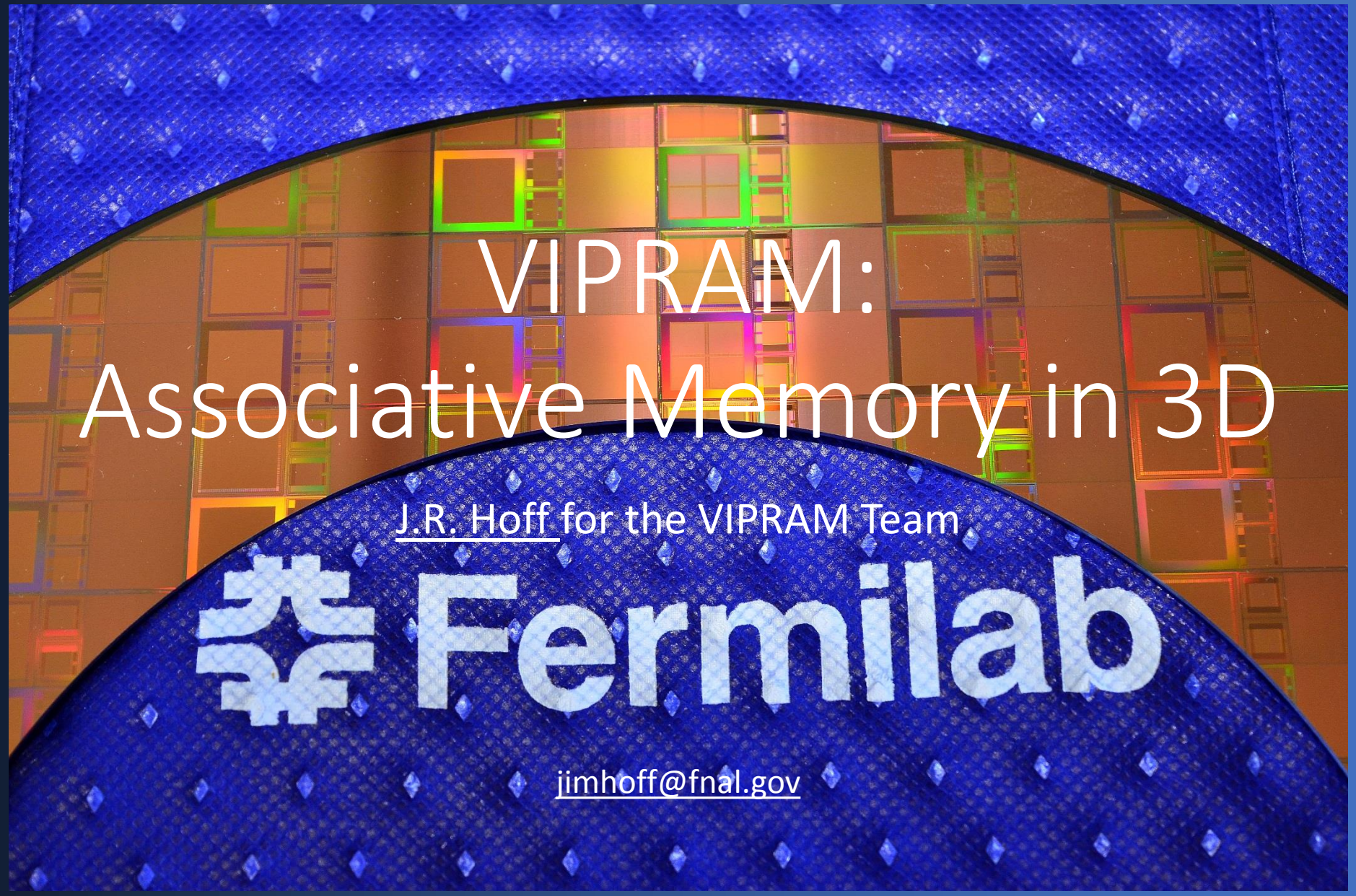
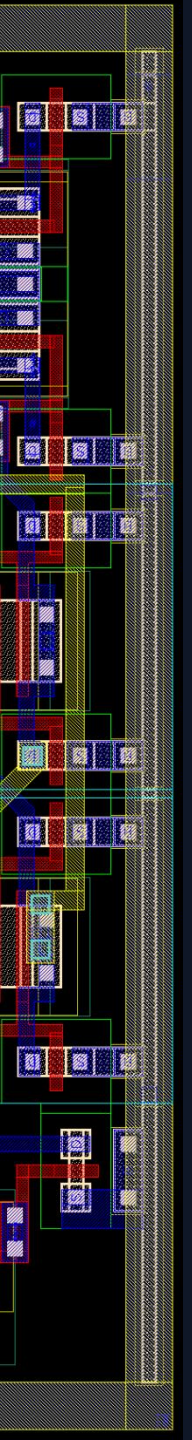


VIPRAM: Associative Memory in 3D

J.R. Hoff for the VIPRAM Team

jimhoff@fnal.gov

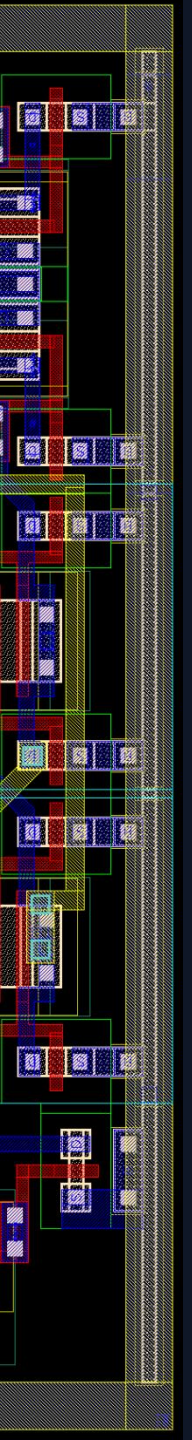


What is Pattern Recognition for Track Finding?

Honestly…it's Friday.

If you don't know the answer to this question by now, you might be in the wrong room…

Besides…I'm an engineer in a room full of physicists. We do not necessarily see the same thing. (Take a look at Wednesday afternoon's talks or speak to Ted or Sergo.)



What is an Associative Memory?

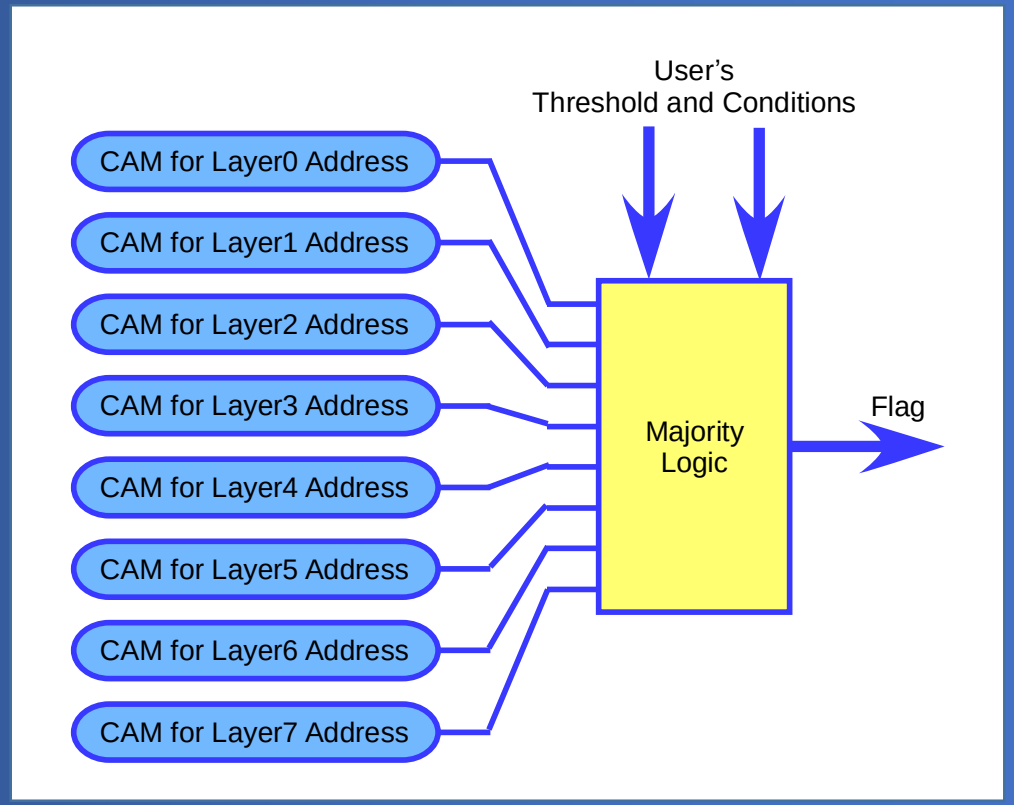
A memory element (i.e. sequential logic) capable of determining whether or not a given piece of data is contained within one of its internal locations.

Technically an Associative Memory can respond to a range of matches. A CAM or Content Addressable Memory responds to a "perfect" match. Therefore, technically a CAM is an Associative Memory, but a given Associative Memory is not necessarily a CAM.

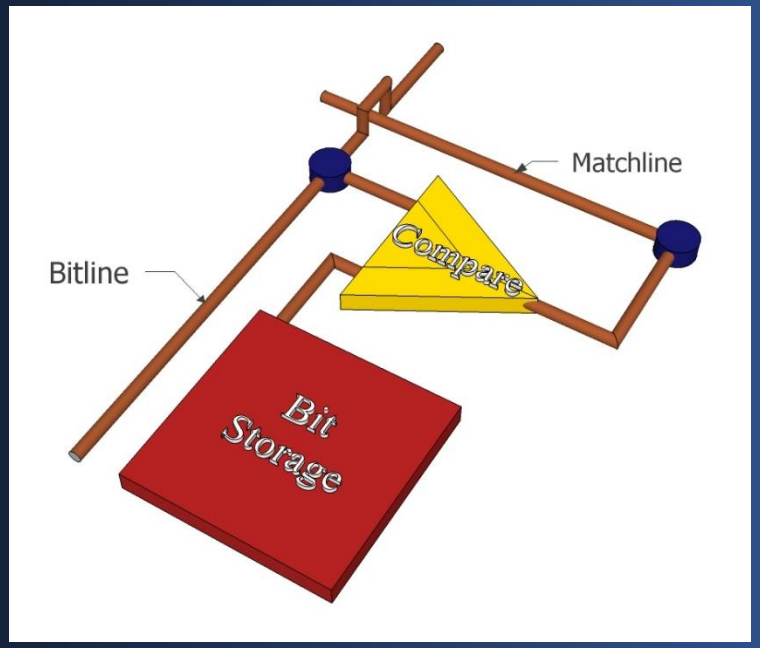
Our Associative Memory Architecture

What is the underlying architecture?

