A Low-Power Monolithically Stacked 3D-TCAM

Mingjie Lin and Jianying Luo
Department of Electrical Engineering
Stanford University, CA 94305
{mingjie, jyluo}@stanford.edu

Yaling Ma
Department of Mechanical Engineering
Clemson University, SC 29631
yalingm@clemson.edu

***Abstract*—This paper presents three techniques to reduce the power consumption in ternary content-addressable memories (TCAMs). The first technique is to use newly developed monolithically stacked 3D-IC technology for the implementation, because vertical stacking can drastically reduce interconnect length in both matchlines and searchlines, hence reducing signal path delay and power consumption. The second technique is to replace the conventional SRAM memory in a TCAM with an array of programmable vias (or electrolyte non-volatile memory). Special programming circuitry is designed to read/write memory bits from/to the programmable via array because they do not simply store data in the form of low and high voltage levels. We also devised a new TCAM cell design to further reduce power consumption in TCAMs by taking full advantage of 3D-IC technology. A 1024×144-bit TCAM using the proposed schemes is implemented with 1.0-V 65nm CMOS technology. Our analysis and simulations have shown that the proposed monolithically stacked 3D-TCAM can reduce the total dynamic power consumption by almost 3.5 times and increase TCAM cell density by about 4 times in comparison with a conventional 2D-TCAM chip of the same capacity.**

I. Introduction

The dedicated comparison circuitry in TCAMs as illustrated in Fig. 1 significantly increases both silicon chip area and power consumption, two design parameters that designers strive to reduce. The power problem is further exacerbated by the widening demand for large-capacity TCAMs and the CMOS technology scaling. Despite of many successes, reducing power consumption without sacrificing speed or area remains a serious challenge in designing large TCAMs.

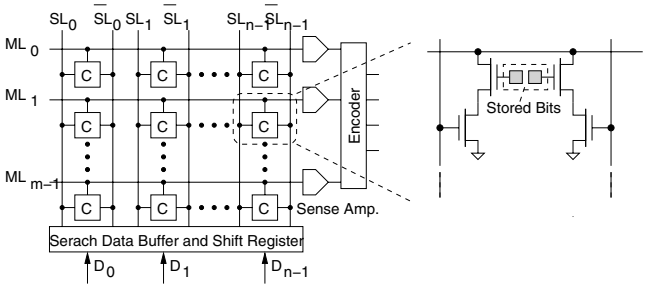

Fig. 1. Schematic of a small TCAM example.

The high power consumption in a TCAM results from the frequent signal toggling in searchlines and matchlines during each search clock cycle, which can be modeled as

$$P_{\text{total}} \approx \alpha C_{\text{total}} V_{\text{DD}} V_{\text{swing}} f, \tag{1}$$

where α represents the switching activity. Equation (1) clearly indicates that only a small number of parameters can be varied in order to reduce power. Specifically, the capacitive loadings to the matchlines and searchlines, the necessary voltage swing, and the transition activity factor should all be minimized. Many schemes, typically involving some optimizations of matchline and/or searchline structures, have been proposed to reduce power consumption in TCAMs. Notable examples include reducing the voltage swing on the matchlines, using current-based techniques to indirectly reduce the match-line voltage swing, pipelining matchline, selectively precharging matchlines, and bank-selection schemes [1], [2], [3], [4].

One conceptually appealing approach to reduce the dynamic power consumption of TCAMs is to stack the memory bit array on top of the searching circuitry that would be implemented with multiple active layers of conventional CMOS technology. One such example is depicted in Fig. 2. Besides the obvious benefit of denser TCAM cells, vertical stacking drastically reduces the interconnect lengths in both matchlines and searchlines, hence reducing signal path delay and power consumption [5]. More importantly, several architectural changes can be specially tailored to 3D-TCAM to further improve its performance.

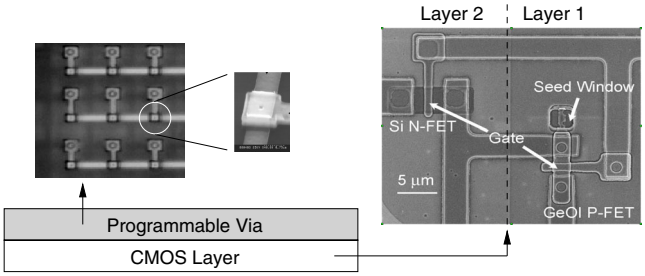

Fig. 2. Active layers of a monolithically stacked 3D-TCAM.

