Assessment of MRAM Technology Characteristics and Architectures

Rajagopalan Desikan* Stephen W. Keckler Doug Burger

Computer Architecture and Technology Laboratory
Department of Computer Sciences
*Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712-1188

Department of Computer Sciences
Tech Report TR-01-36
The University of Texas at Austin

Feb 5 2002

ABSTRACT

High memory latency and limited off-chip memory bandwidth have proven to be one of the major hurdles restricting the performance of computer systems on high-performance computing applications. In an effort to alleviate this problem, microprocessor designers have continued to increase the amount of on-chip memory. Until recently, fast memory has been in the form of SRAM caches and performance on some applications has been enhanced by increasing cache capacities and the number of levels in the cache hierarchy. Alternate technologies, such as embedded DRAMs can enhance on-chip memory capacity at some cost of higher latency. An emerging memory technology is magnetic RAM (MRAM) which can be fabricated on top of a conventional processor core. MRAMs offer potential advantages in density and non-volatility over conventional MOSFET memory structures, and may motivate alternate memory system architectures to exploit these characteristics for higher performance and reliability. In this report, we analyze the characteristics of SRAMs, embedded DRAMs, and MRAMs, and examine how MRAM technology may be employed in high performance computer systems.

1 Introduction

Until recently, the lowest levels of microprocessor memory hierarchies have been implemented using Static RAM (SRAM). Located on-chip, this memory supplies high-bandwidth and low-latency access to computation elements, and with shrinking feature sizes, larger level-1 caches and now level-2 caches are appearing on the same dice as the processors. Still, limited SRAM capacity necessiates more off-chip memory access, and this combined with an asymmetric scaling of processing power and pin speed, has exacerbated and increased memory latency. Furthermore, SRAM is susceptible to failures from external radiation sources and is likely to consume significant power from subthreshold leakage in future technologies. Capacity can be increased dramatically using embedded DRAM (eDRAM) technologies, but this technology does not help reduce bandwidth limitations to large capacity off-chip memories and static power consumption.

This document examines a newer magnetic RAM (MRAM) technology as an alternate or supplement to SRAM and eDRAM. MRAM has the potential advantages of small dimensions, high bandwidth, and non-volatility. Furthermore, current research is examining methods of stacking multiple layers of MRAM above a silicon substrate to achieve substantial densities in a single package. A number of MRAM memory designs have been proposed recently [14] [17] [15] [10], but it is presently unclear which design will prove the most feasible. The purpose of this document is to compare the features and scalability of MRAM memories to evolutionary SRAM and eDRAM technologies, and to present preliminary memory system concepts constructed from MRAM technologies.

2 Competing Memory Technologies

Advances in fabrication technology have resulted in a number of memory technologies emerging as viable alternatives to on-chip SRAM, including embedded DRAM and MRAM. In this section, we compare the various device-level and operating characteristics of these technologies in an effort to understand each technologies advantages and disadvantages. The features for comparison are taken from published technical papers. As such, an apples-to-apples comparison is impossible and the results of this section are intended to provide a qualitative, rather than a quantitative comparison of the technologies.

2.1 Device-level Characteristics

The device-level characteristics of memory technologies include cell area, power, and reliability. These characteristics are heavily influenced by the physical characteristics of the device and the process technology.

2.1.1 Cell Area

The cell area plays an important role in determining the viability and use of a given memory technology. Historically, DRAM designs have concentrated on achieving minimum cell area, while SRAM designs have concentrated more on achieving high speed. The cell area can be specified using the process technology independent metric λ, which is equal to half the minimum feature size at a particular technology. The standard 6T SRAM cell occupies approximately $550\lambda^2$ [11], while recent embedded DRAM designs have demonstrated $80\lambda^2$ cells [16]. Off-chip DRAMs have smaller cell area, and recent designs have demonstarted 20-$40\lambda^2$ cell area [9] [20]. Several different MRAM cell designs have been proposed with the cell area ranging from $12\lambda^2$ for the GMR MRAM architecture,

to as small as $6\lambda^2$ for the cross point MTJ MRAM architecture [14]. However, demonstrated MRAM cell areas have been $80\lambda^2$ [10].

2.1.2 Power

With shrinking feature sizes and increased transistor counts, power has become a first order design constraint. Dynamic power is increasing due to clock rate increases, while static power is increasing because of larger leakage currents resulting from lower threshold voltages. Leakage power is becoming a major challenge in SRAMs, which has led researchers to explore various strategies to reduce SRAM power consumption [5] [8]. Power consumption is directly related to performance, and high performance designs consume more power. Using Ecacti [13], we computed the power consumed by the 256 KB (2 Mbit) cache used in the Pentium 4 as 1 W [13]. An 8 Mbit eDRAM design dissipates 10 W at 1 GHz, 0.175μm technology [16]. Initial MRAM designs have demonstrated 24 mW power consumption while running at 20 MHz in a 0.6μm technology for a 256Kb part [10]. To compare these numbers, we divide the power consumption by the size and frequency of the part to obtain the the size and frequency independent metric, J/bit. These comparisons are shown in Table 1 and demonstrate that dynamic power for SRAM and eDRAM technologies is comparable, while it is potentially higher for MRAM technology.

We expect the power consumption of MRAMs to reduce as technology scales and as power/performance trade-offs are made. For example, SRAMs and eDRAMs can be designed for low power by sacrificing some performance [4] [7]. However, MRAMs face additional challenges due to the asymmetric power requirements of read and write operations. A recent study has shown the write power to be 8 times the read power [15].

2.1.3 Reliability

Integrated circuits are becoming increasingly susceptible to noise as feature sizes continue to shrink. Thus, reliability is becoming an important issue in both logic and memory designs. The reliability of a particular memory technology will depend greatly on the method the technology uses for storing information. SRAMs store the information using flip-flops, and a typical 6-T SRAM stores both the data bit and its complement using a feedback arrangement. Thus, it is more robust than eDRAMs which store their information in the charge of a capacitor node. MRAMs store their information using the direction of spin of electrons, and are thus less susceptible to strikes from external energized particles. However, this does make them susceptible to stray magnetic fields from other circuits.

2.2 Operational Characteristics

The operational characteristics of a memory system include the access times (read/write) and achievable bandwidth. These characteristics depend on the physical device characteristics, and the architectural organization of the memory technology.

2.2.1 Read and Write Times

The Read time is defined as the interval from when the address is placed on the address bus and when the data is available at the sense amplifiers. Read time is an important component of the latency of a given memory technology, particularly in caches. A number of low latency SRAM and eDRAM designs have been proposed recently. The load-use access latency of the 256 KB level-2

cache used in the Pentium 4 is 7 cycles at 1.5 GHz [6]. If we consider 4 cycles as time taken to access the level-2 cache array, then the read time is equal to 2.7 ns. A 3.7 ns access, 1 Gb eDRAM was demonstrated recently at 0.18 μm technology [16]. MRAM speeds are nowhere close to this, with the cycle time being approximately 50ns in an initial 256Kb design at 0.6 μm technology [10].

For DRAMs, architectural techniques like page mode access are used to reduce the access latency. However, page mode access results in the first access to DRAMs being much slower than the later accesses to the same row. Embedded DRAM designs also suffer from the same drawback, and this can significantly hurt the performance of applications which have irregular memory accesses. Embedded DRAMs also need to be refreshed periodically, and cannot be read when the refresh operation is in progress. MRAMs do not need to be refreshed and hence do not suffer from this drawback.