A two-tiered associative memory structure *with pattern match memory* and composed of CAMs and the ability to monitor them. It searches for “identical” matches within each individual CAM and then flags a road match when a user-definable number of CAMs have matched.



Key Components of a CAM..

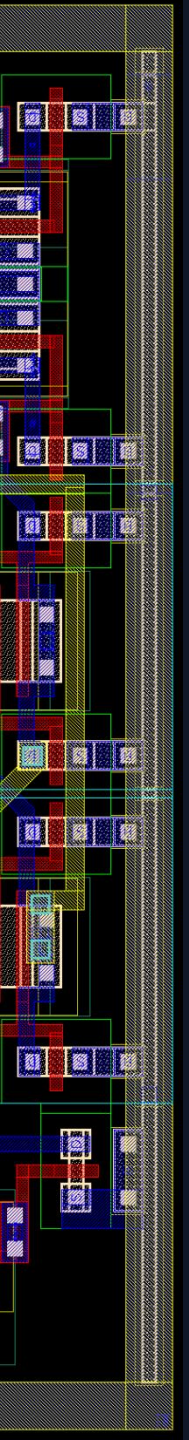


It is a classic storage element (SRAM) paired with a Comparator.

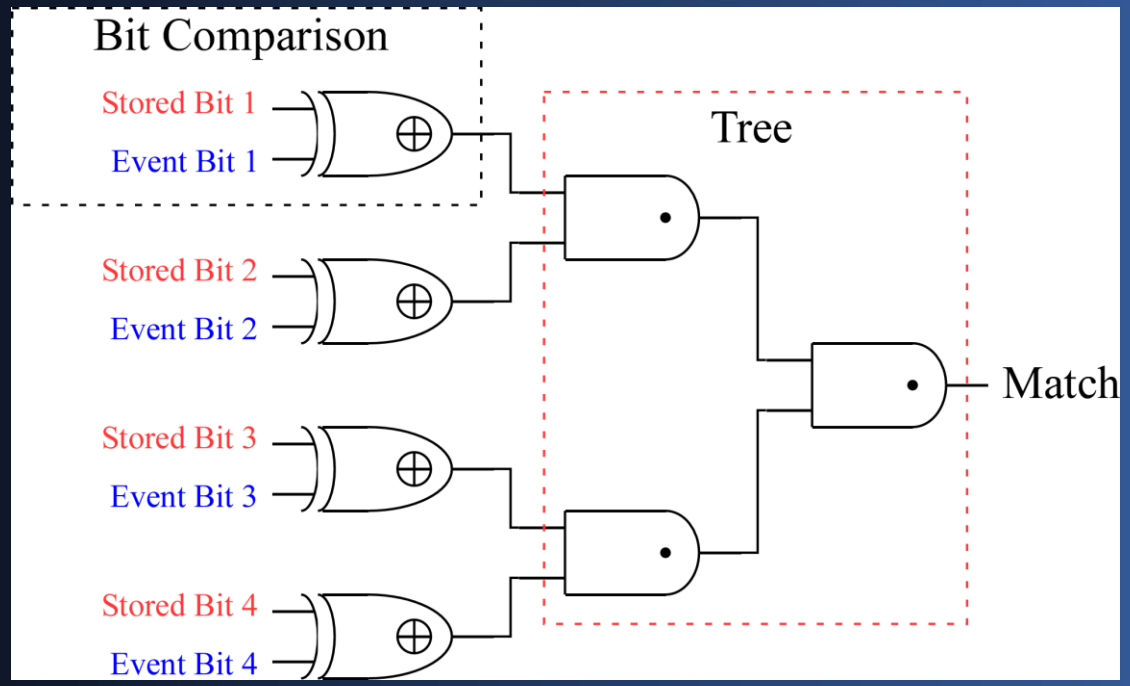
Going back to the definition of Associative Memory:

It is a memory capable of determining (the Comparator) whether a given piece of data (the Bitline) is contained in its location (the Bit Storage).

