LOW-POWER, HIGH-BANDWIDTH AND ULTRA-SMALL MEMORY MODULE DESIGN

by

Qawi IbnZayd Harvard

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Electrical and Computer Engineering

Boise State University

May 2011

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE APPROVAL

of the dissertation submitted by

Qawi IbnZayd Harvard

We have read and discussed the dissertation submitted by student Qawi IbnZayd Harvard, and we have also evaluated his presentation and response to questions during the final oral examination. We find that the student has passed the final oral examination, and that the dissertation is satisfactory for a doctoral degree and ready for any final modifications that we may explicitly require.

Date

R. Jacob Baker, Ph.D.
Chair, Supervisory Committee

Date

John Chiasson, Ph.D.
Member, Supervisory Committee

Date

Robert Hay, Ph.D.
Member, Supervisory Committee

Date

Sin Ming Loo, Ph.D.
Member, Supervisory Committee

Date

Pinaki Muzumder, Ph.D.
External Examiner

BOISE STATE UNIVERSITY GRADUATE COLLEGE

FINAL READING APPROVAL

of the dissertation submitted by

Qawi IbnZayd Harvard

To the Graduate College of Boise State University:

I have read the dissertation of Qawi IbnZayd Harvard in its final form and have found that (1) the modifications required by the defense committee are complete; (2) the format, citations, and bibliographic style are consistent and acceptable; (3) the illustrative materials including figures, tables, and charts are in place; and (4) the final manuscript is ready for submission to the Graduate College.

Date

R. Jacob Baker, Ph.D.
Chair, Supervisory Committee

Approved for the Graduate College:

Date

John R. Pelton, Ph.D.
Dean of the Graduate College

ACKNOWLEDGEMENTS

I’d like to thank God for the continued blessings. I want to thank my Mother for her compassion, oldest sister for her support, middle sister for her hospitality, older brother for his companionship, and youngest sister for her encouragement. To my Father, I love you Pops. See you when I get there.

It is important for me to convey a clear message on how influential Dr. Baker has been to my success. Dr. Baker ensured that my idea of success was not a measure of material wealth. Instead, Dr. Baker showed me that success is related to a person’s contribution to their field of expertise and a person’s contribution to society. Thank you, Jake!

I’d also like to acknowledge Dr. Chiasson for being a great educator. Dr. Chiasson spent endless hours helping me understand complex subjects, and he always did it with a smile and with patience. Thank you, John. I’d like to acknowledge Dr. Robert Hay and Dr. Sin Ming Loo for their contributions towards my research. Thanks, guys.

During my journey, I have had the support of many individuals, and I’d like to acknowledge them now: Dr. Pinaki Muzmumder, Dr. Robert Drost, Ahmed Oteafy, Ahsan Hidayatullah, Rizwan Latif, Hamid Partovi, Alex English, Dr. Amy Moll, David Mullner, Giacomo Ghidini, Jennifer Ambrose, Marty Lukes, Dr. Jim Browning, Dr. Bill Knowlton, Dr. Thad Welch, and Angus McDonald.

AUTOBIOGRAPHICAL SKETCH OF THE AUTHOR

Qawi Harvard began his work as an Electrical Engineer as a Junior at Boise State University. He worked with Dr. Jeff Jessing developing silicon bulk micromachine processing techniques used to create surgical blades. As a Senior, Qawi joined Dr. Jake Baker's Analog and Mixed Signal Group to help develop analog to digital converters.

Qawi received his Bachelor's degree in Electrical Engineering from Boise State University, with an emphasis in CMOS circuit design, and a minor in Applied Mathematics, in 2003. That same year he joined Micron as a Product Engineer. He was responsible for moving memory products from conception into production. At Micron, Qawi specialized in CelluarRAM memory products destined for mobile applications.

Qawi left Micron in 2006 to pursue a career in circuit design at Qimonda. At Qimonda, Qawi designed high frequency input receivers, high bandwidth column paths, and clock and data recovery circuits for graphics applications. Qawi left Infineon as a Senior Design Engineer to return to academia in 2009.

During the course of his academic pursuits, Qawi worked as a Memory Consultant for Sun Microsystems Research Laboratory. As a memory consultant, Qawi was responsible for defining next generation memory solutions for server applications. After two successful projects at Sun, Qawi joined a start up in Cupertino, CA to develop advanced embedded memories for SoC applications.

ABSTRACT

The main memory subsystem has become inefficient. The performance gained has come at the expenses of power consumption, capacity, and cost. This dissertation proposes novel module, DRAM, and interconnect architectures in an attempt to alleviate these trends. The proposed architectures utilize low-cost interconnects and packaging innovations to substantially reduce the power, and increase the capacity and bandwidth of the main memory system.

This dissertation develops the theory behind a low-cost packaging technology to create an 8-die and 32-die memory module. The 32-die memory module measures less than 2 cm^3.

This dissertation also proposes a 4 Gb DRAM architecture utilizing 64 data pins to supplement the memory module design. This DRAM architecture is inline with ITRS roadmaps and consumes 50% less power while increasing bandwidth by 100%. The large number of data pins is made possible with the use of a low power capacitive-coupled interconnect.

As part of the capacitive-coupled interconnect, this dissertation proposes a receiver circuit designed for the capacitive interface. The designs were fabricated in 0.5 μm and 65 nm CMOS technologies. The 0.5 μm design operated at 200 Mbps, and consumed less than 3 pJ/bit of energy. While the 65 nm design operated at 4 Gbps, and consumed less than 15 fJ/bit.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE – INTRODUCTION

Main memory architecture has an immediate and future impact on computer system performance. This is true in both server and mobile (laptop, tablet, and smart phone) computer systems. In server platforms, main memory power consumption is approaching that of the central processing unit (CPU), which is the main component of server power consumption. Once main memory consumes the largest amount of power, main memory energy consumption will become application critical. In mobile platforms, size is at a premium and only a small number of memory components can fit in the device. Main memory power consumption and capacity has become a major concern for computer system designers.

This dissertation researches the server's main memory architecture to determine areas where innovation will have the most impact on increased performance. Capacity, bandwidth, power consumption, and cost are the key areas of interest. The dissertation provides information as to why these areas are key over other figures of merit. A new memory architecture is proposed that includes innovation in advanced packaging, interconnects, and die level architecture. These innovations have a direct impact on immediate and future main memory architectures.

2010 and 2011 have seen the popularization of cloud based computing emerging in the mainstream computer market. The way computer programs are accessed and used is being changed to increasingly use mobile devices such as laptops, tablets, and smart phones. We are seeing a cannibalization of desktop towers by laptops, and a cannibalization of laptops by tablets. The growing trend is that computer form factors are getting smaller and there will be a substantial increase in server usage.

2010 and 2011 saw the release of three industry-changing devices. In April 2010, Apple released its iPad. In December 2010, Google released a pilot program giving access to its web-based OS system. In January 2011, Motorola demonstrated their Atrix smart phone at the Consumer Electronics Show (CES). Apple’s iPad successfully penetrated the mainstream market with a tablet-based computing device. Google’s Chrome laptop is the first computer that uses an operating system that is purely cloud based. The Motorola Atrix is the first smart phone to use a dual-core microprocessor. With the use of parallel processing, the dual-core microprocessor will increase the phone’s processing power. This allows it to perform tasks that are typically available to full computers. The Atrix allows users to dock their smart phone into a small dock connected to a monitor, keyboard, and mouse. The phone instantly boots into a Linux operating system (OS) and allows users to continue their work on the Linux desktop.

These devices allow a smaller form factor for the mobile devices because the cloud gives access to higher performing computers over the Internet. Virtual machines appear to be the computing devices of the future. The Atrix is a prime example of this virtual machine usage. The user can dock the Atrix and use the Linux desktop to log into a virtual Windows 7 machine, virtual Linux machine, or a virtual gaming rig. These new devices will require an increase in servers and server performance to support the increase in virtual machines.

Server and mobile applications will benefit from advances in main memory technology. More memory in a mobile application will allow for more local caching of data, and more memory in a server will result in better virtualization performance. This dissertation researches past, present, and future techniques used to create a main memory architecture and highlights inefficiencies while proposing a new architecture.

This dissertation utilizes previous work developed in my Master's thesis. The work is supplemented with additional research developed in this dissertation.

Contributions of the Dissertation

Increases in main memory capacity and the increase in bandwidth in servers are shown to result in main memory systems that consume large amounts of power. This dissertation proposes several architecture changes that can overcome these power issues while increasing capacity and bandwidth.

The discoveries found by performing this work have also led to innovative packaging solutions that can be moved into the mobile space. Mobile computing is largely becoming the mainstream of today's hardware. Mobile applications have power, bandwidth, and capacity limitations; however, the limitations of mobile and server applications have different attributes.

