Bridging the Processor-Memory Gap:
Current and Future Memory Architectures

Project Report for CS/ECE 752
Advanced Computer Architecture

Michael J. Brim
James D. Speirs

December 13, 2002

Abstract:

For many decades, dynamic random access memory, or DRAM, has been the technology of choice for use as core processor memory storage. Although the functionality and general access characteristics of DRAM have not changed dramatically since its inception, the technology has evolved by continually improving in overall bit density. However, the success of DRAM has also enabled the emergence of the processor-memory latency gap. Much effort has been spent improving processor functionality and redesigning memory hierarchies to limit the effects of this growing gap in performance. In recent years, there has been a rapid onset of different designs proposed to attack the problem at the memory itself.

A summary of current DRAM and SRAM technologies focused on this problem is given, enumerating the specific design characteristics that differ among the proposals. Next, we present two technologies, intelligent RAM and magnetic RAM, which provide a fundamentally different perspective from conventional DRAM architectures for bridging the processor-memory gap. Our study is concluded by a discussion on the future of memory architectures.

1. Introduction

For many decades, dynamic random access memory, or DRAM, has been the technology of choice for use as core processor memory storage. Although the functionality and general access characteristics of DRAM have not changed dramatically since its inception, the technology has evolved by continually improving in overall bit density. Along with the density improvements, DRAM's survival has been aided by its commoditization and established manufacturing processes that reduce overhead costs to virtually nothing, making the entry of any new technology into the market excessively costly from the point of view of systems manufacturers and consumers.

However, the success of DRAM has also enabled the emergence of the processor-memory latency gap. The gap is a result of the continual improvements in processor cycle speed occuring at a much faster rate than improvements in DRAM access latency. A paper by Wulf and McKee [1] is often cited in characterizing this ever-widening gap as "hitting the memory wall," where further increases in processor speed yield little to no performance benefits due to memory access time acting as the primary bottleneck. Many improvements have been made to processors and memory architectures in the past several years in order to avoid "hitting the wall." Multilevel memory hierarchies are a direct result of trying to minimize the need to access core memory during the common execution path. Through the addition of multiple levels of high-speed static RAM (SRAM) caches with high hit ratios, the hierarchies effectively reduce the number of accesses to memory. Similarly, processor designs that implement features such as instruction prefetching, memory speculation, and lockup free caching attempt to hide the memory access latency cost. However, hiding the latency of accessing memory does not solve the other important factor in the processor-memory gap, that of the increased bandwidth required from the faster processors. In reaction to increased bandwidth demands, many recent improvements in DRAM architectures have been proposed.

In this study, we provide an overview of these proposed technologies and the specific limitations at which they are targeted. Furthermore, we present two technologies that provide a fundamentally different perspective from conventional DRAM architectures for bridging the processor-memory gap. Our study is concluded by a discussion on the future of memory architectures. Through evaluating past memory implementations and analyzing current memory architectures, we will speculate what the future may hold in store for large non-disk memory storage devices. Specifically, we seek to provide insight into whether current technologies can adapt to future processor and memory hierarchy architectures, or if these new architectures will require the development of new technologies to meet the increasing performance demands.

The rest of the paper is organized as follows. Section 2 provides an overview of related work and distinguishes our work from previous memory architecture investigations. Section 3 details the current state of SRAM and DRAM technology, providing the motivating factors and distinguishing characteristics for current designs. Section 4 continues with a discussion of memory architectures that fundamentally vary from

current ones employing SRAM and DRAM technologies. Section 5 concludes our investigation by presenting our viewpoints on the future of memory architectures.

2. Related Work

As the processor-memory gap continues to grow, many technologies have been proposed in efforts to overcome, reduce, or hide the gap. As a result, most surveys of currently available technologies are incomplete or outdated very shortly after being produced as newer ideas and technologies are explored. This aspect of continuous change in memory technology was recognized in two very similar surveys released in early 1999 [2, 3]. Both surveys provide common information on the standard operation of DRAM, as well as a reference model using typical datasheet parameters to compare and contrast a wide range of competing technologies. Our work summarizes the findings of both surveys, while also drawing from other DRAM comparisons such as [4, 5]. We do not, however, reproduce the DRAM reference model information, instead focusing on the key design aspects and motivating factors for the various technologies. Other distinguishing aspects of our work include a discussion of proposed memory architectures that fundamentally differ from previous ones and a set of observations for the predicted future of memory architectures.

