Variation Analysis of CAM Cells

Amol Mupid[1], Madhu Mutyam[2], N. Vijaykrishnan[1], Y. Xie[1], M. J. Irwin[1]
[1] *Department of Computer Science and Engineering*
The Pennsylvania State University, University Park, USA
{mupid, vijay, yuanxie, mji}@cse.psu,edu
[2] *International Institute of Information Technology, Hyderabad, India*
madhu_mutyam@iiit.ac.in

Abstract

Process related variations are considered a major concern in emerging sub-65nm technologies. In this paper, we investigate the impact of process variations on different types of content addressable memories (CAM). As CAM structures are used in various on-chip structures such as caches and TLBs, understanding process variation impact of CAM structures is important. To gain insight on the relative importance of process variation effects, we also examine the effect of variability due to temperature and supply voltage changes. Our results show that the NAND-type CAM cells are more susceptible to failures as compared to NOR-type CAM cells. Finally, we propose an architectural technique to mitigate the performance degradation effects of process variation. The proposed technique yields an average of 37.5%, 25% and 12.5% reduction in CAM occupancy for caches with 25%, 50% and 75% variation respectively, over SPEC 2000 CPU benchmarks as compared to a worst-case design in the presence of process variation.

1. Introduction

As technology scales, it is becoming increasingly difficult to control critical transistor parameters for sub 90nm designs. The challenges of fabricating ever-shrinking transistor size have lead to significant variability in the critical transistor parameter such as channel length (L_{eff}), oxide thickness (T_{ox}) and threshold voltage (V_{th}) across a die (intra-die variation) and across a wafer (inter-die variation). Variability occurs due to various manufacturing reasons such as wafer misalignment, lithographic interactions, plasma etch microloading, random dopant fluctuations and imperfections in planarization steps. The deviation of the fabricated transistor parameters from the intended parameters creates significant functional correctness concerns such as stability in memory cells, creation of new critical paths in the design and increased leakage currents. The degree of variability encountered in the new process technologies makes designing for worst-case process margins non-viable option for future designs.

Content Addressable Memories (CAMs) allow data to be searched based on content rather than physical location. This provides a fast method to compare data against a table of previously stored entries and return the address of the matching entries. CAMs have a single clock cycle throughput, which makes them faster than other hardware and software-based search systems. CAMs can be used in wide range of applications where high search speed is critical. These applications include packet forwarding and packet classification in network routers, processor caches, translation look aside buffers (TLB), data compression applications, database accelerators and neural networks.

In this paper, we investigate the impact of process variations on different CAM cells. There are two main types of CAM cells, namely, NOR-type CAM cell and NAND-type cell [14]. For remainder of this paper we discuss only these two types as they are in predominant use compared to other variants. We evaluate these two types of CAM cell structures and contrast their susceptibility towards failures due to process variations. In addition to process variation we investigate the effect of variation in environmental factors like temperature and supply voltage and determine which factor dominates and hence needs mitigation.

The rest of the paper is organized as follows. Core CAM cells and match line structures are presented in Section 2. In Section 3, we analyze both NOR and NAND cells for failures due to process variations and examine the impact of process variations on delay. The results of the analysis are presented in Section 4. We also investigate environmental variations like temperature variations and supply voltage variations. In Section 5, we discuss circuit and architectural techniques to mitigate the impact of process variations. Finally, concluding remarks are presented in Section 6.

This work was supported in parts by grants from NSF CAREER 0093085, NSF Award 0454123 and GSRC.

Proceedings of the 8th International Symposium on Quality Electronic Design (ISQED'07)
0-7695-2795-7/07 $20.00 © 2007 IEEE

2. Core Cells and Matchline Structure

A CAM cell has a bit storage and bit comparison. Figures 1(a) and 1 (b) show the NOR-type CAM cell and NAND-type CAM cells, respectively. Both these cells use SRAM cell for bit storage.