This dissertation also proposes a new architecture and circuits for use in capacitive-coupled interconnects. It is believed that this architecture will improve the chances for its introduction into commercial products. The architecture uses an approach that removes the need for alignment circuits, and creates interconnects with known capacitive values. The known coupling capacitor value greatly simplifies the design of receiver and transmits circuits. A robust receiver design that reduces the energy consumption to 15 fJ/bit at 4 Gbps in 65 nm CMOS is experimentally verified.

Energy consumption per bit transmitted is a parameter used in wireless communication systems. The metric is used because it is easy to transform into a signal to noise ratio. The metric has found its way into the specifications of receiver and transmitter designs not used for wireless communication, and for this reason it is reported in this dissertation. The energy metric is simply the integral over a bit period of the power consumed. In this dissertation, we switch between power consumption in watts to energy consumption in joules per bit.

Finally, this dissertation describes the creation of a memory module that uses inexpensive advanced packaging technologies to solve capacity, bandwidth, and power limitations. The solution allows memory capacity to increase without impacting system level power consumption or limiting the memory channel bandwidth. A 4 Gb low power DRAM architecture is produced that leverages the advantages of high I/O bandwidth to reduce the power consumed. Use of both the DRAM architecture and the new memory module form factor substantially increases performance metrics. The following section details findings of a literature review.

Literature Review

The inception of the multi-chip module has placed a limit on the area available for the integration of multiple chips, with memory taking up to 50% of the available space. Three-dimensional integration became popular in the early 90s as a solution to the space constraints of multi-chip modules. Val and Lemoine proposed a memory cube consisting of 8 SRAM chips in 1990 [1]. The SRAM input/output pads were moved to the edge of the die through the deposition of a thin layer of conducting material. The chips were glued together and stacked eight high with the thin film hanging over all four sides of the cube forming metallic tabs. Val and Lemoine proposed the use of an additional layer of conducting material that could be deposited on each side of the cube to connect the SRAM die together. The innovations proposed by Val and Lemoine led to additional research into the area of cubing memory products. The downfall of these innovations was that the technology was too far ahead of its time.

IBM researchers, Bertin, Perlman, and Shanken used technology developed at Irvine Sensors to manufacture 18 – 20 DRAM chips into a cube [2]. A thin-film was used to transfer the DRAM pads to one edge of the cube. An additional thin-film level was created to connect the cube to a ceramic substrate. The cube technology developed in the early 90s was too far ahead of its time and could not gain popular support without a true need in the computer industry. Therefore, integration issues were not addressed. Cost and performance efficiency were not developed for the technologies, only proof of concept. This, along with the integration of cache onto the same die as the microprocessor, limited the exposure of the technology.

The computer industry is reaching another impasse relating to scaling limits of semiconductor memory products. Researchers are again turning to three-dimensional integration to circumvent the limitation.

Engineers at Samsung have demonstrated an 8 Gb four high DRAM stack. This design addressed the capacity, bandwidth, and power constraints of DRAM that is limiting computer system performance [3]. The design used through silicon vias (TSV) to connect the die together.

TSV integration into the market place will occur once several key milestones are reached. A cost of ownership of $150 per wafer is an aggressive milestone created by a consortium of industry and academic experts working in the 3DIC field [4]. Along with cost, injected substrate noise, area penalty, and height limitations are additional challenges associated with TSV technology.

The semiconductor industry has large amounts of capital behind the development and incorporation of TSV for the three-dimensional integration of silicon chips, so, it is only a matter of time before the technology becomes mainstream [4]. The module developed in this dissertation overcomes the cost issues seen by TSV incorporation by using inexpensive packaging solutions.

Dissertation Organization

Chapter One introduces the reader to several key innovations in server and mobile platforms. These innovations ensure the increase in mobile and server usage.

Chapter Two discusses the problem of power consumption in mobile and server platforms, and directly relates them to system level performance. A breakdown of the power consumption of a server is shown and the impact of main memory power consumption is summarized. The relationship between the main memory power consumption is correlated to performance, bandwidth, and capacity. Chapter Two provides the introduction to the architectures of the server, memory module, and memory architecture necessary to provide a definite problem statement.

Chapter Three discusses a present and historical perspective of three-dimensional integration of multiple memory chips. Through silicon vias are reviewed, and a low cost stacking technology is introduced. The research in these areas is brought together to develop a nano-module. The steps required to develop the nano-module are thoroughly discussed in Chapter Three. The nano-module can be simplified with a memory chip that places its interface on the edge of the die.

Chapter Four summarizes the invention of a memory die that contains a large number of data pins placed on the edge of the die. Several key innovations are introduced in the memory architecture that allow for a substantial bandwidth increase while reducing the power of the memory die. The large number of data pins on the die will consume a substantial amount of power if conventional receivers and transmitters are used.

Chapter Five introduces a low power capacitive-coupled receiver design that is capable of reducing the energy consumed by the receiver to less than 100 fJ/bit at 4 Gbps. Limitations of conventional capacitive-coupled interconnects are highlighted in Chapter Five. A new approach is introduced, simulated, manufactured, and the experimental results of two silicon chips are presented.

Chapter Six summarizes the information found in this dissertation. A direction of future work is also presented in Chapter Six.

CHAPTER TWO – MAIN MEMORY LIMITATIONS

The ubiquitous nature of computer systems has affected how we live our lives. Personal computers (ranging from smart phones to desktops) provide a varying list of abilities to the end user. More often these devices are used to access data contained on the Internet. Accessing the data requires a network that transmits data to and from datacenters. Datacenters provide data storage and compute resources through a network connection. The increase in Internet usage has resulted in an increase in datacenter capabilities [5].

Server Power Consumption

Originally, datacenters were sparse enough that their effect was negligible. As the number of datacenters increased, they began to make a substantial impact on society. One of the major negative contributions of datacenters is their consumption of energy. Energy efficiency of datacenters is affected by the increase in energy costs, increased emissions from electricity generation, and increased strain on existing power grids [5]. All of these considerations are elevated in importance due to the increase in datacenter capacity.

Datacenter’s energy consumption doubles every five years. U.S. estimates show that $4.5 billion dollars were spent in 2006 to supply 61 billion kilowatt-hours (kWh) of electricity (1.5 percent of U.S. consumption) to the nation’s datacenters [5]. This trend has led power consumption to become a major interest to server manufacturers. Figure 2.1 shows the cost of powering and cooling the datacenter approaching the cost of the server hardware [6].

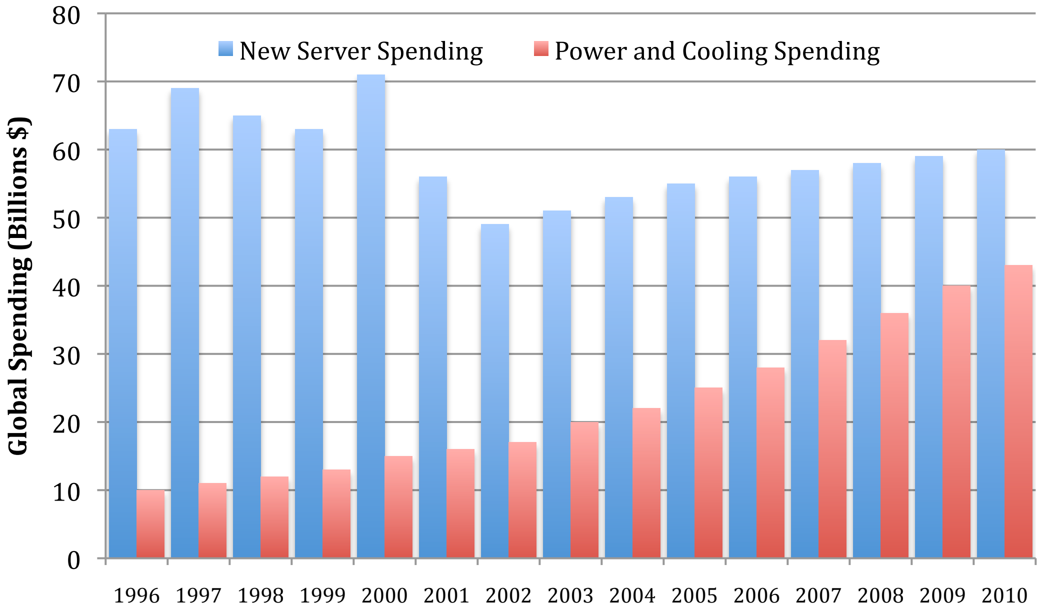


Figure 2.1 – Global Spending for Datacenters

Power reduction can be achieved at various levels of the power supply chain. In this dissertation, we focus on reducing power consumption at the server component level, specifically, at the main memory level. More than 50% of the power consumed by a server is consumed in the central processing units (CPU) and its main memory. Figure 2.2 gives a breakdown of server power from its various components [6].