II. Monolithically Stacked 3D-TCAM

Our proposed 3D-TCAM is implemented with monolithically stacked 3D-IC technology, whereby active devices are lithographically built in between metal layers. The main advantage of such approach is that, in principle, it can achieve comparable vertical via density and scale at the same rate as the base CMOS technology. Although this approach has yet to be mature for the TCAM application, preliminary test results of a 3D prototyping chip (Fig. 2) have been quite encouraging

978-1-4244-1684-4/08/$25.00 ©2008 IEEE

and have shown that forming transistors on a dielectric with low thermal budget is quite feasible [6].

Moreover, since most chip area in a 2D-TCAM is occupied by stored memory bits, we use the programmable via structure [7] [1] to replace the SRAM cell array in order to further improve the TCAM cell density. The primary challenge of using such phase-changing memories is that unlike SRAM cells (or DRAM cells), they do not simply store data in the form of low and high voltage levels. Instead, data are stored as low and high resistance values, which motivates us to design new circuitry for the comparison operation and the matchline-sensing scheme.

III. The Programmable Via Array

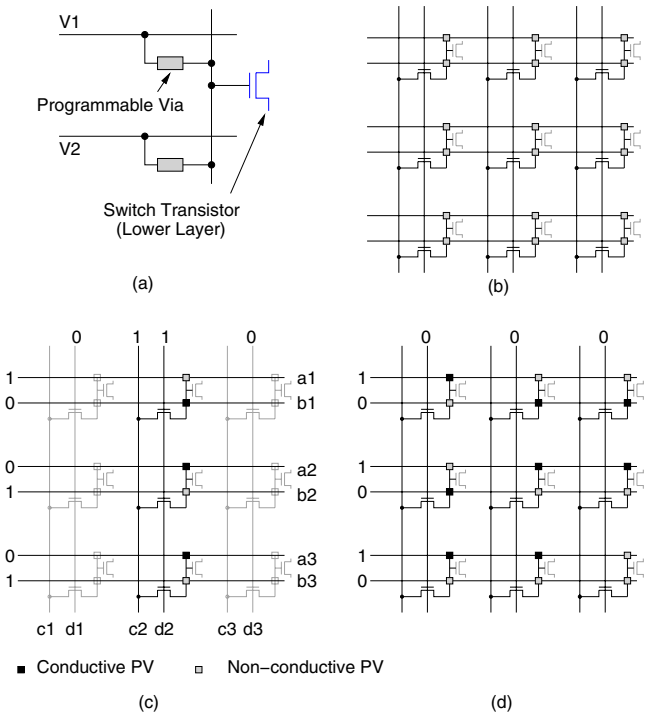

Fig. 3. (a) A memory cell containing two programmable vias and a NMOS access transistor. (b) Programming circuitry of a 3×3 programmable via array. (c) Illustration of programming operation. (d) A 3×3 programmable array with stored values.

A conventional 2D-TCAM typically has many SRAM bits, each of which is relatively large in size and therefore results in long interconnects. This motivates us to replace this SRAM bit array with programmable via array (PVA), an emerging non-volatile memory [7]. Programmable via uses the switching effect of a nano-scale metal junction, in which the stimulation of electro-chemical reaction in a solid electrolyte results in the stretching or shrinking of a metallic bridge, thereby creating or dissolving an electrically conductive channel. The programmable via can be located on the silicon substrate on which transistors have been formed. Since it can be formed within the area of a via hole between two metal layers, the area required for its arrangement at the crosspoint of two wires is only $4\lambda^2$. Further, the potential area savings for this device are actually even greater because multiple PVs can be stacked one upon another.

As mentioned in Section I, a PVA memory stores data as low and high resistance values. Therefore, special circuitry must be designed to program the programmable array. Fig. 3(a) depicts that two programmable vias will function as a single SRAM cell to store one bit and Fig. 3(b) shows the proposed programming circuitry. Each memory bit is associated with one NMOS transistor for programming. The array of programmable via can be programmed column by column. For example, in Fig. 3(c), in order to program the second column, we first turn off all NMOS accessing transistors for other columns by applying "0" to the vertical line $d1$ and $d3$. Next, we apply high voltage to line $c2$ and different "0" and "1" pair to the lines $a1 - b1$, $a2 - b2$, and $a3 - b3$. Depending on the voltage difference at its two ends, each programmable via will become conductive or remain insulated. When reading data from this programmable via array as illustrated in Fig. 3(c), all vertical lines $d1$, $d2$, and $d3$ will be first turned off. Depending on the conductivity of the programmable via, the voltage level controlling the switch transistors are either "1" from $a1 - a3$ or "0" from $b1 - b3$.