Write times usually do not vary greatly when compared to read times for SRAMs and eDRAMs. Writes for MRAMs on the other hand require more current than reads, and hence are likely to be slower. However, writes are not usually on the critical path, and architectural techniques like store buffers can often be used to hide any extra write latency.

2.2.2 Achievable Bandwidth

The achievable bandwidth from a given memory technology depends on the architectural organization of the memory system. Off-chip DRAMs have traditionally been organized as high latency high bandwidth devices, while SRAMs have been organized as low latency moderate bandwidth devices. As embedded DRAMs can be fabricated along with logic, they reduce the latency while providing high bandwidth. Recent studies have demonstrated eDRAMs with 128-Gbytes/sec throughput at 1 GHz [16]. The 256 KB level-2 cache in the Pentium 4 chip from Intel is implemented using SRAMs and achieves 48-Gbytes/sec throughput at 1.5 GHz [6]. Embedded DRAMs are likely to have higher bandwidth than SRAMs due to the smaller cell size, which results in more input-output wires. MRAMs have cell size similar to eDRAMs and can also take advantage of vertical interconnects. Moreover, the ability to fabricate multiple layers of MRAM memory will also result large amount being available very close to the processor. Hence, they are likely to have a higher bandwidth than eDRAMs.

2.3 Characteristics Summary

In this section, we have compared some of the important characteristics of the state of the art in SRAM, eDRAM, and MRAM technologies. As seen from these characteristics, each memory technology has some advantages and some disadvantages over the other technologies. Table 1 summarizes the characteristics of the three technologies evaluated in this section, using high level published data.

As mentioned earlier, it is hard to compare these numbers directly due to various factors like difference in process technology, frequency, size, and architectural organization. For example, even though the MRAM cells are roughly 7 times smaller in Table 1, actual MRAM cell size is larger than SRAM cell size because of the smaller process technology used in the SRAM cell. However, we expect the MRAM cell area to scale with technology. The comparison of power also presents similar difficulties. We have tried to normalize power by defining the J/bit metric which takes into account the part size and frequency. Using this metric, we see that the MRAM power consumption is higher than SRAM and eDRAM power consumption. However, we have not taken the process technology into consideration, and we expect MRAM power consumption to reduce when it is scaled to smaller technologies.

Table 1: Preliminary estimates of SRAM, eDRAM, and MRAM

	SRAM	eDRAM	MRAM
Technology	0.18μm	0.18μm	0.6μm
Cell Size	$550\lambda^2$	$80\lambda^2$	$80\lambda^2$
Power	1 W at 800 MHz, 2048 Kb	10 W at 1 Ghz, 8 Mbit	24 mW at 20 MHz, 256 kb
Normalized Power (J/bit)	1.3×10^{-15}	1.2×10^{-15}	4.6×10^{-15}
Read Time	2.7 ns	3.7 ns	50 ns
Bandwidth	40-100 Gbytes/sec	128-1000 Gbytes/sec	N.A.

3 MRAM Design Issues

Several attributes differentiate MRAM technologies from solid-state technologies such as SRAM and eDRAM. This section outlines those features that are likely to affect memory system architectures employing MRAM.

Feasible capacity per layer: While the raw density of MRAM bits is a function of the minimum magnetic flux, estimation of number useful bits per layer of MRAM depends on a number of other factors. First, some significant fraction of the wiring layers must be reserved for power distribution in the MRAM layers. Second, routing wires must be available to connect the MRAM bits to sense amplifiers located in the active layers, which are layers containing the active devices like MOSFETS. Third, if MRAM layers are to be stacked, then space devoted to vias must be allocated to connect from the upper layers to the active layer. Finally, while some off-chip connections could be made from the periphery of the chip, achieving substantial off-chip bandwidth may require area bonding. Thus, vias may be required to connect the active layers to package pins and the outside world. Although many of these interconnect wires will likely be on different metal layers than the active magnetic bit and word lines, the effective density of the MRAM layers will be determined in part by how the MRAM bits can be connected into the network.

Another potentially fruitful research area is in methods of stacking multiple active silicon layers in a 3-D structure. This could enhance the utility of dense MRAM structures as the sensing circuits for each MRAM layer could be placed nearby. However, the vias for communicating between the layers as well as distributing power and off-chip bandwidth may still limit some of the useful area in the stack. Current areas of research include vertical bus structures [12] and 3-D SOI [21].

Non-uniform access times: Physical proximity in future systems will have a substantial impact on communication latency. For example, on a conventional silicon integrated circuit with projected clock rates and wire dimensions, the latency across the diameter of a chip in a 35nm technology is likely to exceed 30 cycles [1]. This non-uniformity in access times has serious implications for on-chip memory systems; current work is examining non-uniform cache architectures (NUCA) to address the capacity and latency trade-offs [3]. MRAM memory systems are likely to be subject to the same constraints. Data located in MRAM layers close to the surface of the active silicon circuits are likely to be accessed faster than layers higher up in the stack if the vertical interconnects introduce appreciable delay. MRAM access latency must also be incurred if the lateral distance between the data and the location in the active region where it is needed is large. In essence, the two-dimensional routes in today's integrated circuits become three-dimensional routes with MRAM

stacks. It is conceivable that average distances between sources and sinks of data would decrease when migrating from a 2-D to a 3-D topology [21].

Bandwidth: Two components of bandwidth are important in MRAM memory systems. First is the bandwidth to the processing elements in the active silicon substrate. This bandwidth is a function of horizontal and vertical wire dimensions (cross-sectional area) and the speed at which the MRAM cells and wires can be switched. If vertical wire dimensions from the MRAM layers can scale with the horizontal wire dimensions found in the active region, then the bisection bandwidth between the memory and the processors could exceed that of conventional 2-dimensional topologies. The second bandwidth component is the connection between MRAM and the outside world. This bandwidth will be significantly smaller than the internal bandwidth as external communication channels will likely employ transceivers implemented in the active region. Thus incoming data would go to the active region and then up into the MRAM. The bandwidth is then limited by the speed and bisectional area of the connections that can be made to the active region.

Asymmetric reads and writes: Because writes to the MRAM arrays require changing the magnetic polarity, power consumption and delay are likely to be much larger than for reads. Scheuerlein *et al.* report write power to be 8 times that of read power [15]. While writes are typically much less frequent than reads, particularly since caches filter both reads and writes from deeper regions of the memory hierarchy, this large asymmetry between read and write cost will likely affect the memory hierarchy design. Multiple writes might be buffered and aggregated into a single write to conserve power and delay. This approach is similar in nature to proposed non-volatile memory system architectures, such as eNVy [18].

Non-volatility: Using magnetic rather than electrical storage provides a number of potential benefits. Magnetic storage may be less vulnerable to atmospheric radiation sources (such as ionized neutrons and alpha particles) but its susceptibility to other noise sources, such as stray electrically induced magnetic fields from the active layers, is unknown. Clearly, magnetic non-volatility has uses for storage in mobile devices. However, it may also be useful in high performance systems that are never powered down because power consumption is reduced when the memory system is not in use. In contrast, leakage current forces DRAM memories to be refreshed and SRAMs to dissipate power even when not being accessed. Leakage currents in both of these technologies are likely to increase with shrinking features sizes [5].

