WIDE I/O ARCHITECTURE UTILIZING PROXIMITY COMMUNICATION

by

Qawi IbnZayd Harvard

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Electrical Engineering

Boise State University

December 2009

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Qawi IbnZayd Harvard

Thesis Title: Wide I/O DRAM Architecture Utilizing Proximity Communication

Date of Final Oral Examination: 8 October 2009

The following individuals read and discussed the thesis submitted by student Qawi IbnZayd Harvard, and they also evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination, and that the thesis was satisfactory for a master's degree and ready for any final modifications that they explicitly required.

R. Jacob Baker, Ph.D.	Chair, Supervisory Committee
Sin Ming Loo, Ph.D.	Member, Supervisory Committee
Thad Welch, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by R. Jacob Baker, Ph.D., Chair of the Supervisory Committee. The thesis was approved for the Graduate College by John R. Pelton, Ph.D., Dean of the Graduate College.

ACKNOWLEDGMENTS

I would like to thank God for placing me on this path and giving me the ability to function at this level.

I'd also like to acknowledge Dr. Jake Baker for his support. Dr. Baker has been a dominant contributor to my success, and this thesis would not be possible without his support. I would like to thank Dr. Kris Campbell for the financial support that she provided. I would like to thank Dr. Robert Drost for his guidance during my summer internship, which led to the idea for this thesis.

I’d like to thank my thesis committee for thoroughly reviewing the research and this thesis. Thanks Dr. Sin Loo and Dr. Welch. I would like to thank Ms. Welch for her assistance in proof reading this thesis.

I would like to thank my family for their continued support and encouragement. I love you very much.

ABSTRACT

The bandwidth and power consumption of dynamic random access memory, used as the main memory of a computer system, impacts the computer's execution rate even with the existence of a memory hierarchy. DRAM manufacturers focus on density increases due to the innate price per bit decline of main memory while processor manufacturers continually focus on boosting performance by increasing the number of instructions completed per second. This leads to a performance gap between the microprocessor and DRAM.

Proximity communication promises to increase the I/O density of DRAM products while reducing the power consumption of the main memory system. This thesis develops and discusses the design of a memory system employing 4 Gb DRAM chips with a 64-bit wide communication bus using proximity communication. Technological roadblocks are analyzed and novel solutions are proposed. The proposed 4 Gb DRAM architecture can reduce the power consumption of a main memory system by 50% while increasing the bandwidth by 100%. The 4 Gb chip developed in this thesis measures 68.88 mm^2 and has an array efficiency of 59.9%. These estimates are comparable to the 2012 International Technology Roadmap for Semiconductors' estimates of 74 mm^2 and 56%, respectfully.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iv

ABSTRACT v

TABLE OF CONTENTS vi

LIST OF FIGURES ix

CHAPTER 1—INTRODUCTION 1

1.1 Organization 2

1.2 Contributions of This Thesis 3

CHAPTER 2—PROXIMITY COMMUNICATION 4

2.1 Advantages 5

2.1.1 I/O Density 6

2.1.2 Removal of Off-Chip Wires 7

2.1.3 Removal of ESD Structures 8

2.1.4 Removal of the On-Die Termination 8

2.1.5 Ease of Testability 9

2.2 Challenges 10

2.2.1 Electronic Sensors for Measuring Misalignment 10

2.2.2 Electronic Re-Alignment 14

2.3 Summary 17

CHAPTER 3—DRAM TRENDS 19

3.1 Memory Gap 19

3.2 The DRAM Market and Technology 22

3.2.1 Selling Price 22

3.2.2 Wordline Scaling 25

3.2.3 Bitline Scaling 27

3.2.4 Contact Resistance Scaling 28

3.3 DRAM Generations 28

3.3.1 Power 29

3.3.2 Bandwidth 32

3.3.3 Bus Loading 35

3.3 Summary 38

CHAPTER 4—A 4 Gb DRAM ARCHITECTURE 39

4.1 Creating a 256 kb Array 41

4.1.1 Memory Array 44

4.1.2 Periphery Circuitry 46

4.2 Creating a 256 Mb Array 49

4.3 Creating a 1 Gb Array 51

4.4 4 Gb DRAM Architecture 53

4.5 Summary 54

CHAPTER 5—A PROXIMITY COMMUNICATION DRAM ARCHITECTURE 55

5.1 Architecture Decision 55

5.1.1 Pad Moving and Centralization 56

5.1.2 512 Mb Bank Structures 62

5.2 Side Mount Architecture 63

5.3 Challenges 65

5.3.1 Number of Metal Layers and Global I/O Routing 65

5.3.2 Local I/O Routing 67

5.4 Slice Architecture 71

5.5 Summary 73

CHAPTER 6—CONCLUSIONS 77

BIBLIOGRAPHY 79

LIST OF FIGURES

Figure 2.1 Cross-sectional view of proximity communication 4

Figure 2.2 Transmit and receive circuit diagram 5

Figure 2.3 I/O density..... 6

Figure 2.4 Six degrees of misalignment..... 11

Figure 2.5 Electronic sensors 12

Figure 2.6 Vernier scale. 13

Figure 2.7 Simulation versus silicon data 14

Figure 2.8 Receiver and transmitter array..... 15

Figure 2.9 One dimensional steering..... 16

Figure 2.10 Transmit array and receiver array..... 17

Figure 3.1 Relative performance change..... 19

Figure 3.2 An alternative performance comparison..... 21

Figure 3.3 Historical price decline 23

Figure 3.4 DRAM spot prices 24

Figure 3.5 Wordline scaling..... 26

Figure 3.6 Wordline resistance..... 26

Figure 3.7 An evolutionary view of DRAM. 29

Figure 3.8 Server architecture 30

Figure 3.9 Current consumption..... 31

Figure 3.10 Array pre-fetch..... 33

Figure 3.11 DRAM pre-fetch and bandwidth evolution 34
Figure 3.12 ITRS DRAM bandwidth 35
Figure 3.13 Bandwidth of a memory channel 36
Figure 4.1 1 Gb DDR/DDR2 chip 39
Figure 4.2 512 Mb DDR2 chip 40
Figure 4.3 512 Mb DDR3 chip 40
Figure 4.4 4 Gb DDR3 chip 41
Figure 4.5 $6F^2$ memory cell 42
Figure 4.6 Developing a 256 kb memory array 43
Figure 4.7 Cross sectional view of DRAM 44
Figure 4.8 Determining the size of a 256 kb memory array 46
Figure 4.9 Schematic of a CMOS wordline driver 47
Figure 4.10 Schematic of the bitline sense amplifier 48
Figure 4.11 Creation of a 256 Mb bank 49
Figure 4.12 Global periphery circuitry and area allocation 50
Figure 4.13 Creation of a 1 Gb memory array 51
Figure 4.14 Expanded view of the 1 Gb memory bank 52
Figure 4.15 40nm 4 Gb DRAM chip 53
Figure 5.1 The 4 Gb DRAM. 56
Figure 5.2 Moving the communication channel 57
Figure 5.3 Centralized column and centralized row 58
Figure 5.4 Changing the architecture. 60
Figure 5.5 Initial proximity communication enabled DRAM architecture. 61

Figure 5.6 Physical size of the memory bank. 62
Figure 5.7 4 Gb DRAM w/ proximity communication. 64
Figure 5.8 Half-bank structure 66
Figure 5.9 16:1 separation of the 64k bitlines. 67
Figure 5.10 Local I/O routing within a half-bank. 68
Figure 5.11 New global I/O structure in the array column path 69
Figure 5.12 Page decode region 71
Figure 5.13 SLICE architecture 72
Figure 5.14 Control and Data SLICE blocks 73
Figure 5.15 Energy per bit and chip bandwidth 75

CHAPTER 1—INTRODUCTION

The performance gap between the computer's processor and its main memory has been growing over the past two decades. The major performance measurements of main memory manufacturers have remained density and die size over this time. Increasing these performance measurements places a physical limit on the latency of the main memory. More and more bits are placed into a fixed silicon area, which increases the associated parasitics. This physical limitation keeps the latency of the main memory scaling at roughly 7% over the past two decades, while processor performance has been scaling at roughly 50%. This differential is termed the “memory gap” and refers to the growing performance disparity between the processor core and its main memory.

Processor manufacturers have made several architecture changes that have enabled their performance to continue to scale with Moore's Law. Multiple computer cores, multiple instruction threads, increased size and levels of cache, along with speculative accessing, have made memory stalls almost transparent to the computer user. However, main memory scaling has continued on a limited trend. Main memory manufacturers increase their density per unit area by developing longer bitlines, longer wordlines, decreased unit cell size, and feature size scaling. Main memory manufacturers moved away from their bandwidth limitations by using DRAM pre-fetch and high speed input/output circuits. Unfortunately, the new pre-fetch architectures did not begin taking hold until 1995. This places memory bandwidth scaling decades behind processor bandwidth scaling.

Proximity communication is a new I/O technology that uses capacitors to electrically connect two chips. The off/on chip communication technique has the ability to substantially increase the main memory bandwidth and not impact the power consumption. Proximity communication may not reduce the access latency, which is a technology driven parameter, but it will increase the chip-to-chip bandwidth. This thesis develops a wide I/O DRAM architecture that uses proximity communication.

The architecture has the ability to substantially increase main memory bandwidth, while reducing the power consumption. This is achieved by allowing a single DRAM chip to provide a full cache line of memory (64 Bytes).

1.1 Organization

This thesis provides a feasibility study of a DRAM architecture that incorporates capacitive proximity communication to increase the I/O count. Chapter 1 provides a brief introduction and thesis outline. Chapter 2 discusses the innovations and performance measurements of capacitive proximity communication.

Developing a feasible DRAM architecture that incorporates proximity communication is only possible once the complexities of the DRAM market are understood. Chapter 3 reviews the DRAM market and ensures a proper understanding of why DRAM architectural decisions are made. Chapter 4 uses the information obtained to develop a 4 Gb DRAM architecture that meets forecasted approximations of a chip designed for production in 2012.

The information provided in Chapter 4 allows for the development of a 4 Gb DRAM architecture that incorporates proximity communication. The feasibility study

performed in Chapter 5 discusses architecture decisions made to remedy technological road blocks for a wide I/O DRAM architecture. Chapter 6 concludes this thesis, provides an overview of the major findings, and discusses future work.

1.2 Contributions of This Thesis

This thesis provides an in depth discussion of capacitive coupled proximity communication and its integration into a computer systems main memory. A 4 Gb DRAM architecture utilizing proximity communication was developed that is realizable with existing technology and meets 2012 ITRS predictions. Challenges associated with incorporating proximity communication into DRAM were characterized and several innovations were developed that alleviated these challenges.

A new global I/O routing structure was discussed that promises to increase the number of data signals that can be read and written to a memory array. While the slice architecture developed in this thesis promises to increase the modularity of memory systems. This thesis provides the ground work for developing a new memory hierarchy that utilizes proximity communication.

CHAPTER 2—PROXIMITY COMMUNICATION

Proximity communication is a wireless chip-to-chip communication technology that occurs when two chips are placed within close proximity of each other. Two chips are placed face to face and their bonding pads are allowed to come within close proximity of each other without touching. The arrangement effectively creates a capacitive connection between the chips. The main concept of capacitive coupled proximity communication is illustrated in Figure 2.1 [1].

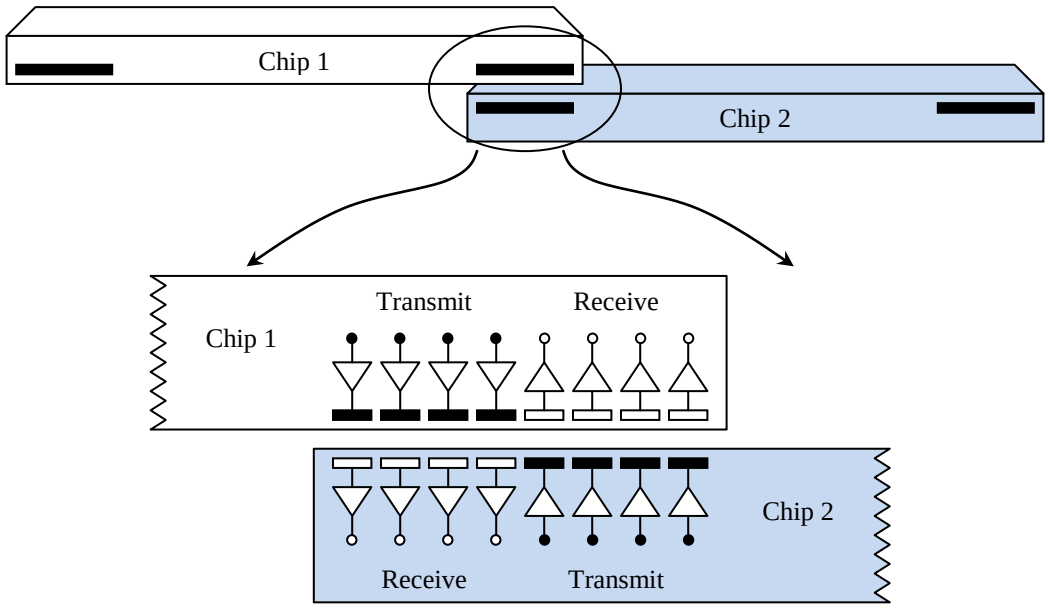


Figure 2.1 Cross-sectional view of placing two chips within close proximity of each other [1].

The metal-insulator-metal parallel plate capacitor, formed by using proximity communication, is used to transmit and receive electrical signals between the two chips. Figure 2.2 shows the typical transmit and receive circuits used to communicate across the chip-to-chip coupling capacitance (C_s), and the parasitic capacitances (C_{pt}, C_{pr}) associated

with the receive and transmit circuits [1]. This novel technology is allowing research engineers to exploit the major advantages of using capacitive coupled proximity communication compared to current and future chip-to-chip communication technologies.

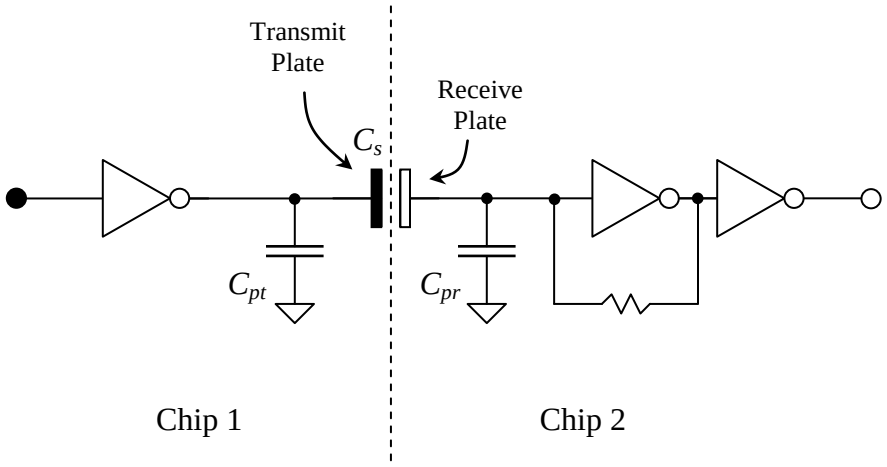


Figure 2.2 Transmit and receive circuit diagram used in proximity communication [1].

2.1 Advantages

In 1994, David Salzman (then of Polychip Inc.) and Thomas Knight (then of MIT Artificial Intelligence Laboratory) coauthored a technical paper. It was presented at the 1994 International Conference of Multichip Modules and titled "Capacitively Coupled Multichip Modules." In this paper they discussed the feasibility of proximity communication [2]. The advantages, integration, and future work required to develop proximity communication into a viable interconnect technology was also presented.

Since the introduction of this revolutionary idea, several other research teams have successfully demonstrated the viability of capacitive and inductive proximity communication [3]. Inductive proximity communication is proving to be an additional research avenue for developing proximity communication that has several of the same

advantages of capacitive coupled proximity communication. For the purposes of this thesis, the term proximity communication will refer only to capacitive coupled proximity communication.

It is helpful to view the chip-to-chip communication channel as a multi-lane highway with traffic flowing in both directions. Proximity communication would increase the number of cars that can travel in one highway lane (Figure 2.3). The reported increase in I/O density per square millimeter shows approximately 20 times (2000%) the number of I/O channels, or conversely, the silicon real estate required for I/O channels can be reduced by 20 times. The low value of capacitance used in proximity communication channels improves the practicality of this nascent technology.

2.1.1 I/O Density

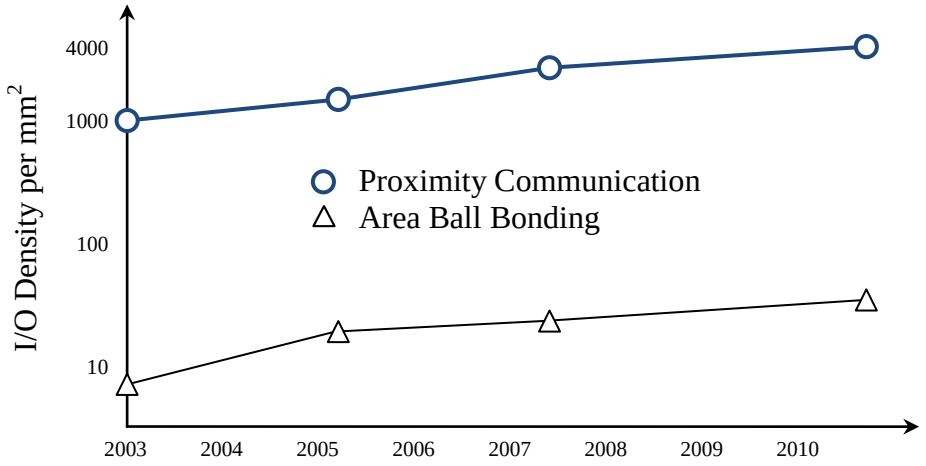


Figure 2.3 Proximity communication decreases the size required for a single I/O. The reduced communication channel provided by proximity technology allows for an order of magnitude increase in the I/O density [1].

When the spacing between pads, permittivity of the isolation material, and overlap area are known the value of the chip-to-chip coupling capacitor can be

determined. In order to understand the possible I/O density scaling, an analysis of the coupling capacitor and its effects on system performance is required.

The capacitance associated with two parallel plates each having an area, A, and spacing, d, is estimated using $C = \varepsilon_0 A/d$. A permittivity equal to 8.9 x 10^{-12} F/m (free space) can be used for initial estimates, realizing that it is possible to multiply this value by the relative permittivity for a better estimate.

As an example, using a pad size of 100 μm x 100 μm and a chip-to-chip separation of 1 μm, it is possible to have a maximum coupling capacitance of 10pF/mm^2. Selecting a minimum coupling capacitor of 50fF per I/O channel (signal) allows for a maximum of 200 signals/mm^2. The number of signals per mm^2 will increase when materials with a relative permittivity greater than 1 are used.

2.1.2 Removal of Off-Chip Wires

Removal of off-chip wires eliminates the complexity associated with the continual scaling of wire bonding processes. The initial benefit of removing the off-chip wires is the inductive behavior associated with the transmission channel is removed. The removal of the wires, and thus the inductance, also allows for a fixed impendence to be delivered to the proximity channel [2]. An additional benefit to removing the off-chip wires is that the challenges associated with the wire bonding process are also eliminated.

One of the challenges is the wire diameter shrinkage with each successive generation. The reduction of the cross sectional area for current flow results in an increase in the wire resistance with each generational shrinkage. The transmitter circuits and termination circuitry required to support this ever increasing resistance are required

to scale their power consumption to keep the bandwidth constant over these evolving generations. The removal of these wires allows for the removal of the area consuming termination circuitry along with reducing the power consumption (energy) of the transmitter circuitry.

2.1.3 Removal of ESD Structures

Once the proximity communication pads are formed, a normal passivation process can be used to cover the pads. The passivation will behave as the dielectric between the two plates of the proximity capacitor. The passivation now covers any exposed metal, preventing the possibility of electrostatic discharge (ESD) thus enabling the removal of the ESD circuitry. The removal of the ESD structures saves space, but it is the elimination of the associated ESD parasitics that is the main benefit.

ESD circuitry uses reverse biased diodes that prevent large voltages from entering the chip and destroying electrically sensitive circuits. The sizes of the ESD diodes are directly proportional to its ability to protect the internal circuitry. ESD diodes contribute a considerable capacitance to the chip-to-chip communication channel and thus slow down communications. In other words, the parasitic capacitance due to the ESD structures reduces the overall bandwidth of the channel. In addition, the power consumption resulting from charging and discharging the parasitic capacitance increases. Other benefits of removing the ESD structures include the reduction in silicon real estate and the ability to remove on-die terminations, discussed next.

2.1.4 Removal of the On-Die Termination

Current DRAM architectures are required to operate with an on-die termination resistance that matches the characteristic impedance of the transmission line formed by

the electrical path on the printed circuit board. Because the transmission line is so short in proximity communication, the transmission line effects become less important. The necessity to have a resistive termination is removed and a capacitive termination can be used (e.g. the input of a receiver).

The energy required to drive the transmission line can be substantially reduced because the load becomes purely capacitive and not resistive. As the energy per communication channel begins to drop, so does the area requirements for the communication circuitry.

2.1.5 Ease of Testability

Current multichip modules may require decapsulation, removing bonding wires, replacing the chip, and rewiring the chip to replace a defective part. Proximity communication does not rely on physical wires to connect the chips together. This allows the chips to be physically glued in place and unglued as needed. The chips can then be easily replaced and swapped which increases the ease of testability. Being able to replace defective chips and simply swap silicon components at the system level should increase system yield and reduce the system cost [2].

The 1994 conference paper by Knight and Salzman [2] discusses the advantages of using proximity communication because the technology increases chip bandwidth, increases I/O density, avoids on/off chip wire bonds, eases chip replacement at the system level, enhances system level testability, enables smaller chip sizes, and removes the need for ESD protection. Since this 1994 conference paper, several companies and universities have aided in bringing the proximity communication technology to fruition.

Sun Research Laboratories have published several journal papers that discuss their findings and solutions in the area of Proximity Communication [4 – 6].

2.2 Challenges

Creating a product that utilizes proximity communication is only possible if the practical technological roadblocks are solved. These include mechanical misalignment between the two metal plates, supplying power to the chips, and thermal removal solutions. Mechanical misalignment is the largest technological roadblock due to the fact that large misalignment can cause the proximity communication channel to not function or, even worse, misalignment can introduce errors in the communication channels.

Researchers at Sun Laboratories have been working on determining ways to overcome the mechanical misalignment challenge. Precise electrical sensors were developed to detect misalignment, and electronic circuits are used to electrically reposition the proximity communication pads [6].

2.2.1 Electronic Sensors for Measuring Misalignment

There are varying sources of mechanical misalignment. System vibrations, the initial placement of the proximity chips, and thermal expansion are major sources of mechanical misalignment. If the misalignment is not corrected it will set a limit to the size and pitch of the metal plates used for proximity communication. Mechanical misalignment can have six degrees of misalignment. The six axes of misalignment are separation, two types of tilt, two types of translation, and rotation. Figure 2.4 shows the six axes of misalignment between two proximity pads.

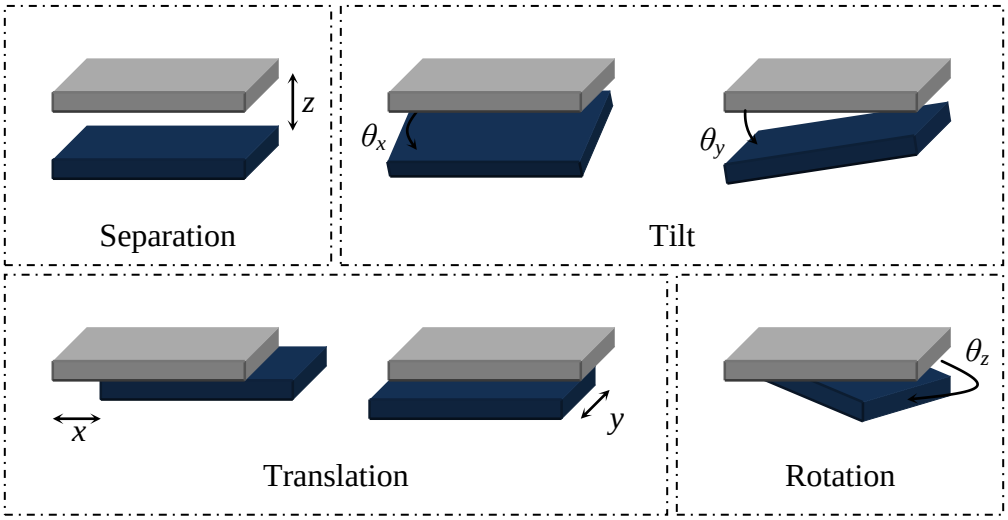


Figure 2.4 The mechanical misalignment of two proximity pads can take on six degrees of misalignment [6].