IV. The TCAM Core Cell

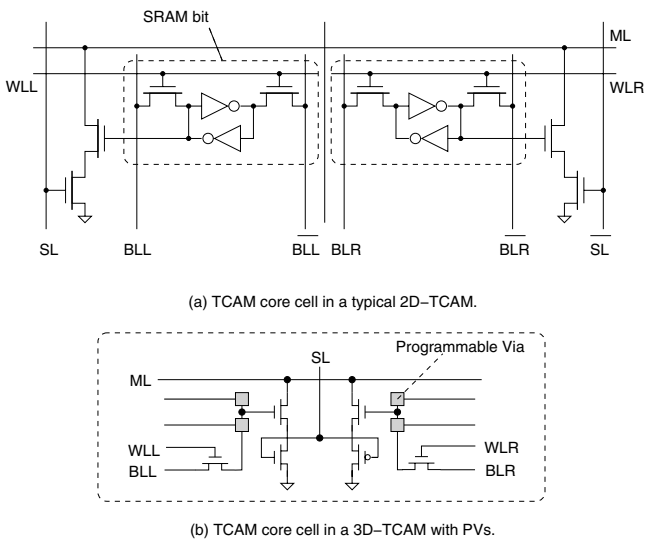

Fig. 4. Logic schematic of a 16-T NOR-type TCAM core cell. The cell is shown using 2 SRAM bits as storage with usual SRAM access transistors and bit lines.

A TCAM cell serves two basic functions: bit storage (as in RAM) and bit comparison (unique to TCAM). Fig. 4(a) shows a NOR-type TCAM cell in a typical 2D-TCAM. To encode ternary states, we need to have two SRAM cells in each TCAM cell and store two bits, denoted as D and $\overline{D}$. One bit, D, connects to the left pulldown path and the other bit, $\overline{D}$, connects to the right pulldown path, making the pulldown paths independently controlled. We store an X (don't care) state by setting both D and $\overline{D}$ equal to logic "1", which disables both pulldown paths and forces the cell to

[1] This is similar to the nanoBridge proposed in [8].

match regardless in the inputs. As shown in Fig. 4(a), the bit storage in a conventional 2D-TCAM is an SRAM cell where cross-coupled inverters implement the bit-storage nodes D and $\overline{D}$. This schematic also includes the NMOS access transistors and bitlines which are used to read and write the SRAM storage bit. The NOR cell implements the comparison between the complementary stored bit, D (and $\overline{D}$), and the complementary search data on the complementary searchline, SL (and $\overline{SL}$), using four switches, which are all typically comparison transistors with minimum size to maintain high cell density. These transistors implement the pulldown path of a dynamic XNOR logic gate. Each pair of transistors, forms a pulldown path from the matchline, ML, such that a mismatch of SL and D activates at least one of the pulldown paths, discharging ML to ground. A match of SL and D disables both pulldown paths, disconnecting ML from ground.

As shown in Fig. 4(a), the conventional TCAM cell has two complementary searchlines and the comparing circuitry consists of only NMOS transistors. In previous generations of CMOS technology, the gate capacitance of transistors dominates the capacitance of metal wires. As CMOS technology scales down to sub-100nm and further [9], the wire capacitance becomes increasingly more significant relative to the diffusion and gate capacitance of transistors, which motivates us to redesign the TCAM cell with single searchline to cut down the power consumption in searchlines. To achieve this objective, we need to use both NMOS and PMOS transistors. TCAM cell typically avoids using PMOS transistors because its switching speed is slower than NMOS and high TCAM cell density requires using the minimum width for all gate sizing. In our proposed 3D-TCAM, the delay improvement due to 3D stacking can afford us to use PMOS.

V. Pipeline granularity of precharging scheme

Matchlines and searchlines are the two key structures in TCAMs. Fig. 1 illustrates how NOR cells are connected in parallel to form a NOR matchline. A typical NOR search cycle operates in three phases. First, in searchline precharge stage, the searchlines are precharged low to disconnect the matchlines from ground by disabling the pulldown paths in each TCAM cell. Second, in the matchline precharge stage, with the pulldown paths disconnected, transistor precharges the matchlines high. Finally, in the matchline evaluation stage, the searchlines are driven to the search word values, triggering the matchline evaluation phase. In the case of a match, the corresponding matchline stays high as there is no discharge path ML voltage, to ground. In the case of a miss, there is at least one path to ground that discharges the matchline. The matchline sense amplifier (MLSA) senses the voltage on matchline, and generates a corresponding full-rail output match result.