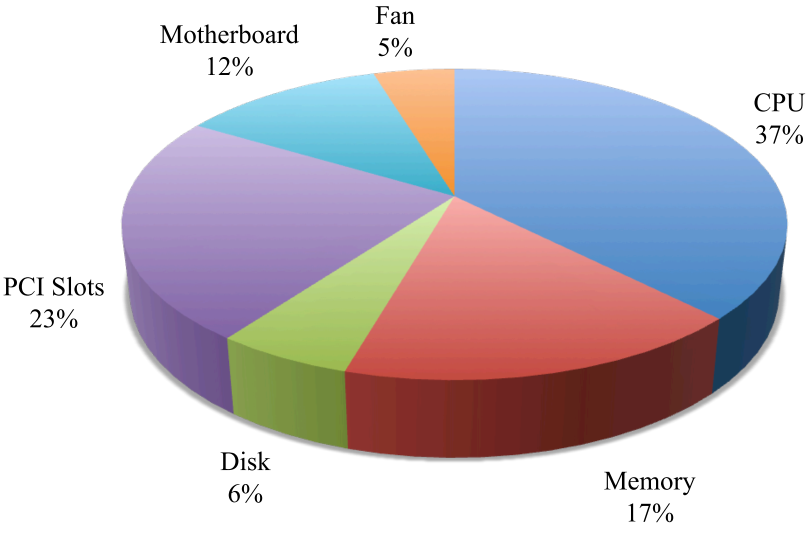


Figure 2.2 – Power Consumption Breakdown of a Server

Power vs. Performance

CPU power consumption was trending up with performance gains over the past few decades. It was only when the power wall was reached that designers began to look critically at power consumption. When the cost of cooling a microprocessor became impractical for commodity processors, designers began to reduce performance gains. Figure 2.3 shows how microprocessor performance scaled at roughly 52% per year from 1986 to 2002 [7]. In 2002 performance scaling reduced to roughly 20% per year. The reduction of performance gains can be attributed directly to power consumption.

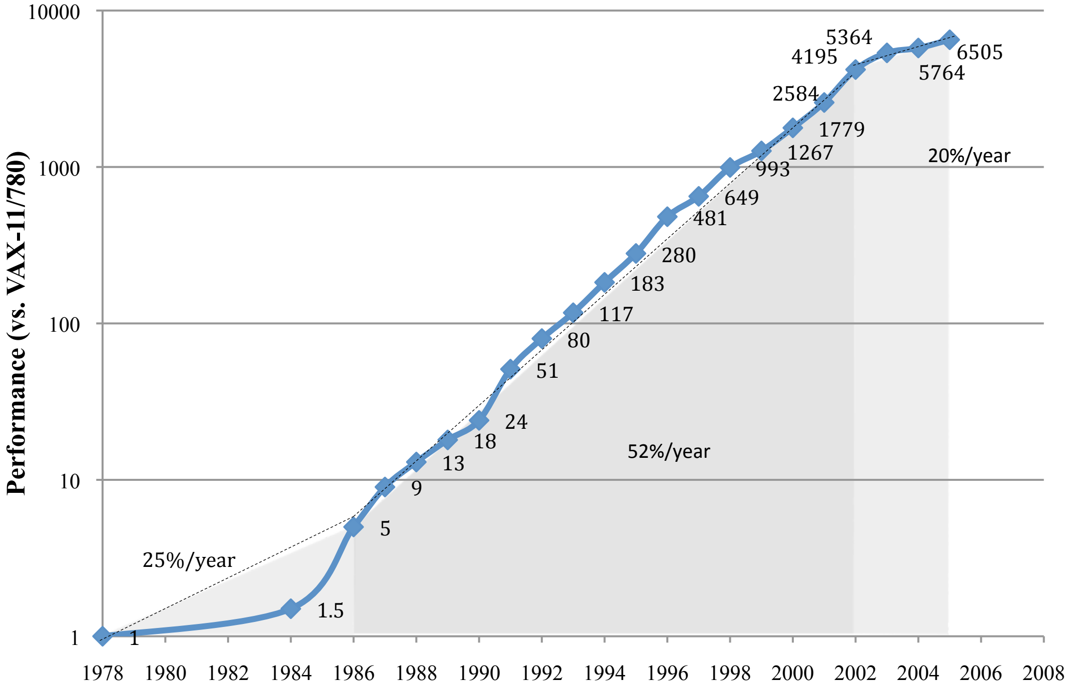


Figure 2.3 – Growth in Processor Performance

In 2004, Intel released the Prescott microprocessor that was a single core processor using a clock frequency of 3.6 GHz. At the time, processor performance was directly related to clock frequency. Processor manufacturers kept increasing their clock frequency to increase performance. This required the use of a larger amount of transistors. The dynamic power consumed by the transistors can be derived from the following formula.

$$P = CV^2 f_{CLK}$$

The above equation is used to determine the power consumed when charging a capacitor at a frequency of f_{CLK}. The equation states that the dynamic power (*P*) consumed when charging a capacitor (*C*) is equal to the capacitance value times the voltage (*V*) times the frequency.

Before the Prescott microprocessor, voltage scaling was used to successfully contain power consumption to an acceptable value. Voltage scaling, in this dissertation, refers to the process of reducing the external voltage supply at each process shrink (reducing the power by reducing *V* in the equation above).

Voltage scaling allowed performance to continue increasing while not impacting power consumption by a substantial amount. The limit of voltage scaling was reached when the voltage could not be reduced enough to compensate for the increase in clock frequency. As microprocessors began consuming 100 watts of power or more, the clock frequency started to decrease as seen in Figure 2.4 [7]. This sparked the commercialization of multi-core processors.

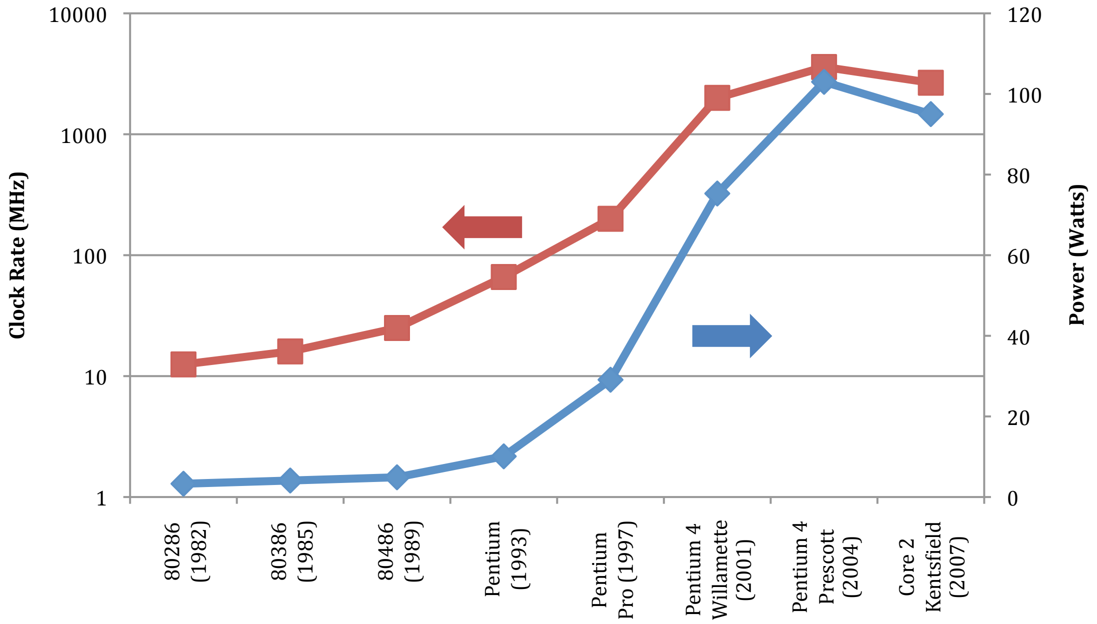


Figure 2.4 – Clock Rate and Power for Eight Generations of Intel x86 Processors

Dynamic Random Access Memory (DRAM) Power Consumption

DRAM is currently demonstrating the same type of power limitations as previously seen in processors. Like processors, voltage scaling in DRAM has masked the issue of a power limit. Unlike microprocessors, market segmentation also plays a major role in masking the power limit. For these reasons, power consumption has taken the back seat to other performance metrics (bandwidth and capacity).

Initial DRAM products used a power supply of 15 volts (1970’s) [8], while current generations are using 1.5 volts [9]. DRAM, manufactured using complimentary metal oxide semiconductors (CMOS), cannot scale their external voltage much further down than 1.0 V. As current consumption continues to rise at each subsequent generation of DRAM, we can visualize the existence of a DRAM power limit. Figure 2.5 shows how DRAM manufacturers are touting voltage scaling as a power reduction metric while not mentioning the practical voltage-scaling limit [10].