3. Current RAM Technologies

3.1 Dynamic RAM

The design and operation of DRAM is widely documented, and thus is not included for brevity. Instead, our focus in this report is to identify currently available DRAM memory chips, describing the technological motivations and resulting design innovations for each chip type. We also discuss any potential limitations of the technology that are inherent in the chip's design.

3.1.1 Synchronous DRAM (SDRAM)

Synchronous DRAM, or SDRAM, is widely used in the desktop and server computer marketplace as main memory. SDRAM was developed as the first DRAM to operate synchronously with the system clock. Previously, DRAM had operated asynchronously, but the ever-increasing clock rate of processors eventually made it technically infeasible to provide adequate memory bandwidth to the processor [3]. SDRAM allows memory requests to be submitted to the DRAM on every clock cycle, and thus enables multiple outstanding requests to memory, a feature unavailable using asynchronous DRAM. SDRAM also takes advantage of spatial locality of requests by allowing burst mode operation, where a sequential number of bytes in the same memory array row can be pipelined onto the data output bus by only updating the column address on each cycle. A final characteristic of SDRAM to improve access latency is the splitting of the memory array into multiple banks, which reduces the access time to that of a single bank access while allowing improved throughput from interleaved bank requests. A design feature of SDRAM that inherently limits its performance is the need to precharge data that has

recently been accessed and refresh the charges of unaccessed data. Both of these operations require use of the sense amplifiers in the DRAM, such that new data requests can only be serviced while these operations are not taking place.

3.1.2 Enhanced SDRAM (ESDRAM)

Enhanced SDRAM [2, 3, 4] introduces a solution to the sense amplifier sharing problem inherent in ordinary SDRAM by placing a SRAM buffer after the sense amplifiers for each memory bank to hold the contents of the last memory row accessed. Further read requests to the same row or burst operations are then performed using the data cached in the SRAM. This permits the sense amplifiers to be freed up for concurrently doing data precharging, refresh operations, or subsequent memory accesses. Since write requests require use of the sense amplifiers, ESDRAM can service read requests concurrently with other reads or writes, but cannot service multiple write requests concurrently. Furthermore, ESDRAM offers little benefit when there are multiple request streams that alternate between accesses to differing rows of the same bank. In such cases, the extra latency introduced into the access time by the addition of the SRAM and extra control logic is not amortized over multiple requests that hit in the row buffer. However, ESDRAM has a slightly faster internal DRAM access latency than standard SDRAM, which helps to overcome the effects of the increased control latency.

3.1.3 Virtual Channel SDRAM (VC-SDRAM)

Virtual Channel SDRAM [2, 3] is similar to ESDRAM in that it uses SRAM caches to buffer data from recent accesses. VC-SDRAM also helps to eliminate the multiple stream problem found in ESDRAM, as it is in fact designed specifically to overcome that problem. VC-SDRAM splits each memory bank's sense amplifiers into four segments. These segments can then be cached in any of the 4N SRAM buffers known as channels (where N represents the number of memory banks). The channels in which to place segments are chosen dynamically in order to minimize conflicts from requests to the same bank. VC-SDRAM will in general perform similarly to ESDRAM in the absence of multiple request streams, but has an obvious performance advantage where multiple streams are present, such as in current multitasking operating systems.

3.1.4 Fast Cycle DRAM (FCDRAM)

Fast Cycle DRAM [4] was designed to reduce overall access time and provide for better pipelining of memory requests to individual memory banks. FCDRAM accomplishes its goals by dividing memory banks into smaller blocks, where each block has reduced access latency. FCDRAM further hides the overhead of precharging through request pipelining. The benefits of FCDRAM are costly, however, in that splitting banks into blocks greatly reduces the capacity available on the chip when compared to standard SDRAM.

3.1.5 Double Data Rate SDRAM (DDR SDRAM)

Double Data Rate (DDR) SDRAM [2, 3] is an extremely simple modification to standard SDRAM that effectively doubles the bandwidth provided by the DRAM. DDR refers to the ability of the DRAM to receive or drive data onto the data bus on both the rising and falling edges of the clock. The invention of DDR SDRAM has led to the practice of referring to standard SDRAM as Single Data Rate, or SDR SDRAM. Although DDR SDRAM provides the ability to send or receive data on each clock edge, the usefulness of the enhancement is still critically dependent upon the presence of multiple accesses to the same row, such as in burst operations, or accesses interleaved among banks. If such access sequences are not present, the limiting factor of data rate will be the internal DRAM access latency.

