

On System Level Performance of DRAM/Logic Merged Technology

Yong-Bin Kim* Tom Chen**

* Microelectronics Division
 Samsung Electronics Co. San Jose, CA, USA

**Department of Electrical Engineering
 Colorado State Univ. Fort Collins, CO 80523, USA
 chen@longs.lance.colostate.edu

Abstract

This paper assesses the performance of DRAM/Logic merged technology at the system level. The evaluation has been done for the six state-of-the-art microprocessors without changing memory hierarchies, and two graphics chip applications. The analysis shows that average gain of 9.29% can be achieved had these processors been implemented using DRAM/logic merged technology, and DRAM/logic merged technology provides higher performance than the conventional graphics chip application approach by taking advantage of at least 2X wider bandwidth for Massively Parallel Processor(MPP) case by factor of ten.

1 Introduction

State-of-the-art core CPU is a fast superscalar, superpipelined, RISC running at higher than 100MHz clock rate. Keeping such a CPU fed at full speed takes something on order of 1-2GB/sec of memory bandwidth. For the densest DRAM technologies the DRAM interconnects provide typically something considerably below 50MB/sec of bandwidth per part [1]. With the demand for better graphics, the bandwidth requirement on DRAM is increasing significantly, from several tens of MB/sec in 1990 to over 500MB/sec [2]. There are less than two dozen off-chip interconnects available on a DRAM part, and for the densest DRAM technologies these interconnects provide typically around 50MB/s of bandwidth per part even though the actual bandwidth is much higher. This is in contrast to the 1 to 2 GB/sec bandwidth needed to support the core of most modern microprocessor. Therefore, how to get high bandwidth from one or two chips with limited I/O becomes an issue. These issues make new technology like DRAM/Logic merged technology more attractive to integrate both DRAM and logic on the same chip. This concept is being more emphasized in graphics controller chip design, because a total integration with a graphics controller and macro based DRAM frame buffers is the promising approach with respect to cost, performance, and power. DRAM macros provide bandwidth by wide data bus and on-chip speed. Just as there are reasons the embedded DRAM will run fast, there are reasons the logic next to it will run slowly due to limitations of DRAM process such as higher transistor threshold voltage. In this paper, considering the logic performance overhead on DRAM process and memory capacity, system level performance is analyzed for six state-of-the-art CPU chips and graphics chips.

2 CPU Performance Analysis

In recent years, increases in memory subsystem speed have not kept pace with the increase in core CPU speed, causing the core CPU execution rates to become increasingly limited by the cache performance. The cache performance barrier has been overcome by employing hierarchical structures (two or more level) and associativity in memory structure. Other than memory hierarchy and associativity, system performance study considering the cache performance involve other important factors such as cache miss rate, miss penalty, memory bandwidth. Therefore, the important factors to be considered for system performance analysis are:

1. Cache size
2. Cache miss rate
3. Cache miss penalty
4. Cache degree of associativity
5. Memory hierarchy
6. Cache cycle time
7. Memory bandwidth

Some of these factors are not mutually exclusive. There are some interaction between these factors. For instance, cache miss rate is a function of cache size and its degree of associativity. Cache miss penalty has something to do with the memory bandwidth.

In this research, system level performance will be evaluated on DRAM/logic merged technology using the same architectural parameters and the same memory hierarchies of the CPUs. The variable factors are processor cycle time considering the logic performance overhead on DRAM process and memory related factors as described above. Memory miss rate is a function of memory size and set-associativity. The first level(L1) cache miss rate data for cache size and set-associativity has been calculated by Gee et al. for Spec92 benchmarks [3].

$$MR_1 = .0087 + .0969e^{-x/1.40} + .0631e^{-x/7.18} + .0196e^{-x/47.93} \quad (1)$$

$$MR_2 = .0087 + .0789e^{-x/1.17} + .0593e^{-x/8.09} + .0136e^{-x/98.86} \quad (2)$$

$$MR_4 = .0082 + .0849e^{-x/1.20} + .0303e^{-x/7.51} + .0235e^{-x/23.72} \quad (3)$$

$$MR_5 = .0036 + .1718e^{-x/9.63} + .0396e^{-x/7.25} + .0130e^{-x/74.53} \quad (4)$$

Fig.1 L1 cache miss rate fitting Equations

x , MR_1 , MR_2 , MR_3 , and MR_4 represents cache size, cache miss rate for one-way, two-way, four-way, and eight-way set associativity, respectively. For a given cache size and degree of associativity of the first level cache, the miss rate(MR) is predicted by extrapolating the curve fitted using the data. The fitting equations for different cache size and set-associativity are shown in Fig.1. Cache miss rate decreases as the degree of associativity increases. However, greater associativity can come at the cost of increased memory hit time. Since the speed of CPU is tied directly to the speed of a cache hit, CPU cycle time should be increased considering the multiplexer delay for the set associative cache. The mux delay is assumed to be 10% of the clock cycle time for two-way set associativity configuration. Hill et al. [4] found 2% difference in hit time for custom CMOS caches as the degree of set associativity increases from two-way to four-way. Extrapolating this trend, the clock cycle time for N-way, CKT_{N-way} can be found as follows.