Scalability: Transistor devices are expected to continue to scale at an aggressive rate, with feature sizes shrinking by a factor of four over the next 12 years. This scaling will likely realize at least an order of magnitude in density of SRAM and DRAM memory arrays. MRAM is in its infancy and bit area in current prototypes is approximately $7.1\mu m^2$ at $0.6\mu m$ technology [10], which is 1.7 times larger than an SRAM bit in today's $0.18\mu m$ technology [11]. If the areal density in the disk drive industry can serve as a guide, MRAM bits should shrink substantially, and wiring will likely become the limiting factor.

Sensing: MRAMs use the resistance of the Magnetic Tunnel Junction (MTJ) cell for information storage [14]. The information is retrieved by sensing the current through the MTJ cell for the two different resistance values. For stable MRAM operation, a magneto-resistance (MR) ratio of 40% or higher is generally needed, although some recent sensing schemes work well with a lower

ratio [19]. With a higher MR ratio, the sense amplifiers can be placed farther away from the cells. With a low MR ratio, if the sense amplifiers are placed far away from the MRAM cell, the signal transition might be too weak to detect. This fact should be kept in mind while designing MRAM banks, as reading large banks may prove challenging.

4 Possible MRAM Applications

The characteristics of MRAM that separate it from other technologies include density, bandwidth, asymmetric access times, and power. In this section, we suggest several applications for MRAM technologies for high performance computer systems and describe at a high level how the attributes of MRAM may motivate different memory architectures and organizations.

4.1 High Bandwidth Caches

A conventional architecture may employ MRAM as an independent level in the cache hierarchy. This technology assessment indicates that MRAM cannot deliver the speed or bandwidth of a small SRAM array and thus is not suitable for a level-1 cache. However, the capacity and latency is likely amenable to a level-2 or perhaps a level-3 cache. In a chip-multiprocessor architecture, the MRAM stack could be partitioned so that the layers above one of the processors and its level-1 cache would constitute that processor's principal level-2 cache. This organization would enhance physical locality for each processor to its memory system and decrease communication distances among the processors, thus reducing inter-processor communication latency. It also remains to be seen if a single memory controller will be enough to orchestrate accesses to multiple MRAM banks. If multiple controllers are required, designs must consider the space and time overhead associated with them.

A second factor when organizing an MRAM stack as a cache is that distant layers will require longer access times than nearer layers. This non-uniform access time motivates an architecture that attempts to place most frequently accessed data nearest to the processor and less frequently accessed data (but still data that is useful to cache) in layers further from the processor. We are currently examining non-uniform cache architectures for a two-dimensional silicon substrate and policies for data placement and migration. The results from these experiments can be applied to MRAM architectures. SRAM row buffers in the active layer can be employed to cache data from the MRAM banks in the higher layers.

4.2 MRAM Main Memory

With on-chip memory increasingly occupying more real estate on a processor die, MRAMs might completely replace DRAM banks as main memory in a computer system. If multiple layers of MRAM can provide enough capacity, then external DRAM banks can be completely eliminated, with users buying MRAM processors with more memory when the need arises.

Even if we do not completely eliminate external DRAM banks, the properties of MRAM can be leveraged in some other interesting ways.

- As an alternative to automatic management of data through caching, MRAM could be effectively used by the operating system or application as a software controlled cache. For example, the capacity of MRAM may allow the operating system to migrate hot pages of data from disk or main memory into MRAM storage.

- An operating system could map a subsection of the file system to non-volatile storage to speed reboot or accelerate applications with frequent disk access. This technique, and the previous one, may be applicable to programs with enormous data sets and to enhance out-of-core computational algorithms.
- An application may employ MRAM storage as separate from the cache hierarchy, perhaps as high bandwidth storage to selected parallel data structures.
- Due to the high cost of writes to MRAM, the operating system can elect to store only clean pages in MRAM and store the dirty pages in DRAM instead.

4.3 Data Structures for Reliability

The non-volatile properties of MRAM may enable its use in improving reliability through more efficient logging and recovery. Today's systems maintain stable update logs (if they support logging) on disks and require substantial disk access to recover from a crash. With high bandwidth access to stable storage, logging can be more frequent and efficient. Furthermore, recovery times will decrease due to the speed and bandwidth available to restore state from MRAM-based logs and checkpoints.

5 Modelling MRAMs

To evaluate the potential merits and demerits of MRAM technology, we have begun to develop an analytical model to explore the design space. The evaluation can be broadly broken up into two parts : a low-level model to give access times and area overhead for MRAMs, and a high-level full system simulator which uses these values to evaluate system performance. In this section, we describe the low-level area and timing model as well as the high-level architectural full system simulator model we have developed for this purpose.

5.1 Area and Timing Model

Our current memory model incorporates multiple layers of MRAM memory on top of an active silicon substrate. We assume that they are no active devices in the MRAM layers, and hence all the control circuitry resides on the base active layer. Preliminary architecture models include a collection of microprocessors also in the active layer as in shown in Figure 1. The wordlines for each bank in this model are routed from the active layer up to the MRAM layer along the z-axis, and then routed horizontally from the sides along the x-axis. The bitlines are routed along the y-axis in the MRAM layer, and routed vertically along the z-axis to connect with the devices in the active layer.

Figure 2 shows the read operation for an MRAM cell in the present configuration. During a read, the bitlines are held high and all the word lines are initially held high. Then the wordline corresponding to the word being read is pulled low. Current flows in the MRAM cell along the path shown in Figure 2. The magnitude of this current depends on the resistance of the MTJ stack. Hence, the voltage drop seen in the bitline will depend on the MTJ stack resistance. The sense amplifiers sense this voltage drop to detect the value stored in the MRAM cell. The diode in the MRAM cell should have high resistance when it is reverse biased, and hence thin film diodes are used in the MRAM cell. The diode performs the following two functions:

