TCAM Core Design in 3D IC for Low Matchline Capacitance and Low Power

Eun Chu Oh and Paul D. Franzon
ECE Dept., North Carolina State University, 2410 Campus Shore Drive, Raleigh, NC, USA 27606

ABSTRACT

Ternary Content Addressable Memory (TCAM) has been an emerging technology for fast packet forwarding, commonly used in longest prefix match routing. Large table size requirements and wider lookup table data widths have led to higher capacity TCAM designs. However, the fully parallel characteristic of TCAM makes large TCAM design more challenging and limits its capacity due to intensive power consumption. This paper proposes 3D IC technology as a solution to reduce the power consumption by reducing the interconnect capacitances of TCAM. In 3D IC, multiple wafers are stacked on top of each other, and the tiers are vertically connected through 3D vias. 3D vias reduce metal interconnect lengths and parasitic capacitances, resulting in power reduction. In this paper, 3D vias are used to replace matchlines, whose transition during parallel search operations is a major source of high power consumption in TCAM. An analysis of parasitic interconnect capacitance has been done using a quasi-static electromagnetic field simulation tool, Ansoft's Q3D Extractor, on a TCAM memory core in both conventional 2D IC structure and 3D IC structure with the process parameters of the MIT Lincoln Labs 0.18um FDSOI process. Field analysis and spice simulation results using a capacitance model for interconnects show that a 40% matchline capacitance reduction and a 23% power reduction can be achieved by using a 3-tier 3D IC structure instead of the conventional 2D approach.

Keywords: TCAM (ternary content addressable memory), 3D IC (3-dimentional integrated circuits), integrated circuit interconnect, vertical interconnect, inter-tier via, 3D via, matchline capacitance, interconnect capacitance, low power TCAM, CAM

1. TCAM INTRODUCTION

In content addressable memory (CAM), data is accessed based on its content rather than its location specified by an address. CAM, especially fully parallel CAM, provides fast search functionality that is widely used in various applications such as lookup tables, processor caches, translation lookaside buffers (TLB), data compression, databases, and artificial intelligence [1-3]. Recently, CAMs with larger capacities and higher speeds have been in demand in especially high speed network applications including packet routers using longest prefix matching routing algorithms. In such applications, packets need to be classified by the most specific match from a large routing table, with possible "x" (don't care) bits. Ternary CAM (TCAM) is more suitable than binary CAM for this application as TCAM is capable of storing and searching for "x" as well as "0" and "1;" each TCAM cell stores two bits to represent three states. Table 1 shows the encoding of the TCAM bit.

Table 1. Encoding of TCAM bit

B1	B2	B
0	0	"x"
0	1	"0"
1	0	"1"
1	1	-

Smart Structures, Devices, and Systems III, edited by Said F. Al-Sarawi, Proc. of SPIE
Vol. 6414, 641405, (2007) · 0277-786X/07/$15 · doi: 10.1117/12.695915

1.1 TCAM architecture

A general TCAM architecture consists of a memory core, an address decoder, a comparand data driver, a bitline data driver, a searchline (SL) pre-discharger, a matchline (ML) precharger, a ML sense amplifier, and a priority logic encoder, as shown in Fig. 1. In a data searching operation, the search data is sent to the TCAM core to compare the comparand data with all the stored data simultaneously, and the address of the matched data is sent to the output if there is a match. In case multiple matches are found, the priority logic encoder prioritizes the addresses and sends the one with highest priority. Due to the fully parallel characteristic of the TCAM search operation, power dissipation is high, and lower power TCAM still remains a challenging part of TCAM design.

