International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Reduction of Dark Silicon through Efficient Power Reduction Designing of Un-core Components in 3D CMPs

Angel Ann George Yesudasan

Student, IETE Member, Dept. of ECE, Christ Knowledge City, APJ Abdul Kalam Technological University, Kerala

---

Abstract – *THIS review presents an effective architecture that can be used in CMPs to reduce dark silicon. The design targets on the various homogeneous and heterogeneous structures used in un-core components of chip multiprocessors in order to reduce heat dissipation and increase efficiency. The up comings and shortcomings of existing circuit techniques like source biasing, dual V partitioning, VTCMOS and MTCMOS are compared with the newly designed un-core components structures in CMP. On the basis of this design, it is proposed that heterogeneous 3D CMPs are the best to compensate the dark silicon age in VLSI industry.*

Key Words: **Chip Multi-Processors (CMPs), VTCMOS (Variable Threshold CMOS), MTCMOS (Multi-Threshold CMOS), STT-RAM (Spin Torque Transfer RAM), NoC (Network on Chips), Non-Volatile Memories (NVMs)**

1. INTRODUCTION

VLSI technology is the core of every chip multiprocessors. In other words, it is the basic building block of electronics industry. It has been very challenging for the researchers to develop the CMPs under certain constraints like power and temperature. Several techniques and methods have been developed to reduce power consumption without losing the efficiency to an extent. A number of circuit techniques have been developed to reduce leakage current and heat dissipation. But neither of these techniques reduces power dissipation and heat due to hardware structures like interconnects in the chips and un-core components like memory modules.

Now focus has shifted to energy conservation in un-core components like cache hierarchy and chip interconnects. Researchers have shown that 50% of total power dissipation in CMOs are due to leakage power in these un-core components. As the number of transistors increases, the Dennard scaling breaks down. This is a challenge in VLSI industry. But this can be solved if we are able to make use of maximum power that is being supplied to the device because at present stage, a huge part of power supplied to the chip is lost due to heat dissipation. Reducing heat dissipation leads to reduction in dark silicon.

Studies usually focus on the core components like memory to solve this problem. Here the un-core components like on-chip interconnects and cache hierarchy system are taken into consideration for minimizing the total power budget. 3D integrated circuits and heterogenous cache hierarchy are combined together in the CMP architecture to destroy the dark silicon age.

2. LITERATURE REVIEW

2.1 Source Biasing and Dual V Partitioning

Source biasing is the process of applying a positive voltage to the source terminal of a device in OFF state during stand by state. When source biasing is done, the body effect occurs. As a result, the threshold voltage of the system raises which leads to a significant reduction in sub-threshold leakage current. Consequently, the gate-source voltage changes to negative value. The overall effect is that the device in OFF state continues to be in a more "strong OFF" stage. Thus, during standby modes, the leakage currents can be reduced. It is the basic technique for reducing the leakage current.

A dual VTH partitioning scheme provides the designer with transistors that are either fast (with a high leakage) or slow (with a low leakage). In modern process technology, multiple threshold voltages are provided for each transistor. Therefore, a circuit can be partitioned into high and low threshold voltage gates or transistors, which is a trade-off between propagation delay and reduced leakage current

Both of these techniques do not take care of the heat dissipation in the wire interconnects in the CMPs.

2.2 VTCMOS and MTCMOS

In variable threshold CMOS (VTCMOS) technique, the substrate bias voltage is dynamically varied to control the threshold voltage of MOS transistors. Since the subthreshold leakage current drops exponentially with increasing threshold voltage, the leakage power dissipation in the standby state can be significantly reduced with this circuit design technique. Multi-threshold CMOS is a very effective circuit technique to reduce the leakage current of a logic circuit in the standby mode. In this technique, low VTH transistors are used to design the logic circuit for which the switching speed is essential and high VTH transistors (also called sleep transistors) are used to effectively isolate the logic circuit from VDD and GND in the standby state and thus effectively reduce the standby subthreshold leakage power dissipation. I

Even though they provide effective isolation, the effectiveness of VTCMOS technique reduces as the channel length becomes smaller, or the VTH values are lowered. VTCOMS is intrinsically more problematic for reliability since

the high voltage across the oxide decreases the lifetime of the device. They suffer from higher standby subthreshold leakage power dissipation, which can become a great problem for portable systems where a larger amount of this leakage power is dissipated during the long standby period.