Side Bar: CAM Architectures

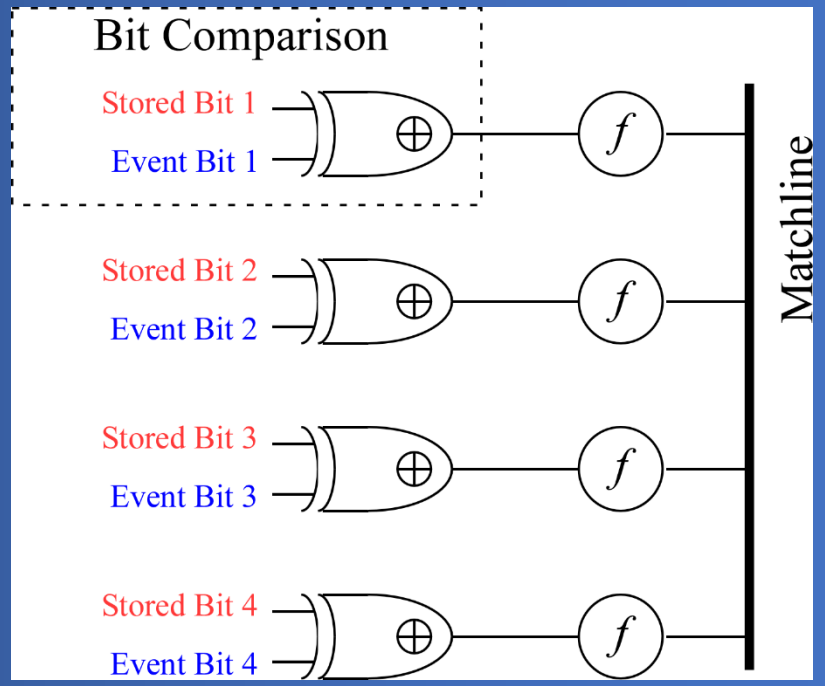


Tree Structure



AMchips

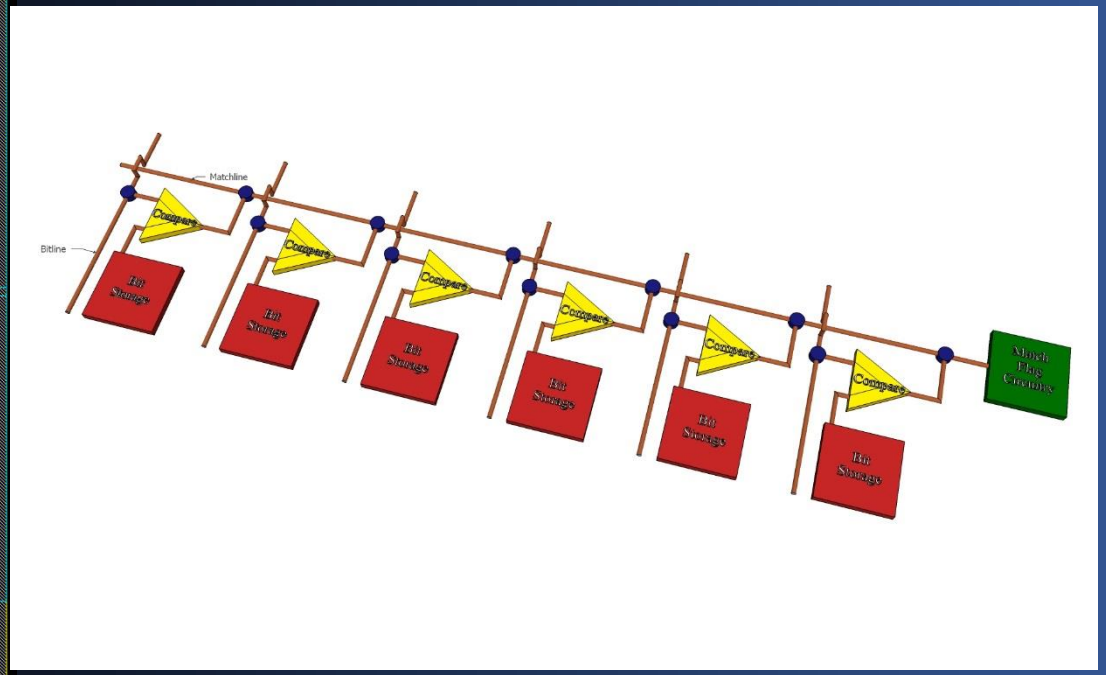
Matchline Structure



VIPRAM

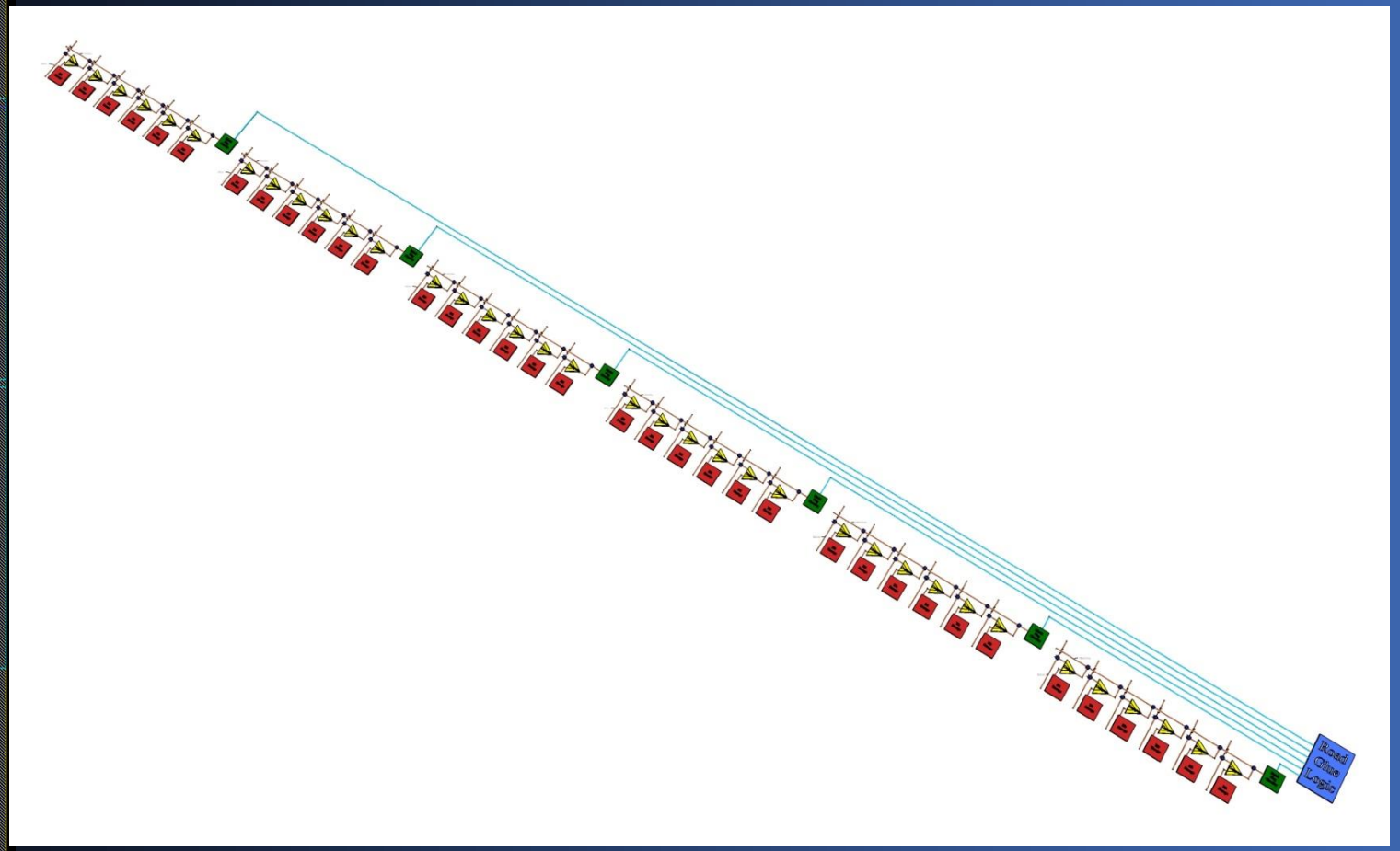
Fermilab

Moving towards OUR architecture



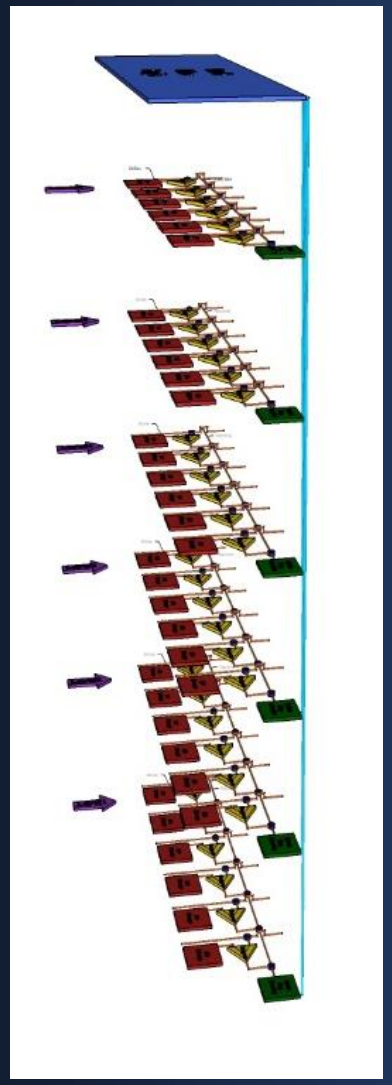
- First, you put a few CAM bits together and create something a little more useful. The picture shows a 6-bit CAM. Typical numbers we talk about are 12-bit, 15-bit or 18-bit.
- Second, the green box. Our CAMs are made to remember matches. Event data arrives to the VIPRAM unordered, so the likelihood that the layer data for a particular track will arrive to the VIPRAM_L1CMS from all of the different detector layers at the same time is vanishingly remote. We are required to remember each CAM match until all the layer data in a road match finally gets there.

Now we see our architecture…

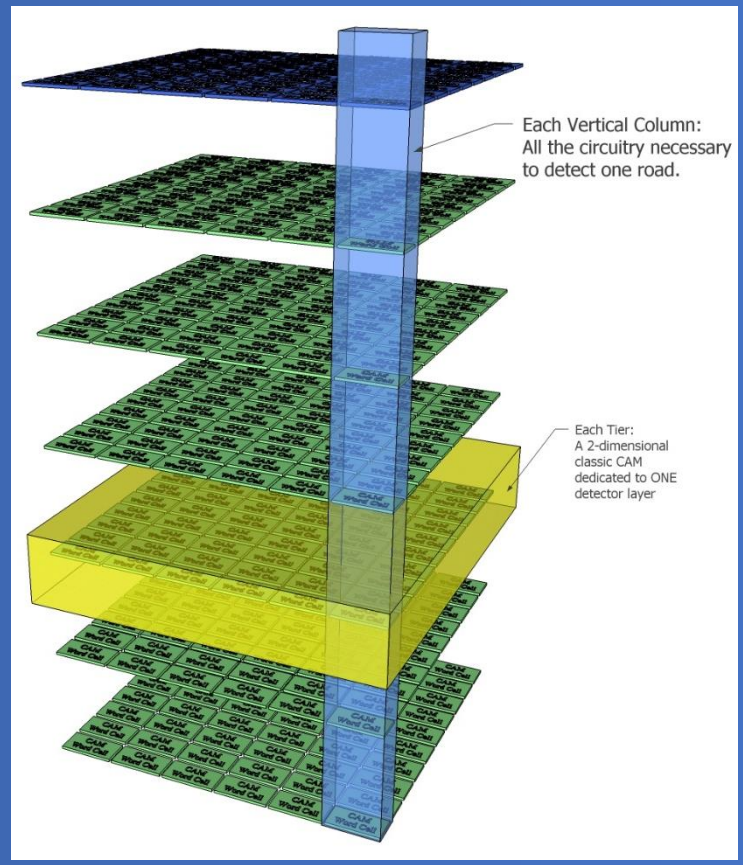


- Several independent CAMs are brought together.
  - Each is connected to a its own detector layer.
  - Each remembers its own matches.
- Finally, the CAM matches are monitored and the road flag fires when an appropriate number of CAMs match.

You mentioned 3D….

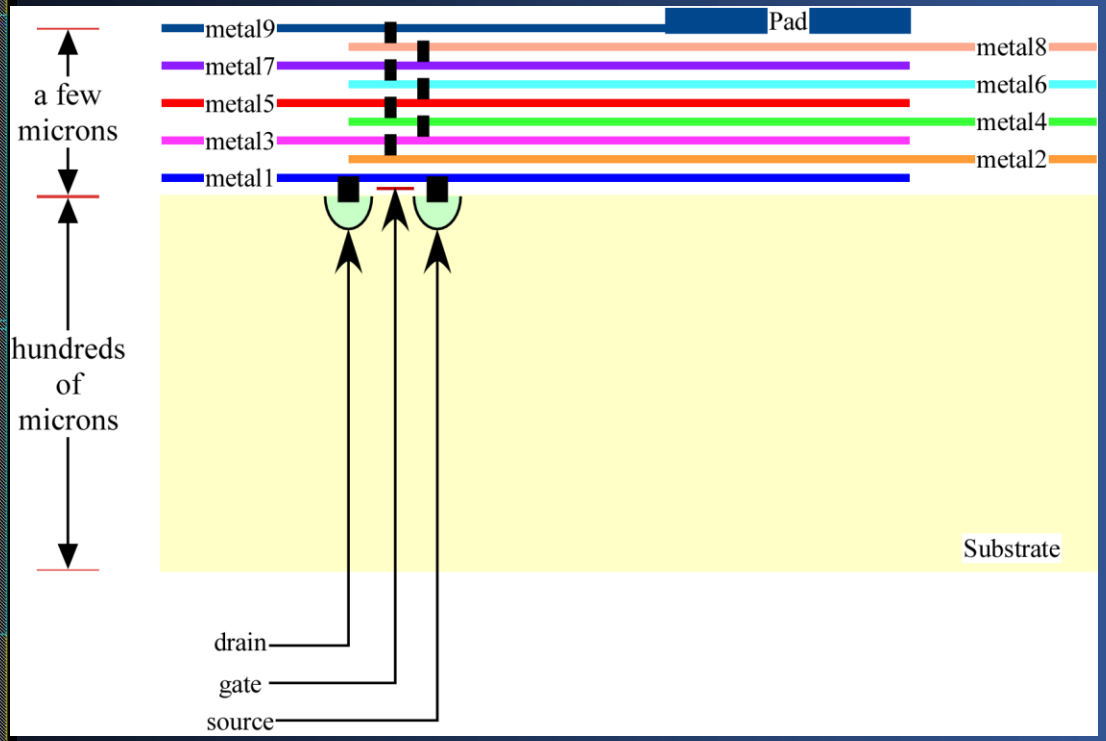


- Just about any circuit structure can be converted into 3D. It is just a question of whether or not you gain anything by it.
- With 3D you get more transistors per acre
- With 3D you get more route layers
- With 3D you get another degree of freedom in *placement* and *route*.



