The Pennsylvania State University

The Graduate School

College of Engineering

3D NUCA CACHE ARCHITECTURE ANALYSIS

A Thesis in

Computer Science and Engineering

by

Yogitha Puttasiddaiah

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2008

The thesis of Yogitha Puttasiddaiah was reviewed and approved* by the following:

Yuan Xie
Assistant Professor of Computer Science and Engineering
Thesis Advisor

Sencun Zhu
Assistant Professor of Computer Science and Engineering

Raj Acharya
Professor of Computer Science and Engineering
Head of the Department of Computer Science and Engineering

*Signatures are on file in the Graduate School

Abstract

It's a well known fact that memory performance is not keeping up with processor performance. Where Processor performance is increasing at roughly 60% every year, memory performance is less than 10% [3]. Shrinking feature sizes further worsen the effect of interconnect delay making it a critical obstacle in improving the memory access time. The widening gap between processor performance and memory has forced today's researchers to look new avenues to counter the problem. The first and most successful step to alleviate this gap is three-dimensional designs. Memory can be stacked directly on top of a microprocessor through 3D integration resulting in significant reduction in wire delay between the processor and the memory. Studies have shown significant performance, power and area benefits of such an approach compared to the conventional two-dimensional designs. Another approach to attack memory access time is to adopt on-chip network-based communication, the concept of Non-Uniform Cache Architectures (NUCA). NUCA architecture divides memory space into multiple banks which have different access latencies depending on their location relative to the processor. NUCA employs a packet-switched network between banks thus the access times are a function of where the data blocks are found.

The first contribution of this work is we implement two major extensions to the CACTI 6.0 cache modeling tool. First, we add the ability to model three-dimensional cache. Second, we add the ability to model MRAM and PRAM memory technologies. The

second contribution is a detailed comparison of the four different mainstream memory technologies, SRAM, DRAM, MRAM and PRAM in terms of performance, power and area in an architecture that combines the benefits of 3D and NUCA. The work also highlights that the 3D NUCA L2 architecture generates much better results than the conventional two-dimensional (2D) designs.

Table of Contents

Chapter 4 Results

List of Figures

List of Tables

List of Abbreviations

NUCA Non Uniform Cache Access

3D Three Dimension

SRAM Static Random Access Memory

DRAM Dynamic Random Access Memory

MRAM Magnetoresistive Random Access Memory

PRAM Phase Change Random Access Memory

UCA Uniform Cache Access

Acknowledgements

I would like to express my deep and sincere gratitude to my advisor, Dr. Yuan Xie. His wide knowledge, logical way of thinking and pro student outlook has been of great value for me. I cannot thank him enough for being patient with me during testing times. Throughout my thesis work, he provided encouragement, sound advice, good teaching and lot of good ideas.

I take this opportunity to thank my colleague Guangyu Sun for guiding me through the initial stages of the research, would have been lost without him. I thank Xiangyu Dong and Aditya Yanamandra for their invaluable inputs.

I owe a colossal debt of gratitude to my husband whose numerous sacrifices and constant encouragement made my dream of going to grad school possible. My deepest thanks to my family for their unflagging love and support throughout my life; I am indebted to my father, for his care and love. Although he is no longer with us, he is forever remembered.

Chapter 1

Introduction

This chapter provides the forward to the thesis report. To its ending, it sketches out, in Section 1.1 and 1.2, what motivated the research presented in this report and what the research was wished to achieve. Section 1.3 rounds the chapter off by outlining the arrangement of the report.

1.1 Motivation

With technology scaling, increased interconnect cost make it crucial to think about better ways of develop integrated circuits [2].Three-dimensional integration makes it possible to stack memory directly on top of a microprocessor [7]. Previous studies have studied and confirmed the performance benefits of such an approach [4]. Three-dimensional die stacking has received huge interest in the area of computer architecture [4]. With 3D integration it is possible to build circuits using multiple layers of active silicon with low-latency, high-bandwidth and very dense vertical interconnects [4]. Stacking memory directly on top of a processor is a natural way to attack the Memory Wall problem [8]. Multi-core processors will include huge and intricate cache hierarchies, they are expected to raise the size of both L2 and L3 caches [1]. In 65nm technology, up to 77% of the delay will be accounted to the interconnect [1]. Conventional architectures assume that monolithic memory has a single, uniform access

time, thus any increase in cache size in turn increases the access time. To overcome this hurdle the concept of NUCA is very promising. Memory space in NUCA is divided into multiple banks, which have different access latencies according to their locations with respect to the processor. The banks are connected through mesh-based packed switched network [5]. The benefits of combining 3D stacking and NUCA includes higher packing density, smaller footprint, improved performance due to reduced average interconnect length, lower power consumption and reduced average memory access time. The down side of excessive stacking is increase in junction temperatures especially in the inner layers of the die. These increased temperatures can render some of the above benefits of 3D integration worthless which highlights the requirement of thermal aware designs. This work we studies and compares, in detail the extent of benefits the above approach has on each of today's most used memory technologies, SRAM, DRAM, MRAM and PRAM with respect to area, power, performance and thermal issues in L2 cache memory.

1.2 Goals

The overall goal of the research presented through this work is to study the benefits and drawbacks of 3D, NUCA architecture compared to the traditional 2D, UCA architecture and analyze how the current memory technologies fare in this new architecture in terms of area, power and performance. The finer details of the aims of this work is listed below as follows

- Demonstrate the advantages of 3D die stacking in-terms of reduction in area, power consumption, wire length and cycles per memory access.
- Study the extent of the affect of rise in peak temperatures due to 3D die stacking
- Determine the improvement in cache latency achieved with NUCA
- Measure and compare the percentage of improvement in terms of area consumption of the four above mentioned memory technologies
- Measure and compare the reduction in dynamic and leakage power consumption for the four technologies
- Measure and compare the improvement in the performance, i.e. time per access for the four technologies

1.3 Report Structure

The remainder of this report can be pictured to constitute the following sections:

Chapter 2 briefly lists the background required for the chapters that follow. It familiarizes with the concept of 3D, NUCA architecture and the four memory cell technologies, SRAM, DRAM, MRAM and PRAM.