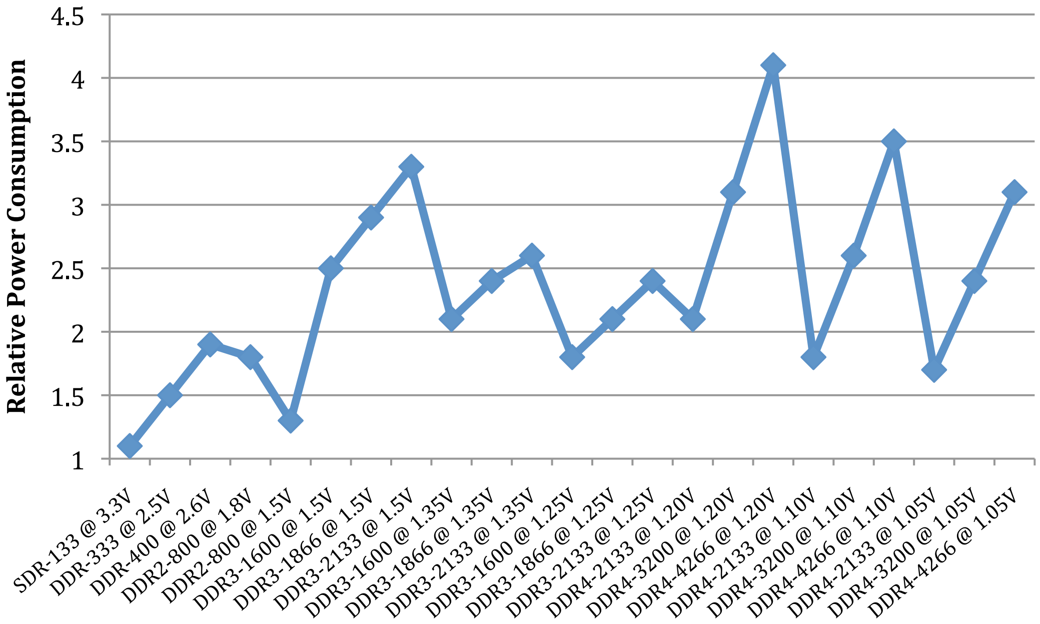


Figure 2.5 – Power Consumption Versus Maximum Frequency in DRAM

Market diversification has led to three major markets for DRAM products: mobile, desktop, and servers. Currently, mobile memory is designed with power in mind. It consumes less power than desktop memory, but suffers additional cost premiums and performance reductions to achieve the reduction in power. Server memory does not have memory designed specifically for that market segment. Instead, enterprise applications use desktop DRAM. The low component count of desktop DRAM (relative to enterprise) is designed with power as a secondary metric. This causes server applications to be outfitted with inefficient DRAM products; this contributes to the datacenter's energy efficiency problem.

Power consumption of DRAM will continue to increase at a rate greater than that of microprocessors. Figure 2.6 shows the increase in power as a function of both bandwidth and capacity [6]. The results are misleading, in that, it shows bandwidth having a greater impact than capacity on power consumption. The reason the figure is misleading is that it only analyzes the differences in 2 GB and 4 GB memory modules. Current server architectures can utilize over 300 GB of DRAM [11].

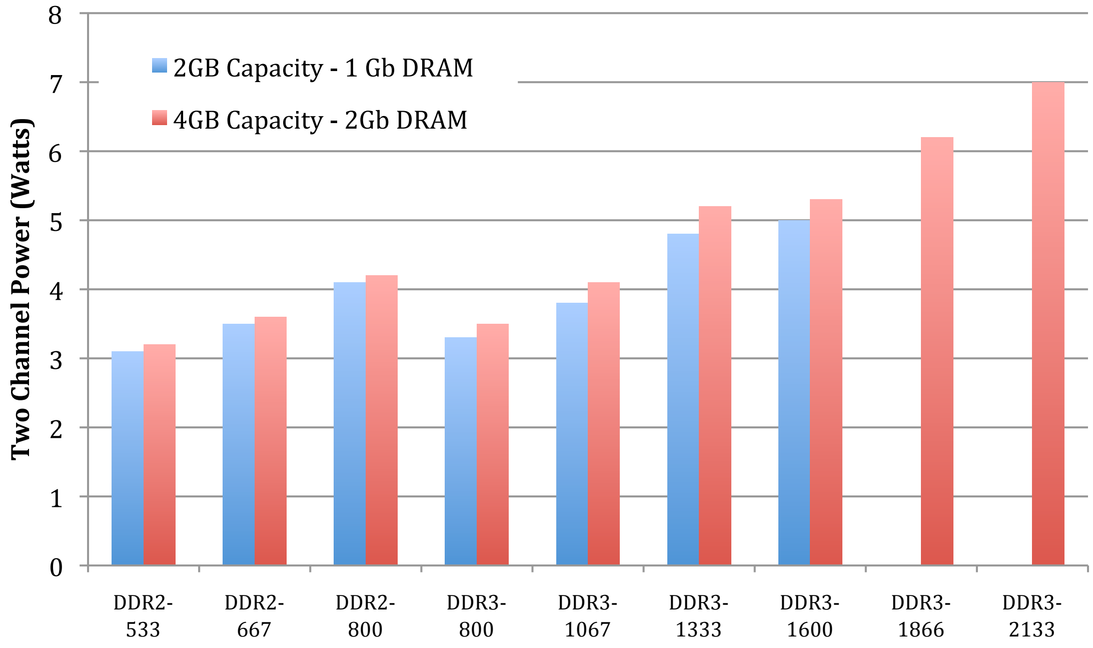


Figure 2.6 – Power Consumption of DRAM Versus Capacity and Bandwidth

A server configuration that uses large amounts of DRAM has a significant impact on power consumption. Servers utilize 16, 32, and up to 64 dual inline memory modules (DIMM) to achieve the required main memory capacity. Depending on its usage, each DIMM can consume up to 20 W of power. An extreme case of power consumption is achieved when 64 DIMMs are used. This could result in an average power consumption of 640 W (10 W per module) and a peak consumption of 1.28 kW (20 W per module). Reducing the power consumption of main memory is left up to server and CPU architects. Servers use complicated usage schemes to reduce the power consumed by main memory.

Capacity vs. Power

Table 2.1 shows the effects of both bandwidth and capacity for server configurations utilizing 64 GB of DRAM [6]. Depending on the configuration of the DRAM, which provides 64 GB of memory, the DRAM will have power approaching that of the microprocessor.

Table 2.1 – Power Consumption of Various 64 GB DIMM Configurations

Sample Card	Freq (MHz)	DIMM Configuration	DIMM Tech/Capacity	Power/DIMM (Watts)	64GB System Power (Watts)
Card A	1066	Quad Rank x4	2Gb/8GB	15.5	124
Card B	1333	Quad Rank x8	2Gb/8GB	10.6	84.8
Card C	1333	Dual Rank x4	1Gb/4GB	10.6	169.6
Card D	1333	Quad Rank x4	2Gb/16GB	20.5	82
Card E	1600	Quad Rank x8	2Gb/8GB	10.1	80.8
Card F	1600	Quad Rank x4	2Gb/8GB	19.1	152.8

Increasing the performance of datacenters requires an increase in memory capacity. Reaching 64 GB of main memory in a server, referenced in Table 2.1, requires the use of high capacity memory modules. 64 GB memory modules typically house dual or quad die packages and occupy all 36-component slots on the module. Figure 2.7 shows memory capacity allowing more transactions (increased performance) to be completed per minute in a database [12].