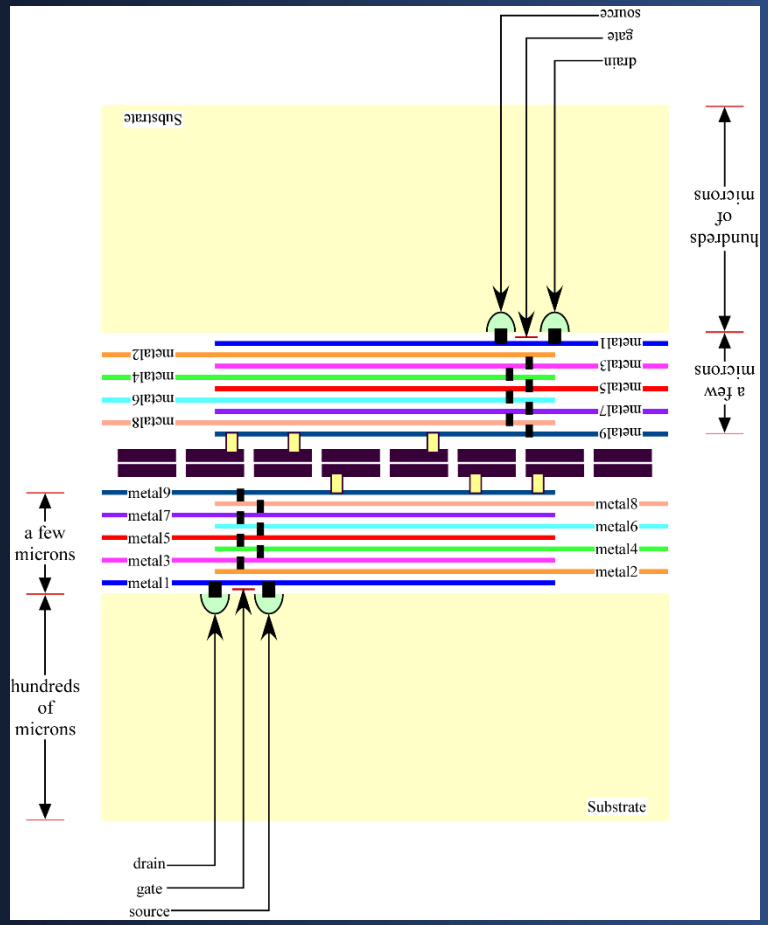
Let’s back up a bit…

Ordinary 2D VLSI



1. A silicon substrate, typically very lightly doped and relatively thick – 300 μm is common.
2. Transistors created in the substrate through ion implantation
3. Connections are made to transistor sources, gates and drains by “contacts” between polysilicon or doped silicon and metal1.
4. 9 layers of metal connected to one another through vias between one layer and its nearest neighbor.
5. Pads provide an ohmic connection between the top metal layer and the outside work through an opening in the oxide.

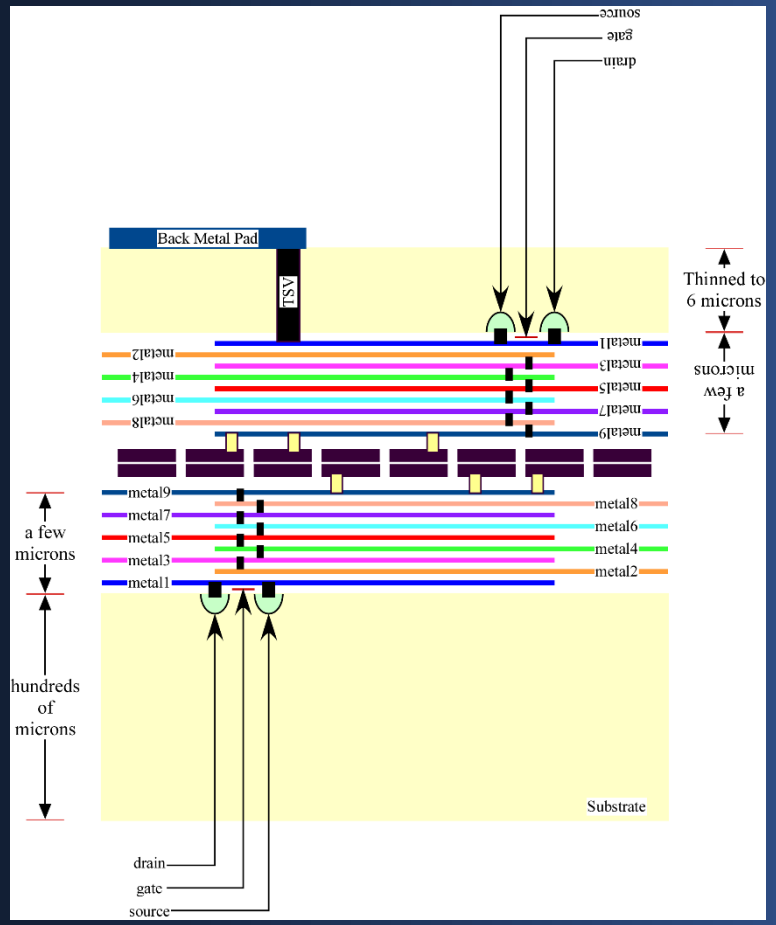
2-Tier 3D



- First, two unique 2D VLSI wafers are created in a garden-variety VLSI process (or two different garden-variety VLSI processes). Nothing fancy here.
- Second, no pads. They simply are not fabricated during the 2D VLSI process. All that is required to accomplish this is to inform the foundry. No big deal.
- Third, a new metal layer is added to each wafer above the top 2D VLSI metal (the Zip Layer). This is added by the 3D fabricator.
- One wafer is flipped over on top of the other and with the help of a little pressure and temperature, the Zip Layers on each tier are adhered to one another.
- This is called a Face-to-Face bond.

Fermilab

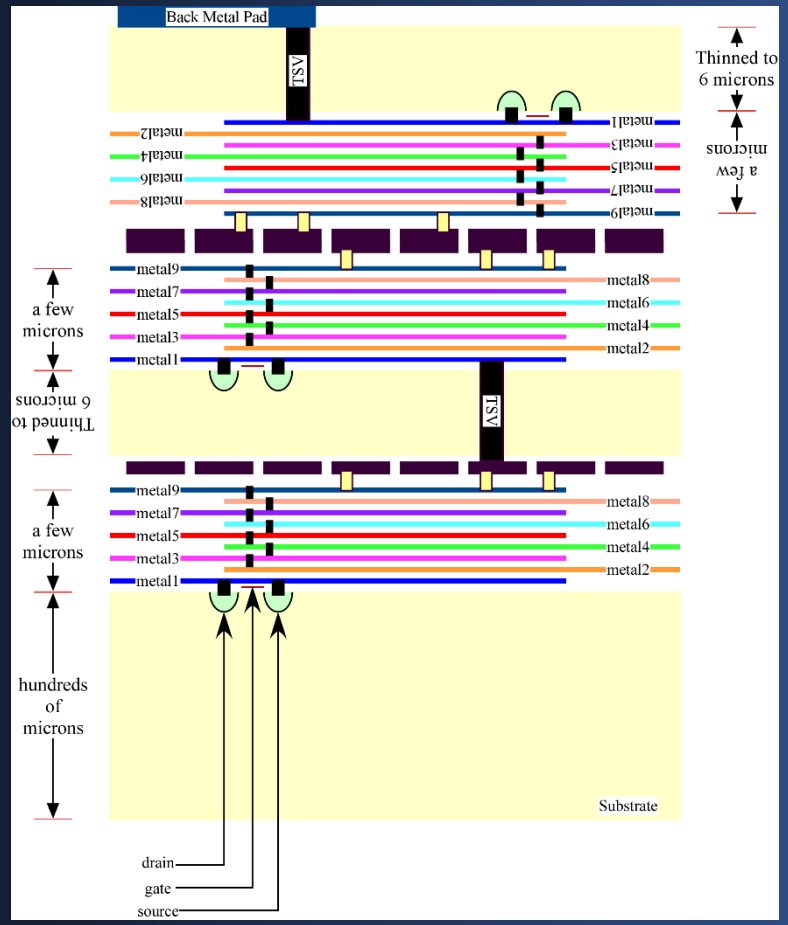
Wait a minute!!! No pads! How do we connect?



- Chemical Mechanical Polishing thins the top tier down to 6 μm. (That's right…all the mechanical support is in the bottom tier. Mechanically the top tier would make a tissue look huge and strong.)
- Cavities are opened in the thinned substrate with the Bosch Process. (A time-multiplexed etch alternating between a plasma etch and a passivation layer deposition.) The cavities are etched down to the metal1 layer (i.e. the lowest metal layer on the top tier.)
- Tungsten is deposited in the cavity (plus a spacer for isolation).
- Finally, Aluminum back metal is deposited to form pads and (if you need it) simple routing.

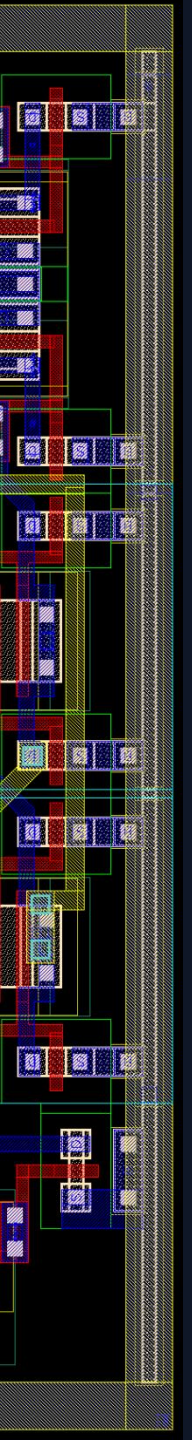
Fermilab

Can we continue? 3 layers? 4?



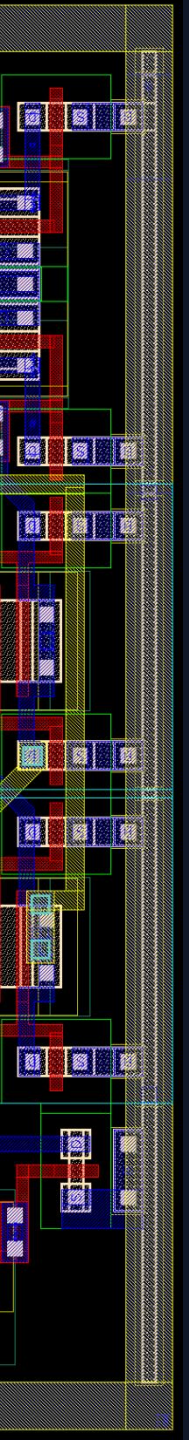
- In a word, yes. If the circuit can derive a benefit from it.
- After the initial Face-to-Face bond, subsequent bonds are Face-to-Back.

Fermilab



Are we alone in this? Who is working on it?

- Tezzaron – (2004) 8051 processor/memory stack
- Intel – (2004) 3D version of the Pentium 4. Of interest to us is that it was a 2-Tier 3D design with inter-tier communication through the bond interface and I/O and power connections through TSVs and back metal pads. Remember this.
- Intel – (2007) The Teraflops Research Chip and experimental 80-core design with stacked memory. Traditional approach consumed up to 25 watts of power. The 3D approach consumed 2.2 watts.
- Georgia Tech – (2012) 3D-MAPS a 64 custom core chip.
- Other companies – Ziptronix, SanDisk, ZyCube, Amkor, ALLVIA, austriamicrosystems, CMC Microsystems, Xilinx, Altera, Applied Materials, DARPA, DuPont, IBM, Honeywell, Qualcomm, Samsung, Sony, TI, etc, etc, etc.

