Design Considerations of Die-Stacked DRAM Caches

Rou-Li Melody Wang, Yun-Chao Yu, and Jin-Fu Li
Department of Electrical Engineering
National Central University
Taoyuan, Taiwan 320

***Abstract*—Multiple-channel die-stacked DRAMs have been used for maximizing the performance and minimizing the power of memory access in 2.5D/3D system chips. Stacked DRAM dies can be used as a cache for the processor die in 2.5D/3D system chips. Typically, modern processor system-on-chips (SOCs) have three-level caches, L1, L2, and L3. Could the DRAM cache be used to replace which level of caches? In this paper, we derive an inequality which can aid the designer to check if the designed DRAM cache can provide better performance than the L3 cache. Also, design considerations of DRAM caches for meet the inequality are discussed. We find that a dilemma of the DRAM cache access time and associativity exists for providing better performance than the L3 cache. Organizing multiple channels into a DRAM cache is proposed to cope with the dilemma.**

I. Introduction

Three-dimensional integration technology using through-silicon via (TSV) enables multiple layers of dynamic random access memory (DRAM) to integrated with processors. Stacked DRAMs can be used for a cache [1]. DRAM cache can be divided into SRAM-tag and Tags-in-DRAM designs [2]. SRAM-tag DRAM cache stores tags in a separated SRAM structure, which needs large area cost for the SRAM. Tags-in-DRAM DRAM cache places the tags in DRAM to avoid the high area overhead of SRAM.

Fig. 1 shows a conceptual diagram of a processor system-on-chip (SOC) and a Tags-in-DRAM DRAM cache integrated with 2.5D/3D technology. In the DRAM cache, a cache line consists of Tag bits and Data bits. The processor SOC has a DRAM controller integrated with a Cache controller. When a cache line is read, the Tag bits are compared with the read address to check if a hit occurs. The access time, associativity, and hierarchy in the memory system of the DRAM cache have heavy impact on the system performance. Furthermore, the DRAM cache can be used to replace which level of caches. In this paper, we analyze those issues and discuss design considerations of DRAM caches.

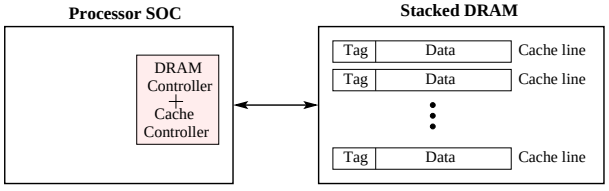

Fig. 1. Conceptual diagram of a processor SOC and a DRAM cache integrated with 2.5D/3D technology.

II. Design Considerations of DRAM Caches

Modern multi-core processor SOCs usually have three-level caches [3]. Subsequently, we derive an equation which can be used to determine how characteristics of a DRAM cache should be met if the DRAM cache is used to replace level-3 cache. Fig. 2(a) shows a conceptual diagram of a processor SOC with three-level caches, L1, L2, and L3. Assume that the access times of L1, L2, and L3 are t_1, t_2, and t_3, respectively. Also, the hit rate of L1, L2, and L3 are h_1, h_2, and h_3, respectively. Therefore, the effective access time of the memory hierarchy can be expressed as

$$\begin{aligned} T_{eff} &= h_1t_1 + (1-h_1)h_2t_2 + (1-h_1)(1-h_2)h_3t_3 \\ &\quad +(1-h_1)(1-h_2)(1-h_3)t_m, \end{aligned} \tag{1}$$

where t_m denotes the access time of main memory. Consider that the DRAM cache is used to replace the L3. As Fig. 2(b) shows, the processor SOC with L1 and L2 caches and DRAM cache serving L3 is integrated with the processor SOC using 2.5D/3D integration technology. The effective access time of the memory hierarchy of 2.5D/3D system chip can be expressed as

$$\begin{aligned} T'_{eff} &= h_1t_1 + (1-h_1)h_2t_2 + (1-h_1)(1-h_2)h'_3t_{dc} \\ &\quad +(1-h_1)(1-h_2)(1-h'_3)t_m, \end{aligned} \tag{2}$$

where h'_3 and t_{dc} denote the hit rate and access time of the DRAM cache, respectively.

