nature of a CAM. The Identical Tier Architecture simply repeats the IO sections of the AMchip design for every tier. Therefore, the Identical Tier architecture actually multiplies the power consumption by the number of tiers in the design. This is

of particular concern for interior tiers that will be forced to dissipate their power through adjacent tiers.

2. **Speed** – The speed of the new device must be greater by a significant margin, than what is possible with the AMchip03 architecture. The Identical Tier Architecture fundamentally cannot be faster than the AMchip0x that it is based upon. Each tier of the Identical Tier Architecture can be thought of as one chip in a daisy chain. In the existing design, this is implemented on a printed circuit board, which presumably has greater load capacitances. The speed of the Identical Tier Architecture should be (at best) the same as an AMchip on a board and could be even slower due to thermal issues for the interior tiers. Again, the issue here is related to power consumption.

3. **Limited Improvement** – The 3D AM is required to increase the pattern density of the associative memory chips, and the Identical Tier Architecture would do this. However, stacking N tiers one above the other only increases the pattern density by a factor of N. There is no room in the architecture for exploiting the unique capabilities of 3D to optimize the pattern density, as well as reducing power consumption or perform much thermal management. Moreover, the control and interface logic, as well as IO structures in the AMchip, are simply repeated on each tier. This is not the best utilization of resources.

For these reasons, the Identical Tier Architecture is (so far) not th**e preferred approac**h for long term. Work, however, will continue, and the architecture is not being dismissed entirely. In fact, it was the first architecture we have seriously considered for obvious reasons. For example, it is still possible to consider it as a near term (intermediate step) solution given all its advantages. This will enable us to explore the 3D approach in a rather short time frame and get to a better understanding of various issues such as power, speed, and fabrication.

3.2.3. "True 3D" Implementation of VIPRAM

The obvious alternative to an "all layers identical" 3D architecture is an "all layers NOT identical" 3D architecture - in other words, permitting the designers the freedom to move functionality from one tier to another, to eliminate certain functions from a layer and to optimize a given tier for a given function. This is the true 3D architecture. However, the cost of a design is always a factor, whether or not it can be produced and for each unique tier in a cost-effective way, since the design, mask set , fabrication costs, and yield for each tier are unique and substantial. A good True 3D design will need to strike a balance between "all layers NOT identical" and "all layers NOT unique."

There are several possible approaches to a true 3D CAM chip, and it will be the subject of some R&D to discover which of the approaches is the best one. This is one of the objectives of this research. It should be remembered that, while the terms CAM and 3D CAM are used regularly to describe this work, those terms exist for historical reasons.

In fact, what have been discussed in this proposal are not simple CAMs. They are tracking associative memories or PRAMs – Pattern Recognition Associative Memories. Like CAMs, their basic operation is to compare candidate addresses to stored address patterns. Unlike CAMs, they do not flag matches until they have matched candidate addresses from multiple sources (tiers representing layers) to an array or road of stored address patterns. This gives rise to a very natural 3D progression. Perhaps one way to see this more clearly is to view the Associative Memory architecture in Figure 6 in a different way: rotated by 90 degree. The idea is to have a dedicated CAM tier for each detector layer, where the incoming hits are matched to the stored hit locations, and have one control/interface tier to collect and associate the hit matching information

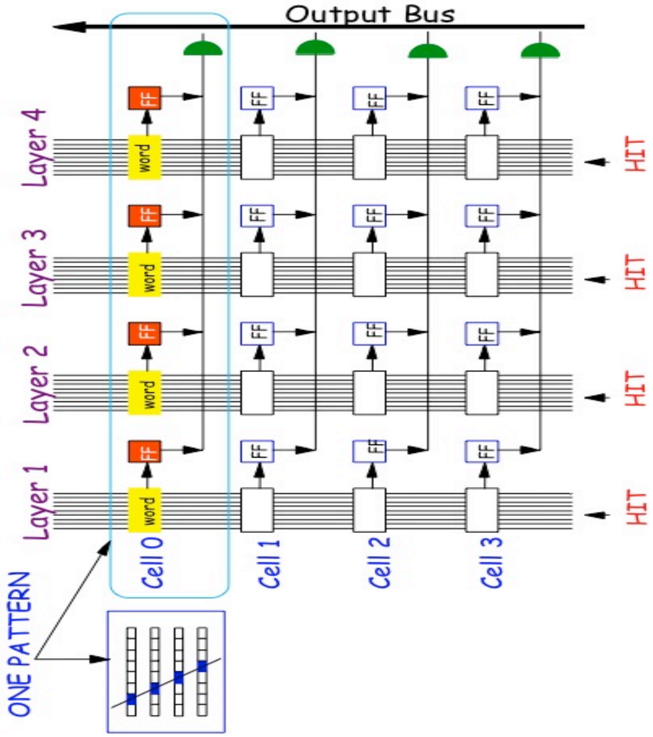


Figure 6. Associative memory architecture

How CAM works

To take full advantage of the 3D technology, it is useful to provide an overview of the fundamental architecture of both the CAM as well as the Associative Memory. We will first take a look at the basic architecture of conventional CAM, identify the uniqueness of the HEP AM, and describe how we might take advantage of the 3D technology to enhance the AM performance.

Conventional CAMs store an array of address patterns that a user wishes to compare to a stream of candidate addresses [9]. Each new candidate address is presented to the chip where it is compared simultaneously to each stored address pattern in the array. If there is a match, a flag is raised. This simple algorithm is somewhat complicated by the fact that more than one stored address pattern can flag a match. In such a case, a priority encoder must select one of the matches as the CAM's chosen match.

A CAM has a regular architecture with a few basic components such as CAM cell, search-lines (SL), match-lines (ML) and match-line sense amplifiers (MLSA) [9]. A CAM cell serves for two basic functions: bit storage and bit comparison which can be a NOR or NAND-type cell. When multiple cells are connected in parallel to form a CAM word the match-line of each cell is shorted to the ML of adjacent cell. In the case of AMchip03, the design is for 6 detector layers and each has about 16 bits, therefore the total number of CAM bits is about 100. For large CAM bits like this, it can create a long ML line which has parasitic resistance and capacitance and contributes to the power consumption. Figure 7 shows a CAM model consisting of 4 words, with each word containing 5 bits arranged horizontally. A CAM search operation begins with loading the search-data word into the search-data registers followed by the pre-charging all match-lines high, putting them all temporarily in match state. Next, the search-line drivers broadcast the search word onto the search-lines, and each CAM cell compares its stored bit against the bits on its search-lines. If there is a match, the match-lines remain high and in case of a miss, match-lines discharge to ground. Match-line sense amplifiers detect whether ML has a match or a miss condition. Finally, the encoder maps the matching location to its matching address [10] (more references here).

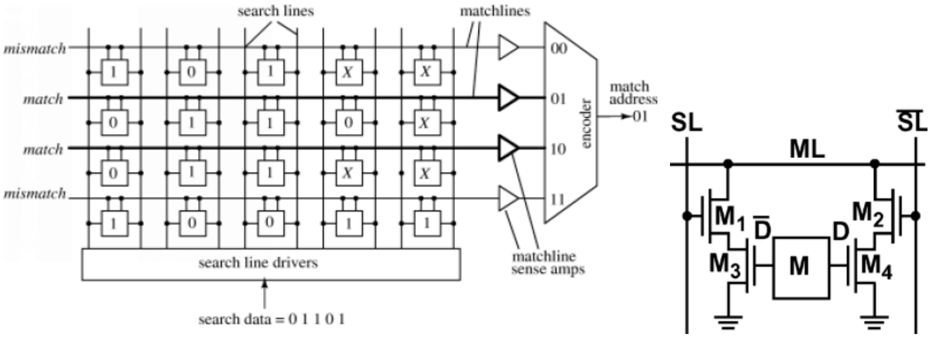


Figure 7. CAM model for 4 words with each word containing 5 bits.

CAM compares input search data against a table of stored data and returns the address of the matching data. CAMs have a single clock cycle throughput making them faster than other hardware or software based search systems. However, the speed of a CAM comes at the cost of increased silicon area and power consumption. As CAM size increases, so does the power consumption. Thus, power reduction is the main challenge in CAM design without sacrificing speed or area.

The unique requirement for AM for HEP application

As mentioned earlier, the CDF SVT Associative Memory chip [11] (from now on we will call it PRAM, Pattern Recognition Associative Memory, to emphasis its purpose for HEP) is a departure beyond the conventional CAMs. Like conventional CAMs, PRAMs store address patterns and look for matches between incoming hits and stored hit locations for a given detector layer. At this level, the match is expected to be exact (Binary CAM, instead of Ternary CAM) and an array of Match Flags is the typical output. A PRAM has an array of Match Flag Latches which capture and hold the results of the match until reset (for next event). As the hits from the various layers of the detector for the same event arrive, the PRAM is looking for more than simple matches from one candidate address to one or more stored address patterns. PRAM organizes stored address patterns into roads which are linked arrays of several stored address patterns from different detector layers. Each stored address pattern in a road is from a different layer in the detector system and these linked arrays represent a path or road that a particle might traverse through the layers of the detector (hence the name "road"). The ultimate goal of the PRAM is to match real particle trajectories to those roads. Like a conventional CAM, a PRAM flags a match when a candidate address matches a stored pattern address for a given detector layer. However, before the PRAM does anything with that match, it must find matches in all the elements (layers) that constitute a road. Pattern masking also exists in PRAM, but rather than being bit masks on an individual candidate address, they are actually layer masks that will allow the flagging of a road match even if there is a missing detector layer. Therefore, priority encoding remains necessary though its implementation is at a different level.

In a PRAM it is logical to divide the Pattern Address Array into banks of address pattern by layer number. This effectively divides the PRAM into N parallel conventional CAMs, one for each detector layer. In standard 2D integration, such a division increases the design size due to the routing necessary to link each Match Flag in a road. In 3D, that routing area can be virtually eliminated.

As mentioned earlier, Tezzaron Semiconductor develops multi-tiered 3D memory arrays. While these are not CAM arrays, CAM arrays and memory arrays share much in common. Using a true 3D approach, Tezzaron divides the functionality between the tiers. Being cost conscious, they limit to two types of tiers. The top tier is the control tier and it contains the IO logic, the sense arrays, the decode logic and the address line drivers. The remaining tiers are DRAM tiers, connected to the control tier by through silicon vias. Tezzaron takes the further step of fabricating the two different types of tiers in two different CMOS processes. The control tier is optimized by using CMOS high-speed processes that create high-performance transistors. The DRAM tiers use a high-density NMOS process that creates high-quality capacitors. The end result is a faster, denser memory without any changes to the design[1]. This approach is also remarkably similar to the proposed logical, layer-by-layer division of a particle physics PRAM. The top tier or Control tier, houses the IO and Road Glue Logic. The lower tiers or CAM tiers– one per detector layer – house the individual CAM arrays.

[1] See http://www.tezzaron.com/memory/FaStack_memory.html

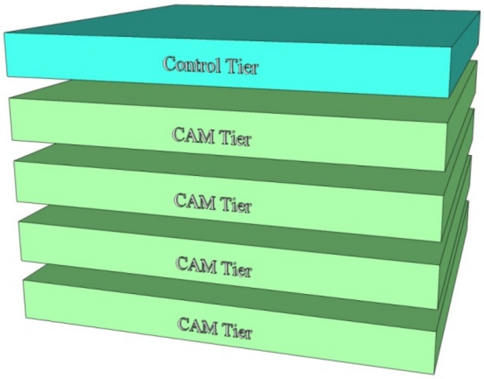


Figure 8. Control and CAM tiers for 3D implementation of VIPRAM.

Natural Progression From 2D to 3D: VIPRAM

The simple CAM operation of comparing a candidate address to a stored pattern maps well into 2 dimensions. The candidate addresses are driven along bit lines in one direction. Individual bit lines are compared to stored bits and the match line is either pulled to a zero (no match) or it remains a one (match). This is a CAM word cell, as shown in Figure 9 and 10, where only 6 CAM bits are shown for simplicity (note that for the actual VIPRAM design, there will be 15 or 16 CAM bits per cell).

