The new variable resolution Associative Memory for Fast Track finding aka AMchip04

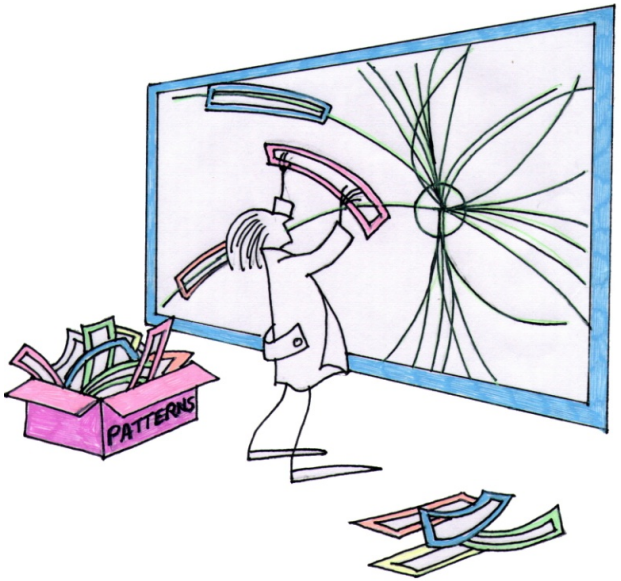
AM CMS-FTK meeting
Sep 10-11, 2012, Pisa

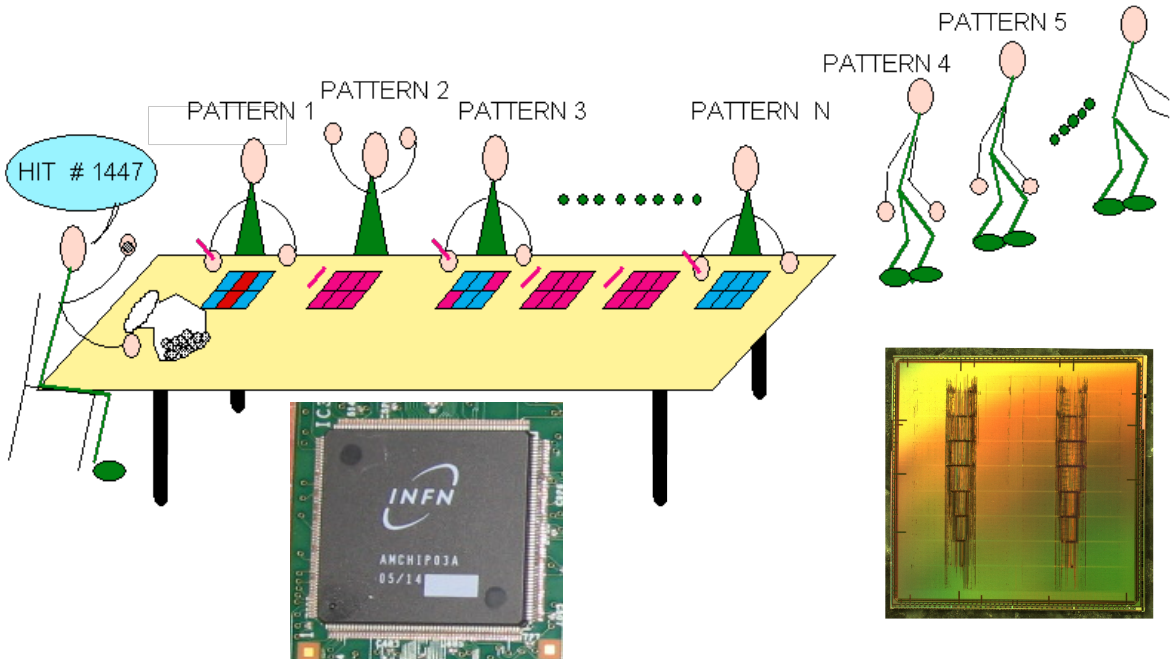
Alberto Annovi

Istituto Nazionale di Fisica Nucleare
Laboratori Nazionali di Frascati

FTK algorithm:
Pattern recognition & Track fitting

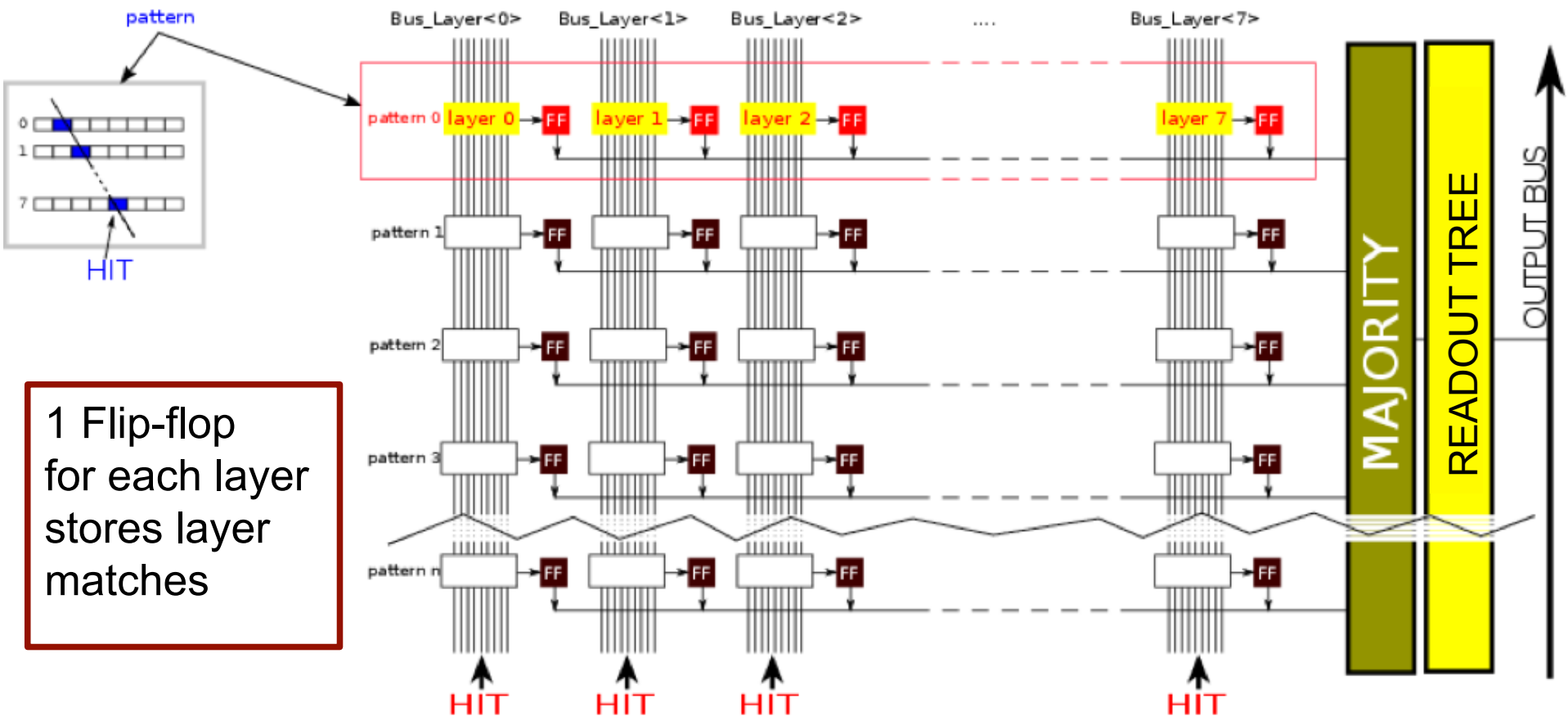
- Pattern recognition – find track candidates with enough Si hits





  - O(10^9) prestored patterns simultaneously see the silicon hits leaving the detector at full speed.
  - Based on the Associative Memory chip (content-addressable memory) initially developed for the CDF Silicon Vertex Trigger (SVT).

AM working principle



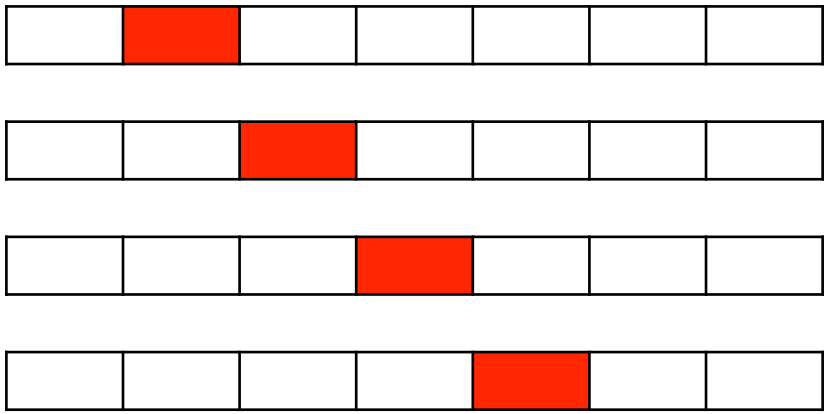
1 Flip-flop for each layer stores layer matches

All patterns compared in parallel with incoming data. Look for correlation of data received at different times. (Feature unique to AMchip)

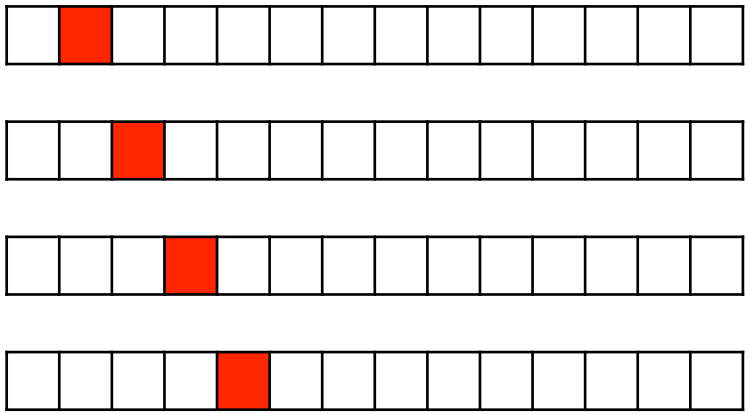
Fast pattern matching. Flexible input: position, time, objects…