Chapter 3 introduces the experiment setup and model description.

Chapter 4 depicts the results of the all the experiments conducted. The results present the detailed analysis of the comparison of 3D, NUCA architecture, with conventional 2D designs and studies the performance of the four memory technologies in terms of area, power and performance.

Finally, chapter 5 comments and concludes the results of the above sections and discusses the potential opportunities for future work.

Chapter 2

Background

2.1 NUCA Architectures

In NUCA architectures a large cache is normally partitioned into many smaller banks. An inter-bank network is responsible for communicating addresses and data between banks and the processor. The latest large NUCA caches implement packet-switched on-chip grid network. The latency for a bank is determined by the delay to route the request and response between the particular bank that contains the data and the processor [5]. Each bank is associated with a router. The average delay for a cache access is calculated by counting the number of network hops to each bank, the wire delay attached on each hop, and the cache access delay within each bank [5].

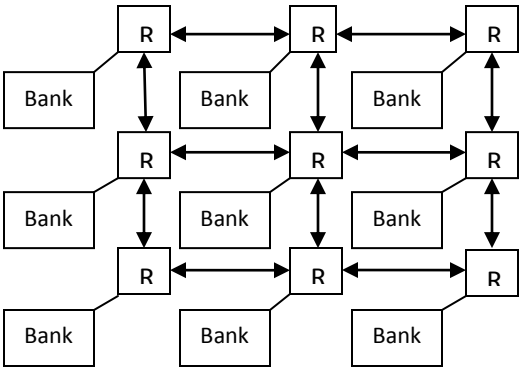

Figure 2.1: NUCA mesh [1]

Higher the number of partitions smaller is the delay and power within each bank, but would result in greater delays and power on the network due to overheads associated

with each router and decoder. At the same time increased partitions results in more routers on the network thus reducing the possibility of two packets conflicting at a router [5].

2.2 3D Design

Too many partitions, i.e. number of banks on a two dimensional plane does not prove to be beneficial. The large chip area necessitates the use of higher number of routers. A cache access request to a far-away bank would have to pass through a large number of routers increasing the delay substantially. In order to optimize NUCA's performance it is very vital to limit the number of routers the data traverses between source and destination [1]. The optimal solution will be to employ 3D stacking of multiple device layers with direct vertical interconnects tunneling through them thus reducing the number of hops between source and destination.

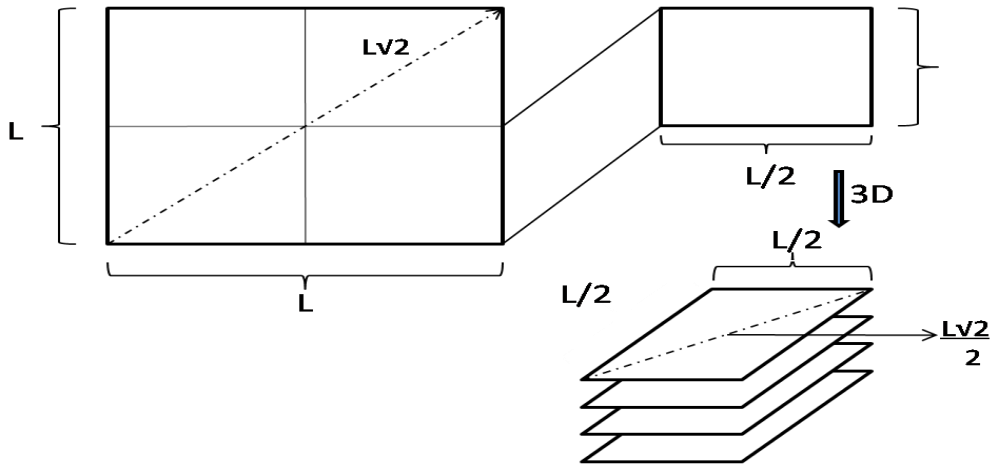

Figure 2.2: 3D designs reduce the wire length by a factor of the square of the number of layers used [1][11].

Joyner et al. [9] have shown that wire length reduces by a factor of the square root of the number of layers in 3D architectures. Some other benefits of 3D die stacking are

- Smaller footprint – less board space
- More transistors in a package – bigger designs
- Faster designs due to reduction in wire capacitance
- Less dynamic power due to reduced wire capacitance

2.3 SRAM

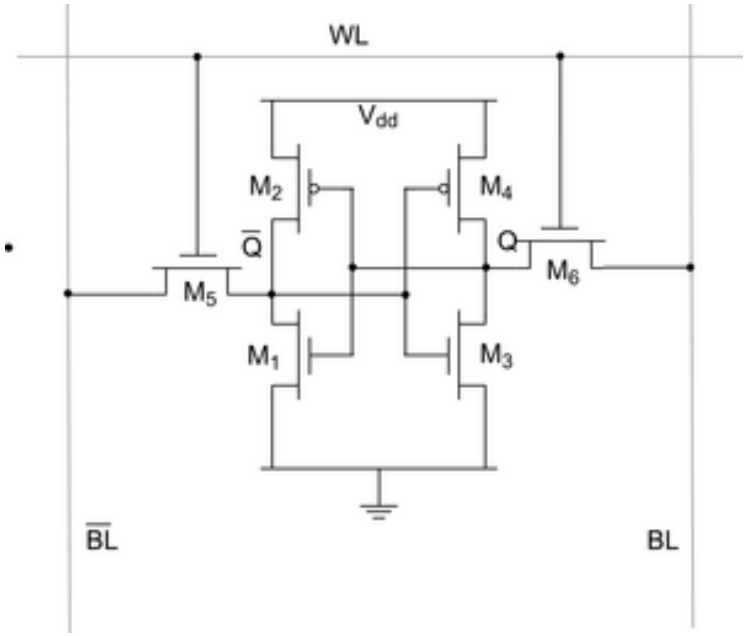

Figure 2.3: SRAM Cell[12]