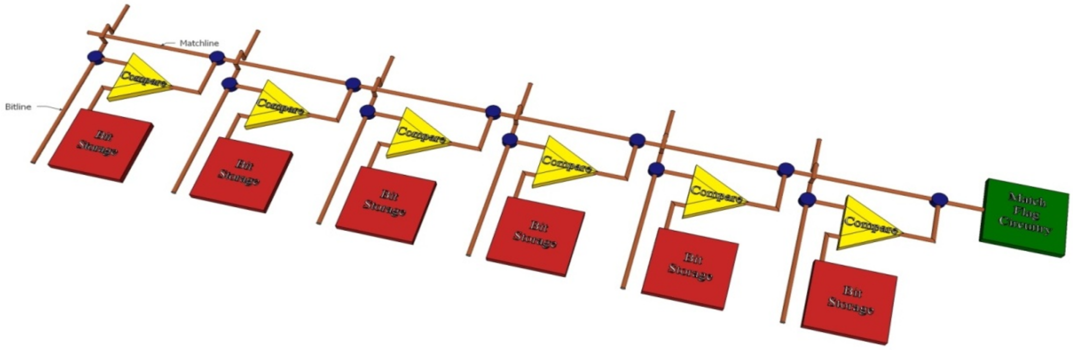


Figure 9. An array of CAM words. This is a “classic Binary CAM” in 2D.

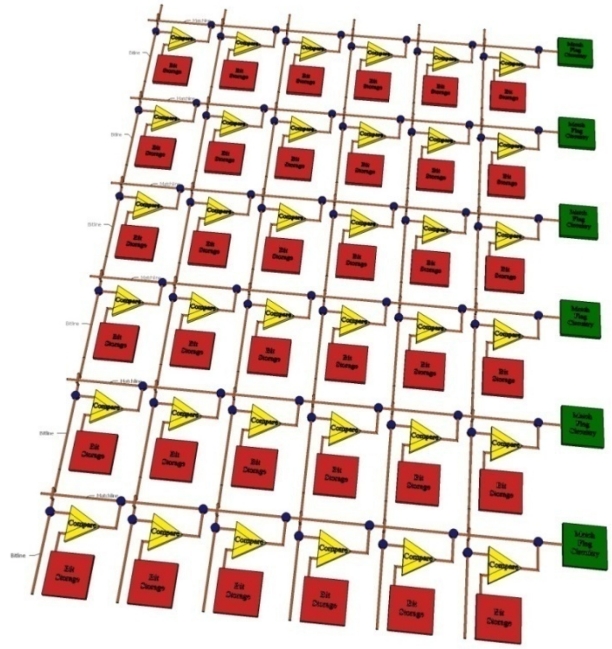


Figure 10. Multiple words in 2D CAM words.

Individual patterns are stored “horizontally” in the rows of red bit storage squares. Candidate addresses are driven “vertically” along the bit lines and match flags are similarly arrayed vertically in the green boxes as shown in Figure 11. PRAM operation is somewhat more complicated.

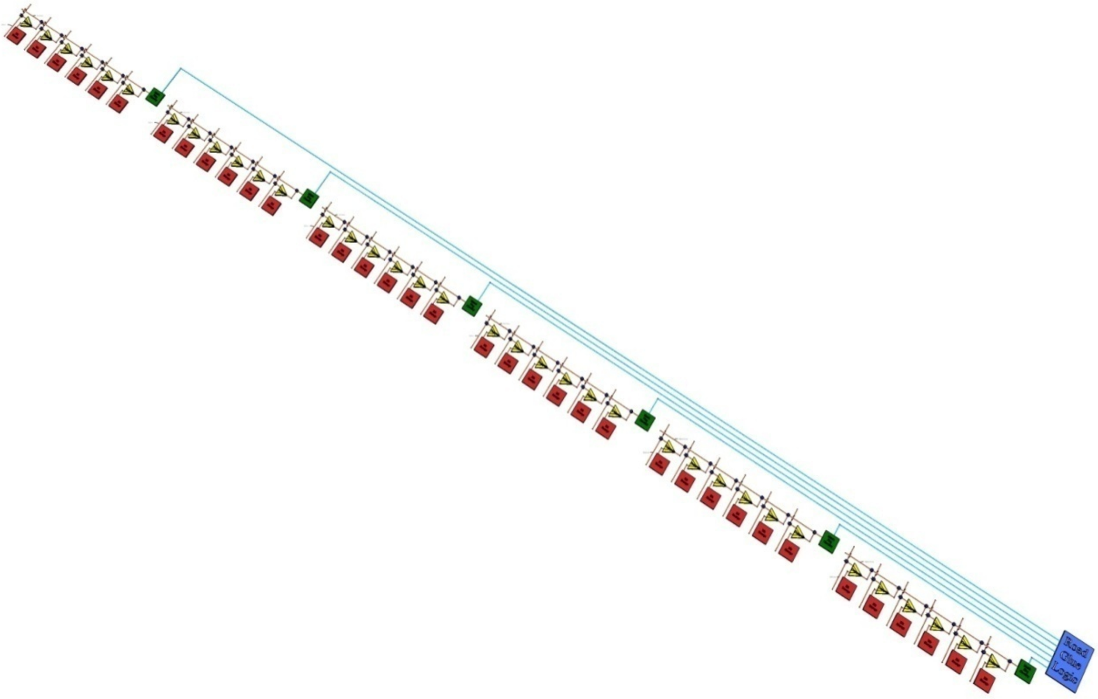


Figure 11. CAM word cells in 2D.

In a PRAM, individual patterns are stored horizontally like they are in a classic Binary CAM, but now several patterns – one per detector layer - are stored. Candidate addresses are driven vertically like they are in a classic CAM, but now several candidate addresses – one per layer – are driven. In 2D, a PRAM can be thought of as an array of classic CAMs laid out side-by-side column-wise with an extra set of Road Glue Logic connecting each row. This is shown in Figure 12. Each column is a classic Binary CAM dedicated to one particular detector layer. Each row is all the circuitry necessary for one complete road detection.

A significant improvement can be obtained when going in a third dimension. There is no need to spread the different patterns out horizontally; instead stack them vertically with one layer per tier. There is no need to complicate and bloat the horizontal routing of signals; instead route the individual match lines to a Road Glue Logic on a top tier. The result (the design of VIPRAM) looks like Figure 13, where the vertical blue tube represents one independent pattern or road.

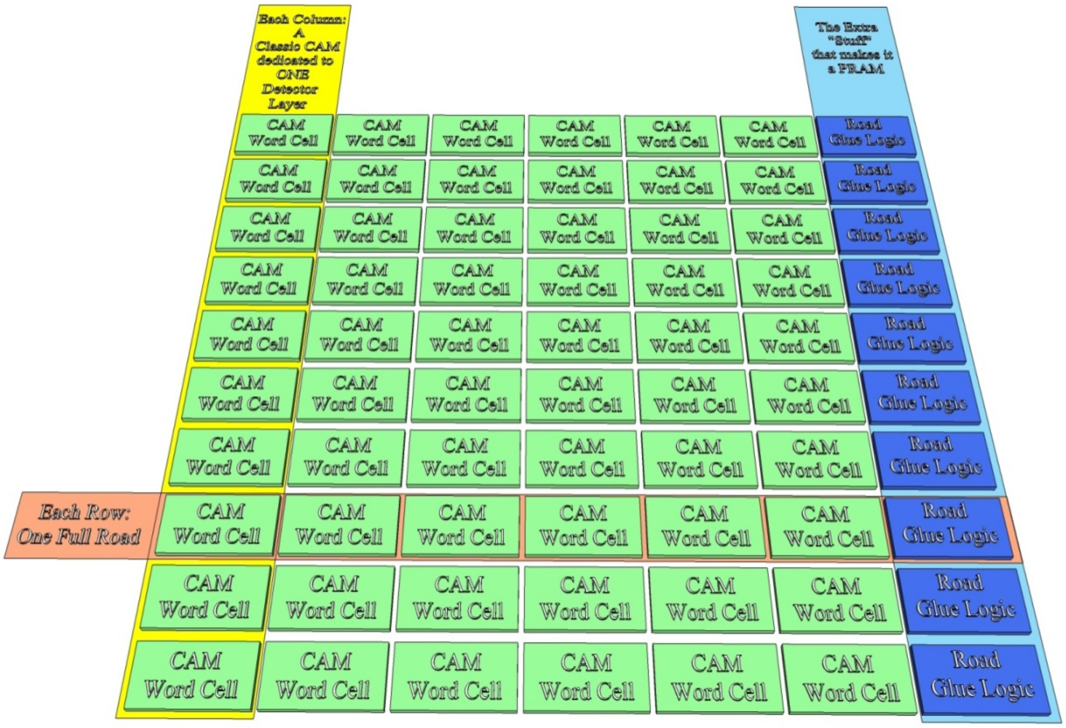


Figure 12. A PRAM in 2D.

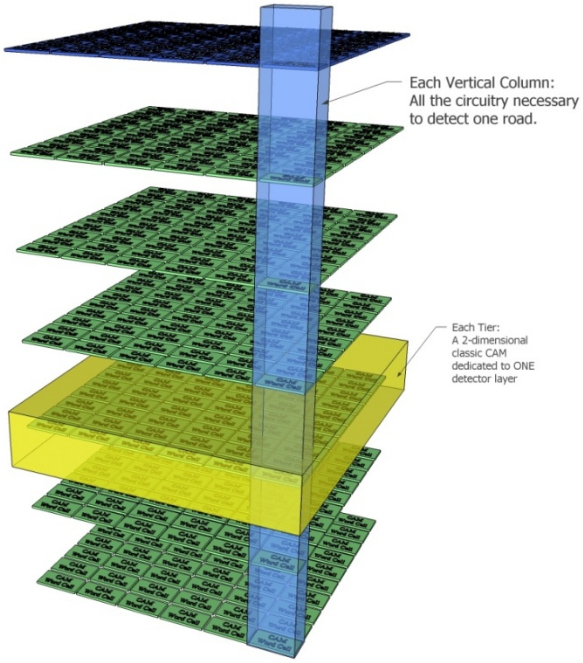


Figure 13. A PRAM in 3D where vertical blue tube represents one independent road

Each tier now resembles a classic CAM in two dimensions. Each tier accommodates only one detector layer of data. Candidate addresses (i.e. hits) of a particular detector layer are driven along the bit lines of one tier and one tier only. Hit patterns are stored on the tier that corresponds to their detector layer. Each match line from a given layer is driven vertically and directly into the Control tier (the details on how match lines are driven vertically from the identical CAM tiers to the control tier is described in Appendix 8). In a 2D footprint the size of a single CAM word cell, all of the road detection circuitry can be implemented over a few tiers. This means that in an area that once contained "R" road patterns, a 3D PRAM can process "R x L" road patterns, where L is the number of detector layers[2]. This is one of the main architectural advantages for HEP PRAM using 3D technology.

Previous work in 3D CAM architectures looked into vertical integration along the CAM match line[3]. In other words, each tier became a single CAM bit of an address word and the match line –i.e. the logical signal that combines the results from all the CAM bits – was passed upward from layer to layer. The top tier became a two dimensional array of match flags and the area consumed by each pattern would be roughly the area of a single CAM bit. This approach is logical, and it works for simple CAMs. It provides the greatest possible pattern density. It falls somewhat short, however, in two respects. First and foremost, in the case of particle physics Pattern Recognition Associative Memories, there would be an excessive amount of routing required to bring the individual bits of the various different detector layer addresses to all the different tiers. Second, it is anticipated that there will be at least 15 address bits for each CAM cell per layer and this would require 15 or more tiers in a 3D design.