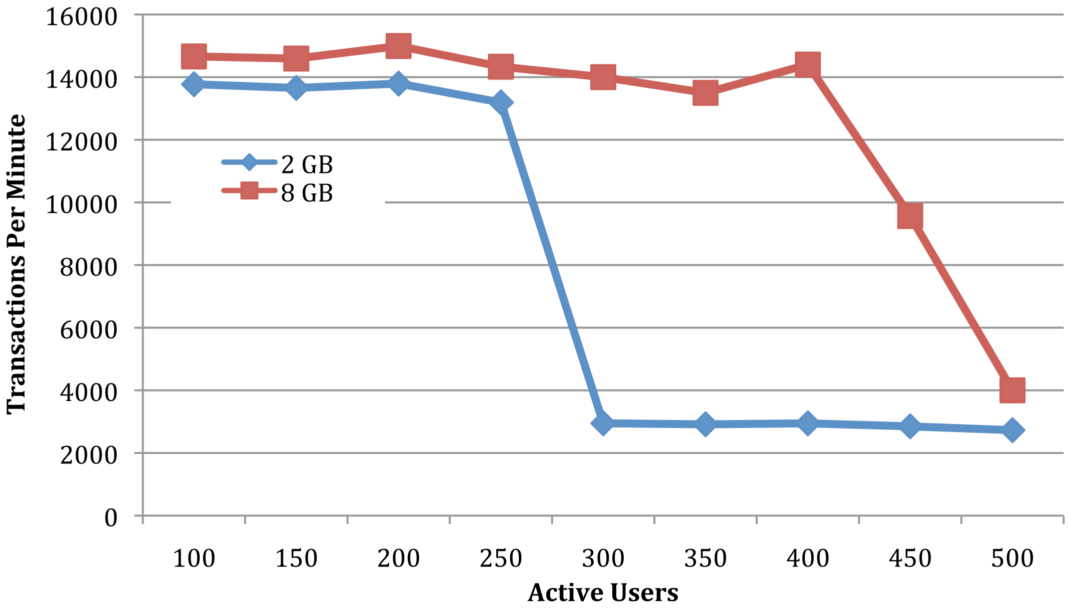


Figure 2.7 – Memory Capacity Impacts System Performance

Current computers are balanced with respect to performance, power, capacity, and cost. Blindly increasing capacity without ensuring that the other factors are considered will cause the innovation to not be applicable to mainstream markets. Instead, new technology that does not maintain system balance will only find its place in low volume and high premium markets.

Table 2.1 and Figure 2.7 show that increasing memory capacity in a server also increases system performance and system power consumption. This causes some servers to not utilize all of the memory slots available due to power consumption.

Bandwidth vs. Power

Another effect of increasing server main memory is that the memory bandwidth begins to degrade as you occupy more memory slots. Figure 2.8 shows that the bandwidth of the memory channel is inversely proportional to the number of occupied memory slots [13].

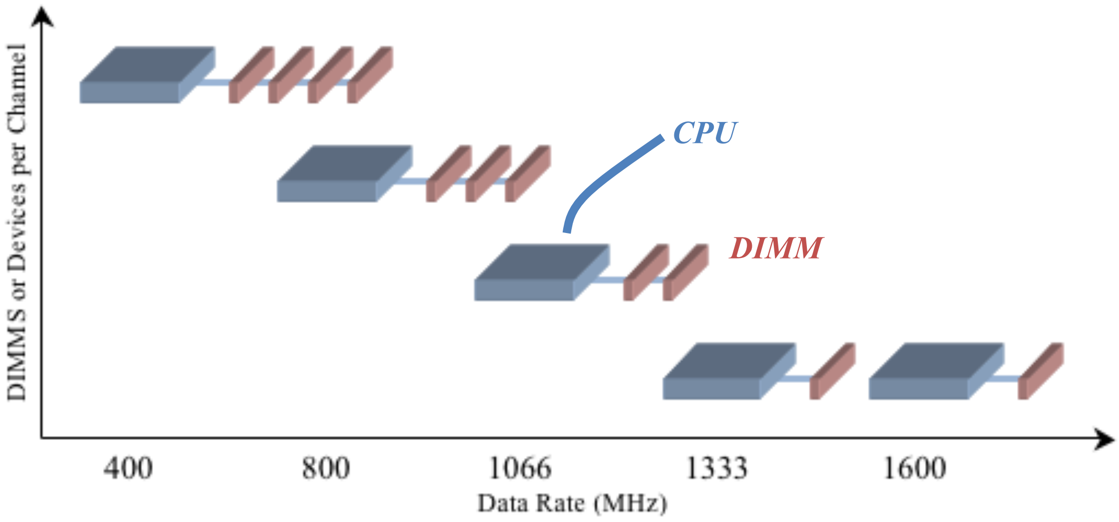


Figure 2.8 – Devices per Channel Versus Bandwidth

An advanced memory buffer (AMB) is added to the memory module or the motherboard to obviate the bandwidth limitation. The AMB can be thought of as a high-speed serial link between the memory and the microprocessor. Current AMB chips consume the same amount of power as one memory module [14]. For this reason, the AMB was removed from the memory module and placed onto the motherboard. This resulted in one AMB per memory channel. The buffer on board (BoB) configuration is becoming the mainstream approach to solving the capacity issue. Another option of increasing the memory capacity, without degrading bandwidth of the memory channel, is to populate the module with a large number of components. This increases the cost of the memory module.

Bandwidth of a processor's main memory has an impact on its performance. Analyzing the roofline model of Figure 2.9 shows this effect [7]. The roofline model was developed to allow CPU architects a quick and easy way to determine what was limiting a computer's performance. It uses benchmark data to plot both the peak memory bandwidth and the peak floating-point operations. The roofline model shows that memory bandwidth limits computer performance by creating stalls, i.e., making the processor wait for its memory. Once memory bandwidth increases to a point that completely satisfies the memory bandwidth criteria for a processor, the processor becomes the bottleneck. In this way, we can show why memory bandwidth improves system level performance.

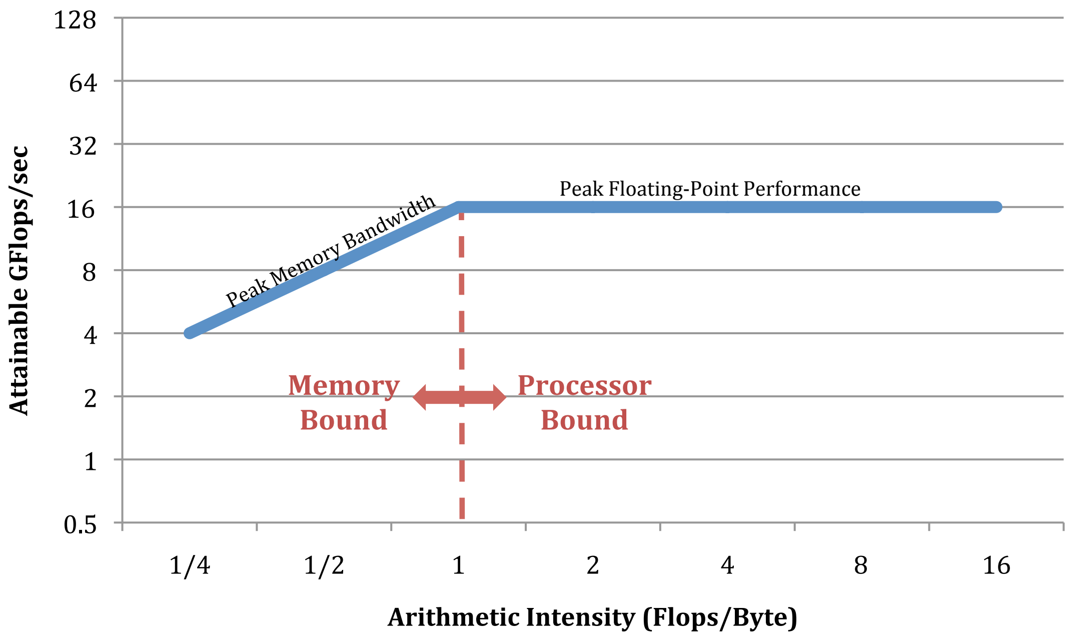


Figure 2.9 – Roofline Model

The peak memory bandwidth shown in Figure 2.9 is defined as the maximum transfer rate to/from main memory and the microprocessor. This value can be obtained through memory specifications or from benchmarks, such as the steam memory benchmark. Peak floating-point performance describes the maximum number of floating-point operations that can be completed in a second and varies for different processors. Arithmetic intensity is a term used to describe the number of floating-point operations in a program to the number of data bytes access by a program from main memory [7].

The memory hierarchy uses on-chip cache to supply the required bandwidth to the processing core. The main memory (DRAM) is housed off-chip and has a lower bandwidth compared to on-chip cache. This is because the main memory is bandwidth limited by the memory channel. The transmission lines of the memory channel have a multi-drop architecture that creates impedance discontinuities.

Each impedance discontinuity causes a reflection into the memory channel. These reflections will degrade the transmitted signal to a point where the receivers on the memory controller, or the DRAM chip, cannot recognize the transmitted or received symbol. This is another reason why the memory hierarchy has been split between on-chip cache and local off-chip main memory.

Server Architecture

Servers contain the same hardware that you would find in a desktop computer, except at a larger scale. Figure 2.10 shows the Intel S5520UR server motherboard, which is used in video servers, web servers, and high performance computers [15].