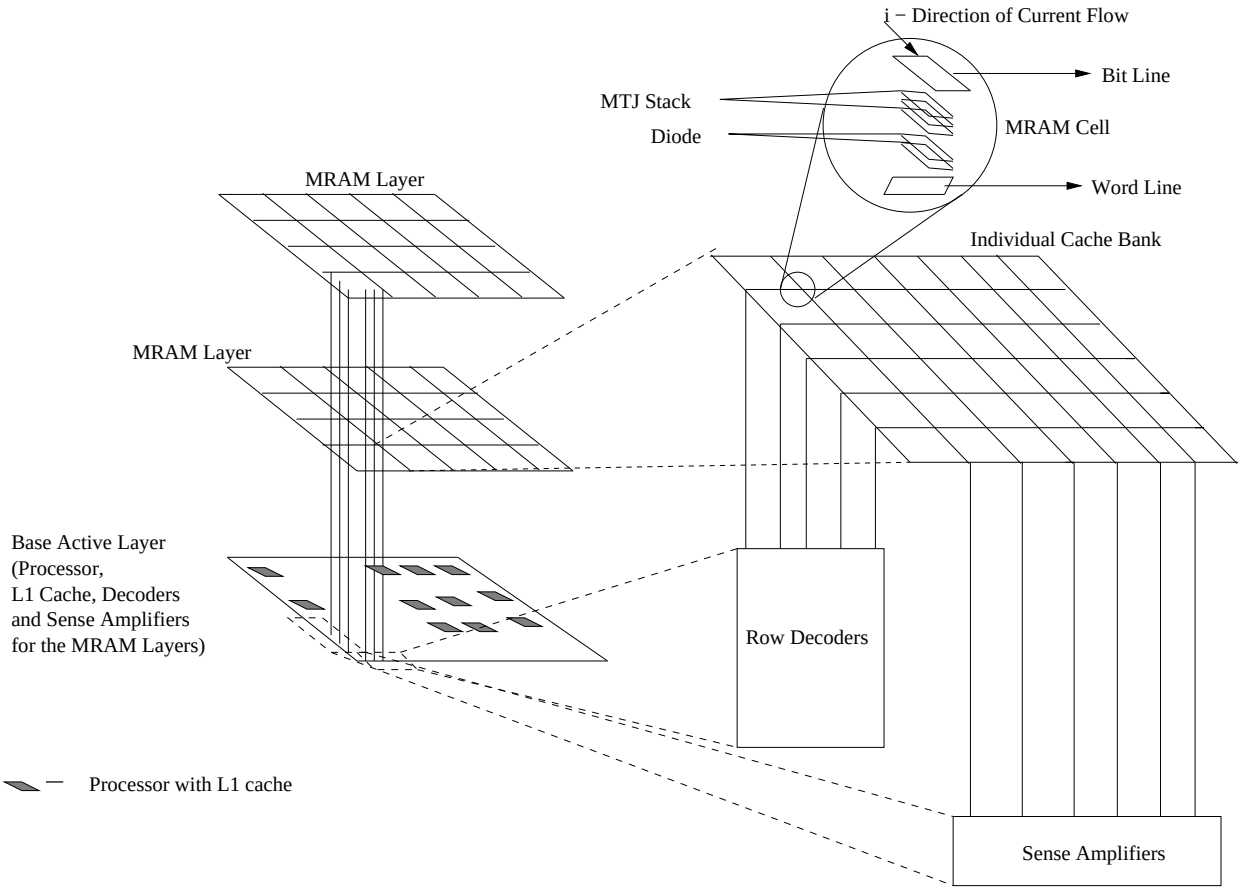

Figure 1: MRAM Model

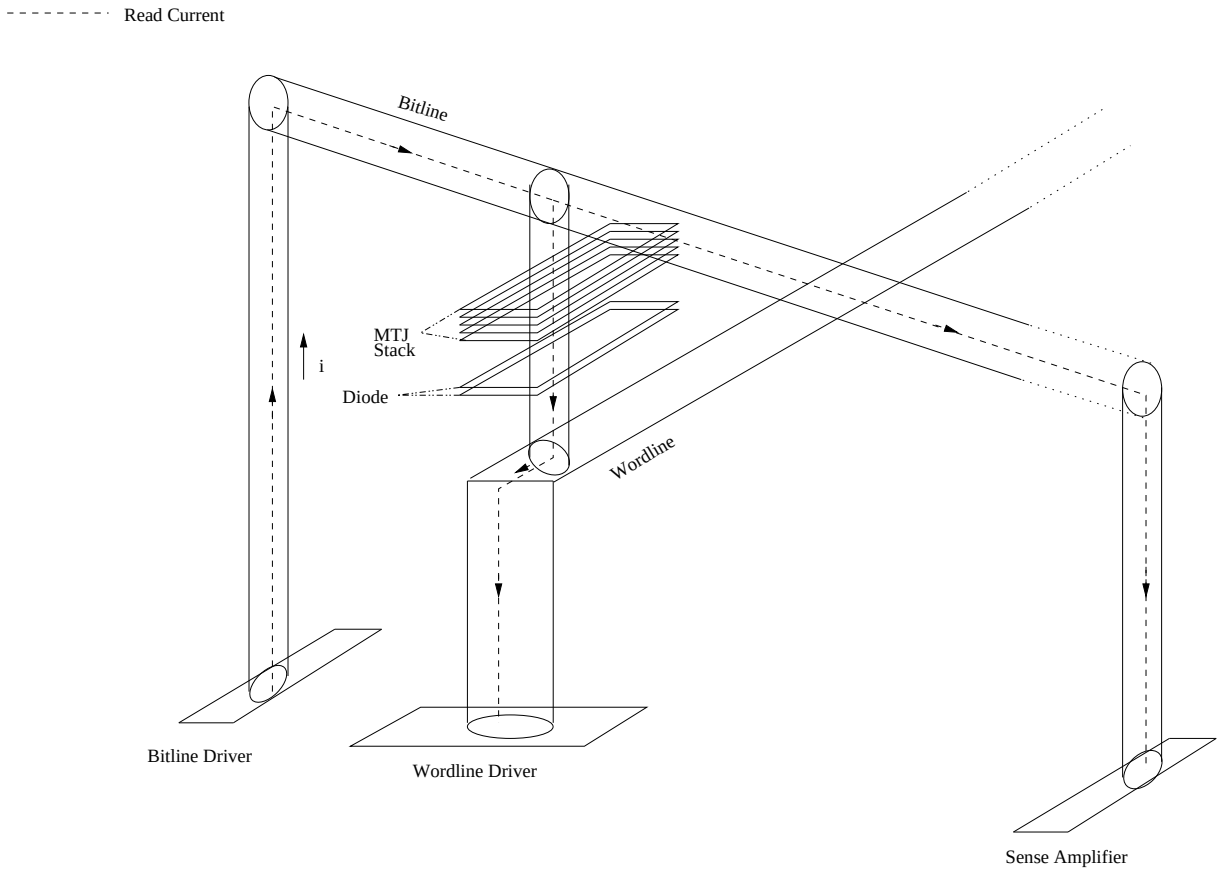

Figure 2: MRAM Cell Operation

1. It prevents voltage drop in the bitline due to leakage current in the MRAM cells whose wordlines are high.

2. It prevents damage to the MTJ stack by not allowing the write current in the wordline to flow through the MTJ stack.

5.1.1 Area Modeling

The area model takes the x and y dimensions of a bank as input and estimates the via overhead in each MRAM layer. If the die size and the technology are given, it also estimates the number of banks that can be accomodated on the die. If we give the number of layers as more than one, than the area model gives estimates for three cases:

1. Each layer has its own set of word and bit lines

2. Each layer shares the word lines but has its own set of bit lines. This amounts to folding the MRAM layer along the x axis.

3. Each layer shares the bit lines but has its own set of word lines. This is equivalent to folding the MRAM layer along the y axis.

The area model also estimates the area overhead associated with the sense amplifier and the row decoder logic in the active layer. The active layer area currently accounts for only row decoders, wordline drivers, bitline drivers and sense amplifiers. We do not account for the tag area now as presently it is not clear where the tags (if any) will be stored in MRAM or SRAM structures. The row decoder consists of the decoder driver, the first stage NAND gates, and the second stage NOR gates. The wordline driver consists of two inverters. The area for these structures is derived using CACTI [13].

The following questions remain open with respect to area modeling:

- If MRAM is treated as a cache, the tags could be stored in the active devices layer or they could be stored in the MRAM layer along with the data. If the tags are stored in the active layer, then the model must account for this area occupied in the active layer. If instead, the tags are stored in the MRAM layer, then the MRAM data capacity will be reduced. The benefit of storing tags in the active layer is faster lookup. Data lookup could be started in parallel with tag lookup, and the request could be squashed if the tags don't match. The downside of course is the active layer area used up by the tags.