Expected Areas of Improvement in the True 3D Architecture

As described previously, True 3D architecture offers an increase in pattern density over 2D designs. There are other expected improvements as well:

1. **Power/Thermal** – Only the control tier will operate at full speed – i.e. data from all detector layers pushing through the tier. CAM tiers will operate at 1/L speed (where L is the number of tiers or detector layers) – i.e. only data from a particular detector layer will be pushed onto a particular CAM tier. Therefore, the sensitive internal layers should have lower power consumption.
2. **Speed** – The Control Tier of the VIPRAM contains a 2 dimensional array of Road Glue Logic cells. Moreover, there should be much less routing on both the Control Tier as well as the CAM tiers within this 2 dimensional array, comparing to the AMchip0x 2D design. The extra routing space and the regularity of the 2 dimensional arrays of Road Glue Logic cells can be exploited for speed. For example, the actual road

[2] Actually, this is conservative. "R" road patterns in a 2D PRAM took up the area of "RxL" CAM words plus the area of R Road Glue Logic Cell plus extra area for routing. Therefore, we should fit *more* than RxL roads in the same area occupied by R roads in a 2D PRAM.

[3] Y-J. Hu, J-F Li, and Y-J Huang, "3-D Content Addressable Memory Architectures", 2009 IEEE International Workshop on Memory Technology, Design and Testing, p. 59 (2009)

addresses can be placed on the periphery. A detected road would activate one "row" address and one "column" address which, taken together, would constitute the unique address of the detected road. Previous work in pixel readout architectures can be used to maximize the VIPRAM's speed.

Brief Summary of the Advantages of the True 3D approach

Here is a list of potential advantages of the True 3D approach:

1. **Simplicity** – conceptually, the design is still rather simple, with one control tier and multiple CAM tiers. The Tier ID issue is solved using a simple and proven "diagonal via" trick, without the need of any extra transistor. The number of CAM tiers that can be stacked will be determined by the number of "diagonal vias", see Appendix 8.
2. **Established Approach -** the architecture and 3D process involved will be very similar to the Tezzaron 3D DRAM stacking. The design will benefit from Tezzaron's extensive experiences.
3. Naturalness – HEP tracking is naturally a 3D task, using a CAM tier to represent a detector layer is natural and unique for 3D application for HEP PRAM.
4. Less Routing – much less routing congestion for both Control Tier and CAM tier, and large reduction of interconnect lengths.
5. Pattern Density – will increase by a factor that is larger than N (# of CAM tiers). This is because all the control, interface and majority logic (~ 50%) will be moved to the Control Tier, leaving more space for CAM implementation on the CAM tiers.
6. Flexibility in implementation – Control/CAM Tier can be designed in different technologies to optimize performance
7. Easy for collaboration – CAM tier optimization can be more focused and independent from the Control tier design as long as interface specification is agreed upon followed (full custom design, optimize for size). This allows better division of tasks.
8. Much more room for power and thermal management/optimization using 3D. All the known power reduction techniques could be used within this architecture (see Appendix 8.
9. Uniformity – the design is such that the TSVs and CAM cell blocks are uniform and evenly distributed across the CAM tier.
10. Architecture allows stacking of a minimum of two tiers for prototyping and proof of principle.
11. More detector layers can be implemented without decreasing pattern numbers per CAM tier

> **A rough estimate shows that with this approach, one could gain at least two orders of magnitude in the total number of possible patterns over the AMchip03 (see Appendix 8). One of the goals of the present R&D is to quantify this gain.**

The one true disadvantage of the true 3D VIPRAM approach is the need of two mask sets and separate designs which would be more costly.

4. The VIPRAM 3D Stacking Process

4.1. Overview

A wide range of 3D integration approaches is possible and available, including simple chip stacks, silicon chip carriers and interposers, chip-to-wafer stack, and full wafer-level integration. Each has advantages and trade-offs, and each is driven by commercial demands. Special process flow may be needed to meet the needs of specific applications. The process choice will impact the overall design, planning and ultimately the cost and will need to be studied carefully. For example, the orientation of the die in the 3D stack has important implications for the design. The "face-to-back" method is based on bonding the front side of the bottom die with the back side (usually thinned) of the top die. In order to construct such a stack, a handle wafer must be utilized. The "face-to-face" approach focuses on joining the front sides of two wafers. Alignment tolerances can have a direct impact on the density of connections achievable in the 3D stack, and thus the overall performance. The dimensions of the TSV are important to 3D designers since they directly impact exclusion zones where designers cannot place transistors and in some cases, back-end-of-line wiring as well. The dimensions of TSV are dependent on the 3D process used to fabricate them and are a function of silicon thickness, aspect ratio and sidewall taper, and other process considerations. The specific process flow used to fabricate the TSV is of importance since the method of via processing drives specific design rules, such as "vias-first" and "vias-last process. For more background information on the 3D process involved, please see Appendix 9 or the Handbook of 3D Integration: Technology and Applications of 3D Integrated Circuits (edited by Philip Garrou, Christopher Bower, Peter Ramm).

These considerations all have direct implications on design and will be important in both the selection of 3D processes and the optimization of circuits within a given 3D process. Since the 3D technology is an emerging technology, there are many different processes out there and it is desirable for a prototype to take an established approach that has a high probability of success. For this proposal, we choose to follow the Tezzaron's 3D stacking approach for its 3D DRAMs (for details, please see Appendix 9), and some of the cost estimate shown is based on the actual Tezzaron's experience with this approach.

4.2. The Issue of Yield Management

(This is a placeholder for yield discussions… more later. One good example is the paper "Techniques for producing 3D ICs with high-density interconnect", by S. Gupta etc).

Any successful approach to 3D integration must address both technical and economic issues. True 3D integration requires very high density vertical interconnects and the possibility of fabricating 3D ICs creates a desire to build much larger semiconductor systems. This, in turn, raises the specter of significant yield difficulties.

The yield problem could be mitigated by improving repair and redundancy schemes that are enabled by the same wiring improvements that enable 3D integration.

The inherent benefits of the 3D interconnect, more routability and closer proximity to a larger number of transistors, provide the fundamentals to solving the yield issue. The key to yield lies not in eliminating the bad transistors, but in providing fast and easy access to the good transistors. This is the Tezzaron approach to yield issues. More description here on yield issues ….

4.3 3D Process Options for VIPRAM Prototyping

MPW Prototyping

Multi-project prototyping to keep the cost down is a viable option for the VIPRAM project. The 3D MPW runs that have been available to Fermilab will be available in the foreseeable future. For the discussion below, we will assume wafer-to-wafer stacking. Other options are also available.

These runs are two-tier single mask set processes. This means that the typical user will only get one tier stacked upon another and that both tiers will be placed on the same reticule in the layout mask set. One of the two tiers is actually flipped in the layout and the two tiers are interconnected via face-to-face bonding. Therefore, when two wafers are brought together, one of them is flipped and placed on top of the other before the bond is formed. This results in the “flipped” layout being “unflipped” in fabrication. In Figure 18, note that A Left and A Right as well as B Left and B Right tier are placed symmetrically across the vertical axis of the reticule.

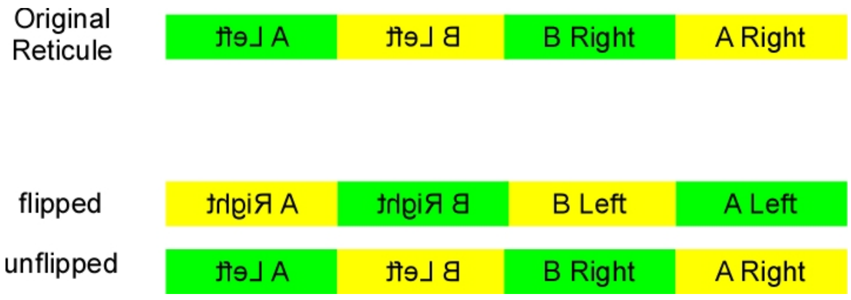


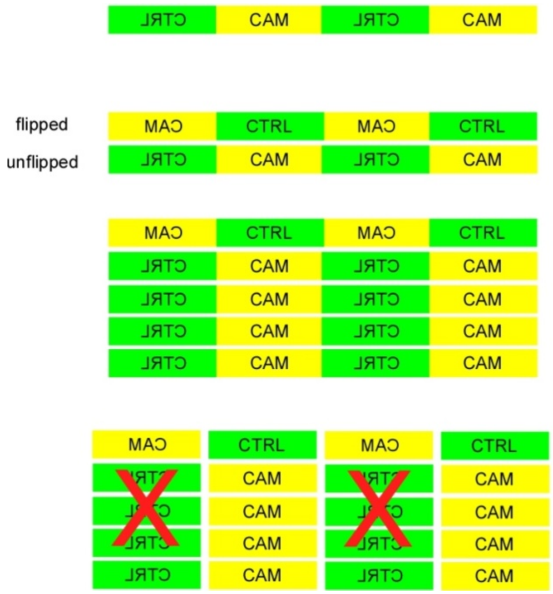
Figure 18. Reticules in wafers.

After dicing, fifty percent of the 3D stacks are upside-down and backwards. These stacks will most probably be worthless. However, fifty percent of the stacks will be available for testing. In its simplest form, 3D MPW runs will give the VIPRAM project the opportunity to test a single Control Tier with a single CAM tier, and they will do so at a very reduced cost. It is also possible (for more money) to reserve two spaces on a 3D MPW run. With two spaces, it is possible to fabricate a flipped Control Tier opposite a normal CAM Tier and then a flipped CAM Tier opposite a second normal CAM Tier. The flipped Control Tier and the normal CAM Tier are symmetric across the vertical axis as are the flipped CAM Tier and the second normal CAM Tier. The two pairs are also themselves placed symmetrically across the horizontal axis. This allows for a back-to-back bonding that effectively stacks three CAM tiers under one Control Tier. Of course, there is a 75% yield hit when this procedure is done, but it does allow for multi-tier (more than 2) stacking in 3D MPW runs.

The cost for such a run is difficult to predict at this time. A collaboration of MOSIS, CMP and CMC has been formed, recently, based on the success of the 3D Consortium started by Fermilab and is moving 3D MPW runs to the next level, providing regularly scheduled, publicly available 3D MPW runs. Their first submission offering is tentatively scheduled for the fourth quarter of 2010. The details, the frequency of runs and the cost breakdowns are not yet publicly available. Regardless, the VIPRAM project should probably be regarded as a public 3D MPW run. That is to say, if a 3D MPW fits into our schedule and if the project feels that something can be learned from a two-tier design, we should consider them.

Single Mask Set Prototyping

It is also possible to use a single mask set in a dedicated run. This will allow the VIPRAM project to stack a virtually unlimited number of tiers at a consistent 50% yield hit. Figure 19 shows the steps.



1. Start with 25-30 wafers of reticles containing both CTRL Tier and CAM Tier (Single Mask Set). “Flip” the CTRL Tier’s layout in the reticle before the 2D wafers are processed.
2. Flip one wafer, place it over another and create a Face-to-Face bond between the two wafers.
3. Stack additional wafers (without flipping) and form Face-to-Back bonds with the first wafer pair. Repeat this as needed to form the reqired number of tiers.
4. Dice and throw away the “inverted” stacks. Results in an automatic 50% loss to yield.

Figure 19. Single mask set prototyping.

The approximate cost breakdown for such a project in 130nm CMOS was informally quoted by Tezzaron Semiconductor in July of 2010 as follows:

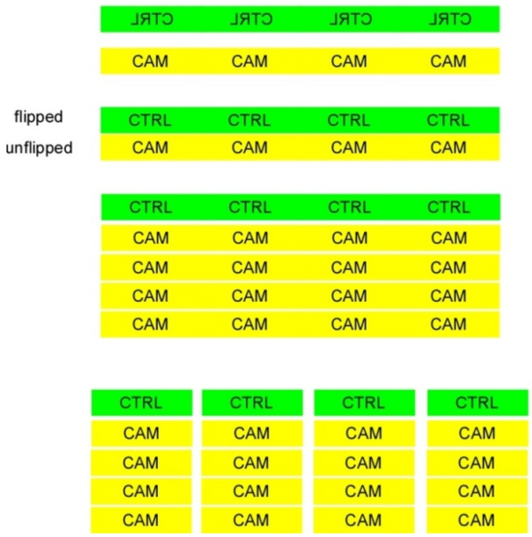
Mask Cost	$150,000
Wafer Cost (25 wafers)	$75,000

3D Fabrication	$35,000
Total	**$260,000**

This would result in approximately 500 3D stacks. This number takes into account the 50% yield loss due to the single mask set, but it does not take into account any yield loss due to fabrication or mask issues.

