A STATIC POWER REDUCTION TECHNIQUE FOR TERNARY CONTENT ADDRESSABLE MEMORIES

Nitin Mohan and Manoj Sachdev
Electrical and Computer Engineering, University of Waterloo, Ontario, Canada
nitinm@vlsi.uwaterloo.ca, msachdev@uwaterloo.ca

Abstract

Ternary content addressable memories (TCAMs) are attractive for high-speed packet forwarding and classification in network switches and routers. Traditionally, the static power in TCAMs has been a small fraction of the total power due to high activity of TCAMs. However, technology scaling and architecture-level techniques are reducing the dynamic power of TCAMs. The technology scaling is also increasing the off-current of transistors. Hence, the static power is becoming a significant portion of the total power consumption in TCAMs. This paper presents a technique to reduce the static power in SRAM-based TCAMs without affecting the speed of operation. We analyze the circuits and present the trade-offs of using this power-reduction technique. The simulation results show a significant reduction in the static-power (up to a factor of 11) for an SRAM-based TCAM in 0.13 µm technology.

Keywords: *TCAM; low-power; static-power; scaling.*

1. INTRODUCTION

Content addressable memories (CAMs) are special kind of random access memories (RAMs) that also support parallel data-search operation. Traditionally, CAMs have been attractive for applications such as translation look-aside buffers (TLBs) in virtual memory systems, tag directories in associative cache memories, database accelerator, data compression, and image processing. However, the recent research and development of CAMs is primarily driven by the demand for high-speed table lookups in network switches and routers [1]. A lookup operation in a CAM is performed by comparing an input keyword with all the stored entries in parallel. If a stored data-word matches with the input keyword, its location is provided as the search result. In the presence of multiple matches a priority encoder selects the highest priority address. The parallel search feature of CAMs provides very high-speed table lookup with a fixed latency.

CAMs can be divided into two categories: binary CAMs, and ternary CAMs (TCAMs). The binary CAMs can store and search only binary data. Hence, their applications are limited to exact-match searches. A more powerful and feature-rich TCAM also supports partial-match searches. The TCAMs can store and search ternary states: '0', '1', and 'X' (don't care). The 'X' state can be used as a wild card in a search operation. It also allows a search operation in a given range of values. These features are particularly suitable for classless inter-domain routing (CIDR), and packet classification in a single search operation [1]. The high-speed parallel search operation of CAMs leads to very high power consumption. Several methods have been proposed in the past to reduce the dynamic power of the TCAMs. The static power of TCAMs has been ignored due to high activity of TCAMs and low transistor-leakage in the popular CMOS technologies. However, technology scaling and recent improvements in the TCAM architectures have increased the contribution of the static power. Hence, static power reduction techniques are needed to control the power consumption in TCAMs.

Rest of the paper is organized as follows. Section 2 explains the importance of static power in the next generation of TCAMs. Section 3 describes a static power reduction technique proposed by this work. Section 4 presents the simulation results and discussion. Finally, section 5 concludes the paper with main points and future directions.

2. STATIC POWER IN TCAM

The technology scaling reduces transistor dimensions and hence the gate delay by 30% with each new generation of process technology [2]. This increases the operating frequency by 43%, and reduces the energy by 65% and dynamic power by 50% [2]. For a given supply voltage, a reduction in the gate oxide thickness increases the electric field across it. This may cause the gate oxide to break down. Hence, the power supply voltage is also scaled for reliability reasons. However, a smaller power supply voltage reduces the overdrive voltage ($V_{GS} - V_t$) of

CCECE 2004- CCGEI 2004, Niagara Falls, May/mai 2004
0-7803-8253-6/04/$17.00 ©2004 IEEE

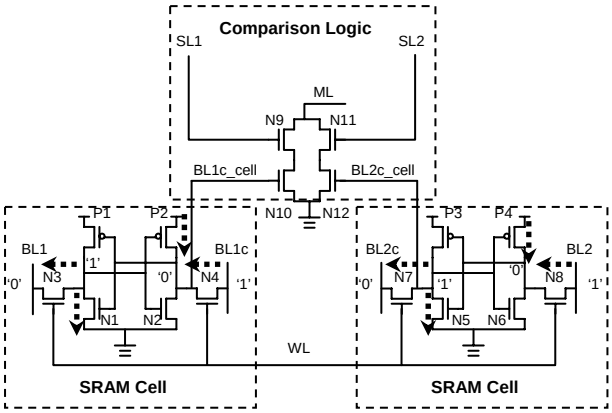

Figure 1: Leakage current paths in a conventional 16T TCAM cell

the transistors. The performance is maintained by reducing the transistor threshold voltage (V_t). A reduction in V_t increases the off-current of a transistor exponentially. It has been estimated that V_t decreases by 15% per generation, and the off-current of a transistor increases by 5 times each generation [2]. As a result, the static power of the circuits implemented in advanced CMOS technologies is no longer a negligible component of the total power consumption. Recent architecture-level techniques enable only a small portion of TCAM chip which further reduces the dynamic power of TCAMs [3][4]. In addition, during read, write and no activity periods, most part of the TCAM chip remains inactive. Therefore, the static power will become a significant portion of the total power consumption in the next-generation of TCAMs.