Simulation results of chip-to-chip misalignment show that the largest tolerable error in misalignment is a function of the pad size and pitch [6]. The misalignment tolerance sets a limit to the electronic sensor's precision. Novel circuit techniques were used to create electronic sensors embedded in the same silicon as the proximity communication circuits [6].

The circuit seen in Figure 2.5 is the rectifier circuit used to measure chip separation. A clock signal is sent through the proximity capacitor and the displacement current is measured on the receiving chip. The rectifier circuit allows for precise measurement of the displacement current that rejects all parasitics. Placing the separation sensor in multiple positions around the chip allows for the measurement of the tilt angles θ_x and θ_y. The rectifier circuit allows for chip-to-chip separation and tilt angles to be measured with a resolution of 0.2 μm which is well under the misalignment tolerance.

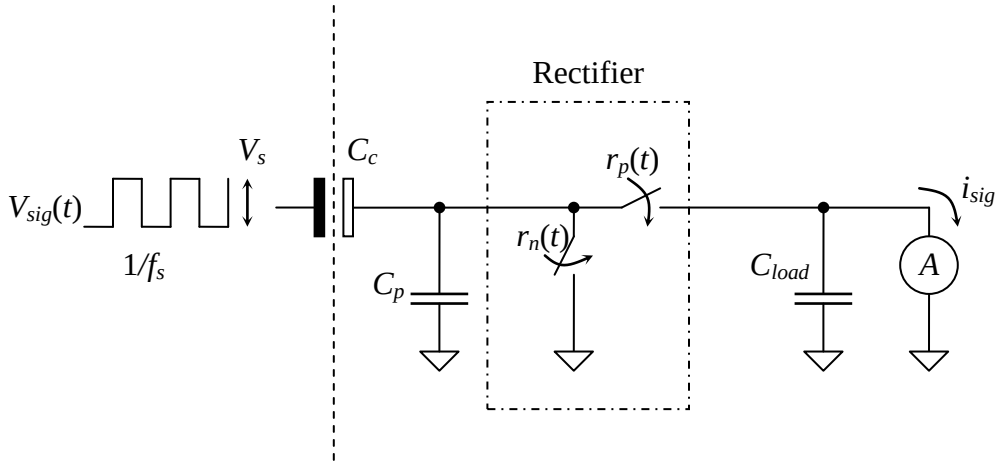


Figure 2.5 Electronic sensor used to measure capacitance values [6]. Multiple placements of the sensor allow for the measurement of the tilt angles.

In order to measure *xy* translation along with rotational misalignment, as seen in Figure 2.4, a sensor technology was developed [6]. A small movable scale that moves orthogonally to the measurement axis is termed a vernier scale. Figure 2.6 shows the vernier structure developed consisting of several metal plates on each of the proximity chips.

The vernier scale works by transmitting alternating data on each of the transmission pads. The overlap of the metal plates will eventually destroy the received data on one of the receiver metal plates. The position of the unknown data is used to determine the translation between the two chips. As seen with the chip separation sensor, it is possible to place multiple vernier scales around the chip to determine the rotational misalignment.

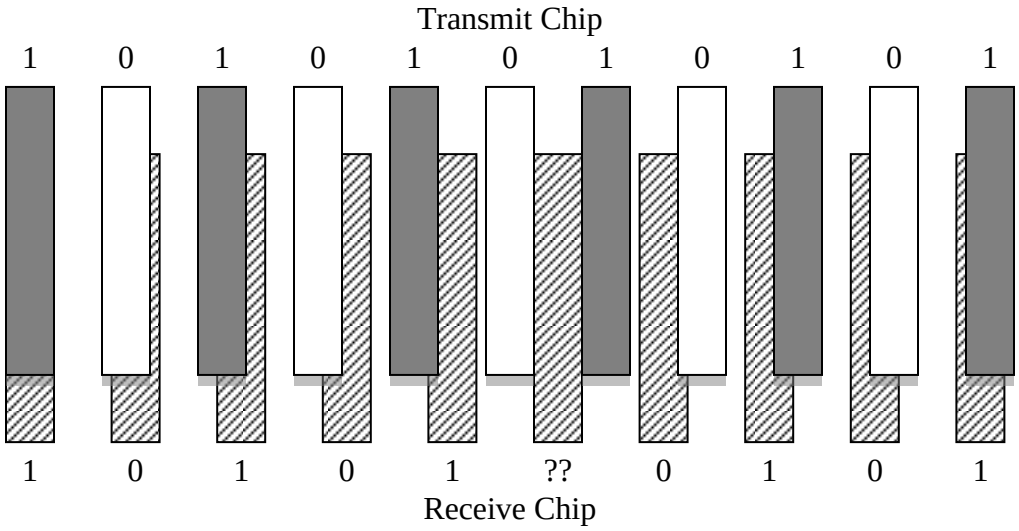


Figure 2.6 Metal plates used to create a vernier scale to measure *xy* translation [6].

Sun Research engineers have manufactured several test chips that contained a proximity channel along with the electrical misalignment sensors. The simulation versus silicon results are displayed in Figure 2.7. Two chips were placed face-to-face to enable proximity communication between chips. One of the proximity communication chips was held fixed while the other chip was placed on a movable arm which allowed six dimensional positioning so that each dimension of misalignment could be measured.

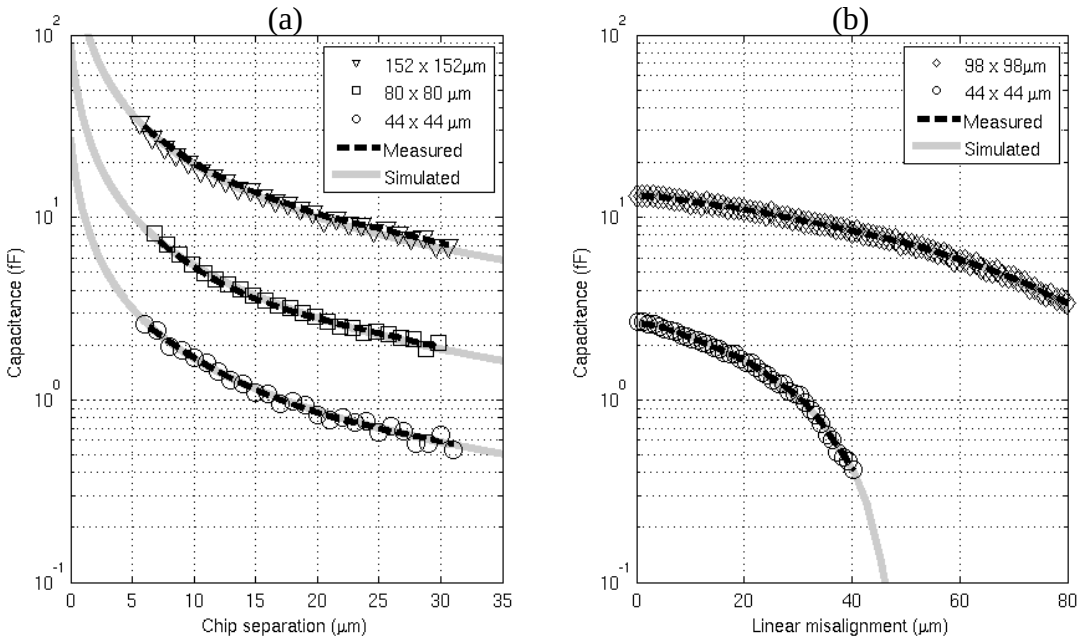


Figure 2.7 Simulation versus silicon data for chip separation (a) and linear misalignment (b) measured with on-chip electronic sensors [6].

The work presented provided the first step to overcoming a major challenge in bringing proximity communication to production. The electronic sensors allowed for precise measurement of six dimensions of misalignment. The research team developed an electronic re-alignment scheme that provided the second step needed to overcome proximity misalignment.

2.2.2 Electronic Re-Alignment

The receiver and transmitter array shown in Figure 2.8 is the ground work of electronic realignment circuitry. Developed by Sun Researchers, this array has the ability to electrically reposition the transmitter pads to align the transmitter and receiver pads. The receiver pads sit on 50 µm centers, while the smaller transmitter pads sit on 12.5 µm centers. In Figure 2.8, nine transmitter pads and one larger receiver pad are used to create a single proximity communication channel.

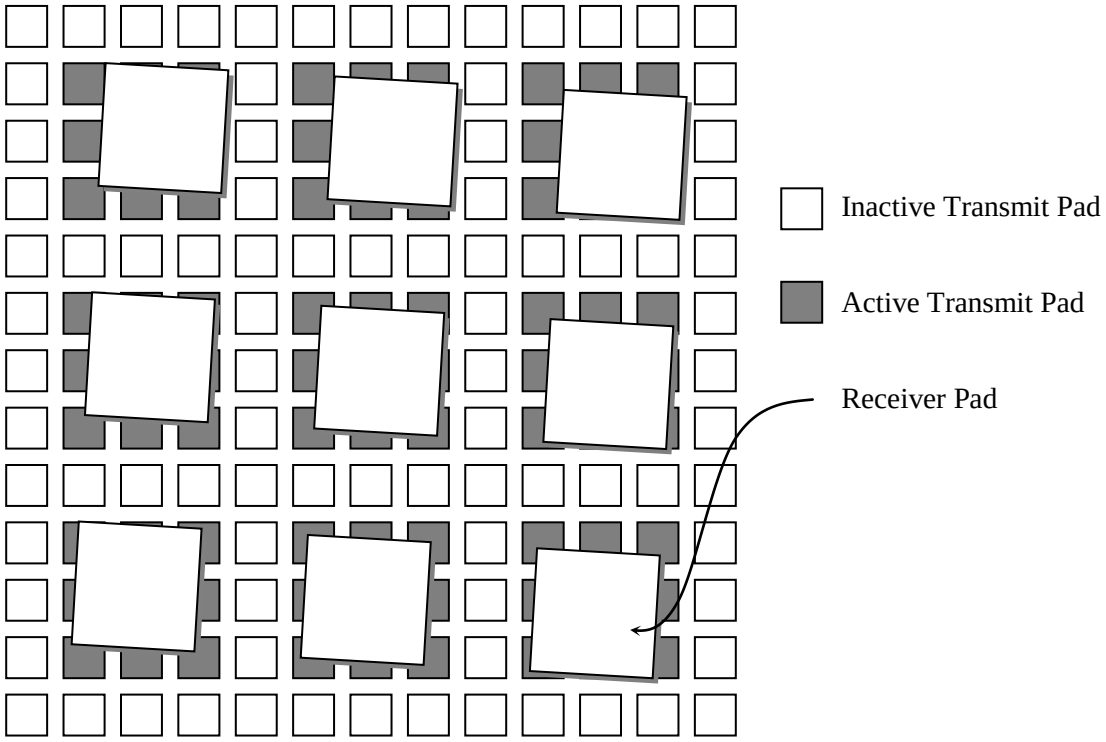


Figure 2.8 Receiver and transmitter array used to electrically align the proximity channel once the misalignment is determined [4].

Multiplexors can be used to electrically steer the transmit data towards the directed receiver pads. The addition of electronic realignment circuitry will increase the power consumption needed to drive a proximity channel. A novel circuit using NMOS-only pass gates as multiplexors was developed to decrease the power consumption for electrical realignment [5]. The concept of one dimensional steering circuitry can be seen in Figure 2.9.

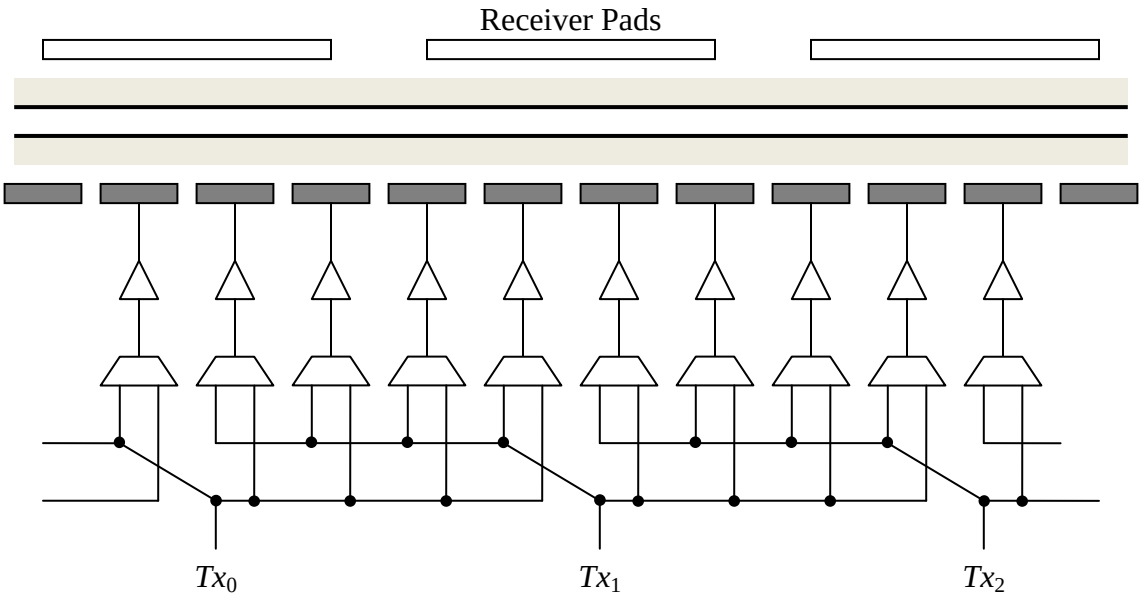
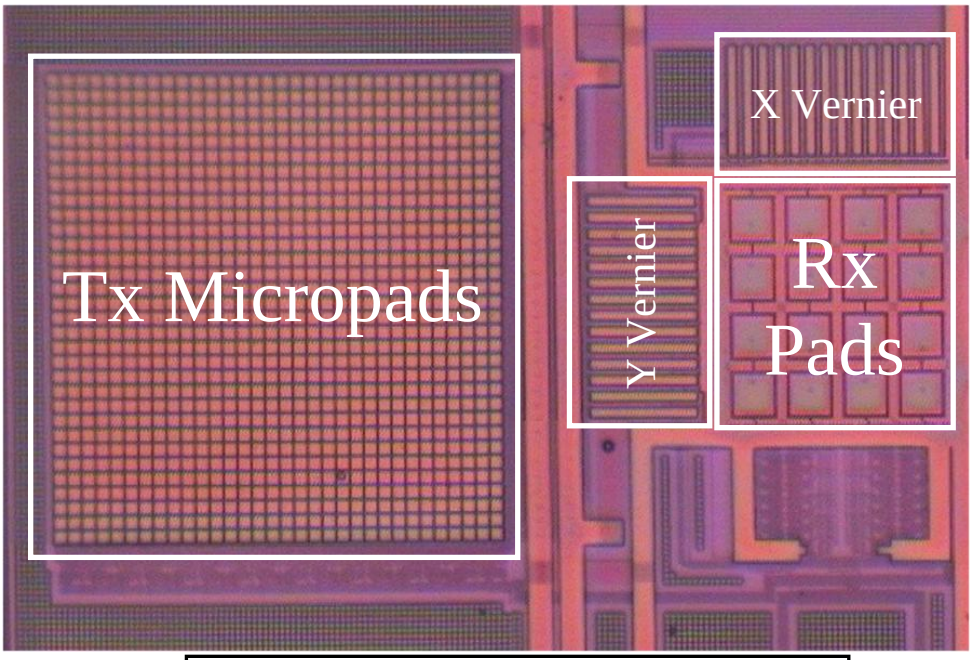


Figure 2.9 One dimensional steering circuit used to realign the receiver and transmitter pads [4]. The T_{x1} transmitter signal spans eight micro pads allowing for a 50µm shift in the transmitter pads. A two dimensional steering circuit can be made by using two one dimension steering circuits.

Figure 2.9 shows an abstract view of electrically steering the transmitter pads to the correct receiver pads. For two-way electronic steering, the T_{x1} signal is gated at a multiplexor by a control signal. Testing the implementation of the electrical alignment circuitry consisted of creating a test chip using a 16 x 16 array of receiver pads measuring 200 µm by 200 µm. The transmit array consisted of 1024 micro transmit pads that measured 400 µm by 400 µm. The test chip showed precision alignment at a resolution of 6.25 µm over a range of ± 100 µm. The test chip and realignment range is shown in Figure 2.10.

(a)



(b)

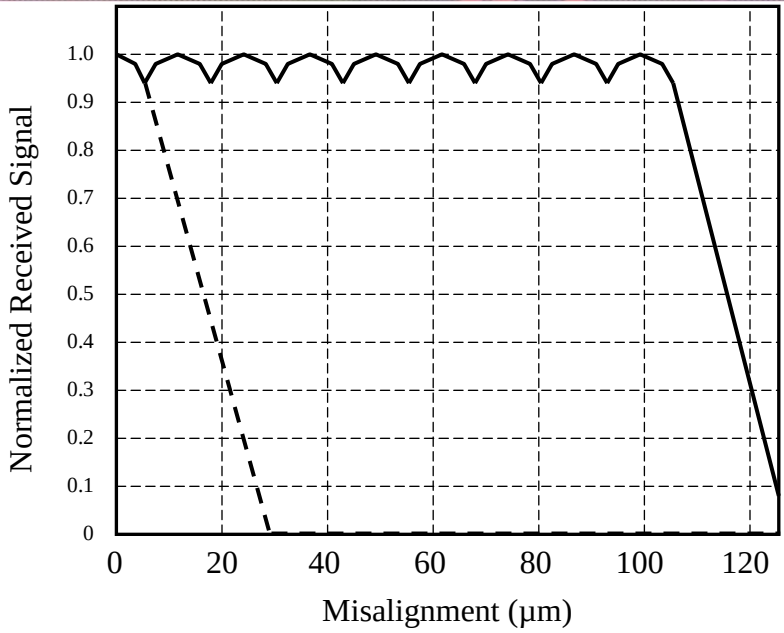


Figure 2.10 Die showing the transmit array and receiver array (a) and the simulated signal versus misalignment (b) [4].

2.3 Summary

Proximity communication is an up and coming chip-to-chip communication technology that has many advantages. The increase in I/O density, the removal of off-chip wiring, the

removal of ESD structures, and the removal of on-die termination circuitry allows for a higher bandwidth per channel and a reduction in the power consumed compared to existing technologies.

Mechanical misalignment is a major challenge to bringing this technology to fruition. Electrical sensors along with electrical realignment can be used to overcome the challenges of mechanical misalignment. Several test chips were developed [4 – 6] in varying technologies that gave physical evidence of the advantages of proximity communication and showed how novel circuit techniques can be used to overcome the challenges posed by mechanical misalignment.

CHAPTER 3—DRAM TRENDS

The development of a new DRAM architecture requires a thorough understanding of the DRAM market. This chapter discusses the history, current, and future trends of the DRAM market. Manufacturing costs are the major driving force of the DRAM market due to the market's competitive nature. Unlike other markets, the end user of DRAM dictates the majority of decisions made.

3.1 Memory Gap

The performance gap between microprocessors and DRAM has been in existence over the past three decades, continuously widening since 1980. Often this performance gap is referred to as the memory wall or the memory gap and can be seen in Figure 3.1.

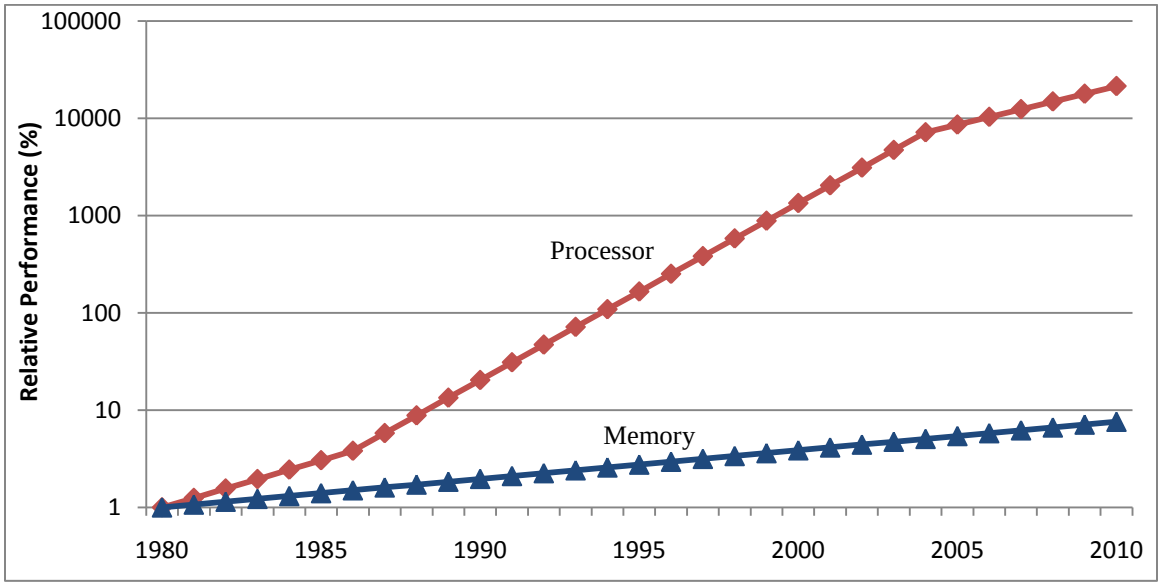


Figure 3.1 Relative performance change of the microprocessor and DRAM [7]. The microprocessor performance is a measure of the VAX MIPS, while the DRAM performance is a function of the access latency.

Figure 3.1 shows the relative performance gains of both the computer processor and the DRAM memory with respect to benchmarks in 1980. The processor reference is the VAX-11/780 which completed 500,000 instructions per second. This performance measurement is termed one VAX MIPS and is the standard relative benchmark for processor performance. The DRAM memory benchmark begins in 1980 with the 64 kb DRAM part. The memory plot shows the average DRAM access latency improving at a rate of roughly 7% per year. Relative processor performance is in stark contrast, increasing roughly 50% per year from 1986 – 2004 and roughly 20% during other times. The performance difference between processors and DRAM must be fully understood before attempting to develop a DRAM architecture that will improve the relationship between the processor and the main memory.

Gordon Moore's 1965 article titled "Cramming more components onto integrated circuits" was the genesis of Moore's Law [8]. Moore noticed that the complexity of minimum component costs doubled every two years. Over the years Moore's Law has been refined to state that the number of transistors that can be placed on a single integrated circuit will double every two years. Another way to look at Moore's Law is a scaling factor of $\sqrt{2}$, or a 41% increase in the number of transistors on an integrated circuit per year.

Reviewing Figure 3.1, it is clear that the processor performance is scaling roughly with Moore's Law while the DRAM performance has a stunted performance increase relative to Moore's law. The genesis of this disparity is due to the use of transistors being different for the processor versus the DRAM. Processor designers use the doubling of

transistors every two years to increase the number of instructions that can be performed per second. DRAM designers used the doubling of transistors every two years to increase the density of the DRAM chips. Figure 3.2 shows an alternative view of the "memory gap."

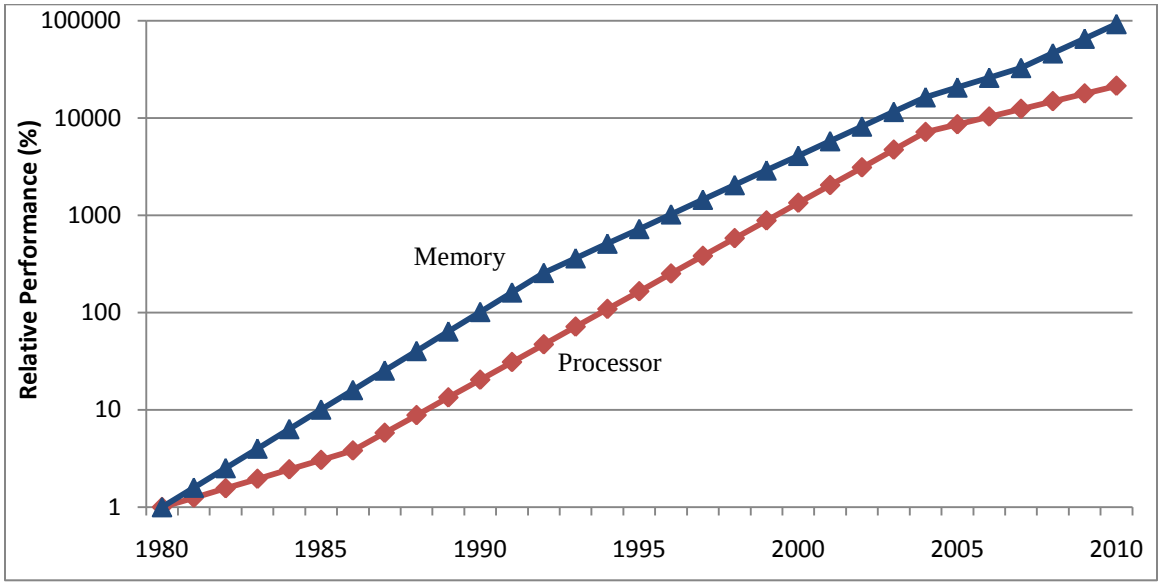


Figure 3.2 An alternative performance comparison between the microprocessor and DRAM [7]. The microprocessor performance is a measure of the VAX MIPS, while the DRAM performance is a measurement of the relative density of the DRAM.