2.3 Optimized 3D Stacked Memory Architecture

SMART-3D is a 3D-stacked memory architecture with a vertical L2 fetch/write-back network using a large array of TSVs. They ameliorate latency by exploiting the excessive, high-density bandwidth of TSV between the processor last level cache and the 3D DRAM. Upon each L2 miss, the SMART-3D architecture fetches an entire page of data but keeps the caches organized by 64B lines to avoid complicating coherency. This adaptive SMART-3D design helps to mitigate the false sharing problem in a multi-socket system.

They lower the energy consumption in the L2 cache and 3D DRAM and reduces the total number of row buffer misses. For single-threaded memory intensive applications, the speedups range increases. But they deal only with un-core components.

2.4 STT MRAM Technology in NVMs

In STT-MRAM (also called STT-RAM or sometimes ST-MRAM and ST-RAM) is an advanced type of MRAM devices. STT-MRAM enables higher densities, low power consumption and reduced cost compared to regular (so-called Toggle MRAM) devices. The main advantage of STT-MRAM over Toggle MRAM is the ability to scale the STT-MRAM chips to achieve higher densities at a lower cost. High-performance SRAM technology has been widely used in the on-chip caches due to its standard logic compatibility, high endurance, and fast access time features.

Even though it has ability to scale the STT-MRAM chips to achieve higher densities at a lower cost, the amount of current needed to reorient the magnetization is at present too high for most commercial applications.

3. PROPOSED ARCHITECTURE

In the proposed architecture for future CMPs, emerging technologies such as non-volatile memories (NVMs) and 3D techniques are taken into account to combat dark silicon. The architecture model assumed is based on a 3D CMP with multi-level hybrid cache hierarchy stacked on the core layer. Each cache level is assumed to be implemented using a different memory technology.

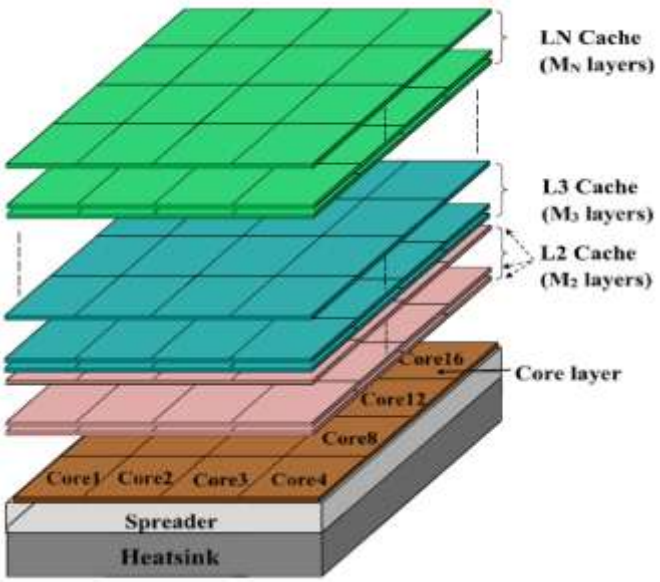

Fig -1: Heterogeneous Cache Hierarchy Model

Based on AMAT (Average Memory Access time), evaluation of the cache systems performance and system power consumption when the stacked cache levels in the homogenous hierarchy are made from SRAM, EDRAM, STT-RAM, or PRAM.

Technology	AMAT	Power Consumption
SRAM	0.09	1
EDRAM	0.16	0.62
STT-RAM	0.3	0.37
PRAM	1	0.22

Fig -2: Comparison of AMAT and system power consumption

Power consumption is normalized with respect to the SRAM, whereas AMAT is normalized with respect to the PRAM. Based on these views, SRAM is the fastest and a higher power-hungry option and it is better to be used in lower levels of the cache hierarchy to support faster accesses. According to the observations in Fig 2, use SRAM in the L2 cache level, EDRAM in the L3 cache level, STT-RAM in the L4 cache level, and PRAM in the L5 cache level as shown in Fig 1.