In a conventional 16T TCAM cell, the leakage current paths are shown by dashed arrows in Figure 1. The leakage currents through the access transistors (N3, N4, N7, and N8) will be absent if BL1c = BL1c_cell and BL2c = BL2c_cell. Therefore, the average leakage current of a conventional 16T TCAM cell is the sum of four inverter leakage currents and two access transistor leakage currents. In contrast to a binary CAM cell, which consists of one SRAM cell, each TCAM cell consists of two SRAM cells. The extra SRAM cell is needed to store the mask bit ('X'). Therefore, the static power consumption of a TCAM chip is two-times higher than that of a binary CAM for the same storage capacity.

3. STATIC POWER REDUCTION TECHNIQUE

The subthreshold leakage of a MOS transistor with $V_{GS} = 0$ and $V_{DS} = V_{DD}$ can be given by the following expression [5]:

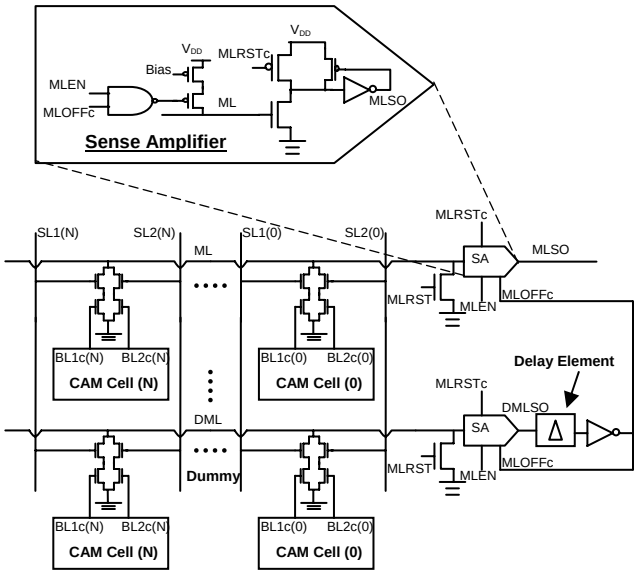

Figure 2. Current-race sensing scheme [6]

$$I_{SUB} = I_0 \exp(-\frac{k_1\sqrt{\phi_S} - k_2\phi_S - \eta V_{DD}}{nV_T})$$

The above expression suggests that a reduction in V_{DD} decreases the subthreshold leakage current exponentially by reducing the drain induced barrier lowering (DIBL) [5]. Therefore, the static power in TCAMs can be significantly reduced by lowering the V_{DD} in the storage (SRAM) cells.

The effect of V_{DD} reduction on the TCAM search speed depends on the operation of match line (ML) sense amplifier. In this work, we have used a recently proposed current-race scheme for ML sensing [6]. The current-race sensing scheme is shown in Figure 2. Initially, all the MLs are discharged to ground. The input keyword is placed on the search lines (SLs), and all the MLs are charged with current sources. The MLs of the matched words charge faster than those of the mismatched words. It is due to the fact that mismatch words have one or more conducting paths from ML to ground. A dummy word with forced match condition is also included to generate a reference signal (MLOFFc) which turns-off all the current sources when match detection is completed. As soon as the voltages of the dummy ML and the matched words reach the threshold voltages of their respective sense amplifiers, their outputs change from '0' to '1'. The outputs of the mismatched words remain at '0'. A delay element is used to ensure that voltages of all the matched MLs reach the threshold voltages of their respective sense amplifiers even in the presence of process variations.

