Energy-aware cache schemes: a survey

Chentao Zhao

Department of Electrical and Computer Engineering

Michigan Technologic University, Houghton, United State

chentaoz@mtu.edu

Abstract-Cache consumes large part of energy in the modern processor. When we design an energy aware computer system, it is necessary to explore the cache schemes to adopt methods reducing the cache power consumption. This survey introduces a new 4 level taxonomy, summarizes recent energy aware cache schemes and techniques.

Keywords: SRAM, power reduction, cache scheme, taxonomy.

I. INTRODUCTION

Low consumption caches are widely required in nowadays computer design. In embedded system designing, we should consider not only the performance but also the system power consumption; Cache is widely used as the instruction and data memory in modern computer architecture; it has the high speed performance, however, cache memory occupies large processor layout areas, consuming roughly 20% to 60% of the processor power. In this survey paper, we will study various advanced cache schemes to reduce the cache memories' overall power consumption.

II. RELATED WORK

Related Survey has been performed by Robert P el at in 1982[1]. Their study looks very tutorial; it is a very prospective survey in that decades, but only fundamental principle and limited amounts of schemes was covered. The power was not concerned too much in there study.

Other cache relative surveys were usually made for memory classification, cache design and cache tuning. [2][3]

In conventional memory classification work, cache is equal to their physical prototype SRAM (Static Random Access Memory); it is well introduced as high speed memory in the memory group, but is not oriented to the energy aware cache field.

In related cache design work, performance concern is prior to power concern; power consumption of caches is sometimes revealed in the transistor and circuit level. Some fundamental power reduction schemes are also introduced; for example, the choice between 6TRAM (6 Transistor Random Access Memory) based caches and CAM (Content Addressable Memory) based caches is common topic in the cache design survey paper and textbook. However, few of

these surveys cover the most recent advanced schemes of cache power reduction such as drowsy cache scheme, hybrid cache scheme or MTJ cache scheme.

In cache tuning work, main stream simulation tool and tuning methods are introduced. These methods are typically practical for cache design and calibration for low power use, some circuit level schemes are mentioned [2] as additional introduction; However, the study is not fit for overview of whole picture in the cache power reduce technology field.

Different from these 3 kinds of survey mentioned above, this paper introduces a study mainly focusing on different kinds of the cache schemes of power reduction. In first section, a multi-level structure taxonomy is introduced in order to provide an overall view of the energy aware cache schemes. Then we will figure out the main idea of typical schemes under different species and discuss their performance.

III. THE TAXONOMY OF EVERGY-AWARE CACHE SCHEMES

This study introduces new taxonomy for low power cache schemes, which includes:

1) Transistor level cache power reduction
   - Improves the cache power consumption by improving the transistor performance.
   - Highly depends on the Very Large Scale Integration (VLSI) technique innovation.
   - Device is the basic unit in this level.
2) Circuit level cache power reduction
   - Improves the cache power consumption by improving the memory circuit design.
   - Depends on the transistor performance and the circuit design and tuning methods.
   - Memory cell circuit is the basic unit in this level.
3) Architecture level cache power reduction
   - Improves the cache power consumption by adjusting the memory system structure.
   - Depends on the RAM cell performance.
   - Function block is the basic unit in this level
4) Compiler level cache power reduction
   - Improves the cache power consumption by software control strategy.
   - Can be configured according to particular project feature.
   - Strategy sequence varies from case to case;

The taxonomy can be illustrated in figure 1;

One of the benefit of this 4 level taxonomy is, we can classify any particular energy aware cache scheme into 1 level by its main feature.

Multi-level structure also indicates that the evolution of the low level technique will generally bring huge impact to the whole scheme field, while high level scheme should be independent with the low level structure and have some potential to be implemented on different low level platform.

Generally this taxonomy would be a useful guide for both cache scheme innovation and cache scheme application. In the context below, we will introduce typical energy-aware cache schemes under each branches. Principle feature will be covered and there will be brief discussion about each scheme.