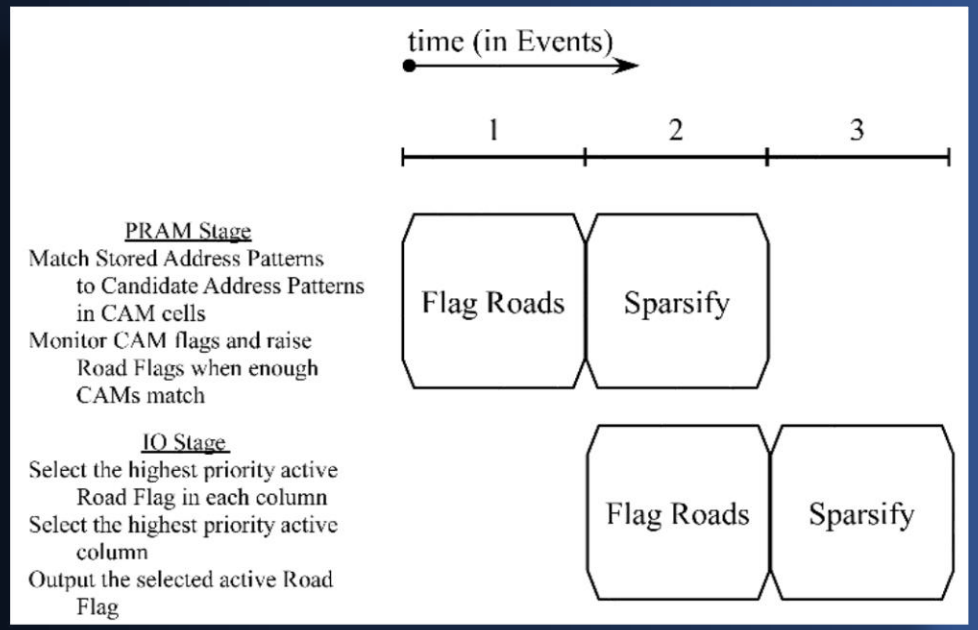


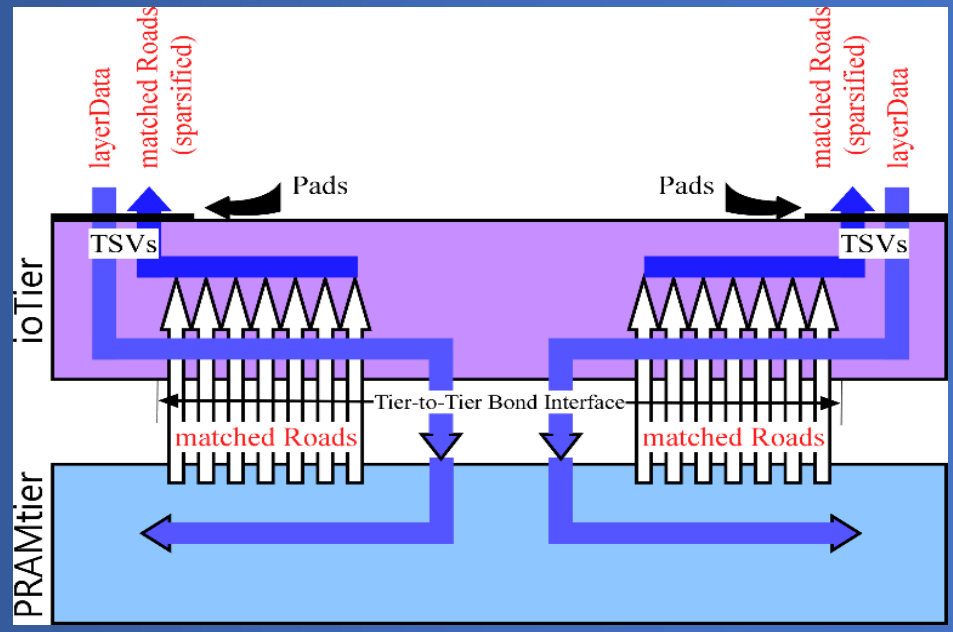
So what do WE get out of it???

Fermilab

VIPRAM_L1CMS:
A Pipelined, Vertically Integrated Architecture

- VIPRAM_L1CMS is a simple, two-staged pipelined device that
  - Stage 1: performs pattern recognition to generate road flags from current event data
  - Stage 2: sparsifies the road flags generated in the previous event and outputs them

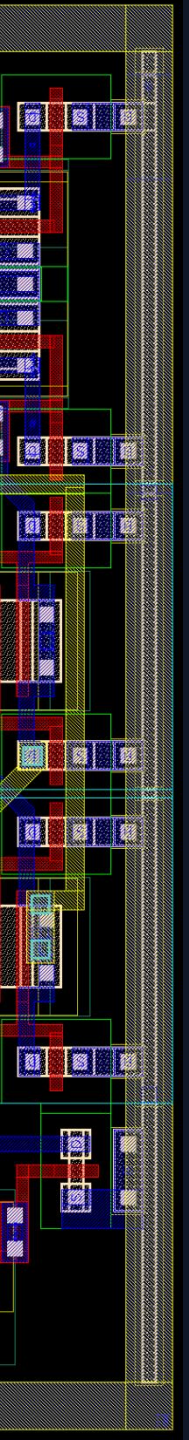




What makes VIPRAM_L1CMS unique is the simple fact that these two pipeline stages are located on two different tiers of a 3D, Vertically Integrated Circuit.

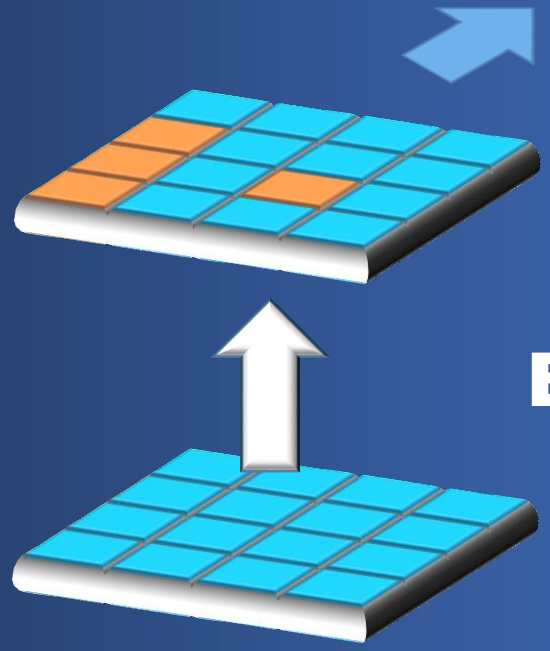
Fermilab

Using Vertical Integration Advantageously



Sparsified Road Matches

ioTier

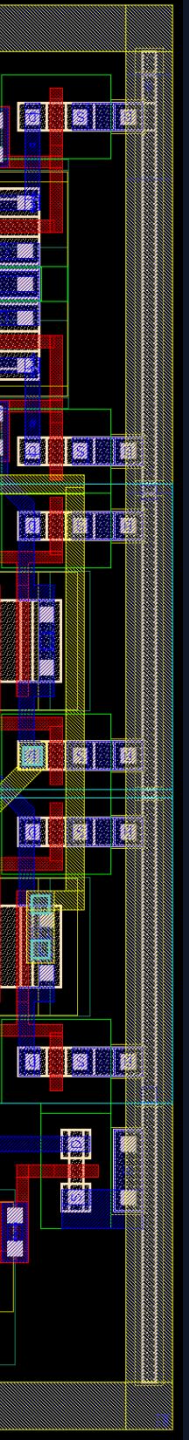


End-of-Event Signal

PRAMtier

Layer Data

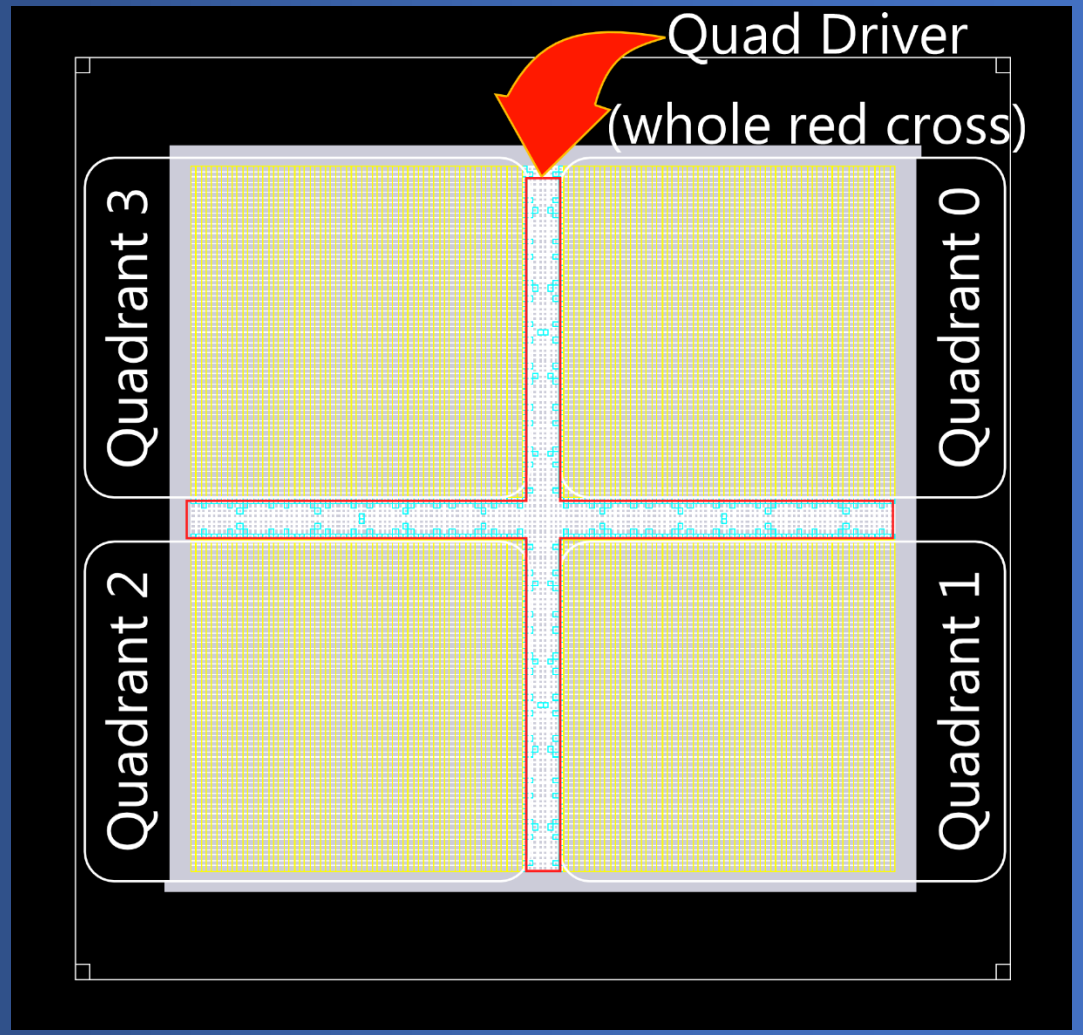
Fermilab



The PRAM Tier

- 4 Quadrants
  - 64x64 road cells
    - 8 CAM Cells
    - Majority Logic
    - Flag Logic

Notice the Quad Driver. This exists because of 3D integration. We are able to strategically drop our layer data down wherever we want to. We choose to do so exactly in the center of the chip and drive outward from there, equalizing delay.