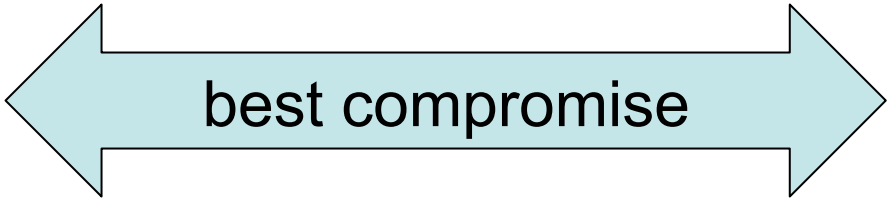
Generatig the pattern bank

Wide patterns



Thin patterns





High efficiency
with less patterns (hardware)
BUT more fakes

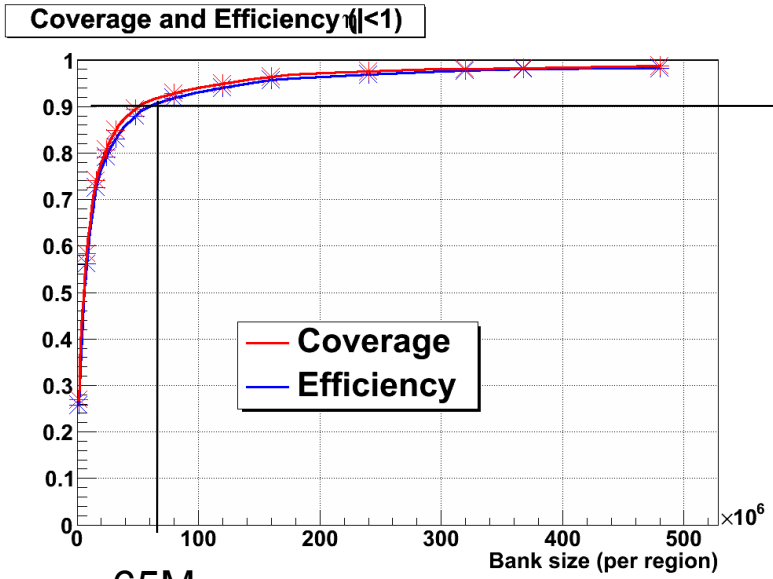
More patterns (hardware)
for same efficiency
less fakes

Pattern efficiency

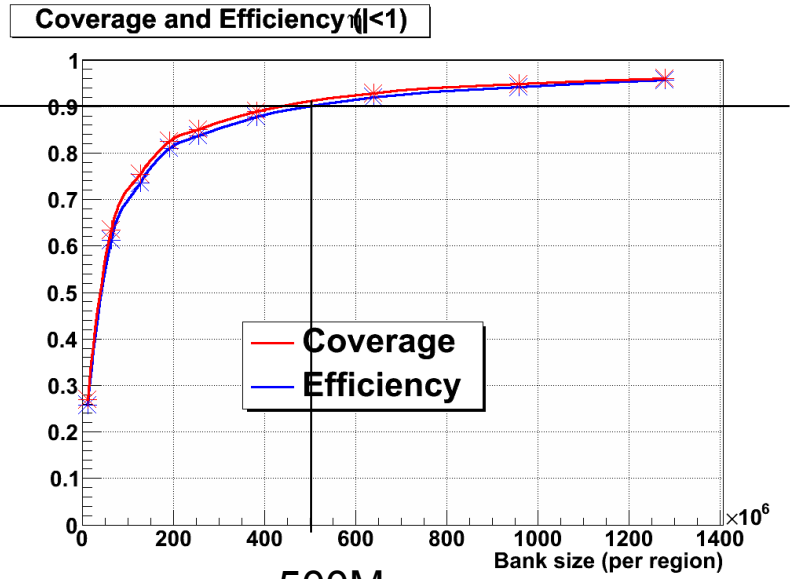
ATL-UPGRADE-PROC-2011-004

Pattern size
r-ϕ: 24 pixels, 20 SCT strips
z: 36 pixels

Pattern size (half size)
r-ϕ: 12 pixels, 10 SCT strips
z: 36 pixels



65M



90%

500M

# of patterns in Amchips (barrel only, 45 ϕ degress)

<# matched patterns/event @ 3E34> = 342k

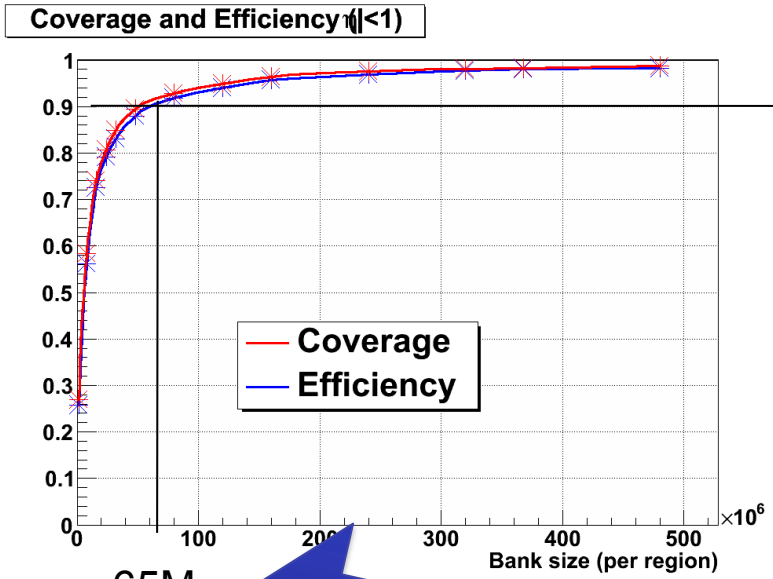
<# matched patterns/event @ 3E34> = 40k

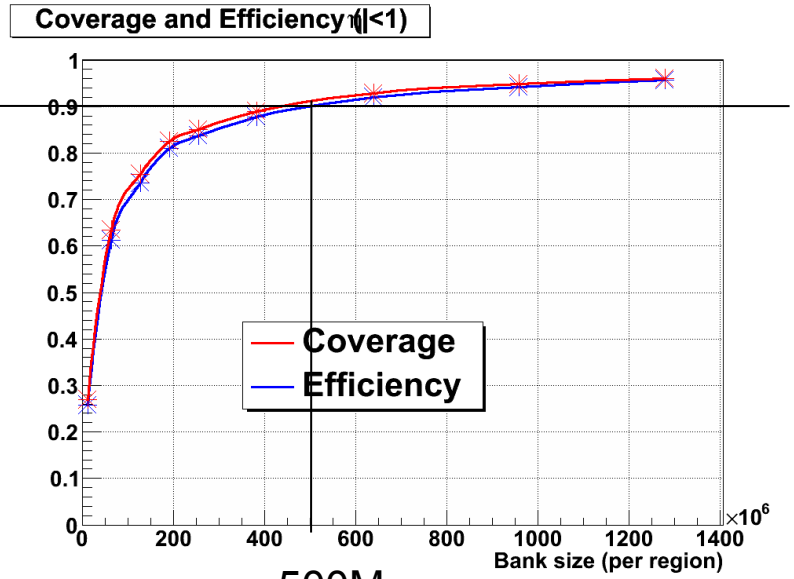
Pattern efficiency

ATL-UPGRADE-PROC-2011-004

Pattern size
r-ϕ: 24 pixels, 20 SCT strips
z: 36 pixels

Pattern size (half size)
r-ϕ: 12 pixels, 10 SCT strips
z: 36 pixels





90%

65M

500M

o

(barrel only, 45 ϕ degress)