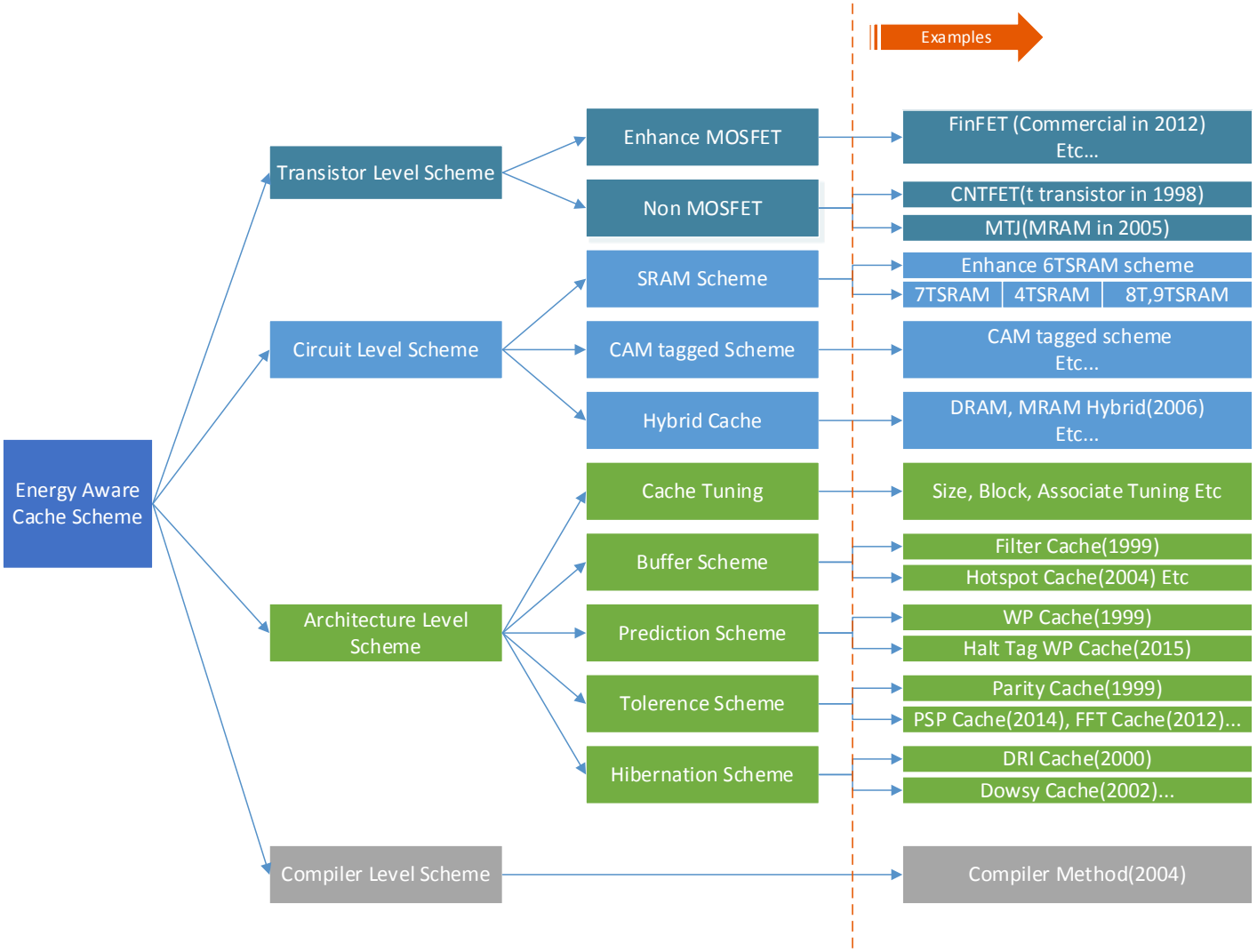

Figure 1: energy aware cache schemes taxonomy

IV. SURVEY OF ENERGY-AWARE CACHE SCHEMES

4.1 Transistor Level

Metal-Oxide-Semiconductor Field-Effect Transistor (MOSFET) is main stream of the cache transistor species; According to the moor's law, the number of transistors per square inch on integrated circuit doubles every year, new semiconductor fabrication technique will introduce smaller processor dimension, then reduces the on chip power consumption including the cache.

For conventional MOSFET below 90nm manufacturing technology, leakage current contributes the largest portion of the power consumption. Thus the most direct way we reduce the cache overall consumption is reducing the transistor leakage power.

Two direction of scheme to reduce the transistor leakage power have been studied; one is to enhance the leakage performance of the MOSFET, this kind of scheme can be described as enhance MOSFET scheme; the other strategy is to replace the MOSFET with other promising transistor species which has instinctive lower leakage performance for principle or material reason, this kind of schemes can be summarized into Non-MOSFET schemes;

Enhance MOSFET cache scheme

One of the popular transistor schemes in this species is FinFET technology, which was introduced as novel principle by University of California at Berkeley and NKK Corporation, japan in 1999 and implemented for commercial processor production by Intel in 2014 [4][5][6]. FinFET represents the species of MOSFET known as multi-gate MOSFET, each gate is named as a 'fin' made of silicon because the similar shape as a fish fin. FinFET achieves lower gate leakage together with very short gate length; in addition, its layout and fabrication is very similar to conventional MOSFET, which is then named as planar gate MOSFET.

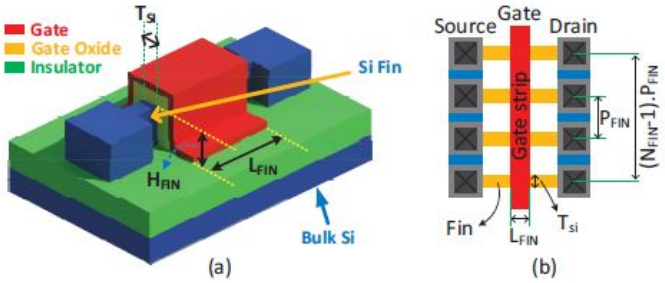

Figure 2: FinFET structure (from Chenming Hu, Univ. of California)

Similar species includes ultra-thin body silicon on insulator (UTB-SOI), which is alternative to the multi-gate MOSFET with deep-sub-tenth micron technology. UTB-SOI was introduced by the same study group as FinFET [7][8]. Both of these enhanced MOSFET schemes have the prospect to be fabricated as cache with lower consumption than conventional planar gate MOSFET.

Enhance MOSFET schemes share some feature: the schemes usually improve both the performance of the processor and memory, in other words, this type of technology usually improve the power consumption without performance compromise. Thus for benchmark, we need to take account the leakage performance of the transistors. The problem of the enhance CMOS schemes are the change of the circuit design since more gates will be involved, for SRAM cell of the cache particularly, we may have additional auxiliary circuitry overhead. However, it is still the most feasible transistor level scheme to improve the cache power consumption performance.

Non-MOSFET Cache scheme

Non-MOSFET Cache scheme introduces new transistors other than MOSFET; various novel and realized transistors have been introduced to reduce the leakage using innovation ideals from wide technique fields. Here we mainly introduce two typical types of Non CMOS Cache schemes: Carbon Nanotube Field Effect Transistor (CNFET) cache and Magnetic Tunnel Junction (MTJ) cache.