- The vertical interconnects for control and data signals to the higher level MRAM layers will have to pass through all the layers below. If the vertical lines are not shared among the vertical layers, a larger portion of the area in lower layers will be occupied by these vertical wires. If we decided to put more MRAM memory banks in the higher layers to take advantage of the area available, then it might be harder to fabricate. Putting the same number of memory banks in all layers results in easier fabrication. Using our area model, we computed the capacity of different MRAM layers assuming lesser number of vertical interconnects for the higher layers, and found that the difference in capacity between the first and the fourth layers is 0.68%.

- Different size vias could be a solution to reduce access time to higher MRAM layers. The larger area available in the higher layers could be used for fabricating wider vias with lower resistance, thus reducing access time to the higher layers.

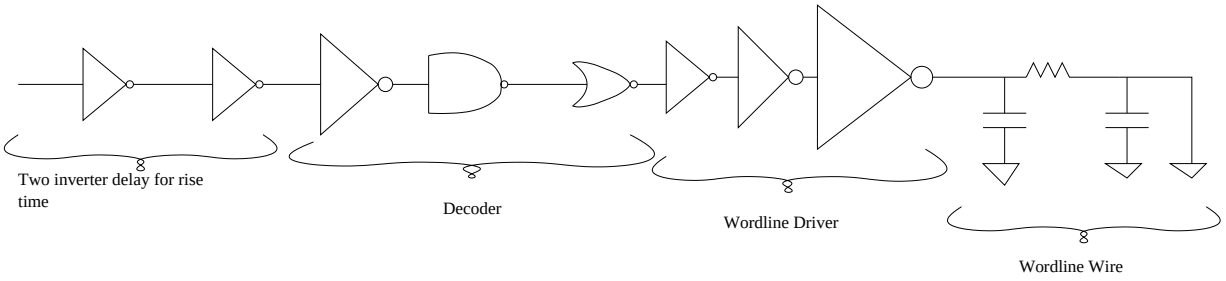

Figure 3: Components of Wordline Delay

5.1.2 Timing Model

In our timing model, the base unit of access is a bank. A bank is like a small cache with its own set of decoders and sense amplifiers for retireving data from MRAM cells in the bank.The delay model computes the time taken by a read and write operation to one bank of the MRAM array. Thus the granularity of read and write access time is assumed to be at a bank level in our model. The time for the request to reach the decoder circuit for the bank is presently not considered, although it will be incorporated in the future. Also not considered is the time taken for the data to reach the component (processor, L1/L2 cache) requesting the data from the sense amplifiers of the bank. The model currently gives the worst case access time for a bank. Our model is similar to the CACTI cache model [13]. The delay model has the following components:

- Decoder
- Wordlines
- Bitlines
- Sense Amplifiers

The wordline architecture is shown in Figure 3. Depending on the address, one NOR gate will have a high output while others will have low output. We use an odd number of inverters in the wordline driver. This results in all but one wordline going high. The total wordline delay is thus composed of the decoder driver delay, the decoder delay (consisting of the NAND and NOR gates), the wordline driver delay, and the delay due to the resistance and capacitance of the vertical and horizontal wires.

The bitline delay is influenced by multiple capacitances The capacitances considered are the capacitance of the thin film diode, the MTJ stack, and the bitline wire. The bitline architecture is shown in Figure 4. The magnitude of the current flowing through the bitline will depend on the resistance of the MTJ. Ideally, we would start the bitline current when we access a bank. Once the bitline current reaches I_{ref}, we will pull the selected wordline low to detect the bit stored. Note that this model requires all wordlines are maintained at V_{ref} before sending current through the bitline. The delay in this case would correspond to the sum of delay for the bitline current to reach I_{ref}, the time for the diode to start conducting when the wordline starts going low, and the time it takes for this change in current to propogate to sense amplifier. The leakage current due to the capacitance of the diodes which are not turned on will also influence the delay.

We currently do not know the architecture of the current source and the sense amplifier. We also do not know the resistance and capacitance of the thin film diode, and the capacitance of the tunnel junction. Hence, in our delay model, we assume that the bitlines are continuously powered