It is clear that the DRAM out performs the processor when comparing the relative density of DRAM chips versus the relative performance of microprocessors. While increased processor performance may come at the expense of increased current consumption, the increase in DRAM density comes at the expense of access latency. For this reason DRAM access latency will not scale with microprocessor performance, but the density does.

3.2 The DRAM Market and Technology

DRAM density scaling trends have required manufacturers to develop significant innovations in the manufacturing process. The reason why density has become the main scaling figure for DRAM manufacturers rather than access latency has to do with the competitive nature of the DRAM market.

3.2.1 Selling Price

The selling price of main memory DRAM is not set by the manufacturing costs. It is common for DRAM manufacturers to produce a part with profit margins less than 5%, but if there is significant volume a part with 5% profit margin might be manufactured. Small fluctuations in the manufacturing cost can reset the profit margin to 0% or even negative, due to the relatively low selling price of DRAM parts.

Figure 3.3 shows a price per bit decline of 36% per year from 1974 to 2005. This is one of the primary reasons why DRAM manufacturers use the increase in transistor count to increase their density. The increase in density also allows DRAM manufacturers to keep increasing the selling price of their products, thus remaining profitable. As the average price per bit decrease remains at 36% per year, the density increases at roughly 41% per year.

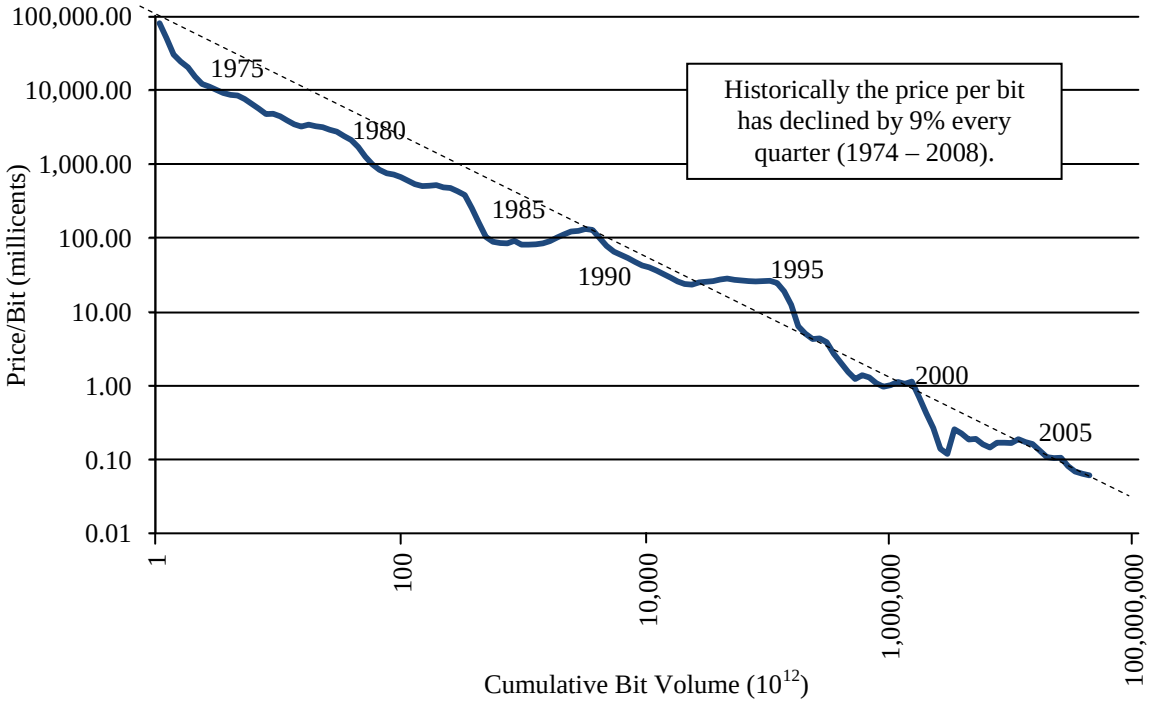


Figure 3.3 Historical price decline of the price per bit for DRAM products as a function of density [9].

The highly competitive DRAM market has its selling price eventually decided by the end user as manufacturers are required to decrease their price to remain competitive. Figure 3.4 shows this concept as the average selling price of DRAM products eventually reaches a minimum. The time it takes for a commodity product to reach a selling price minimum is decreasing over time, as can be seen in Figure 3.4. This is due to the DRAM market setting the selling price rather than the manufacturing cost setting the selling price.

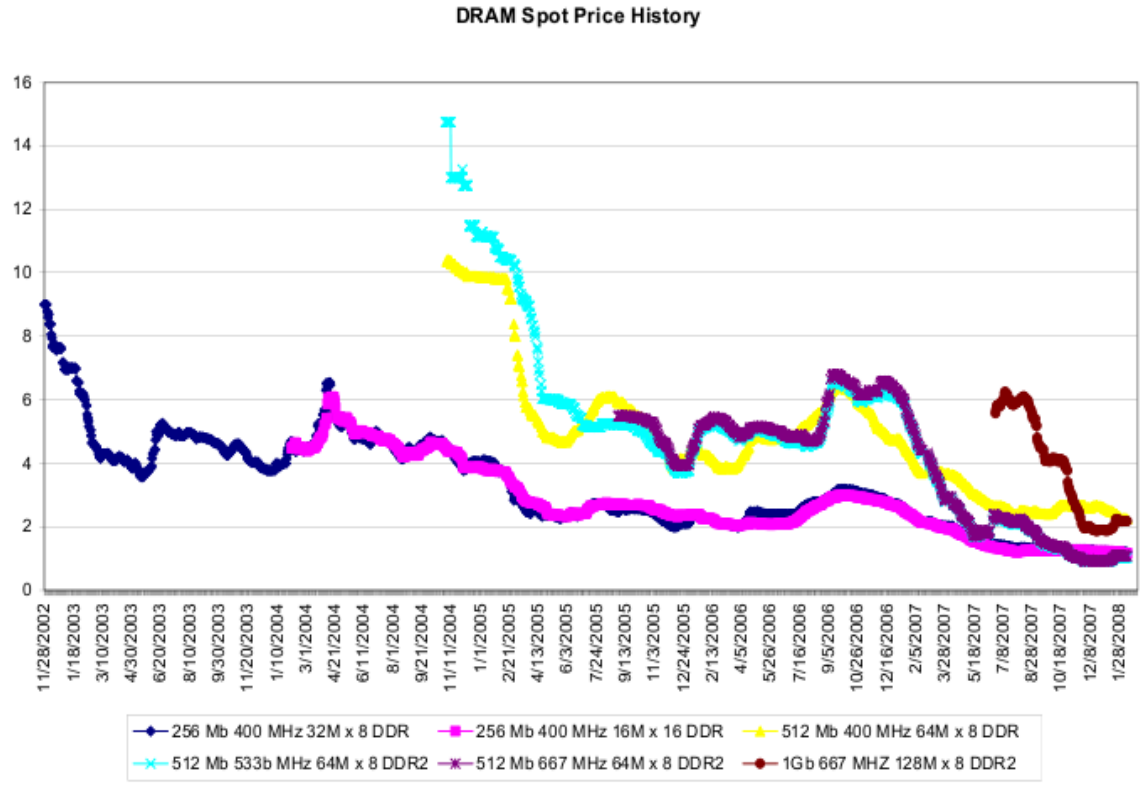


Figure 3.4 DRAM spot prices for varying products from 2002 – 2008 [10].

Oversupply, yield excursions, competitive advantages, and many other market dependencies will vary the selling price of commodity DRAM products. Due to the 36% price decline per bit of DRAM parts, memory manufacturers must increase their density at a rate greater than 36% per year to remain in business. Due to this, cost is the major reason why DRAM manufacturers have placed a premium on density scaling rather than latency scaling.

The following example emphasizes the impact of the selling price decline and the commodity nature of the DRAM market on the manufacturer's decision making process. If a 1 Gb DRAM part costs $2.00 today, then two years later a 2 Gb part would cost $1.64. The 36% reduction in selling price per bit per year is the intrinsic decay of the

selling price of DRAM products. Several manufacturers may produce a 2 Gb part and set the selling price at $1.64. If one manufacturer decides to reduce the price by $0.10, all other manufacturers will have to follow suit. This type of shallow pricing margin leads to large fluctuations in the profitability of memory manufacturers. Added to that, the pricing changes that occur during seasonal transitions are enough to change the profitability of all DRAM manufacturers. These transitions follow economic and seasonal cycles. This is why the DRAM selling price is often referred to as cyclical.

3.2.2 Wordline Scaling

Main memory DRAM typically uses a 512 wordline by 512 bitline memory array as the largest continuous memory array. The 256 kb memory array requires the addition of bitline amplifiers and wordline re-drivers at the edge of the array to break up the parasitics when multiple 256 kb arrays are used to create larger arrays. The physics associated with driving the wordline and bitline parasitics fundamentally sets the access latency of the DRAM chip.

The selling price decline associated with DRAM manufacturing places density scaling as a major tool to combat the price per bit decline. Density scaling is the process of increasing the number of memory bits per unit area, and is made possible through successive technology shrinks. Placing more memory bits into a fixed area requires wordline lengths to decrease at a rate of 41% per year.

The wordline scales roughly with, or slightly behind, the technology shrink. The parasitic increase of wordline scaling can be seen as two major parasitics, capacitance and resistance. Figure 3.5 shows that as the wordline width scales to smaller sizes, the cross sectional area for current flow reduces and this increases the wordline resistance.

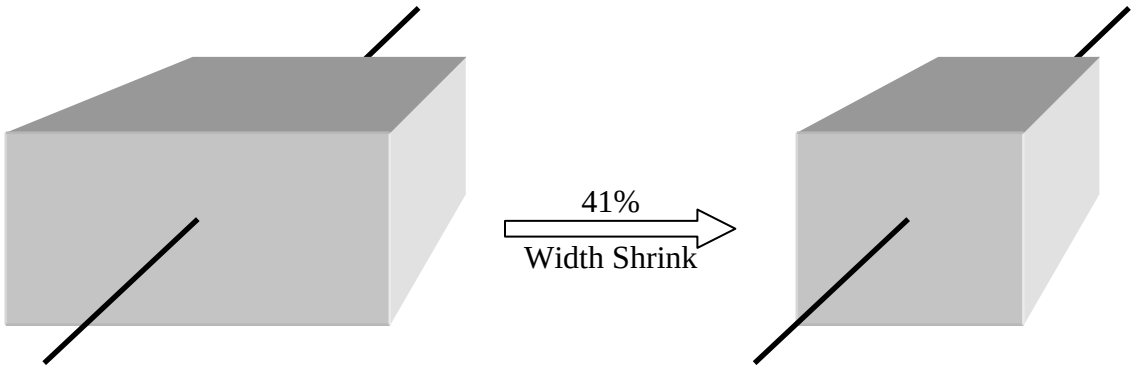


Figure 3.5 Wordline scaling reduces the cross sectional area of the wordline. The increase in wordline resistance requires additional innovation to keep the access latency scaling at 7% per year.

There are many technology innovations that have allowed wordlines to continue shrinking with technology scaling without increasing the wordline resistance, and thus the access latency. A recent innovation is the use of a titanium nitride barrier between the tungsten wordline cap and the polysilicon gate structure, Figure 3.6 [11].

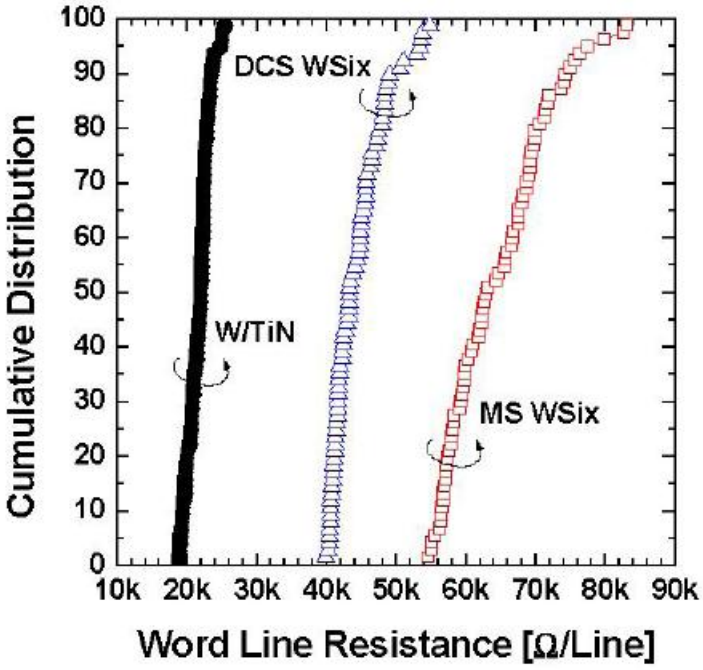


Figure 3.6 Cumulative distribution of the wordline resistance for varying silicide structures [11].

The capacitance associated with wordline scaling is less of a concern due to the fixed number of cells attached to each wordline segment (512 cells). The scaling continues to reduce the dielectric spacing between metal plates of a parasitic parallel plate capacitor. The metal plates that form a parasitic capacitance with the wordline can be adjacent wordlines, bitline contacts, or the silicon substrate. As the aspect ratio of the tungsten cap is increased to reduce the wordline resistance, the parasitic capacitance will increase.

The same way that density trumps latency, any new innovations in wordline technology will be used to increase the array efficiency of DRAM rather than to further reduce the access latency below the standard 7% per year scaling. Unlike the wordline, the bitline can be made of all tungsten because it is not used as the gate of a MOSFET (Metal Oxide Semiconductor Field Effect Transistor). The metal bitline allows for a reduction of bitline resistance to about half of the wordline resistance, thus reducing the column access latency with respect to the wordline access latency [12].

3.2.3 Bitline Scaling

The figure of merit often used to characterize the effects of bitline scaling is the bitline capacitance. The bitline capacitance is made up of two main components: parasitic capacitance of the metal wire and parasitic capacitance of the drain to bulk depletion capacitance formed by the bitline contact. Of the two parasitic capacitance values, the drain to bulk depletion capacitance is the dominant of the two. This limits the number of contacts that can be placed on each bitline. This ensures a proper signal-to-noise ratio for a high volume production (6σ variation possible). For this reason bitline scaling is not as critical as wordline or contact scaling.

3.2.4 Contact Resistance Scaling

The inclusion of metal contacts for the access transistor has allowed for a reduction in the memory contact resistance. Leading edge technology is reporting a contact resistance that falls below 400Ω [12]. This, along with other advances, has allowed the column access latency (t_{CL}) to remain at 11 – 12.5 ns for a 56 nm technology.

$$t_{CL} \approx command\ propagation + data\ access$$

The column access latency refers to the time it takes for an external command to begin receiving data from an open page. This time is essentially equal to the column latency plus the time it takes a command to be received into the DRAM part plus the time for the data to pass through the data-path. This is the initial latency seen at the memory controller when accessing data from an open page, and it affects the overall bandwidth of the memory.

3.3 DRAM Generations

Significant DRAM innovations have been introduced at the beginning of each new generation. In 1997, for example, DRAM manufacturers switched to a fully synchronous DRAM. The new memory was dubbed synchronous data rate (SDR) DRAM and allowed memory manufacturers to begin using a global clock signal along with command signals to control the DRAM chip. Figure 3.7 shows the history of DRAM products.

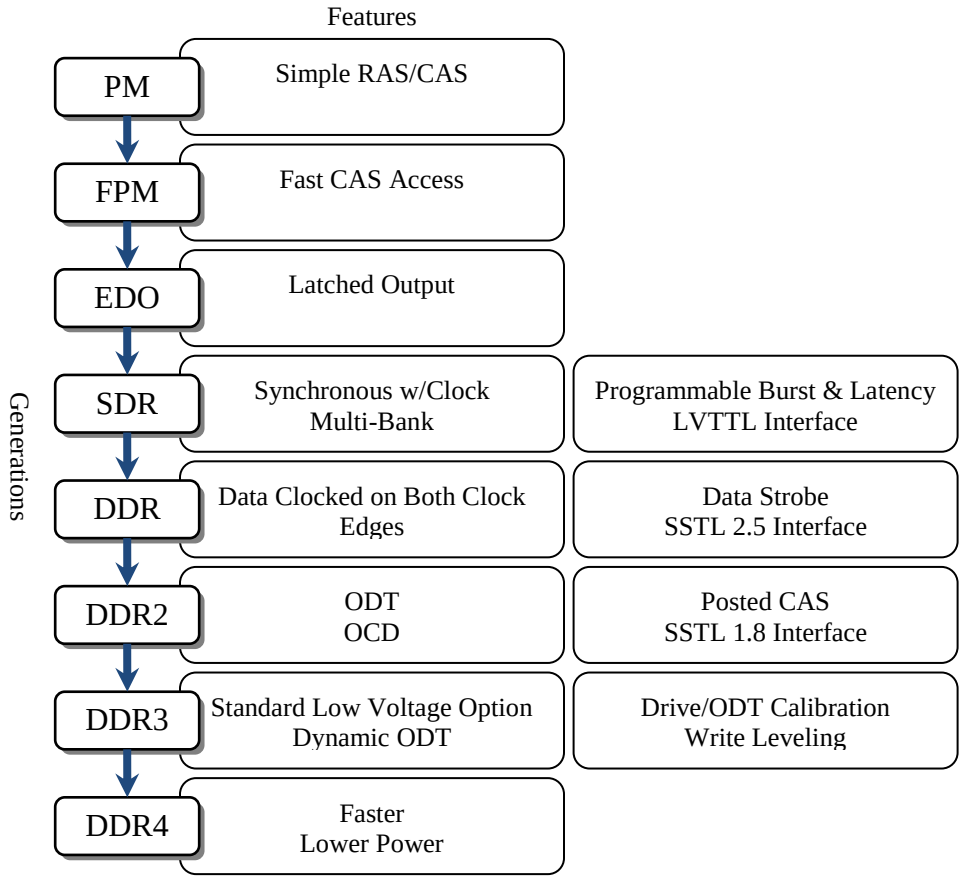


Figure 3.7 An evolutionary view of DRAM generations [14].

A generational approach was adopted that allowed SDR memory to operate at a bandwidth of 133 – 200 MHz. At the system level, the major changes that occur with each generation are related to power, bandwidth, and bus loading.

3.3.1 Power

The reduction of power consumption is a benefit for any product that uses DRAM chips. Server data centers hold DRAM power consumption higher in the list of important characteristics due to the sheer number of DRAM chips used in a server. The power consumption of each chip directly adds to the cost of running the data center due to cooling and power costs. The power reduction that occurs in DRAM generation changes

is not specifically unique to DRAM manufactures. This is an industry standard that requires semiconductor chips to scale their power supply voltage down in an attempt to reduce power.

Data centers have become a necessary part of business operation due to their ability to provide a large amount of computer resources. The computers of a data center, called servers, typically have specialized processors and more density than a home-use PC. Current server motherboards are adding 32 memory slots for each server. The processor utilizes an on-chip memory controller to control the multiple memory slots. Each memory controller is assigned a memory channel which is used to connect to a certain number of module slots as shown in Figure 3.8.

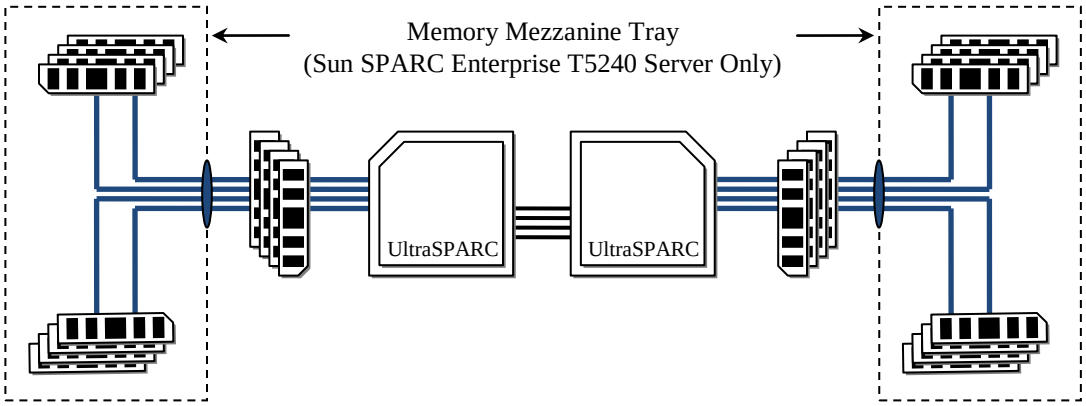


Figure 3.8 Server architecture of a Sun SPARC Enterprise server showing the memory capacity and number of memory channels [15].

Figure 3.8 shows the architecture of an enterprise server which uses Sun SPARC microprocessors. Each processor utilizes four memory channels for a total of eight memory channels for the server. A memory access is performed by turning on each memory channel and accessing one of the modules connected to each memory channel. Each module accesses nine DRAM chips in parallel, with a total of 72 DRAM chips

being accessed in parallel during a memory access. Each DRAM chip can dissipate anywhere from nothing to roughly 500 milliwatts (for a 1.5V DDR3 memory module [16]). The 72 chips would therefore produce 36 watts of power per access. This type of power consumption per access converts to approximately 20% of the total server power usage and the number is rising due to the increase in the number of DRAM chips required per server. Pressure is put on DRAM manufacturers to reduce the power consumption of DRAM chips over a generational process.

The easiest way to reduce power consumption is to reduce the power supply with every generation switch. This allows the current consumption to increase while at the same time allowing the power consumption to be reduced. Figure 3.9 shows the initial offering of a new generation having significant current reductions over the previous generation. However, the current consumption profile for each generation trends upwards as the bandwidth scales higher.

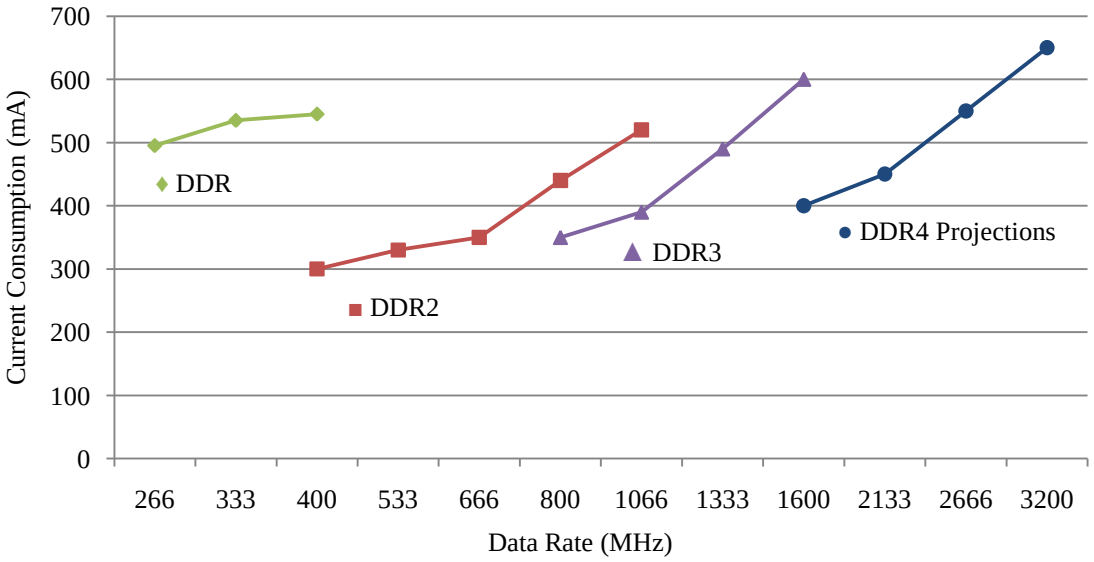


Figure 3.9 Current consumption of 1 Gb memory chips operating in x16 configuration for each of the proceeding DRAM generations, along with a prediction of the next generation [17].