Carbon nanotube is used to fabricate the field effect transistors especially to improve the gate leakage performance [9][10][11]. SRAM cells also have been developed using CNFET and it is thought to be a promising candidate for electronic devices in nanometer scale [11]. In fact, circuit simulation results indicate the CNFET leakage instinctive reduction can reduce the SRAM cell leakage power as much as 90% compared with that of a FinFET and as much as 99% compared with the conventional MOSFET [11]. However, the cost of the nanotube technique is the bottleneck of the CNFET stepping into commercial stage.

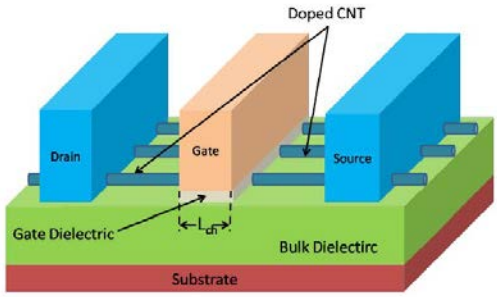

Figure 3: CNFET structure (from [10])

MTJ devices have different quantum principle from MOSFET: intrinsic spin of the electron and its associated magnetic moment make contribution to the gate function, other than the diffusion effect; First MTJ RAM was fabricate by IBM in 2005, known as MRAM[12].From then on, it became more and more popular for its unique characteristic. There are two kinds of MTJ devices with different working principles: (Spin Transfer Torque) STT-MRAM and (Spin Orbit Torque) SOT-MRAM, they are separately the device as figure4.

In MTJ RAM cell, data is stored as magnetic form other than electric charge form. Thus static leakage can be basically prevented in this RAM. Other benefits such as instinctive non-violate feature and high storage density make MRAM very popular in nowadays memory technology. Generally speaking, MRAM introduces more write latency than conventional SRAM, However, for the read operation of the large RAM size (typically 8M), the MRAM can be even faster than SRAM [13]. As the high level cache especially shared cache demands higher and higher density, MTJ and MRAM scheme is feasible way to perform better leakage, especially in the Multi-core SOC case.

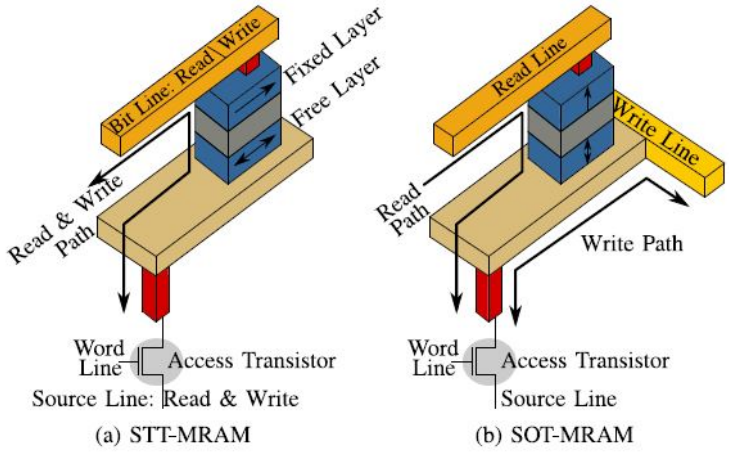

Figure 4: MTJ structure (From [12])

4.2 Circuit level

In circuit level, we are talking about circuit design for low power consumption memory cell.

Typical 6TSRAM cell is the basic circuit used for cache memory. Other similar RAM cells are introduced to improve the performance of the memory in certain circumstances, including 4TSRAM, 7TSRAM, 8TSRAM and 9TSRAM. For these xTSRAM series memory cell, similar power strategies have been implemented, then we summarize these schemes into SRAM schemes;

Different from direct cache, associated cache contains an addition tagged array; basically, this array can be realized by SRAM circuits or alternate circuits called Content Addressable Memory (CAM); CAM tagged cache scheme introduces different power pattern, so we separate CAM tagged schemes from the SRAM schemes.

Hybrid cache means we combine the SRAM cells with non-violated memory units such as MRAM, DRAM and PRAM (Phase-change Random Access Memory); as mentioned before, the multiprocessor structure introduces large size shared cache, which can be optimized by Hybrid Cache Architecture design (HCA) instead of using low density SRAM circuit. Here I classify HCA into a cache scheme species.

SRAM scheme

6TRAM means for 1 bit storage, a RAM cell needs 6 transistors to form the memory cell circuit. Typical 6TRAM cell and the RAM array is illustrated as figure5.