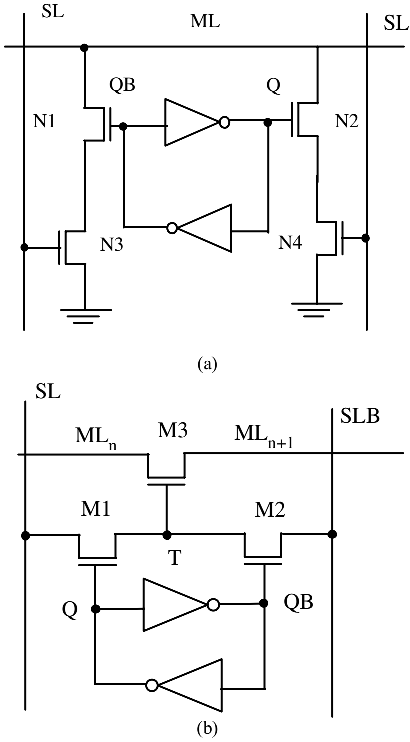

Figure 1. (a) NOR-type CAM cell and (b) NAND-type CAM cell. For simplicity, SRAM access transistors and Bitlines are omitted.

2.1 NOR Cell

The comparison between the stored bit, Q (and QB), and the complementary search bit SL (and SLB) is performed using four comparison transistors, N1 through N4. These transistors are usually minimum-sized to maintain high cell density. The transistors N1/N3 and N2/N4 form two separate pull down paths for the matchline, ML. A mismatch between stored bit, Q and search bit SL activates at least one of the pull down paths. Hence matchline, ML, discharges through one of these pull down paths on a mismatch. In the event of a match of Q and SL, both the pull down paths are disabled and hence ML is disconnected from ground. To form a CAM word multiple NOR CAM cells are connected in parallel by connecting the ML segment of each cell to the ML segment of the neighboring cells. When connected in such a fashion, the pulldown paths are in parallel and resemble the pulldown paths in a NOR logic gate. The matchline, ML, is precharged high and discharges through one or many pulldown paths depending on the number of bits mismatching in that particular word. A match for the entire word occurs only when all the cells in that word result in a match. The corresponding matchline, ML, does not discharge. The main feature of NOR-type CAM matchline is its high speed of operation. The worst case occurs when there is one-bit miss in a word and the matchline discharges through two series transistors in the cell, which forms the pulldown path.

2.2 NAND Cell

The comparison between the stored bit Q (and QB), and the complementary search bit SL (and SLB) is performed using three comparison transistors, M1 through M3. These transistors are usually minimum-size to maintain high cell density. Consider the case of a match when SL=1 and Q=1. This switches *on* transistor M1, that provides a weak logic '1' at the node T, which in turn switches *on* M3. In the event of a mismatch, node T gets logic '0' and switches *off* transistor M3. To form a CAM word multiple NAND cells are connected in a serial fashion. When connected in this fashion, even a single mismatch in any of the bits, disconnects the matchline, ML. The matchline, ML, is precharged high and discharges through a series of NMOSs only in the event of a complete match. NAND-type CAM operation is slower compared to NOR-type as evaluation path consists between 8 and 16 transistors as opposed to 2 transistors. NAND-type CAM saves power [13] as a mismatch prevents ML from discharging and hence is used in low power designs.

3. Failure Analysis of CAM Cells due to Process Variation

Intra-die variations result in mismatches in device parameters of identically designed transistors on the same die. These mismatches modify the strength of individual transistors resulting in various failures. Process variation in CAM cells may result in *search time failure*, *match failure*, and *SRAM bit failure*. We now discuss these failures in details.