Because of strong thermal correlations between a core and cache banks directly stacked on the core, the core and the cache banks in the same stack is called a core set in this architecture (Fig 3).

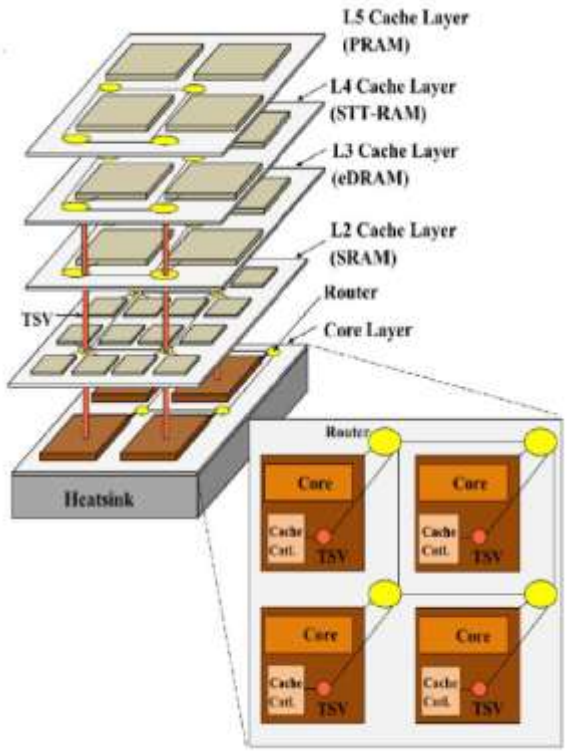

Fig -3: 3D CMP architecture with heterogeneous cache hierarchy

The total power consumption of a CMP mainly comes from three on-chip resources: cores, cache hierarchy, and interconnection network. CMPs with a large number of cores (more than eight) require building architectures through a scalable network-on-chip (NoC). The total power consumption of a 3D CMP can be calculated as the sum of the power of individual on-chip resources (core and un-core components). P denotes power.

P(TOTAL) = P(CORES) + P(UNCORES)

P(TOTAL)=P(CORES)+P(CACHEHEIRARCHY)+P(INTERCONNECTION)

Maximizing performance under power constraint is an important target in digital system design in these days. The peak power dissipation during the entire execution must be less than the maximum power budget.

The cache power consumption in multi-program workloads can be reduced by allocating each program its own dedicated cache banks in its own core set privately (Fig 4(a)). Larger classes of multi-threaded applications are based on barrier synchronization and consist of two phases of execution: a sequential phase, which consists of a single thread of execution, and a parallel phase in which multiple threads process data in parallel. The parallel threads of execution in a parallel phase typically synchronises on a barrier.

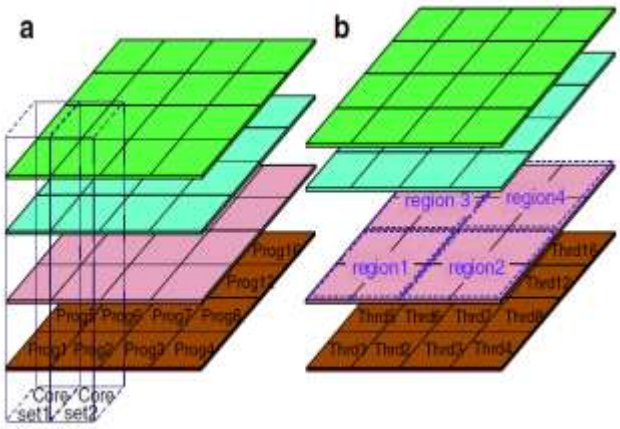

Figure -4: The style of using cache hierarchy in a (a) multi-program workload and (b) a multithreaded workload

In parallel phase, all threads must finish execution before the application can proceed to the next phase. In multi-threaded workloads, cache levels are shared across the threads. In parallel phase, threads share regions at each layer of the cache levels in the hierarchy. For example, for a performance-maximization problem with respect to power budget, first, a fixed region 1 as shown in Fig 4(b), is dedicated in each level to the threads, as an initialized value. Then based on power budget and other constraints in the optimization problem, the number of regions can be increased or keep it fixed in each level in order to obtain the maximum performance.