3.3.2 Bandwidth

Bandwidth scaling trends are increasing at a rate of roughly doubling every three years or 26% increase every year. This 26% scaling did not occur until the incorporation of DRAM pre-fetching in 2000 [18]. DRAM pre-fetching refers to accessing multiple memory cells in parallel and then serializing the data as it is transmitted off the chip. Bandwidth scaling is possible through building upon previous generation's innovations.

Fast Page Mode (FPM) DRAM allowed an accessed row to remain open while the column address was changed repeatedly. The 32 data pins used on Fast Page Mode parts output (or input) data during each column address. Micron® developed a technology that allowed the column address to change before the data pins had completed their access. This innovation was termed Extended Data Out (EDO) DRAM and allowed a 5% - 10% increase in column cycle time.

DRAM bandwidth is often reported as the maximum data rate possible. Before the innovations of Extended Data Out DRAM, the bandwidth was limited by the column cycle time, due to the fact that the column address could not change until the data access was complete. In 1995 Extended Data Out DRAM was operating asynchronously with a CAS cycle time of 20ns. During the lifetime of Extended Data Out DRAM, synchronous innovations were proving technological superiority over asynchronous parts.

Using a master clock signal in DRAM to control the timing of input and output signals is referred to as a synchronous DRAM. There exists at least one patent that discusses the use of using both the rising and falling edges of the clock signal in Burst Extended Data Out memory, which was an initial attempt to introduce double-data rate (DDR) concepts to DRAM [19].

Figure 3.10 shows that initial SDRAM offerings coupled the off chip bandwidth to the bandwidth of the DRAM memory core. The physics associated with increasing the density in memory chips placed a limit of roughly 200 MHz for the array column path cycle time. In contrast, 1997 microprocessors were performing at 600 MHz, while DRAM chips only offered a bandwidth of 67 MHz.

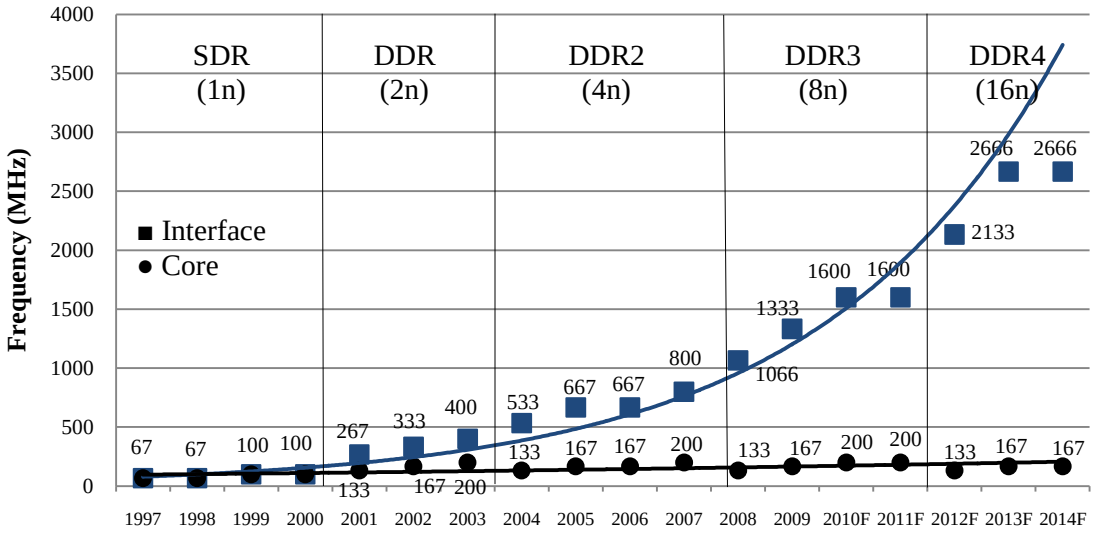


Figure 3.10 Array pre-fetch of two and higher has allowed off chip bandwidth to increase at a rate of 26% per year since 2000, while the core frequency does not scale [18].

In order to remedy this performance gap, the memory controller gradually increased the width of the memory channel from 16 bits wide in 1980 to 64 bits wide in 1993. Meanwhile DRAM manufacturers began using an array pre-fetch that allowed the entire memory word to be accessed in parallel, and later serialized. The inclusion of array pre-fetch along with double data rate DRAM allowed the off-chip bandwidth to be decoupled from the array core.

Double data rate refers to using both the positive and negative edges of the clock signal for a data access, which inherently doubles the data bandwidth. Figure 3.11 shows

that DDR memory required a pre-fetch of 2n and a serialization technique for each data pin. This increase in array pre-fetch allowed the memory core to remain operating at its prescribed 133 – 200 MHz range while the data pins were increasing their bandwidth.

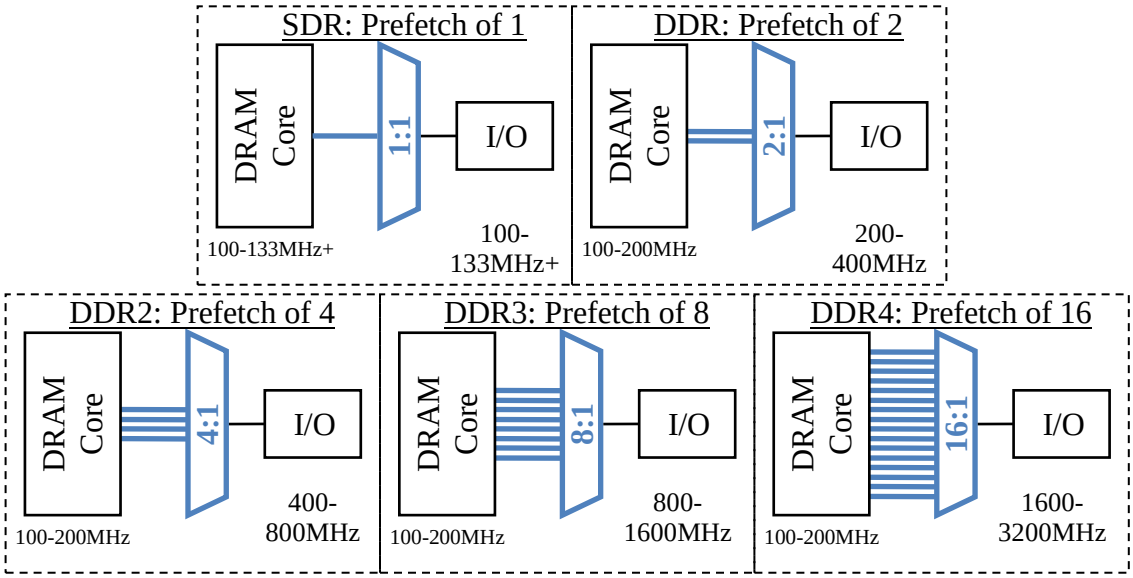


Figure 3.11 DRAM pre-fetch and bandwidth evolution [14].

Although there was an order of magnitude difference in DRAM data rate versus microprocessor clock rate, the bandwidth of DRAM parts began to scale at roughly 26% per year in 1997.

DDR3 parts currently use an 8n pre-fetch to reach a 1.6GHz data bandwidth. Figure 3.12 shows the International Technology Road-map for Semiconductors (ITRS) predicts a continual 26% per year scaling for the bandwidth. If the column path remains capped at 200 MHz, a pre-fetch of 16n will be required for a 2133 MHz DRAM part.

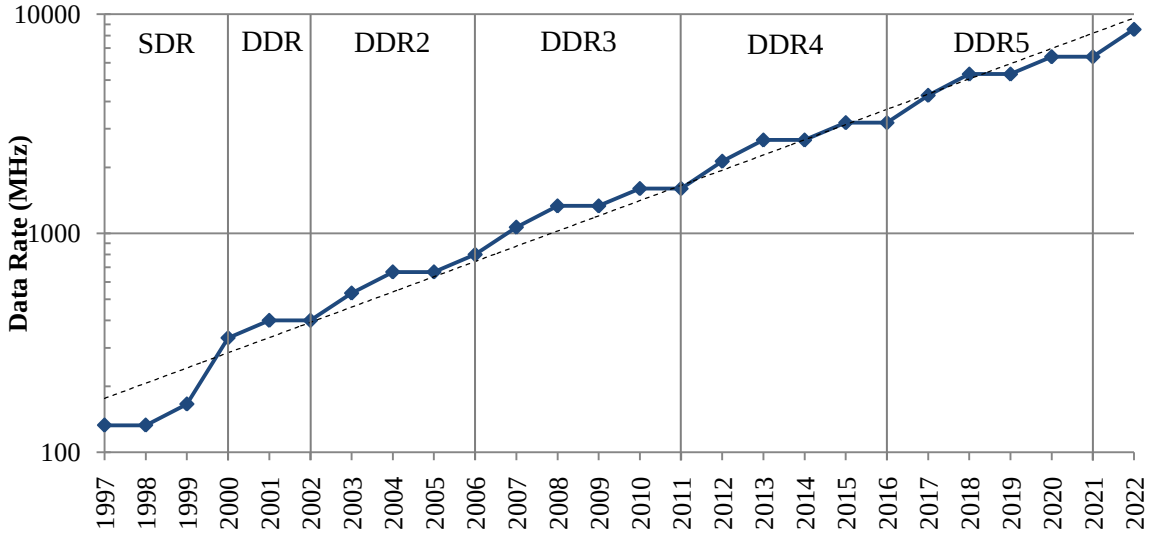


Figure 3.12 The ITRS DRAM bandwidth roadmap [20].

An important point to note is that as the memory array increases its frequency from 133 to 200 MHz there is also an increase in the energy required to get that increase in data rate. It is because of this that memory manufacturers leave the 200 MHz array operation for the end of a generation's lifetime. In this way, an introduction of a 133 MHz part operating with an increased pre-fetch will have a larger power reduction, since the same page size is opened in each generation.

3.3.3 Bus Loading

DRAM innovations have allowed for the generational progression of DRAM. The technology scaling has allowed for increased density on each DRAM die but the system requirements have increased substantially, allowing the possibility of 1024 Gbs of DRAM (128 MB). This requires memory modules to adopt new technologies that allow for an increased DRAM chip count per DRAM module.

The Sun memory configuration seen in Figure 3.7 allows for four memory modules per memory channel. This sets an empirical limit to the bandwidth, as seen in Figure 3.13, of the memory channel due to the load that each occupied Dual Inline Memory Module (DIMM) slot places on the memory channel.

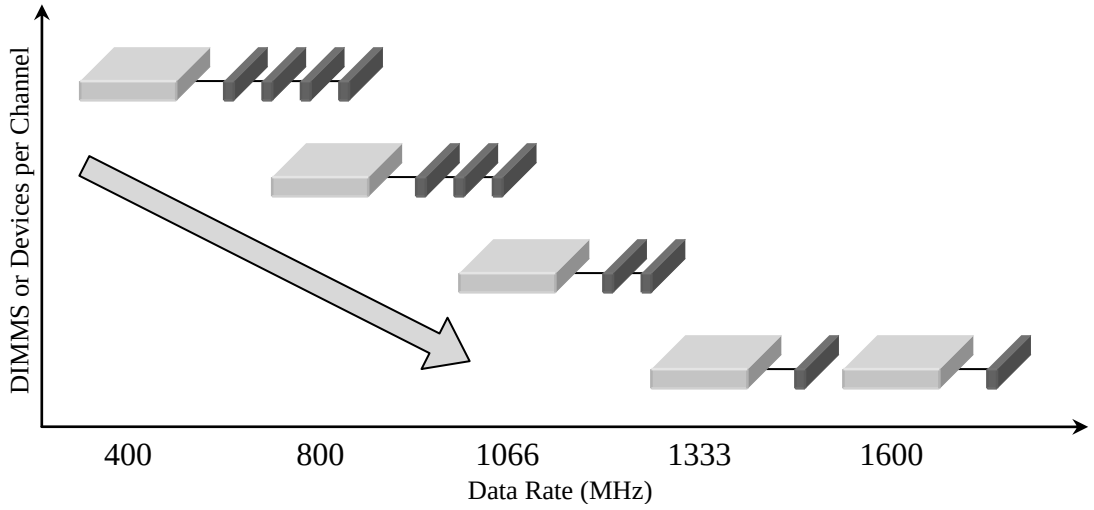


Figure 3.13 The number of occupied memory slots reduces the bandwidth of a memory channel [18].

The Series Stub Terminated Logic (SSTL) is utilized in many modern computers due to the provision of easy memory upgrades. The series stub terminated logic refers to terminating electrical signals at each memory module with a resistive pull up device that prevents transmission line reflections from interfering with transmitted data on the shared memory channel.

The resistive termination network, along with module loading, places a bandwidth limitation on the shared memory channel. For this reason there is a capacity limit using the SSTL industry standard. Memory manufacturers have developed several technologies to overcome the module capacity limitations. There are several classes of memory

modules, some tailored to high capacity systems while others are tailored towards the main memory market.

Unbuffered DIMMs are the standard memory modules that are used in home computers which have a limited number of DIMM slots per memory channel. To overcome the capacity and bandwidth limitations the Registered DIMM was created which allowed 3 RDIMMs to occupy a memory channel without reducing the bandwidth. The term registered refers to the addition of a discrete register being added to the memory module to reduce the load on the clock and command signals. The register is used to buffer the address, command, and clock signals to the DRAM chips on the memory module. This allows for higher bandwidth (1066 – 1333 MHz) to be achieved with 3 DIMMs per memory module.

The Fully-Buffered DIMM (FBDIMM) was created to increase the maximum capacity of a memory channel, past the attainable capacity using RDIMMs, without affecting the bandwidth of the memory channel. The FBDIMM employs the use of a memory buffer on the module. The advanced memory buffer is a high speed serial interface that uses serial-deserialization (SERDES) techniques to achieve a higher bandwidth. The drawback of the advanced memory buffer is that it consumes almost the same amount of power as a standard unbuffered DIMM and also carries with it a large price premium. The FBDIMM technology was adopted on DDR2 memory modules and required a new memory controller design to be incorporated into the computer system.

The lessons learned and advancements made with the FBDIMM were considered for the newest memory module. The Load Reduced DIMM (LRDIMM) is a new technology developed by Inphi® and Micron®. The new memory module consumes less

power than the FBDIMM and allows 9 memory modules per memory channel. Each memory module carries with it a price premium and additional power consumption. Newer module technologies allow for higher memory capacity and higher bandwidth, but the solutions require much more power than a conditional unbuffered DIMM.

3.3 Summary

The genesis of the memory gap was determined to be the price decline per bit of DRAM products which requires DRAM manufacturers to focus on density increases. The increase in density places a physical limit of column cycle time to be less than 200 MHz. The generational approach used by DRAM products relies on array pre-fetch to increase the chip's bandwidth.

CHAPTER 4—A 4 Gb DRAM ARCHITECTURE

Developing an accurate DRAM architecture survey is only possible after analyzing the market trends in Chapter 3. ITRS roadmaps show that a 4 Gb memory product will be manufactured in a 40 nm process in 2012. This chapter details the creation of a 4 Gb DRAM architecture that falls in line with 2012 predictions.

Figures 4.1 – 4.4 show the architectural progression of DRAM products. Each DRAM chip has a relatively standard architecture. The DRAM architecture consists of several major blocks. Using an analogy of the human skeleton to describe two of the major blocks found in DRAM architecture produces the SPINE and RIB analogy. The SPINE of the DRAM refers to the center periphery circuitry that spans the entirety of the chip. The RIB refers to the global wordline circuitry that branches off of the SPINE and vertically separates the array banks.

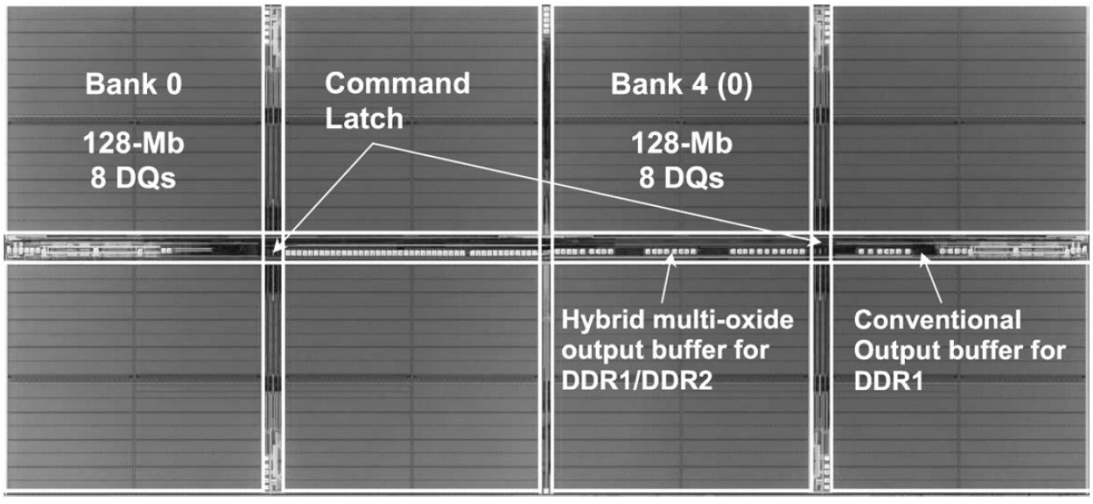


Figure 4.1 Micrograph of a 1 Gb DDR/DDR2 chip. Operating with a pre-fetch of 4n requires 32 global data lines for each 128 Mb bank [21].

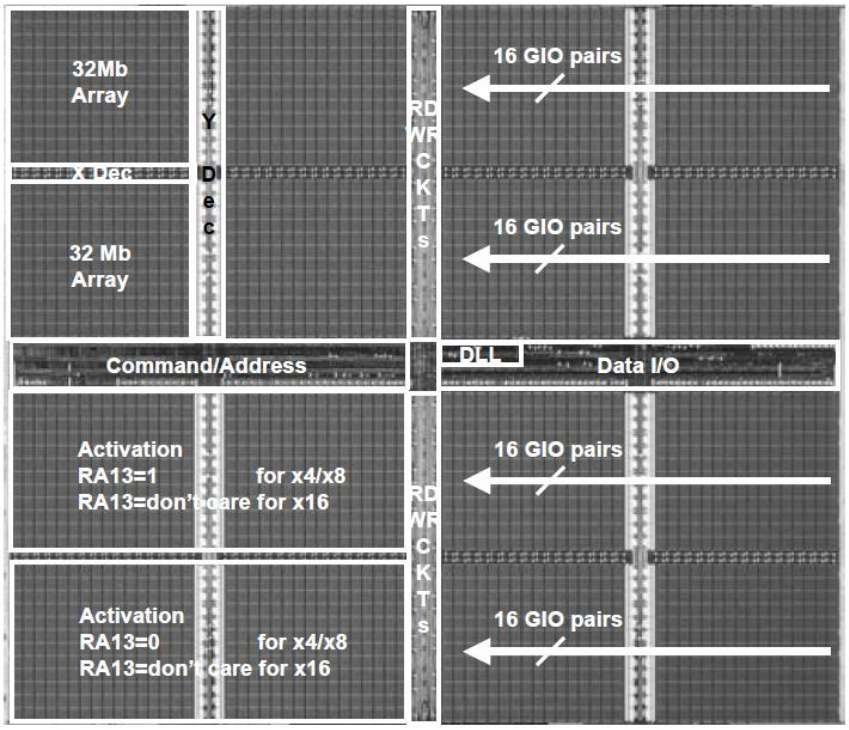


Figure 4.2 Micrograph of a 512 Mb DDR2 chip showing 32 data lines per 128 Mb bank [22].

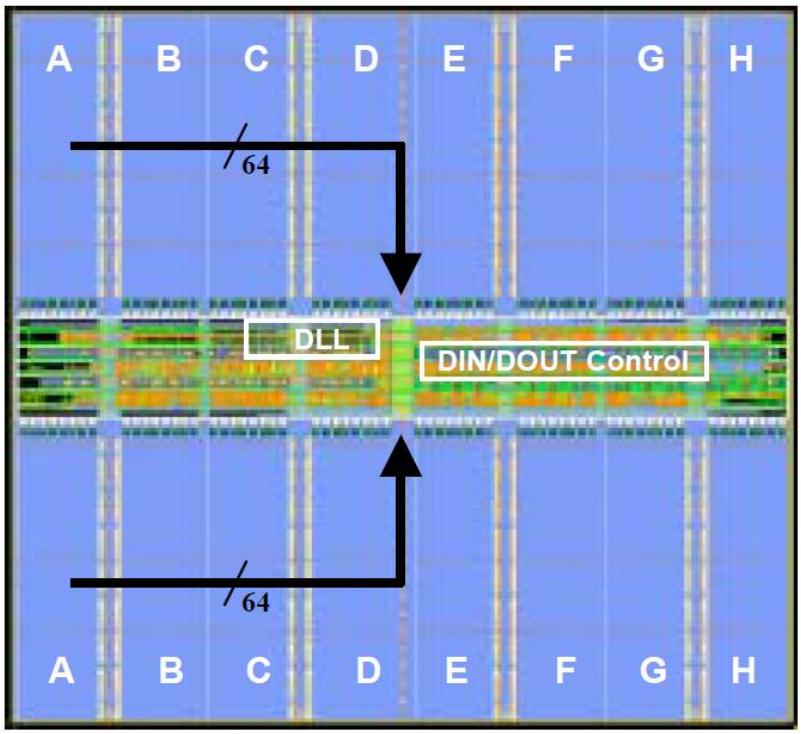


Figure 4.3 Micrograph of a 512 Mb DDR3 chip showing 64 data lines per 128 Mb bank [23].

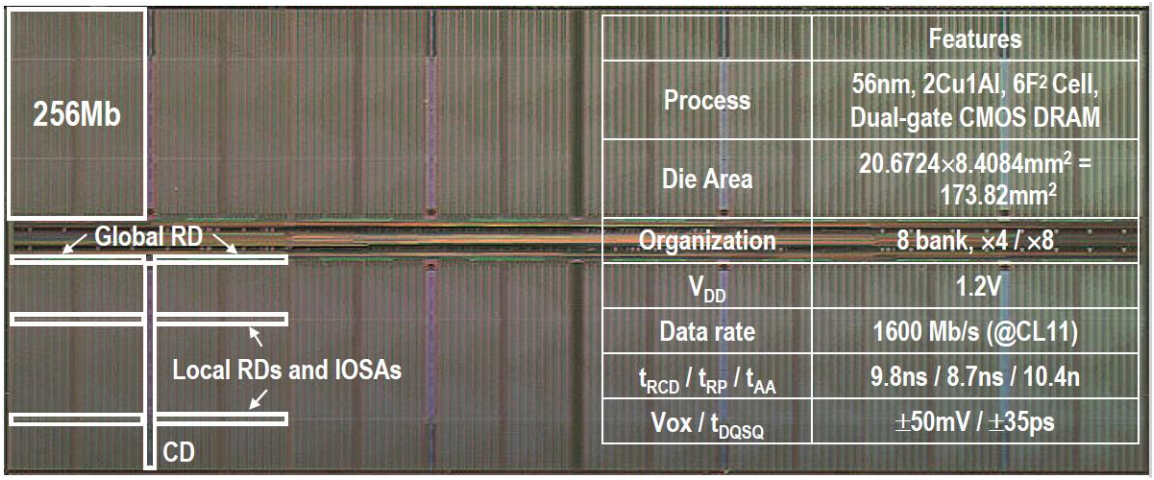


Figure 4.4 Micrograph of a 4 Gb DDR3 chip. Each 256 Mb bank requires 128 data lines [24].

The center of the chip is occupied by pad rows and other periphery circuitry, while the RIB houses the global row circuitry. Figures 4.1 – 4.3 show that the increase in density and array pre-fetch has required the global column structures to sit horizontally in the array.

4.1 Creating a 256 kb Array

The array development begins with the memory bit. This allows for the proper development of the array architecture and accurate size estimates. The minimum feature size used in the memory cell is the technology node. For a 40 nm process the minimum feature size is 40 nm. The smallest DRAM memory cell will have an area equal to $6F^2$, where F is the minimum feature size. Figure 4.5 gives a top down view of a standard $6F^2$ memory bit with the schematic superimposed on the layout. The capacitive memory element is vertically integrated into the memory element.