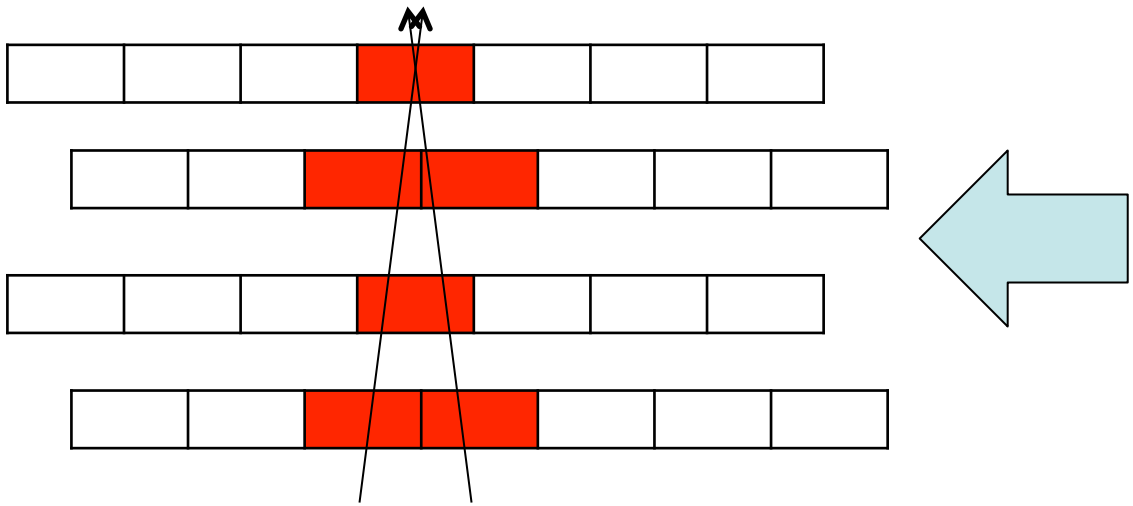
Want this

<# matched patterns/event @ 3E34> = 342k

atched patterns/event @ 3E34> = 40k

discretiziation effects

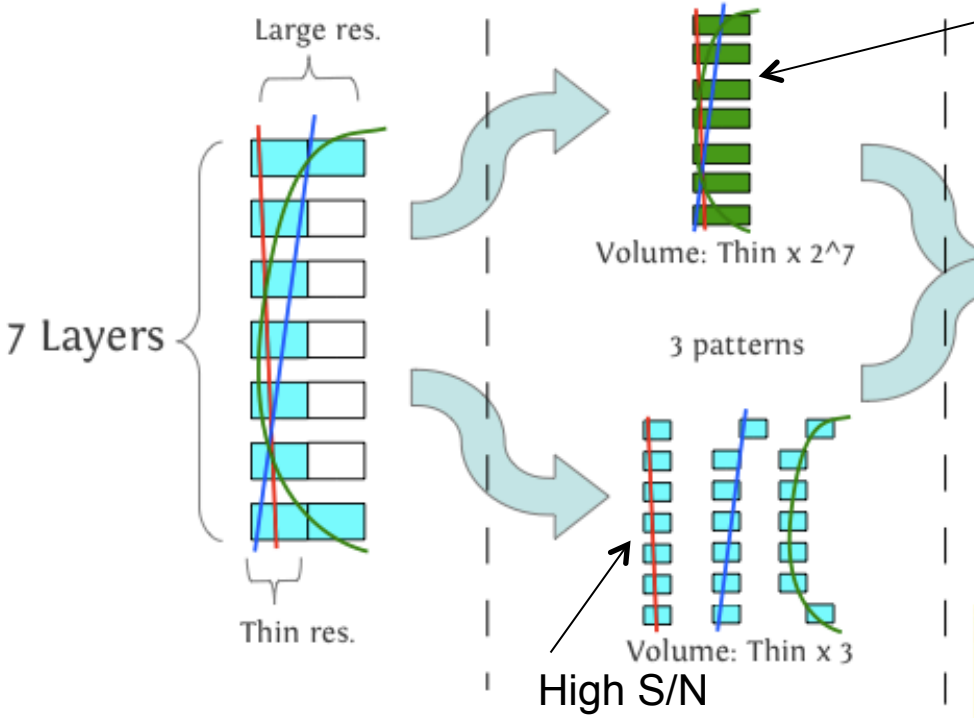
Layers are not aligned



Would use 4 patterns locations instead of 1 without variable resolution

AMCHIP04: VARIABLE RESOLUTION

A new "Variable Resolution Associative Memory" for High Energy Physics
ATL-UPGRADE-PROC-2011-004
doi:10.1109/ANIMMA.2011.6172856

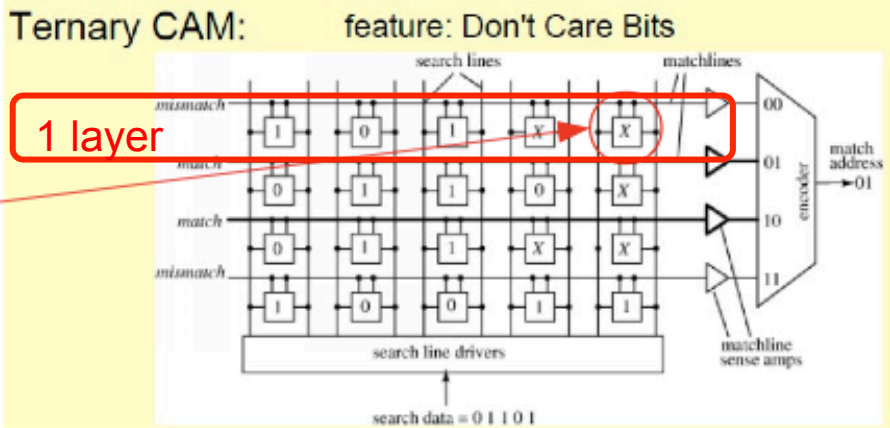


Good rejection and occupy only one pattern location.

Per-pattern choice of optimal resolution.

We can use **don't care** on the least significant bit when we want to match the **pattern layer @ Large resolution** or use all the bits to match it **@ Thin resolution**

Coincidence window is programmable layer by layer and pattern by pattern

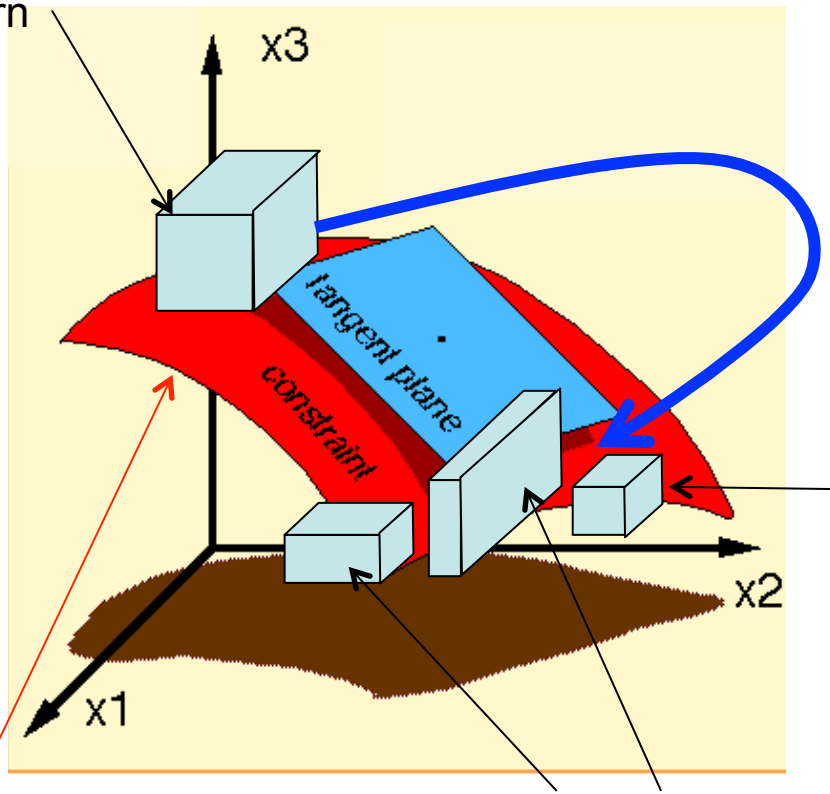


The patterns: a different point of view

5 strip + 3 pixel layers
→11 coordinates
→11D hit coord. space

Large pattern

The pattern bank:
• cover the track manifold with patterns.
• covered space outside manifold → fakes.
• variable resolution → dramatically improves S/N



A factor of 2 on each side
→a factor 2^11 less volume
→O(1/2048) less fakes!!
… forgetting correlated hits

Thin pattern

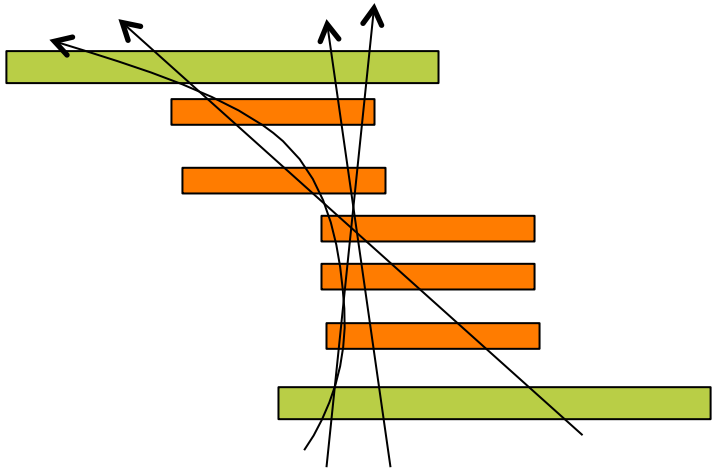
Fixes resolution patterns → fixed aspect ratio

5D track manifold

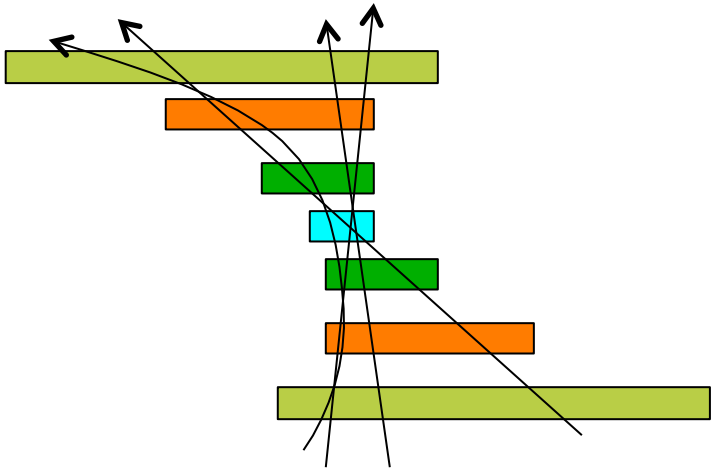
Variable resolution patterns

Many bits variable resolution

1 bit variable resolution



3 bit variable resolution



1 pattern
Volume 4*
Volume 2^(7*2)*4* = 2^16

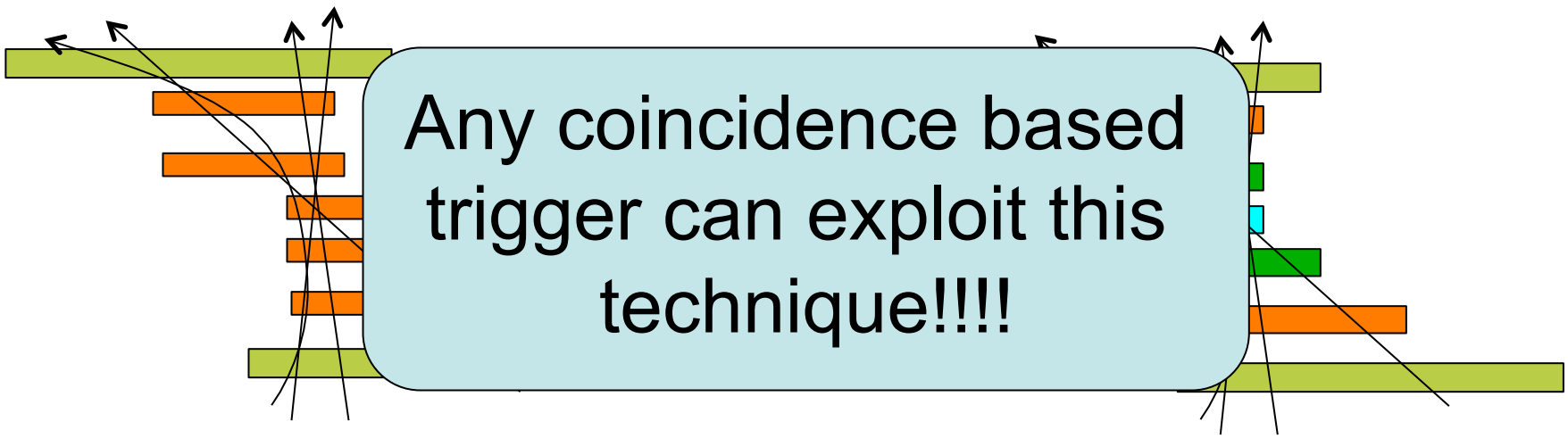
1/16 less volume
less fakes!!!