3.1 Search Time Failure

The CAM cell search time is defined as the time taken for a pre-specified voltage drop ($0.1V_{dd}$) on the matchline, ML. The matchline sense amplifier (MLSA) detects this voltage drop and generates a full swing output *match* result. Due to process variations, the strengths of various transistors in cell are modified. Consider a case of mismatch in NOR-type cell, a decrease in strength of transistors (due to high V_{th} in N1, N2, N3 and N4 in Figure.1) that discharges matchline

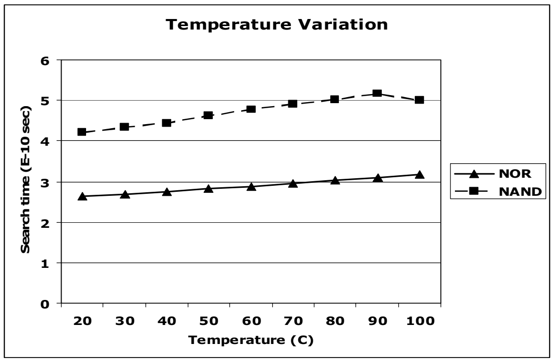

Figure 2. Effect of Temperature Variation.

during search operation will lead to smaller voltage drop on the matchline, which may result in wrong evaluation. This type of failure is termed as *search time failure.* Similarly in the case of a match in NAND-type cell, a decrease in strength of comparison transistors (due to high V_{th} in M1, M2 and M3 in Figure 1) will reduce the voltage drop on the matchline, which may again resultin wrong evaluation. As a number of transistors (8 to 16) are in series formation on a NAND matchline, it is more likely to be affected by strength modification of the comparison transistors.

3.2 Match Failure

In a NOR-type cell, if the cell is storing a '0' (Q=0 and QB=1) SL=0 and SLB=1, then ML, is supposed to stay high as there is no path to ground. If process variation increases the leakage current (due to low V_{th} in N2 and N3 in Figure 1), then there will be a voltage drop on matchline, which may result in wrong evaluation by sense amplifier. Similarly in the case of a match in NAND-type cell (SL=1 and Q=1), a decrease in strength of comparison transistors (due to high V_{th} in M1, M3 in Figure 1) and increase in leakage current (due to low V_{th} in M2 in Figure 1) may not generate sufficient voltage for transistor M3 to turn ON. This will prevent the matchline from discharging and may result in wrong evaluation. This failure is termed as *match failure.*

3.3 SRAM bit Failure

In both NOR-type and NAND-type cell, 6T SRAM bit is used for storage. [11] presents three types of failures associated with SRAM bit, namely, *read failure*, *write failure* and *access time failure*. These failures are observed in both types of CAM cells. Since SRAM failure analysis take these into account, we deal only with failures specific to CAMs namely, *search time failure* and *match failure* in this work.

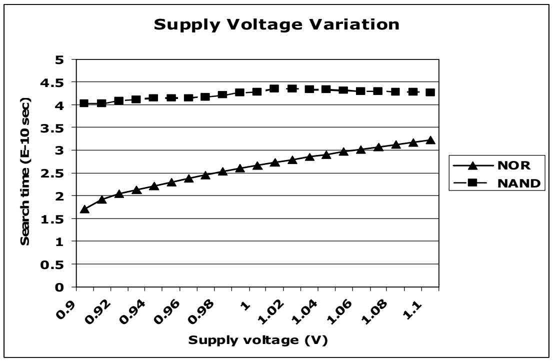

Figure 3. Effect of Supply Voltage Variation.

The following section describes our circuit design and simulation of both types of CAM structures.

4. Experimental Results

In order to investigate delay under different variations, we choose to setup experiment involving 64 NOR-type CAM cells in a row. The number of NAND-type cells in a row is limited to 16 to prevent quadratic delay degradation [12].

We implemented our design in 65-nm PTM technology [1] with nominal supply voltage of 1V. The optimal transistor sizing for custom designed core cell of each type was determined through our circuit simulation as stability of the core cell is an important precondition for this study.

In addition to our analysis of failures due to process variation, we examine the effect of temperature (environmental variation) and supply voltage variations. This is particularly important as this enables us to conclude whether process variation necessarily dictates any change in the CAM design. We present results of temperature and supply voltage variation prior to process variation and then draw a conclusion.