$$CKT_{N-way} = (1.08 + 0.02\log_2(N))CKT_{1-way} \quad (5)$$

where N is the degree of associativity and equal to or greater than 2. CKT_{1-way} of the merged system has to take the logic performance overhead on DRAM process into account for the cycle time of the merged system cache. Routing area penalty also has to be paid for the merged system since there are less number of metal layers available in DRAM process. The penalty has been also analyzed by Y.B. Kim et al. [5] along with logic performance penalty on DRAM process. The wiring capacitance per net is increased by 20% on DRAM process comparing to logic process. The absolute capacitance increase per net is 7.6ff which is equivalent to 0.2 fan-out since the average capacitance per fanout is about 40ff. Considering this extra penalty, The clock cycle time for one way set-associativity on the merged technology becomes

$$CKT_{D1} = \frac{(0.18571 + 0.11179 \times (X + 0.2))}{(0.22286 + 0.07786 \times X)} \times CKT_1 \quad (6)$$

where X is the number of fanout. Wider bandwidth takes better advantage of spatial locality. Therefore, wider bandwidth reduces miss penalty effectively. miss penalty is almost inversely proportional to bandwidth for most of the cases. In this experiment 8 times bandwidth increase is assumed because there is a good chance to increase 8X bandwidth if the memory capacity is increased by eight times at the same die area exploiting DRAM device density. The average local hit rate of L2 cache versus degree of associativity for a state-of-art micro processor is shown in Fig.2. The average hit rate can be found by taking the complementary rate(1 minus average hit rate). Since the second level cache size is larger than the first level, the second level cache hit time is longer than the first level cache by at least factor of three or eight including the set associativity overhead(mux delay). In this research, the second level cache hit time is assumed to be eight clock cycles. The performance of two different systems can be compared in terms of execution time for a given task, and CPU execution time for a program is simply

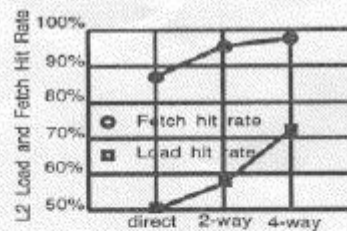


Fig.2 L2 cache hit rate.

expressed in the following way.

$$CPU \text{ Execution Time} = IC \times \left(CPI + \frac{\text{Number of Memory Access}}{\text{Instruction}} \times MR \times MP \right) \times CKT$$

where CPI is cycle per instruction and CKT is clock cycle time. For a given test micro-code, instruction count(IC), memory references, Miss Penalty(MP) and CPI are fixed. The variable hardware parameters for system performance analysis on DRAM/logic merged technology are Miss Rate(MR) and Clock Cycle Time(CKT) taking the memory capacity increase and logic performance overhead on DRAM process into account. Considering the cache miss rate and bandwidth enhancement and logic overhead on DRAM process, the CPU performance gain has been assessed for six different state-of-the-art processors and the results are shown in Table I. The upper half of the table represents the current implementations and the lower half represents the equivalent implementations using merged technology. The two implementations are compared to obtain the performance gain of the merged technology in the table.

3 Graphics Chip Performance

As DRAM/logic merged technology permits very significant amount of logic to be placed on conventional DRAM chips, meaning that the bandwidth available from internal memory arrays can be utilized by one or more CPUs placed directly on the chip called "Processor-In-Memory(PIM)" architecture. While this is valuable for conventional computer, it becomes even more so for Massively Parallel Processors(MPPs), particularly for embedded applications where the cost of replicating all the glue and bandwidth recovery circuitry can become an excessive part of the overall system cost. Peter et. al. [1] integrated 8 complete processing elements, each with its own 64KB DRAM, 16 bit CPU, and peripheral circuits including controlling, addressing, refreshing, parity, and redundancy elements. Peter et. al. [1] said, for future generations of PIM chips they will be able to adjust relatively smoothly the usage of surface area of a chip die that contains both logic and DRAM from 100% logic and 0% DRAM to 0% logic and 100% DRAM. Fig.3 shows the CMOS chip technology prediction reported by Semiconductor Industries Association(SIA). The number of CPUs can

Table 1. Performance analysis result

	Alpha 21164	HP 9000	Intel P6	MIPS R10000	Power PC 820	Ultra Sport
L1 cache	-	-	-	-	-	-
Freq.	100MHz	180MHz	133MHz	200MHz	133MHz	167MHz
Size	16K	-	16K	64K	32K	32K
Ways	direct	-	4 way	2 way	2 way	2 way
L2 cache	-	-	-	-	-	-
Size	96K	256K	256K	-	-	-
Ways	3 way	direct	4 way	-	-	-
Merged cache	-	-	-	-	-	-
L1 cache	-	-	-	-	-	-
Freq.	82MHz	148MHz	109MHz	165MHz	109MHz	138MHz
Size	128K	-	128K	512K	256KK	256K
Ways	direct	-	4 way	2 way	2 way	2 way
L2 cache	-	-	-	-	-	-
Size	1M	2M	2M	-	-	-
Ways	3 way	direct	4 way	-	-	-
Per. Gain	6.05%	4.10%	0.92%	17.75%	9.19%	17.76%
σ Average	-	-	-	-	-	9.29%