1 pattern
Volume 1/4*
Volume 2^12

Many bits variable resolution

1 bit variable resolution

3 bit variable resolution



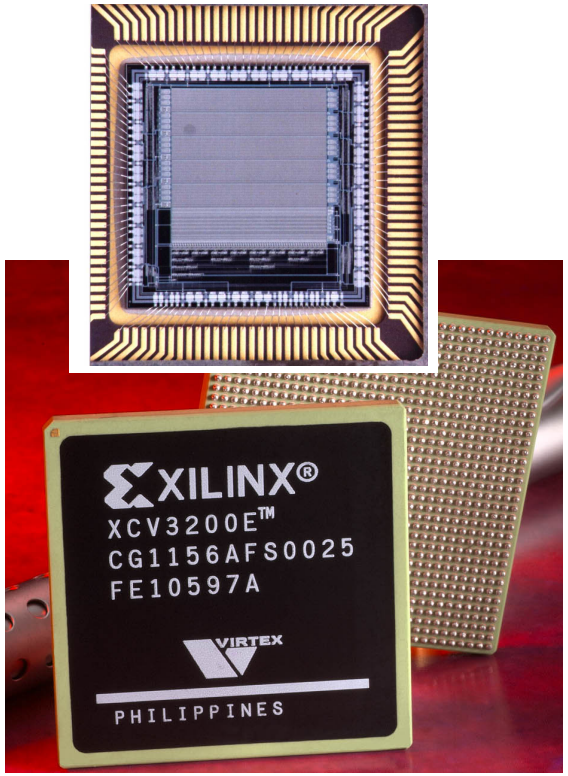
Any coincidence based trigger can exploit this technique!!!!

1 pattern
Volume 4*
Volume 2^(7*2)*4* = 2^16

1/16 less volume
less fakes!!!

1 pattern
Volume 1/4*
Volume 2^12

AM chips from 1992 to 2005



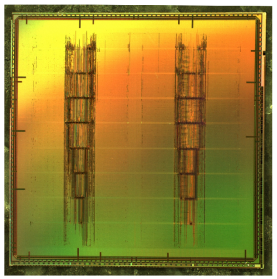
- (90's) **Full custom VLSI chip -** 0.7μm (INFN-Pisa)
- **128 patterns, 6x12bit words each**
- **384k patterns (SVT total)**

F. Morsani et al., "The AMchip: a Full-custom MOS VLSI Associative memory for Pattern Recognition", IEEE Trans. on Nucl. Sci., vol. 39, pp. 795-797, **(1992).**

On the opposite side: **FPGA** for the same AMchip

P. Giannetti et al. "A Programmable Associative Memory for Track Finding", Nucl. Intsr. and Meth., vol. A413/2-3, pp. 367-373, **(1998).**

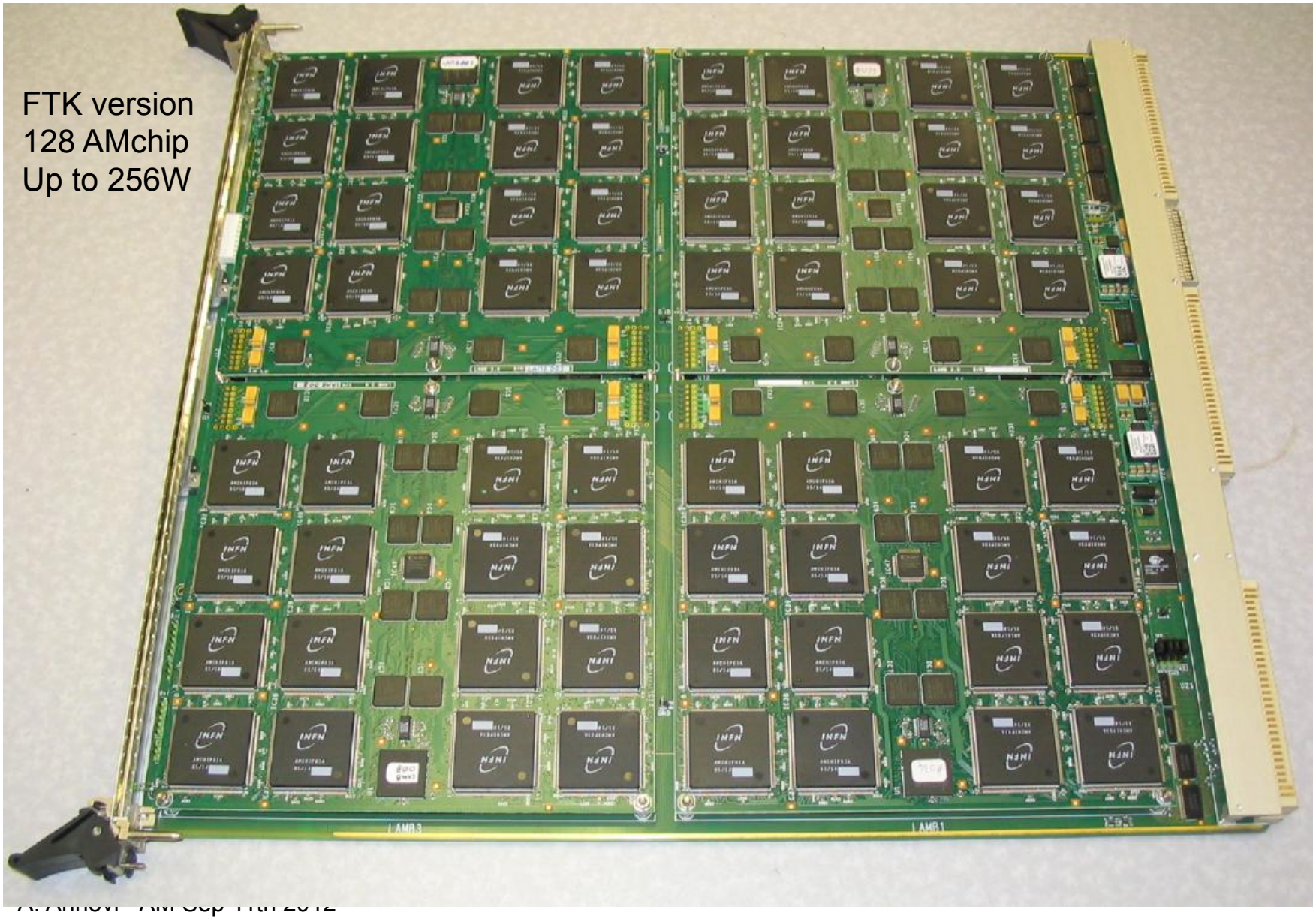
G Magazzu' I progetto standard cell presented @ LHCC (1999)



In the middle: **Standard Cell 0.18** μm (INFN-Pisa-Ferrara)→ 5000 pattern/chip Amchip SVT upgrade total: 6M patterns

L. Sartori, A. Annovi et al., "A VLSI Processor for Fast Track Finding Based on Content Addressable Memories", **IEEE Transactions on Nuclear Science**, Volume 53, Issue 4, Part 2, Aug. **2006** Page(s):2428 - 2433

AMchip03 array (AM board)



AMchip Comparison

	AMchip03	AMchip04	effect
Technology	180nm	65nm LP	X8 pattern density
Clock freq.	50MHz	100MHz	Faster, higher power cons.
Die size	10x10mm^2	12x12mm^2	X1.5 patterns (prototype 3.5x4mm^2)
Core voltage	1.8V	1.2V	Lower power cons.
Core power	1.3W	2W	At 40MHz and 100MHz respec.
Full custom	No	Yes	X2 pattern density
Layers	6	8	¾ pattern density
Patterns/chip	5k	80k	8k in prototype
Ternary layers	N/A	3 to 6	Better S/N with variable resolution
Bits/layer	18	15	
Input hit b/w	4.3	12	Gbit/s
		2 event buf.	readout 1st, load 2nd event

The hard part: FTK goal 1 billion pattern for LHC phase I (80k patt * 16k chips)

push pattern density to the limit, keep power under control despite x16 patterns x8/6 layers and 40MHz --> 100MHz