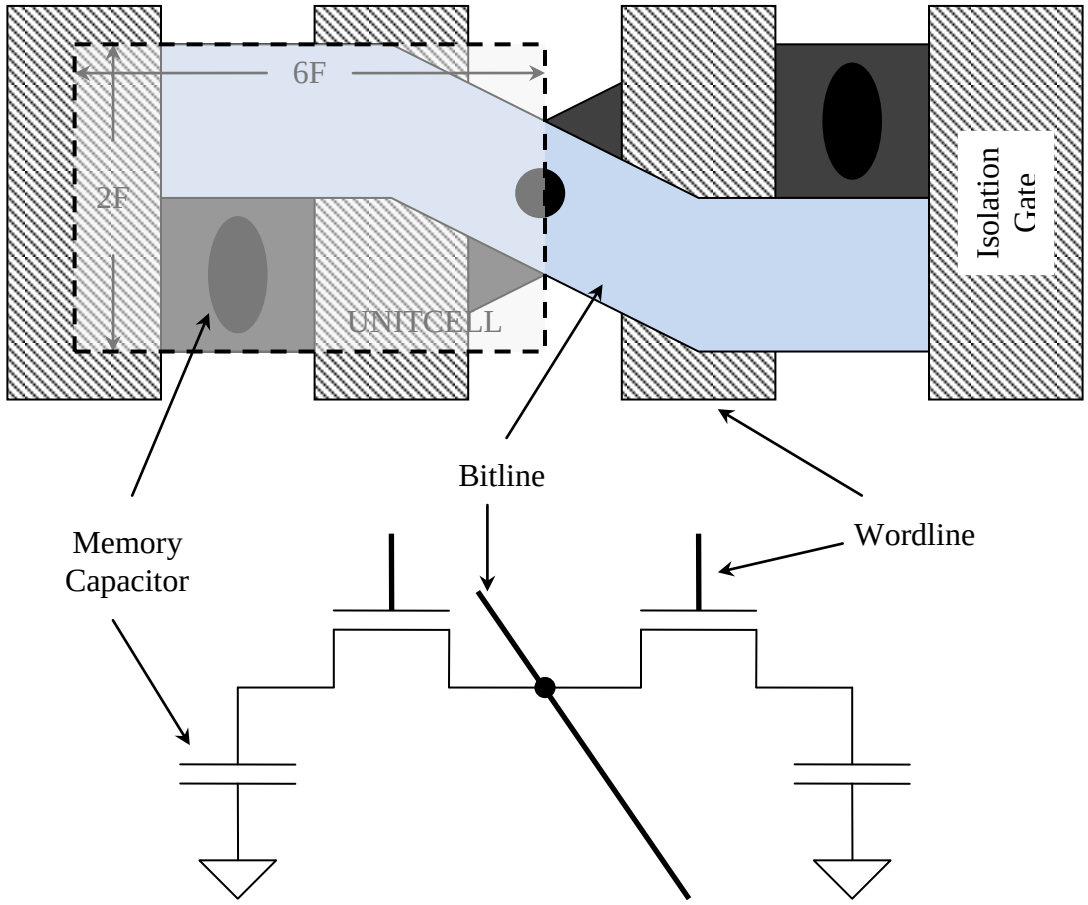


Figure 4.5 Layout and schematic of a $6F^2$ memory cell showing the major levels [25].

The one transistor – one capacitor (1T-1C) memory cell uses an NMOS switch as an access device with a capacitor as the memory element. To reduce the size of the memory cell two 1T-1C memory cells share their bitline contact as seen in Figure 4.5. The double memory cell element is often referred to as a memory bit or simply an mbit and is the fundamental building block for creating an array. Using our estimated technology node of 40 nm as the feature size it is possible to gain an accurate prediction of the cell size as 0.0096 μm^2 per memory bit.

The 256 kb memory array is the smallest continuous memory array available. It consists of 512 wordlines and 512 bitlines. In order to create the memory array, the unit

cell developed in Figure 4.2 is arrayed 512 times wide and 512 times high. Figure 4.6 shows the premise and implementation of creating a 256 kb memory array using the layout of a $6F^2$ unit cell.

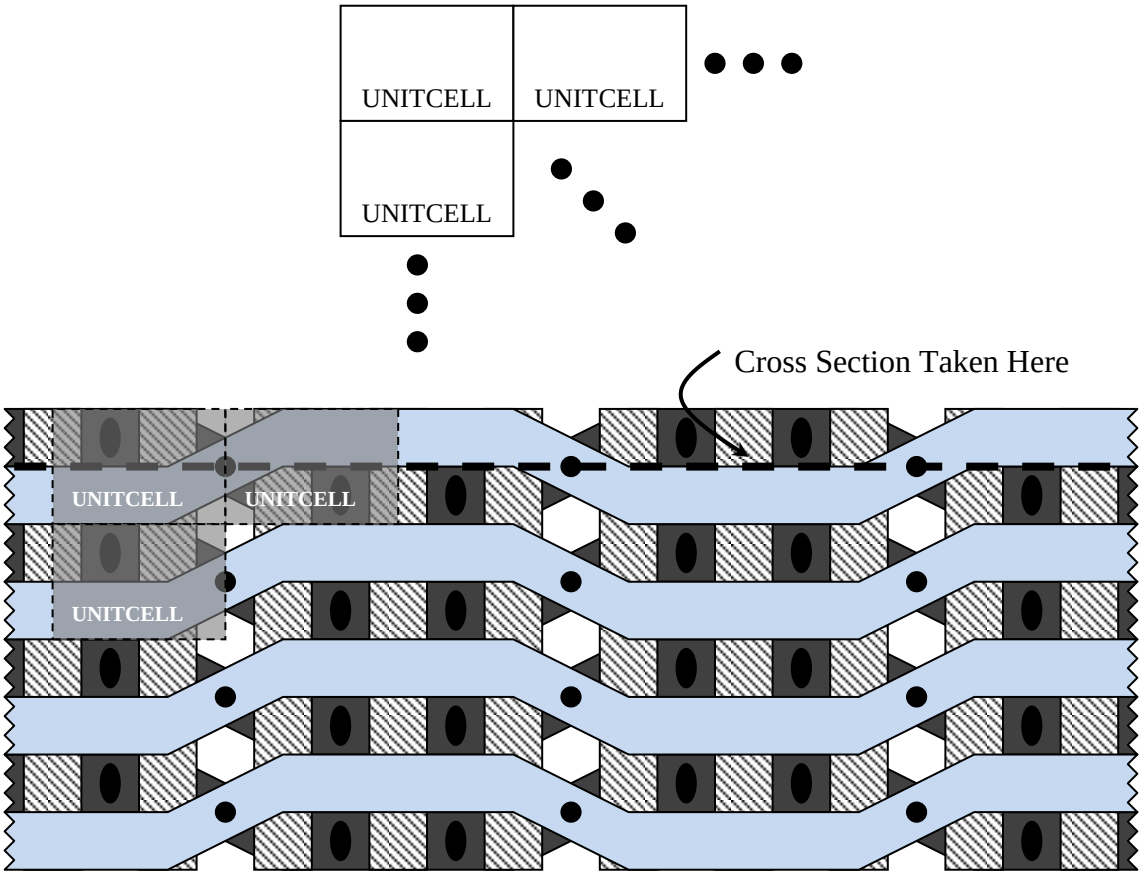


Figure 4.6 Developing a 256 kb memory array using the memory bit.

Using a scanning electron microscope it is possible to view a cross section of the 256 kb memory array at the cross sectional line drawn in Figure 4.6. The cross sectional view seen in Figure 4.7 shows the cross section of the memory array when the cross section is taken perpendicular to the wordlines.

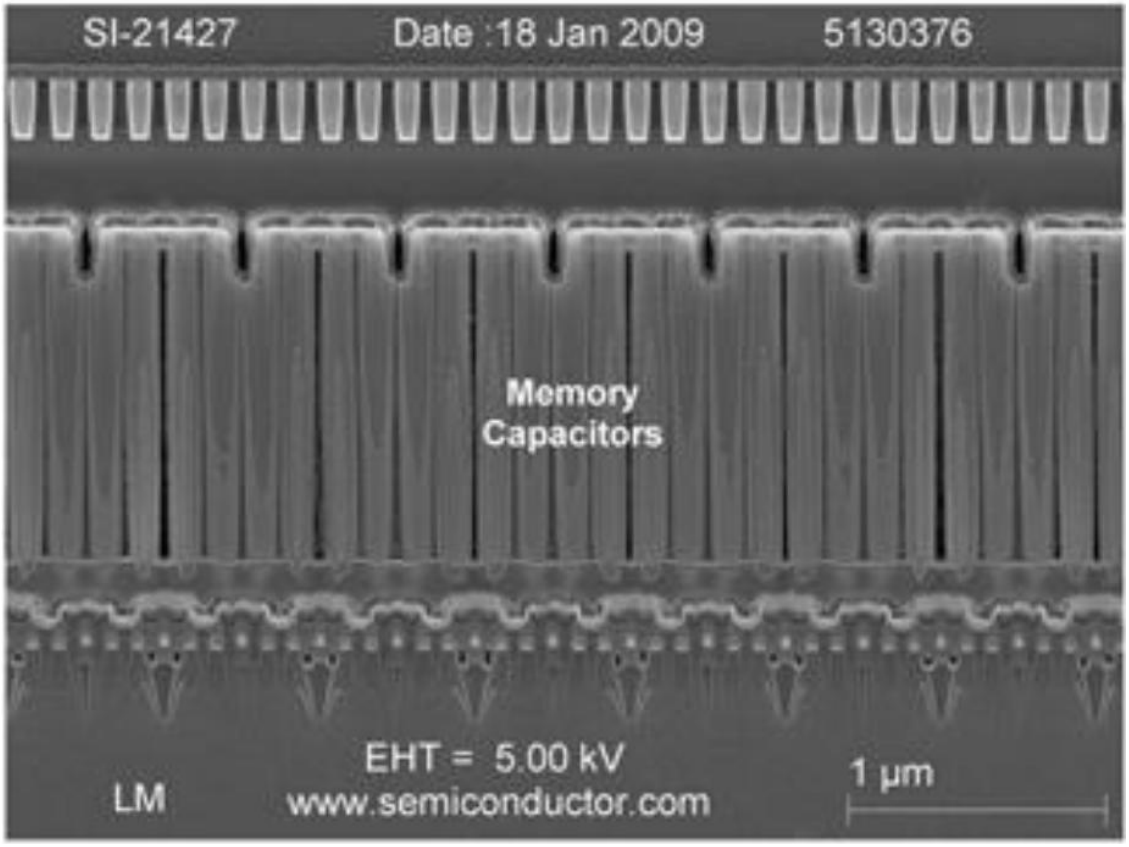


Figure 4.7 Cross section scanning electron microscope picture of a DRAM memory array [26]. From top to bottom: metal 1, memory capacitors, bitline, wordline, substrate.

4.1.1 Memory Array

DRAM manufacturers define their metal layers starting at the tungsten bitline which is often referred to as metal 0 (M0). A two metal layer DRAM chip contains a tungsten metal 0 and two metal layers above the capacitor (metal 1 and metal 2). Current DRAM manufacturing processes are transitioning from aluminum metal layers to copper metal layers [26].

DRAM densities of 1 Gb and below traditionally use two levels of metal above the capacitor, and 2G bit densities are expected to use three levels of metal [28]. This

specification requires that the metal 1 layer have a tight pitch in the array. This requires DRAM manufacturers to employ an aspect ratio of three or higher for the metal 1 layer.

The memory capacitor sits below the metal 1 layer in Figure 4.7. A large amount of work is invested into developing a memory cell that has a value of 20 to 30 fF value. To continuously keep this capacitor value constant as technologies shrink, new capacitor structures are used to increase the surface area of the parallel plates. Currently a cylindrical structure is used to create a large surface area for the capacitor.

The bitline used by DRAM manufacturers is tungsten metal with three times the resistivity of copper. The tungsten bitline is also referred to as metal 0 when used in the periphery of the DRAM chip. A metal 0 to metal 1 contact has a large aspect ratio to reliably connect metal 0 to metal 1. This aspect ratio is a limiting factor for how tall the memory cell can be made.

The cross sectional view of the memory cell's wordline can be seen below the bitline in Figure 4.7. The DRAM chip was manufactured in a 50 nm process and has a wordline pitch of 100 µm. As the technology node continues to shrink, the transistor leakage increases due to the reduction in the wordline width. The length of the access device cannot continually shrink with the technology if the refresh rate is to remain constant. The wordline was sunk into the silicon substrate to combat this effect. This type of wordline structure is referred to as a recessed access device and it is used to increase the length of the access transistor.

Figure 4.8 shows the 512 bitline by 512 wordline memory array created by arraying the $6F^2$ memory cell. We can determine the size of the 256 kb memory array because the area of the memory cell is known. Several redundant wordlines and bitlines

are added to the 256 kb memory array which slightly increases the area. The redundant elements are used to map out defective memory bits, thereby increasing the yield of the DRAM chips. The number of redundant elements is determined by the defect density of a process technology.

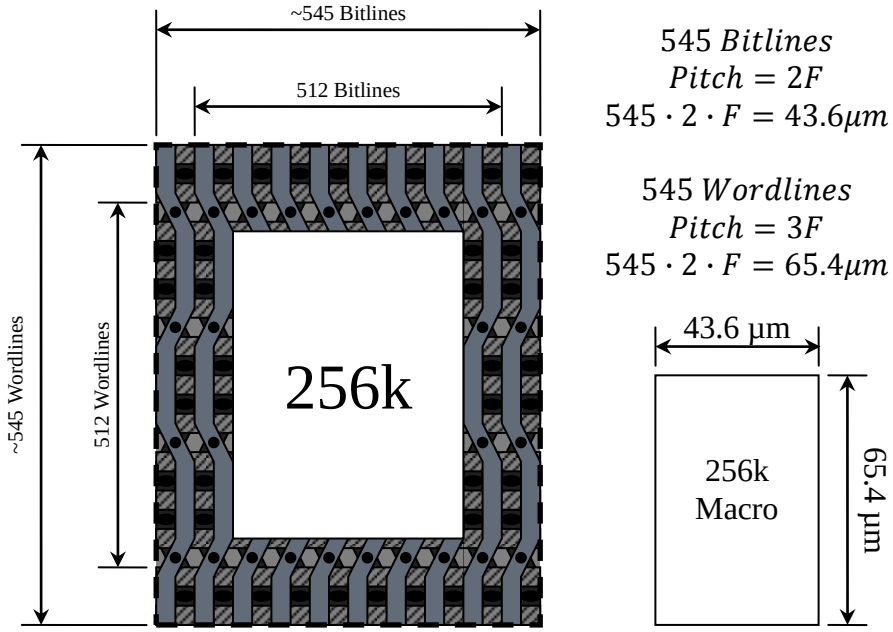


Figure 4.8 Determining the size of a 256 kb memory array.

4.1.2 Periphery Circuitry

Wordline drivers and bitline sense amplifiers (BLSA) are the periphery circuitry used to access the memory bits in the 256 kb memory array. The wordline drivers drive the wordlines high and low during an access. An example CMOS wordline driver is seen in the Figure 4.9.

Each 256 kb memory array requires 512 bitline sense amplifiers and 512 wordline drivers. When 100F space is allocated for the space of the wordline drivers it is possible to get an accurate measurement of the periphery area. Reviewing the $6F^2$ memory bit it is

clear that each wordline spans 3*F*, therefore each wordline driver should span 3*F* as well. Layout tricks can be used to allow the wordline drivers to fit into a larger pitch.

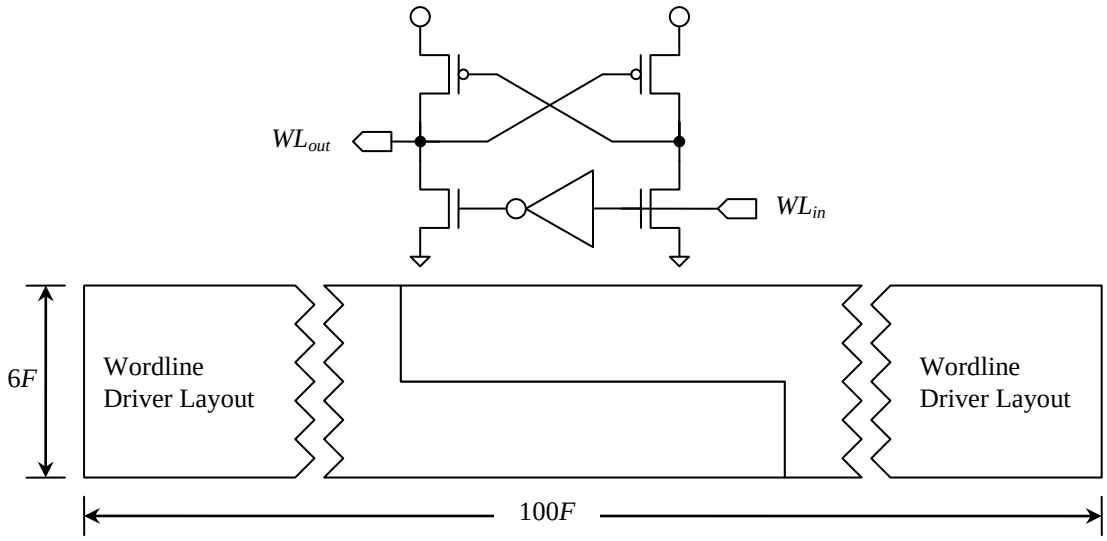


Figure 4.9 Schematic of a CMOS wordline amplifier along with the depiction of the layout tricks required to meet the pitch of the array [29].

A space of 100*F* is allocated for the bitline sense amplifier region of the 256 kb memory array. The $6F^2$ memory cell shows that each bitline spans 2*F*. The bitline sense amplifier uses the bitline and an unused bitline (reference bitline). Therefore, each bitline sense amplifier spans 4*F*. Layout tricks can be used for the bitline sense amplifier that allows each amplifier to sit on a larger pitch. Figure 4.10 shows a schematic representation of the bitline sense amplifier periphery circuitry.

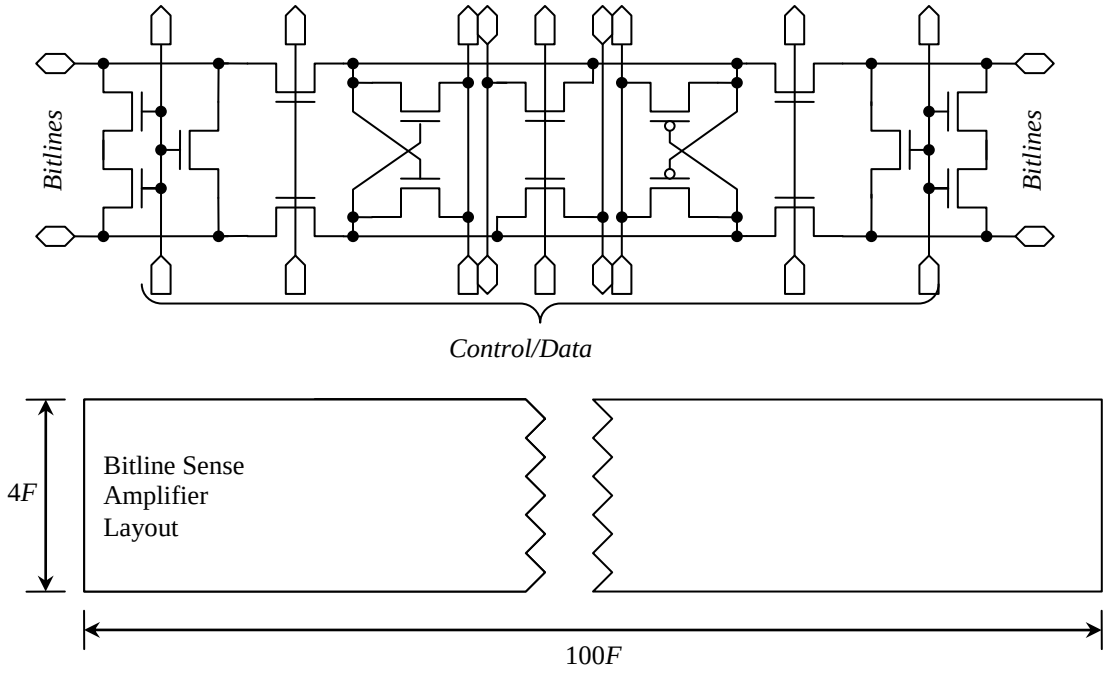


Figure 4.10 Schematic of the bitline sense amplifier along with the space allocation [29].

Using the bitline sense amplifiers in an open memory array allows for a denser memory array compared to the folded architecture [29]. The open array allows for a memory bit to be placed at every cross section of a bitline and wordline. The folded architecture places a memory element at every other crossing of the bitline and wordline. This allows an open architecture to have a more dense memory array.

In comparison to the folded memory architecture, the open architecture has a poorer noise profile. The open architecture does not have the shared activation noise on the bitline and its reference. There are circuit design innovations that reduce the effects of the open architecture noise performance [29].

4.2 Creating a 256 Mb Array

The process of creating a memory array and wrapping it with periphery circuitry is the central theme of creating a memory chip. A 256 Mb memory array is created in Figure 4.11 using the 256 kb memory array.

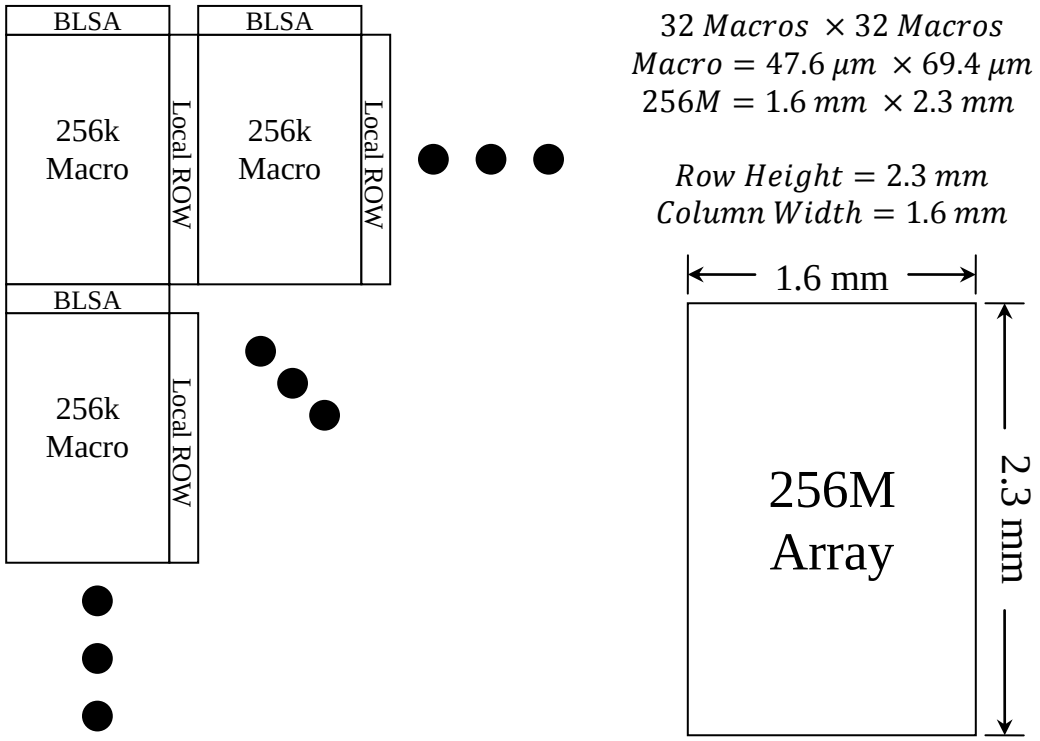


Figure 4.11 Creation of a 256 Mb memory array using 1024 (32 x 32) 256 kb memory arrays along with area estimates.

The area of each block is recorded at creation and allows for an accurate area estimate of the 256 Mb memory array. Figure 4.11 shows the 256 Mb memory array measures 1.6 mm by 2.3 mm. The rounding from 2.22 mm to 2.3 mm is 80 µm. The size of the memory array is exact, but the $100F$ allocation for the periphery circuitry might be a poor approximation. This 80 µm is divided up between 32 bitline sense amplifiers and allows an additional 2.5 µm per bitline sense amplifier. This 2.5 µm is an additional

62.5*F* added to the 100*F* allocated, which makes this approximation fairly useful when determining the size of the memory blocks.