In an attempt to further the benefits of DDR SDRAM, the Joint Electron Device Engineering Council (JEDEC) 42.3 Future DRAM Task Group is developing a standard for DDR2 [4, 5], the next generation DDR SDRAM. The DDR2 specification is greatly different from the original JEDEC DDR specification. Unlike conventional SDRAM, DDR2 SDRAM does not support burst operations. Rather, it defines that the maximum write access and all read accesses require four data cycles (two external clock cycles). Although this change has no impact on data bus utilization, there is a direct effect on the address bus, since many more column addresses need to be transmitted for large bursts. One reason cited [5] for not supporting variable burst operations is that it removes the requirement to support transaction interruption. Another interesting part of the DDR2 specification is the posted-CAS enhancement, represented by the Additive Latency (AL) parameter. Posted-CAS allows a CAS to directly follow a RAS on the address bus, resulting in better overall address bus utilization. The AL parameter is used to set the delay for initiating the CAS relative to the start of the RAS. Finally, to maximize the effects of interleaving data accesses among banks for concurrency, the specification suggests the remapping of the processor address space by the memory controller such that temporally close accesses map to independent memory banks. The DDR2 Task Group is also considering proposals for using the techniques of ESDRAM and VC-SDRAM in conjunction with the basic DDR2 design, in the hopes to further reduce the average access latency seen by processor requests. In [5], a performance comparison simulation study shows that the ESDRAM and VC-SDRAM based improvements yield better performance than base DDR2 SDRAM.

3.1.6 Direct Rambus DRAM (DRDRAM)

Direct Rambus DRAM [6] attacks the problem of reducing the processor-memory gap by completely redesigning the interface to main memory. Mitra [3] terms the interface redesign a “revolutionary” change, since it includes a new specification for all the different parts of the memory system, including controller, buses, and DRAMs. The design goal of a DRDRAM based system is to provide the required processor bandwidth by having a significantly faster bus than SDRAM that is smaller in width. Initial proposals defined the bus rate at 400MHz and width at 16 bits, using DDR signaling to achieve an effective maximum bandwidth of 1.6 GB/s. In order to support the high frequency bus, average DRAM access time must be significantly faster than that of SDRAM. A reduced access time is achieved by having many more banks in a DRDRAM

than in a SDRAM, such that almost all overhead is hidden by the concurrency available from interleaving accesses. The large number of banks also required redesign of the organization of banks with respect to the control and data buses. In a Rambus channel, each RDRAM is attached to the common bus in serial, and two separate clocks are used, one for data movement in each channel direction. Although Rambus technology is licensed, many processor vendors have committed to producing Rambus-based systems, thus relaxing the difficulty in competing with SDRAM based systems. Initial entries into the market for Rambus-based systems have not been competitive with SDRAM systems on a price-performance basis, but recent systems employing Rambus technology have performed much better than their SDRAM counterparts.

3.2 Static RAM

Our summary of current SRAM technologies focuses on those somewhat related to DRAM technologies already discussed, typically through shared mechanisms for improving performance. As with our previous discussion for DRAM, we do not include information on general organization or operation, assuming reader familiarity with these ideas.

3.2.1 Pipelined Burst SRAM (PBSRAM)

Pipelined Burst SRAM [2] draws from improvements that were successful in DRAM. PBSRAM, currently used for L2 caches, shares many similarities with SDRAM, including synchronous operation and burst mode. Unlike traditional SRAM, PBSRAM does not split data addresses into row and column portions which are sent in consecutive cycles on the address bus, instead choosing to keep the entire address intact to increase the speed of access. A further enhancement available with PBSRAM is that of late write mode, where write data is delayed with respect to the write request such that the delay is the same as that between a read request and when the read data is output. The advantage of late write mode is better utilization of the data bus, since there is no need to insert null operations when switching between read and write requests as is present in standard PBSRAM. Furthermore, DDR signaling can be used with PBSRAM to provide increased bandwidth.

3.2.2 Enhanced SRAM (ESRAM)

Enhanced SRAM [7] was developed by Enhanced Memory Systems, the same company responsible for ESDRAM. ESRAM is interesting due to the fact that it does not use SRAM, instead using DRAM for a much denser, less power consuming memory chip that boasts the performance and pin compatibility with PBSRAM. In order to provide SRAM performance, ESRAM takes advantage of PBSRAM's similarity to SDRAM and applies the same caching technique to reduce initial access latency to just 13 ns.