4.1 Temperature Variation

One of the important environmental factors, which need to be considered for CAM design, is temperature. We examine the effect of temperature on the delay of both types of CAM cells over a range of 20^0C-100^0C.

The mobility of MOSFETS reduces with an increase in substrate temperature due to increased scattering at higher temperature [8]. At the same time threshold decreases with increase in temperature due to change in fermi potential ($Ø_f$) [9]. The effect of mobility has been found to be more dominant and temperature increases delay both in devices and interconnects [8]. From Figure 2, the maximum delay increase over the nominal temperature of 25 ^{0}C is 19.48% and 20.79% for NOR-type and NAND-type cells, respectively.

4.2 Supply Voltage Variation

Supply voltage is one of the most important environmental factors that cause variations in operating conditions. As technology scales the supply voltage is typically scaled down to reduce power consumption.

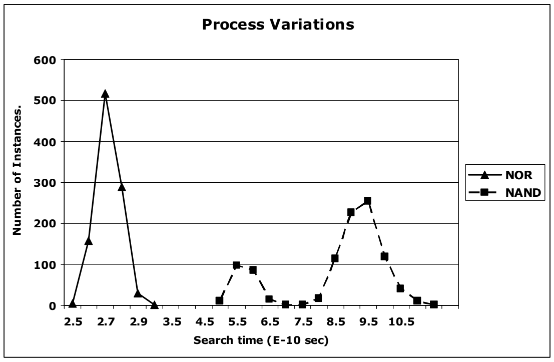

Figure 4. Effect of Process Variation.

The 65-nm technology used in our simulation has a nominal supply voltage of 1 V. 10% variation in supply voltage is expected from the nominal value [4]. The effects of supply voltage variations on delay of both types of CAM cells are demonstrated in this section.

Search time is defined as the time taken by matchline to drop to a pre-specified voltage (0.9V in our case). Two opposing trends are present here. First, an increase in the required voltage drop from Vdd to 0.9V, this increases the search time. Second, an increase in the speed of transistors, particularly {N1, N3}, {N2, N4} in Figure 1(a) and M3 in Figure 1(b), decreases the search time. For NOR-type cell, the worst case has only two such transistors and hence the first trend dominates. This results in an increase of the search time with increasing supply voltage. For NAND-type cell, the worst case involves a series of transistors. In this case, initially first trend dominates. After supply voltage of 1.02V the second trend dominates (Figure 3). The maximum delay variability over the nominal supply voltage of 1V is 20.95% in NOR-type, whereas it is merely 1.54% in NAND-type CAM cell.

4.3 Process Variation

The intra-die variations can lead to variability in the behavior of the different portions of identically designed components. These variations have significant impact on the memory cells because they are typically designed using minimum sized transistors for density reasons. Based on the underlying physical phenomenon causing the variation, the variation can be totally random (as in the case of threshold voltage fluctuations due to random dopant fluctuation [6]) or spatially correlated (due to lithographic variation [6]). The critical parameter variations include channel length (L), gate oxide thickness (T_{ox}), and threshold voltage (V_{th}). In this work, we consider the threshold voltage (V_{th}) variation as the major source of intra-die variation, because the effect of other parameter variations can be translated as effective variation in threshold voltage [11].

We evaluate both types of CAM design using Monte Carlo simulations in HSPICE. The deviation in V_{th} is assumed to follow a normal distribution with 20% standard deviation.

As we can see from Figure 4, the deviation is much larger in the case of NAND-type cell as compared to NOR-type cell. In NOR-type cell, the match line discharges through two transistors (in the worst-case scenario) and only if these two transistors are adversely affected, the delay increases.