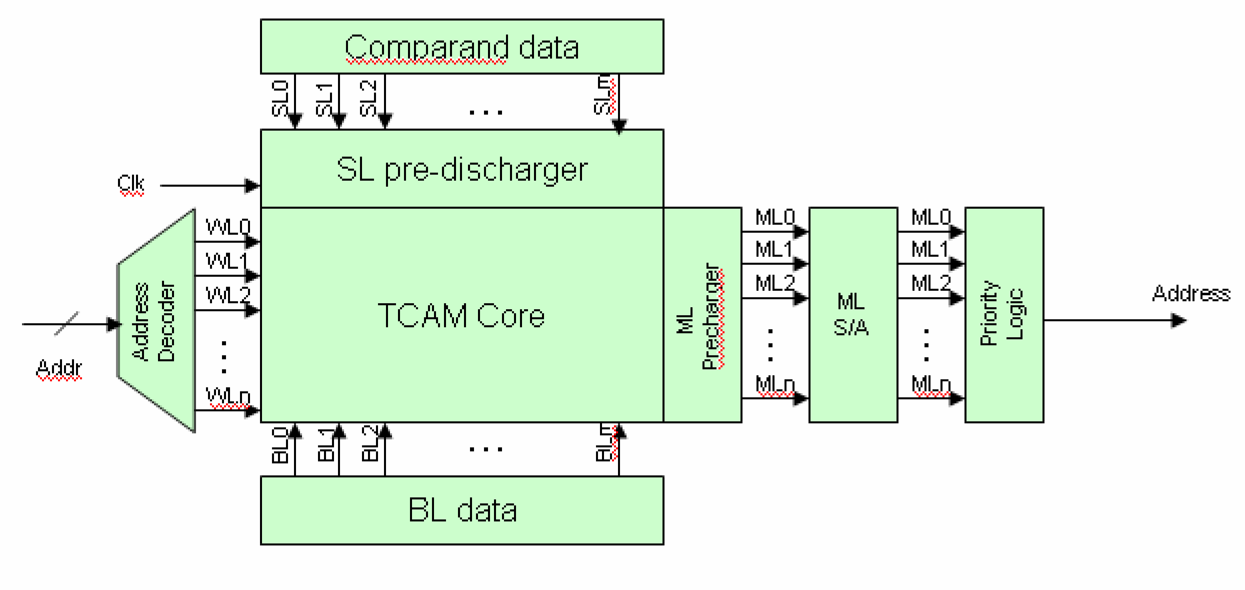


Fig. 1. Functional block diagram of typical TCAM

1.2 TCAM cell operation

The high power consumption of TCAM is largely due to the comparison function used in the search operation. The TCAM cell consists of a storage circuit part and a comparison circuit part used in the search operation as shown in Fig. 2. The TCAM core is implemented as an array of TCAM cells with horizontally running wordlines (WLs) and matchlines (MLs), and vertically running bitlines (BLs) and searchlines (SLs) as shown in Fig. 3. All the cells in the same row share the same WLs and MLs, and all the cells in the same column share the same BLs and SLs.

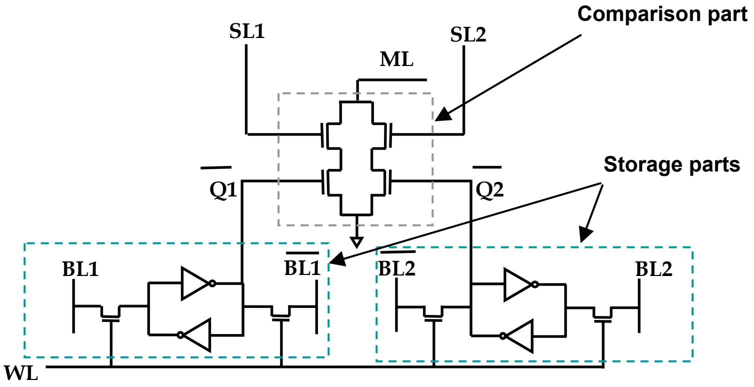


Fig. 2. Conventional 16T TCAM cell

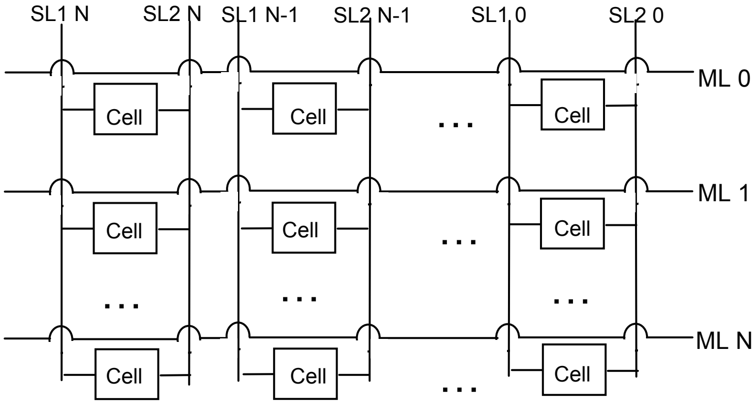


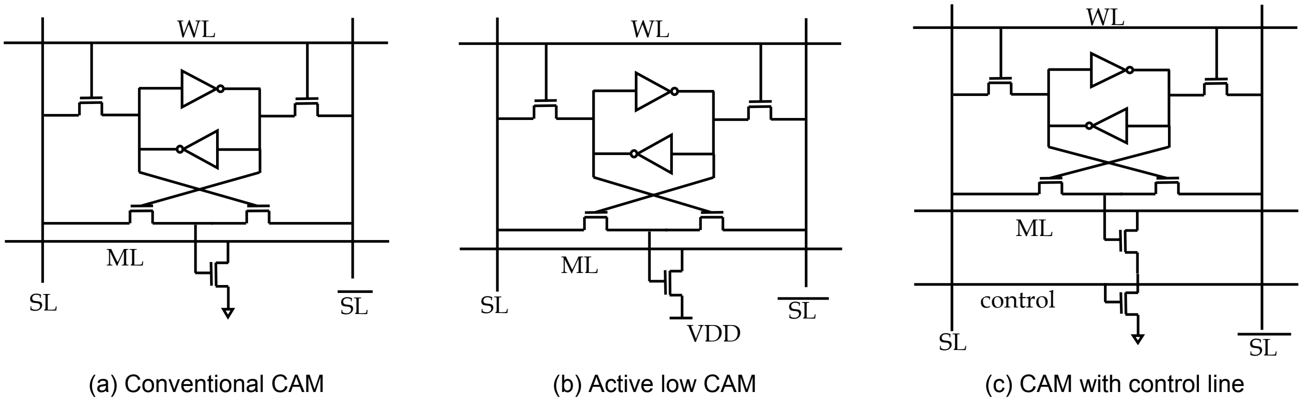
Fig. 3. TCAM core

TCAM write operations are implemented in a fashion similar to SRAM implementations. When WL is high, data is sent through BLs and stored in a latch. Search operations are performed when WL is low. For fully parallel TCAM, search data is broadcast through SLs, and each TCAM cell compares the search data against the stored bit. Before every search, MLs are precharged high. If the word matches, ML remains high indicating that the data is stored in the CAM. It is considered a word match when all the cells in the word match. If any of the cells in the word do not match, the word is considered a mismatch and ML discharges to ground through the comparison pass logic, which is the case for most of the words in lookup table as only a few words in the table match with the search data. Also, it should be noted that SLs are predischarged to ground before every search to avoid static current through comparison logic. In fact, approximately 50% of the SLs are charged back during evaluation. The frequent transition of MLs and SLs with large capacitances is the major source of power consumption in TCAM [7]. In this paper, we take the 3D IC approach to reduce ML capacitance, the part of the system that consumes the most power, to reduce the power consumption.

