Microprocessors and Microsystems 42 (2016) 100–112

Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

Exploration of temperature-aware refresh schemes for 3D stacked eDRAM caches

Young-Ho Gong[a], Jae Min Kim[b], Sung Kyu Lim[c], Sung Woo Chung[a,*]

[a] *Department of Computer Science, Korea University, Seoul 136-713, Korea*
[b] *Samsung Electronics DS, San#24 Nongseo-Dong, Giheung, Gyeonggi-do 446-711, Korea*
[c] *School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA*

ARTICLE INFO

Article history:
Received 26 August 2014
Revised 20 May 2015
Accepted 2 January 2016
Available online 8 February 2016

Keywords:
Cache
eDRAM
Refresh interval
3D Microprocessors
Temperature

ABSTRACT

Recent studies have shown that embedded DRAM (eDRAM) is a promising approach for 3D stacked last-level caches (LLCs) rather than SRAM due to its advantages over SRAM; (i) eDRAM occupies less area than SRAM due to its smaller bit cell size; and (ii) eDRAM has much less leakage power and access energy than SRAM, since it has much smaller number of transistors than SRAM. However, different from SRAM cells, eDRAM cells should be refreshed periodically in order to retain the data. Since refresh operations consume noticeable amount of energy, it is important to adopt appropriate refresh interval, which is highly dependent on the temperature. However, the conventional refresh method assumes the worst-case temperature for all eDRAM stacked cache banks, resulting in unnecessarily frequent refresh operations. In this paper, we propose a novel temperature-aware refresh scheme for 3D stacked eDRAM caches. Our proposed scheme dynamically changes refresh interval depending on the temperature of eDRAM stacked last-level cache (LLC). Compared to the conventional refresh method, our proposed scheme reduces the number of refresh operations of the eDRAM stacked LLC by 28.5% (on 32 MB eDRAM LLC), on average, with small area overhead. Consequently, our proposed scheme reduces the overall eDRAM LLC energy consumption by 12.5% (on 32 MB eDRAM LLC), on average.

1. Introduction

As the process technology scales down, microprocessor performance improves dramatically, especially due to the reduced gate delay. However, the wire delay is not much reduced compared to the gate delay, since it is more affected by wire length. Among components in a microprocessor, caches occupy the largest area since microprocessors have multi-megabyte last-level cache (LLC). The large cache area leads to long wire delay that accounts for a substantial portion of cache access time [6]. To alleviate the long wire delay, many studies have proposed 3D microprocessor design which uses vertical interconnection by through-silicon-vias (TSVs) between dies [23–25].

There are two representative 3D microprocessor design alternatives [23]: (i) cache-on-core and (ii) core-on-core. In the case of cache-on-core 3D microprocessor design, LLC dies are vertically stacked on top of the CPU core die. Since TSV-based 3D interconnection provides shorter wire length than 2D wire interconnection, it reduces the cache access time of cache-on-core 3D microprocessors significantly, compared to conventional 2D microprocessors. In the case of core-on-core 3D microprocessor design, each die consists of functional blocks of a CPU core as well as a fragment of LLC. In each die, the functional blocks and the fragment of LLC are connected by 2D wire interconnection. In order to improve performance, functional blocks (such as arithmetic units, register file, and etc.) are split into each die. Furthermore, the functional blocks far from each other in the 2D microprocessor design are connected vertically to improve performance [17]. However, the performance improvement of core-on-core 3D microprocessors is insignificant compared to that of cache-on-core 3D microprocessors, since the wire delay reduction between functional blocks is not so large compared to that in the case of cache access [16]. In addition, core-on-core 3D microprocessor design is more likely to face thermal problem than cache-on-core 3D microprocessor design, since vertically stacked functional blocks are much hotter than vertically stacked caches [18,20]. Note that the increased power density of 3D microprocessors causes higher on-chip temperature [28]. Thus, many recent studies have focused on cache-on-core 3D microprocessor design [1,5,24,25].

In conventional 2D microprocessors, most caches consist of SRAM cells to provide high performance. However, the high

[*] Corresponding Author. Tel.: +82 2 3290 3571.
E-mail addresses: kyh555@korea.ac.kr (Y.-H. Gong), joist@gmail.com (J.M. Kim), limsk@ece.gatech.edu (S.K. Lim), swchung@korea.ac.kr (S.W. Chung).

http://dx.doi.org/10.1016/j.micpro.2016.01.010
0141-9331/© 2016 Elsevier B.V. All rights reserved.

performance comes at the expense of large area and high leakage power. Especially, among components in the conventional 2D microprocessor, SRAM based LLC has the largest area and the highest leakage power due to its large capacity. To alleviate large area and high leakage power of SRAM based LLC, many studies have proposed embedded DRAM (eDRAM) based LLC [5,13,34]. Compared to SRAM cells, eDRAM cells have smaller area and less leakage power, since they consist of less number of transistors than SRAM cells. Hence, recent studies have shown that using eDRAM stacked LLC is a promising approach for 3D microprocessors rather than using SRAM stacked LLC [5,34].

Different from the SRAM cells, eDRAM cells need to be refreshed periodically to preserve their data. The eDRAM cell *retention time* (the maximum time the stored data can be retained without any refresh operation) varies depending on temperature [32]. As eDRAM cell temperature gets higher, its leakage current also increases. Eventually, the eDRAM cell needs more frequent refresh operations to preserve its data.