The Static Random Access Memory (SRAM) uses six transistors to store each memory bit. As the word *static* signifies, it does not need to be refreshed periodically as it uses

bistable latching circuitry to store bits. It is volatile in nature thus data is lost when the memory is not powered [9]. Four transistors form two cross-coupled inverters which stores each bit in the SRAM. The two access transistors control access to the cell for read and write functions [9]. SRAM is fast but expensive. Since it uses six transistors per cell it is less dense compared to other memory cell technologies.

Assuming that current content of the memory is logic high, at Q. The reading operation begins by pre-charging the bit lines to logic high, setting the word line and enabling the access transistors. Next, the BL is left at pre-charged value and BL' is discharged through M1 and M5 to logic low, thus the values in Q and Q' are pushed to the bit lines. Transistors M4 and M5 pull the bit line to logic high [9]. Writing operation begins with the value to be written being applied to the bit lines, and then asserting the word line. Since the bit line input drivers are stronger compared to the transistor in the cell, they can override the state of the cross-coupled inverters thus latching in the value [9].

2.4 DRAM

Dynamic Random Access Memory (DRAM) stores each bit of data in a separate capacitor. It needs to be periodically refreshed to keep the capacitors from discharging and loosing the information stored. DRAM is simple in structure unlike SRAM and consists of only one transistor per cell making it dense. It is volatile, since it loses its data when the power supply is removed [9].

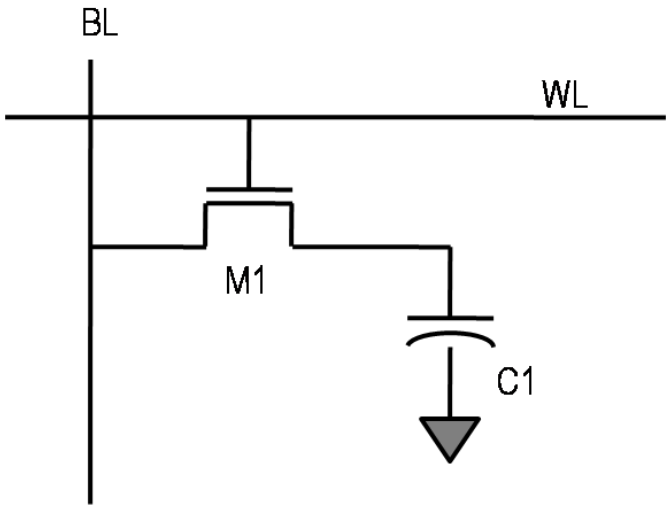

Figure 2.4: DRAM Cell

The basic operational concept of a DRAM is simple compared to that of SRAM. During a write cycle, the data value is placed on the bit line and the word line is raised. The cell capacitor either charges or discharges depending on the data value being written. For a read operation, the bit line is pre-charged first. On asserting the word line, a charge redistribution takes place between the bit line and the storage capacitance resulting in a voltage change on the bit line. The direction of which determines the value of the data stored [10](text book).

2.5 MRAM

In a Magnetoresistive Random Access Memory (MRAM), data is stored by magnetic storage elements and not as electric charge or current flows. Two ferromagnetic plates form the storage elements. Each of these plates can holds a magnetic field separated by a thin insulating layer [10]. The direction of one of the plate is fixed, called the reference. The direction of the plate

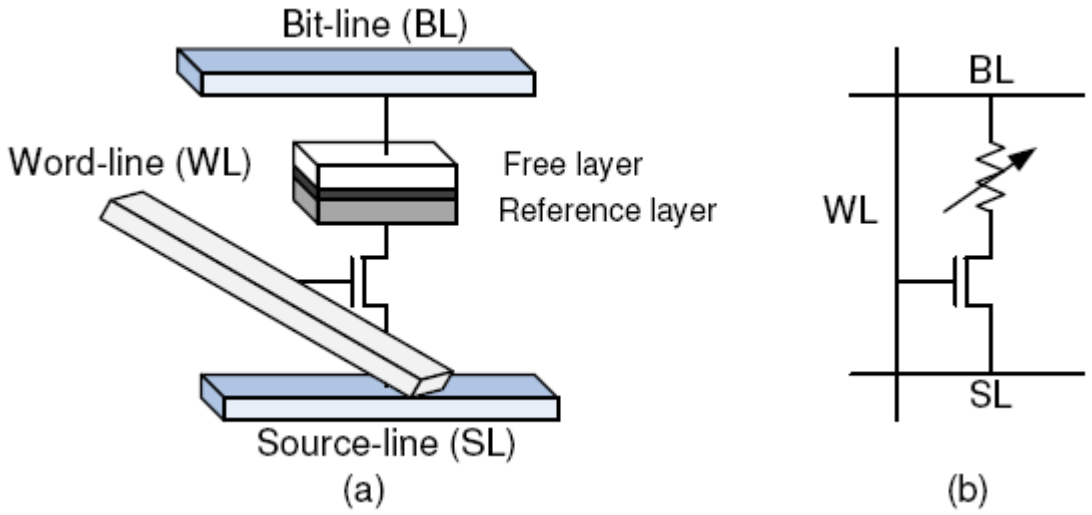

Figure 2.5: MRAM Cell. (a) Structural View. (b) Schematic view [7].

will change depending on the driving current [7]. MRAM is dense, non-volatile and the leakage power consumption is very low. The performance of MRAM falls between SRAM and DRAM.