4. New Approaches in RAM Technology

4.1 Rational Behind Alternative Thinking

Processor speeds continue to grow each year and although much advancement has been made in SRAM and DRAM technologies over the years, the underlying theories remain the same. The gap between processing speed and memory access speed grows at approximately 50% every year. In addition, the DRAM chips of today are becoming so large that managing access to its system is an increasingly difficult task. Designers have developed cleaver ideologies to mask the latency that natural occurs when accessing memory off the main processor chip. There will come a time when these designs will no longer be sufficient as the speed of the modern microprocessor continues to grow in accordance with Moore's Law. While we observe processor speedup of approximately 60% per year, there is only an increase of DRAM access speed of 10% per year. Researches have begun to turn their focus away for the DRAM module itself and taken a look at what else can be conceived to tighten this ever-growing gap of processor clock speed versus memory access latency.

4.2 Intelligent RAM (IRAM)

One possibility that has been devised would merge the processor and memory schemes into a single chip. This approach would immediately solve relieve the latency that occurs when accessing off chip DRAM. The new developments of faster internal DRAM access cannot overcome the time delay that occurs while traversing the wiring and interconnects contained on system boards. By eliminating this separation of processor chip and memory module, researches have increased both efficiency and bandwidth while lessening the overall latency.

4.2.1 Obvious Advantages

Currently, most L1 and L2 caches imbedded in a processor design utilize the speed and accuracy benefits SRAM has to offer. Off chip memory, however, is typically constructed using DRAM due to the higher density offered which in turn creates a greater data storage capacity [8]. Available storage area is the advantage of using DRAM and has caused researchers to focus more heavily on achieving greater access speed on these less costly, higher capacity modules. Integrating the processor into the DRAM construction varies greatly increasing the SRAM contained on the silicon chip. By adapting the processor to memory we can take advantage of the large density of DRAM and expand the bus connection between the two, thus creating bandwidth increases of a factor of 100 [10]. Rather than using the addressing schemes commonly implemented in today's architecture, which utilizes one of several mapping techniques coupled with a multiplexor for address translation, we could basically reinvent the processor to memory interaction. Greater bandwidth would allow us to option to create addressing schemes that do not require multiplexor passing and could use a more direct approach to utilize the vast amounts of DRAM memory available.

This “backwards” design strategy would also offer less access latency due to the close proximity of the processor to the memory device itself. By shortening the wires of the interconnects and alleviating mulitplexing with new addressing schemes the processor can issue address requests and receive data much more quickly. Present designs in testing and use can reduce the overall L2 to memory latency by a factor of 5 to 10 [10]. In addition to the clear advantages of bandwidth increase and latency reduction, the integrated modules display much greater efficiency in regards to overall power consumption. By eliminating the need to traverse board space and drive requests/replies across high capacitance buses, the IRAM design yields less power usage for external references [8]. In fact, the greater density of the DRAM module itself reduces the number of external references required due to its greater storage capacity than that of SRAM.

4.2.2 Inherent Disadvantages

Intelligent RAM creation may seem wonderful in concept and even present great benefits in its realization, but there are several drastic drawbacks to this strategy. The most drastic of these would occur during a system’s upgrading. Today’s system design allows for the processor chip and memory modules to be replaced separately in relation to one another. Adding more DRAM memory to a system or even increasing a system’s processor speed by replacing the microprocessor have become such simple tasks that even the consumer with little computer knowledge can accomplish this task. However, if the memory and processor are joined into a single unit, the upgrading will not be as commonplace. The higher costs of the IRAM units will make consumers less likely to upgrade as frequently and at the same time offer less individual system customization.

The IRAM module would be very beneficial for higher-end machines in which reducing the memory latency to a minimum is essential. The commercial consumer markets are not drawn to this architectural scheme due to the high costs and lack of choices in memory size versus processor speed. Another major drawback of this idea is that the basic design of today’s processor chip and DRAM module are constructed of different material with different interfaces in mind. Creating IRAM modules often requires simplifying the processor architecture and fabrication process or creating new DRAM architectures with greater levels at a higher cost. For this reason, IRAM is typically used for devices such as graphics or audio cards. These applications utilize a simplified and universal connection mechanism, PCI or AGP, and implement smaller amounts of memory combined with processors that are highly specific in nature with a narrow range of instructions to decode. IRAM technology is used today for such devices and may never be used to replace our typical processor chip with a separate main memory.