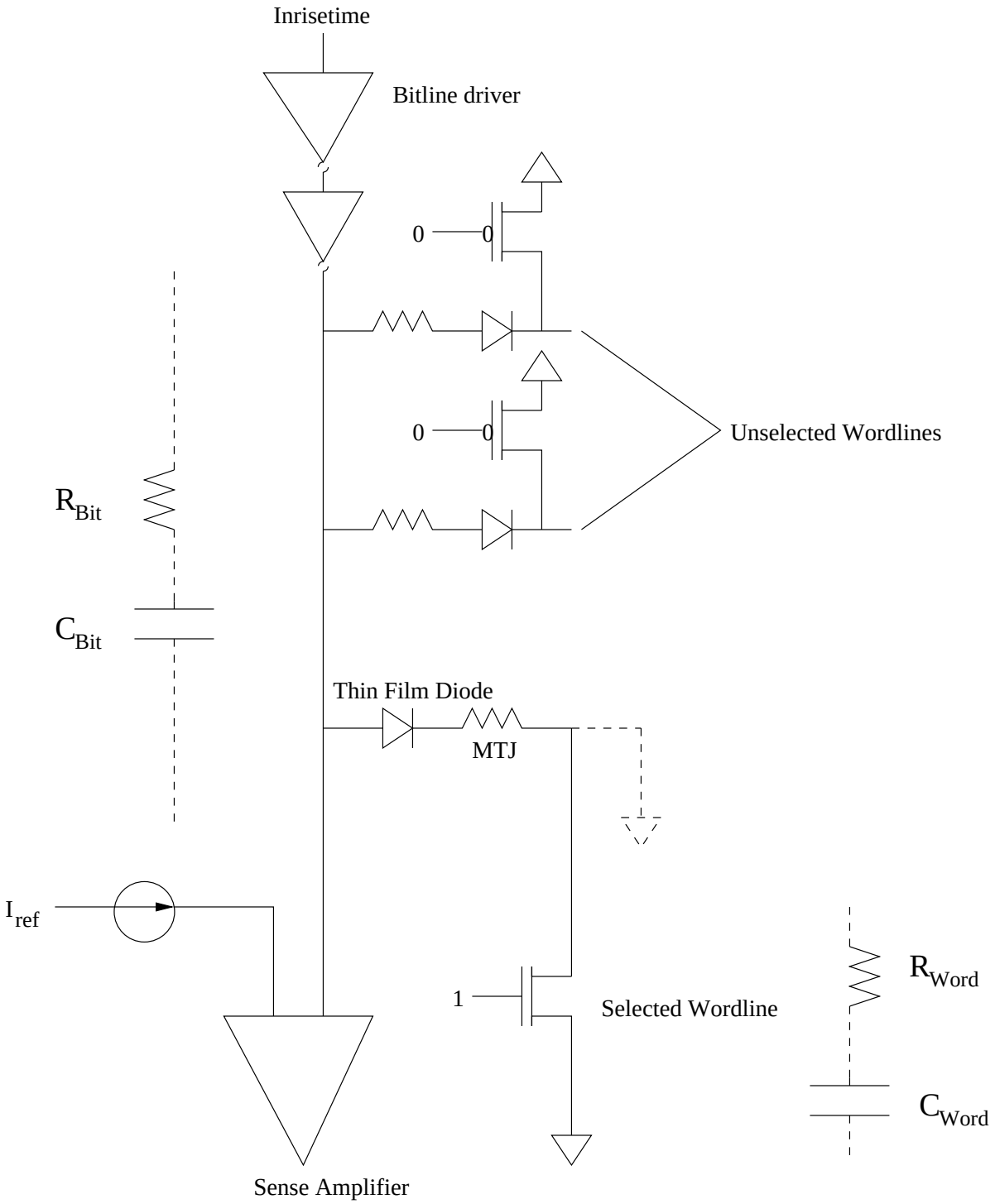

Figure 4: Components of Bitline Delay

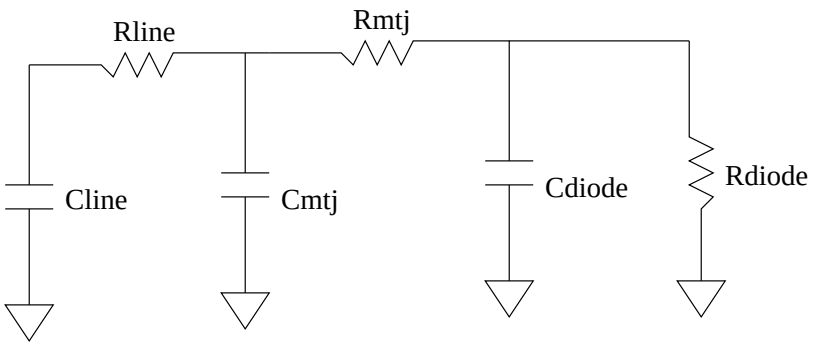

Figure 5: Bitline Delay Circuit Model

with I_{ref}. We also assume that the wordline will have to be pulled all the way to ground before the sense amplifiers can reliably detect the value stored. Thus the total delay is equal to the delay for the wordline to be pulled to zero, and the delay of the change in current to propagate down the wire from the selected wordline to the sense amplifier. The resistance and capacitance of wires for both the wordline and the bitline delay have been obtained using analytical wire model of Agarwal et. al [2].

The area of the MRAM cell is assumed to be 5 λ x 5 λ, which we feel is the cell size which will make MRAM competitive with other memory technologies. However, the interconnect size may be larger than the cell size, and thus we might be forced to stagger interconnects to achieve compact banks. The smaller cell size substantially increases the MRAM memory available in each layer. However, it also increases the amount of active area needed for the decoders and the sense amplifiers in the active substrate. To reduce the active area, we might have to build very large MRAM banks. We are currently exploring Gigabit DRAM decoders for use in MRAM banks. One way to address a large bank using a small number of bits is to increase the granularity of addressing to a word or higher. The number of independently addressed banks should match the number of average concurrent requests which we expect the processor(s) to generate.

To save power, it might be possible to mutiplex adddress bits before sensing the cells. This can be achieved using a large number of narrow banks. When an data request is received, the request can be directed to the bank holding the data and the data in one row can be transferred to the whoever is requesting it. This way, we avoid multiplexing the output of the sense amplifiers.

5.2 Architecture Models

In Figure 6 we show one possible MRAM architecture. In this architecture, the MRAM memory is treated as the level 2 cache. We logically divide MRAM memory banks into two namely, macro banks and micro banks. The microbanks are the actual physical MRAM banks. The level 1 cache is connected to a macrobank controller. Blocks of the level 1 cache are mapped to different macrobanks, and these mappings are maintained by the macro bank controller. Depending the address, the macrobank controller directs the request to one of the micro bank controller.

The tags for the MRAM are maintained in the active layer in this model. The micro bank controller does a tag match of the incoming address, and on a match checks to see if the buffer contains this data. If the requested data is not found in the buffer, it sends a request to the appropriate bank. The layers of MRAM could be organized in a number of ways, and the organization will greatly influence the active area consumed by the control circuitry and the access time.

The data from the MRAM can be read in large chunks and stored in buffers in the active layer.

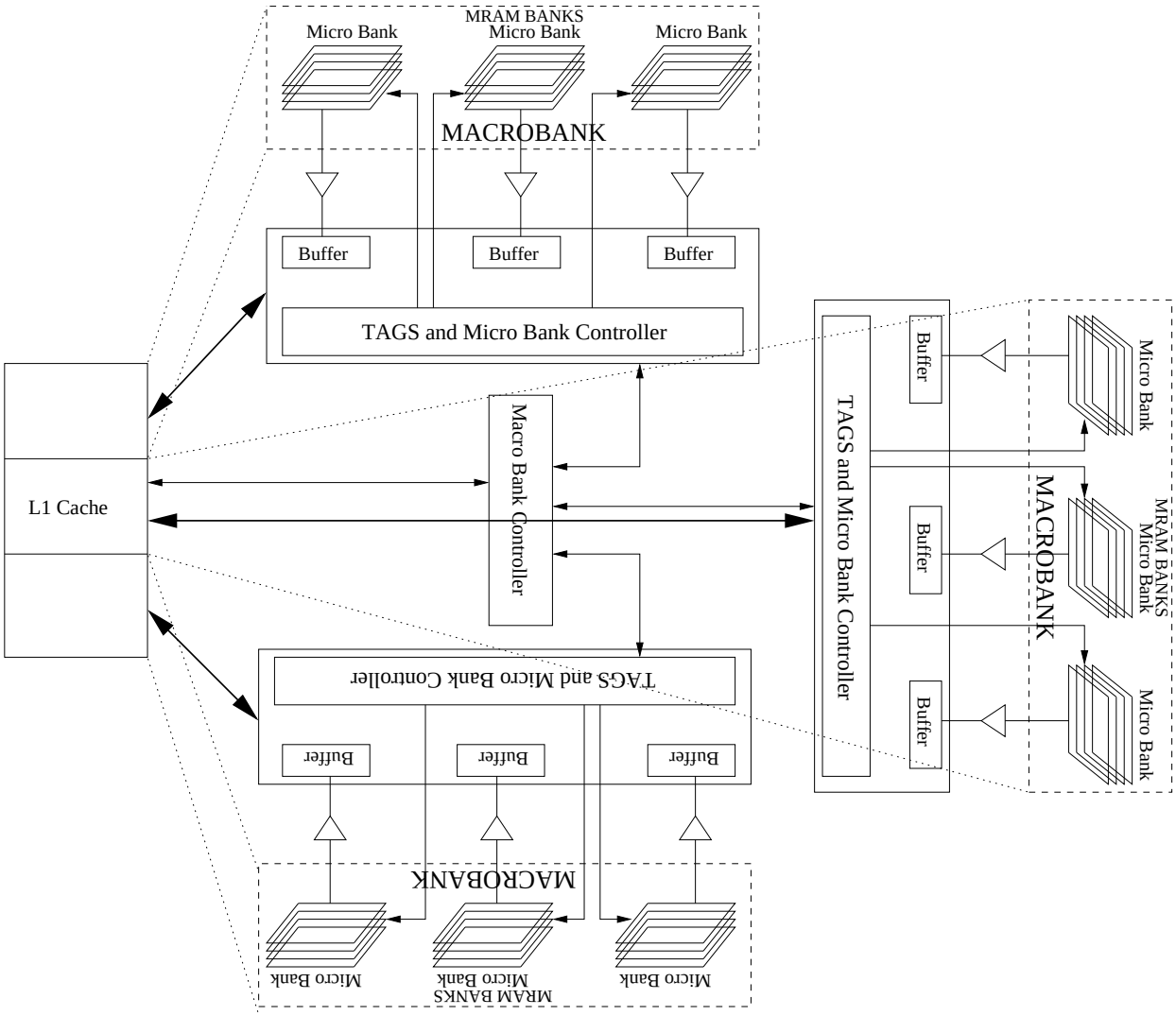

Figure 6: Architecture Model 1

Future requests to the same bank can be serviced by the buffer if the data is found in the buffer. The advantage of this organization is that the data is only read if it is known to be available in the cache, and hence we do not waste power unnecessarily reading data and discarding it. However, this organization increases latency as the data request is initiated after the tag match and not in parallel. The tags will likely be stored using SRAM memory, and hence we need to consider leakage energy due to the SRAM cells also in this scheme.