A very small negative voltage difference is applied on the bit line with respect to the source line for a read operation. This will cause a small current to pass through the magnetic element. The current is compared to a reference to decide if a logic high or logic low is stored in the memory cell [7]. A positive voltage difference is applied between the source line and bit line for writing a logic '0' and vice versa for writing a logic '1'. The amplitude of the current and the duration of the writing pulse depends on the size of the magnetic element [7].

2.6 PRAM

Phase Change Random Access Memory (PRAM) uses a complex chalcogenide alloy as its memory element which acts as a programmable resistor [11]. The alloy exists in a stable

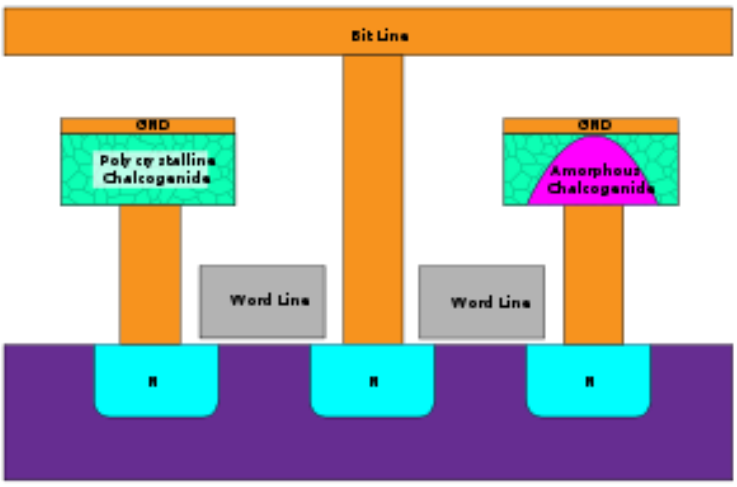

Figure 2.6: Cross section of two PRAM Cells one in low resistance crystalline state and, other in high resistance amorphous state [12].

polycrystalline phase in its natural state which has low electrical resistance. There is an other state this alloy can exit, a meta-stable high resistance amorphous phase [11]. The low electric resistance state is used to store a logic high, and a high resistance state to store a logic low. PRAM is a very dense, non-volatile, fast, low power memory technology.

Writing logic '0' involves current pulse with strength enough to melt a small volume of the cell. This molten region is quenched quickly into an amorphous state of high resistance (logic low state) [11]. Logic '1' programming uses a pulse smaller amplitude and longer width. A low voltage pulse is used to detect the cell response for a read operation.

2.7 Basic Comparison of the SRAM, DRAM, MRAM and PRAM

Table 2.1 lists the properties of SRAM, DRAM, MRAM and PRAM in terms of Density, Power consumption, and speed on a conventional two-dimensional, Monolithic memory design.

		SRAM	DRAM	MRAM	PRAM	
	Density	Low	High	High	High	
	Dynamic Power	Low	Medium	High	Very Low	
	Leakage Power	High	Medium	Very Low	Very Low	
	Speed	Fast	Slow	Fast Read speed, Very slow write speed	Medium	
	Non-Volatility	No	No	Yes	Yes	

Table 2.1: Comparison [7]

Chapter 3

Analytical Setup

3.1 Model 3D NUCA

This chapter outlines the experimental setup and the approach adopted to analyze and compare how SRAM, DRAM, MRAM and PRAM memory technologies fare in a 3D, NUCA architecture in terms of area, power and performance.

As cache size increases, the interconnect delay starts to dominate in terms of access time and power. Experiments have proven that 3D technology can boost the access time and bring down the power consumption by reducing the global wire length. CACTI cache access modeling tool is a widely used tool for cache evaluation to determine the area, power and performance for given cache configuration. While prior versions of CACTI have supported 3D cache modeling, they have only supported uniform cache access architecture (UCA). This work extends CACTI tool version 6.0 to model 3D cache with NUCA architecture. The enhancement also enables the tool to model MRAM and PRAM memory technologies. The modified CACTI 6.0 takes the following parameters as input: cache size, memory type, number of layers, cache block size, cache associativity and technology size, and outputs the cache configuration optimized for area, power and performance.

In the extended CACTI 6.0, Cacti 6.0 is first employed to iterate over a number of bank organization to find the optimum number of banks for a user input cache size and memory type. Once the number of optimum number of banks is found they are then divided between the numbers of layers specified by the user. A router is associated with each bank and the average cache access delay is calculated by counting the number of network hops to each bank, wire delay for each hop (horizontal, vertical and through hop), router delay and the access delay with each bank. The access delay within each bank is determined by employing CACTI 5.0.