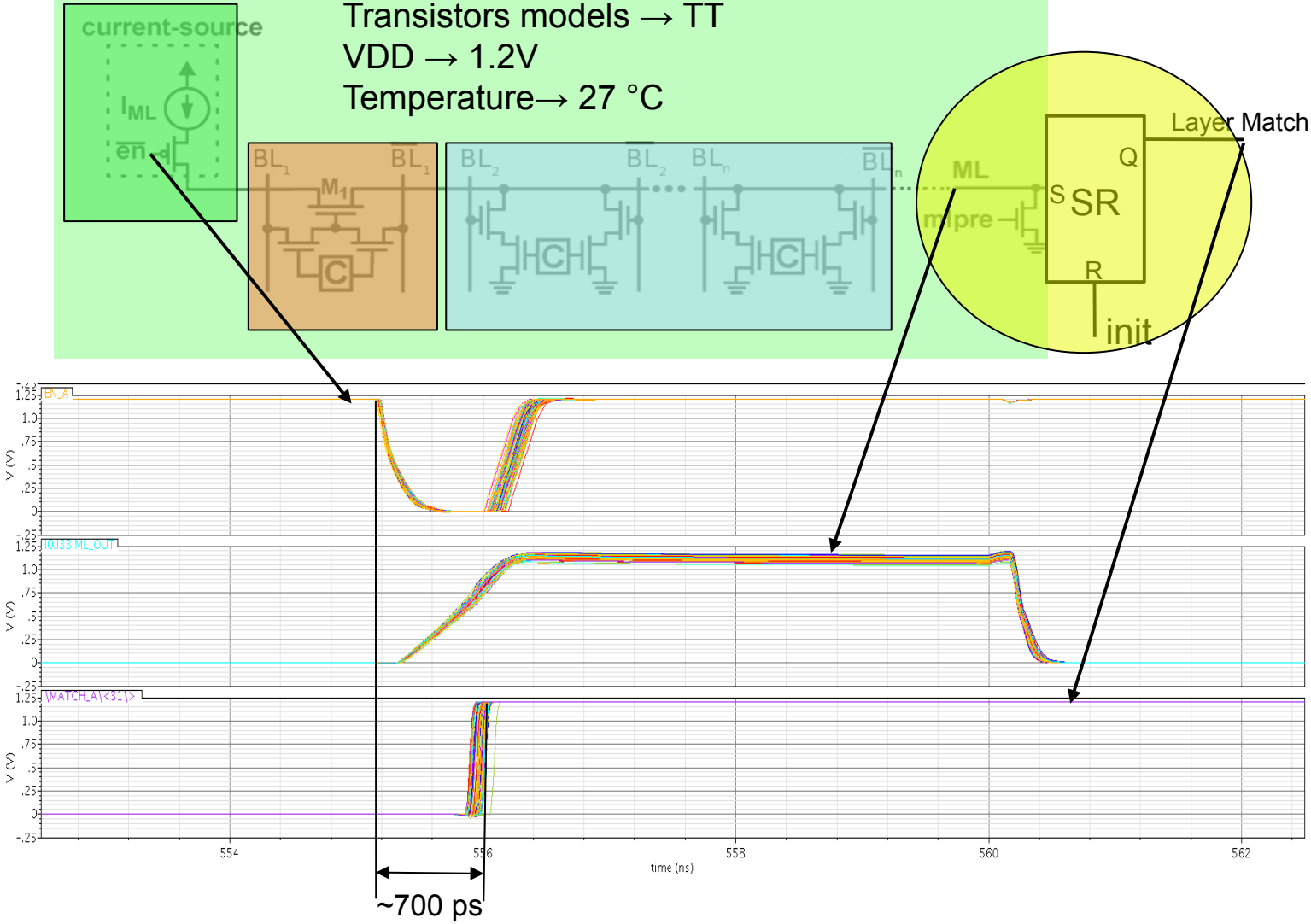
would mean x50 power consumption with same design & technology

AM chip04 functions / specs

- Store pre-calculated trajectories (patterns)
  - Each pattern: 8 positions (numbers or words) one for each layer
- Compare patterns with incoming data
  - Detectors hits for one event
- For each event readout patterns
  - with enough hits 8/8, 7/8 or 6/8
- For each pattern readout:
  - Pattern address (ID) + bitmap of fired layers
- Configuration and pattern loading through JTAG interface

CAM layer timing diagram

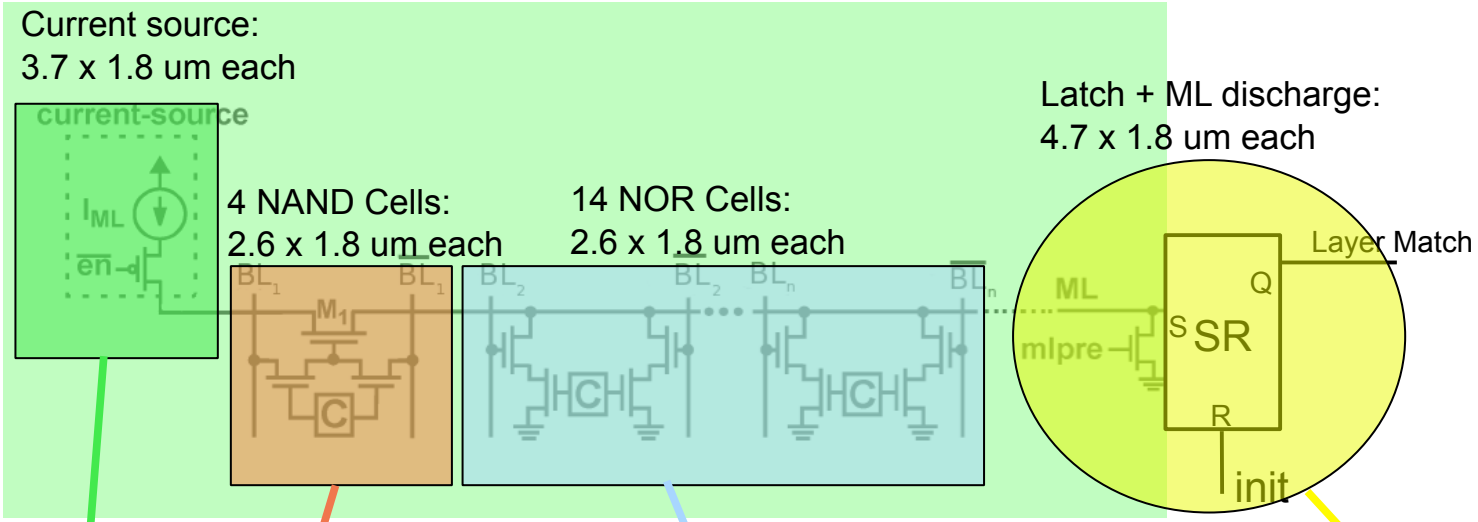
Simulation done in nominal conditions:
Transistors models → TT
VDD → 1.2V
Temperature→ 27 °C



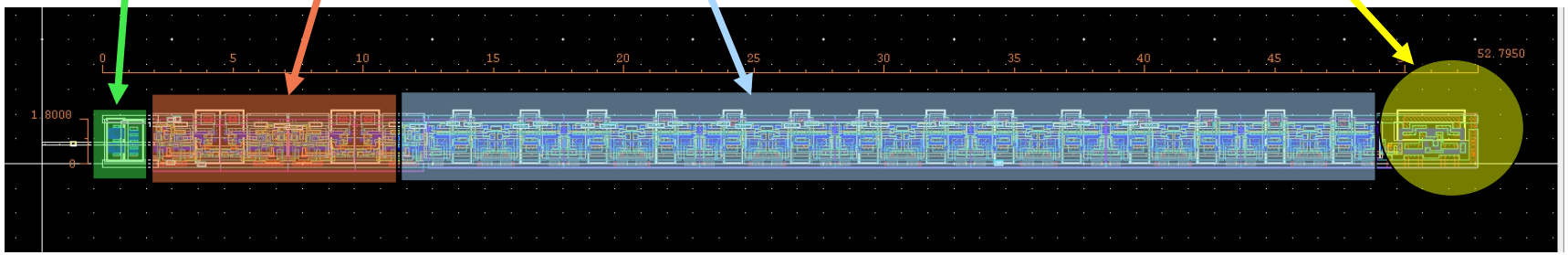
Associative Memory Layer

To save power we have used two different match line driving scheme:

- *Current race scheme (dummy layer timing)*
- *Selective precharge scheme*



Full Layer layout : w=~53µm X h=1.8 µm



Current race and selective - precharge schemes

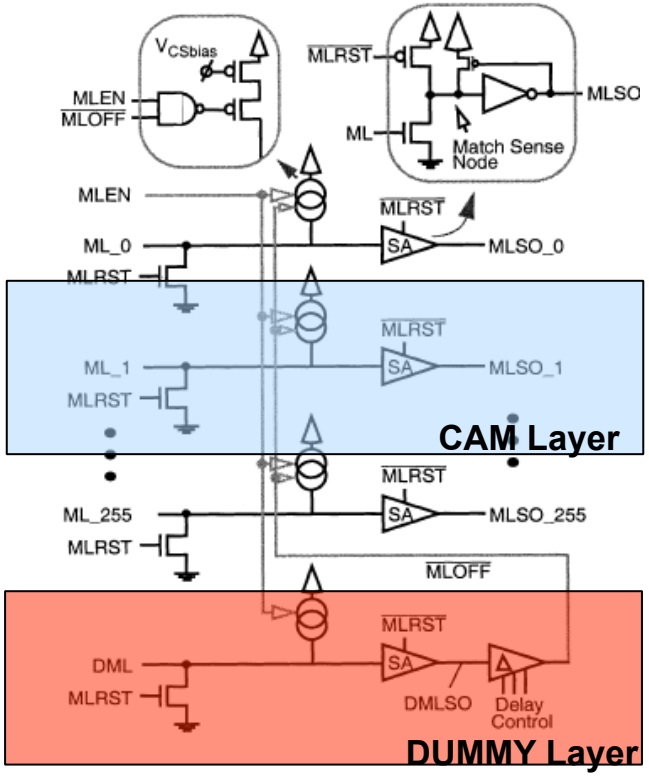


Fig. 5. Current-race ML sensing scheme.

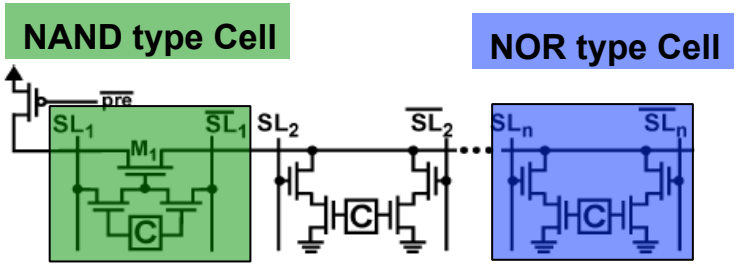


Fig. 16. Sample implementation of the selective-precharge matchline technique [43]. The first cell on the matchline is a NAND cell, while the other cells are NOR cells. Precharge occurs only in the case where there is a match in the first cell. If there is no match in the first cell, the precharge transistor is disconnected from the matchline, thus saving power.