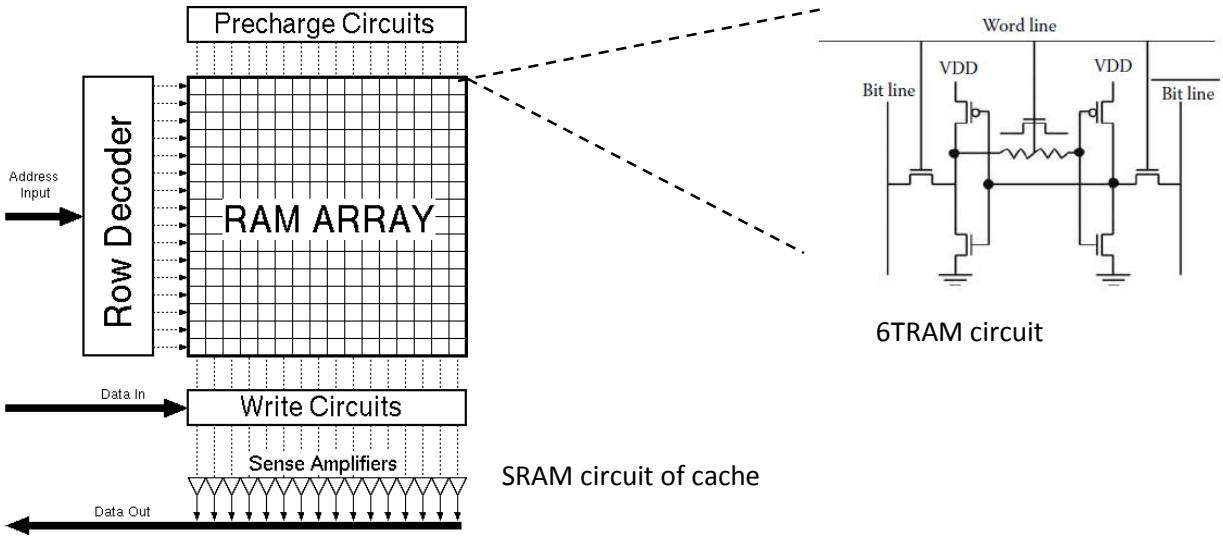

Figure 5: typical 6TRAM cache circuit

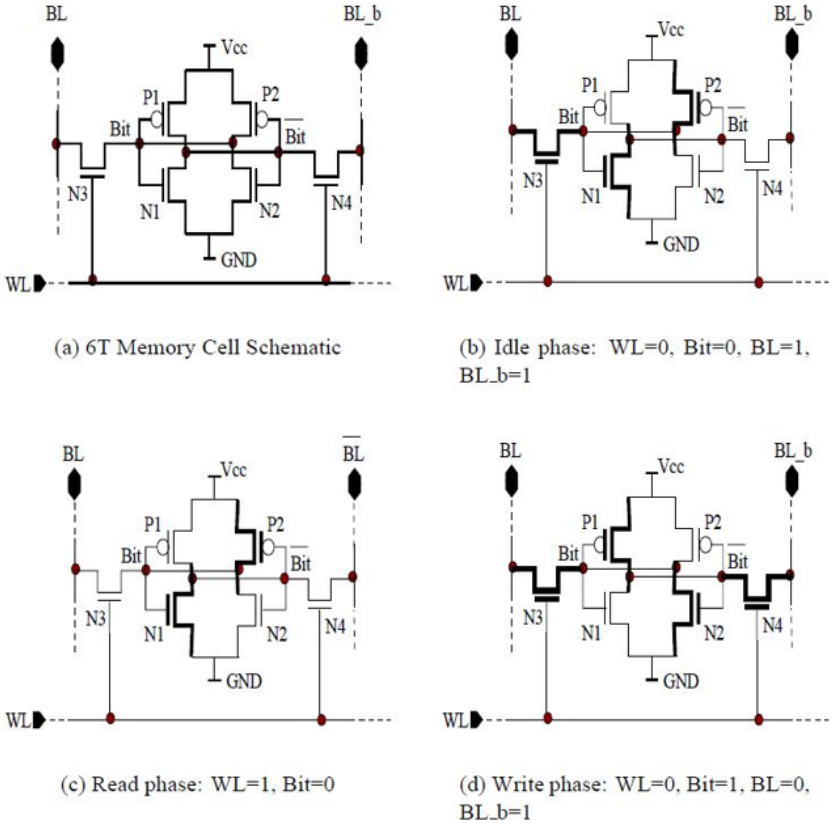

Figure 6: 6TRAM leakage activity(from [14])

In 6TRAM circuit, leakage happens during every read/write operation and even idle states. The transistors with leakage power consumption are bold in figure6.

7TSRAM, 8TSRAM and 9TSRAM introduce addition transistors to feedback the access condition based on 6TSRAM circuit in order to improve the RAM performance. Novel 4TSRAM is also introduced. Less transistors bring higher storage density and less latency; however, 15% more leakage is introduced as penalty [15].

Besides leakage, load and store operation also introduce dynamic power consumption. Generally, the dynamic power consumption can be summarized as the equation:

$$\boldsymbol{P_{dynamic} = \alpha \times C \times f \times V^2}$$

Where α is the access rate, C is the transistor capacitance, f is the system frequency, V is the supply voltage. Since C is determined by the transistor and the wire characteristic,α and f are hardly adjustable since performance issue, traditionally, Dynamic Voltage Scaling (DVS) is introduced to reduce the power consumption of the memory. DVS dynamically adjust the supply voltage level upon the working circumstances. As overhead, DVS scheme needs additional auxiliary circuits to monitor the working condition and adjust the voltage.

CAM tagged scheme

The selection between SRAM and CAM for the tagged array strategy has been discussed from the day cache was introduced, the reason is that the consumption component is different for RAM and CAM strategy, as figure below [16]. For a RAM tagged cache, 35% power is consumed by comparator and 14% power is consumed by decoder; For a CAM tagged cache, power consumption is mainly distributed on the physical address bus (wire capacitance consumption), which indicates we can get better dynamic power performance when the cache has less switching factor, which usually happens on the multicore or multi-tread processor system. Here switching factor means the percentage of the signal switching from cycle to cycle [16].