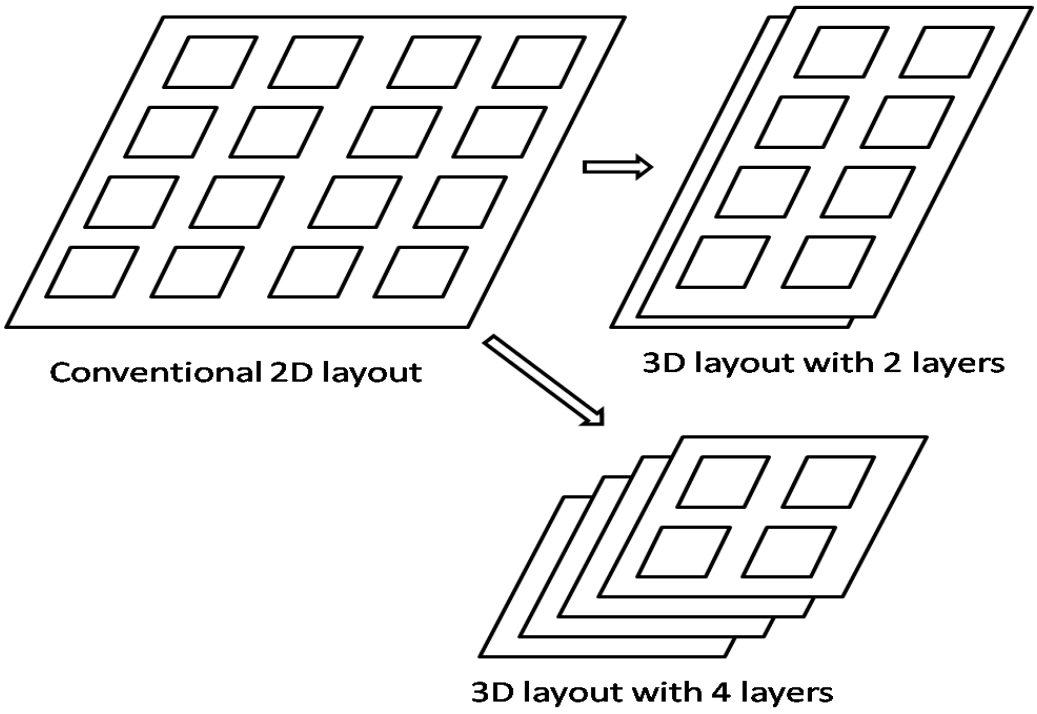

Figure 3.1: 3D NUCA

Modeling MRAM and PRAM cache

Area Model

MRAM, PRAM and SRAM have very similar electrical interfaces from perspective of the peripheral circuitry [7][11]. All the three are similar in the sense they have word lines to pick the destined memory cell and bit lines to transfer data. The major difference with between SRAM and the other two technologies is that, SRAM has two differential bit lines while bit lines for MRAM and PRAM are single ended. Due to this reason MRAM and PRAM cannot use SRAM-based sense amplifier directly, as they need reference signals for their sense amplifiers [7][11]. But as the cache size grows the overhead of the reference signal is negligible. Due to the organizational likeness of MRAM and PRAM to SRAM, merely changing all the SRAM related parameters, will modify CACTI to support MRAM and PRAM.

Timing Model

The cache is divided into the following components, H-tree input, decoder, word line, bit line, sense amplifier, comparator and H-tree output. The read access time is calculated by adding the delays of these individual components. The original delays of H-tree and decoders of the CACTI is retained, the bit line delays for MRAM and PRAM are obtained from HSPICE model [7] and PCM model [11] respectively. Since the reference signal potentially increases the sensing delay we add an additional amplifier delay constant of around 20% to the original SRAM sense amplifier delay in CACTI to adopt it to MRAM

and PRAM requirement [7]. Similarly write access time is calculated by summing up the delays of the following components, H-tree input, decoder, word line and writing pulse.

Energy Model

Dynamic Energy

The dynamic energy calculation is carried out by replacing the power estimation part of CACTI with the values obtained by MRAM HSPICE model and PRAM PCM model

Leakage Energy

Due to the non-volatile nature of MRAM and PRAM cells, the standby leakage power consumption is zero. We still need to take into account the active leakage power consumed by the MRAM and PRAM cells and the leakage in the peripheral circuitry.

Chapter 4

Results

This section compares SRAM, DRAM, MRAM and PRAM in terms of density, power consumption and speed. All caches simulated are 16-way associative, 64-byte block, L2 caches at 32nm technology.

4.1 Area and Density

The cost of memory system is directly proportional to the density of the memory media. A SRAM cell is consists of 6 transistors against 1 transistor in both in DRAM, MRAM and PRAM. The MRAM cell is about 1.7 times that of a DRAM cell at 90nm technology. But with further scaling and extra peripheral circuits needed for DRAM begins to dominate thus reversing the difference. For all the four technologies approximately 50% reduction in area is observed with doubling the number of layers.

Table 4.1 lists the area comparison of 32MB cache with 8MB banks.

	Cache	1-layer	2-layer	4-layer	
	32MB SRAM	67.74	33.87	20.02	
	32MB DRAM	23.93	14.17	7.11	
	32MB MRAM	15.32	7.94	4.22	
	32MB PRAM	5.38	2.69	1.35	

Table 4.1: Area Comparison

Table 4.2 compares the packing densities of the four technologies, PRAM has the maximum packing density

	Cache	SRAM	DRAM	MRAM	PRAM	
	22mm2	8	30	51	98	
	68mm2	32	91	128	250	
	244mm2	128	293	512	710	

Table 4.2: Density Comparison

4.2 Power Consumption

4.2.1 Dynamic Energy

Energy consumption per operation for 32MB cache, with 4 banks of 8MB each is listed in Table 4.3. A shift from one layer to two layers shows a reduction in dynamic energy for read operation by 17% in SRAM and PRAM, 20% in DRAM and 30% in MRAM. A shift from two layers to four layers further reduces the energy consumption for SRAM by 9%, DRAM by 15%, MRAM by 20% and PRAM by 13%.

	Cache	1-Layer	2-Layer	4-Layer	
	32MB SRAM	0.134	0.110	0.100	

	32MB DRAM	0.403	0.325	0.265	
	32MB MRAM	Read 0.214, Write 4.280	Read 0.148, Write 2.970	Read 0.118, Write 2.368	
	32MB PRAM	0.116	0.096	0.083	

Table 4.3: Dynamic Energy comparison

4.2.1 Leakage Power

Table 4.4 list the leakage power consumption for the three technologies. For uniform evaluation power per unit area is used in the comparison. Due to the non-volatile property of MRAM and PARM cell the standby leakage is very less compared to the other two technologies. The low leakage power consumption in MRAM more than compensates for the high power requirement for write operation. The high leakage power requirement for SRAM limits its usage to smaller caches.