According to Eqs. 1 and 2, we can obtain that

$$T'_{eff} - T_{eff} = k[(h'_3t_{dc} + (1-h'_3)t_m) - (h_3t_3 + (1-h_3)t_m)],$$

where $k = (1-h_1)(1-h_2)$. If $T'_{eff} - T_{eff} \leq 0$, then the DRAM cache provides better performance than the L3. Since $k > 0$, we can obtain the following inequality

$$\begin{aligned} & h'_3t_{dc} + (1-h'_3)t_m \leq h_3t_3 + (1-h_3)t_m \\ \Rightarrow \quad & h'_3t_{dc} - h'_3t_m \leq h_3t_3 - h_3t_m \\ \Rightarrow \quad & h'_3(t_{dc} - t_m) \leq h_3(t_3 - t_m) \\ \Rightarrow \quad & \frac{h'_3}{h_3} \geq \frac{t_m - t_3}{t_m - t_{dc}} \end{aligned} \tag{3}$$

According to Eq. 3, we can have the following observations. Typically, the access time of DRAM cache is smaller than that of L3. That is, h'_3/h_3 is larger than 1. Furthermore, if t_m is much larger than t_{dc}, h'_3/h_3 is approximated to 1. This

implies that the DRAM cache can more easily provide better performance since $h_3' \approx h_3$.

For example, the Sparc T5 16-core processor SOC with three levels caches, L1, L2 and L3, reported in [3]. L3 is a 16-way set associative cache with 64-byte cache lines. The operation frequency of the SOC is 3.6GHz and the hit latency of L3 is 51 cycles. Thus, the access time of L3 is $51 \times 0.27 = 13.77ns$, i.e., $t_3 = 13.77ns$ Assume that the access time difference between the DRAM cache and L3 is $\triangle = t_{dc} - t_3$. Fig. 3 shows the ratio h_3'/h_3 with respect to different values of t_m for different value of $\triangle$. We see that if the value of $\triangle \rightarrow 0$, the curve approaches the point $h_3'/h_3 = 1$. On the other hand, if the t_m is much larger than t_{dc}, h_3'/h_3 approaches to 1. Those imply that the DRAM cache can provide better performance benefits more easily. Consequently, there are two approaches for maximizing the performance benefits in designing a DRAM cache to replace the L3 cache in SOC chips. The first approach is to minimize the t_{dc} and the second one is to increase the hit rate.

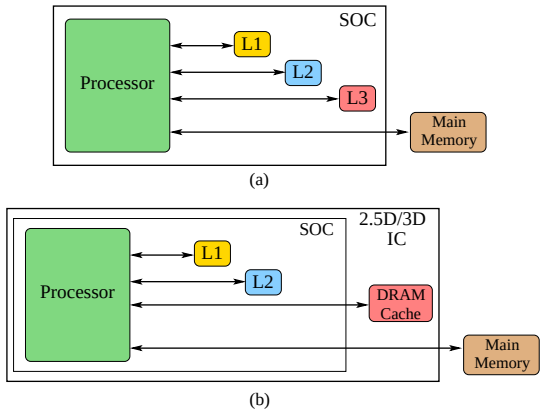

Fig. 2. (a) A conceptual diagram of a processor SOC with three-level caches. (b) A conceptual diagram of a processor SOC with two-level caches integrated with a DRAM cache using 2.5D/3D technology.