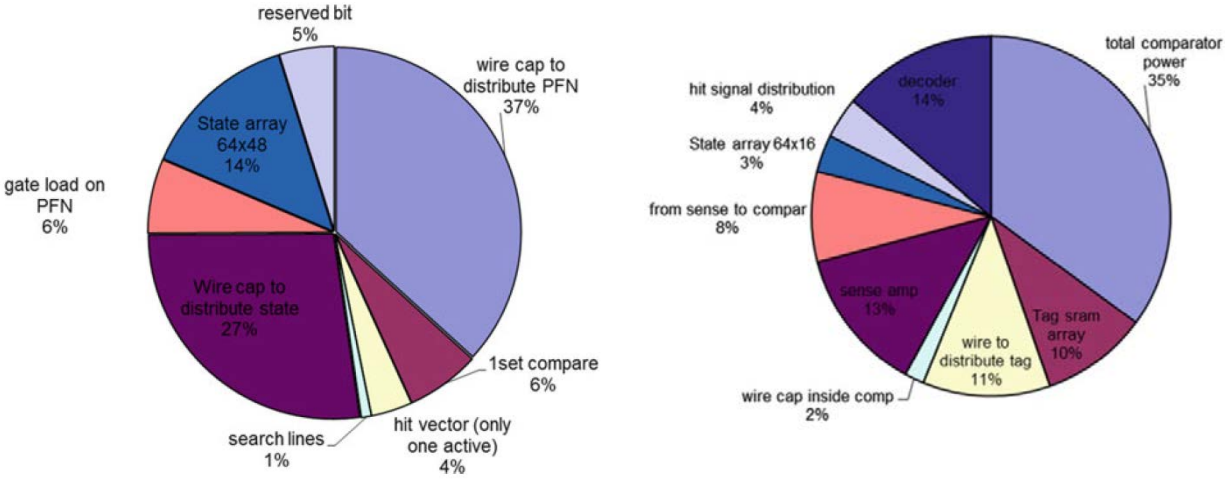

Figure 7: Different power pattern: RAM and CAM(From [16])

Hybrid Cache scheme

Hybrid Cache scheme has a specific background, that is the popularity of the multi-core system. multi-core introduce large Last Level Cache (LLC), conventional SRAM are not fit in this application since its poor density and high power consumption; on the other hand, Non-violated memory reduce the latency gap from the SRAM, such as MRAM we mentioned above. Since LLC can tolerant more latency, hybrid scheme has been widely explored in may project and the relevant synchronizing circuits have been introduced. [13][17][18][19]

In fact, mostly recent research have indicated the hybrid cache have better performance when the cache size is relatively large, as figure8[20]; we can see this “relative large” level is about 5 MB for MRAM, 15MB for DRAM and 45MB for PRAM. It is quite reasonable size considering the LLC of the share cache.

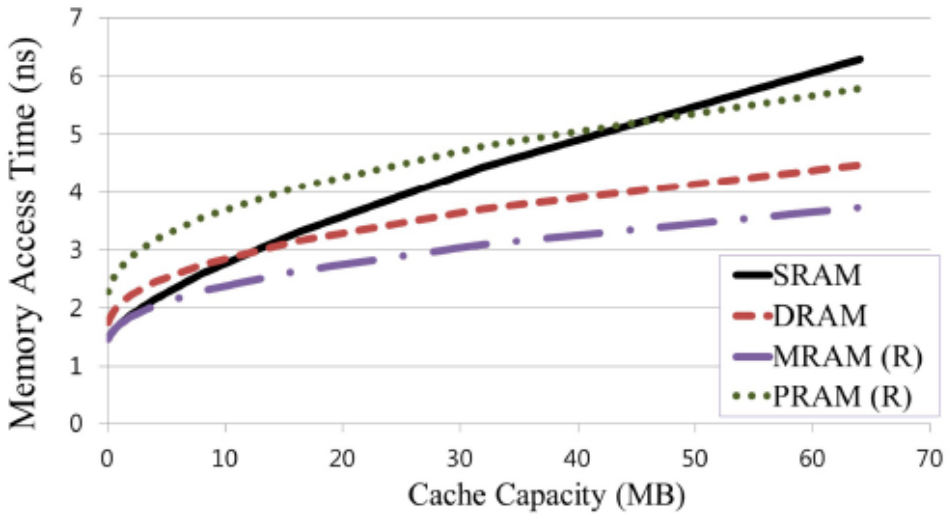

Figure 8: RAM access time versus cache size (From [20])

4.3 Architecture level

In the architecture level, we consider the overall energy performance of the cache with the machine structure. Basically, the consumption of the cache can be summarized as equation below:

$$\text{Pcache} = N(C \times V^2 + \beta \times k \times E_m + \alpha \times 2 \times E_m) + V \times I_L$$

Where, N is access rate; the C is the cache capacitance per access switching; β is the ratio of cache-writes to the total number of cache access; α is the cache miss rate; k is the write-policy dependent coefficient; E_m is the energy required by lower one level memory access. All of these are machine structure relevant parameters, the others are supply voltage and the leakage current.

Conventional cache tuning

Before we introduce cache architecture technique, we should notice that in embedded system design, cache system can be tuned to achieve the optimized power performance.

Papers has been published to introduce the main stream cache tuning Methods [1], in this paper, cache tuning are defined as the process of determining the best cache configuration in the design space for a particular application or an application's phrase [1]. Cache configuration includes cache size, block size and associativity. Proper configuration of the cache can effectively reduce the cache energy consumption. For this purpose, various tuning methods are introduced in this paper, including Offline static cache tuning, in which cache tuning performed at design time through simulation methods, and online dynamic cache tuning, in which hardware monitor resources are required.

Cache tuning is a directly way for us to optimize the cache power performance in system design project. The advantage is obvious, we optimize both the performance and power consumption, in other words, achieve power consumption with little performance compromise; without any extra hardware overhead but simple modify cache configuration. The bottleneck is also obvious that we cannot achieve too aggressive power performance, in other words, we are confined by the existing hardware and architecture limitation.

In order to solve this problem, we should explore innovative architecture in instruction level. We can divide all this architecture innovation into three classes according to its scheme feature: buffer scheme: prediction scheme, which is trying to predict the economic way we access the cache; Tolerant scheme, which is trying to complement the performance penalty caused by aggressively voltage reduction; hibernated scheme, which is trying to reduce the leakage energy of the idle cache zone.

Buffer scheme

Buffer scheme introduces a direct-mapped buffer or filter cache between the processor and the I-cache. Paper published by Johnson K et al, 1999 used experiment illustration the filter cache can effectively reduce the cache power, but introduces new buffer access latency as well. The power reduction was measured as about 51% across a set of 19 multimedia and communications applications [21].