In Figure 7, we show another architecture model for organizing MRAM memory. The main diference in this organization is that the tags are stored in the MRAM layer rather than the active layer. This organization results in less active area usage but reduces capacity in the MRAM layer as now the tags also need to be stored along with the data. We also cannot determine if there is a hit or miss without sending request to the MRAM layer.

In this model also we have macro and micro bank controllers. the macro bank controller sends the request to the appropriate micro bank controller, which in turn determines the correct bank and the correct layer to send th request to, and sends the request. Once the tags and the data

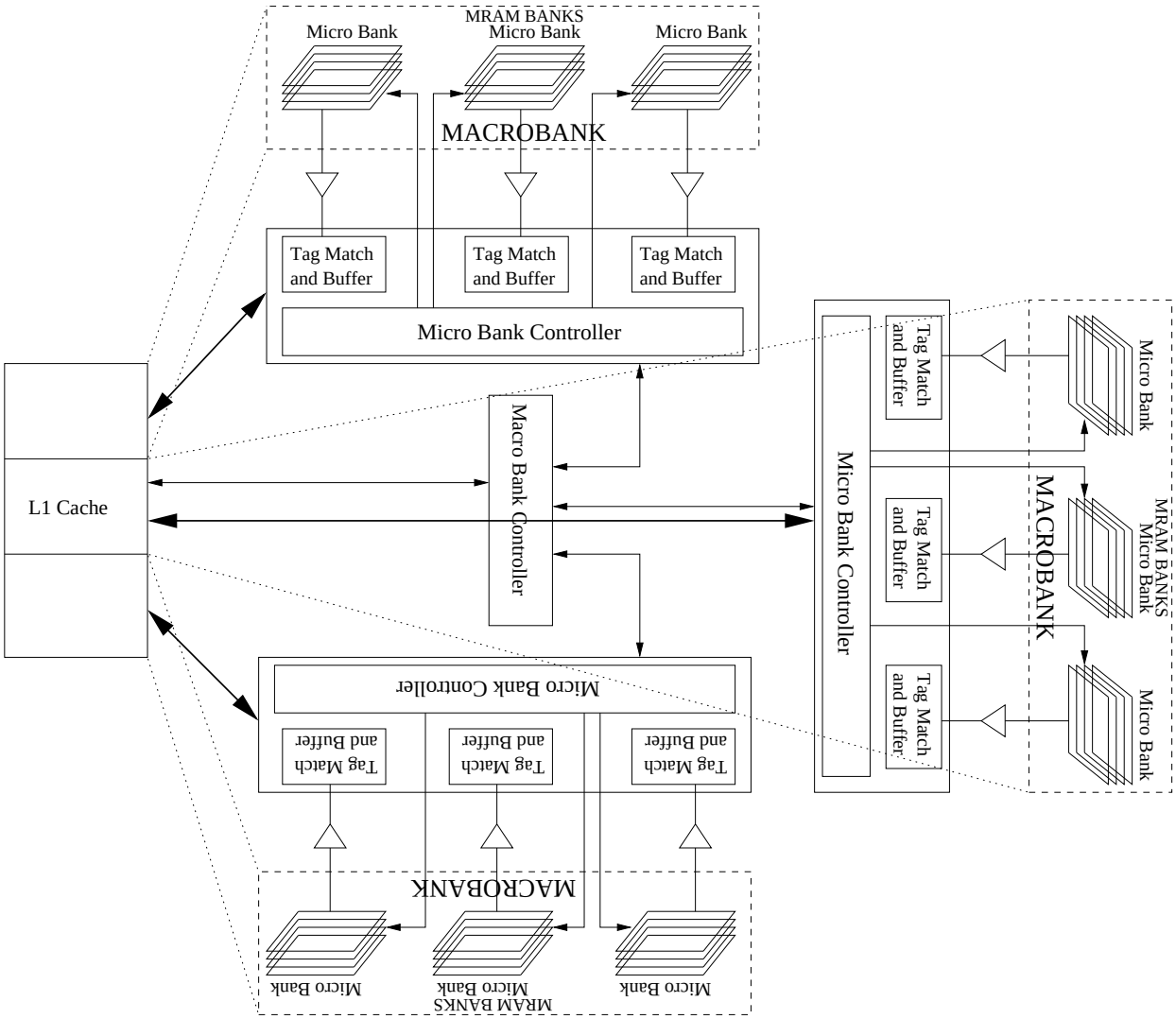

Figure 7: Architecture Model 2

come back from the MRAM layer, tag matching is performed and the requested data is sent to the level 1 cache on a hit. The data is also cached in the buffer to service future references to adjacent blocks.

In this model, it would might be good to have direct mapped caches to avoid reading a lot of data from MRAM memory and later discarding it. Also, we expect MRAM caches to have very large capacity, and hence associativity may not influence the performance very much.

The main factors to consider while evaluating the two models are the following:

1. Area: The tag area overhead in the first model might be significant enough to make the model infeasible. Assuming 48 bit addresses, and 256 MB level 2 MRAM cache, the tags will occupy 4 % of the area.

2. Latency: The latency of the first scheme will likely be higher as the tags need to be checked first before sending the request to the MRAM banks.

3. Power: Power will play an important role is determining the feasibility of either scheme. The first scheme saves power by avoiding unnecessary reads on a cache miss. Also, in the case of cache with high associativity, we need to read only one block of the set which matches. However, the tags will be implemented using SRAM memory, and hence will resultage in leakage. The second does not waste power leaking as the tags are also implement using non-volatile MRAM memory. However, the data and the tags need to be read in parallel to ensure smaller latency, and this will result in higher power consumption on a miss.

6 Future Work

In the preceding sections, we have examined described MRAM memory and its potential applications. We have also described the low level area model developed for modeling MRAM memory. In this section, we describe our goals for the future.

6.1 Low Level Area Model

1. The area model presently considers only read and writes to have same access time. MRAM writes will take longer as the current requirement is more. This needs to be incorporated in the model.
2. The bitline delay is computed for a step input and a voltage reference. If the sense amplifiers instead use a current reference, appropriate modifications have to be made.
3. The wordline drivers should be cascaded to achieve minimum area delay.
4. The vertical interconnects are modeled as regular wires. This model might be too simplistic.
5. Bitline and wordline access is assumed to occur concurrently. It might be the case that bitline access might have to be delayed.
6. Area overhead of bitline current source needs to be considered.
7. Multiplexing to reduce number of sense amplifiers needs to be incorporated.
8. Area and timing overhead of output drivers needs to be modeled
9. The decoder sizes are incremented in address multiples of 3 like in CACTI. This should be changed to linear increase for accurate area estimates.