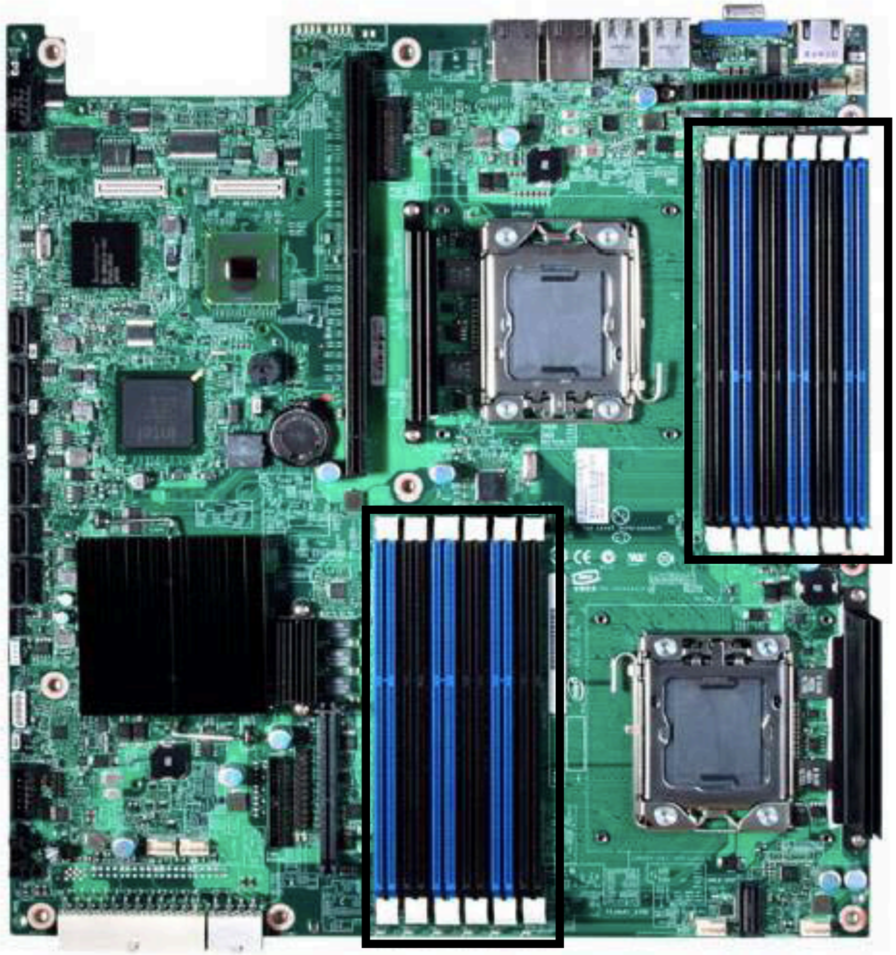


Figure 2.10 – Intel® Server Board S5520UR

The Intel motherboard can be used with one or two Xeon 5560 processors and up to 12 DDR3 memory modules. The memory controller is integrated onto the processor and allows communication to the memory modules. The controller communicates to six modules through three memory channels, as seen in Figure 2.11 [15].

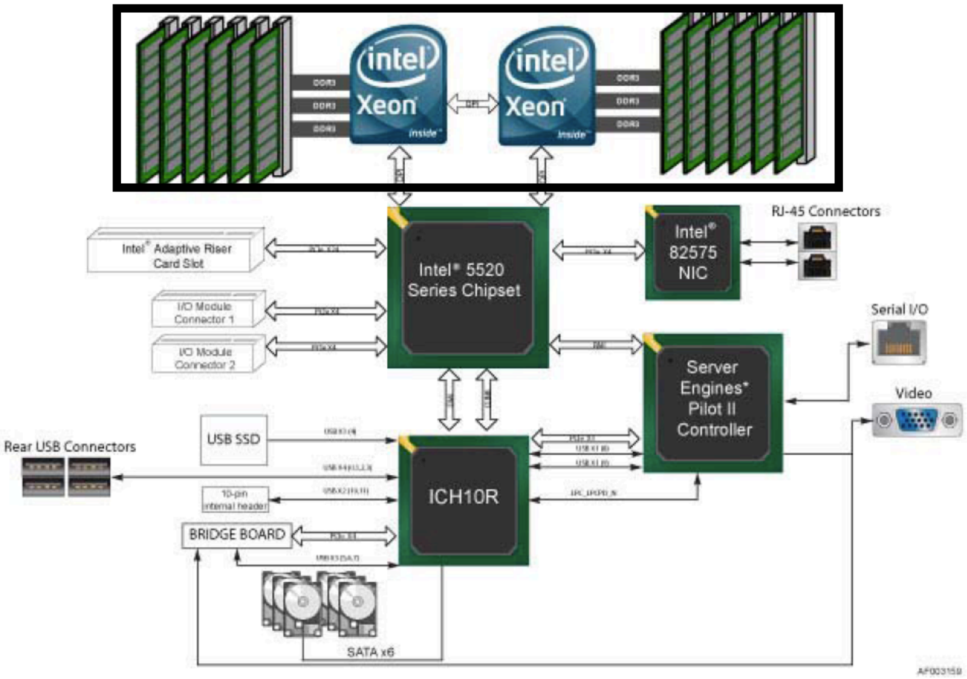


Figure 2.11 – Intel® Server Board S5520UR Functional Block Diagram

Each memory channel consists of 144 signals for a total of 432 signals routed between each processor and its main memory. Creative routing is required to route the three memory channels between the processor and its main memory. Modern motherboards are approaching 20-layer printed circuit boards (PCB) to achieve this complex routing. The distance between the processor and the main memory is large enough to require the creation of transmission lines. As discussed previously, the multi-drop configuration used in the memory channel limits the sustainable bandwidth in the memory channel.

The increasing memory bandwidth requirements of processors have required fewer memory modules to occupy each memory channel. Processor designers solved this issue by adding more memory channels to the processor; the effect was an increase in power and additional layers used on the motherboard PCB. Adding additional layers to the motherboard increases the cost to manufacture the board. This gives evidence that current memory channel architecture is only sustained by increasing the power consumption and cost of computer systems.

Registered DIMM (RDIMM) Architecture

The third generation of Double Data Rate (DDR3) DRAM is transitioning to become the mainstream memory that is used in server computers. The addition of a register is required to reduce the load on address and command signals being sent from the memory controller. Servers also require the existence of an additional parity option to ensure error-correcting capabilities for the module. Standard DDR3 modules come in a variety of configurations (capacity, bandwidth, and component count). Figure 2.12 shows the outline of a high capacity DDR3 RDIMM [9].

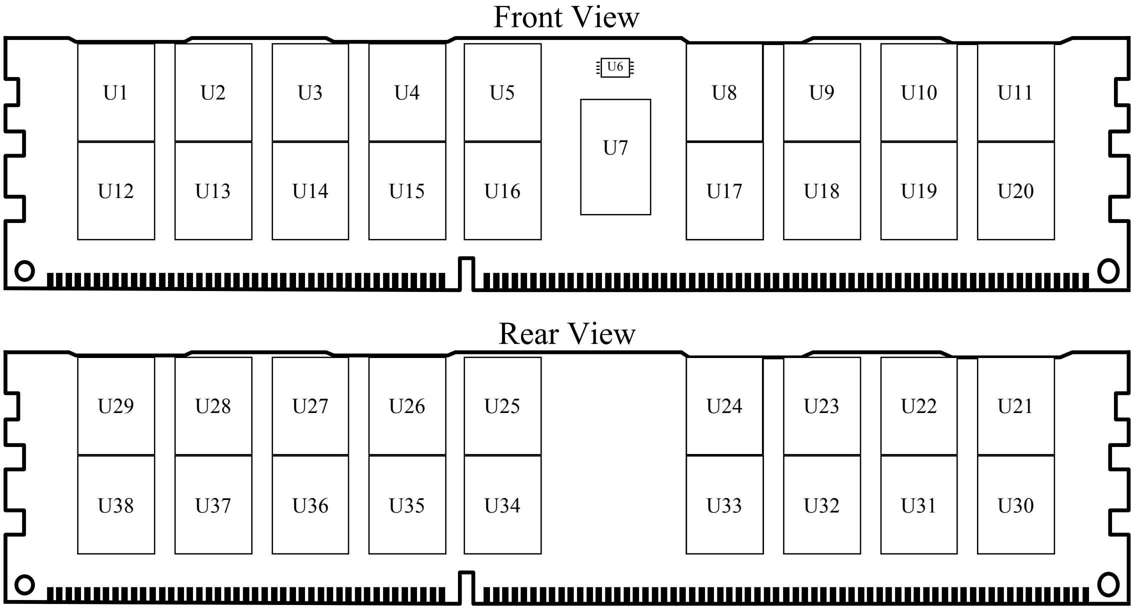


Figure 2.12 – Typical High Capacity Memory Module

When devices are planar, the maximum number of components that can fit on a module is 36. The memory modules are required to have greater than ten layers on the PCB design to achieve the high capacity memory module; this increases the cost to manufacture the module. Reaching 72 components per module requires the use of multiple memory die per component. This type of configuration would use dual die (for 72 components) or quad die (for 144 components) memory components.