Dual Mask Set Prototyping

The most expensive type of prototyping is a dual mask set dedicated run. This type of run contains all the steps necessary for a fabrication run. Please note, however, that the current expectations for the VIPRAM is that in its final design it will be in 65nm CMOS and not in the 130nm CMOS priced here. In other words, no one should mistake this price for a quote of the final price of the VIPRAMs. In the Dual Mask Set, one mask set is used for each of the Control and CAM tiers. Extra wafers are fabricated for the CAM tiers depending on how many CAM tiers are necessary per Control tier. One Control tier is bonded face-to-face to one CAM tier and then subsequent CAM tiers are bonded to the original pair via face-to-back bonding. This is shown in Figure 20.



1. Start with N Control Tier wafers and 4*N CAM Tier wafers. The Control Tier wafers should have their 2D layouts flipped across the vertical axis.

2. Flip one Control Tier wafer, place it over one CAM Tier wafer and create a Face-to-Face bond between the two wafers.

3. Stack additional wafers (without flipping) and form Face-to-Back bonds with the first wafer pair. Repeat this as needed to form the reqired number of tiers.

4. Dice. No necessary process-related yield loss.

Figure 20. Dual mask set prototyping.

There is no process-related yield loss because all fabricated stacks are right-side-up and properly oriented. The cost breakdown for 1500 3D stacks in a 130nm CMOS process is as follows assuming one Control Tier to four CAM Tiers:

Control Tier Mask Cost	$150,000
CAM Tier Mask Cost	$150,000
Control Tier Wafer Cost (6 wafers)	$18,000
CAM Tier Wafer Cost (24 wafers)	$72,000

3D Fabrication	$35,000
Total	**$425,000**

A dual mask set dedicated run is at this moment of questionable benefit. Perhaps, in the future, the benefit of such a run will be obvious, but for now it seems that money and effort would be better spent on a single mask set dedicated run followed by an exploration of deeper sub-micron processes. This would allow us to do research in 3D CAMs at a reduced cost and then, knowing what we can expect from 3D, really push the pattern density with smaller feature sized CMOS processes.

5. Integration of VIPRAM, FPGA and RAMs (to be done)

As mentioned earlier, as the AM bank density and size increases, the number of fired roads will increase at high luminosity due to higher occupancy. The direct consequence of this is that more full resolution detector hits associated with roads have to be retrieved and transferred from the AM stage to the track fitting stage, which demands much higher bandwidth between the two. Since the track fitting stage typically has to be done on a separate module from the pattern matching stage, this would pose a significant design challenge at both the board and system level (for details, see FTK proposal [1]). It is highly desirable for the AM stage and track fitting stage to be implemented in such a way that the two are very close to each other, or preferably within the same chip. The track fitting stage can be viewed as the second stage of pattern recognition using full resolution information. The resulting integrated chip would be much more powerful for fast pattern recognition. In addition, both the board and system level design would be significantly simplified if this level of integration can be achieved. This is the motivation behind the integration of a bank of memory, a computational FPGA and a VIPRAM into a single package or even a single chip if possible.

While industry has been pushing hard to develop technologies that can achieve this type of integration, and there are quite a few promising techniques already developed, at this point we do not yet know which one would be the best choice. One example is silicon interposer approach developed for advanced packaging (see Figure). The details of the integration will be worked out in the future, with considerations at chip level, board level as well as at system level. A detailed cost analysis at both chip and system level will be needed. This will be one of the main subjects of the R&D. The actual integration design work is for Phase II.

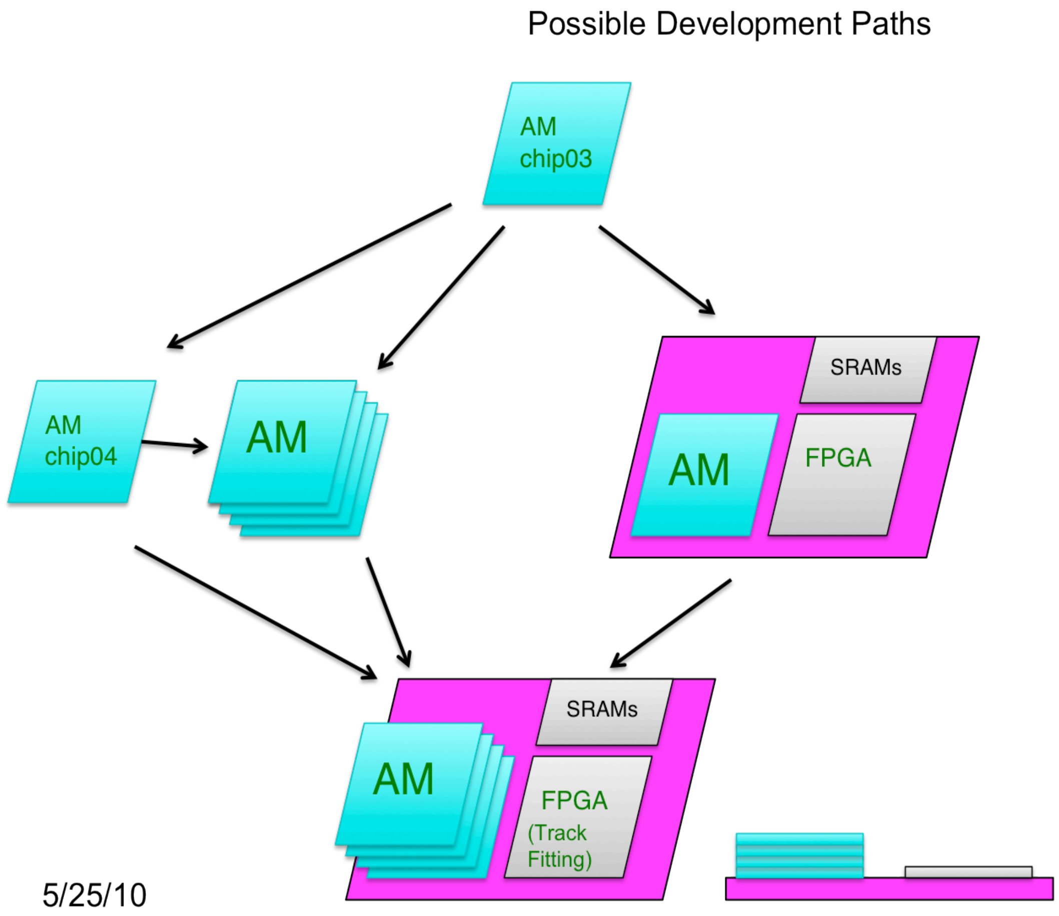


Figure 21. 3D AM and FPGA integration.

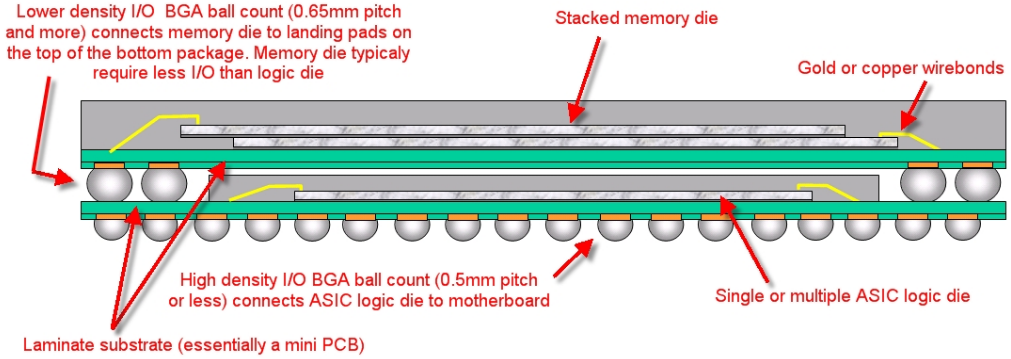
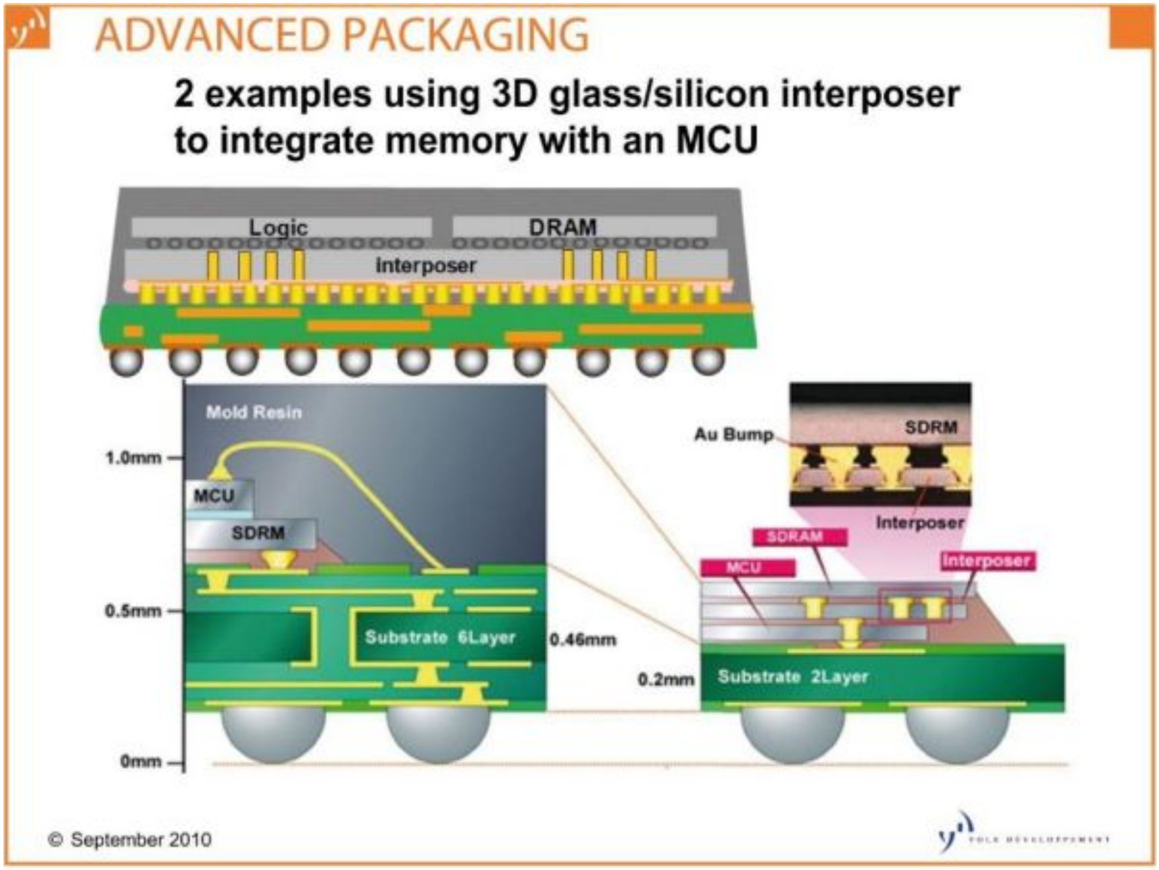

Lower density I/O BGA ball count (0.65mm pitch and more) connects memory die to landing pads on the top of the bottom package. Memory die typicaly require less I/O than logic die
Stacked memory die
Gold or copper wirebonds
High density I/O BGA ball count (0.5mm pitch or less) connects ASIC logic die to motherboard
Single or multiple ASIC logic die
Laminate substrate (essentially a mini PCB)

Figure 22. System in package integration (from PoP wiki)



ADVANCED PACKAGING

2 examples using 3D glass/silicon interposer to integrate memory with an MCU

© September 2010

6. Project Organization, Prototyping Plan and Deliverables

6.1. Overview

6.2. 3D AMchip Stacking & VIPRAM prototyping

To be discussed with collaborators soon. Will need to come up with a coherent near term and long term R&D plan, for R&D in 2D, possible 3D stacking of identical tier design, and true 3D VIPRAM design…

6.2.1. Prototyping, Simulation and Testing Plan

6.3. 3D Integration of AM, FPGA and RAMs

As for AM + TF (or AM + TSP) integration prototyping, one could first start with single AMchip03, and integrate with one FPGA and RAMs … the first thing to do could be to actually design a normal PCB for the integration, by selecting the right FPGA and RAMs. This board could be used for develop testing capabilities for the prototype integration chip.