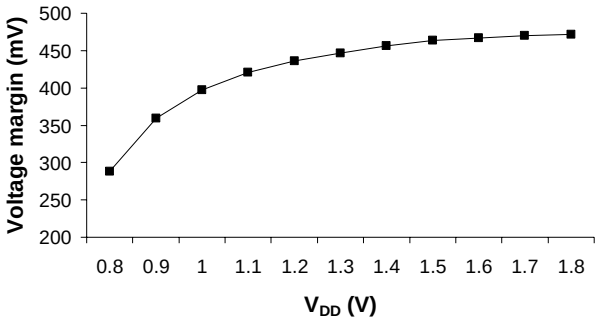

Figure 3: Voltage margin variations with V_{DD}

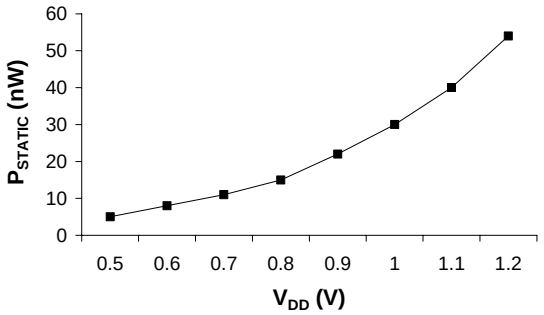

Figure 4: Static power variations with V_{DD} for a 144-bit TCAM word in 0.13 µm CMOS technology

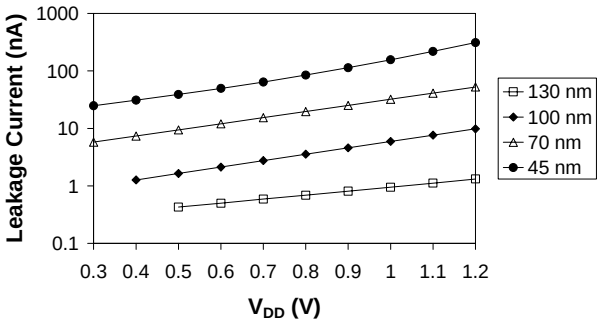

Figure 5: Leakage current variations with V_{DD} for various technology nodes

The V_{DD}-reduction scheme requires two power supply voltages. The ML sense amplifiers are powered by a higher V_{DD} to maintain the search speed. The storage memory cells are powered by a smaller V_{DD} to reduce the static power consumption. Typically, each row has only one ML sense amplifier, and the static power of the ML sense amplifiers is much smaller than the static power of the memory cells. Therefore, the higher power supply voltage at the ML sense amplifiers does not affect the total static power consumption significantly.

Lowering the V_{DD} in the SRAM cells reduces the overdrive voltage ($V_{GS} - V_t$) of transistors N10 and N12 in the comparison logic (Figure 1). This reduces the 'ON' current of the ML pull-down path. The reference signal (MLOFFc) in the current-race sensing scheme is generated by the dummy ML, which imitates a matched word. Since there is no ML to ground path for the dummy word, a reduction in the 'ON' current does not affect the search speed. Moreover, the current sources are turned off by the reference signal (MLOFFc). Hence, the power consumption in the current-race sensing scheme is also unaffected by the V_{DD} reduction. However, the V_{DD}-reduction scheme reduces the separation gap between the 'ON' and 'OFF' currents of the ML pull-down path. As a result, it becomes difficult for the ML sense amplifier to differentiate between the two conditions: (i) match, and (ii) 1-bit mismatch. In the current-race sensing scheme, the smaller gap between 'ON' and 'OFF' currents corresponds to a smaller voltage margin in the ML sense amplifier. The voltage margin is defined as the difference of the threshold voltage of an ML sense amplifier and the maximum voltage of an ML with 1-bit mismatch [6]. Higher voltage margin can handle larger variations in the threshold voltage (V_t) of NMOS transistors.

4. RESULTS AND DISCUSSION

Figure 3 shows the voltage margin variations with V_{DD} for the current-race sensing scheme simulated in TSMC 0.18 µm CMOS technology. As expected, the voltage margin decreases as V_{DD} is reduced.

We also simulated a 144-bit word of an SRAM-based TCAM in Cadence using UMC 0.13 µm CMOS technology design kit. Figure 4 shows the static power variations with V_{DD}. A reduction in V_{DD} from 1.2V to 0.5V reduces the static power consumption by a factor of 11. With the technology scaling, short channel effects will become more severe, and this technique will become more effective. We simulated the leakage current reduction with V_{DD} for different technology nodes using Berkeley predictive technology model (BPTM) [7][8]. The results are shown in Figure 5 on logarithmic scale. The leakage current reduction is linear on the logarithmic scale. It confirms the exponential dependence of V_{DD} on the leakage current. As technology scales down from 130 nm to 45 nm, the leakage current increases rapidly. Figure 5 also shows that the slope of the leakage current variation with V_{DD} increases as the technology scales down from 130 nm to 45 nm. Hence, the V_{DD} reduction technique becomes more effective in more advanced technologies as expected from the earlier discussion.