Improved Buffer scheme is introduced by Chia-Lin Y. et al, 2004. The cache architecture is named as HotSpot Cache since a dynamic steering block which looks like is added to perform higher utilization of the filter. The energy reduction provided by simulation is announced to be 52% on the average, which is similar to the filter cache, however hotspot cache has no cost of performance degradation [22].

The main principle filter cache or HotSpot cache works for reducing the power consumption is that this architecture is efficient in reducing the cache access rate, the processor fetch the most frequently used instruction from the filter cache, instead of access the larger L1 cache.

Prediction scheme

Prediction Scheme optimizes the way we access the cache memory cell data array, so that we reduce the cache capacitance switching activity for each access.

Koji I et al, 1999 introduce a methods of accessing a single cache predicted instead all the ways in a set to reduce the energy cost in set-associative caches. This architecture, named as way-predicting cache choose one way before starting the normal cache accessing process. Way is chose by an algorithm of MRU (Most Recently Used), if the predicting is not correct, the cache search the other way. This way-prediction scheme announced 70% power reduction without any performance impair [23].

In 2015 Neethu.M et al improve the way predicting methods by introduce halt tags for the early detection of probably ways in decode cycle of the cache. The prediction circuit offer the MRU way as long as more than one halt tag exists concurrently. If the predicted way missed, ways with halt tag hit are accessed at the same time; in this way we can fast detect the miss when no halt tag hit exist. This extra halting circuit improve the prediction accuracy of way prediction (WP) methods, Using this methods in conventional caches, experiments illustrate a 5% more power reduction than WP architecture [24].

Tolerant scheme

Aggressive voltage strategy can impact the vulnerable cache storage and introduce soft errors, which is mainly Single Event Upsets (SEUs) and Single Event Multiple Bits Upsets (SEMUs) in the cache memory. In other words, highly tolerant scheme is fundamental for aggressive voltage reduction scheme.

To reduce these errors impair, Error Detection Codes (EDC) and Error Correction Codes (ECC) are widely used in modern processors and caches. We call these cache parity caches. Lots of study(Seongwoo K et al 1999) has been performed in this cache information integrity fields, the main target is to increase the cache ability to detect and correct and correct the soft errors at the same time to achieve as low performance and power overhead as possible [25] .

Studies has been made to improve the performance of the EDC/ECC methods. Hamed F. et al introduced Per-Set Protected Cache (PSP Cache) of L1 Cache in 2014, which minimize the number of redundant bits remaining the protection feature of EDS/ECC. In their cache architecture, all data are accessed simultaneously share a single EDC instead of occupying their own one. This improve bring 73% less energy overhead than parity caches [26].

In 2012, Abbas B et al introduce another fault tolerant cache architecture implementing an outside block to handle the errors caused by ultra-low-voltage, called Flexible Defect Map (FDM). This new cache architecture was named Flexible Fault-Tolerant Cache (FFT-Cache). In FFT Cache a portion of the cache can be treated as redundancy when the system detect the ultra-low-voltage condition. In this way, this architecture try to use minimum number of cache access overhead to tolerate the maximum amount of defect. Benchmark results show this architecture can achieve dramatically energy saving (66% dynamic power with 48% static power for L1 cache and 80% dynamic power with 42% static power for L2 cache) combined with ultra-low-voltage of 375mv. The compromise is just 5% performance impair with 13% area overhead [27].

Hibernated scheme

In this scheme species, instruction makes the memory units hibernate or Shut down some section of the cache during low load and then reactive those when high load so that static leakage power consumption can be reduced.

Michael P et al in 2000 introduced Gated-Vdd technique to reduce the leakage consumption, they introduced a Dynamic Resizable (DRI) i-cache architecture, in which the number of sets increase or decrease due to usage times of this I cache set. That is equivalent the size of the i-cache is dynamically controlled by the miss rate. The simulation results indications 62% on average improving in power performance and only 4% impact on the speed [28].

Krisztian F et al introduced drowsy caches architecture, aiming at reducing leakage power. They chose another way other than turn off the cache partition but to keep the cache with few use in a low power "drowsy" mode, and wake up when the cache need to be accessed. Extra circuit had been introduce into the SRAM to realize this drowsy and wake up action. With drowsy caches technique, 24% power reduction can be achieved. Since wake up action introduces extra latency, the performance is impacted about 1.2% when more than 74% caches partition is at drowsy states [29].

Chuanjun Z et al introduced an improved data cache architecture based on the observation fact that in data cache, a major portion of data access involves frequent values, which can be managed in an encoded cache with only a few bits occupation. Thus long addresses of frequent values can be encoded into few bits and occupy only a small array of the data cache. Larger portion of the data cache can be implemented shutting off action. With this Frequent-value data cache architecture 33% static energy reduction can be achieved compared with conventional 32-bit per word cache. No miss rate as overhead [30].

4.4 Compiler level

Cache power reduction schemes are also studied at higher compiler level. In 2012, W.Zhang et al introduced a cache leakage power reduction circuits combined with compiler support. In this scheme, the cache memory have 3 states: active, state-preserving and state-destroying with the power set to be 1.0V, 0.3V, and 0V sequentially. These states are shifted by compiler, which insert relative instruction during compiling [31]. The scheme diagram is as figure below:

For the compiler level cache schemes, we can see the obvious advantage is that, we can configure the cache according to different project requirement in a flexible way. The overhead is that this scheme relays on specific instructions to switching caches states [31]. Extra hardware overhead (similar to drowsy caches scheme) also should be taken into account.