6.4. Deliverables (over three years)

6.5. Share of Responsibilities

6.6. Cost Estimate

To be discussed with all collaborators…

7. References (check them! … will add more references here).

[1] "FTK: A Hardware Track Finder for the ATLAS Trigger," Technical Proposal, 2010.

[2] L. Sartori, "Current Amchip R&D: mini@sic," *AM Mini-Workshop* presentation, 2010.

[3] A. Annovi, M. Bettini, M. Bucciantonio, P. Catastini, F. Crescioli, M. Dell'Orso, P. Giannetti, D. Lucchesi, M. Nicoletto, M. Piendibene, G. Volpi, "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.

[4] A. Bardi, M. Dell'Orso, S. Galeotti, P. Giannetti, G. Iannaccone, E. Meschia, F. Spinella, "The tree search processor for real-time track finding," *Nuclear Science Symposium, 1998. Conference Record. 1998 IEEE* , vol.2, no., pp.969-973 vol.2, 1998.

[5] 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.

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

[7] A. Annovi, A. Bardi, M. Bitossi, S. Chiozzi, C. Damiani, M. Dell'Orso, P. Gianneti, P. Giovacchini, G. Marchiori, I. Pedron, M. Piendibene, L. Sartori, F. Schifano, F. Spinella, S. Torre, and R. Tripiccione, "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] W.R. Davis, J .Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A.M. Sule, M. Steer, and P.D. Franzon, "Demystifying 3D ICs: The Pros and Cons of Going Vertical," *Proceeding in the IEEE Design and Test of Computers*, vol. 22, no. 6, 2005..

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

[10] K. E. Grosspietsch, "Associative Processors and Memories: A Survey," *IEEE Micro*, vol. 12, no. 3, pp. 12-19, 1992.

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

[12] G. Kasai, Y. Takarabe, K. Furumi, and M. Yoneda, "200 MHz/200MSPS 3.2W at 1.5V Vdd, 9.4 Mbits Ternary CAM with New Charge Injection Match Detect Circuits and Bank Selection Scheme," *Proceedings of IEEE Custom Integrated Circuits Conference*, pp. 387-390, 2003.

[13] M.M. Khellah, M. Elmasry, "Use of Charge Sharing to Reduce Energy Consumption in Wide Fan-in Gates," *Proceeding in the IEEE International Symposium of Circuits and Systems*, vol.2, pp. 9-12, 1998.

[14] C.A. Zukowski, and S. Y. Wang, "Use of Selective Precharge for Low-Power Content-Addressable Memories," Proceeding in the IEEE International Symposium of Circuits and Systems, vol.3, pp. 1788-1791, 1997.

[15] I. Arsovski and A. Sheikholeslami, "A Current-Saving Match-Line Sensing Scheme for Content-Addressable Memories," Proceeding in the IEEE International Solid-State Circuits, pp.304-305, 2003.

[16] K. J. Schultz, and P. G. Gulak, “Architectures for Large-Capacity CAMs,” IEEE Transaction in VLSI, vol. 8, no. 2-3, pp.151-157, 1995.
(to add: Bob Patti’s review paper from 3D Handbook, original TSP paper … more).

8. Appendix

8.1. Introduction to 3D Technology and Process Involved for VIPRAM

8.1.1. Overview

A 3D chip is generally referred to as a chip comprised of 2 or more layers of active semiconductor devices that have been thinned, bonded, and interconnected to form a “monolithic” circuit. These layers, or tiers, can be fabricated in different processes. Performance is improved by reducing interconnect resistance, inductance, and capacitance for higher speed, by reducing I/O pad count, reducing the interconnect power and crosstalk, and by reducing the circuit form factor. 3D integrated circuits can have increased circuit density due to multiple tiers and at the same time be very thin since the individual layers can be very thin. Moreover, the technology provides the freedom to divide functionality among tiers to create new designs that simply are not possible in 2D [8].

There are four key technologies needed for 3D circuit integration: bonding between layers, wafer thinning, through wafer via formation and metallization, and high precision alignment. The latter is an integral part of the bonding of the two wafers carried out at the foundry or the bonding facility and will not be further discussed here. The other three aspects are of crucial importance.

To establish an electrical connection between two or more tiers and to enable connections to external bond pads a via needs to be formed through the silicon wafer and metalized, a so-called Through-Silicon Via (TSV). There are two main strategies for the via formation, “via first” and “via last”. The “first” and “last” makes reference to the chronology of semiconductor and TSV formation. “Via-first” means the TSV is formed before the semiconductors are formed on the silicon, and conversely, “via-last” means TSV formation after semiconductor formation. The 3D technology is based on through-silicon vias (TSVs) in wafers, in such a way that every integrated circuit can be considered as a two-sided device, where connections can be made to the top or bottom or both sides of a chip. This is the definition of 3D integrated circuits, with multiple levels of transistors and increased routing levels.

8.1.2. Face-to-back versus Face-to-face Stacking

There are different strategies for the bonding of the various tiers. Normal 2D Integrated Circuits are created by a series of crystal growth, ion implantation, and metallization steps that form the desired circuits in the top few microns of a semiconductor wafer that is typically several hundred microns thick. The top few microns that contain the circuitry is called the "face" and the opposite side that contains nothing but semiconductor wafer is called the "back", as shown in Figure 14.

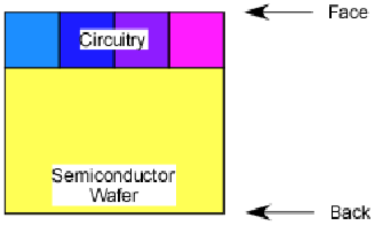


Figure 14. Illustration of a single tier.

When stacking normal 2D integrated circuits to form 3D integrated circuits, they can be stacked "face-to-face" or "face-to-back". "Back-to-back" is also possible, but for various technical reasons, it is not used as often. Face-to-Back and Face-to-Face stacking places different requirements on the designers.

Face-to-Back requires TSVs for the tier-to-tier interconnects. This means that by one of several methods, a hole is made through the silicon wafer and a via is placed in that hole to give access to the circuitry on the face-side from the back-side. Then tier-to-tier interconnections are made by stacking tiers on top of another by connecting the face side of one chip to the back-side of the other. Note that, as far as the 2D design is concerned, both tiers are oriented in the same way.

Face-to-Face stacking, on the other hand, makes the tier-to-tier interconnection in the upper metal layers of each tier. No TSV is necessary for the tier-to-tier interconnect, though they will be necessary for connections to pads on the back and to the outside world. Unlike face-to-back stacking, face-to-face stacking requires that one of the layers be deliberately flipped in its design so that when it is flipped in its 3D fabrication, the appropriate tier-to-tier interconnections align with one another as shown in Figure 15.

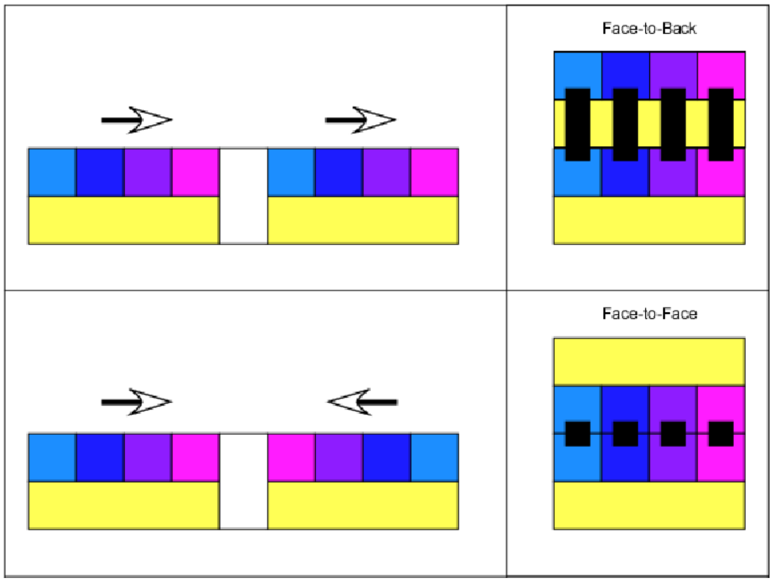


Figure 15. Face-to-Face and Face-to-Back bonding.

Face-to-Face Stacking of two tiers is comparatively straightforward. Unfortunately, Face-to-Back Stacking of two tiers is not straightforward. Typically, Face-to-Back Stacking is reserved for stacking subsequent tiers onto two tiers already stacked by Face-to-Face stacking. The reason for this has to do with the manner in which stacking is done.

As just stated, normal 2D Integrated Circuits are created by a series of crystal growth, ion implantation, and metallization steps that form the desired circuits on the top few microns of a semiconductor wafer that is typically several hundred microns thick. From the perspective of 2D circuit design, the several hundred microns of unused semiconductor wafer under the circuit is just substrate. The substrate is a source of parasitic electrical connection between different circuit elements. Other than that it serves little purpose and is frequently ignored, especially by digital designers. In 3D design, however, the substrate is the mechanical support structure for the circuit. It is to be used when needed and removed when no longer necessary.

To make a Face-to-Face Stack, Tier 1 is held by its substrate. This is the “handle wafer”. Tier 2 is inverted and the Face-to-Face Stack is made.