6.2 Architecture Model

The low level area model only gives the accesses times of one bank along with the number of banks that can be accomadated at a particular technology, and the active area overhead for the control circuitry. It does take into account the actual placement of the banks and the overhead of the routing signals in the active layer. A high level architecture model needs to be built which considers this along with the individual bank access times. The model should allow the user to configure the MRAM banks in different ways, and from the configuration compute the total access time for a particular address.

For evaluating MRAM systems, the high level model will have a uni or mutiprocessor model running benchmarks which exercise the MRAM memory system. The performance of the benchmarks will be used to determine the optimal MRAM configuration.

7 Conclusion

In this document we have compared the key technological features of MRAM to conventional SRAM and embedded DRAM, and articulated distinct issues in the design of MRAM based systems. We believe that the most important challenges are as follows:

- While MRAM compares favorably in term of anticipated bit density to SRAM and eDRAM, interconnects between the magnetic cells may reduce the inherent benefits of density.
- While the cross point MTJ MRAM architecture requires only one diode per cell [14], the sensing circuits will require more active devices like MOSFETs. The speed and bandwidth of MRAM memory will depend on where the sensing circuits can be placed and the wire latencies between the MRAM bits and these circuits.
- The power and delay associated with writing into the MRAM array will likely prevent it from being used as a purely random access memory.
- The non-volatility provides a unique opportunity for scientific computations, perhaps to help reduce standby power consumption and power currently spent on disk accesses.
- Novel circuit designs will be required to exploit MRAM memory organizations.

While MRAM could become another layer in a traditional memory hierarchy, its bandwidth and volatility attributes suggest that it may be used in new ways for high-performance and reliable computer systems. We believe that there are fruitful opportunities for using MRAM as a high volume, high bandwidth memory for local and persistent data, as a staging area for persistent data in out-of-core computations, and as a storage for persistent logs to enhance recovery times and reduce logging overheads. Examination of these and other techniques is ongoing.

References

[1] V. Agarwal, M. S. Hrishikesh, S. W. Keckler, and D. Burger. Clock rate versus IPC: The end of the road for conventional microarchitectures. In *Proceedings of the 27th Annual International Symposium on Computer Architecture*, pages 248–259, June 2000.

[2] V. Agarwal, S. W. Keckler, and D. Burger. The effect of technology scaling on microarchitectural structures. Technical Report TR2000-02, Department of Computer Sciences, University of Texas at Austin, Austin, TX, Aug. 2000.

[3] V. Agarwal, C. Kim, S. W. Keckler, and D. Burger. The effect of system-level interconnects on microarchitecture design and simulation. Submitted to IEEE Transactions on VLSI Systems, August 2001.

[4] F. Hamzaoglu et al. Dual-V_t SRAM cells with full-swing single ended bit line sensing for high-performance on-chip cache in 0.13μm technology generation. In *Proceedings of the 2000 International Symposium on Low Power Electronics and Design, ISLPED 2000*, pages 15–19, Jul 2000.

[5] H. Hanson, M. S. Hrishikesh, V. Agarwal, S. W. Keckler, and D. Burger. Static energy reduction techniques for microprocessor caches. In *2001 International Conference on Computer Design*, pages 276–283, Sep 2001.

[6] G. Hinton, D. Sager, M. Upton, D. Boggs, D. Carmean, A. Kyker, and P. Roussel. The microarchitecture of the pentium 4 processor. *Intel Technology Journal*, 1, February 2001.

[7] K. Itoh, T. Watanabe, S. Kimura, and T. Sakata. Reviews and prospects of high-density RAM technology. In *Semiconductor Conference, 2000. CAS 2000 Proceedings. International*, volume 1, pages 13–22, June 2000.

[8] S. Kaxiras, Z. Hu, and M. Martonosi. Cache decay: Exploiting generational behavior to reduce cache leakage power. In *Proceedings of the 28th Annual International Symposium on Computer Architecture, ISCA 2001*, pages 240–251, Jun 2001.

[9] T. Kirihata et al. A 113mm^2 600Mb/sec/pin 512Mb DDR2 SDRAM with vertically folded bitline architecture. In *2001 IEEE International Solid-State Circuits Conference*, pages 382–383, 468, Feb 2001.

[10] P. K. Naji, M. Durlam, S. Tehrani, J. Calder, and M. F. DeHerrera. A 256kb 3.0v 1T1MTJ nonvolatile magnetoresistive RAM. In *2001 IEEE International Solid-State Circuits Conference*, pages 122–123, 438, Feb 2001.

[11] H. Pilo et al. An 833mhz 1.5w 18mb CMOS SRAM with 1.67Gb/s/pin. In *2000 IEEE International Solid-State Circuits Conference*, pages 266–267, Feb 2000.

[12] P. Ramm, D. Bonfert, H. Gieser, J. Haufe, F. Iberl, A. Klummp, A. Kux, and R. Wieland. Interchip via technology for vertical system integration. In *Proceedings of the International Interconnect Technology Conference*, pages 160 – 162, June 2001.

[13] G. Reinman and N. Jouppi. Extensions to CACTI, 1999. Unpublished document.

[14] R. E. Scheuerlein. Magneto-resistive IC memory limitations and architecture implications. In *Proceedings of the International NonVolatile Memory Technology Conference*, pages 47–50, May 1998.

[15] R. E. Scheuerlein, W. Gallagher, S. Parkin, A. Lee, S. Ray, R. Robertazzi, and W. Reohr. A 10ns read and write non-volatile memory array using a magnetic tunnel junction and FET switch in each cell. In *2000 IEEE International Solid-State Circuits Conference*, pages 128–129, Feb 2000.

[16] O. Takahashi et al. 1-GHz fully pipelined 3.7-ns address access time 8k x 1024 embedded synchronous DRAM macro. *International Journal of Solid-State Circuits*, 35:1673–1679, Nov 2000.

[17] S. Tehrani, B. Engel, J. M. Slaughter, E. Chen, M. DeHerrera, M. Durlam, P. Naji, R. Whig, J. Jenesky, and J. Calder. Recent developments in magnetic tunnel junction MRAM. In *IEEE Transactions on Magnetics*, volume 36, pages 2752–2757, Sep 2000.

[18] M. Wu and W. Zwaenepoel. eNVy: A non-volatile, main memory storage system. In *Proceedings of the 6th Symposium on Architectural Support for Programming Languages and Operating Systems*, pages 86–97, Oct 1994.

[19] K. Yamada, N. Sakai, Y. Ishizuka, and K. Mameno. A novel sensing scheme for a MRAM with a 5% MR ratio. In *2001 Symposium on VLSI Circuits Digest of Technical Papers*, pages 123–124, Mar 2001.

[20] H. Yoon et al. A 4Gb DDR SDRAM with gain-controlled pre-sensing and reference bitline calibration schemes in the twisted open bitline architecture. In *2001 IEEE International Solid-State Circuits Conference*, pages 378–379, 467, Feb 2001.

[21] R. Zhang, K. Roy, C.-K. Koh, and D. B. Jones. Exploring SOI device structures and interconnect architectures for 3-dimensional integration. In *Proceedings of the Design Automation Conference*, pages 846 – 851, Jun 2001.