Scheme from: "*Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey*", Kostas Pagiamtzis and Ali Sheikholeslami IEEE Journal of Solid-State Circuits, Vol. 41, NO. 3, March 2006

Scheme from: "*A ternary content-addressable memory (TCAM) based on 4T static storage and including a Current-Race sensing scheme*", Ali Sheikholeslamiet Al. IEEE Journal of Solid-State Circuits, Vol. 38, NO. 1, January 2003

Power consumption rough estimates

We use the nominal simulation condition:
Transistor models : Typical
Power supply : 1.2 V
Temperature : 27 °C
Frequency :100 MHz

Compatible with first measurements

These values do not take into account the standard cells part of the chip and the on chip power supply network distribution parasitic and other parasitic.

Memory state	Mean (mW)	Max (mW)	RMS (mW)
Write	23.04	557.57	28.12
Quiescent	21.89	22.09	21.88
Don't match	63.36	720.32	113.41
Match 1 out of 16 patterns	70.96	814.18	123.55
Match 1 out of 8 patterns	79.20	868.03	140.03
Match 1 out of 4 patterns	91.87	1045.44	172.34

Appoximate consumption 80mW / 8kpattern / 100MHz ~= 100μW / kpattern / MHz
+ plus standard cell logic

Ternary CAM Cell with two NOR type cells

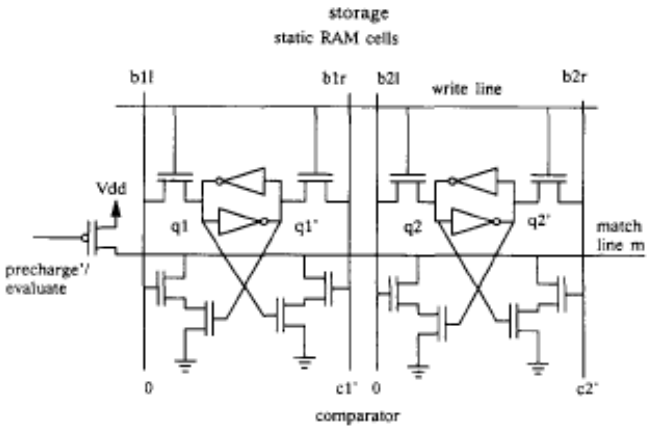


Fig. 8. Two adjacent static binary CAM cells.

storage scheme
stored values

	q1q2
0	01
1	10
*	00

(a)

retrieval scheme
presented values

presented ternary value	encoded value in the bit lines of two binary static CAMS. c1c2	b1l	b1r	b2l	b2r	binary CAM equivalent operation l	r
0	01	0	1	0	0	0	M*
1	10	0	0	0	1	M	0
*	11	0	0	0	0	M	M

*M is the masking of a bit operation common in commercial binary CAMS.

(b)

Fig. 9. Encoding and retrieval schemes for don't-care in two static binary CAM's cells with masking capability. (a) Encoding scheme. (b) Retrieval scheme.

Can use from 3 to 6 ternary cells per layer.
Variable resolution 1-8 up to 1-64.

Images from: "*Encoding Don't Cares in Static and Dynamic Content-Addressable Memories*", Sergio R. Ramirez-Chavez, IEEE Transactions on circuits and system-II: Analog and Digital Signal Processing, Vol. 39 NO. 8, August 1992

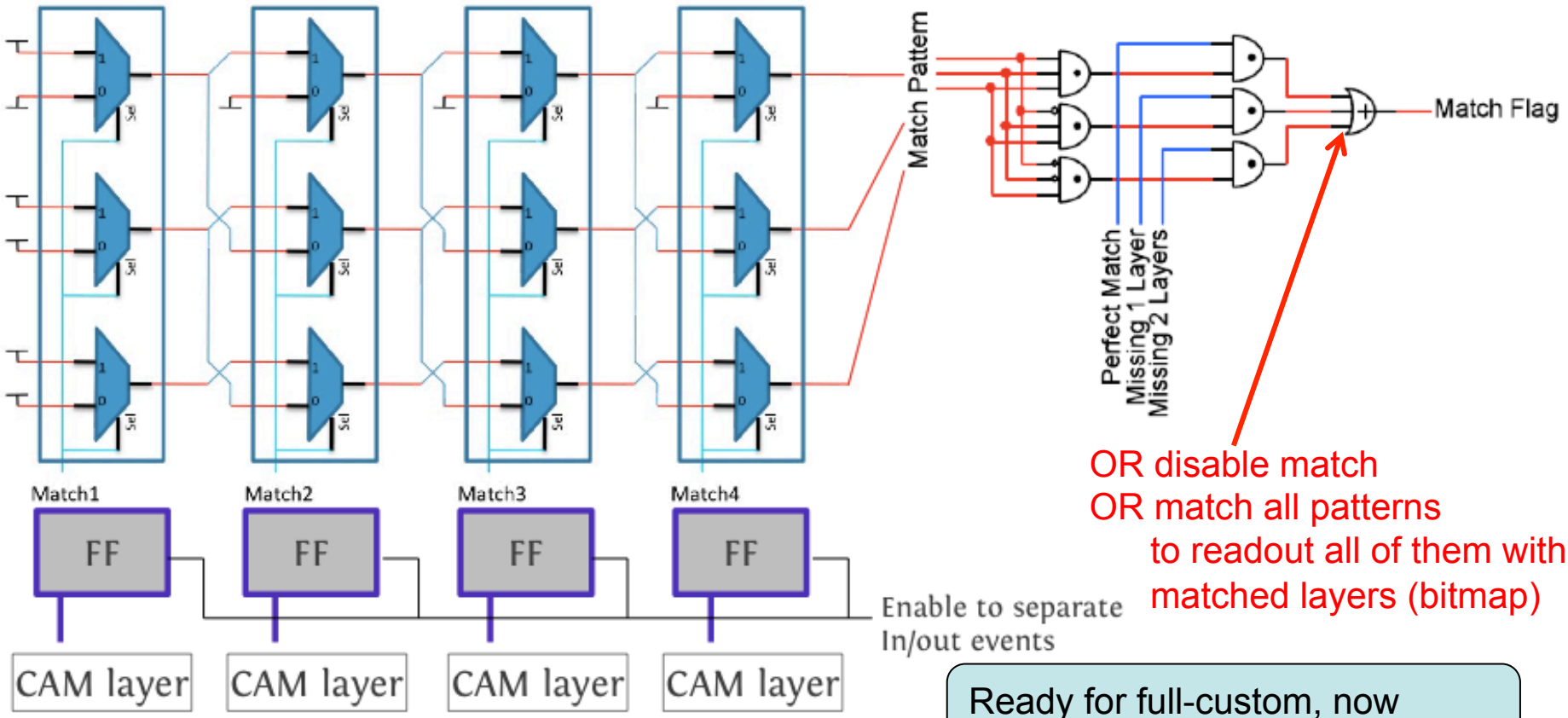
CAM cell configuration

- 18 CAM bits per layer: 4 NAND and 14 NOR
  - NOR pairs can make a ternary cell
- Default 12 bits + 3 ternary (minimum)
  - 15 bits per input bus (maximum)
  - (14:7) NOR, (6:3) 4 NAND, (2:0) 3 NOR-pairs
- 6 bits + 6 ternary (maximum)
  - Use only 12 bits per input bus
  - (11:10) NOR, (9:6) 4 NAND, (5:0) 6 NOR-pairs
- Ternary cells (NOR pairs) mapped to LSBs
- NAND cells are mapped to LSBs after the ternary cells, when they don't match small power consump.

AMCHIP04: MAJORITY LOGIC

Pattern match logic is made by identical logic for each layer:
Receives in input 3 bits, if not matching shift down the output.

At the end of the chain the 3 bits are compared with the majority requirements: perfect match, 1 or 2 missing



OR disable match
OR match all patterns
to readout all of them with
matched layers (bitmap)

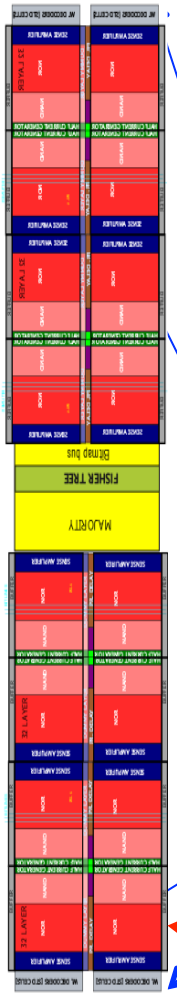
Ready for full-custom, now
syntethized with standard cells

Logic proposed by J. Hoff (Fermilab)

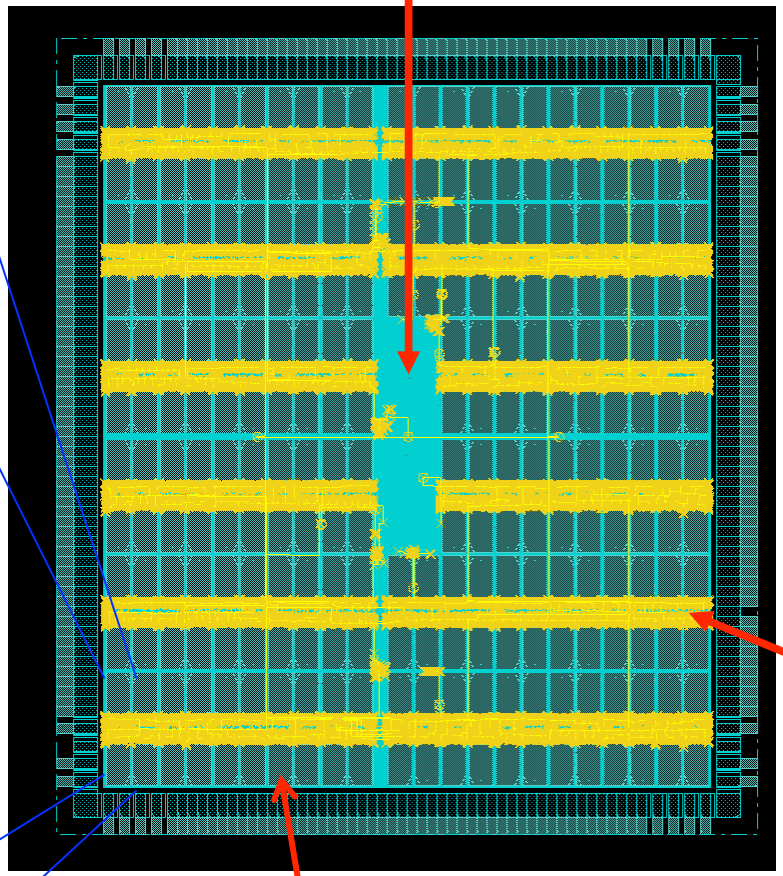
Layer matches from CAM layers are stored in a FF just before resetting the CAM layers.
We can load an event in the CAM layers while we are reading the patterns found in the previous event.