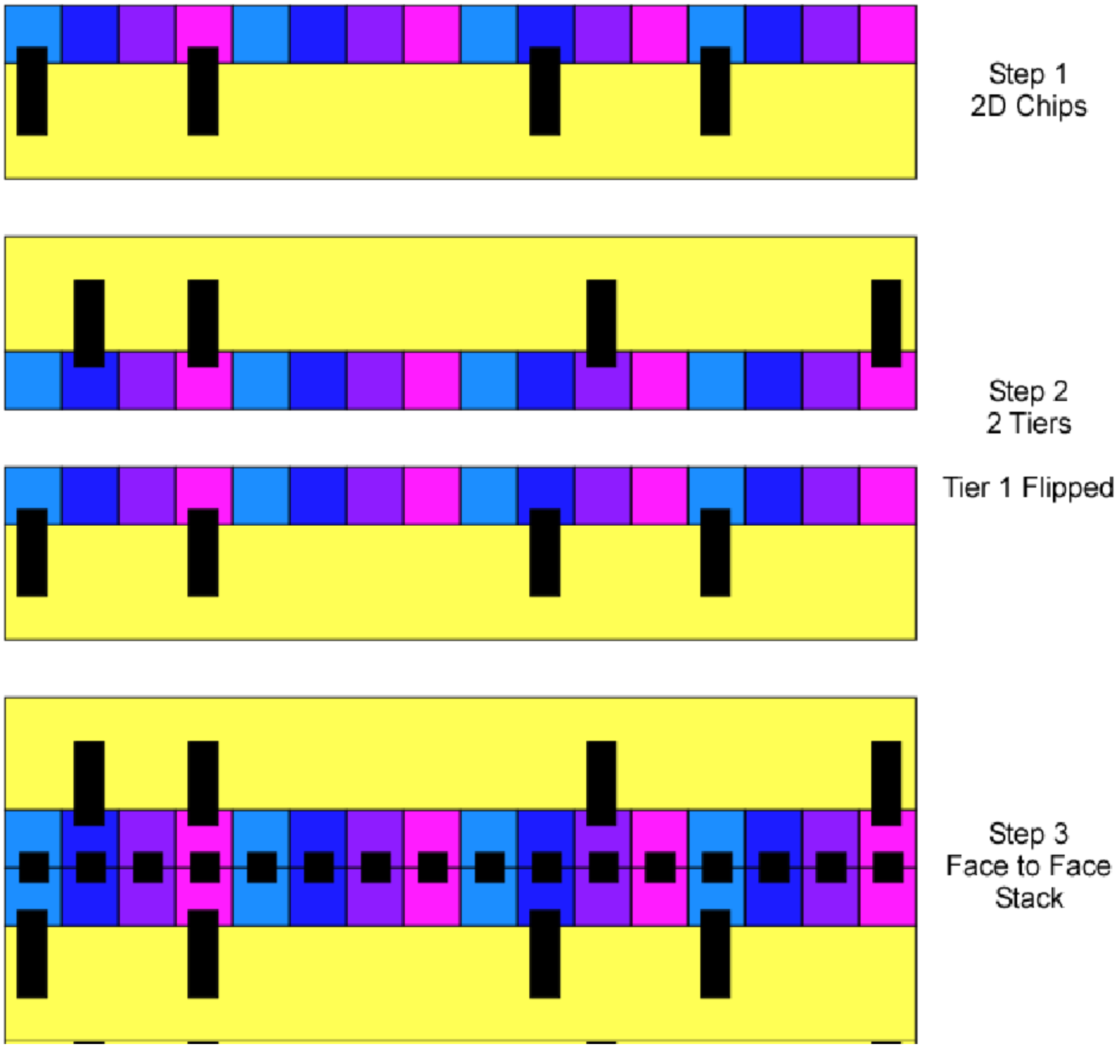


Figure 16. Steps for Face-to-Face stacking.

To add a third tier, the face-to-face stack of Tier 1 and Tier 2 is used as the handle wafer for the Face-to-Back Stacking of Tier 3. The face of Tier 3 is stacked onto the back of the tier 1-2 face-to-face stack. To accomplish this, first, the unnecessary substrate of tier 2 is thinned until the TSVs are exposed. Tier 3 is then aligned and connected to tiers 1-2 as shown in Figure 16 and 17.

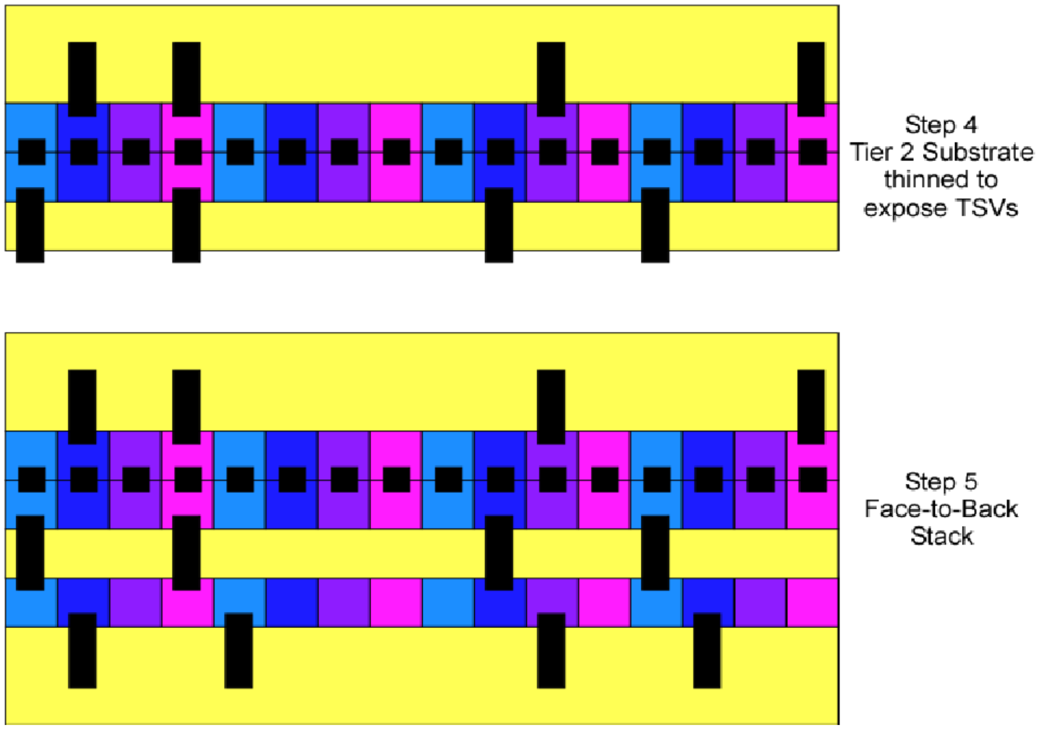


Figure 17. Steps for Face-to-Back stacking.

The reason that simple Face-to-Back stacking is so difficult is that there is no obvious handle wafer for Tier 1. This becomes easier to visualize when you realize that these figures are not drawn to scale. The circuitry occupies only a few microns on the top of each tier. The substrates are hundreds of microns thick. In order to perform a single Face-to-Back stacking, Tier 1 would have to be thinned to expose its TSVs. That would leave a thin sheet of silicon and oxide and metal with no mechanical strength or stability. That sheet would somehow have to be attached with sub-micron accuracy to the Tier 2 wafer. Face-to-Face followed by Face-to-Back works because the substrate of tier 1 is never thinned. It is the handle wafer for the first Face-to-Face stacking and for the subsequent Face-to-Back stacking.

In this proposal, to stack the identical AM tiers, the current working assumption is that the face-to-back process is a viable approach. To solve the problem of the face-to-back stack process, one could use a dummy handle wafer to initially bond with the first AM tier in a “face-to-face” fashion, then followed by the face-to-back bonding with the other AM tiers. After the stack is finished, the handle wafer can be removed or even stay with the stack. As for the true 3D approach with Control and CAM tiers, there is no problem because the Control tier and the first CAM tier can be bonded face-to-face, followed by other CAM tiers in a face-to-back configuration.

8.2. Tezzaron ‘s FaStack Technology (from Tezzron’s web site)

(for details, also see Bob Patti's paper "3D Integration at Tezzaron Semiconductor Corporation", Handbook of 3D Integration 2008).

For this proposal, we choose to follow Tezzaron's 3D stacking approach, the FaStack Technology, for its 3D DRAMs to implement the 3D VIPRAM. Tezzaron's FaStack technology creates fast, dense, highly integrated 3D chips. The heart of the process is wafer-level stacking. The device circuitry is divided into sections, which are built onto separate wafers using standard processing. The wafers are then post-processed for through-silicon interconnection, creating hundreds of thousands of vertical "SuperVia" connectors. The wafers are aligned with a precision of 0.5 micron, then bonded, thinned, and diced into individual devices. The FaStack process addresses the thermal stress issues with 3D stacking by ultra-thinning the wafers to prevent thermal buildup, and the copper used in the bonding process provides additional relief. FaStack devices have many advantages over their single-layer counterparts. They are much denser and their short vertical interconnects allow them to operate at higher speeds with lower power budget. In addition, FaStack allows disparate elements to be processed on separate wafers for simpler production and greater optimization.

A semiconductor wafer is usually about 750 microns thick, but its electrical activity is confined to a surface layer from 4-10 microns thick. The functional part of a wafer is thus a tiny proportion of its thickness; the rest of the wafer only provides structural support. The Tezzaron FaStack process uses most of the structural base of the first silicon wafer, but keeps less than 15 microns of each additional wafer in the stack. This produces multi-layer chips that fit easily into standard packaging.

The FaStack process begins by building hundreds of thousands of vertical interconnect structures (Super-Contacts) into the circuitry during normal wafer processing ("via-first"). The wafers are then metalized by coating them with a 0.5 micron SiO2 insulating glass layer and then a 1.0 micron Cu metal bond point layer with a proprietary layout design. Using a thermal diffusion bonding process at less than 400 ^{o}C, two metalized wafers are aligned with their front sides facing one another and then bonded together. The structural base (back side) of the upper wafer is then thinned to less than 10 microns using a combination of conventional wafer grinding, spin-etching, and chemical-mechanical polishing. The thinning exposes the Super-Contacts that were built into the wafer. The back side of the thinned wafer, with its exposed Super-Contacts, can be metalized with bond points and bonded to the front side of a third metalized wafer. Thinning, metalizing, and bonding are repeated as desired. Once the wafer stacking process is completed, one side of the stack is thinned to the Super-Contacts and padded out for I/O; the other side is back-lapped to remove excess silicon.

Tezzaron built the first working 3D IC prototypes (six different devices) in 2004. In 2008, Tezzaron began producing custom stacked components under contract and now provides stacking services for a number of customers. The Super-Contact density can reach 160,000 per mm square (typical designs use ~ 10,000 per mm square). The alignment precision for 200 mm wafers has a 3-sigma process tolerance of +- 1 micron, precision with +- 0.3 micron is typical. Ultra-thinning reduces the wafer thickness to as little as 8 microns, uniform to within +- 0.5 micron. The FaStack's aggressive wafer thinning prevents excess thermal buildup and allows the stack to behave as one thermal unit, and copper bonds facilitate heat dissipation. The FaStack units are thin enough to mount in standard packages. The high-density interconnects take full advantage of short vertical paths. The prototype FaStack 8051 processor runs at either 5X the speed of a normal 8051 or 10% of the power.

Note that Tezzaron's first key breakthrough in 3D development was the "Super-Via", a vertical copper structure that adapts standard process flow wafers to Tezzaron's 3D stacking process. Super-Vias can be post processed into any wafer merely by adding metallization and additional dielectric, so they do not require a full manufacturing line or direct involvement of an outside foundry. Further, the Super-Via interconnect provides alignment marks, thinning control, interconnect, and bonding surfaces in a single structure. It further adds intrinsic cooling capabilities with its vertical copper structures. Since the early development and success with the Super-Via flow, Tezzaron has been developing a new second generation of interconnect. This second generation interconnect does involve the primary wafer foundry, but it adds more design flexibility while drastically decreasing the 3D interconnect footprint. The size of the Super-Via was 4.0 x 4.0 micron in its first incarnation; the second generation is 1.2 x 1.2 micron, while face-to-face bonding has a size of 1.7 x 1.7 micron. Minimal pitch is 6 micron, < 4 micron and 2.4 micron respectively.

FaStack's wafer stacking offers benefits to a variety of applications. Sensor arrays, for example, achieve unprecedented density by moving the support circuitry to a different layer than the sensors themselves. "System-on-Chip" (SoC) devices built with FaStack reduce power consumption, footprint, and interconnect delays. Microprocessors built with FaStack incorporate a huge, fast memory cache on a separate layer. FaStack also enables enormous improvements in memory technology and allows seamless integration of differing substrates. As 3D processing moves into the mainstream, entirely new products will merge to capitalize on this technology.

8.3. Bob Patti's "Diagonal Vias" idea

Bob Patti of Tezzaron had dealt with the Tier Self ID problem long time ago with his 3D memory design and he has come up with a clever and simple solution that exploits 3D advantages and uses no extra transistors. This idea has been used extensively for Tezzaron's 3D DRAM stacking with the Control + DRAM tiers design, and we plan to follow the same approach for the 3D VIPRAM design for vertical communications between the Control tier and the CAM tiers. This solution is called the "Diagonal Via" and was patented about 10 years ago (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).