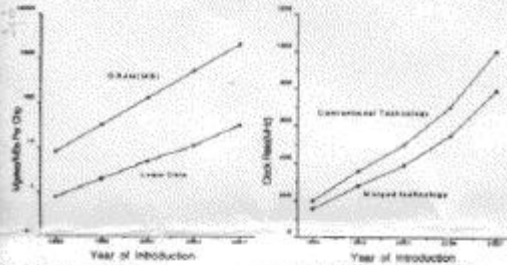


Fig.3 CMOS technology projections

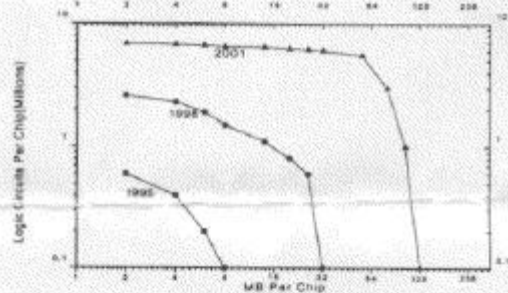


Fig.4 DRAM/logic merged chip capacity estimate

be estimated assuming there are typically 12K gates per CPU for typical fixed point engine. Fig.4 shows the potential merged chip configuration. The clock rate for merged technology in Fig.3 reflected logic performance overhead (21.9%). The maximum clock frequency for 1995 is 156.3MHz (6.4ns cycle time). Fig.4 shows the hardware availability on the merged chip for a given die size. For instance, if there are 8MB then there should be one CPU on the chip assuming 1 CPU has 12K logic circuits. On the other hand, if there is 2MB then 80 CPUs can be implemented on the same die area. For computer intensive tasks, the number of CPUs that 1MB memory supports is between 3 and 5. For low end tasks, the number of CPUs that 1MB memory supports is above twenty (20) [6]. Fig.5 shows different performance curves for different ratios of memory to number of CPU. One instruction was assumed to be executed at average 2.5 cycles. Fig.5 shows the merged chip performance as a function of memory size. For the memory to cpu ratio of 2M to 1, for example, the performance increases as memory size increases initially since memory size is a performance limiting factor up to memory size of 6MB. This means there are lot of CPUs that are not effectively

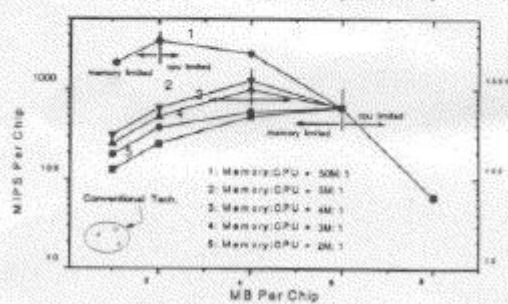


Fig.5 Merged chip performance vs. memory size

used because the supporting memory capacity is not enough. However, the performance decreases beyond 6MB point since the number of CPU available on the chip becomes limiting factor. Even though all of the available CPU are used, the absolute number of CPUs on the chip is not enough comparing the large memory size available.

As shown in Fig.5 the optimum performance is obtained at 4 to 6 MB memory range for high end applications. The system performance of conventional massively parallel systems is also plotted in Fig.5. The system performance using merged technology is better than the conventional approach by at least factor of 10 mainly due to bandwidth increase and on chip memory capacity increase.

4 Conclusion

This paper presents the performance analysis of Logic/Dram merged technology on system level considering the logic gate performance overhead and all the important cache factors which affects on the system level performance.

Without changing CPU architecture and with keeping the cache memory hierarchies the average performance gain for six state-of-the-art microprocessors is 9.95%. If the merged process technology is fully developed so that the logic performance overhead and routing penalty in the logic part are minimized, the performance gain will be even higher than 9.95%.

The performance of the graphics chip integrated using the merged technology has been evaluated. DRAM/logic merged technology provides higher performance than the conventional graphics chip application approach by taking advantage of at least 2X wider bandwidth for Massively Parallel Processor(MPP) case by factor of ten.

References

- [1] Peter M. et.al, "Combined DRAM and Logic Chip for Massively Parallel Systems", *Sixteenth Conf. on Adv. Rec. in VLSI*
- [2] A. Kanuma, "The Best DRAM Approach for Graphics Application", *ISSCC 95* pp.96-97,1995.
- [3] Gee, J.D., et.al, "Cache Performance of the Spec 92 benchmark suite", *IEEE Micro 13:4* pp.17-27, 1993.
- [4] Hill, M. D., "A Cache for Direct Mapped Caches", *Computer 81:12(December)* pp.25-40, 1988.
- [5] Y.Kim, T.Chen, "Assessing Merged DRAM/Logic Technology", *ISCAS 96*, 1996
- [6] Peter M. Kogge, "EXECUBE-A New Architecture for Scalable MPPs", *Proceedings of International Conference on Parallel Processing* pp.77-84, 1994.