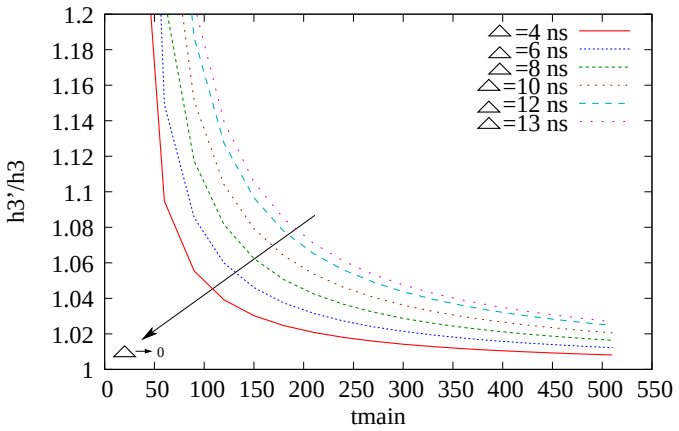

Fig. 3. Ratio of $\frac{h_3'}{h_3}$ with respect to different values of t_m for different differences of t_{dc} and t_3.

The hit rate of DRAM cache is related to the associativity. Consider a 1GHz double-data-rate DRAM cache which has multiple channels and 72-bit words. Each channel 4 bank groups and 8 banks per bank group, and channel size is 576MB. Assume that a cache line has 8 words, i.e., 576 bits. If the associativity is direct mapping and burst length=8, the access time of a read hit operation can be expressed as $t_{command}+t_{RAL}+t_{AC}+4ns$, where $t_{command}$, t_{RAL}, and t_{AC} denote the command issuing time, the random access latency, and the output access time, respectively. Since a cache line has 8 words and the DRAM cache is double data rate operated at 1GHz, 4ns is needed to read 8 words. On the other hand, if the associativity of DRAM cache is M-way set associative, the access time of a read hit operation can be expressed as $t_{command} + t_{RAL} + t_{AC} + 4Mns$. Since each time M cache lines should be read, $4Mns$ is needed to read $8M$ words. Clearly, the access time of DRAM cache is increased with the associativity. If the set associative is implemented, the hit rate h_3' is increased, but the access time t_{dc} is increased as well. On the contrary, if the direct mapping is implemented, t_{dc} is decreased, but h_3' is decreased. To cope with this dilemma, one possibility is to take advantage of the feature of multi-channel DRAMs. We can organize multiple channels into a DRAM cache to minimize the access time. For example, if we organize l channels into a DRAM cache with M-way set associativity, then the access time of a read hit operation can be expressed as $t_{command} + t_{RAL} + t_{AC} + (4M/l)ns$ since the access preparation time $t_{command} + t_{RAL} + t_{AC}$ of each channel can be overlapped. Thus, we can increase the associativity to increase hit rate and reduce the access time t_{dc} of DRAM cache.

III. Conclusion

In this paper, we have derived an inequality which can be used to aid the designer to check if the designed DRAM cache can provide better performance than the L3 cache. Also, design considerations of DRAM caches for meet the inequality have been discussed. We see that a dilemma of the DRAM cache access time and associativity exists for providing better performance than L3 cache. Organizing multiple channels into a DRAM cache has been proposed to cope with the dilemma.

Acknowledgments

This work was supported in part by the Ministry of Science and Technology (MOST), Taiwan, R.O.C., under Contract NSC 102-2221-E-008-108-MY3 and MOST 104-2220-E-008-009.

References

[1] G. H. Loh, "3D-stacked memory architectures for multi-core processors," in *35th International Symposium on Computer Architecture (ISCA)*, 2008, pp. 453–464.

[2] M. K. Qureshi and G. H. Loh, "Fundamental latency trade-off in architecting DRAM caches: Outperforming impractical SRAM-tags with a simple and practical design," in *45th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2012, pp. 235–246.

[3] J. Feehrer, S. Jairath, P. Loewenstein, R. Sivaramakrishnan, D. Smentek, S. Turullols, and A. Vahidsafa, "The Oracle Sparc T5 16-core processor scales to eight sockets," *IEEE Micro*, vol. 33, no. 2, pp. 48–57, Apr. 2013.