As was done before, global periphery circuitry is added to the 256 Mb array. The global row circuitry is allocated 300 µm (width) and is as tall as the 256 Mb array. The global row circuitry is used to house the redundant row fuses, decode the global row address, create the row and bitline sense amplifier control signals, and drive the master wordline signal. Using 300 µm of space for the global wordline circuitry is a large estimate but will allow for additional space savings if the layout of the block is below 300 µm.

The global column circuitry is used to drive the global data lines, decode the column address, and create the global control signals. Figure 4.12 gives a block level representation of the 256 Mb memory array along with the global periphery circuitry.

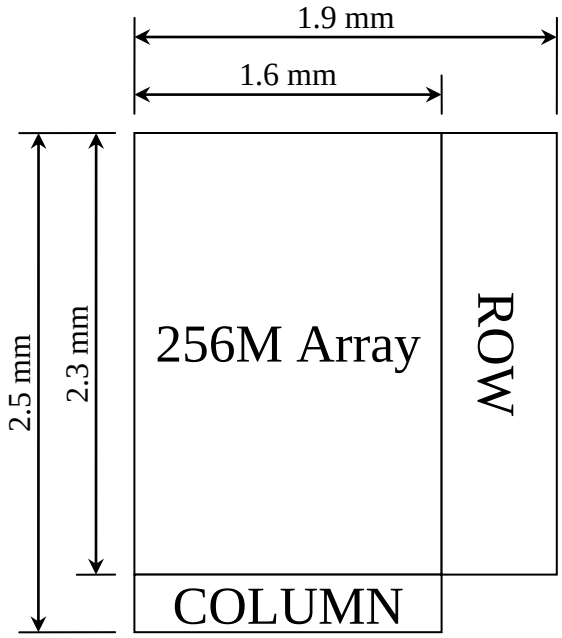


Figure 4.12 Wrapping the 256 Mb memory array with global periphery circuitry and area allocation.

4.3 Creating a 1 Gb Array

Four 256 Mb memory arrays can be used to create a 1 Gb memory array, Figure 4.13. It is possible to create multiple 1 Gb memory arrays by using different column and row structures.

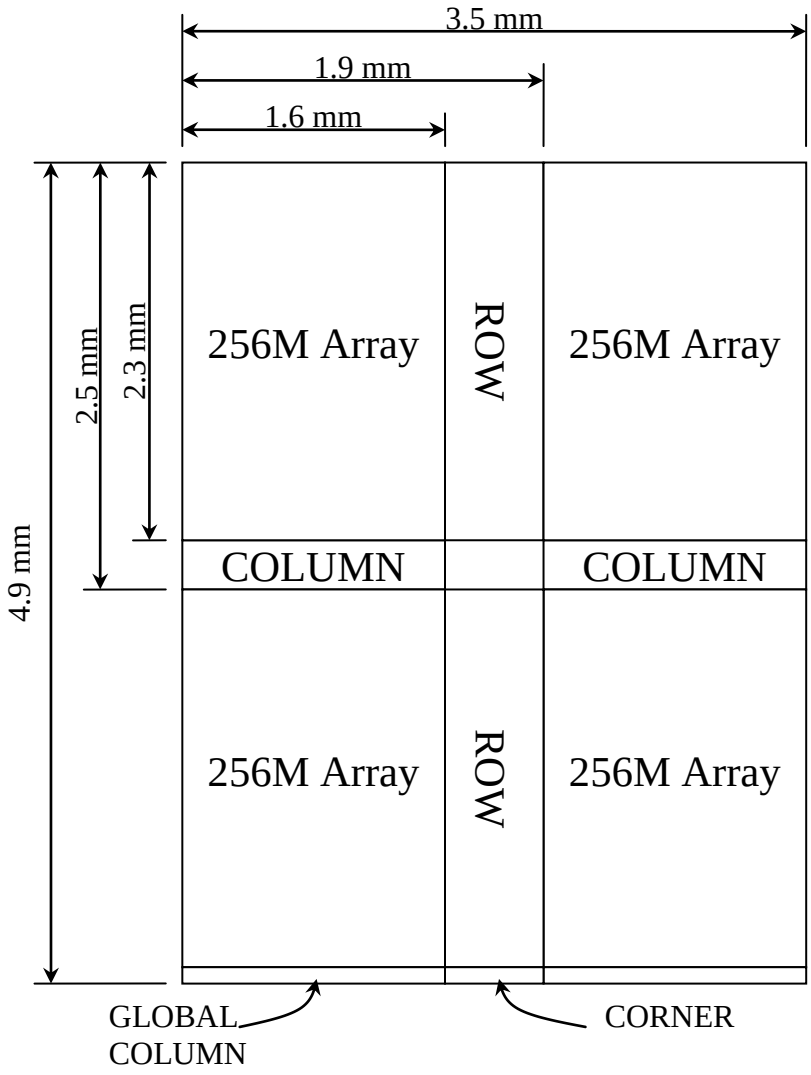


Figure 4.13 Creation of a 1 Gb memory array, addition of periphery circuitry, and area allocations.

The 1 Gb memory array shown in Figure 4.13 measures 3.5 mm by 4.9 mm. An exploded view of the periphery circuitry required to access the array is shown in Figure 4.14.

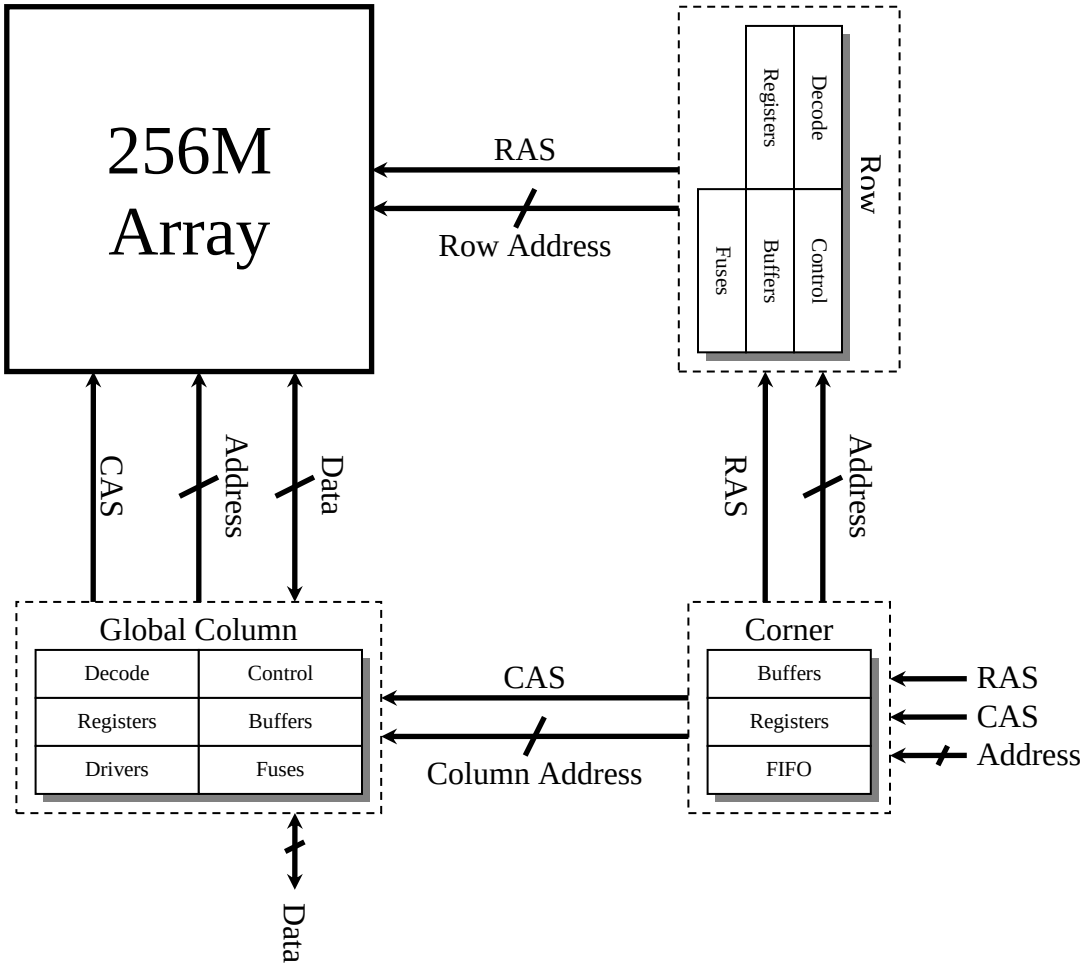


Figure 4.14 Expanded view of the 1 Gb memory bank.

The corner circuitry consists of command buffers, address registers, address buffers, and a write FIFO that can be used for changing the burst length. The global column circuitry houses redundancy fuses, check circuitry, address registers, data drivers, control logic, decode circuitry, and buffers. The row circuitry contains the same components as the global column circuitry except for the data drivers.

4.4 4 Gb DRAM Architecture

Using the 1 Gb memory array along with additional periphery circuitry, it is possible to create the 4 Gb memory architecture seen in Figure 4.15. An accurate chip size and array efficiency estimate is possible when using the area allocations developed in this chapter. Figure 4.15 shows the 4 Gb chip has an array efficiency of 57.7% (1.7% greater than forecast by the ITRS [30]) and a chip size of 71 mm^2. The ITRS road map predicts a chip size of 74 mm^2 for a 40 nm 4 Gb chip in production in 2012. The chip architecture falls within reasonable chip architectures for future DRAM products and emphasizes the feasibility of developing a 4 Gb DRAM architecture.

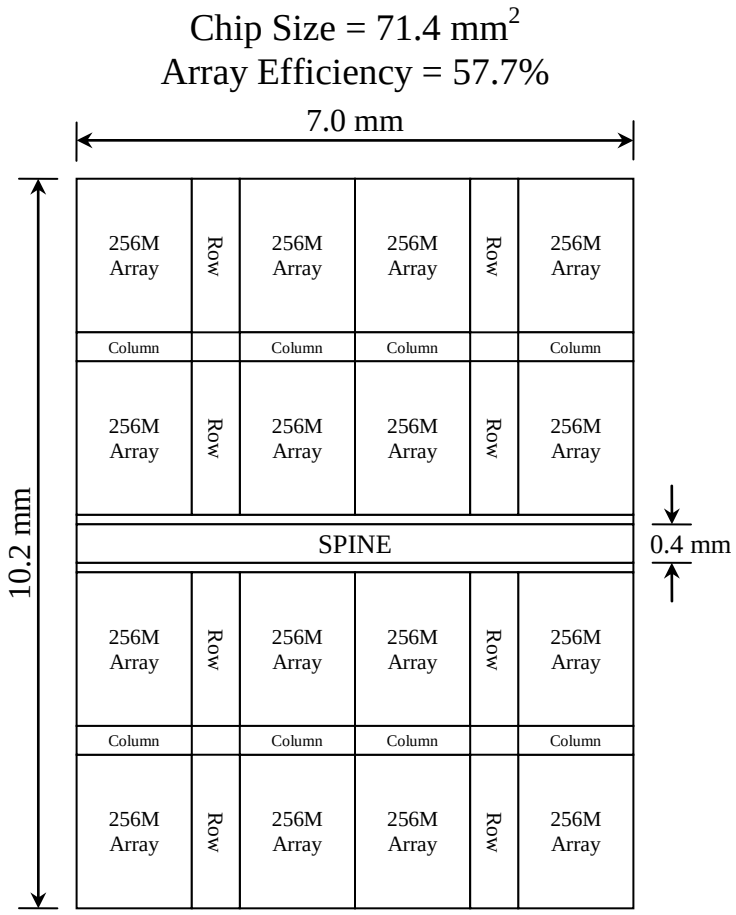


Figure 4.15 Chip size and array efficiency estimates of a 40nm 4 Gb DRAM chip.

4.5 Summary

Using the DRAM trends discussed in chapter 3 along with a $6F^2$ memory cell size, a 4 Gb memory chip architecture was developed with a 40 nm feature size. The predictions made in the architectural decisions allow for a chip architecture that falls in line with the ITRS road map. This chip architecture will allow for the development of a wide I/O DRAM chip architecture that is suitable for a capacitive coupled proximity communication interface.

CHAPTER 5—A PROXIMITY COMMUNICATION DRAM ARCHITECTURE

The market analysis performed in Chapter 3 allowed for understanding of DRAM architectural decisions. Using this understanding, a 4 Gb DRAM chip was produced in Chapter 4 that falls in line with industry projections for a 2012 production product. In this chapter, the 4 Gb DRAM architecture is modified to create a wide I/O DRAM architecture suitable for a capacitive coupled Proximity Communication I/O interface.

The main memory used by personal computers and servers operate with eight data pins. Graphics and embedded memory products operate with a higher data pin count at the expense of increased power consumption. This chapter produces a scalable data pin architecture suitable for proximity communication.

The architecture developed in this chapter should operate with lower energy per bit consumption than main memory DRAM and allow for a scalable I/O count with Proximity Communication.

5.1 Architecture Decision

Using the 4 Gb DRAM architecture as a starting point it is possible to create a wide I/O interface that integrates a proximity communication channel. Deciding on the initial DRAM architecture to use is based on the chip size and array efficiency. The initial requirements for a small chip size and high array efficiency allow cost to direct the decision. The 4 Gb DRAM architecture developed in Chapter 4 is shown in Figure 5.1 and is used as the starting point for developing a wide I/O DRAM architecture.

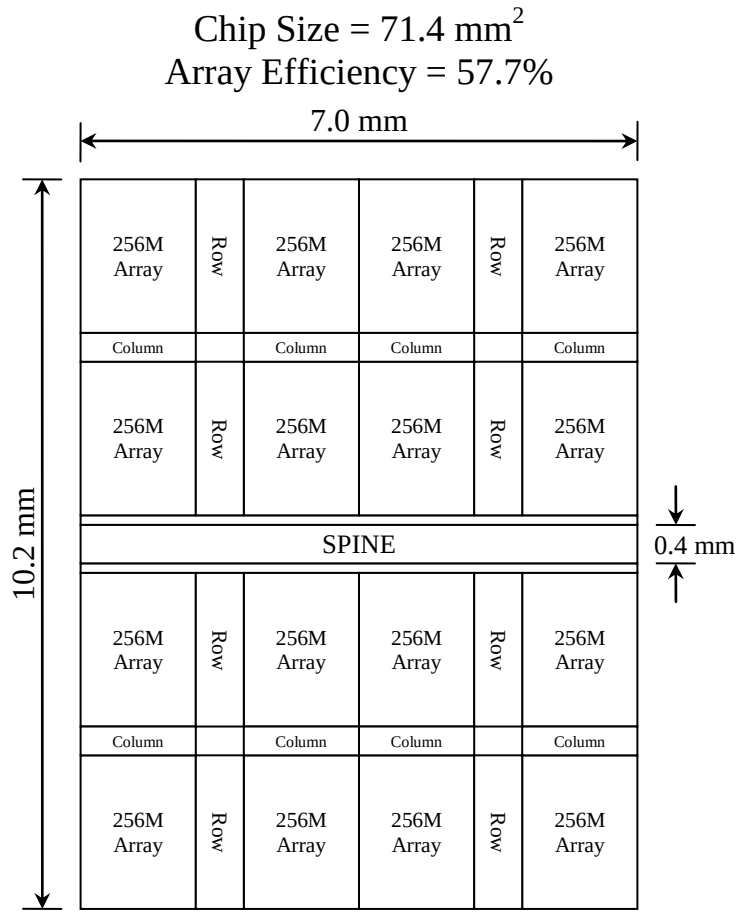


Figure 5.1 The 4 Gb DRAM architecture developed in Chapter 4.

5.1.1 Pad Moving and Centralization

Proximity Communication requires the communication channel to reside on the edge of the chip. Moving the I/O channel to the edge of the DRAM chip is the initial change to the standard DRAM architecture. This allows two chips to be placed face-to-face and enables Proximity Communication. Moving the communication channel to the edge of the DRAM chip creates several interesting challenges when performing an architectural feasibility study. Figure 5.2 shows how the 4 Gb DRAM architecture changes when the SPINE is moved to the edge of the chip.

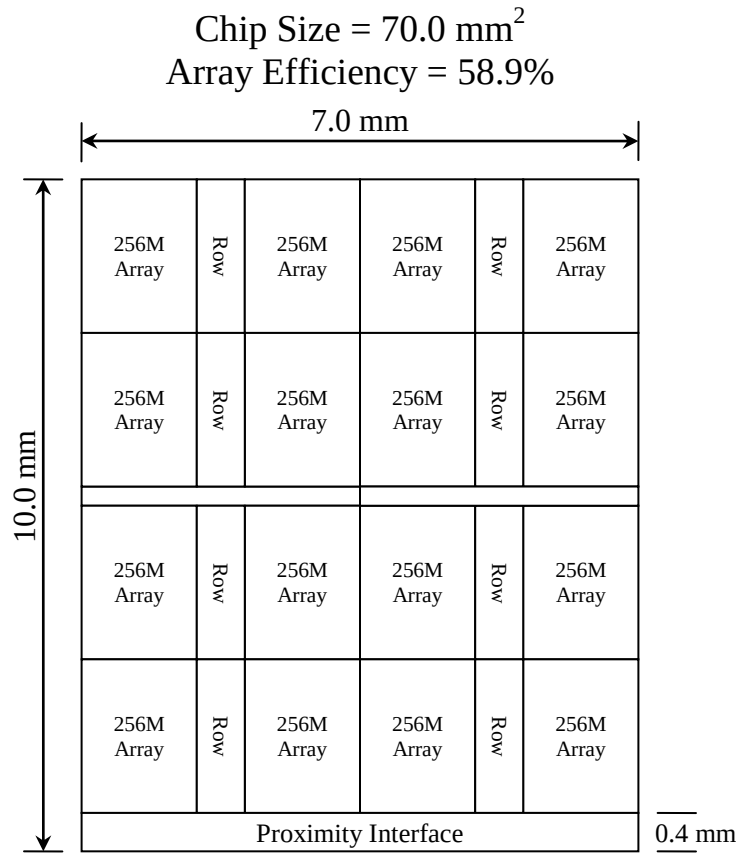


Figure 5.2 Moving the communication channel to the edge of the DRAM chip and reducing the number of column circuitry to reduce the chip size and increase array efficiency.

Figure 5.2 shows that when the SPINE is moved to the edge of the chip, data and command signals will need to be buffered to all memory arrays. Buffering the signals requires the addition of buffers in the middle of the chip. The additional circuitry required in the center of the chip allows the local column circuitry to be moved to the center of the chip. This centralized column scheme can reduce the chip size as seen when comparing Figures 5.1 and 5.2.

The removal of the local column path creates a longer global I/O structure. This will limit the bandwidth of the column path. If we double the number of I/O pins, which is easy to do with the increased I/O density allotted by Proximity Communication, it is

possible to reduce the bandwidth of the column path by 2 and keep the off-chip bandwidth of the DRAM chip constant. This advantage of Proximity Communication allows for a singular global column and row structure in the new DRAM part that is not possible with conventional I/O structures.

Going with a centralized global column and global row structure has the added benefit of increasing the array efficiency and reducing the overall chip size as can be seen in Figure 5.3.

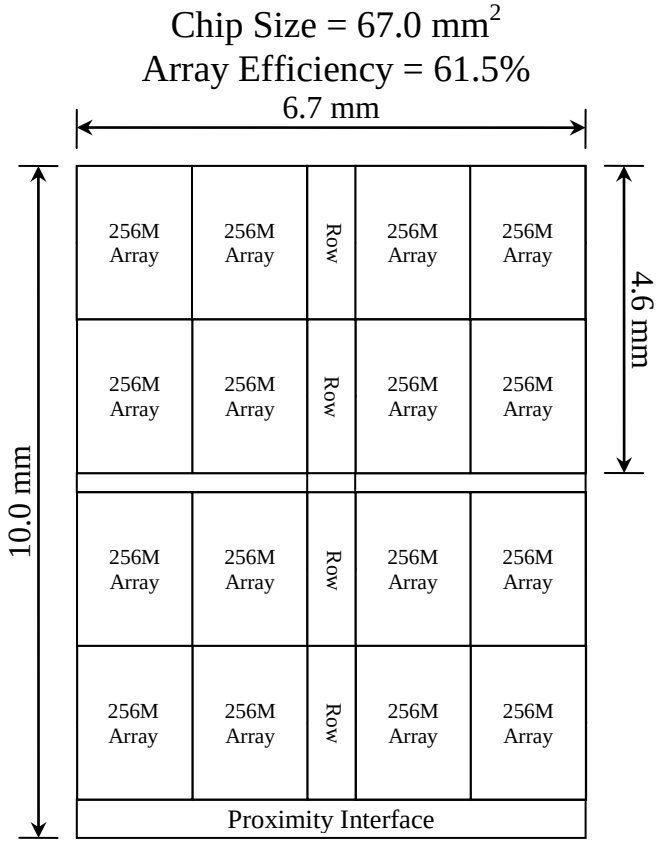


Figure 5.3 A 4 Gb DRAM architecture employing a centralized column and centralized row, this change reduces the chip size and increases the array efficiency.

Developing the initial DRAM architectures enables a feasibility study of architectures that sheds light on the challenges of moving the pad row to the edge of the

die. The requirement of additional buffers to drive signals to the center of the chip is the first challenge. Before the integration of Proximity Communication, the buffers assigned to drive the off-chip wires allocated some power to driving the signals from wire bonds to the center of the DRAM chip. The integration of Proximity Communication changed the wires that drive signals to the center of the DRAM chip from external wires to the on-die wires. Driving on-die wires requires less power than driving off-chip wire bonds. The initial challenge associated with moving the communication channel to the edge of the die can be viewed as a lateral performance change rather than a limiting factor.

The architecture shown in Figure 5.3 uses a modified 256 Mb memory array. The modified array was developed by using the 256 kb structure in a 64 by 16 array rather than the standard 32 by 32 array seen in Figure 5.2. This modified 256 Mb array enables the use of a centralized row structure.

Conventional DRAM chips operate with eight internal memory banks. Eight internal banks enable eight wordlines to be active at once, one in each bank. With eight open wordlines it is possible to perform sequential column accesses to any of the eight banks as long as the accessed address resides on an open wordline. This is a trick used to remove the large row access latency common on DRAM parts.

Figure 5.4 shows the representation of eight internal banks in the newly developed DRAM architecture. Local column path circuitry was added to the architecture to enable a higher memory core operation frequency by reducing the local I/O metal parasitics.

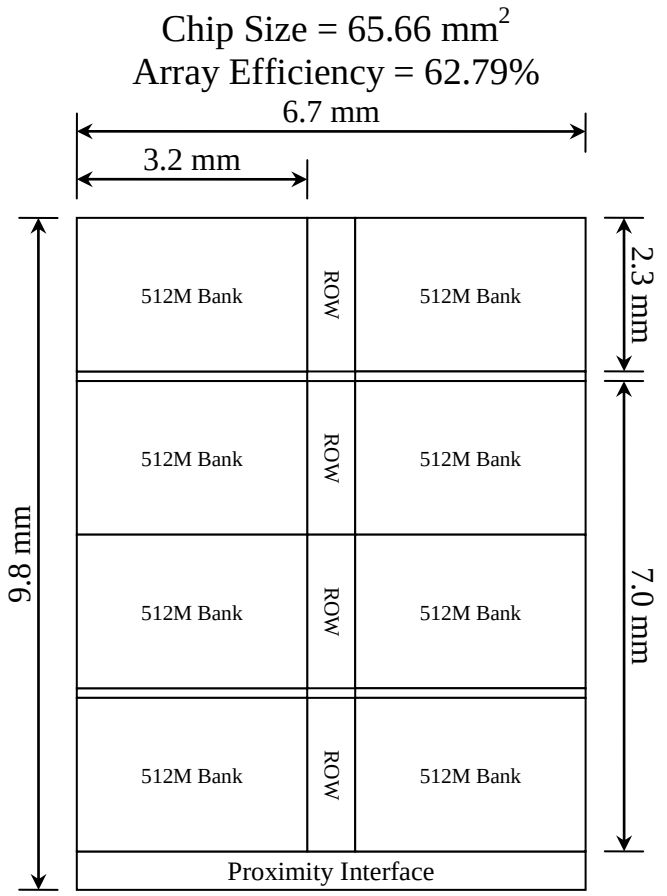


Figure 5.4 Changing the architecture to reflect eight internal memory banks and adding local column circuitry.

The architecture decisions implemented have allowed for the integration of a Proximity Communication channel. Changing to a centralized column and row circuitry has allowed for a 8.0% decrease in chip size and a 8.8% increase in array efficiency. Setting the centralized structures early in the architecture definition allows for enough research and development time to implement this new architecture. To ensure a practical architecture we can allocate additional space for the proximity channel on the edge of the chip to add margin to the architecture, Figure 5.5 shows the initial Proximity Communication enabled architecture.