1.3 Low power CAM cell

Power consumption is a limiting factor in the capacity of TCAM [4]. There have been a number of efforts to achieve low power CAM. In this section, a few different circuit techniques are discussed. Note that these are not TCAMs, but the same techniques can be applied to TCAM as well.

In active-low CAM, shown in Fig. 4(b), ML is pre-discharged instead of being precharged by replacing the gnd with vdd on the pull down logic. This may result in low power consumption due to the lower voltage swing of ML [10]. Also, to reduce the searchline power consumption, an additional transistor which controls the matchline without any dependency on the searchline can be added to the discharge path of ML, as shown in Fig. 4(c) [12]. This incurs an area penalty due to the additional transistor. Another way to reduce power consumption is selective search, which performs partial matching with a small subset of the input data first, and then does the complete matching with the partially matched rows only. The CAM cell would be the same as the conventional cell, shown in Fig. 4(a). As most of the rows are not matched, this would save power, but there is a delay penalty as the search becomes a serial process [11].



(a) Conventional CAM

(b) Active low CAM

(c) CAM with control line

Fig. 4. Low power CAM cells

This paper proposes 3D IC technology as a solution to reduce the power consumption by reducing the interconnect capacitances of TCAM. The lower power circuit technique for TCAM cells discussed above can be applied as a circuit technique with conjunction to the 3D IC technology. For simplification purposes, the conventional TCAM cell design is chosen for the analysis in 3D IC.

2. 3D IC TECHNOLOGY

As device size scales down, wires are packed closer to each other and long interconnect, with its high parasitic capacitance, becomes more significant as a limiting factor to the power consumption and performance of the chip [5]. This paper proposes three dimensional integrated circuit (3D IC) technology as a solution to reduce the power consumption by reducing the interconnect capacitances of TCAM. In 3D IC, multiple dice are stacked on top of each other with direct vertical interconnection using 3D vias. Highly capacitive interconnects can be replaced by the 3D inter-tier vias to achieve low power design.

The 3D IC process developed by Massachusetts Institute of Technology Lincoln Laboratory (MITLL) offers three tiers of 0.18um fully depleted silicon on insulator (FD SOI) circuit fabrication. The 3D circuit integration technology offers higher density vertical interconnection, better circuit-to-interconnect ratio, and lower power consumption [6]. In this 3D process, the base tier, Tier 1 is on an FD SOI substrate, and Tier 2 is flipped, aligned and bonded to Tier 1 having the Si substrate removed, as shown in Fig. 5. Tungsten 3D via interconnects the top-level metal layer of Tier 1 and the top-level metal layer of Tier 2. Tier 3 is assembled in a similar way to Tier 2 except that the 3D vias connect the first-level metal layer of Tier 2 and the top-level metal layer of Tier 3.

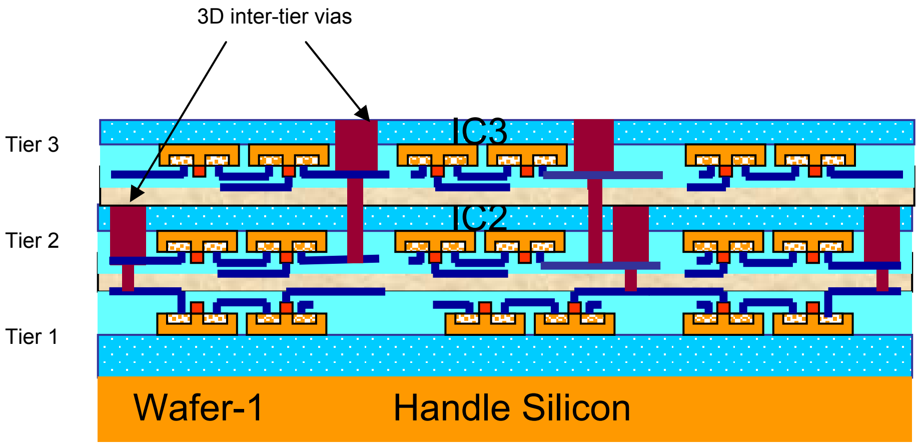


Fig. 5. 3D circuit integration [6]

In this process, tungsten is deposited to form 3D vias. This method of implementing the vias makes them larger than typical vias. The resulting 3D vias are 1.5 um x 1.5 um and 8 um tall, and their capacitance is comparable with metal capacitance. For instance, the parasitic capacitance of metal 1 on Tier 1 with the width of 1 um and length of 10 um would be similar to the coupling capacitance between two 3D inter-tier vias.

Thermal consideration in 3D circuitry is a challenging topic that is still being actively researched. In general, circuits on Tier 1 would have a better thermal conductivity as Tier 1 has the silicon substrate attached, which can be attached to an IC package for heat conduction. However, circuits on Tier 2 and Tier 3 have lower thermal conductivity with Tier 3 having the lowest thermal conductivity. Thus high power circuits on Tier 3 will have the tightest thermal budget, so putting circuits with less power dissipation on Tier 3 is recommended. Also, thick back-metal layers on Tier 2 and Tier 3 can help to extract the heat generated on those tiers through tungsten plugs.

3. DESIGN METHODOLOGY

3.1 Partitioning methodology of circuits to multiple tiers

A new issue to be considered for circuit designs in 3D IC is how to partition the design into a number of design parts to be allocated on different tiers. This is an important issue and the advantage of using 3D IC may greatly depend on how to partition the circuits or systems. There are three ways to partition for a system with arrays, like TCAM: (a) partitioning within a small building block or a cell, (b) partitioning by unit of a large functioning block, and (c) partitioning by unit of a cell. Which design partitioning approach is most appropriate depends on the type and the characteristics of the circuit or system.