High-capacity memory modules require the use of memory ranks. A memory rank describes the use of treating a set of DRAM components as one physical bank. Using ranks adds additional termination power to the memory module. As one rank is being read or written to, the other rank must be in a termination state. The module could be treated as one entire rank, allowing for a reduction in power compared to a multi-rank memory module. Depending on the manufacturer and the mode of operation, DDR3 memory modules containing 36 components can consume 1 watt (W) while powered down and 20 W while in active mode. This power consumption adds a limit to the number of memory modules that can occupy a server. Reaching 64 GB of main memory places a severe power constraint on servers.

Memory capacity increases in server platforms results in an increase in costs and power consumption, as previously explained. The termination resistors are responsible for a large portion of the power consumed by the main memory. The termination resistors are included in the DRAM die and are termed on-die termination. A typical termination resistance of the memory module is 60 Ω. The power consumed by the termination resistor is 0.009 W per data I/O pin. Each DRAM die has eight I/O, and the number of DRAM die per module is 36. If a read or write to DRAM occurs, the termination power of one module is 2.7 W. If a server has 16 DIMM, then the power consumed during a read or write by the termination resistors is 43.2 W. DDR3 memory can operate with a bit period of 625 ps, and a burst length of eight. The resulting energy consumed by the termination devices, for a read or write, is 216 nano-joules (nJ).

The power consumed by the main memory of a server is reaching a point where DRAM power consumption will be greater than the power consumed by the processors. This will make the main memory the largest consumer of power of the server, and thus the datacenter.

DRAM Architecture

DRAM manufacturers have prioritized memory cell size reduction over all other metrics over the years. This leads to the cost per bit reduction of 9% per year [16]. Manufacturers are forced to place 9% more memory bits into a fixed area to remain profitable. Due to this, the devices and wires used to access the memory cells are also being shrunk at each new process technology. This causes the resistance of both the bitlines and the wordlines of the memory cell to increase. DRAM manufacturers focus on the processing aspect of creating the tiny memory cells rather than increasing the bandwidth of the memory devices.

The cost of manufacturing DRAM is also a major concern due to the standardization of main memory. Multiple manufacturers create identical products, which results in a very competitive market space. Die size and process complexity (number of processing steps) are minimized so that costs can remain low. The main memory market is cyclical with deep troughs and high peaks. During the troughs, the costs must be low enough that the manufacturer can remain financially viable, or they go out of business.

Die size reductions have required the use of high array (ratio of the memory arrays to the chip's size) efficiencies (~ 58%). DRAM manufacturers have developed an

8,192 bit page size to reduce the area required for row circuitry. When a read, write, or refresh occurs, the entire page is required to be latched into the bitline sense amplifiers. The bitlines are driven to full logic levels by the bitline sense amplifier. Assuming the maximum activation rate is 100 ns, the bitline capacitance is 200 fF and the page size is 8,192 bits; the sense amplifiers consume ~ 20 mW of power per DRAM die. With 36 components per module, and assuming 16 modules, the power consumed by the bitline sense amplifiers is 11.5 W. This number is directly proportional to the size of the page. If a page is reduced by ½, then the power consumed by the bitline sense amplifiers can also be reduced.

Originally, the page size being large was a benefit to processor designers because of spatial and temporal locality. An accessed page remained open, which negated the row access time when subsequent accesses were to the open page. The introduction of multi-core and multi-thread processors have made temporal and spatial locality less relevant. Independent threads running on a processor have no correlation between their memory addresses. Memory controllers are required to become more complicated so that the requests to main memory can be staggered to preserve temporal and spatial locality. As the number of independent threads increases with more cores and more threads per core, the complexity of memory scheduling will increase.

Problem Statement

The main memory subsystem has become inefficient. Power consumption, capacity, and cost are all moving in the wrong direction to sustain performance gains of computer systems. This dissertation proposes a new main memory architecture that utilizes inexpensive innovations, including interconnect and packaging, to substantially reduce the power, increase the capacity, and increase the bandwidth of the main memory system.

CHAPTER THREE – NANO-MODULE

In this chapter, we turn to low cost packaging technologies to propose the creation of an 8-die package and a 32-die memory module. These new form factors are designed for high capacity systems, such as server computers. The technology can be leveraged for use in mobile platforms. The previous chapter described current and past solutions used to maximize memory capacity in server applications. These included the use of a quad-die package and a very low profile DIMM (VLDIMM). The use of advanced packaging, borrowed from multi-chip module (MCM) developers, allows for the creation of three-dimensional integrated circuit (3DIC) configurations.

DRAM memory modules are limited to 36 component slots (18 per side) due to planar size limitations of the module. Dual and quad die components are used to increase the capacity further. If quad die packages are used for all 36 components, then the total number of memory die that can occupy one module is 144 die. This places a capacity limit because only one module can occupy a high bandwidth memory channel.

Brief History of 3DIC

3DIC technologies were first introduced in the early 1990s. These innovations realized the planar limit to scaling and instead turned to three dimensions to increase the capacity. These innovations are termed 3DIC. MCM engineers were the first to realize the planar limitation when integrating many die. Val and Lemoine proposed a memory stack that comprised multiple SRAM die [1]. Figure 3.1 shows a visual depiction of the concept they discussed.

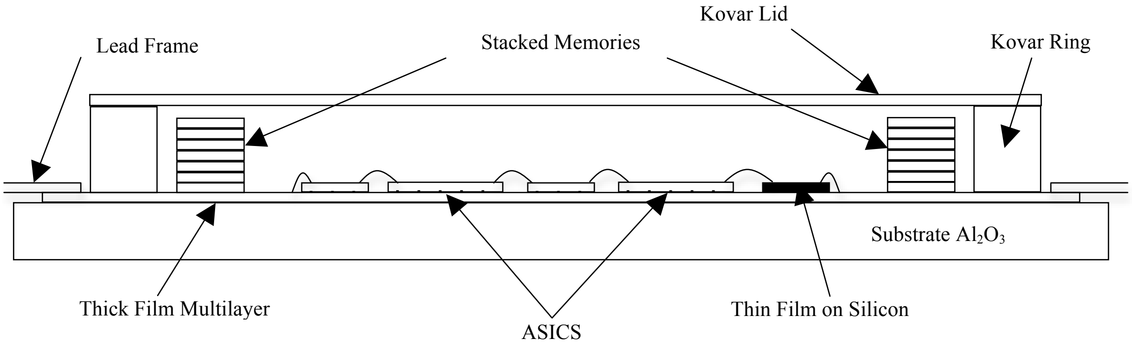


Figure 3.1 – Ultra Dense MCM

The stacked memories are tested and created by first placing the bare die into a recess in a ceramic substrate. After placement, the top of the memory and the ceramic are nearly coincident. A thin film of aluminum is screen printed onto the top of the ceramic to serve as a bonding site for wire bonds. An automatic bonding step is used to wire bond the bare die to the ceramic substrate. The authors show that the bare die can then be electrically tested and/or burned-in.

After testing, both the ceramic component and die component are stacked using identical structures. Insulating glue is used between each layer in the stack. A trimming step is used to trim the ceramic structure away from the bare die. This step leaves the wire bonds exposed on all sides of the cube. This technique is not applicable to modern device sizes. At Boise State, we experimented with thin and thick film deposition and found that the required feature sizes were too small for these types of films.

Val and Lemoine describe a metallic plating step that is used to place metal on each side of the cube. A laser is used to create interconnects by removing excess metal on the sides of the cube. This step provides electrical connection of the entire memory cube. The lateral wiring can be made to form bonding sites. This will allow the cube to be placed onto a MCM that has receiving metal sites available.

Several years after this initial memory cube technology was introduced, researchers described multiple variants of this technology. Irvine Sensors was active in this area and contributed many innovations. In 1993, researchers from Irvine Sensors used IBM interconnect technology to create a 20-DRAM chip memory cube [2]. This new innovation did not require the use of a wire bond step, trimming step, or lateral metal deposition step. Instead, they used a thin film metallic layer to route the memory die's I/O to the edge of the chips.

A tab of metal was created at the edge of the memory die for each I/O. The tab was used to create a conduction path when all 20 die were glued in a memory cube. A substrate was developed that contained landing sites for the cubes exposed metal tabs. The memory cube was simply placed on the substrate, which contained routing from the landing sites to wire bond sites at the edge of the substrate. Figure 2 shows the resulting 3D memory cube.