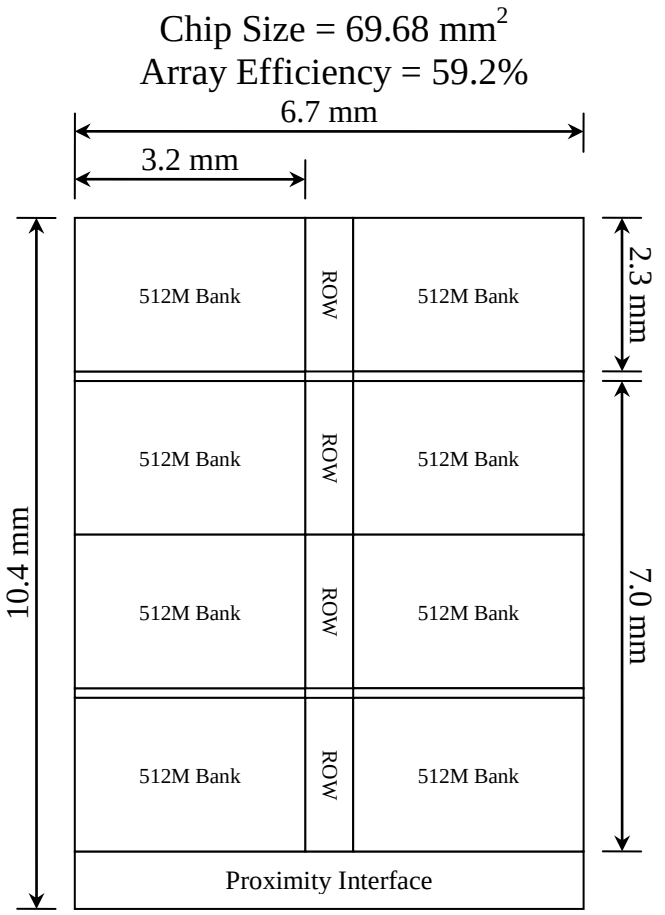


Figure 5.5 Initial Proximity Communication enabled DRAM architecture.

The initial wide I/O architectures allow for an immediate sense of the challenges associated with integrating Proximity Communication on a DRAM chip. The choices made thus far have been with respect to chip size and array efficiency, while the challenges have been in trying to buffer the signals into the DRAM chip. To alleviate some of the initial challenges we can place the Proximity Communication channel on the side of the DRAM chip rather than the bottom. Placing the Proximity Communication channel on the side of the DRAM chip requires an additional review of the 512 Mb bank structure.

5.1.2 512 Mb Bank Structures

When a wordline is activated in a memory bank, a page of data is latched into the bitline sense amplifier. The number of bitline sense amplifiers activated is referred to as a page. Current DRAM devices utilize an 8k page size [31]. The power needed to charge 8k bitline capacitance sets the majority of the power consumption of the DRAM chip. While 8k bitline memory bits are read during a single access, only 64 bits are sent off-chip, which leads to a very low energy efficiency. A wide I/O architecture will increase the number of bits that can be driven off of the chip, greatly increasing the energy efficiency of DRAM products.

A 512 Mb bank can be created by arraying 2048 of the 256 kb memory arrays. Figure 5.6 shows the multiple arrangements possible for creating a 512 Mb memory bank.

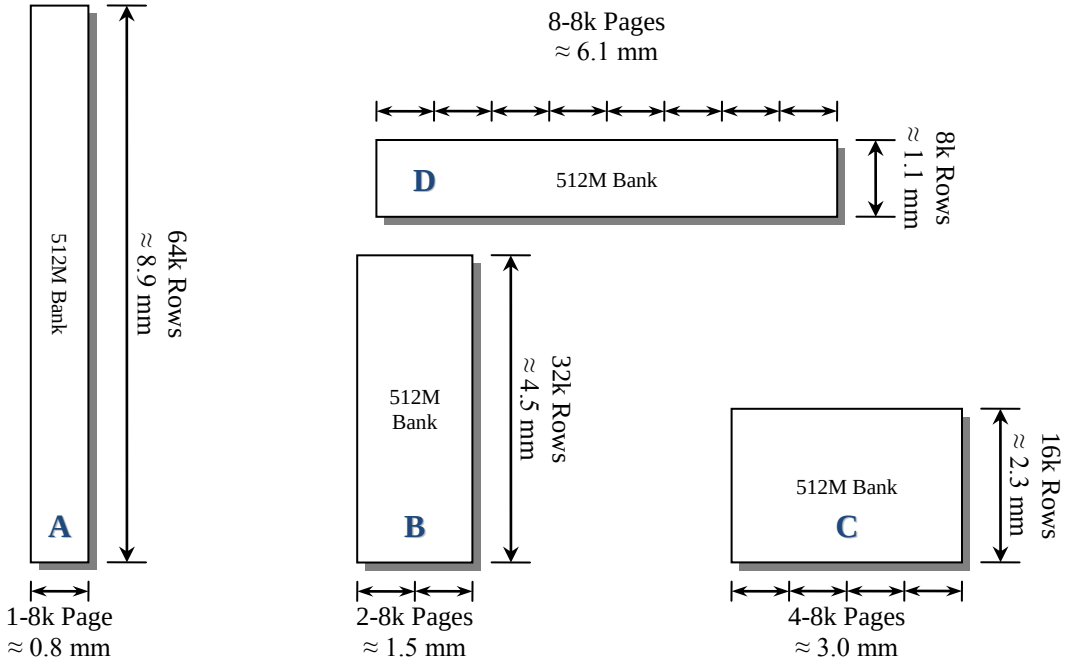


Figure 5.6 Block diagram showing the number of pages available in varying 512 Mb memory arrays along with the physical size of the memory bank.

The 8k page size per memory bank sets the number of possible structures a 512M bit DRAM array can have. Each bank structure in Figure 5.6 has a different level of practicality. The 8k column by 64k row structure is not practical because the local I/O data lines are required to drive ~9 mm to the local column path which houses the I/O re-drivers. Keeping the global I/O metal lines short allows for a higher bandwidth on an open page. For this reason, the 32k column and 64k column structures (C and D in Figure 5.6) are preferred when developing a 512M bit memory array.

5.2 Side Mount Architecture

Using the bank architectures developed in the previous section, several chip level architectures were developed. Chip size and array efficiency were used as the initial selection criteria for the development of a Proximity Communication enabled DRAM architecture. Placing the Proximity Communication channel on the side of the chip rather than the bottom of the chip allowed for a reduction of the global I/O signal lengths.

The initial side mount architecture used the C bank structure shown in Figure 5.6. This allowed for both a centralized column structure and two global row structures. This enables an increase in the bandwidth of the column circuitry due to the reduced metal lengths required to drive the global column signals compared to the original Proximity Communication enabled architectures.

The chip size of the initial side mount architecture can be further reduced by going with a centralized global column structure. The increase in global column and global row parasitics is a trade-off to reducing the chip size and increasing the array efficiency. The DRAM architecture in Figure 5.7 is used as the basis for discussing

further challenges that impact the incorporation of Proximity Communication into DRAM architectures.

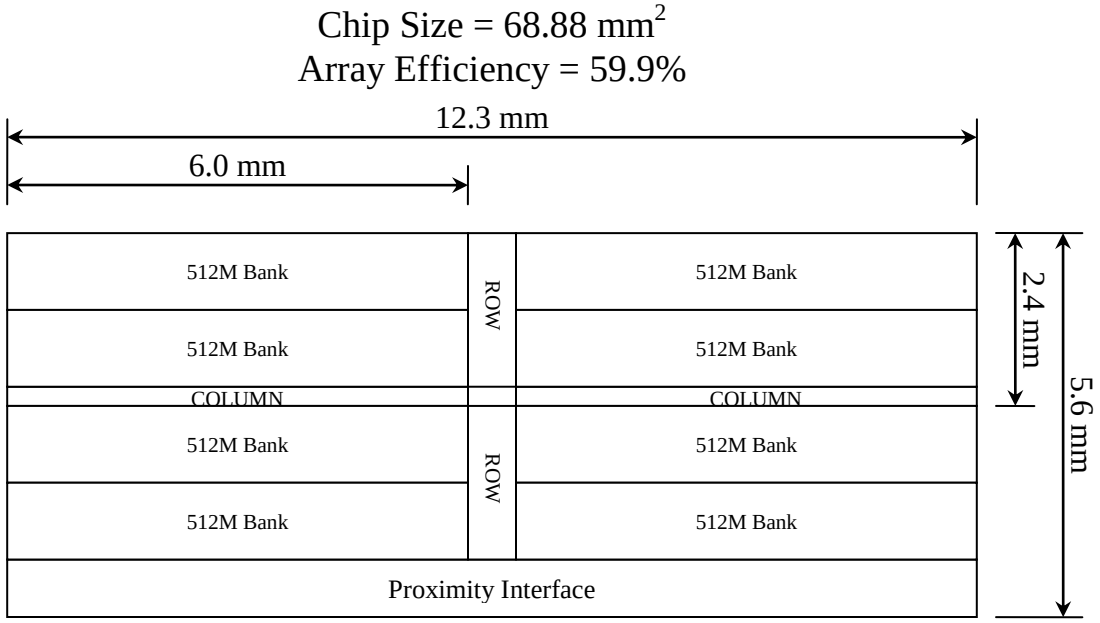


Figure 5.7 A 4 Gb DRAM architecture incorporating Proximity Communication and centralized row and column circuitry.

The DRAM architecture in Figure 5.7 uses a centralized global column and row structure that enables a reduction in chip size. The ITRS prediction for a 2012 40nm 4 Gb DRAM part is 74 mm^2 with an array efficiency of 56% [30]. The side-mount DRAM architecture in Figure 5.8 has a chip size of 68.88 mm^2 and an array efficiency of 59.9%, which falls in line with the ITRS predictions.

DRAM manufacturers have projected the use of three metal layers when the density increases to 2G bit [28]. The 4 Gb DRAM architecture depicted in Figure 5.7 can be used with a DRAM process utilizing only two levels of metal above the memory capacitor. The advantage of using fewer levels of metal along with the reduction of chip size compared to the standard 4 Gb DRAM architecture developed in chapter 4 will

greatly reduce the manufacturing cost associated with a wide I/O Proximity Communication enabled DRAM architectures.

5.3 Challenges

There are three major challenges to enabling a DRAM architecture that utilizes an I/O interface with greater than 32 data pins. The number of metal layers above the capacitor, the global I/O routing, and the local I/O routing are the three major challenges. Understanding this complexity requires the consideration of how a wide I/O architecture changes the way the memory array is accessed. Current commodity DRAM products access 64 bits in parallel. A wide I/O DRAM architecture with 64 data pins operating with a burst length of eight, and therefore a pre-fetch of 8n, requires 512 bits to be accessed in parallel. The challenges associated with a wide I/O architecture are centered on the increase in the number of global data pins.

5.3.1 Number of Metal Layers and Global I/O Routing

Creating a 4 Gb wide I/O DRAM architecture that utilizes only two levels of metal above the capacitor requires a few changes. In a two metal DRAM process, the highest level of metal is used for the global I/O routing because the highest level of metal is typically copper and therefore increases the bandwidth of the global I/O routing. While the metal one layer is used for the global wordline circuitry because the extra delay of driving the larger parasitics has a smaller effect on the total row access latency which is dominated by the wordline parasitics. The current pitch of the metal 2 layer is approximately four times the minimum feature size. In a 40 nm process this pitch is 160 nm.

Dividing the 8k page between two half-banks allows for the reduction of the allocated metal usage for global routing per bank. The 512 data bits accessed from the 8k

page size is split between the two half-banks requiring only 256 bits per half-bank. Figure 5.8 shows the concept of separating the 8k page between two half-banks.

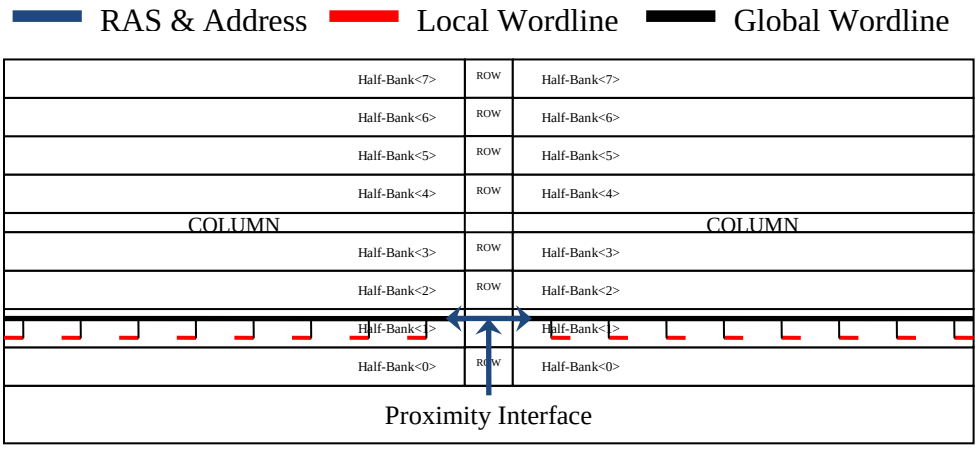


Figure 5.8 Creating half-bank structures and dividing the 8k page between two half-banks.

The architecture developed in the previous section has a half-bank width of approximately 6mm. Increasing the width of the bank, or half-bank, enables more bits to be brought out of the array at once without consuming a large amount of global metal routing. For a 6mm wide half-bank, accessing 256 bits requires a global I/O pitch of 23.4 μm. Each global I/O will have a width of approximately 160nm, due to current DRAM products using four times the feature size for global I/O routing. We can estimate the percent of global routing used for global I/Os as 0.7% of the global routing.

Dividing the banks into two separate half-banks reduces the amount of global metal to approximately 0.7% of the total global metal. The small amount of global routing allows for a possible increase in the amount of global metal used. This enables a possibility of increasing the number of global I/O tracks from 512 to 1024 or possibility higher, enabling a generational approach to a Proximity Communication enabled DRAM architecture.

5.3.2 Local I/O Routing

The 256 Mb half-banks uses 128 256 kb array macros wide and 8 256 kb array macros high. This gives 8 banks, 4k rows, and 64k columns in the 4 Gb memory chip. To distribute the 8K page into two half-banks requires a 4k page size per half-bank. This means that out of the 64k bitlines only 4k fire bitline sense amplifiers fire. To fire only 4k bitlines versus the total 64k bitlines a 16:1 ratio is used to divide the half-bank bitlines. This is accomplished by sending four additional bits with the master wordline to the local wordline drivers. The four bits are decoded at the local wordline driver and perform the 16:1 page decode. Figure 5.9 depicts how the array macros are decoded so that only 4k of the 64k bitline sense amplifiers are fired during a wordline activation.

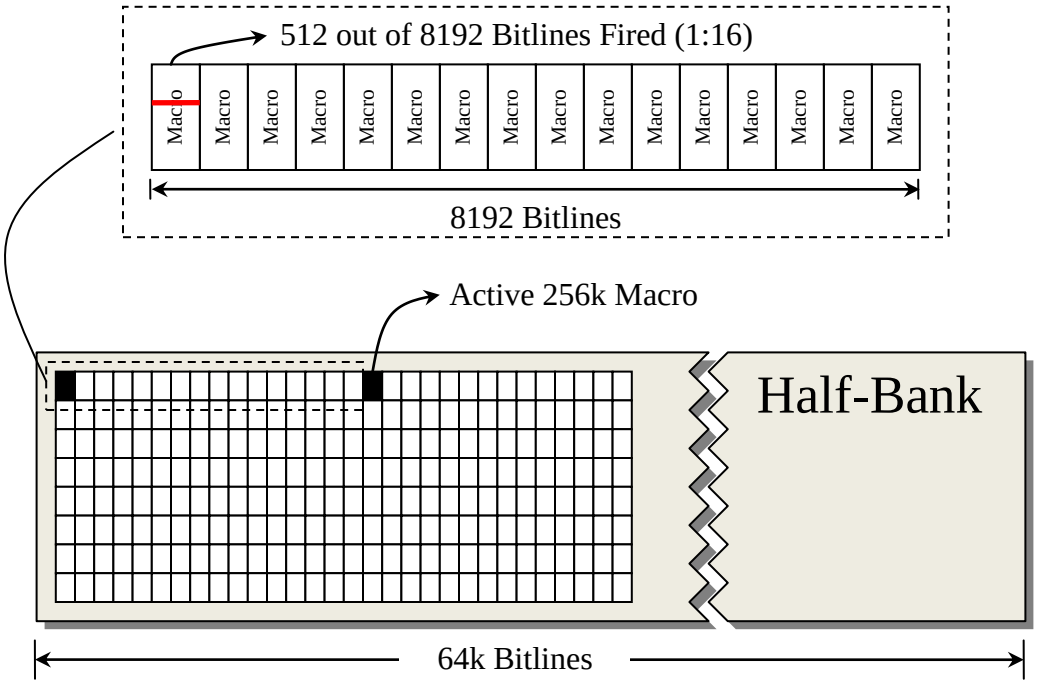


Figure 5.9 Keeping a 4k page size per half-bank requires a 16:1 separation of the 64k bitlines.

One half-bank is responsible for accessing 256 bits. This requires 32 bits to be accessed from each of the activated 256 kb memory macros. This places a challenge on

the local I/O routing due to the limited space and metal available to the local I/O wiring. The 4 Gb DRAM architecture developed in chapter 4 allocated 100F space for the bitline sense amplifier region. The local I/O signals are differential, requiring 64 data lines per bitline sense amplifier (100F). We can reduce this challenge by segmenting the bitline sense amplifiers to above and below the 256 kb memory array. The 32 I/O signals required by the bitline sense amplifier is still a challenge considering only 4 local I/O signals are being used in current DRAM architectures.

In the case of a by 64 architecture Figure 5.10 shows how the data pins are mapped into the local column path and the size of the local I/O routing channels. To keep the global I/O pitch at 23.4 μm, the local I/O signals must be distributed across the half-bank. The 16:1 page decode region is 800 μm wide and is set by the size of 16 256 kb memory macro and the depth of the local wordline drivers. Each of the 32 local I/O signals found in a bitline sense amplifier must span this 800 μm width to keep a 23.4 μm pitch on the global I/O.

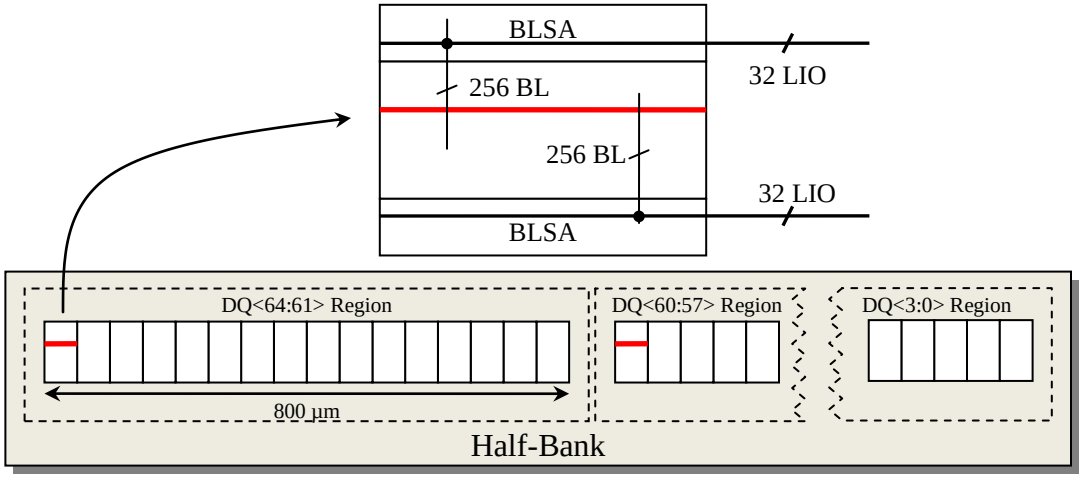


Figure 5.10 Space and data mapping of the local I/O routing within a half-bank.

Typically the local I/O signals are routed in the tungsten metal 0 layer due to no bitlines being used in the bitline sense amplifier region. This places a large parasitic resistance on the local I/O signals which will increase the column cycle time. The increase in the number of local I/O signals, and local I/O trace, creates the second challenge of creating a wide I/O DRAM architecture suitable for Proximity Communication.

To alleviate this local I/O routing challenge, a new global I/O structure can be implemented in the local column path. The new global I/O structure has the ability of allowing the global and local I/O routing to operate at two separate frequencies. Figure 5.11 shows a possible column path structure that can be used to reduce the impact of the local I/O routing challenge.

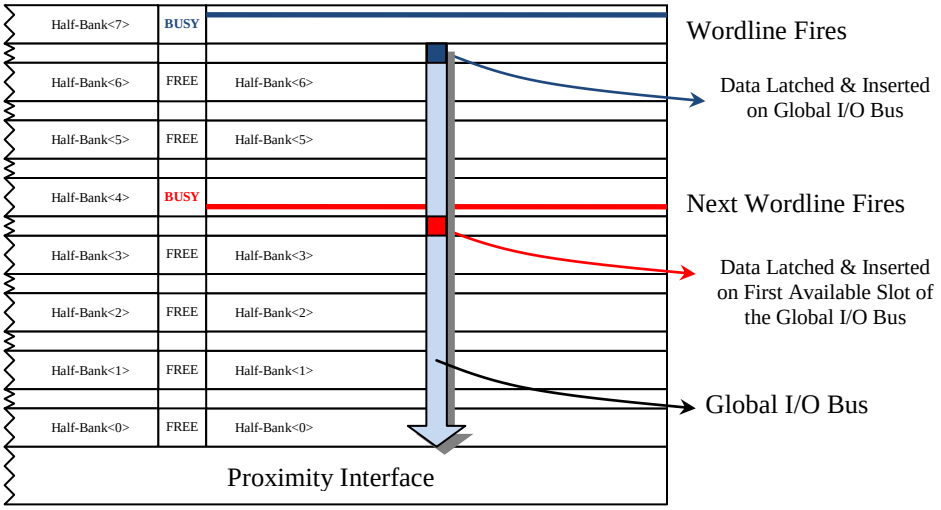


Figure 5.11 The latches in the local column can be used to allow the global I/O signals to operate at higher frequencies and adds another layer of separation between the array parasitics and the column path operation.

Figure 5.11 shows how a new global I/O routing architecture can be used to reduce the number of global I/O routing tracks by implementing a 2:1 or 4:1 serialization

scheme which increases the frequency of the global I/O signals but allows the array to operate at a half or a quarter of the global column frequency.

The new global I/O structure will use current local column circuits with the addition of new control signals to latch local I/O data. The control signals will determine when to place the data onto the global I/O track. This can be performed with a 2:1 or 4:1 serialization scheme. The global I/O lines are able to operate free of array parasitics, which increases their bandwidth and reduces their loading. The local I/O latches will monitor the global I/O bus and determine when an available slot is open for data transmission. Once data is being accepted by the global track, the local latches will update a slot with its stored data.

This may not be necessary for a wide I/O DRAM architecture that uses 64 data pins, but it can find use in subsequent generations that operate with 128 or 256 data pins. A column path protocol can be developed that allows for multiple banks to be accessed and data stored in the local I/O channels. Busy, ready, and data insertion requests can be used to allow the global I/O routing to operate at a higher frequency, while the memory array remains operating at frequencies below 200 MHz.

The wide I/O DRAM architecture will access one half page of memory per half-bank. This is performed by firing a wordline and using additional page decode bits to select one of 16 256 kb memory macros. The unused memory macros will have their bitlines equilibrated to a fixed voltage. It is possible to use the unused bitlines to route local I/O signals through the unused 256 kb memory macros. Figure 5.12 shows how the page decode region can be mapped through unused memory macros and how this can reduce the local I/O routing challenge.