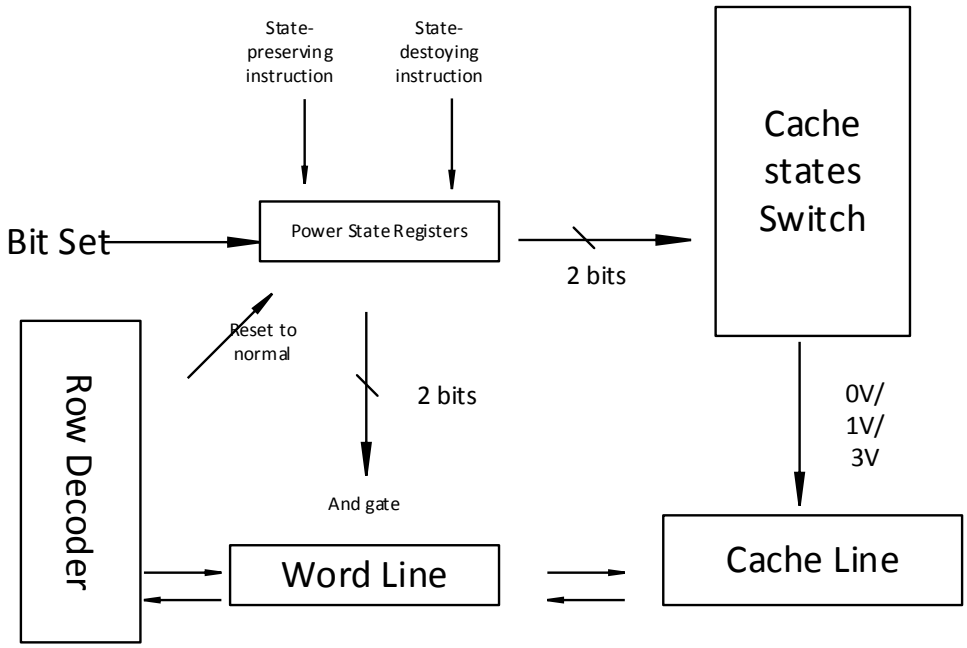

Figure 9: typical compiler level scheme

V. CONCLUSION

This paper presents the survey of the recent low-power-reduction cache techniques and schemes. For combing various new techniques and schemes, a brand new 4 level taxonomy is introduced for energy aware cache schemes. Then we introduce typical cache schemes from low level to the top with this taxonomy. Under this frame, various cache power reduction research can be well classified; we can explore the overview and prospect the advancing direction of the energy aware cache techniques. We also explore typical principle and performance of various cache schemes, then summarizing the prospect and bottleneck of these techniques.

Reference

[1] Cook, Robert P., Cathy J. Linn, Joseph L. Linn, and Terry M. Walker. "Cache memories: A tutorial and survey of current research directions." In Proceedings of the ACM'82 conference, pp. 99-110. ACM, 1982.

[2] Zang, Wei, and Ann Gordon-Ross. "A survey on cache tuning from a power/energy perspective." *ACM Computing Surveys (CSUR)* 45, no. 3 (2013): 32.

[3] Kang, Yong Hoon. "Memory Systems for Nano-computer." In CMOS Processors and Memories, pp. 165-196. Springer Netherlands, 2010.

[4] Huang, Xuejue, Wen-Chin Lee, Charles Kuo, Digh Hisamoto, Leland Chang, Jakub Kedzierski, Erik Anderson et al. "Sub 50-nm FinFET: PMOS." In Electron Devices Meeting, 1999. IEDM'99. Technical Digest. International, pp. 67-70. IEEE, 1999.

[5] Turi, Michael A., and Jose G. Delgado-Frias. "An evaluation of 6T and 8T FinFET SRAM cell leakage currents." In Circuits and Systems (MWSCAS), 2014 IEEE 57th International Midwest Symposium on, pp. 523-526. IEEE, 2014.

[6] Karl, Eric, Zheng Guo, James W. Conary, Jeffrey L. Miller, Yong-Gee Ng, Satyanand Nalam, Daeyeon Kim, John Keane, Uddalak Bhattacharya, and Kevin Zhang. "17.1 A 0.6 V 1.5 GHz 84Mb SRAM design in 14nm FinFET CMOS technology." In Solid-State Circuits Conference-(ISSCC), 2015 IEEE International, pp. 1-3. IEEE, 2015.

[7] Hu, Chenming. "Thin-body FinFET as scalable low voltage transistor." In VLSI Technology, Systems, and Applications (VLSI-TSA), 2012 International Symposium on, pp. 1-4. IEEE, 2012.

[8] Samsudin, K., B. Cheng, A. R. Brown, S. Roy, and A. Asenov. "UTB SOI SRAM cell stability under the influence of intrinsic parameter fluctuation." In Solid-State Device Research Conference, 2005. ESSDERC 2005. Proceedings of 35th European, pp. 553-556. IEEE, 2005.

[9] Martel, R. al, T. Schmidt, H. R. Shea, T. Hertel, and Ph Avouris. "Single-and multi-wall carbon nanotube field-effect transistors." Applied Physics Letters 73, no. 17 (1998): 2447-2449.

[10] Kim, Young Bok, Yong-Bin Kim, Fabrizio Lombardi, and Young Jun Lee. "A low power 8T SRAM cell design technique for CNFET." In SoC Design Conference, 2008. ISOCC'08. International, vol. 1, pp. I-176. IEEE, 2008.

[11] Lin, Sheng, Yong-Bin Kim, Fabrizio Lombardi, and Young Jun Lee. "A new SRAM cell design using CNTFETs." In SoC Design Conference, 2008. ISOCC'08. International, vol. 1, pp. I-168. IEEE, 2008.