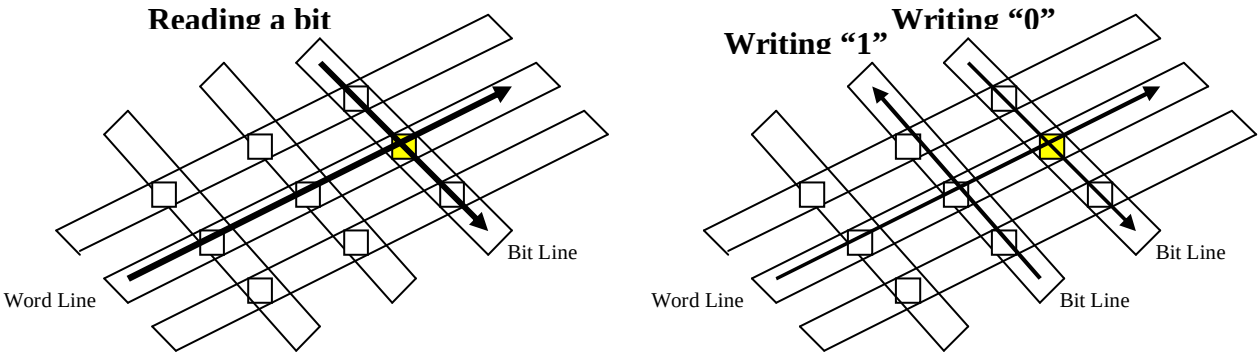
4.3 Magnetic RAM (MRAM)

An architecture currently not available, but showing much promise, is a device that harnesses magnetic storage rather that electron signaling. This new technology may be available on the commercial market as early as next year (2003). Although the units

themselves are new and physical devices are only recently being tested, the theories behind the technology have been around for decades. Instead of using electron "buckets" which must be filled with electrons to signal a "1", or emptied to signal a "0", the MRAM unit utilized magnetic polarity to create a tunneling effect around sensors. IBM first developed these Magnetic Tunnel Junctions in the early 1970's with a fully operational unit in use by 1974 [11]. Even though these were first created nearly three decades ago, it was not until the late 1990's when IBM was able to produce a module capable of storing large amounts of data [11].

4.3.1 Magnetic Tunnel Junctions

The underlying architectural layout resembles that of the current DRAM chip design in that it uses a grid system to represent bit and word junctures. However, instead of an electrical charge being stored at these junctions, a sensory unit is used to detect the magnetic polarity created at the specific point [11]. The figures below represent the reading and writing of bits at each junction. The small squares at each of the junctions represents the insulating material used to create the tunnel barrier between the two overlapping ferromagnetic layers. Arrows are used to represent polarity flow:



Two adjacent layers having the same polarities would create a near zero polarity pull on one another, which in turn would be read as a "0" at that junction. Two layers with opposite polarities would pull at one another around the insulating layer creating a "tunneling" effect around the sensors. This effectively would represent a "1" bit.

4.3.2 Commercial Modules

IBM and Infineon have entered a joint venture in which they hope to produce modules available on the consumer market by 2003; according to their 2000 press release [12]. The new modules are promised to offer the speed of SRAM, the density capacity of DRAM, and the non-volatility of FLASH memory. This breakthrough could revolutionize the memory market as we know it and give way to the realm of "instant-on" computing. The majority of a system's boot up time is spent initializing RAM and loading the operating system. Using MRAM modules that are non-destructive upon reads and system shut downs due to the nature of the magnetic component, we would

effectively reduce this initialization period to nearly zero. The CMOS (Complementary Metal Oxide Semiconductor) could also be replaced by such technology to further reduce the system startup delay time. Although the first available MRAM modules may only be in the 256MB to 512MB range, the advantages of a non-volatile storage device with SRAM characteristics creates an interesting future of the SRAM/DRAM technology of today. The major obstacle will be finding a market for these initially expensive MRAM modules in a market where consumers are attracted to the low-cost, high-capacity DRAM available.