The first partitioning method, partitioning within the cell, breaks down the cell into *n* number of parts for the process providing *n*-tiers. In the case of TCAM with MITLL 3D process, the cell can be broken down into three parts: a storage part, a comparison part, and a second storage part**.** This would reduce the width of TCAM cell circuit by approximately one third, from 20 um to 7 um, shortening the ML metal wire required by one third as well. However, with MITLL 3D IC technology, 3D vias require as much as 4um (width) by 8um (length) by 8um (height) including the metal wire required and the additional parts on the top of the via required by the process, leaving the reduction in ML length less significant. Partitioning methodology within the cell might be a good approach for larger cells or for processes with smaller 3D vias. Due to the large ratio of the dimension of the 3D via with respect to the dimension of the cell, this partitioning method is not as beneficial for a TCAM array.

The second partitioning method, partitioning by unit of block, breaks down the system into *n* number of parts for *n*-tier 3D process. In this case, 3D vias can replace very long wires that interconnect two different blocks at a distance. The benefit in term of capacitance could be large for a particular wire. The benefit would also be great if critical path can be replaced. Yet, the benefit may not be as significant if only non-critical paths are replaced or if the system is already optimized for interconnect length. This partitioning method would be more appropriate for analysis of a complete system. This approach was not considered for matchline capacitance analysis as it focuses on the CAM memory core only, where the significant power consumption of CAM occurs.

The last partitioning method, partition by unit of a cell, breaks down the array into the smallest unit of the array and has the cells allocated on three tiers, as shown in Fig. 6. This is simple and easy to layout. Cell layout can be done on one of the tiers (for example, Tier 1) in 3D IC in a way similar to the way it is done in conventional 2D IC design. The layout is then copied to other layers (i.e. to Tier 2 and Tier 3) using software such as Cadence SKILL™. This way cells are aligned among the three tiers, which makes inter-tier interconnecting a lot simpler. This partitioning method was selected for our TCAM memory array layout in MITLL 3D IC considering the TCAM cell size.