	Cache	**Leakage Power per mm2**	
	32MB SRAM	41.028mW	
	32MB DRAM	11.72mW	
	32MB MRAM	9.32mW	
	32MB PRAM	6.32mW	

Table 4.4: Leakage Power Comparison

4.3 Thermals

In 3D stacking designs thermal issues play a crucial role. Die stacking can increase the power density significantly when highly active regions are stacked on top of each other [6]. The other major reason for high temperatures is that the distance between every new layer and the heat sink is always increasing [6]. The thermal model Hotspot 4.1 was used to for the 3D thermal analysis.

Figure 5 shows the peak temperatures for all the configurations. Stacking SRAM results in the highest thermal increase due, to higher power density of SRAM when compared to DRAM, MRAM and PRAM.

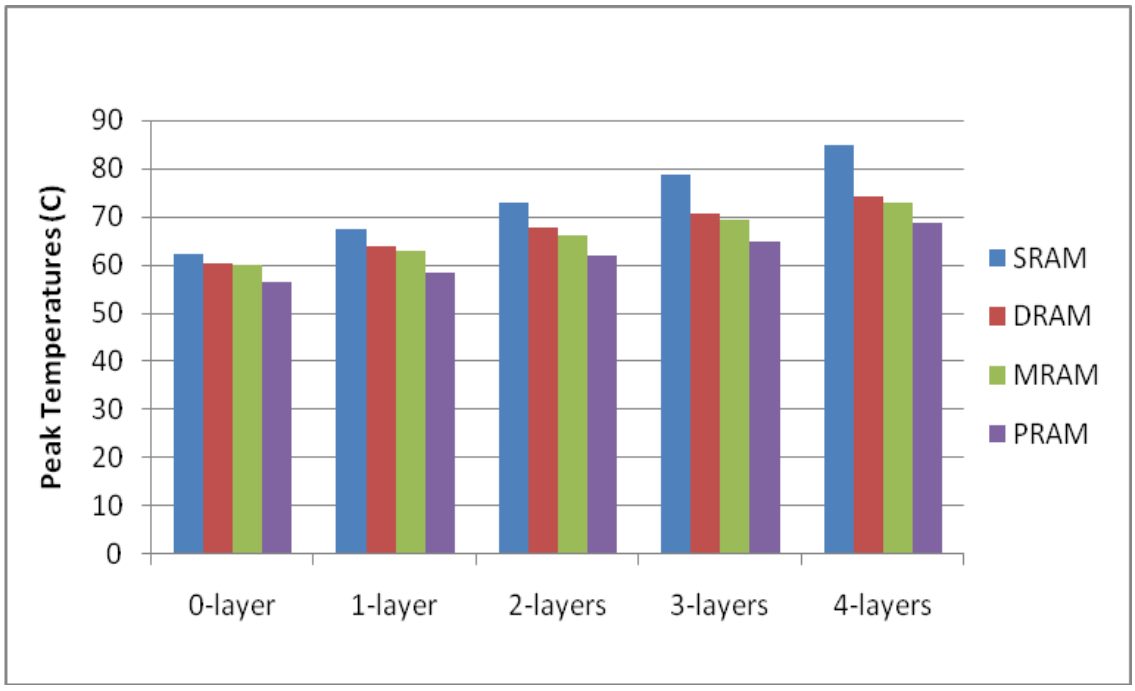

Figure 4.1 Peak Temperature Results for Stacked Cache

4.4 Speed

Table 4.5 shows the results of read latency of SRAM, DRAM, MRAM and PRAM for different stack configurations. Caches with similar areas are chosen to make fair comparison. The results show that SRAM and MRAM are very comparable in terms of read latency but DRAM does not lag far behind. With each increase in layer the latency reduces by a factor of 20 to 25% for SRAM and MRAM and 15 to 20% for DRAM and PRAM

	Cache	1-layer (ns)	2-layer (ns)	4-layer (ns)	
	8MB SRAM	2.01	1.61	1.21	
	32MB DRAM	2.60	2.20	1.80	
	64MB MRAM	3.70	3.4	3.1	
	128MB PRAM	2.42	2.10	1.72	

Table 4.5: Read Latency Comparison (with bus)

The above results are for an architecture where each CPU has a direct bus to each layer. The distance between the core layer and any memory layer is just one hop. Without the bus it would take the CPU one hop to reach layer-1, two hops to reach layer-2 and so on. The access latency results for such a design are listed in Table 4.6. With additional hops to reach each layer the benefit of 3D stacking is not very attractive. With a just a

little extra effort of adding buses from CPUs to layers better results in terms of access time can be obtained.

	Cache	1-layer (ns)	2-layer (ns)	4-layer (ns)	
	8MB SRAM	2.01	1.81	1.76	
	32MB DRAM	2.60	2.40	2.34	
	64MB MRAM	3.70	3.55	3.40	
	128MB PRAM	2.42	2.26	2.02	

Table 4.6: Read latency without bus

Increasing the number of partitions (banks) results in smaller delays and power within each bank, but would result in greater delays and power on the network due to overheads associated with each router and decoder. At the same time increased partitions results in more routers on the network reducing the possibility of two packets conflicting at a router, thus making the design better capable of meeting the high bandwidth requirement of a multi-core system [5]. It is vital to strike a balance to obtain optimal results. Table 4.7 shows the access time variation with different bank counts. Initially the access time reduces with the increase in the number of partitions (smaller bank sizes), but as the partitions become higher the network overhead begins to dominate, thus increasing the overall access time.

	64MB Cache	1-Bank (ns)	2-Bank (ns)	4-Bank (ns)	8-Bank(ns)	16-Bank (ns)	
	SRAM	5.65	4.88	4.64	3.92	4.32	
	DRAM	4.73	3.83	3.38	2.98	3.32	
	MRAM	2.74	2.48	2.03	2.46	2.50	
	PRAM	2.8	2.02	1.995	2.341	2.373	

Table 4.7: Latency Vs number of banks

Table 4.8 shows the read latency based on the physical distance between a bank and the processor.