Matchline and searchline power dissipations are two of the major sources of power consumption in TCAM. Several power-saving techniques, such as low-swing sensing and current-race scheme, are very effective in reducing dynamic power consumption in a TCAM, all of which can be readily

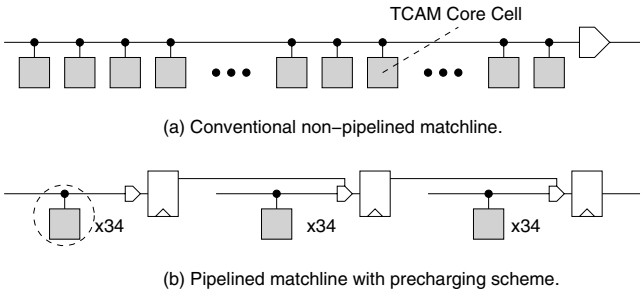

Fig. 5. (a) Non-pipelined matchline architecture. (b) Pipelined matchline architecture broken into segments. A segment is activated only if all previous segments have been matched.

adopted to 3D-TCAM. In this study, we focus on the pipelined precharge scheme. In this approach, the matchline is divided into a number of segments, where a match in a given segment results in a search operation in the next segment but a miss terminates the match operation for that word [2]. To illustrate, Fig. 5(a) shows a simplified schematic of a conventional NOR matchline structure where all cells are connected in parallel. Fig. 5(b) shows the same set of cells as in Fig. 5(a), but with the matchline broken into three matchline segments that are serially evaluated. If any stage misses, the subsequent stages are shut off, resulting in power saving. In conventional 2D planar implementation, this scheme suffers from the increased latency and the area overhead due to pipelining stages. Therefore, in 2D, by itself, a pipelined matchline scheme is not as compelling as basic selective precharge; however, in 3D monolithic stacking, we can show that deep pipelining with finer granularity can be used and achieve significant power reduction while maintaining the equivalent search latency.

To determine the necessary pipelining granularity, we simulated cases with different number of pipeline stages using HSpice assuming a 1024×144 bit TCAM, $50\lambda(H) \times 64\lambda(W)$ for the size of a TCAM core cell in the baseline 2D-TCAM, and $35\lambda(H) \times 30\lambda(W)$ for the size of a 3D-TCAM cell because of replacing of SRAM bits with a programmable via array in a monolithically stacked 3D-TCAM. Additionally, we used a 65nm CMOS technology and the Berkeley Predictive Technology Model (BPTM) for devices and interconnect.

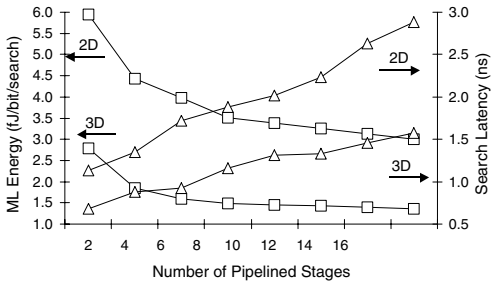

Fig. 6. Results of performance comparison between baseline 2D-TCAM and monolithically stacked 3D-TCAM.

As shown in Fig. 6, for any fixed number of pipeline stages, both search latency and dynamic power consumption are significantly improved due to the reduction of intercon-

nect lengths from monolithic stacking and the utilization of programmable via array in place of SRAM. For example, in the case of 4 pipeline stages, the search latency of a 2D-TCAM is reduced from 1.55 ns in 2D-TCAM to about 0.72 ns in 3D-TCAM, roughly a 2x improvement; the dynamic power consumption improves from 4.5 fJ/bit/search to 1.98 fJ/bit/search. Assuming the application requires 1.55 ns minimum search latency, if we adopt a 4-stage pipelined design, the energy per bit per search is about 4.5 fJ. Meeting the same delay, a 3D-TCAM design can have 10 stages, which can significantly reduce the total dynamic consumption down to 1.5 fJ, a 3X decrease.

VI. Chip Layout and Performance Results

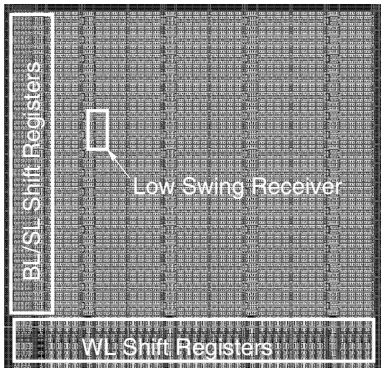

Fig. 7. Layout of the implemented TCAM.