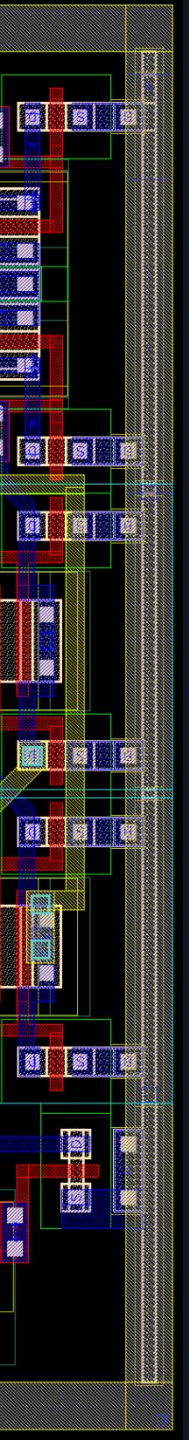


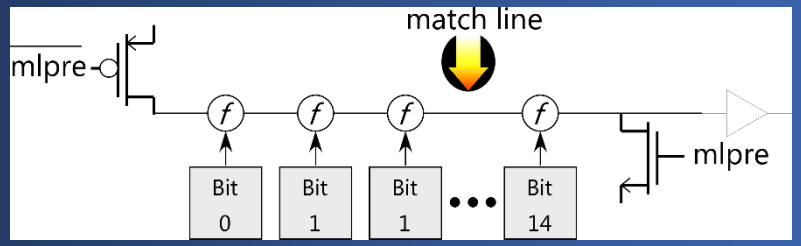
The CAM Cell

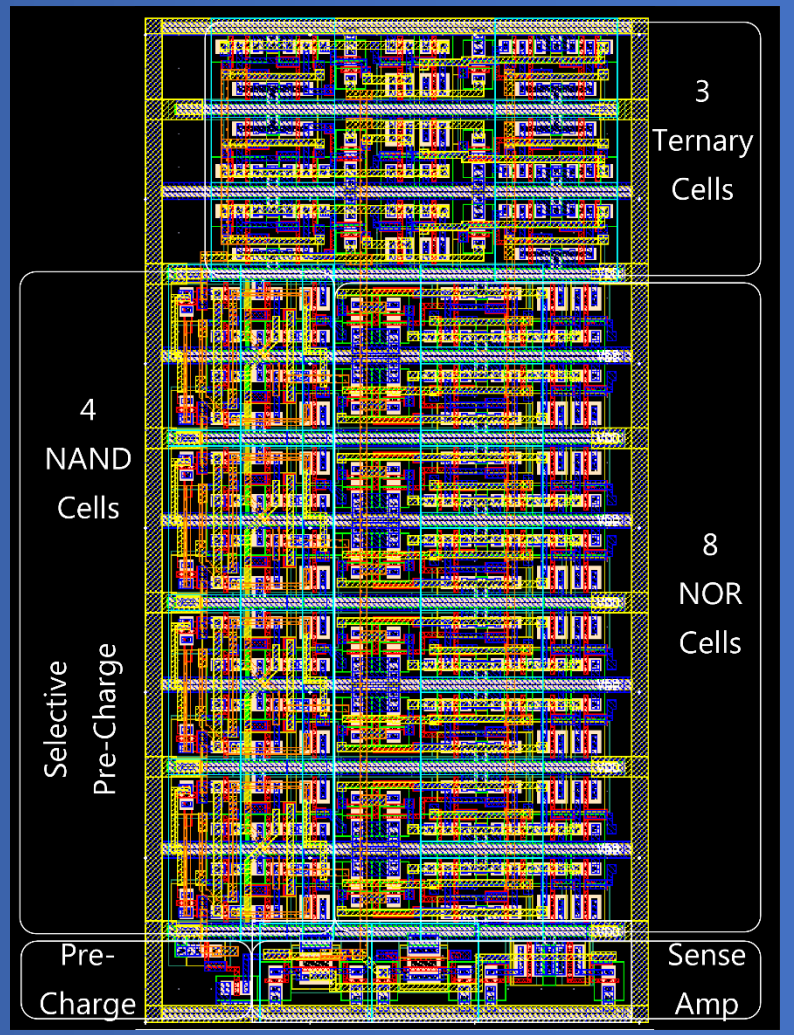
The CAM cell utilizes a classic architecture with a current race scheme and a selective pre-charge scheme.

The sense amplifier is part of a cross-coupled NOR Gate SR-flip flop.

The cell is deliberately laid out to be twice as wide as it is tall.