In order to validate the efficiency of 3D CMP architectures in this work, we employed a detailed simulation framework driven by traces extracted from real application workloads running on a full-system simulator. The traces can be extracted from the GEM5 full-system simulator. For simulating a 3D CMP architecture, the extracted traces from GEM5 can be interfaced with 3D Noxim, as a 3D NoC simulator. GEM5 have to be augmented with McPAT and 3D Noxim with ORION to calculate the power consumption. Maximizing performance under power constraint is an important target in digital system design in these days. The peak power dissipation during the entire execution must be less than the maximum power budget.

From the simulations, it can be proved that the heterogenous cache hierarchy 3D CMPs consume very less power and higher efficiency with less power loss compared to CMPs with homogeneous cache hierarchy due implementation due to 3D NoC layers and stacked heterogeneous cache layers.

4. CONCLUSIONS

In this review, an efficient power model that formulates the power consumption of 3D CMPs with stacked cache layers is proposed as best among other technologies existing nowadays. The proposed model that considers power impact

of un-core beside the cores for the first time is appropriate for heterogeneous and non-heterogeneous CMPs under multi-threaded and multi-program workloads. The saving of power through reducing latencies helps in integrating more transistors in the CMPs at the defined constraints. This new technology solves the problem of dark silicon age into a great extent.

In the future, we can use this model in the optimization problems to minimize power consumption or maximize performance of CMPs under latency and temperature constraints. Moreover, this power model can be used in the prediction functions of machine learning-based power/thermal management techniques for future power-aware CMPs.

REFERENCES

[1] Kao J, Narendra S, Chandrakasan A- "Subthreshold leakage modeling and reduction techniques", IEEE/ACM Int. Conf. Comput.-aided design (ICCAD) in 2002.

[2] Wang W, & Mishra P. "System-wide leakage-aware energy minimization using dynamic voltage scaling and cache reconfiguration in multitasking systems". IEEE Trans. Very Large-Scale Integration in 2012.

[3] Esmaeilzadeh, Blem,E, St. Amant, R Sankaralingam, K Burger. "Dark silicon and the end of multicore scaling". In Proc. Int. Symp. Comput. Archit ,2011.

[4] Woo DH, Seong NH, Lewis DL, Lee H-HS "An optimized 3D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth". In Int. Symp. High Perf. Comput. Arch. (HPCA), 2010.

[5] Loh GH. "3D-Stacked Memory Architectures for Multi-core Processors." In Int. Symp. .Comput. Arch. (ISCA), 2008.

[6] Kultursay E, Kandemir M, Sivasubramaniam A, Mutlu O . "Evaluating STT-RAM as an energy-efficient main memory alternative". In Int. Symp. Performance Analysis of Systems and Software (ISPASS), 2013.

[7] Turakhia Y, Raghunathan B, Garg S, Marculescu, "HaDeS: Architectural synthesis for heterogeneous dark silicon chip multi-processors". In Design Autom. Conf. (DAC), 2013.

[8] Raghunathan B, Turakhia Y, Garg S, Marculescu, "Cherrypicking: exploiting process variations in dark-silicon homogeneous chip multiprocessors" In Design, Autom. Test in Europe Conf. & Exhibition, 2013.

[9] Li S, Ahn JH, Strong RD, Brockman JB, Tullsen DM, Jouppi NP, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures". In Proc. Int. Sym. Microarchitecture, 2009.

[10] Binkert N, "The gem5 simulator" ACM SIGARCH Comput. Archit.News, 39(2), 2011.