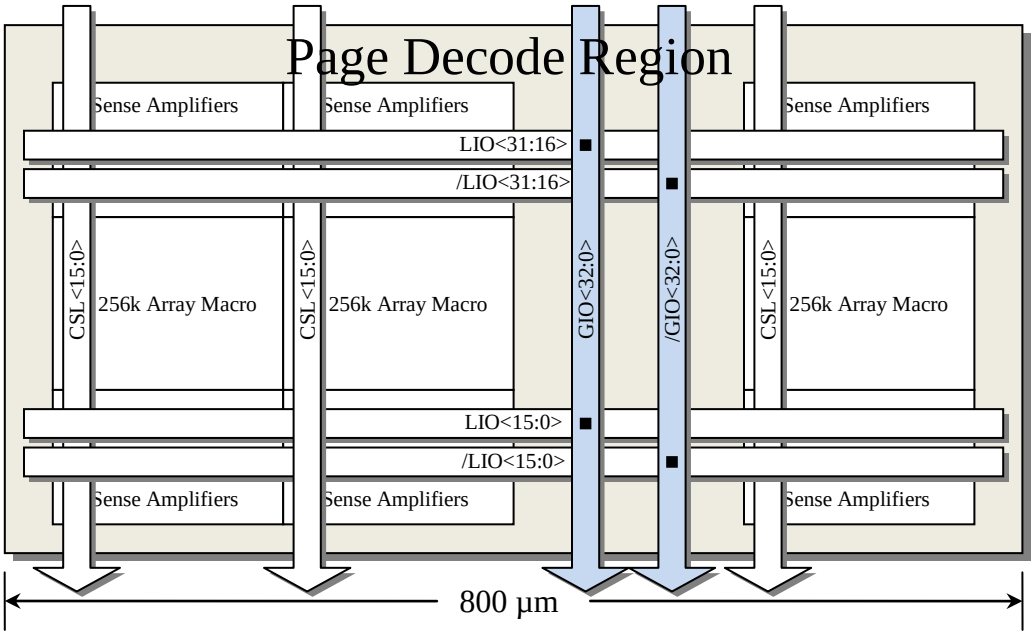


Figure 5.12 Page decode region showing the how the local and global I/O tracks are routed in the array.

5.4 Slice Architecture

The wide I/O DRAM architecture developed in this chapter lends itself well to slice architecture development. The slice architecture is a term used to describe how the building blocks of a chip can be viewed. The architecture can be viewed as a loaf of bread made up of many fairly identical slices of bread. Figure 5.13 shows how the wide I/O architecture developed in this chapter can be sliced up into many identical slices.

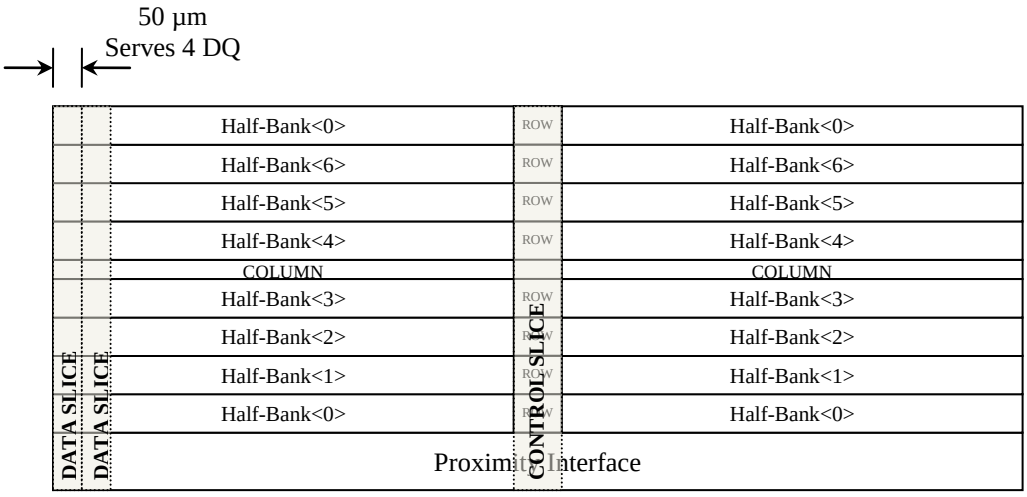


Figure 5.13 Developing slice architectures in the wide I/O DRAM architecture eases the design and verification process allowing for either a shorter design time or an increase in design complexity.

The slice architecture improves the design process by reducing the entire chip into several identical slices. When each slice is treated as its own chip and verified in a full chip manner the design and verification process can be made much simpler. The design and layout engineers need only to design one slice and duplicate that structure many times to create the memory chip. As the I/O density scales with proximity scaling the slice can incorporate more input data pins and increase the serialization circuitry in the data path. Figure 5.14 shows the implementation of a data and control slice depicting power routing, block placement, block size, I/O signals, and control signal routing.

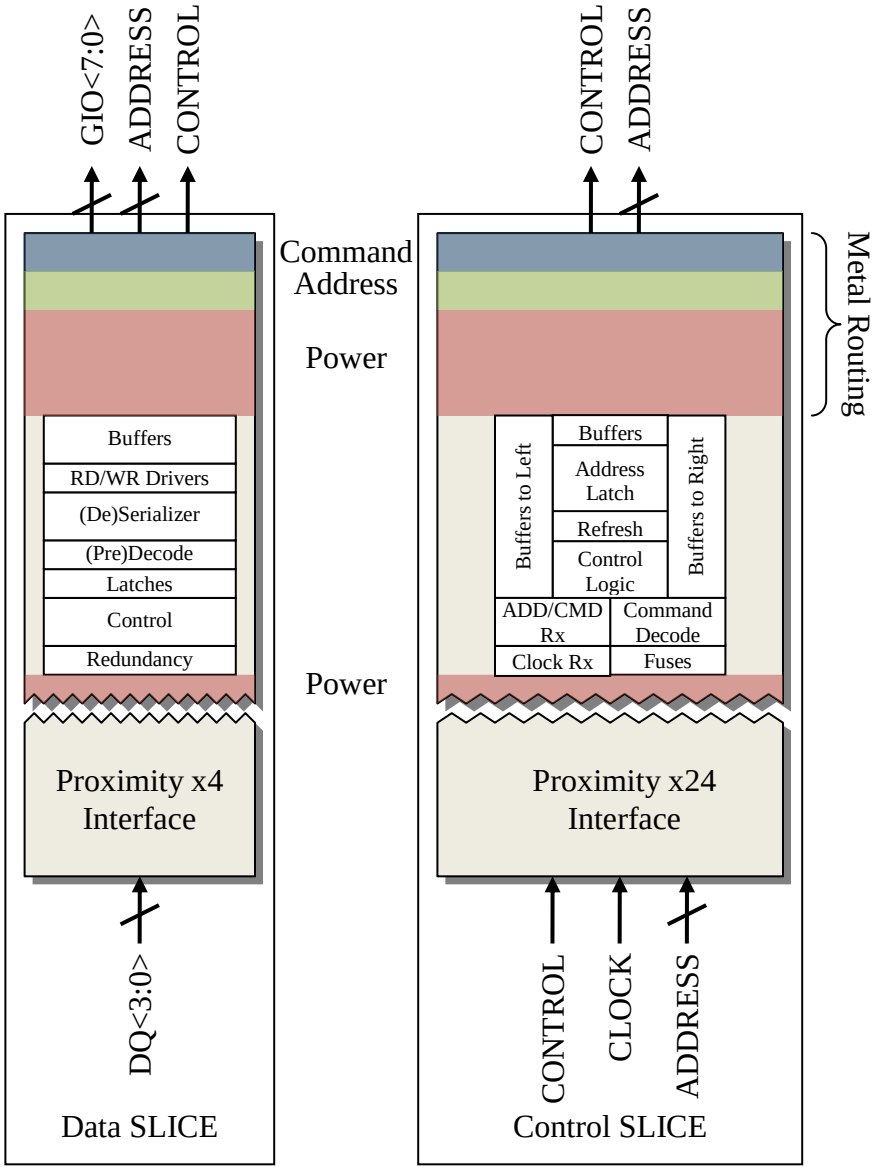


Figure 5.14 Block diagram depicting a Data and Control SLICE used for SLICE architecture.

5.5 Summary

Developing a wide I/O DRAM architecture that is suitable for Proximity Communication requires the communication channel to be moved to the side of the DRAM chip. This enables a Proximity Communication enabled DRAM chip with 8 or 16 data pins that require a limited amount of design changes from current DRAM architectures.

A distributed page and bank structure was developed to enable the possibility of using Proximity Communication with 32 data pins. The developed DRAM architecture placed a page size specification of 8k which allows the array power consumption to remain competitive with current and future DRAM architectures.

Reaching the use of 64 data pins required architectural changes that would not increase the manufacturing cost compared to novel DRAM architectures. Three levels of metal above the memory capacitor is the projection for DRAM densities equal to 2 Gb and above. The wide I/O architecture developed in this thesis allows the metal stack to remain at two levels of metal above the memory capacitor without increasing the chip size. The reduction of projected metal usage enables a significant cost advantage when compared to other DRAM architectures. A new column structure was introduced that will aide in the development of a Proximity Communication enabled DRAM architecture that utilizes 64 data pins.

The wide I/O DRAM architecture utilizing Proximity Communication enables several technological advantages over existing DRAM architectures. Figure 5.16 compares the energy per bit and bandwidth of conventional DRAM modules that use 8 data pins per DRAM chip with the wide I/O architectures developed in this thesis. Fixing the page size and increasing the I/O count through the Proximity Communication wide I/O DRAM architecture allows for an energy efficient DRAM architecture.

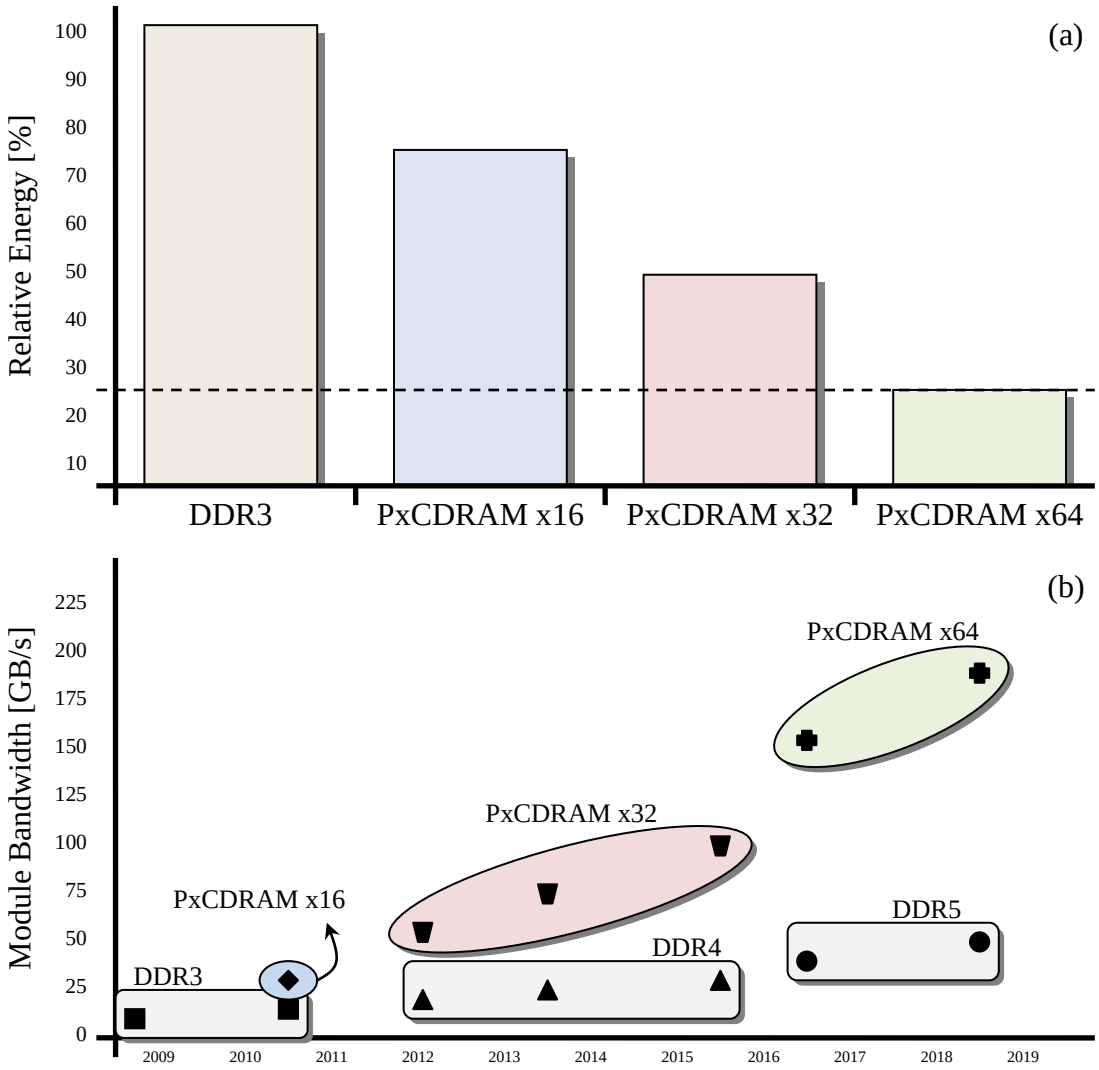


Figure 5.15 Energy per bit (a) and chip bandwidth (b) comparison between current and future DRAM products with the Proximity Communication wide I/O DRAM architectures developed in this Thesis.

Current commodity DRAM chips have poor energy efficiency due to only using 64 data bits of the 8k bits accessed per page. The wide I/O architecture increases the number of bits accessed per page to 512, which significantly increases the energy efficiency of DRAM chips. Figure 5.16 also shows the energy efficiency advantage of Proximity Communication DRAM compared to conventional DRAM architectures.

Although it is possible to only access one Proximity Communication DRAM chip to supply the full 64 bytes of data to the memory controller, it is also possible to increase the amount of data accessed by increasing the memory channel width. The projected bandwidth trends shown in Figure 5.16 clearly show the advantage of using Proximity Communication DRAM over current and future DRAM technologies.

CHAPTER 6—CONCLUSIONS

This thesis proposed a DRAM architecture that incorporates proximity communication to increase the off-chip bandwidth while scaling the number of data pins. Proximity communication allows a wide I/O DRAM architecture to become practical due to the reduced power and increased bandwidth associated with the capacitive communication technology. The research discussed in Chapter 2 showed that proximity communication allows for an increase in I/O density, removal of ESD structures, removal of the resistive termination, and ease of testability. The use of electrical sensors and electrical alignment techniques has enabled proximity communication to become a viable I/O technology.

Developing a DRAM architecture that makes use of this new I/O technology required a thorough analysis of the main memory DRAM market. Chapter 3 reviewed the past, current, and future trends of DRAM products. It was discovered that the selling price per bit decline was the major factor for DRAM architectural decisions. DRAM manufacturers have used transistor scaling to increase the density of their products, resulting in DRAM latency remaining relatively constant compared to processor performance. A wide I/O DRAM architecture that uses proximity communication has the ability to close the performance gap between processors and DRAM.

The information obtained in Chapter 3 was used to develop a 4 Gb DRAM architecture in Chapter 4. Accurate chip size and array efficiency measurements were close to ITRS, and industry, predictions for a 4 Gb chip released in 2012. The

architectural survey discussed in Chapter 4 provided the framework for the development of a wide I/O DRAM architecture that incorporated proximity communication.

Analyzing DRAM architectures using 16, 32, and 64 data pins through the use of proximity communication was performed in Chapter 5. The challenges of creating a wide I/O architecture were found to be in the global and local I/O routing. While a 16 pin architecture utilizing proximity communication required little architecture changes, 32 and 64 data pins required several innovative architectural decisions. The correct use of a split bank and split page allowed for the implementation of 32 and 64 architectures.

A wide I/O DRAM architecture utilizing proximity communication can be used to increase the bandwidth and reduce the power associated with DRAM products. This is done by allowing one chip to supply the full 64 bytes of data required by the memory controller. This is in stark contrast to current architectures which require eight chips to supply the 64 bytes. In the same way, using two wide I/O DRAM architectures can double your bandwidth and reduce the power associated with a main memory access.

Future work includes the production of a memory chip that incorporates several of the ideas proposed in this thesis. A high density memory chip utilizing a wide I/O architecture will be used to investigate the challenges with silicon data. The development of a new column path structure will include novel circuit design techniques that allow for a high bandwidth global I/O structure. The slice architecture theory will be developed with a modular approach. The slice architecture enables the possibility of creating modular DRAM chips that provide just enough memory for the application. This work will allow a wide I/O DRAM architecture to become a viable technology that increases memory bandwidth, and reduces memory power.

BIBLIOGRAPHY

[1] R. Drost, R. Hopkins, I. Sutherland, "Proximity Communication," *Proceedings of the IEEE 2003 Custom Integrated Circuits Conference*, vol. 39, issue 9, pp. 469-472, September 2003.

[2] D. Salzman, T. Knight, "Capacitively Coupled Multichip Modules," *Multichip Module Conference Proceedings*, pp. 487-494, April 1994.

[3] T. Kuroda, "Wireless Proximity Communications for 3D System Integration," *IEEE International Workshop on Radio-Frequency Integration Technology*, pp. 21-25, December 2007.

[4] R. Drost, R. Ho, R. Hopkins, I. Sutherland, "Electronic Alignment for Proximity Communication," *IEEE International Solid State Circuits Conference*, vol. 1, pp. 144-145, February 2004.

[5] D. Hopkins, A. Chow, R. Bosnyak, J. Ebergen, S. Fairbanks, J. Gainsley, R. Ho, J. Lexau, F. Liu, T. Ono, J. Schauer, I. Sutherland, R. Drost, "Circuit Techniques to Enable 430Gb/s/mm^2 Proximity Communication," *IEEE International Solid State Circuits Conference*, pp. 368-369, pp. 609, February 2007.

[6] A. Chow, D. Hopkins, R. Ho, R. Drost, "Measuring 6D Chip Alignment in Multi-Chip Packages," *Proceedings of IEEE Sensors*, pp. 1307-1310, October 2007.

[7] J. Hennessy, D. Patterson, *Computer Architecture A Quantitative Approach*, *4th ed.*, Morgan Kaufmann Publishers, San Francisco, 2007. ISBN 978-0-12-370490-0

[8] G. Moore, "Cramming more components onto integrated circuits," *Electronics Magazine*, pp. 4-6, April 1965.

[9] D. Klein, "The Future of Memory and Storage: Closing the Gap," *Microsoft WinHEC 2007*, May 2007.

[10] B. Pang, Caris & Company http://www.semi.org/cms/groups/public/documents/web_content/p043628.pdf, March 2008.

[11] K. Kim, G. Jeong, "Memory Technologies for sub-40nm Node," *IEEE International Electron Device Meeting*, pp. 27-30, December 2007.

[12] J. Burnim, "On the Scaling of Electronic Charge-Storing Memory Down to the Size of Molecules," *The MITRE Corporation*, November 2001.

[13] Y. Park, S. Lee, J.W. Lee, J.Y. Lee, S. Han, E. Lee, S. Kim, J. Han, J. Sung, Y. Cho, J. Jun, D. Lee, K. Kim, D. Kim, S. Yang, B. Song, Y. Sung, H. Byun, W. Yang, K. Lee, S. Park, C. Hwang, T. Chung, W. Lee, "Fully Integrated 56 nm DRAM Technology for 1Gb DRAM," *IEEE Symposium on VLSI Technology*, pp. 190-191, June 2007.

[14] D. Rhosen, "The Evolution of DDR," *VIA Technology Forum*, 2005.

[15] SUN Microsystems, "SUN SPARC Enterprise T5120, T5220, T5140, T5240, Server Architecture," http://www.sun.com/servers/coolthreads/t5140/wp.pdf, April 2008.

[16] Micron Technology Inc., "TN-41-01: Calculating Memory System Power for DDR3 Introduction," http://www.micron.com/support/part_info/powercalc.aspx, 2007.

[17] Micron Technology Inc. Various Datasheets: http://www.micron.com/products/dram/

[18] Rambus, "Challenges and Solutions for Future Main Memory," http://www.rambus.com/assets/documents/products/future_main_memory_whitepaper.pdf, May 2009.

[19] P. Chiang, M. Fung, "Dual-edge extended data out memory," US PATENT 5,950,223, September 1999.

[20] R. Barth, "2007 Test and Test Equipment," *2007 ITRS December Conference*, December 2007.

[21] H. Fujisawa, M. Nakamura, Y. Takai, Y. Koshikawa, T. Matano, S. Narui, N. Usuki, C. Dono, S. Miyatake, M. Morino, K. Arai, S. Kubouchi, I. Fujii, H. Yoko, T. Adachi, "1.8-V 800-Mb/s/pin DDR2 and 2.5-V 400-Mb/s/pin DDR1 Compatibly Designed 1Gb SDRAM With Dual Clock Input Latch Scheme and Hybrid Multi-Oxide Output Buffer," *IEEE International Solid-State Circuits Conference*, pp. 862-869, April 2005.

[22] C. Yoo, K. Kyung, G. Han, K. Lim, H. Lee, J. Chai, N. Heo, G. Byun, D. Lee, H. Choi, H.C. Choi, C. Kim, S. Cho, "A 1.8 V 700 Mb/s/pin 512 DDR-II SDRAM with on-die termination and off-chip calibration," *IEEE International Solid-State Circuits Conference*," Vol. 1, pp. 312-496, February 2003.

[23] C. Park, H. Chung, Y. Lee, J. Kim, J. Lee, M. Chae, D. Jung, S. Choi, S. Seo, T. Park, J. Shin, J. Cho, S. Lee, K. Kim, J. Lee, C. Kim. S. Cho, "A 512 Mbit, 1.6 Gbps/pin DDR3 SDRAM prototype with C_{IO} minimization and self-calibration techniques," *Symposium on VLSI Circuits*, pp. 370-373, June 2005.

[24] Y. Moon, Y. Cho, H. Lee, B. Jeong, S. Hyun, B. Kim, I. Jeong, S. Seo, J. Shin, S. Choi, H. Song, J. Choi, K. Kyung, Y. Jun, K. Kim, "1.2V 1.6Gb/s 56nm $6F^2$ 4Gb DDR3 SDRAM with hybrid-I/O sense amplifier and segmented sub-array architecture," *IEEE International Solid-State Circuits Conference*, pp. 128-129,129a, February 2009.

[25] F. Fishburn, B. Bush, J. Dale, D. Hwang, R. Lane, T. McDaniel, S. Southwick, R. Turi, H. Wang, L. Tran, "A 78nm $6F^2$ DRAM technology for multigigabit densities," *Symposium on VLSI Technology*, pp. 28-29, June 2004.

[26] C. Wintgens, "The 50-nm DRAM battle rages on: An overview of Micron's technology," http://www.eetimes.com, March 2009.

[27] H. Lee, D. Kim, B. Choi, G. Cho, S. Chung, W. Kim, M. Change, Y. Kim, J. Kim, T. Kim, H. Kim, H. Lee, H. Song, S. Park, J. Kim, S. Hong, S. Park, "Fully integrated and functioned 44nm DRAM technology for 1GB DRAM," *Symposium on VLSI Technology*, pp. 86-87.

[28] K. Kilbuck, "Main Memory Technology Direction," *Microsoft WinHEC 2007*, May 2007.

[29] B. Keeth, R.J. Baker, B. Johnson, F. Lin, *DRAM Circuit Design: Fundamental and High-Speed Topics*, *Second Edition*, Wiley-IEEE, 2008. ISBN 978-0-470-18475-2

[30] International Technology Roadmap for Semiconductor, *2007 Edition*, http://www.itrs.net/Links/2007ITRS/Home2007.htm, 2007.

[31] Samsung Semiconductor Inc. Various Datasheets: http://www.samsung.com/global/business/semiconductor/productList.do?fmly_id=690

[32] J. Handy, "Where Silicon is Headed and Why You Need to Know," Objective Analysis: http://www.media-tech.net/usa-09.html

[33] S. Kadivar, "New Memory Technologies: Evolving Toward Greener Solutions," Samsung Semiconductor Inc.: http://www.samsung.com/us/business/semiconductor/news/downloads/Green_Media_Event_SKadivar.pdf, March 2009.

[34] Hewlett-Packard, "Memory technology evolution: an overview of system memory technologies, technology brief, 8th edition,": http://h20000.www2.hp.com/bc/docs/support/SupportManual/c00256987/c00256987.pdf, April 2009.

[35] T. Jung, "Memory Technology and Solutions Roadmap," *Samsung ANALYST DAY*, 2005.

[36] R.J. Baker, *CMOS: Circuit Design, Layout, and Simulation, Revised Second Edition*, Wiley-IEEE, 2008. ISBN 978-0-470-22941-5

[37] L. Luo, J. Wilson, S. Mick, J. Xu, L. Zhang, P. Franzon, "3 gb/s AC coupled chip-to-chip communication using a low swing pulse receiver," *IEEE Journal of Solid-State Circuits*, vol. 41, Issue:1, pp. 287-296, January 2006.