[12] Gallagher, William J., and Stuart SP Parkin. "Development of the magnetic tunnel junction MRAM at IBM: From first junctions to a 16-Mb MRAM demonstrator chip." IBM Journal of Research and Development 50, no. 1 (2006): 5-23.

[13] Park, Sang Phill, Sumeet Gupta, Niladri Mojumder, Anand Raghunathan, and Kaushik Roy. "Future cache design using STT MRAMs for improved energy efficiency: devices, circuits and architecture." In Proceedings of the 49th Annual Design Automation Conference, pp. 492-497. ACM, 2012.

[14] Mamidipaka, Mahesh, Kamal Khouri, Nikil Dutt, and Magdy Abadir. "Leakage power estimation in SRAMs." Univ. of Cal. Irvine CECS Techn. Report (2003): 03-32.

[15] Mazreah, Arash Azizi, Mohammad Reza Sahebi, Mohammad Taghi Manzuri, and S. Javad Hosseini. "A novel zero-aware four-transistor SRAM cell for high density and low power cache application." In Advanced Computer Theory and Engineering, 2008. ICACTE'08. International Conference on, pp. 571-575. IEEE, 2008.

[16] Mohammad, Baker. "Cache Architecture and Main Blocks." In Embedded Memory Design for Multi-Core and Systems on Chip, pp. 13-28. Springer New York, 2014.

[17] Sun, Zhenyu, Xiuyuan Bi, Hai Li, Weng-Fai Wong, and Xiaochun Zhu. "STT-RAM Cache Hierarchy With Multiretention MTJ Designs." *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on* 22, no. 6 (2014): 1281-1293.

[18] Park, Sang Phill, Sumeet Gupta, Niladri Mojumder, Anand Raghunathan, and Kaushik Roy. "Future cache design using STT MRAMs for improved energy efficiency: devices, circuits and architecture." In *Proceedings of the 49th Annual Design Automation Conference*, pp. 492-497. ACM, 2012.

[19] Ikegami, K., H. Noguchi, C. Kamata, M. Amano, K. Abe, K. Kushida, E. Kitagawa et al. "Low power and high density STT-MRAM for embedded cache memory using advanced perpendicular MTJ integrations and asymmetric compensation techniques." In *Electron Devices Meeting (IEDM), 2014 IEEE International*, pp. 28-1. IEEE, 2014.

[20] Lee, Suji, Jongpil Jung, and Chong-Min Kyung. "Hybrid cache architecture replacing SRAM cache with future memory technology." In Circuits and Systems (ISCAS), 2012 IEEE International Symposium on, pp. 2481-2484. IEEE, 2012.

[21] Kin, Johnson, Munish Gupta, and William H. Mangione-Smith. "The filter cache: an energy efficient memory structure." In *Proceedings of the 30th annual ACM/IEEE international symposium on Microarchitecture*, pp. 184-193. IEEE Computer Society, 1997.

[22] Ali, Kashif, Mokhtar Aboelaze, and Suprakash Datta. "Modified hotspot cache architecture: A low energy fast cache for embedded processors." In *Embedded Computer Systems: Architectures, Modeling and Simulation, 2006. IC-SAMOS 2006. International Conference on*, pp. 35-42. IEEE, 2006.

[23] Inoue, Koji, Tohru Ishihara, and Kazuaki Murakami. "Way-predicting set-associative cache for high performance and low energy consumption." In*Proceedings of the 1999 international symposium on Low power electronics and design*, pp. 273-275. ACM, 1999.

[24] Mallya, Neethu Bal, Geeta Patil, and Biju Raveendran. "Way Halted Prediction Cache: An Energy Efficient Cache Architecture for Embedded Processors." In*VLSI Design (VLSID), 2015 28th International Conference on*, pp. 65-70. IEEE, 2015.

[25] Kim, Seongwoo, and Arun K. Somani. "Area efficient architectures for information integrity in cache memories." *ACM SIGARCH Computer Architecture News* 27, no. 2 (1999): 246-255.

[26] Farbeh, Hamed, and Seyed Ghassem Miremadi. "PSP-cache: a low-cost fault-tolerant cache memory architecture." In *Proceedings of the conference on Design, Automation & Test in Europe*, p. 164. European Design and Automation Association, 2014.

[27] BanaiyanMofrad, Abbas, Houman Homayoun, and Nikil Dutt. "Fft-cache: A flexible fault-tolerant cache architecture for ultra low voltage operation." In*Proceedings of the 14th international conference on Compilers, architectures and synthesis for embedded systems*, pp. 95-104. ACM, 2011.

[28] Powell, Michael, Se-Hyun Yang, Babak Falsafi, Kaushik Roy, and T. N. Vijaykumar. "Gated-V dd: a circuit technique to reduce leakage in deep-submicron cache memories." In *Proceedings of the 2000 international symposium on Low power electronics and design*, pp. 90-95. ACM, 2000.

[29] Flautner, Krisztián, Nam Sung Kim, Steve Martin, David Blaauw, and Trevor Mudge. "Drowsy caches: simple techniques for reducing leakage power." In*Computer Architecture, 2002. Proceedings. 29th Annual International Symposium on*, pp. 148-157. IEEE, 2002.

[30] Zhang, Chuanjun, Jun Yang, and Frank Vahid. "Low static-power frequent-value data caches." In *Proceedings of the conference on Design, automation and test in Europe-Volume 1*, p. 10214. IEEE Computer Society, 2004.

[31] Zhang, Wei, Jie S. Hu, Vijay Degalahal, M. Kandemir, Narayanan Vijaykrishnan, and Mary Jane Irwin. "Reducing instruction cache energy consumption using a compiler-based strategy." *ACM Transactions on Architecture and Code Optimization (TACO)* 1, no. 1 (2004): 3-33.