In the case of NAND-type cell, the worst-case scenario involves the match line discharging through a series of transistors (16 in our simulations). If any of these transistors are adversely affected, the delay increases. Consequently the probability of a transistor being affected adversely due to process variation is higher in NAND-type cell. This basic dissimilarity in the worst-case scenarios for the two types of CAM cells explains the huge disparity in their susceptibility towards process variation. The maximum delay variability over the nominal case of no process variation is 10.5% for NOR-type and 138.5% for NAND-type CAM cells.

For the more commonly used NOR-type CAM cell, both supply voltage and temperature variations cause around 20% increase in delay as compared to 10.5% due to process variation. In the case of power-saving NAND-type CAM cell process variation causes a significant 138.5% delay variability. The increase in delay over nominal case constitutes search time failure. During our simulation we observed that match failures were absent for a reasonable range of variability due to temperature, supply voltage and process variation.

5. Architectural Techniques

This section describes the proposed architectural techniques to mitigate the effects of supply voltage, temperature and process variations.

Since both supply voltage and temperature variation affect at a higher granularity (at cache sub-bank level rather than cache line level), architectural techniques can be easily devised to mitigate these variations. A temperature sensor and a voltage monitor per cache sub-bank would allow us to periodically gauge the temperature and voltage. A look-up table can store the information, which enables taking appropriate action. For instance, the sense amplifiers need to be fired later than usual when the temperature is high. This would help prevent *search time failures*. This does not apply to process variation as intra-die variations affect at a much smaller granularity and hence need more consideration.

Inter-die variation affects all the transistors on a particular die similarly and hence is relatively easy to mitigate. [13] uses a replica matchline to generate timing signals similar to replica wordlines and bitlines used in conventional memory design [10]. The replica matchline loading tracks the inter-die variations in the rest of the array and hence minimizes the effect of inter-die process variation. Considering intra-die variation does not affect all transistors in similar fashion, the above mechanism is inadequate to capture intra-die variations and hence fails to mitigate the effects of intra-die process variations.

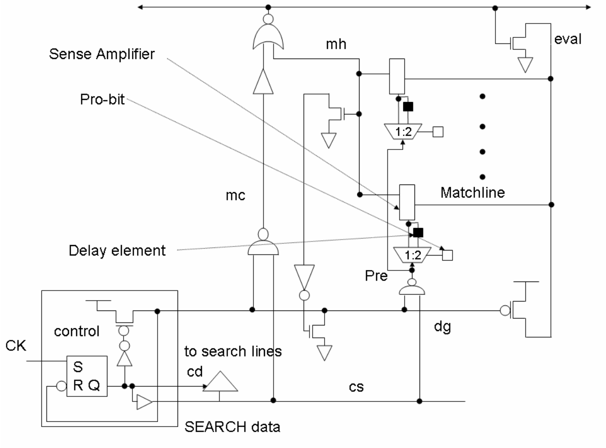

Figure 5. Search self-timing loop

To encapsulate and mitigate the effects of intra-die process variations, we need to look at these effects at the cache line level. Also, since cache layout is regular, the effects of systematic layout-dependent process variations are reduced [7]. We propose the use of a simple architectural add-on combined with an optimized March C-test and double sensing technique [11]. The architectural add-on is a one-bit storage per row called *pro-bit* (process variation indicating bit).

A March C-test involves applying a specific sequence of operations that read and write values of 0s and 1s to different locations in a memory. This sequence has been traditionally used to weed out static faults such as stuck-at faults. Recently, modifications have been suggested to the March test that can also identify the effects of process variation such as destructive readout failures [11]. The pro-bit storage, we employ can be initialized through such a test combined with double sensing technique, that captures search time failures before the operational phase of a cache. Each CAM row is tested using March C-test and double sensing technique. Even if one of the CAM cell in a row indicates search time failure, the row is considered slow and the pro-bit for that row would be ‘1’. Otherwise, it is ‘0’ for normal speed.

The major components of the proposed search self-timing loop are shown in Figure 5 (Waveforms are missing due to lack of space). It shows matchlines, control block, pro-bit storage bits and logic needed to accommodate the variable speed of the CAM rows as well as to generate control signals. Self-timing loop similar to [13] is proposed, to accommodate the use of the architectural add on pro-bit storage discussed earlier.