Prototype Chip Layout

64 patterns
x 8 layers



Control logic



The AMchip has an area of 14 mm²

CAM is organized as 22 column x 12 row matrix of full custom memory blocks

Each block is 64 x 4 layers

Between two rows of blocks there is the majority logic and the readout logic implemented with standard cells

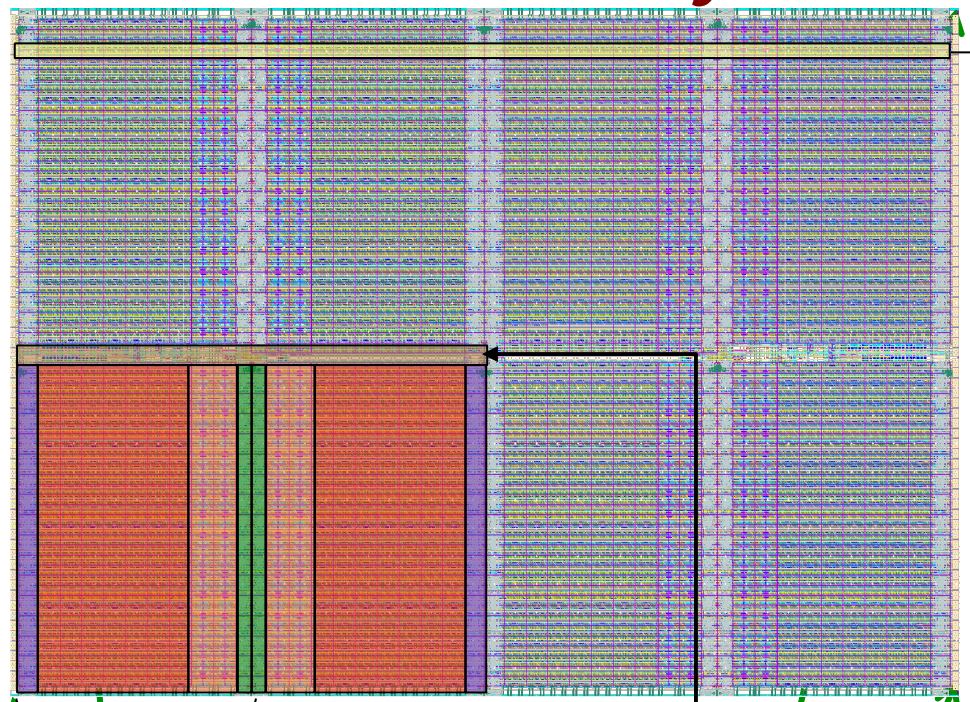
In the center there is the control logic implemented with standard cells

Majority logic and readout logic

size: 3510 μm $\times$ 3985.0 μm

2x128 blocks: 64 half patterns each

Memory Block Layout



4 Layers = 1/2 pattern

Full custom Layout of 64 x 4 CAM layers (half pattern):
w=~226 µm h=~123 µm
compare with pixel size

without including: majority logic, readout logic, and control logic

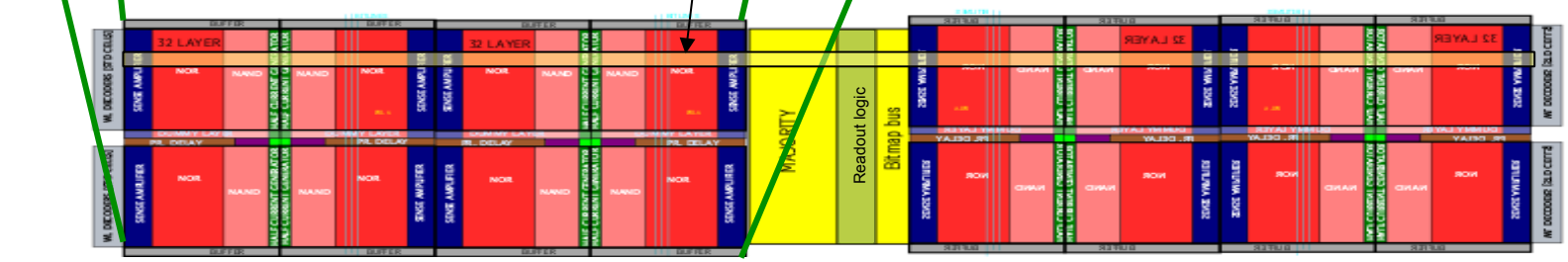
Six metal layers are used to route signals, power supply and ground.

Bit lines are routed vertically while control lines and memory output are routed horizontally

32 CAM layer memory block

Dummy layer

SINGLE PATTERN

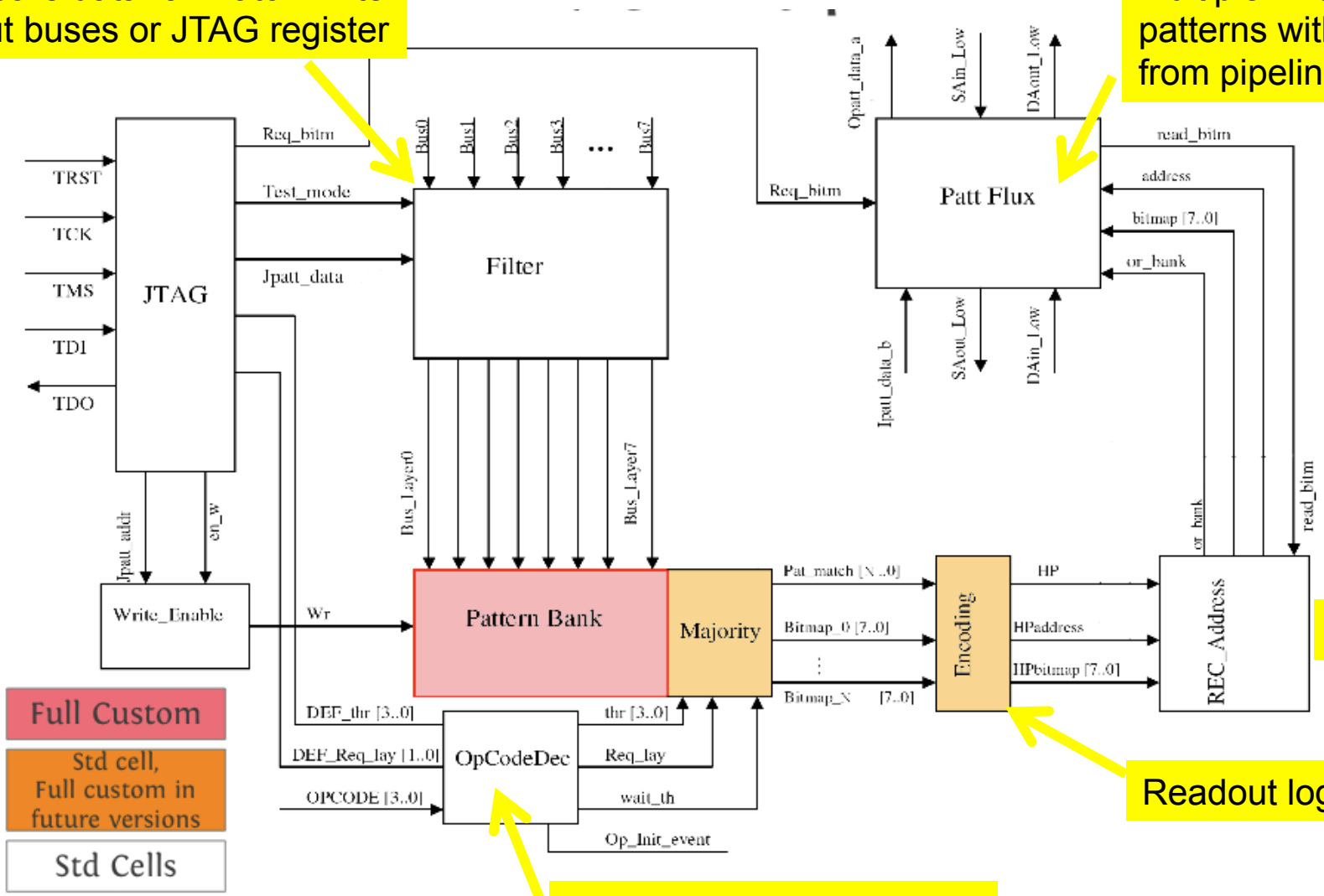


64 patterns (vertically)