Although lower power supply in the TCAM cells reduces the static leakage current, it makes the memory cells more vulnerable to soft errors [9]. Soft errors occur in RAMs when an energetic alpha particle or neutron enters the substrate generating electron-hole pairs. This extra charge may cause the voltage on the storage nodes

to fluctuate temporarily, which may corrupt the stored data. Lower V_{DD} results in smaller charge at each storage node making it more susceptible to soft errors. The critical charge, which is defined as the minimum charge to flip a stored bit, reduces linearly with V_{DD} [9]. Hence, the smaller leakage can be traded-off with the higher critical charge depending on the application requirements. Fortunately, this problem is less severe in CAMs since the capacitance of the storage nodes is higher due to the comparison logic [1]. Hence, lowering the power supply voltage is viable in CAMs. Lower V_{DD} also reduces the static noise margin (SNM) of the storage cell making it more susceptible to noise [10]. In addition, the technology scaling increases the intrinsic device fluctuations [10]. Hence, the V_{DD} reduction scheme requires careful estimation of the noise sources, coupling mechanisms, and SNM.

5. CONCLUSIONS

A static power reduction technique for TCAM is presented. The static power is reduced by lowering the V_{DD} of storage portion of TCAM cells. The simulation results show a significant reduction in the static-power (up to a factor of 11) for a conventional SRAM-based TCAM in 0.13 µm technology. BPTM simulations further confirm that this technique will be more effective in the future generations of CMOS technology (such as 45 nm). We demonstrated that this technique does not affect the speed and dynamic power consumption of the TCAM with current-race sensing scheme. We also observed that this technique reduces the gap between the 'ON' and 'OFF' currents from ML to ground. This sacrifices the robustness of the ML sense amplifiers and makes it difficult to differentiate between the match and 1-bit mismatch conditions. This issue worsens in wider TCAMs with larger number of cells connected to each ML causing larger 'OFF' current. In addition, lower V_{DD} also makes the storage cells more susceptible to soft-errors and noise disturbances. Therefore, V_{DD} of the storage cells should be reduced by taking all the stability issues into account. Moreover, novel ML sensing schemes are needed that can offer robust operation even for a small separation gap between the 'ON' and 'OFF' currents.

Acknowledgements

This work is supported by Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarship (PGS B). The authors would like to thank Bhaskar P. Chatterjee, Wilson W. Fung, and Derek W. Wright for stimulating discussions on this work.

References

[1] M. Wirth, "The next generation of content addressable memories," *Application Note*, MOSAID Technologies, http://www.mosaid.com/semiconductor/papers/NextGenerationofCAMs.pdf.

[2] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, pp. 23-29, Jul.-Aug. 1999

[3] K. Pagiamtzis, and A. Sheikholeslami, "Pipelined match-lines and hierarchical search-lines for low-power content-addressable memories," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 383-386, Sep. 2003

[4] G. Kasai, Y. Takarabe, K. Furumi, and M. Yoneda, "200MHz/200MSPS 3.2W at 1.5V Vdd, 9.4Mbits ternary CAM with new charge injection match detect circuits and bank selection scheme," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 387-390, Sep. 2003

[5] R. X. Gu, and M. I. Elmasry, "Power dissipation analysis and optimization of deep submicron CMOS digital circuits," *IEEE Journal of Solid-state Circuits*, vol. 31, no. 5, pp. 707-713, May 1996

[6] I. Arsovski, T. Chandler, and A. Sheikholeslami, "A ternary content- addressable memory (TCAM) based on 4T static storage and including a current-race sensing scheme," *IEEE Journal of Solid-state Circuits*, vol. 38, no. 1, pp. 155-158, Jan. 2003

[7] Berkeley Predictive Technology Model. [Online]. http://www-device.eecs.berkeley.edu/~ptm/

[8] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu, "New paradigm of predictive MOSFET and interconnect modeling for early circuit design," *Proceedings of the IEEE Custom Integrated Circuits Conference*, pp. 201-204, Jun. 2000

[9] V. Degalahal, N. Vijaykrishnan, M. J. Irwin, "Analyzing soft errors in leakage optimized SRAM design," *Proceedings of the 16th International Conference on VLSI Design*, pp. 227-233, Jan. 2003

[10] A. J. Bhavnagarwala, X. Tang, J. D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE Journal of Solid-state Circuits*, vol. 36, no. 4, pp. 658-665, Apr. 2001