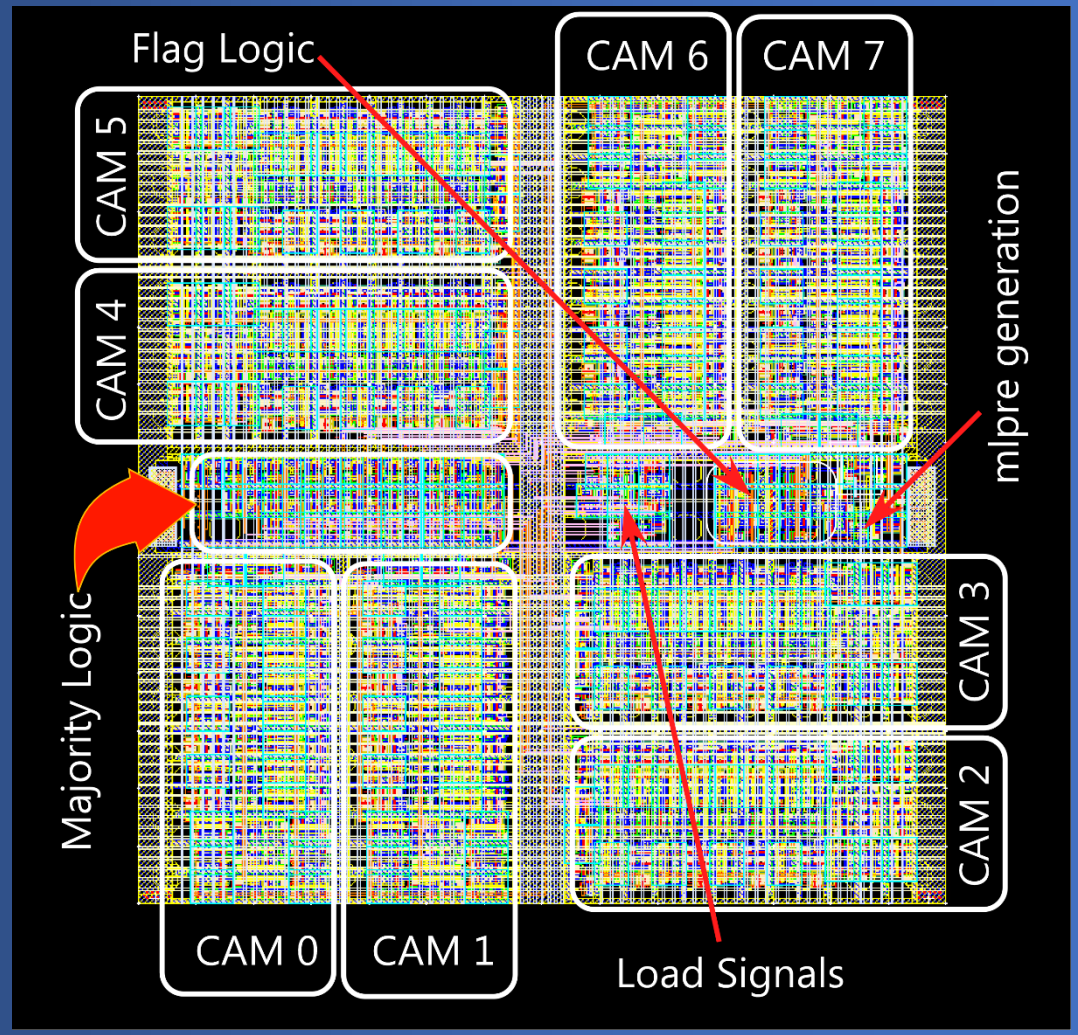


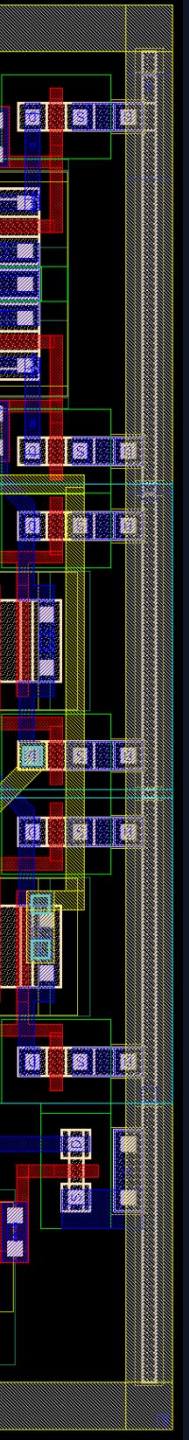


Fermilab

The PRAM Cell

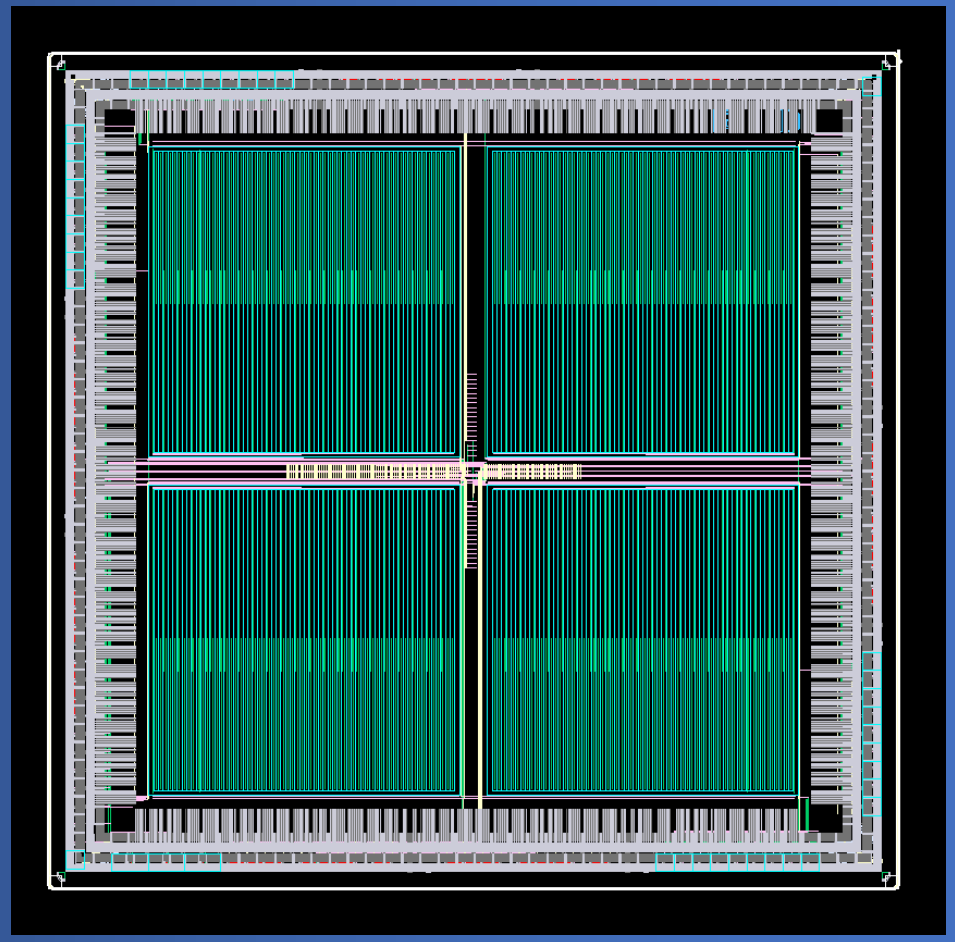
- 70 μm x 70 μm (in 130nm technology)
- Square design permits bump bonding (if desired)
- Square design permits data flow in all directions
  - Layers 0, 1, 6, and 7 flow east to west
  - Layer 2, 3, 4, and 5 flow north to south





The ioTier

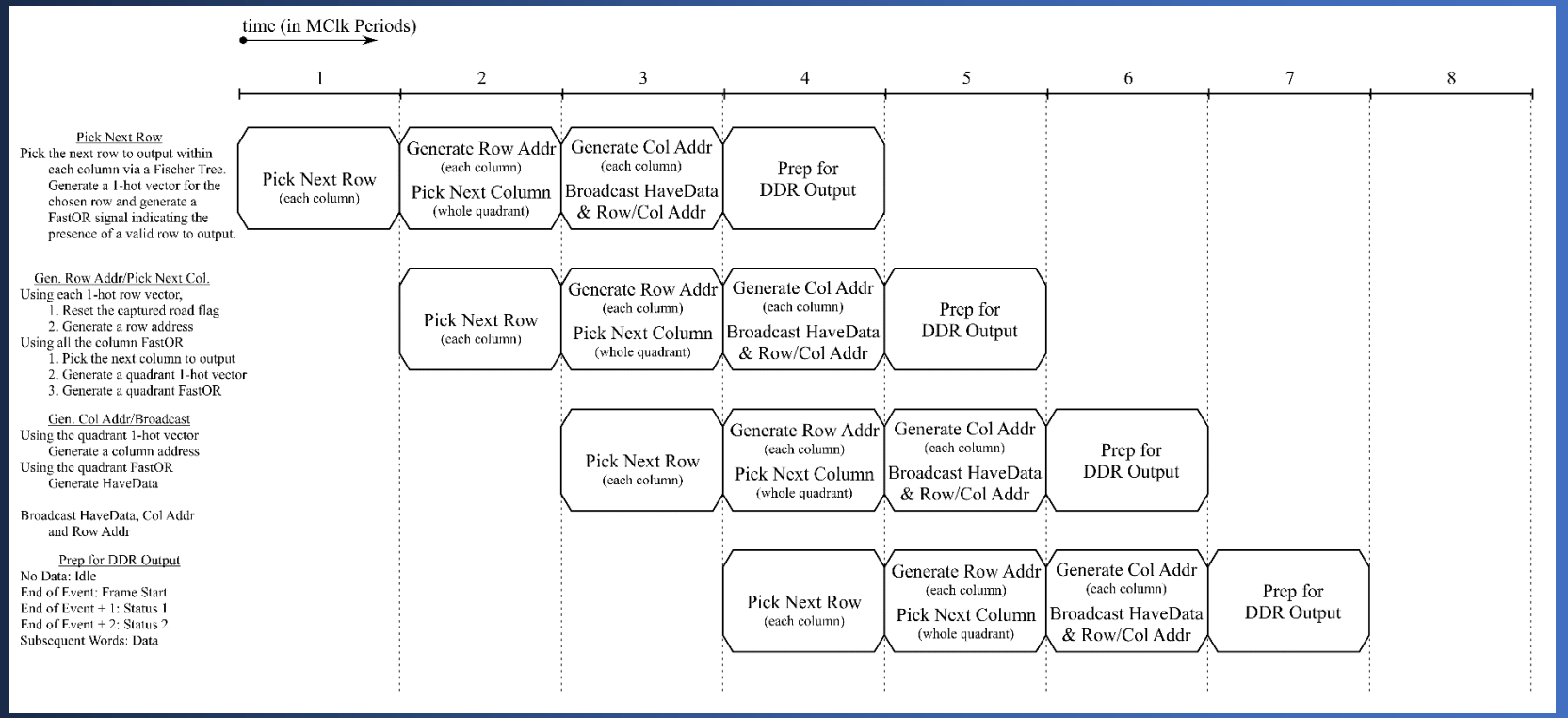
- Shaped a lot like the PRAM Tier (no surprise) except it adds pads. You can’t tell from the image, but those are back metal pads.



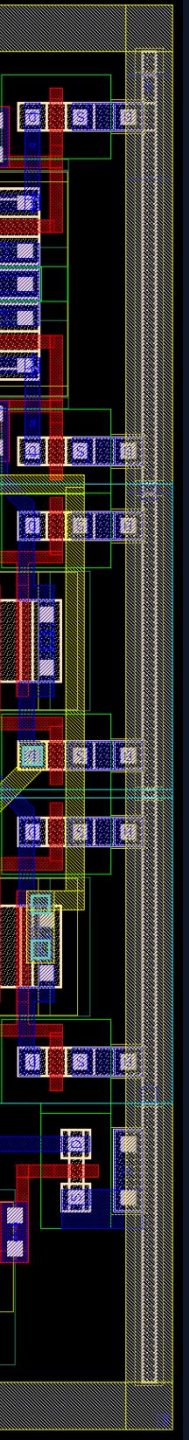
Fermilab

The ioTier

- A pipeline advancing on the master clock edge that exists within the main pipeline that advances on the End-of-Event signal.

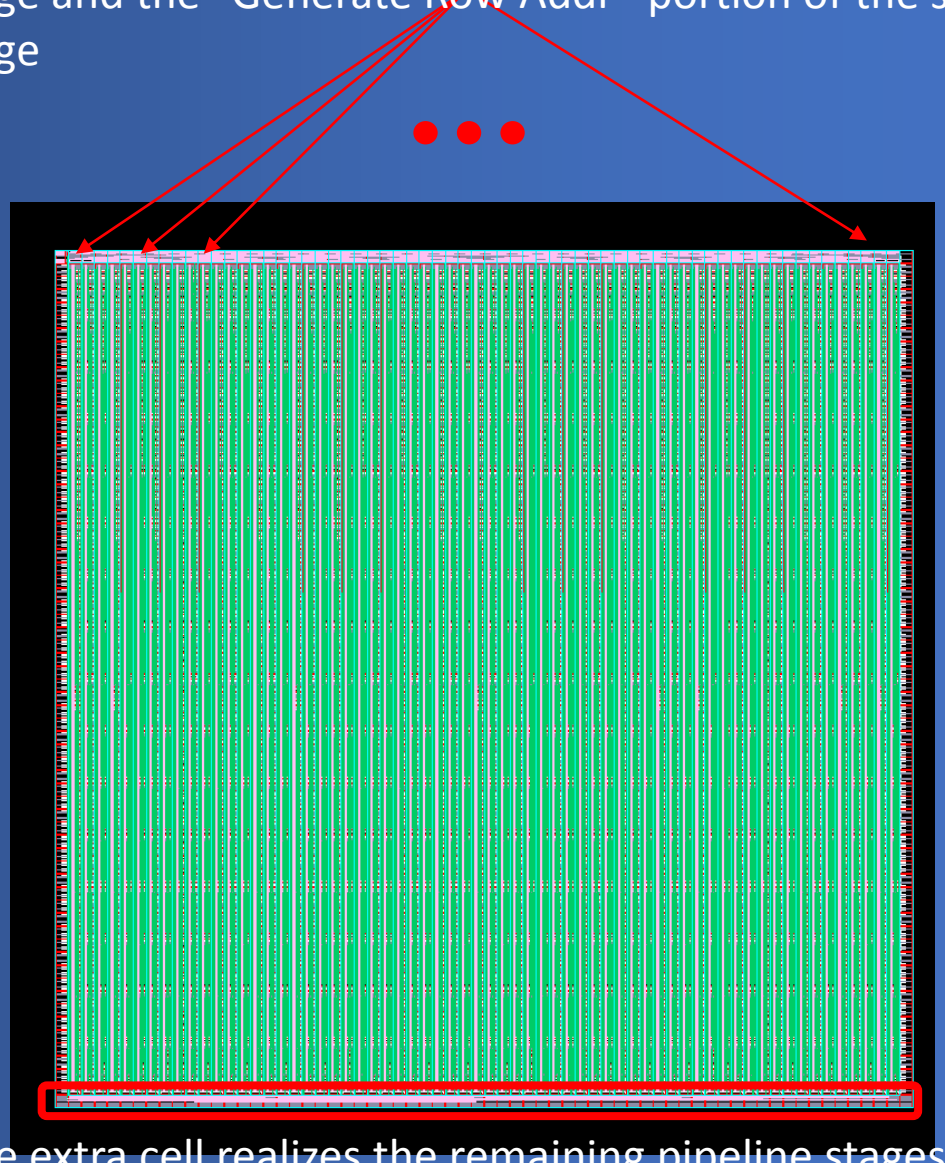


Fermilab



The ioTier Quadrant

64 identical column cells realize the entire “Pick Next Row” stage and the “Generate Row Addr” portion of the second stage