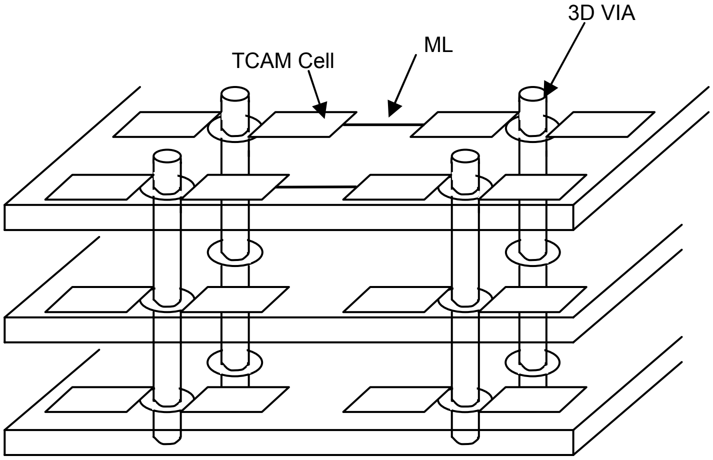


Fig. 6. Three dimensional view of TCAM array in three-tier process

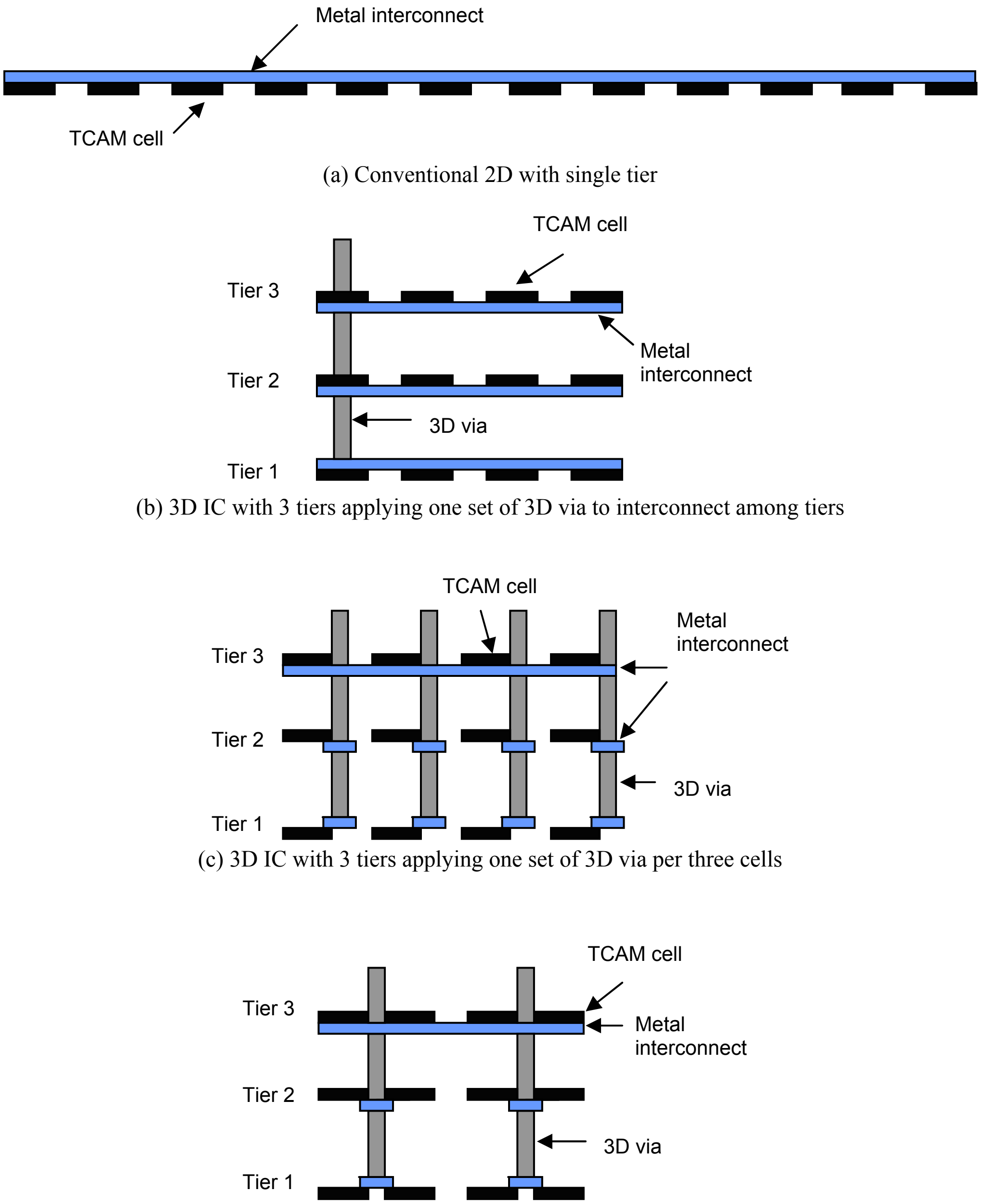
3.2 Matchline layout methodology

Another important factor in 3D design is how 3D vias are placed and what they actually are replacing. Fig. 7 shows cross-sectional views of various matchline layout methods for a 12-bit word. Fig. 7(a) shows the 12-bit TCAM word on the conventional two-dimensional process with a single-tier. ML of every cell in the same word is connected by planar metal interconnect.

The simplest way to convert this 2D design into 3D design with 3 tiers is to partition the ML interconnect into three parts and connect the tiers with a single 3D via set (3D via connecting Tier 1 and Tier 2, and 3D via connecting Tier 2 and Tier 3), as shown in Fig. 7(b). However, it is obvious that this would increase the matchline capacitance as the matchline is only added by additional 3D vias without any reduction.

Another way to connect the ML of each cell is to connect the MLs of only the cells on Tier 3 with metal interconnect and connect the cells on Tier 1 and Tier 2 using 3D vias as shown in Fig. 7(c). As long as the capacitance of 3D via is smaller than the capacitance of metal wire with the length of cell width, 20um, 3D design is beneficial in reduction of ML capacitance.

To reduce the ML interconnect capacitance even further, 3D Vias can be shared by two adjacent cells as shown in Fig. 7(d). In this case, the number of 3D vias required is reduced by one half compared to the ML layout in 7(c).



(a) Conventional 2D with single tier

(b) 3D IC with 3 tiers applying one set of 3D via to interconnect among tiers

(c) 3D IC with 3 tiers applying one set of 3D via per three cells

(d) 3D IC with 3 tiers applying one set of 3D via per six cells

Fig. 7. Cross-sectional views of matchline layout methods for a 12-bit word

3.3 Cell layout methodology

A careful layout can also contribute to reduce the interconnect length and capacitance. A typical TCAM cell layout has two SRAM storage circuits and the comparison logic circuit in between the SRAM storage circuits, as shown in Fig. 8(a). However, this way extra ML interconnect length is required to connect the comparison logic and 3D via as shown in Fig. 8(b). In the new TCAM cell layout, shown in Fig. 9(a), the comparison logic circuit is placed next to two SRAM storage circuits for more efficient sharing of 3D via by two adjacent TCAM cells and to minimize the ML interconnect length.

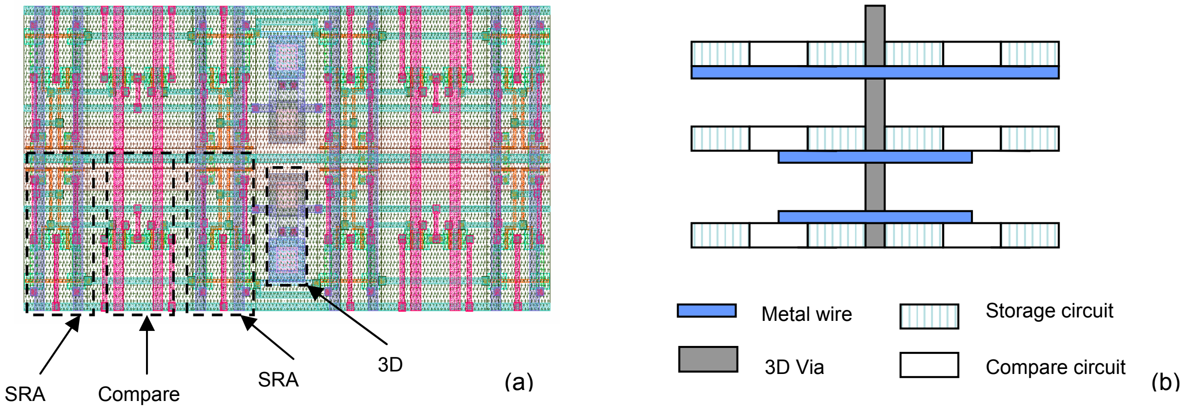


Fig. 8. Layout of 2w x 6b TCAM array in 3D IC with (a) conventional TCAM cell layout and (b) cross sectional view