AMchip TOP level

Prepare data for match/write
Input buses or JTAG register

Multiplex internal
patterns with those
from pipelined Amchips



register

Readout logic

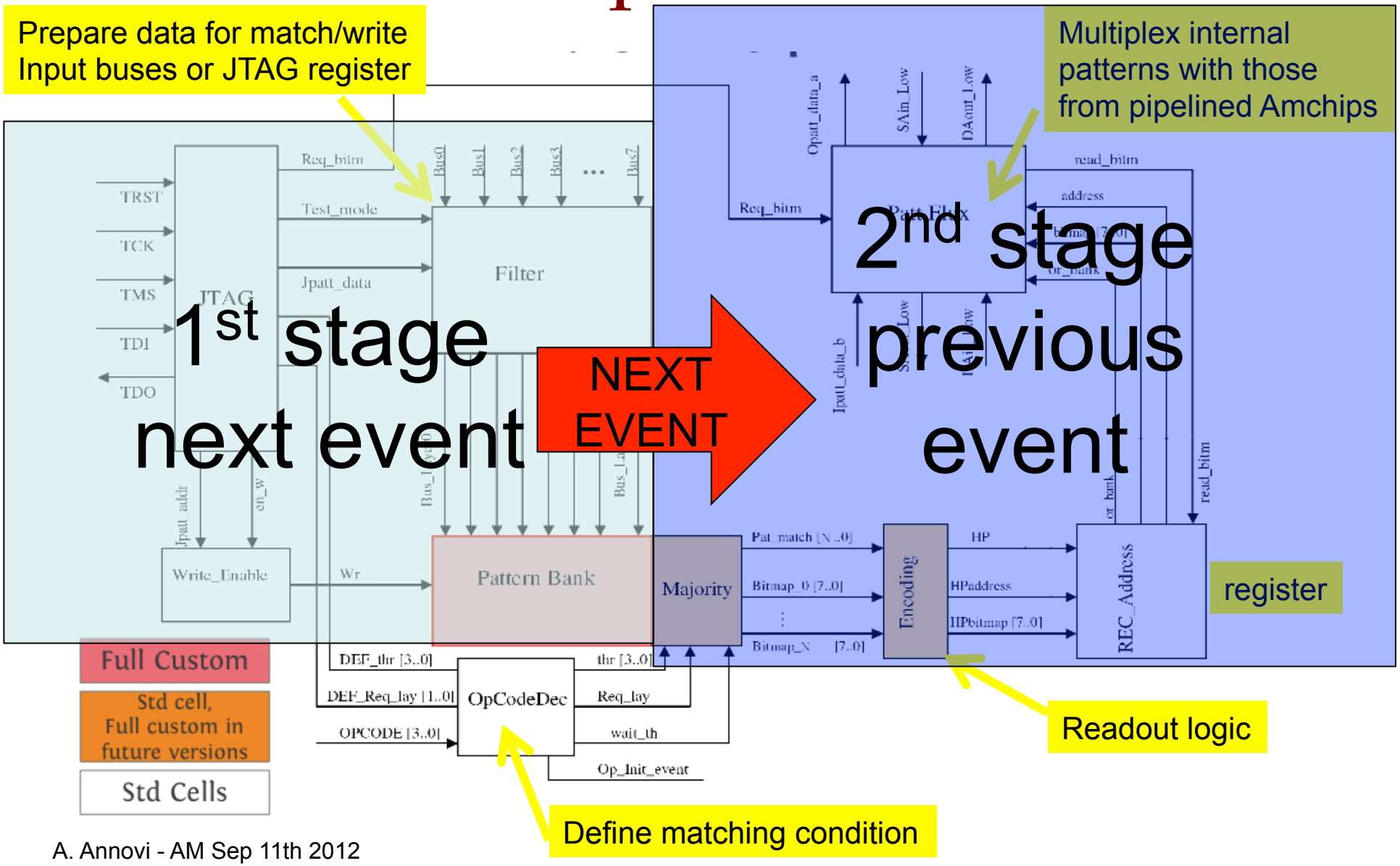
Define matching condition

Full Custom

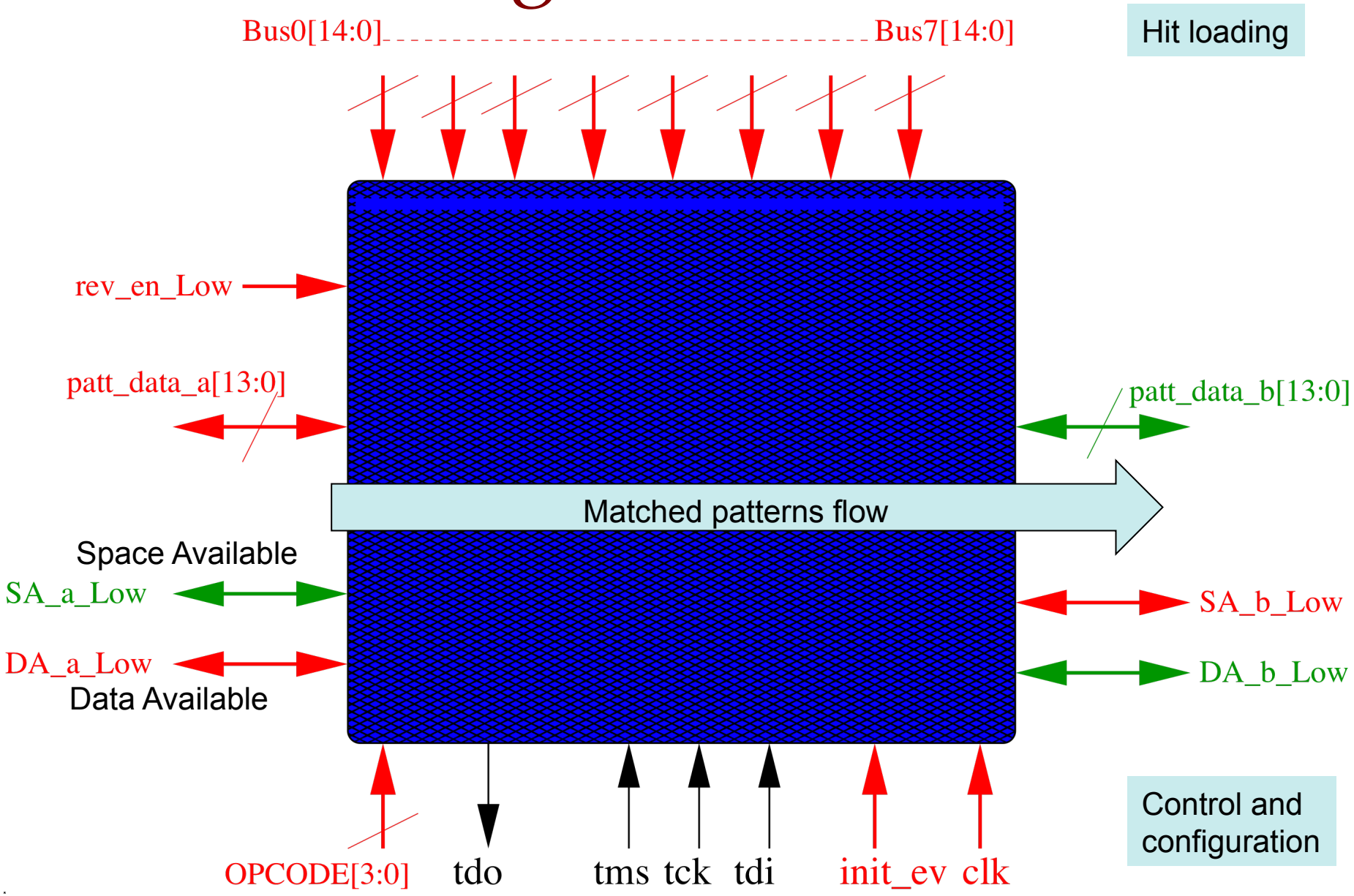
Std cell,
Full custom in
future versions

Std Cells

AMchip TOP level

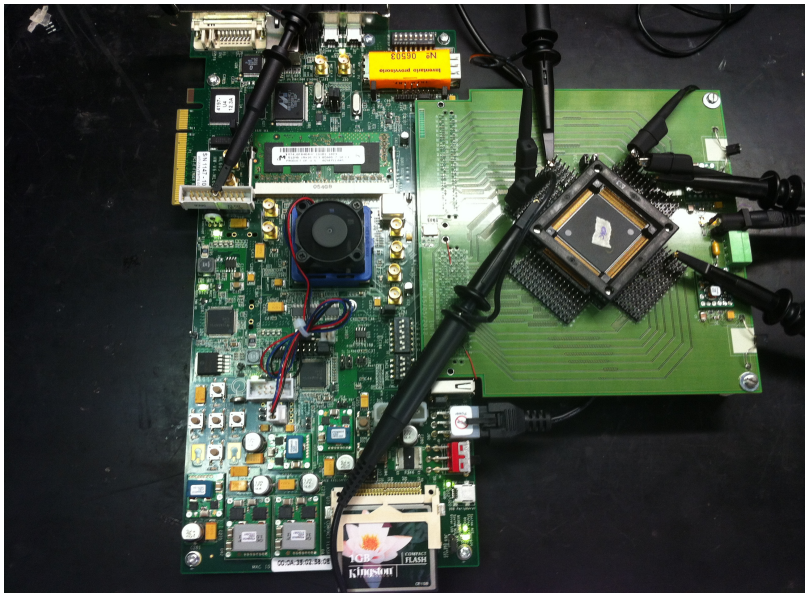


Logical Pinout



First AMchip 04 tests

- AMchip 04 tests in progress
- All test vectors passed sucessufully @ 50 MHz
- Some test performed and passed at 100 MHz
- Next: run all test vectors @100MHz
- Next^2: characterize the device power/speed/ yield



FTK AMchips plans

- AMchip04 (MPW)
  - 8k patterns with 3-6 ternary cells / layer (die 14mm^2)
  - 8x15 bits inputs up to 100MHz
  - 250mW core consumption @ 100MHz
- Miniasic (TBC) 2013
  - IO with 2Gbs serial link / layer
  - FlipChip BGA 21x21 mm^2 (TBC)
- MPW: AMchip05 (TBC) 2013
  - 32k patterns with 2Gbs/layer serial IO
  - FlipChip BGA 21x21 mm^2 (TBC)
- Full mask run: AMchip06 (TBC) 2015
  - 64k patterns with 2Gbs/layer serial IO
  - FlipChip BGA 21x21 mm^2 (TBC)

PQ208 30x30 mm^2

Schedule is aggressive missing contingency…

AMchip05

AMchip05

AMchip04-06 for L1 triggers

- Can we use AMchip04-06 for L1 triggers?
  - In principle yes would save a lot of time/money
  - But 65nm technology and specs choosen for FTK and 2015 production
- Is it a good idea?
  - Depends on your requirements
    - # patterns / chip, IO speed
    - # bits/layer, # layers
- Need to know your requirements now

Summary

- Designed and tested new Associative Memory
- Prototype main goal: verify functionality of new features and full-custom cell
- First application: ATLAS Fast-Tracker
- Special care to minimize power consumption & increase pattern density
- NEW: introduce powerful variable resolution pattern-matching !!!
  - any coincidence based trigger can profit
  - equivalent to a factor 3-5 extra patterns or more