Memory Latency Notes

D. Morano
Northeastern University
dmorano@ece.neu.edu

16th May 2002

1 Introduction

These notes include some information on Dynamic Random Access Memory (DRAM) latencies in current, and possibly future, machines. Some information and background on the increasing problem that DRAM latency performance presents to the whole computer design is also included.

2 Current DRAM Latencies

Cuppu et al [3] presents a good analysis of various DRAM design architectures using the various different DRAM chip technologies that are currently available or are soon to be available. Their analysis largely focuses on the higher-end workstation design point. This work was an extension of prior work by Cuppu et [4]. They show how the chip DRAM timing (precharge, row, column, and data transfer) for Enhanced Synchronous DRAMs (ESDRAM) (among other DRAM chip variants) translates into average access latencies for a realistic memory system design. They show that 100 MHz ESDRAM (a very conservative design choice at the present time) translates into average main memory latencies of between 120 and 170 nsecs for the benchmarks that they investigated (SpecInt-95). For a 1 GHz processor clocks, this results in approximately 120 to 170 clocks of delay before the data from an average cold access (new DRAM page) can be returned to the processor. This result was largely insensitive to the choice of L1 and L2 caches latencies.

Wulf et al [8] (and Hennessy et al [7]) report only an average DRAM latency speed increase of about 7 %. Burger et al [1] give an average speed increase of between 5 % and 10 %. The slow increase in average DRAM latency is largely due to the technology used for the core of DRAMs. The core DRAM technology is largely the same for all DRAM interface variants. There was some improvement with the transition to the newer enhanced SDRAMs that appeared started at the 133 MHz clock frequencies (for example DDR133). Processor clocks increase at a much higher rate and approximate a growth of 40 % to 80 % a year [5]. Extrapolating about five years into the future and assuming a current process clock rate of 2.4 GHz (the latest Pentium 4 processors) and a conservative growth rate of about 40 %, processor clocks can likely be at around 10 GHz. However DRAM average access latencies will only be at about (using the 7 % figure) between 100 nsec and 120 nsec. This represents a latency in processor clocks of 1000 and 1200 respectively.

3 Background

The need to deal with the processor/memory performance gap has been reported by many including Wulf Wulf et al [8], Hennessy et al [7], Burger et al [2], and others.

Cuppu et al [4], Davis et al [6], and Davis et al again later in 2000 [5] have all explored the impact of the newer DRAM interface technologies on program execution time performance. The newer interfaces have also helped with reducing average access latencies but these are considered one-time fixes to the overall problem since the core DRAM technology follows the 7 % growth rate throughout.

References

[1] D. Burger, J. Goodman, and A. Kagi. The Declining Effectiveness of Dynamic Caching for General-Purpose Microprocessors. Technical Report UWMADISONCS CS-TR-95-1261, University of Wisconsin - Madison, Madison, WI, Jan. 1995.

[2] D. Burger, J. R. Goodman, and A. Kägi. Limited bandwidth to affect processor design. *IEEE Micro*, 17(6):55–62, Nov. 1997.

[3] V. Cuppu, B. Jacob, B. Davis, and T. Mudge. High-performance drams in workstation environments. *IEEE Transactions on Computers*, 50(11):1133–1153, Nov. 2001.

[4] V. Cuppu, B. L. Jacob, B. Davis, and T. N. Mudge. A performance comparison of contemporary DRAM architectures. In *Proceedings of the 26th Annual International Symposium on Computer Architecture*, pages 222–233. ACM, 1999.

[5] B. Davis, B. Jacob, and T. Mudge. The new dram interfaces: Sdram, rdram and variants. In *Proceedings of the 3rd Int. Symp. on High Performance Computing*, pages 26–31, Tokyo, Japan, Oct. 2000.

[6] B. Davis, T. Mudge, V. Cuppu, and B. Jacob. Ddr2 and low latency variants, May 2000.

[7] J. L. Hennessy and D. A. Patterson. *Computer Architecture: A Quantitative Approach, 2nd ed.* Morgan Kaufmann, Palo Alto, CA, 1995.

[8] S. M. W. Wulf. Hitting the memory wall: Implications of the obvious. *Computer Architecture News*, 23(1):20–24, Mar. 1995.