	64MB Cache	Latency to closest bank(ns)	Latency to farthest bank(ns)	Average latency (ns)	
	SRAM	3.10	4.90	4.0	
	DRAM	1.80	4.20	3.0	
	MRAM	1.30	3.70	2.5	
	PRAM	1.24	3.44	2.34	

Table 4.8: Latency variations across banks

4.5 Performance

Figures 6,7,8, and 9 show the performance comparison for various layers of SRAM, DRAM, and MRAM in terms of number of operations every 3 million cycles. All the

simulations were performed on SIMICS

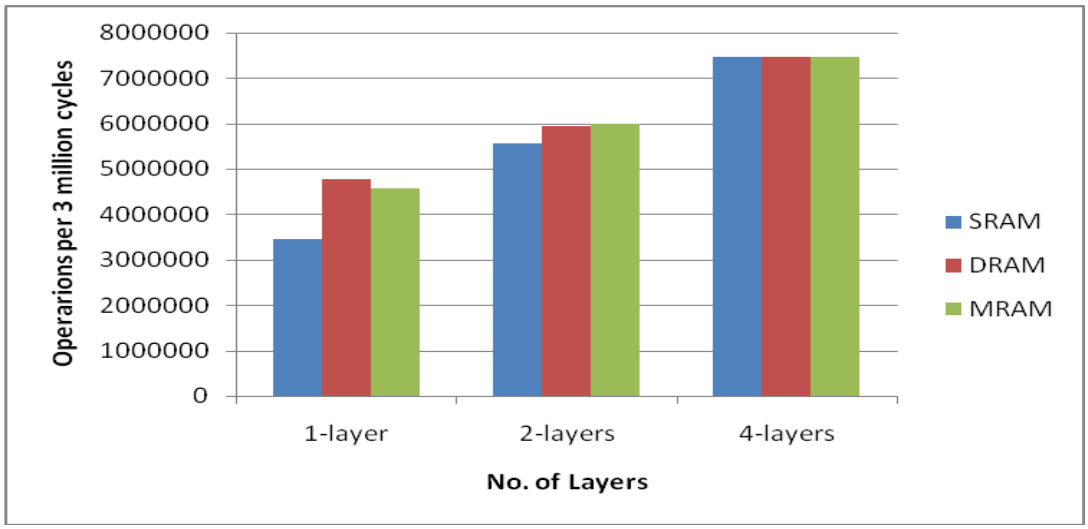

Fig 4.6: Performance -SWIM

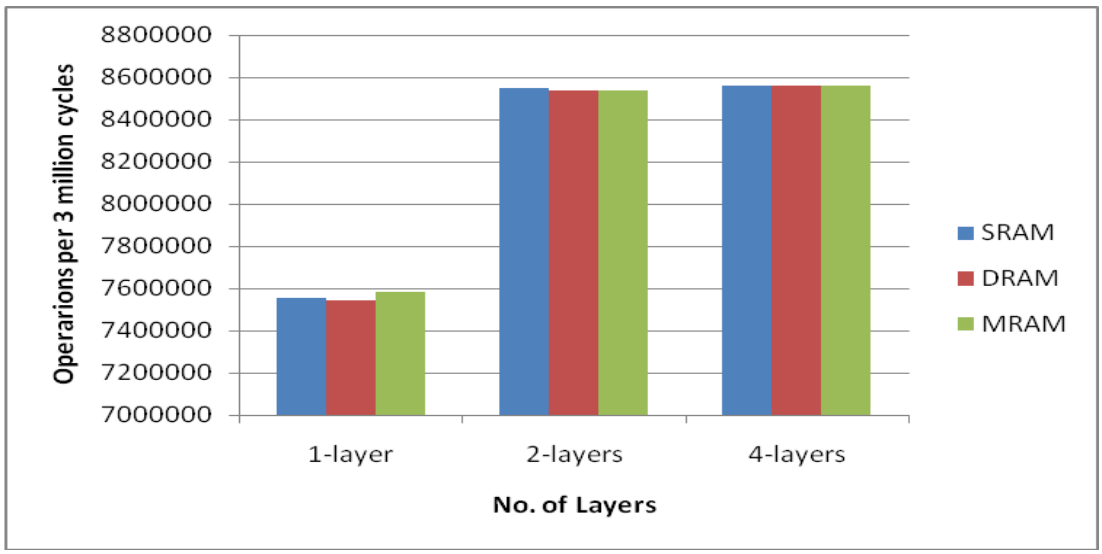

Fig 4.7: Performamce –APSI

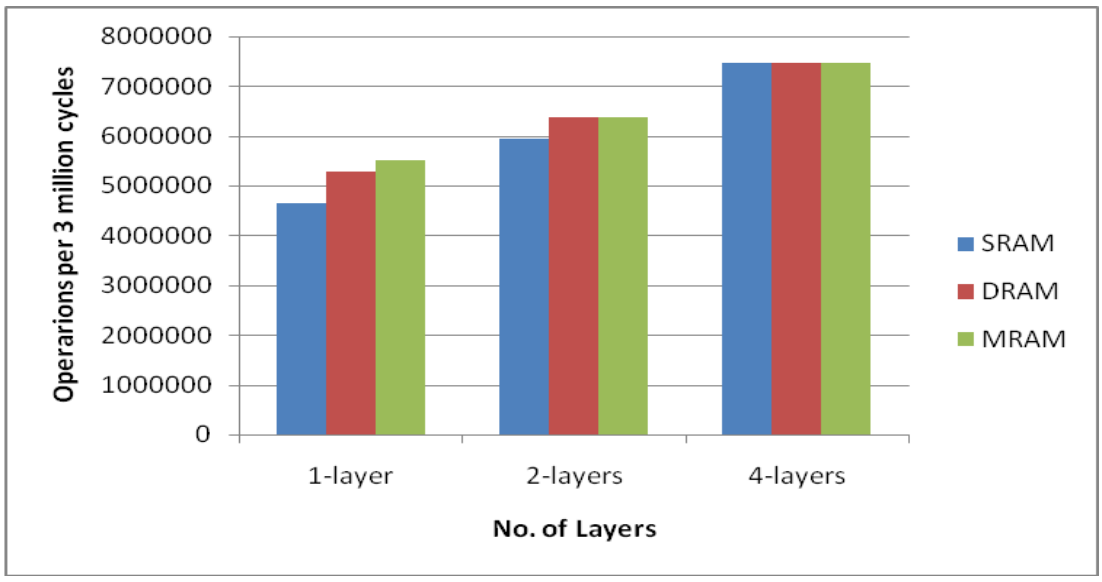

Fig 4.8: Performance - APPLU

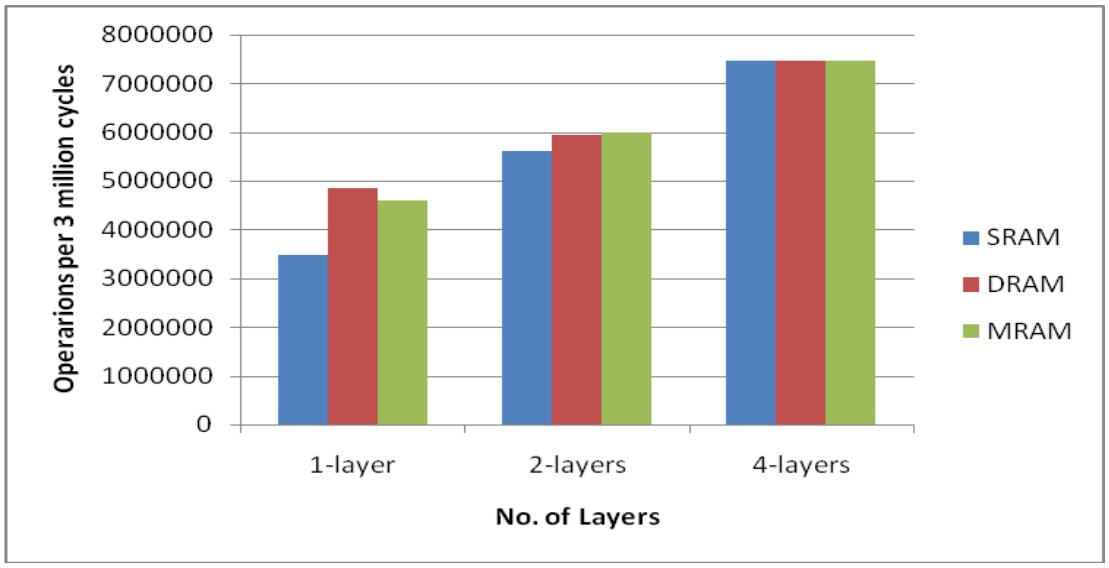

Fig 4.9: Performance - WUPWISE

The simulation results clearly show that with increased number of layers the performance of all the three technologies are very much comparable. But the tie breaker comes when area and power consumption is taken into picture.

Chapter 5

Conclusions

Previous research has proved that stacking memory directly on processors provide significant respite from Memory Wall Problem. This work explores and demonstrates the advantages and disadvantages of three dimensional die stacking with NUCA architecture. The first contribution of this work is we implement two major extensions to the CACTI 6.0 cache modeling tool. First, we add the ability to model three-dimensional cache. Second, we add the ability to model MRAM and PRAM memory technologies. The second contribution is a detailed comparison of the four different mainstream memory technologies, SRAM, DRAM, MRAM and PRAM in terms of performance, power and area in an architecture that combines the benefits of 3D and NUCA. The experiment results demonstrate that MRAM and PRAM are very promising to be universal memory replacement of the future fast, low power and denser cache architectures.