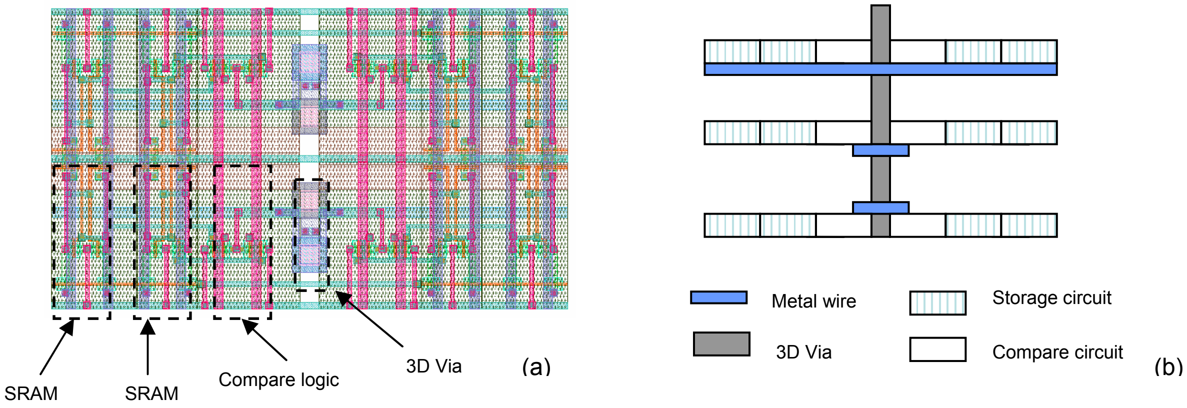


Fig. 9. New layout of 2w x 6b TCAM array in 3D IC with shorter ML interconnect length (a) TCAM cell layout and (b) cross sectional view

4. SIMULATIONS AND RESULTS

For accuracy of parasitic interconnect capacitance analysis, simulations were done using Ansoft's Q3D Extractor, a quasi-static electromagnetic field simulation tool that computes capacitance using the method of moments and the finite element method [9]. For a fair comparison, TCAM memory core in both conventional 2D IC structure and 3D IC structure with identical cell layouts are simulated, except that 3D design allows more degrees of freedom for spacing between interconnects due to the area saved by going vertical. In fact, even with the sparsely packed interconnects in 3D design, the footprint area of the TCAM core in 3D IC using three tiers is about a half of that in 2D IC with a single tier. It is not one third due to the area occupied by 3D vias themselves.

Interconnects of TCAM arrays in the sizes of 2w x 6b, 2w x 12b, and 2w x24b have been analyzed for the capacitance with the process parameters of the MIT Lincoln Labs 0.18um FDSOI process. To obtain the graphical models, the layouts were converted to visual basic files that can create the graphical models, shown in Fig. 10 and Fig. 11 for the various array sizes in 2D design and 3D design, respectively.

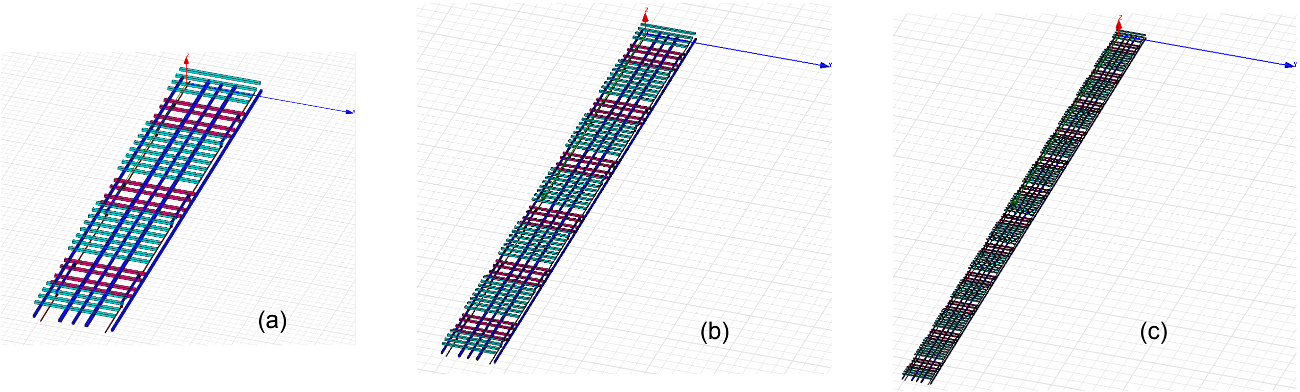


Fig. 10. Q3D model for ML of TCAM arrays in conventional single tier 2D process in a)2w x 6b, b) 2w x 12b, and c) 2w x 24b

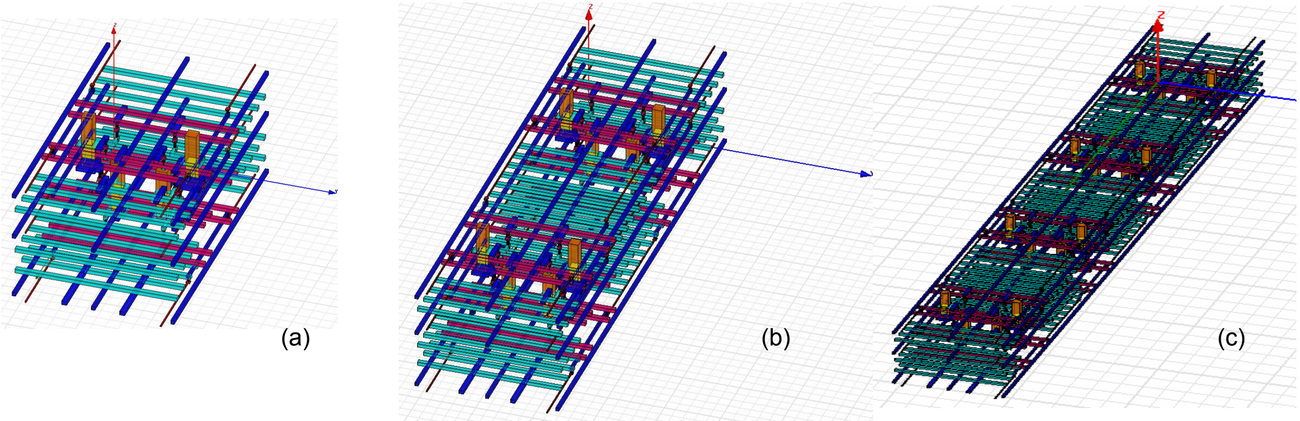


Fig.11. Q3D model for ML of TCAM arrays in three tier 3D process in (a) 2w x 6b, (b) 2w x 12b, and (c) 2w x 24b