The new self-timing loop clocks the sense amplifiers based on the information storage in *pro-bit* and hence prevents *search time failures*.

A search operation has the following sequence of events (refer to Figure 5 for signal names).

1. The control block asserts the **cd** signal after the rising edge of the clock **ck**. This results in search data being driven on the search lines and both **dg** and **cs** are pulled high.
2. As **dg** goes high, it disables the precharge of the local match lines and the PMOS on **eval** is switched off. The pre signal travels the full height of the CAM block and is used to clock the sense amplifiers. The **pro-bit** plays a crucial role as instead of passing the **pre** signal directly to the sense amplifiers it passes it through a delay element if the **pro-bit** storage is a ‘1’, thus preventing a search time failure.
3. Since **mc** travels the full height of the CAM block, it is used to assert the **eval** signal. When it reaches low voltage, match evaluation is initiated.
4. When any sense amplifier detects a hit, **mh** signal is asserted. This switches off the NMOS on the **eval** and all other sense amplifiers. It also pulls the **dg** low.
5. The self-timing loop completes with the falling transition on **dg**. This initiates the precharge of the local match lines, **eval** signal is deasserted and the control block is reset.

This new search self-timing loop enables us to have variable cycle latency for tag matching in the set associative caches in microprocessor. This technique applies to both the types of CAM cells presented earlier. Delay variability due to process variation can be handled in multiple ways.

Increasing the clock cycle duration to accommodate the slow CAM rows is certainly not a viable option. Adopting two cycles tag match latency for all the CAM rows is overly pessimistic. The proposed technique gives a match in single cycle for the CAM rows, which have been identified as normal (*pro-bit* = 0) and takes two cycles for the slow ones (*pro-bit* = 1). In order to study the effect of process variation, here we assume that CAM Tag Cells in L1 data cache are affected by process variation, so that some of the cache sets have higher access latency than the original latency. In the rest of this work, process variation affected data cache set implies data cache set having CAM tag array, which has been identified as slow (*pro-bit* = 1).

We now discuss the access percentages of different L1 data cache sets where some of the cache sets are process variation affected and hence they have high access latency. In order to know the access percentages of different L1 data cache sets, we conducted experiments using Simplescalar tool-set [3] with 22 SPEC 2000 CPU benchmarks (12 integer and 10 floating-point) [2].

In our experiments, we considered three different scenarios where 25%, 50%, and 75% of L1 data cache sets are process variation affected. Figure 6 shows benchmark-wise access percentages. In the figures, "High Latency" indicates the access to a process variation affected cache set and "Low Latency" indicates the access to a normal cache set. Though different benchmarks have different access percentages to process variation affected cache sets, on an average, the access percentage to process variation affected cache sets is almost same as the percentage of cache sets being affected by process variation. For instance, almost 75% of total accesses require single cycle latency in a cache with 25% process variation affected sets, as shown in first graph of Figure 6. This leads to 37.5% percent reduction in CAM occupancy over SPEC 2000 CPU benchmarks as compared to a worst-case design in the presence of process variation. Similarly, the proposed techniques result in 25% and 12.5% reduction in CAM occupancy for caches with 50% and 75% variation.

6. Conclusion

As the technology scales, process variation is emerging as one of the major challenges. Furthermore, memory designs are more susceptible to process variation due to the use of minimum sized transistors. In this paper, we presented analysis of two major types of CAM cells due to process, supply voltage and temperature variations. We also propose a novel architectural technique, which aims at making use of the variable CAM access latency instead of a worst-case design. Our technique yields an average of 37.5%, 25% and 12.5% reduction in CAM occupancy for caches with 25%, 50% and 75% variation respectively, over SPEC 2000 CPU benchmarks as compared to a worst-case design in the presence of process variation.