With the conventional refresh method, all eDRAM stacked cache banks are refreshed once every predefined constant *refresh interval* (the period between the beginnings of two successive refresh operations), which ensures the data integrity even at the highest operating temperature. In fact, the eDRAM stacked cache banks under lower temperature can retain their data with longer refresh interval than the predefined constant refresh interval. In other words, the conventional refresh method causes a lot of unnecessary refresh operations for the eDRAM stacked cache banks under low temperature. Such unnecessary refresh operations eventually result in energy/performance loss.

In practice, recent low-power DRAM chips (which are used for main memory systems) adopt a temperature-aware refresh method, which is called Temperature Compensate Self Refresh (TCSR) [38]. The TCSR changes the refresh interval of the DRAM chip depending on temperature of the whole DRAM chip. However, the TCSR cannot reduce refresh overhead significantly, since the TCSR does not have fine-grained temperature-aware refresh intervals; it has only a few (two or four) refresh intervals depending on temperature of whole DRAM chip. Compared to DRAM chips (for main memory systems), eDRAM stacked caches must have higher peak temperature in runtime due to the impact of 3D stacking on heat radiation ability. According to the temperature analysis of a recent computer system [36], the peak temperature of the DRAM chip is below 60 °C. However, the peak temperature of the CPU core is nearby 80 °C. Assuming that the eDRAM stacked caches are applied to the system, they have much higher peak temperature than DRAM chips. In addition, eDRAM stacked caches have a considerable spatial temperature variation across eDRAM cache banks, since each eDRAM cache bank has different heat radiation ability depending on distance from the heat spreader. In order to effectively reduce refresh overhead of eDRAM stacked caches, we should adopt more fine-grained temperature-aware refresh intervals. Moreover, to further reduce refresh overhead, we should vary refresh intervals of eDRAM cache banks depending on temperature of each cache bank. However, since the TCSR does not use fine-grained temperature-aware refresh intervals and does not apply different refresh intervals to different banks, it is not appropriate for reducing refresh overhead of eDRAM stacked caches.

In this paper, we analyze the retention time of eDRAM stacked cache banks depending on temperature. Since temperature of each eDRAM stacked cache bank varies, the retention time can be different for different cache banks. Note that an eDRAM cell under lower temperature retains its data much longer without any refresh operation. Considering the different retention time of eDRAM stacked cache banks due to temperature changes, we propose a novel temperature-aware refresh scheme for 3D stacked eDRAM caches. We use thermal sensors for detecting temperature of each cache bank. Depending on temperature from thermal sensors, the scheme proposed in this paper applies temperature-aware refresh interval to eDRAM LLC. In addition, our scheme dynamically changes the refresh interval, depending on run-time temperature. As a result, our scheme significantly reduces the number of refresh operations.

The rest of this paper is organized as follows. In Section 2, we present motivational study why the temperature should be considered in the 3D stacked eDRAM caches. In Section 3, we present a review of related works. In Section 4, we propose a novel temperature-aware refresh scheme. In Section 5, we provide our evaluation methodology and evaluation results, in the perspective of temperature, refresh interval, number refresh operations, and energy consumption. Lastly, we conclude our paper and discuss future work in Section 6.

2. Motivation

The thermal problem is more serious in 3D microprocessors than in 2D microprocessors due to the following reasons:

(1) Temperature of each die is likely to increase along with the distance from heat spreader, since the die far from the heat spreader has low heat radiation ability. For example, in Fig. 1, the die in layer 4 has lower heat radiation ability than that in layer 1.
(2) Temperature of each die is affected by temperature of adjacent dies. For example, in Fig. 1, when the die in layer 2 becomes hot, the adjacent dies (layer 1 and layer 3) are also heated up.

Fig. 2(a)–(c) show thermal maps of 8 MB (1-die), 16 MB (2-die stacked), and 32 MB (4-die stacked) eDRAM stacked LLC, respectively; note that the L2 cache is used for the LLC in our paper. Each cache die is 8 MB eDRAM cache composed of eight 1 MB eDRAM cache banks. As shown in Fig. 2(c), the cache die on top of the stack is hottest due to lower heat radiation ability.

The cache banks shown in Fig. 2(a)–(c) have different retention times due to the temperature difference. To analyze how temperature affects eDRAM retention time, we investigate the relation between temperature and retention time. According to [11], retention time of an eDRAM cell is written as follows:

$$T_{RETENTION} \propto \frac{Cs\Delta V_{SN}}{I_{LEAK}} \quad (1)$$

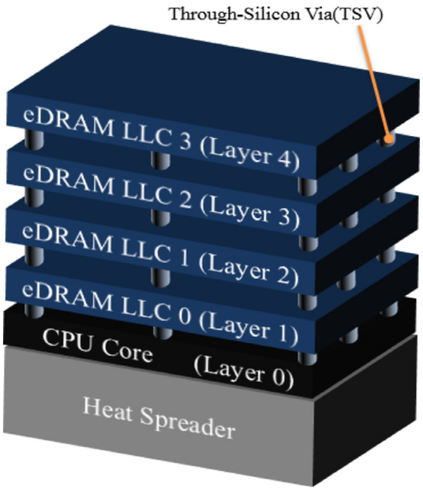

Fig. 1. 3D microprocessor design with eDRAM stacked caches.(Note that this represents a cache-on-core 3D microprocessor design).