Q3D Extractor generates capacitances among all possible combinations of interconnects. The resulting total ML capacitance is shown in Fig. 12. Field analysis using a capacitance model for interconnects shows approximately 40% matchline capacitance reduction consistently for all three arrays, 2w x 6b, 2w x 12b, and 2w x 24b. It makes sense that as the array size doubles, the interconnect dimension doubles, thus, the interconnect capacitance doubles as a result.

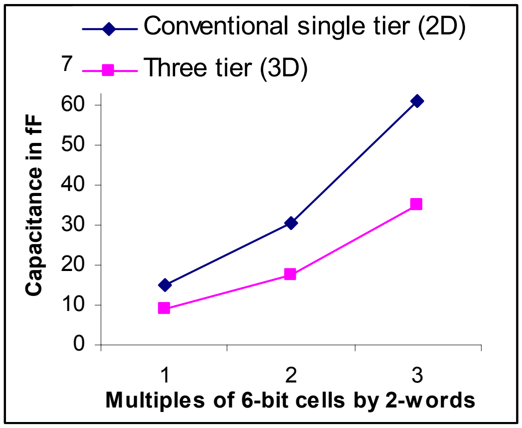


Fig. 12. Total ML capacitance from Q3D simulation results

From the parasitic capacitances extracted by the field simulator, capacitance models for each interconnect are constructed and used for HSPICE simulation for power analysis. To reduce the complexity of the model, parasitic capacitances less than 0.1 fF were neglected. The capacitance models of ML for 2w x 6b in single tier process and three-tier process are shown in Fig. 13 and Fig. 14, respectively. In the single-tier 2D design, the parasitic capacitance due to the substrate dominates over the coupling capacitance with other interconnects, while coupling capacitance dominates over the parasitic capacitance with the substrate in three-tier 3D design. In 3D design, the interconnects on the Tier 2 and Tier 3 are a lot further from the substrate. Thus, only the interconnects on the bottom tier contribute to the parasitic capacitance with the substrate significantly, and ML couples strongly with the neighboring interconnects as shown in Fig. 14.