Bibliography

1. F. Li, C. Nicopoulos, T. Richardson, Y. Xie, N. Vijaykrishnan, and M. Kandemir. Design and Management of 3D Chip Multiprocessors Using Network-in Memory. In Proceedings of ISCA-33, June 2006

2. O. Ozturk, F. Wang, M. Kandemir, and Y. Xie. Optimal Topology Exploration for Application-Specific 3D Architectures

3. C. Liu, I. Ganusov, M. Burtscher, and S. Tiware. Bridging the Processor-Memory Performance Gap with 3D IC Technology

4. G. H. Loh. 3D stacked Memory Architectures for Multi-Core Processors

5. N. Muralimanohar, R. Balasubramonian, and N. Jouppi. Optimizing NUCA Organizations and Writing Alternatives for Large Caches With CACTI-6.0

6. B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. Loh, D. McCauley, P. Morrow, D. Nelson, D. Pantuso, P. Reed, Jeff Rupley, S. Shankar, J. Shen, and C. Webb. Die Stacking (3D) Microarchitecture

7. X. Dong, X. Wu, G. Sun, and Y. Xie. Circuit and Microarchitecture Evaluation of 3D Stacking Magnetic RAM (MRAM) as a Universal Memory Replacement

8. W. A. Wolf, and S, A, McKee. Hitting the Memory Wall: Implications of the obvious. Computer Architecture News, 23(1):20-24, March 1995.

9. J Joyner, P. Zarkesh-Ha, and J. Meindl. A stochastic global net-length distribution for a three-dimensional system-on-a chip.

10. *Digital Integrated Circuits,* Jan M. Rabaey, Anantha Chandrakasan, and Borivoje Nikolic

11. P Mangalagiri, A Yanamandra, Y Xie, N. Vijaykrishnan, M J Irwin, K Sarpatwari, O. O. Awadel Karim. A Low-Power Phase Change Memory Based Hybrid Cache Architecture

12. Wikipedia