4.4 High-Density Storage Systems

The final area of research that may find a market in the near future is that of HDSS systems, or High Density Storage System. We did not focus much research on this area, but what we did find seems worthy of mention. A possibly storage alternative rather than typical magnetic disk technology may be that Holographic storage. In these systems, a specialized crystal is used to hold three-dimensional page images that can be read by two separate lasers pointed in perpendicular to one another. The reading of this crystal surface would create the possibility of storing information in three-dimensional space rather than the two dimensions we find on magnetic disks. Researches believe these components could offer 125 GB to 1 TB of storage space for first generation systems with 40MB/sec to 1GB/sec transfer rates. This technology is aimed at offering an alternative to disk storage, but devices used to replace system main memory could easily be conceived.

5. Future Memory Architectures

5.1 Fundamental Limitations of DRAM

Researches and system architects are constantly attempting to better the existing SRAM and DRAM technology, but the search may lead to no great improvements in this pursuit. DRAM is reaching a platuea in which the gap between memory access speed and system processor speed continues to grow. Without new approaches, the latency due to off chip memory access may soon become the only system aspect preventing faster computing. Creating integrating devices such as Intelligent RAM create a new approach to further utilizing the existing DRAM technology we have available. In order to make great strides in memory latency reduction, new avenues must be perused. Ideas such as Magnetic RAM and Holographic storage appear to be the focus for future systems.

5.2 The MRAM Future

Semiconductor capacitor-based hardware memory architectures have been the prevailing choice for RAM design since nearly the time of the silicon processor and will continue to dominate the commercial RAM market for years to come. Current research displays a new path, a new focus, for the memories of the future. We feel that if Magnetic RAM modules can be conceived and developed to the levels that are currently promised, this

new approach may soon become the mainstream design of RAM. Consumers are not likely to pay the higher costs initially, considering first generation MRAM modules may be manufactured at the same capacity levels of DRAM modules available on the market. However, in our opinion, the clear speed advantage and possibilities of improvement in future modules will pave a path in the memory market. DRAM will continue to be the choice of memory for many years in commercial, academic, and scientific communities, but magnetic-based memory may quickly find a market and show its dominance.

6. References

[1] W. Wulf and S. McKee. “Hitting the Memory Wall: Implications of the Obvious.” *ACM Computer Architecture News*, Vol. 23, No. 1. March 1995.

[2] Matthias Gries. “A Survey of Synchronous RAM Architectures,” TIK-Report No. 71, Computer Engineering and Networks Laboratory, Swiss Federal Institute of Technology Zurich, 19 April 1999.

[3] T. Mitra. “Dynamic Random Access Memory: A Survey,” Department of Computer Science, State University of New York at Stony Brook, 25 February 1999. Available online at "citeseer.nj.nec.com/article/mitra99dynamic.html."

[4] B. Davis, T. Mudge, and B. Jacob. “The New DRAM Interfaces: SDRAM, RDRAM and Variants”. In High Performance Computing, M. Valero, K. Joe, M. Kitsuregawa, and H. Tanaka, Editors, Vol. 1940 of Lecture Notes In Computer Science, pp. 26-31. Springer Publishing, Tokyo, Japan, 2000.

[5] B. Davis, T. Mudge, V. Cuppu, and B. Jacob. “DDR2 and Low Latency Variants.” In *Proc. Memory Wall Workshop at the 26th Annual Int’l Symposium on Computer Architecture*, Vancouver, Canada, May 2000.

[6] Richard Crisp. “Direct rambus technology: The new main memory standard,” *IEEE Micro*, 17(6):18 - 28, November 1997.

[7] Enhanced Memory Systems. “72Mbit DDR ESRAM 2Mx36 Preliminary Datasheet.” Available online at “www.edram.com/products/datasheets/ss2615ds_r1.0c.pdf.”

[8] D. Patterson et al. “A Case for Intelligent RAM”, Computer Science Division/EECS Department,University of California, Berkeley, CA, 2/10/97

[9] D. Patterson et al. “Intelligent RAM (IRAM): the Industrial Setting, Applications, and Architecture,” *ICCD '97 International Conference on Computer Design*, Austin, Texas, 10-12 October 1997.

[10] K. Keeton et al. “IRAM and SmartSIMM: Overcoming the I/O Bus Bottleneck,” *Workshop on Mixing Logic and DRAM: Chips that Compute and Remember*, Denver, CO, USA, 1 June 1997.

[11] “Magneto-Electronics: Magnetic Tunnel Junctions”, IBM Science and Technology at Almaden. Found at:
http://www.almaden.ibm.com/st/projects/magneto/mtj/

[12] “IBM, Infineon partner to commercialize magnetic RAM”. Article by Jack Robertson, ebn United Business Media, December 2000.