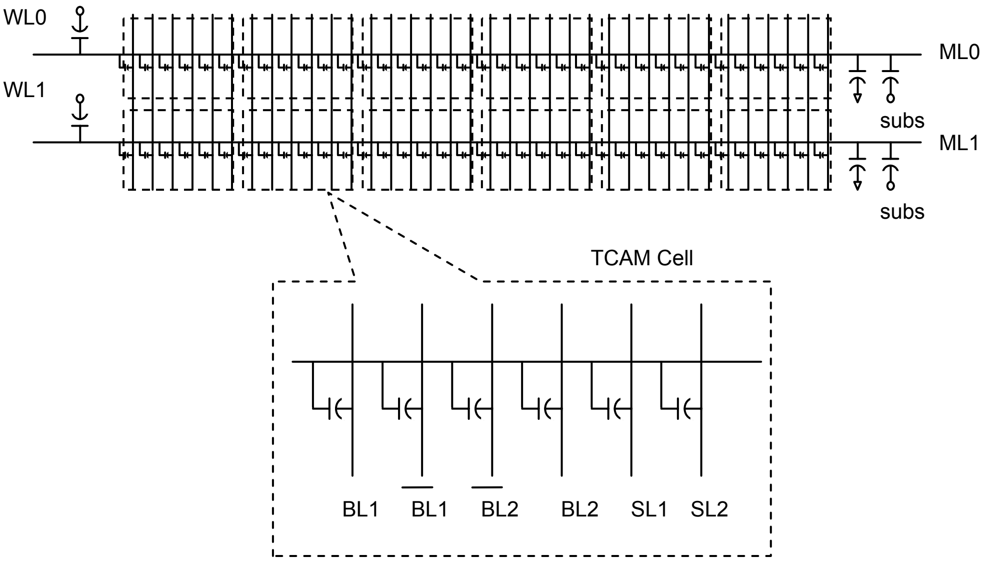


Fig. 13. Capacitance model for ML in 2w x 6b TCAM array in 2D

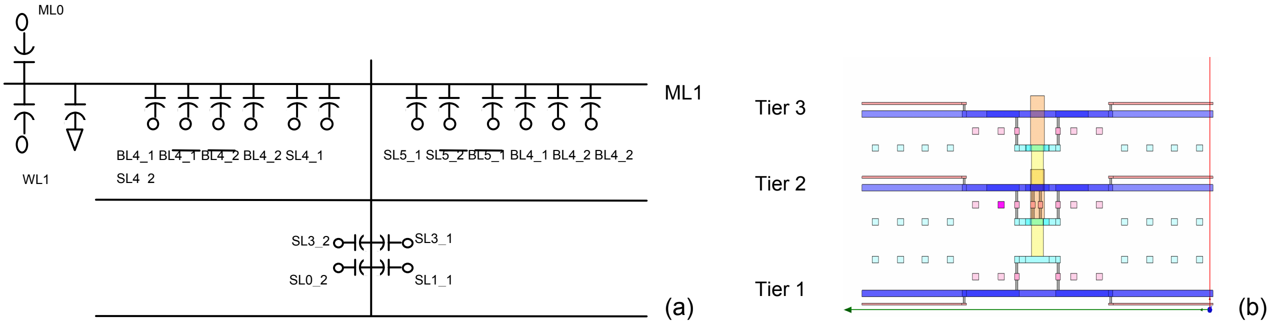

ML0
ML1
WL1
BL4_1 $\overline{\text{BL4\_1}}$ $\overline{\text{BL4\_2}}$ BL4_2 SL4_1
SL4 2
SL5_1 $\overline{\text{SL5\_2}}$ $\overline{\text{BL5\_1}}$ BL4_1 BL4_2 BL4_2
SL3_2
SL3_1
SL0_2
SL1_1
(a)
Tier 3
Tier 2
Tier 1
(b)

Fig. 14. Cross sectional view of ML in 2w x 6b TCAM array in 3D (a) Capacitance model (b) Q3D model

Energy consumption and power dissipation have been analyzed for 2w x 6b TCAM core arrays in both 2D structure and 3D structure during search operation according to equations (1) and (2). A number of different data patterns were written and searched for, and the value has been averaged over various combinations for tens of clock cycle.

$$E = \int i \cdot v dt \tag{1}$$

and

$$P = \frac{\int i \cdot v}{T} dt \text{ , where } T \text{ is the total simulation time.} \qquad (2)$$

HSPICE simulation results show that power dissipation through ML is approximately 60% of the total power consumption. Energy consumption and power dissipation by ML only and the total power dissipation are shown in Table 2. With 40% less ML capacitance, 28% ML power reduction was achieved and 23% total power reduction was achieved by using a three-tier 3D IC structure instead of the conventional 2D approach.

Table 2. Energy consumption for 65ns and power dissipation of 2w x 6b TCAM core at 1GHz

	Conventional single-tier IC (2D)	Three Tier 3D IC	Power reduction in 3D
Energy consumption by ML	2.9 pJ	2.1 pJ	-
Total Energy consumption	8.0 pJ	6.2 pJ	-
Power dissipation by ML	44.6 uW	32.3 uW	28 %
Total power dissipation	123 uW	95.4 uW	23 %

In this paper, only TCAM cores in single-tier and three-tier configurations are considered. Using more design tiers such as four or five tiers, the power reduction as well as the capacitance reduction of ML might be greater [8]. Also, an improvement in the process technology to reduce the size of the 3D inter-tier vias would decrease the interconnect capacitance and power dissipation as well.

5. CONCLUSIONS

With 3D IC technology, lower interconnect capacitance and low power design can be achieved using vertical inter-tier vias. In this paper, TCAM cores for longest prefix matching have been designed using 3D IC fabrication based on the MITLL FDSOI three-tier process and compared with a TCAM core done in a conventional single-tier 2D process. In the 3D IC design the matchlines, among the most power consuming components of TCAM designs, have been replaced by vertical inter-tier interconnects to reduce interconnect capacitance. The interconnect capacitances in both the conventional 2D circuit and 3D circuit were extracted using Ansoft's Q3D extractor, showing that up to 40% matchline capacitance reduction can be achieved in 3D IC. The power consumption of the TCAM array has also been evaluated in HSPICE with the interconnect capacitance models from Q3D, showing up to 28% reduction in ML power dissipation and 23% reduction in TCAM core power dissipation in 3D IC, compared to the TCAM array in the conventional 2D.

REFERENCES

1. F. Shafai, K. J. Schultz, G. F. R. Gibson, A.G. Bluschke, and D.E. Somppi, "Fully parallel 30-MHz 2.5-Mb CAM," IEEE J. Solid state Circuits, vol. 33, pp. 1690-1696, Nov. 1998.
2. J. P. Wade and C. G. Sodini, "A ternary content addressable search engine," IEEE J. Solid state Circuits, vol. 24, pp. 1003-1013, Aug. 1989.
3. T. Ogura, J. Yamada, S. Yamada, and M. Tan-no, "A 20-kbit associative memory LSI for artificial intelligence machines," IEEE J. Solid state Circuits, vol. 24, pp. 1014-1020, Aug. 1989.
4. H. Miyatake, M. Tanaka and Y. Mori, "A design for high-speed low-power CMOS fully parallel content-addressable memory macros," IEEE J. Solid state Circuits, vol. 36, pp. 956-968, June 2001.
5. J. Baliga, "Chips go beyond vertical [3D IC interconnection]," Spectrum, vol. 41, pp. 43-47, Mar. 2004.

6. MITLL low-power FDSOI CMOS process: application notes 3D FDSOI design, revision 2004:1, Dec. 2004.

7. D. Perry and P. Gillingham, “Ternary CAM cell for reduced matchline capacitance,” United States Patent Application Publication, Pub. No. US 2005/0276086, USPTO Class 365, Dec. 2005.

8. J.W. Joyner and J. D. Meindl, “Opportunities for reduced power dissipation using three-dimentional integration,” IEEE international interconnect technology conference, pp. 148-150, June 2002.

9. http://www.ansoft.com/products/si/q3d_extractor

10. G. Thirugnanam, N. Vijaykrishnan, and M. J. Irwin, “A novel low power CAM design,” IEEE international ASIC/SOC conference, pp. 198-202, Sept., 2001.

11. C. Zukowski and S. Wang, “Use of selective precharge for low power content addressable memories,” IEEE international symposium on circuits and systems, June 1997, pp. 1778-1791.

12. T. Jamil. “RAM versus CAM,” IEEE Potentials, pp.26-29, April 1997.