Figure 26 shows a mock-up of a diagonal via showing pads to a tier above and pads to a tier below. In face-to-back bonding, the pads to a tier above would be the upper metal layer bonding interface and the pads to a tier below would be through silicon vias (TSVs). Of course, this is not a real layout. The red and yellow lines, in reality, would be comprised of vertical metal-metal vias and horizontal metal traces. For example, the leftmost red trace might be a via from metal1 to metal2 followed by a horizontal run of metal2 followed by vias from metal2 to metal6 or whatever metal is the bond interface. The yellow trace might be vias from metal1 to metal4 followed by a horizontal run of metal4 followed by vias from metal4 to metal6 or whatever metal is the bond interface. Moreover, in 3D layout, it is possible to stretch the wires into and out of the page. In short, the diagonal via is a compact method for routing signals from a tier above to a tier below or from a tier below to a tie above. In each case, the signals are shuffled one pad to the right and the rightmost pad is routed back to the leftmost pad as shown in Figure 27.

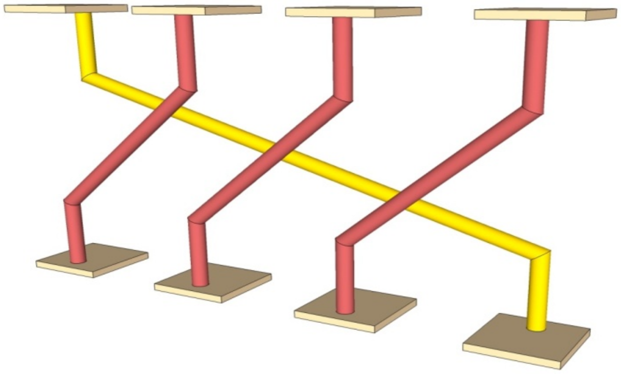


Figure 26. Diagonal vias.

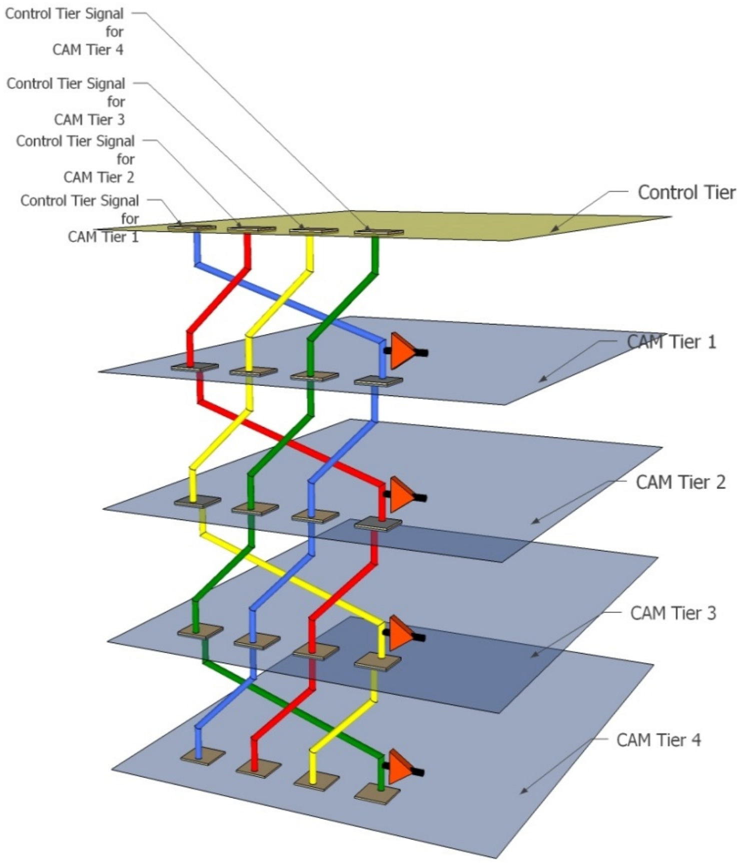
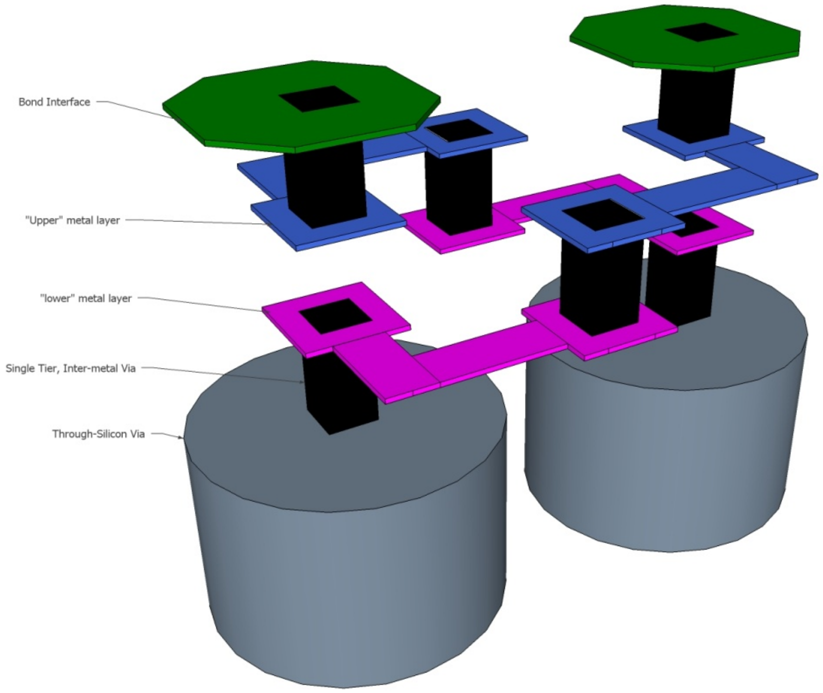


Figure 27. Diagonal vias in multiple tiers.

In this case, the Control Tier is sending data layer/tier specific data to each tier. This same structure works with drivers on each tier and with each tier sending layer/tier specific data to the Control Tier. In a structure with one Control Tier and four CAM tiers, the Control tier sees four vias, one for each CAM tier. The CAM tiers all have exactly the same layout, as is required. The leftmost via on the Control tier is for CAM tier 1. It is obvious that the blue diagonal via takes the Control Tier information and passes it down to the receiver on CAM tier 1. Note that the blue route continues down through CAM tiers 2-4, but it does not ever arrive again at a receiver. Only CAM tier 1 receives the information dedicated to CAM tier 1. Similarly, the rightmost via on the Control tier is dedicated to CAM tier 4. This is the green route in the figure. The signal begins by passing to the left on CAM tiers 1-3, but then it passes to the right on CAM tier 4 and arrives at CAM tier 4's receiver. Again, the only receiver to get this data is the receiver on CAM tier 4.

Diagonal Vias in Greater Detail

Strictly for the curious, Diagonal Vias are not a new technology. They represent no increased fabrication risk. It is simply a clever idea. The following diagram illustrates one possible routing scheme for a two via diagonal via. All of the geometry shown is straightforward, routine VLSI and all of it is drawn on one tier.



On the bottom are two cylindrical through-silicon vias. They are connected to lower metal layers by simple, old-fashioned inter-metal vias. In this picture, the lower metal layer is, in fact, metal1, the lowest metal layer. This is obvious because the through-silicon via to metal via is one simple cube (the black cube that connects the grey through silicon via to the magenta metal). In truth, this lower metal layer could be metal2 or metal3 or almost any metal layer. The only difference would be that the lowest black via in the picture would be slightly more complicated. The lower metal layer routes the signal away from the TSV and brings it up to a higher metal layer via another standard VLSI metal-to-metal via. This second metal layer routes the signal around and uses a third standard VLSI metal-to-metal via to connect the signal to the bond interface.

Following the signals through the diagram, it is obvious that the signals move diagonally in the vertical direction even though the geometries are standard, run-of-the-mill two dimensional VLSI.

8.4. Possible 3D Pad Structures for Tier-ID

One of the problems with the identical tier design with the same mask is the need of Tier self ID. One possible solution developed earlier for this is to use a “Self ID” and then Tier-gated IO, using special 3D Pad Structures. This possible approach is potentially useful in certain cases (could be also combined with the “diagonal vias” approach above). Note that in the case of Control + CAM tier design for the true 3D VIPRAM, there is no need for the special pad structures because the “diagonal via” trick is good enough. In the case of identical tier design, it is possible to only use diagonal via trick with external bias. In any case, the idea of 3D pad structures is still potentially useful and we describe the basic idea here.

3D Pad Structures

Given that different tiers connect to different metals, a simple pad in 2D VLSI can become something more in 3D VLSI. In 3D VLSI, the signal that enters one side of a pad does not have to be the signal that comes out of the other side. A simple 2D VLSI pad is shown below. The blue, horizontal rectangles are metals in the VLSI process and the black, vertical rectangles are metal-to-metal vias.

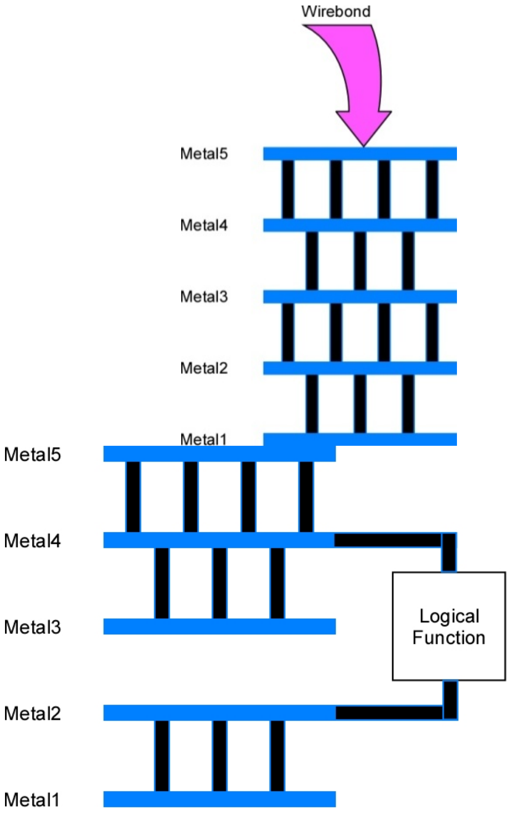


Figure 24. (a) standard 2D pad and (b) a 2D pad with logical function between metal layers.

Note that an opening is made in the oxide at the top of Metal5 (the highest level metal in this hypothetical process) above the pad and any signal on the wire bond is ohmically connected to all of the metal layers below the pad. On-chip wires can be brought in from any metal layer to connect this pad to the internal circuits of the chip. It is also possible to do as in Figure 24b.

Such a pad structure makes little sense in a 2D design. In 3D, however, there can be an “upward-looking” pad above Metal5 and a “downward-looking” pad below Metal1 and this leads to a whole array of novel structures.

3D Parallel Pad

In this pad structure, the logical function in the figure is just a short, and the pads in the same location on each tier are ohmically connected. This pad structure looks like a normal 2D pad structure that simply penetrates all the way through the 3D IC.

Tier-gated Input

In this pad structure, a signal external to the pad dictates whether or not this pad input signal will be made available to this particular tier. With this pad structure the signal on the wire bond can be directed to one or more particular tiers.

Tier-gated Output

In this pad structure, a signal external to the pad dictates whether or not a signal from this tier will be driven off the chip. If not, then the pad’s driving signal from this tier is tri-stated. With this pad structure multiple tiers can have controlled access to the same pads.

3D One-Way IO

This pad structure comes in two flavors. In the first flavor, the upward looking pad is an input pad and the downward looking pad is an output pad. In the second flavor, the upward looking pad is an output pad and the downward looking pad is an input pad.

3D Modified IO

This pad structure is very similar to the 3D One-Way IO structure, but the output is not the same as the input.

Using the 3D Pad structures, it is possible to implement tier self-ID. However, it does require extra transistors and routing.

8.5. Survey of CAM Power Saving Techniques

In this section, we provide a survey of power saving techniques that have been widely and successfully applied in industry for CAM architectures. These techniques can be applied to the 2D designs, (AMchip03 and AmchiP04) and multi-tier chips. Match-line and search-line contribute the majority of power consumption in AM chips, so reducing their power consumption is necessary for obtaining high-density pattern bank.

CAM architecture allows full search within a clock cycle with large power consumption overhead. In order to reduce the power consumption on the current designed AM chips, we provide descriptions of various techniques which can be applied to match-line, search-line and architecture level.