This section presents simulation results of the proposed TCAM architecture including the functionality and power consumption with the effect of parasitics extracted from layout [2]. Timing is determined by simulating all the subcircuits in HSPICE, and power consumptions are determined by simulating the complete TCAM macro netlist in Nanosim to handle large netlist at the transistor-level. In determining typical power consumption, we assumed only one matching location per search in the TCAM and the stored search data are uniformly distributed.

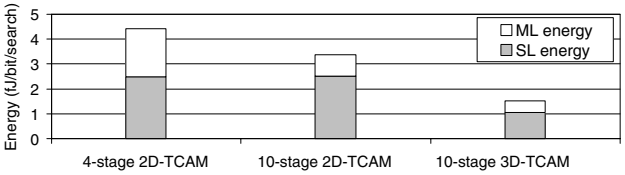

Fig. 8. Breakdowns of dynamic energy consumption in different TCAMs.

Fig. 8 compares the energy consumption of the proposed 3D-TCAM with the baseline 2D-TCAM in units of fJ/bit/search. Both TCAMs contain 1024×144 bits and are implemented with a 1.0-V 65 nm CMOS process. The power consumption of the baseline 2D-TCAM with a 4-stage pipeline is 4.43 fJ/bit/search (1.95 fJ/bit/search in ML and 2.48 fJ/bit/search in SL). Increasing the pipelining stages to 10 for this 2D-TCAM can reduce the total search energy to 3.38 fJ/bit/search with no change in SL energy and a 2.4x reduction in ML energy but increases the search latency from 1.35 ns to 2.02 ns. In contrast, the proposed 3D-TCAM with a 10-stage pipeline only consumes 0.46 fJ/bit/search in ML and 1.05 fJ/bit/search in searchlines, a total reduction of 3.2x. Note in 3D-TCAM, the search-line activity dominates the overall power consumption, as evident in the third bar of Fig. 8. We believe that replacing all flip-flops in our design with low-power flip-flops and/or adding hierarchy to the search-lines can further reduce the energy consumption in MLs and SLs respectively.

[2]Unfortunately, the experimental measurement results of the fabricated testchip is not available because the chip fabrication is still in process.

VII. Conclusion

Most studies on low-power circuit and architecture design in TCAMs have been exclusively based on 2D planar CMOS technology. This work, to our knowledge, is the first study on the utility of 3D integrated circuits and phase-changing non-volatile memory in TCAMs. Our main finding is that besides the obvious performance benefits of larger logic density and better performance due to the shortening of interconnects, 3D-IC can potentially motivate novel designs that may not be compelling in 2D. In this study, 3D-IC enables us not only to employ finer pipeline granularity to further improve power savings without compromising delay performance in 3D-TCAM but also to design a novel TCAM core cell structure with single searchline. It should be noted that our reported results here are still preliminary, we believe additional innovations in architecture and further performance improvements are possible. One immediate next step of our research is the testing of a prototype 3D-TCAM chip with our proposed design.

References

[1] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, 2006.

[2] K. Pagiamtzis and A. Sheikholeslami, "A low-power content-addressable memory (CAM) using pipelined hierarchical search scheme," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1512–1519, 2004.

[3] A. Roth, D. Foss, R. McKenzie, and D. Perry, "Advanced ternary cam circuits on 0.13um logic process technology," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, pp. 465–468, 2004.

[4] C. A. Zukowski and S.-Y. Wang, "Use of selective precharge for low-power content-addressable memories," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, no. 3, pp. 1788–1791, 1997.

[5] M. Lin, A. El Gamal, Y.-C. Lu, and S. Wong, "Performance benefits of monolithically stacked 3-D FPGA," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 26, no. 2, pp. 216–229, 2007.

[6] A. R. Joshi and K. C. Saraswat, "Nickel induced crystallization of a-Si gate electrode at 500C and MOS capacitor reliability," *IEEE Trans. Electron Devices*, vol. 50, pp. 1058 – 1062, April 2003.

[7] W. Wang, A. Gibby, Z. Wang, T. W. Chen, S. Fujita, P. Griffin, Y. Nishi, and S. Wong, "Nonvolatile SRAM cell," in *IEDM Technical Digest. IEEE International*, pp. 27–30, 2006.

[8] S. Kaeriyama, T. Sakamoto, H. Sunamura, M. Mizuno, H. Kawaura, T. Hasegawa, K. Terabe, T. Nakayama, and M. Aono, "A nonvolatile programmable solid-electrolyte nanometer switch," *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, vol. 40, no. 1, pp. 168–176, 2005.

[9] R. Ho, K. Mai, and M. Horowitz, "The future of wires," *Proceedings of the IEEE*, vol. 89, no. 4, pp. 490–504, 2001.