References

[1] Predictive Technology Model, http://www.eas.asu.edu/~ptm/
[2] SPEC 2000 Benchmarks, http://www.spec.org

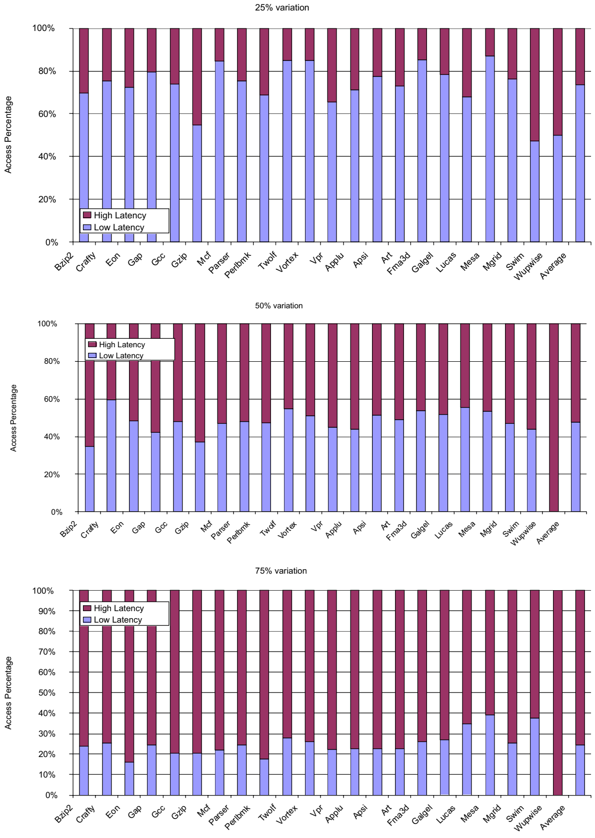

Figure 6. Access percentages for L1 data cache sets with x% of the sets affected by process variation, where x ∈ {25, 50, 75}

[3] Simplescalar toolset, http://www.simplescalar.com
[4] A. Chandrakasan, W. Bowhill and F. Fox, " Design of High Performance Microprocessor Circuits", IEEE Press, 2001.
[5] H. Chang and S. Sapatnekar, "Full-chip Analysis of Leakage Power Under Process Variations, Including Spatial Correlations", DAC, 2005, pp.523-528.
[6] S. Nassif, "Modeling and Analysis of Manufacturing Variations", CICC, 2001, pp.223-228.
[7] L.T. Pang and B. Nikolic, "Impact of layout on 90nm CMOS process parameter fluctuations", Symposium on VLSI Circuits, 2006.
[8] Y. Cheng, E. Rosenbaum, S. Kang, "Improvement on chip-level electrothermal simulator-ILLIADS-T", IEEE International Symposium on Circuits and Systems, 12-15 May 1996 Page(s):432 - 435 vol.4.
[9] J. A. Power, R. Clancy, W. A. Wall, A. Mathewson, W. A. Lane, "An investigation of MOSFET statistical and temperature effects", Proceedings of the 1992 International Conference on Microelectronic Test Structures, 16-19 March 1992 Page(s):202-207
[10] B. S. Amrutur and M. A. Horowitz, "A replica technique for wordline and sense control in low-power SRAMs," IEEE J. Soild-State Circuits, vol. 33, no. 8, pp.1208-1219, Aug. 1998.
[11] A. Agarwal, B. C. Paul, S. Mukhopadhyay and K. Roy, "Process Variation in Embedded Memories: Failure Analysis and Variation Aware Architecture", IEEE Journal of Solic-State Circuits, vol. 40, no. 9, pp 1804-1814,Sep. 2005.
[12] J.-S. Wang, H.-Y. Li, C.-C. Chen, and C. Yeh, "An AND-type matchline scheme for energy-efficient content addressable memories," IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, 2005, pp. 464-465.
[13] 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, no. 11, pp. 1690–1696, Nov. 1998.
[14] 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, March. 2006