Low-Swing Scheme

This method reduces the ML power consumption and potentially increasing its speed by reducing its voltage swing. The reduction of power consumption is linearly proportional to the reduction of voltage swing. Such technique has been applied in [12,13] where the ML swing is reduced from the full supply swing to 300mV. The challenge with such method, that it may require another voltage supply to generate the low-swing voltage.

Selective-Precharge

The match-line spends the same amount of power regardless of the specific data pattern and whether it is a match or miss. Selective precharging technique performs a match operation on the first few bits of a word before activating the rest of the match-line for the remaining bits [14]. Such method can result into significant power savings on the case when there is miss, as the power is reduced to the rest of the match-line. There are two overheads with this technique, (1) the power drawn by the initial pre-charged bits could be higher than the rest of the bits, and (2) the application could have initial matching bits the same among all the words in AM, which eliminates any power savings.

Search-Line Precharge

The power consumed by search-lines can be eliminated by reducing its toggling thus reducing its power. Match-line precharging schemes eliminate the need to precharge search-lines. This technique directly activates the search-lines with their search data without going through a search-line precharge. Since in typical case, the search-line toggles about 50%, this technique can result up to 50% of SL power savings [15].

Bank Selection Scheme

This is an architectural technique for saving power [16]. Power reduction is achieved by bank selection, where only a subset of the AM chip is active on any given cycle. The pattern bank can be categorized into four banks, and when searching data, the bank select bit determines which one of the four banks to activate and search. Since only one of four banks is active at any given time, only ¼ of the comparison circuitry is needed compared to the case where no bank selection is applied and resulting up to 75% of power savings. This technique saves both power and chip area.

8.6. Back-of-envelope Estimate on possible gain in density for VIPRAM

It is worth the effort at this stage to do some very rough back-of-envelope estimate on how many patterns one could possibly expect with this architecture, from purely geometrical (area) point of view, just to get some sense of the scale. Exactly how much one can actually achieve is the subject of the R&D.

A VIPRAM chip with 40 times more patterns than Amchip03 would contain 40 x 5K = 200 K patterns or vertical "blue tubes" (as shown in Figure 13). Assuming a die size of 1 cm x 1cm for a given tier (note that Control tier can be somewhat larger in size), each of the vertical blue tube would, on average, have a footprint in 2D of the size of 22.4 um x 22.4 um ~ 500 um square. This area will need to contain both the diagonal vias and the 15 bits CAM unit. Assume there will be 6 CAM tiers (for 6 detector layers as in AMchip03), so there will be six diagonal vias for each match line on each CAM tier for each blue tube. Note that only one match line will be needed for a CAM tier cell within a given blue tube. With current Tezzaron's FaStack technology, six diagonal vias could occupy about ~ 5 um x15 um ~ 75 um square area, leaving 425 um square area to implement the 15 bits CAM unit. Using the numbers achieved and presented by Matteo at the first AM R&D Mini-workshop in April, 2010, a 15 bits CAM unit would occupy 67.2 um x 2.8 um = 188 um square with full custom design in 90nm. Scale this number up to 130 nm, a 15 bits CAM unit would occupy 380 um square, which is less than 425 um square available. This means that it is possible, from geometrical point of view, for VIPRAM to achieve 40 X AMchip03 patterns using only 130 nm technology with full custom CAM cells for the CAM tiers (or standard cell in 90 nm). This suggests that for initial prototyping, perhaps it is reasonable to already aim for 20-40 x AMchip03 using only130 nm.

When scale down to 65 nm, a 15 bits custom designed CAM unit would occupy only 94 um square. Assuming the diagonal via size stays the same (even though it should become smaller with time), the combined area is about 170 um square. For 1cm x 1cm silicon, this means that the total number of possible patterns (geometrical ratio) could be up to 590K, or a factor of 117 x AMchip03.

Note that for more detector layers, one can add more identical CAM tiers (with slightly increased overhead of diagonal vias) for the VIPRAM design without reducing number of patterns. This is different from the 2D design of AMchip, where the number of patterns will need to be reduced when adding more detector layers. Just like AMchip03, it should be possible to configure a 6 layer VIPRAM to handle 12 detector layers.

The one true disadvantage of the true 3D VIPRAM approach is the need of two mask sets. On the other hand, Tezzaron has chosen the same approach for their 3D DRAM design.

9. Power and Thermal Modeling and Analysis

9.1. Introduction

The advantages for 3D integration are associated with challenges. From a manufacturing perspective, yield is an issue due to the extra processing step for stacking. From the performance and functional perspective, the main challenges are power and thermal issues. Due the increased density of 3D integration, the on-chip temperature profile can increases considerably. Vertical tiers and vias create temperature fluctuations and power noise which can impact chip functionality. Additionally, due to numerous manufacturing steps and increased design complexity, successful 3D integration can result into long and costly design cycles. Thus, it is essential to address power and thermal issues early on in the design cycle before they become an obstacle.

The cost of 3D technology development can be high because of the 3D process steps for vertical stacking. The microelectronics industry, therefore is actively investing effort in addressing the power and thermal challenges with 3D integration to ensure the proper operation of the chips without costly re-spins. Academia has also seized this opportunity to develop the theoretical perception of the power and thermal issues even in 2D technologies. In fact, one of our collaborators has done power and thermal modeling and analysis work in her Ph.D. thesis work ("Power delivery for nanometer technology chips", Ph.D Thesis, Aida Todri, UCSB, 2009. Section 0035, Part0544 168 pages; Publication Number: AAT 3379525). She developed circuit modeling, analysis and power management techniques to alleviate the power-hungry demands of fast switching microprocessors. Given both the academic and industrial advancement for analysis tools, these tools are limited for being applicable to a specific vendor, technology process, oriented for consumer-electronic applications and partial treatment of simultaneous power and thermal issues for 3D integration. We believe that now is the time for HEP community to invest and develop our own in-house tools for addressing the power and thermal issues for HEP applications, and the 3D AM R&D could be the perfect starting point. Since it will take time and real effort to do the modeling accurately, realistically, we do not expect that the 3D AM R&D will critically depend on the power and thermal analysis at the early stage. We do expect that we will benefit from Tezzaron's experiences with its FaStack technology on power and thermal management issues, as well as the 2D AMchip04 R&D effort on reducing the power consumption. The hope is, however, to develop a power and thermal management methodology through the 3D AM R&D (as a by-product) to be used for other 3D design in the future.

Transistors are the basis for chip design. Performing multi-tier chip analysis at transistor level is simply not feasible due to the large number of nodes that cannot be handled by the simulator. We propose to develop circuit models in a bottom-up hierarchical approach. Models will be representative of the actual circuit without capturing all the implementation details. The cell (smallest functional circuit composed of several transistors) level will serve as the basic building block in the modeling scheme. A cell model will capture the impact of power and thermal issues on the cell's functionality and performance. Furthermore, several cells grouped together are utilized to build a larger circuit model. This hierarchical modeling approach continues till we have a complete model for the chip based on the cell models. Accurate model development is essential in capturing the global impact of power consumption and temperature fluctuations in the multi-tier design.

Power and Thermal Modeling for 2D and 3D

We wish to model CAM cells and tiers to analyze and understand the power and thermal behavior of the CAM design in 2D and 3D integrations. These models would serve as the basic building blocks for performing power and thermal analysis, which can guide the designer to modify the system to meet power and thermal constraints.

Given that AMchip03 is designed and tested, we would utilize this design as the starting point for developing our models. To facilitate the analysis of the design, a modeling approach is proposed to represent the underlying design in a hierarchical fashion. The CAM architecture has a regular architecture where the basic building block is the CAM cell. Our hierarchical approach on modeling is a bottom-up approach, where we start by modeling the lower level cells which can be further utilized to represent a larger design. In the following subsections, we describe hierarchical approach by describing the CAM and multi-tier modeling.

CAM Modeling Approach

The CAM cell model is based on the transistor-level description of the NOR/NAND-type cells which ensures their accurate representation. SPICE simulations of the transistor-level circuit model provide an in-depth understanding of the power, voltage and current consumption of the CAM cell. The circuit model can be further represented with respect to several variables such as, supply voltage (Vdd), input transitions, (search-line transitions), stored pattern (to emulate the case of hit and miss state). Varying the values of these variables, we can analyze all the operating conditions of the CAM cell. Such analysis is referred as *corner analysis* in order to capture the cell behavior for all corner cases. Performing a corner analysis at the CAM cell level provides insights to the power consumption and performance of the CAM architecture. The model accuracy will be refined based on the simulation results and measurements from AMchip03. AMchip03 will serve as a vehicle for developing and refining the models.

Multi-tier Modeling Approach

The current AMchip03 has 5120 patterns, corresponding to approximately 500 000 content-addressable memory bits (CAM cells) [7]. The large granularity of the system poses a significant challenge in performing multi-tier chip level power, thermal and performance analysis. Performing SPICE simulation on the entire CAM architecture (millions of CAM cells) is simply not feasible due to the large number of nodes on the circuit which cannot be handled from the simulators. Hence, simple but accurate models are a necessity for performing large scale simulations.

To overcome these challenges, we propose to utilize the simplified CAM level model as described in the previous subsection to build a *macro model* to represent a pattern. A pattern consists of multiple memory bits, i.e. six 16 bit words in AM chip. The pattern macro models would be utilized to represent the *tier model* for the CAM design at a tier level. We also propose to develop TSV models to represent their parasitics impedance and capacitance among the tiers. These parasitics values will be extracted based on the dimension and technology information of the TSVs. Utilizing the hierarchical approach, the *multi-tier model* can be constructed by using *TSV* and *tier models* to represent the 3D design. Our modeling approach will be based on the hierarchical modeling which is illustrated in Figure 23.

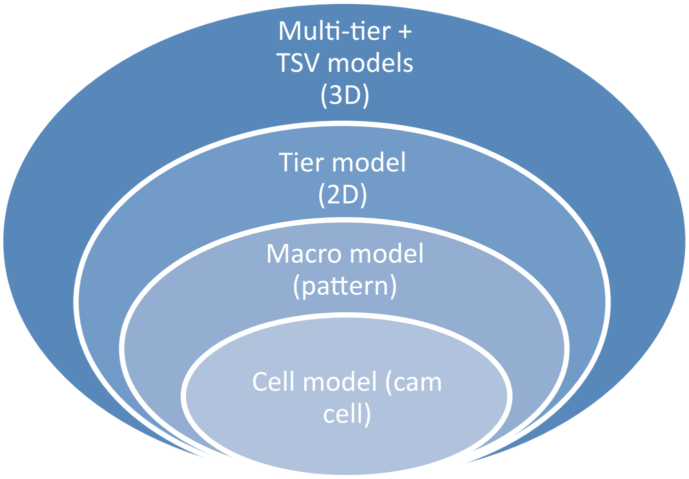


Figure 23. Our modeling approach.

Power and Thermal Analysis for 2D and 3D

We utilize power and thermal analysis methodologies which are widely and successfully applied in the chip design industry. We plan to utilize our models (as described in the previous section) to perform a thorough power and thermal analysis of the design. The objective of such analysis is to identify early on the potential issues of large power consumption and creation of hot spots due to thermal variations.

We propose to perform the power analysis to capture the impact that design choices (i.e. TSV location, TSV dimensions, tier thickness, design style of the glue logic among the tiers) pose on the overall power consumption of the chip. The purpose of such analysis is to provide guidance throughout the design process. Similarly, we propose to perform thermal analysis to estimate the thermal variations inside a tier, TSVs and intra-tiers. Due to the vertical integration, the inner tiers will experience higher thermal variations due to limited thermal dissipation. Thus, by performing such analysis, we expect to identify the thermal issues and find solutions to alleviate them. We plan to utilize the thermal-electrical duality principle to develop thermal models for the analysis.

In this work, we aim to provide the modeling and analysis technique required to perform power and thermal analysis for both 2D and 3D design that can also be utilized for other ASIC design projects.