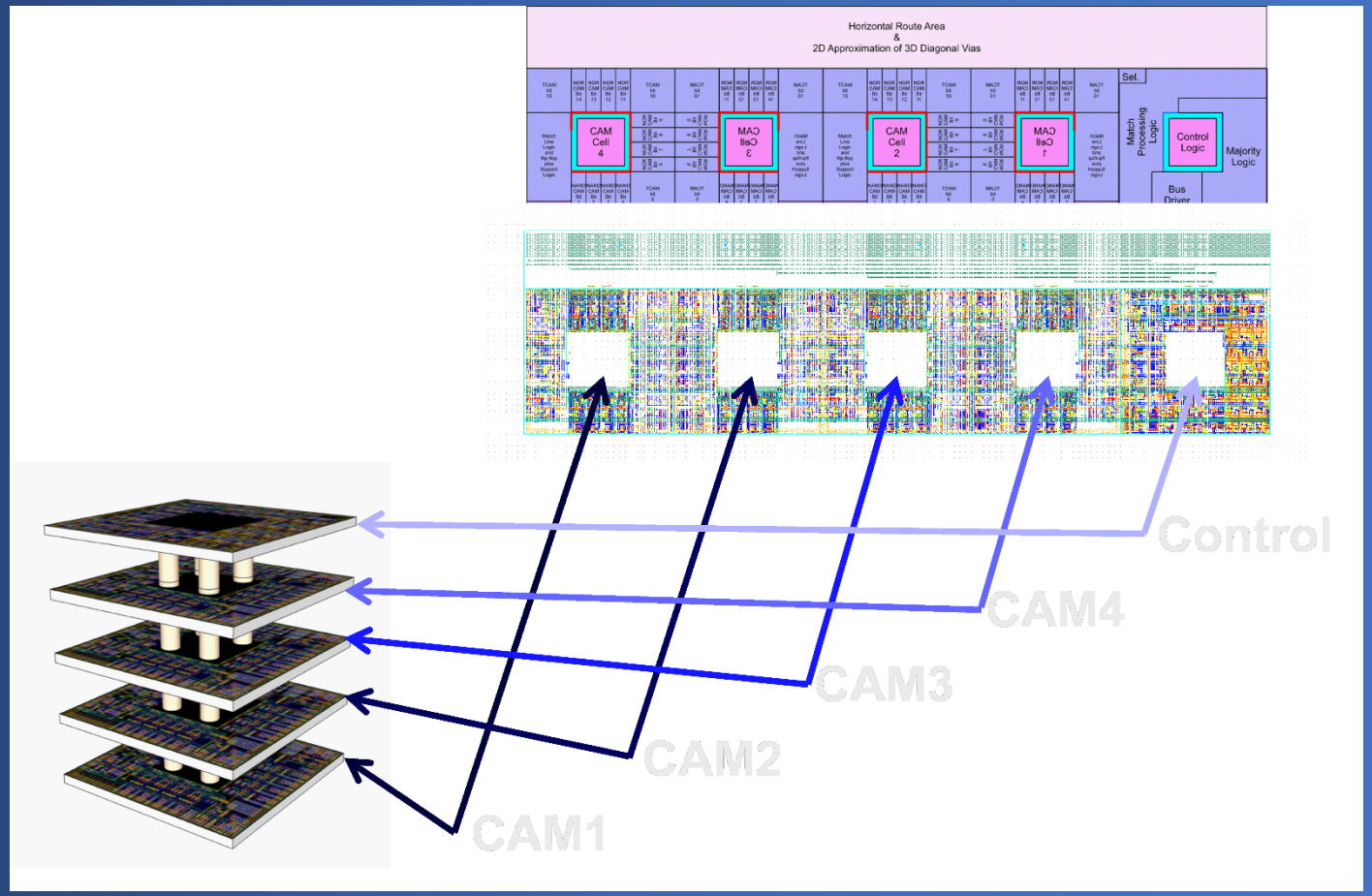
One extra cell realizes the remaining pipeline stages of the quadrant.

Fermilab

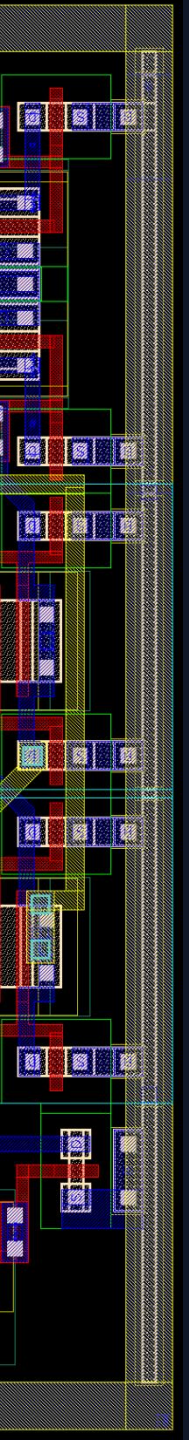
VIPRAM3D:
Generic Research towards deeper 3D stacking

In VIPRAM3D, the CAM and monitoring logic (here called Control) are reshaped. They are in squared rings. The opening in the center of the rings allows space for TSVs and active bond interfaces. Otherwise the CAMs are identical.

Earlier, protoVIPRAM00 was designed and tested. It was pin-for-pin, transistor-for-transistor identical to VIPRAM3D. Testing VIPRAM3D and comparing it to the results of protoVIPRAM00 will give a real comparison between 2D and 3D.

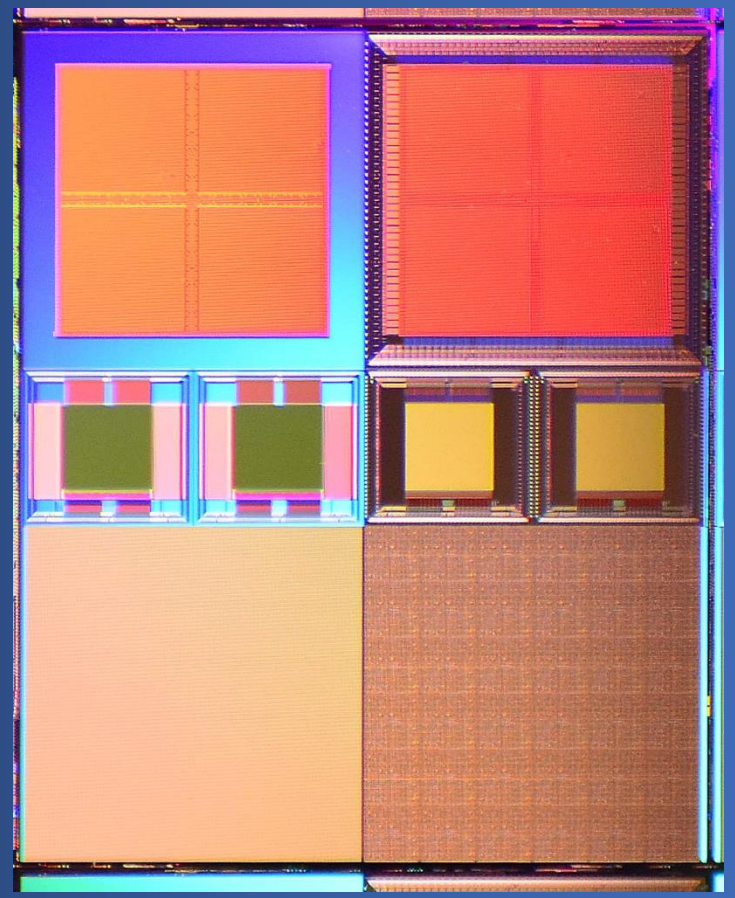


Fermilab

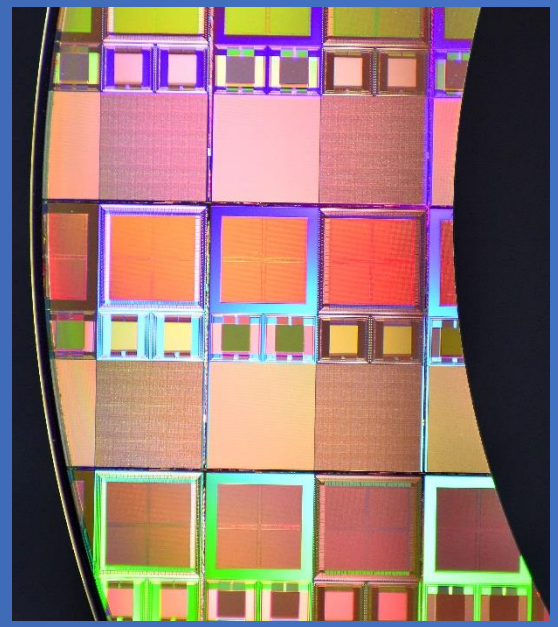


Status

- The first implementation of this architecture was submitted in March of 2016.
- Wafers were returned from Global Foundries in August of 2016
- 3D fabrication began in September of 2016.
- Final chip delivery is anticipated in early 2017.



Annulus remaining after 8 inch center was cored for 3D fabrication



Fermilab