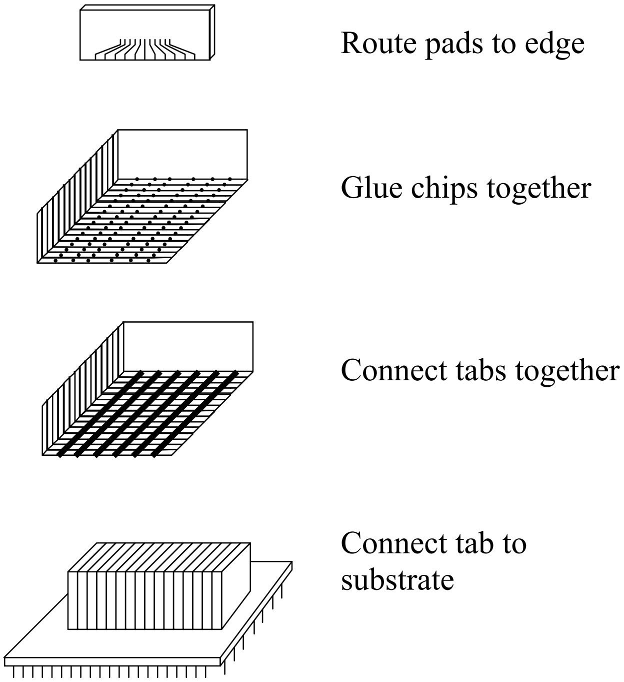


Figure 3.2 – Fabrication Steps of the Irvine Sensor Memory Cube

The substantial gains in memory capacity, due to advances in processing technology in the early 90s, made the requirement for 3DIC integration less important than other areas of research, and its interest wavered. Planar spatial limitations are arising again, and the use of 3DIC configurations are once again being explored. Samsung engineers were the first to demonstrate a 3DIC stack of DRAM die that utilize through silicon via (TSV) technology [3].

Through Silicon Via

TSV interconnects are being pursued by a large number of private and public organizations. Industry wide financial investment into TSV integration increases the chances that the technology will become the main stream for 3DIC.

Samsung engineers reported on stacking four 2 Gb DDR3 DRAM die using TSV technology. 300 TSV were used for both signals and power routing between the four chips. They realized that the use of four identical chips was not necessary. Clock recovery, data path, and I/O circuits were only required on the die that had the physical connection to the substrate. This allowed for lower power consumption and increased die efficiency for the three other chips that did not house the redundant circuits. They used the additional savings to place redundant TSV structures; this increases the yield of the entire 3DIC. Figure 3.3 shows the block diagram of the 8 Gb 3DIC [3].

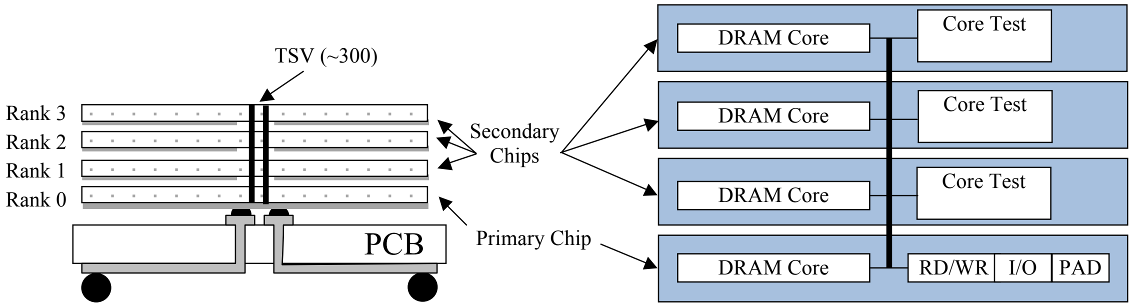


Figure 3.3 – Conceptual View of the 3DIC

TSV technology is suitable for 3DIC because they have a low resistance (< 10 Ω) and take up little silicon area (10 μm^2). However, the technology is limited due to cost and integration challenges. A consortium of industry companies are working together to develop low cost TSV technologies. In 2009, two technologies were selected as front-runners in developing a low cost TSV technology. An aggressive goal for cost of ownership was determined to be $150 per wafer [4]. The price reflects the economic sustainability of TSV technologies in modern wafer processing. Substrate noise injection, reliability concerns, and height restrictions are some of the technical issues facing TSV integration. These reasons make TSV technology a good avenue to pursue, but their integration is costly and uncertain as to when the technology will become mainstream.

The 3DIC using TSV solution that Samsung developed shows how this approach can lead to lower power and higher capacity than traditional quad die packages [3]. There are two issues with this approach: one being cost and the other being the creation of a new memory product. Memory manufacturers typically have one (or two) product offering(s) in a particular product line (DDR3) and density at a time. The Samsung approach would require the creation of an additional two offerings of a particular product line. The two additional offerings would be the bottom device in the memory stack and the three top die in the memory stack. Each of these devices would require a substantial amount of sales to offset the initial investment.

Stacking technologies that have penetrated the mainstream market in high volume use low cost packaging techniques. High density non-volatile memory products are an example of this market trend. NAND memory products extensively utilize low cost stacking technologies. Figure 3.4 shows a typical NAND memory stack.

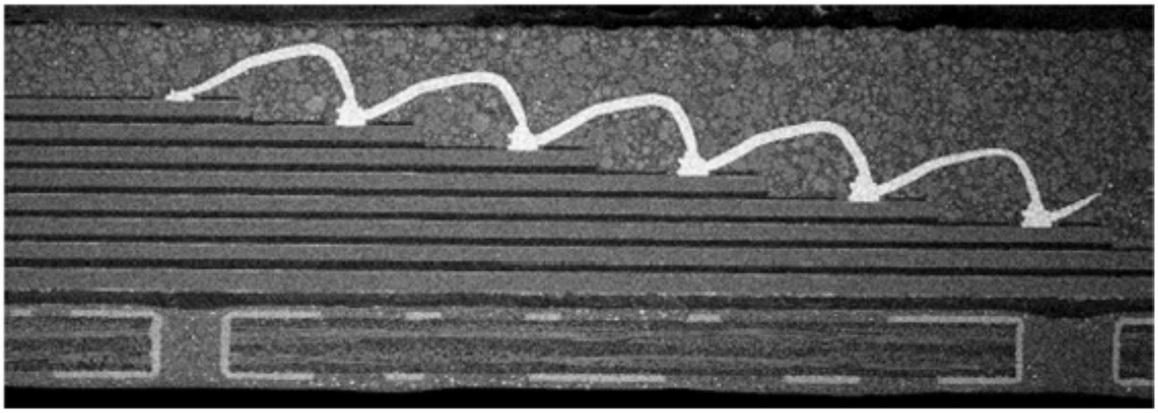


Figure 3.4 – Typical NAND Stacking Technique

The memory devices are stacked in a staggered fashion. Each additional memory die is offset from the previous memory die in the stack. This allows a wire bond to attach to each die's exposed input/output wire bond pad. The wire bond approach allows for a low cost memory solution, but also reduces the bandwidth. The shared wire bond connection adds a large capacitance and inductance to the input/output pad. The parasitic nature of the wire bond creates a bandwidth limit to the input or output signals being driven to/from the memory die. This dissertation proposes a low cost stacking approach that removes the need for the shared wire bond connection, thus increasing the available bandwidth.

Alternative methods have been proposed to increase the bandwidth and functionality of 3DIC components. The proposals, typically, include a high cost premium due to their creative stacking techniques. Figure 3.5 shows an example of this approach to 3DIC configurations [17].

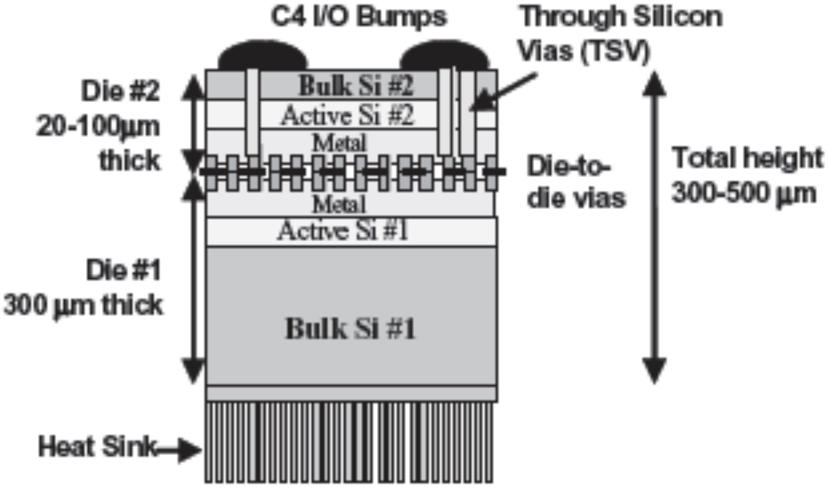


Figure 3.5 – A High Cost Premium 3DIC Configuration

Black et al. thoroughly discuss the benefits of creating a 3DIC configuration that contains a microprocessor along with memory die in the aforementioned reference. Increased bandwidth and lower power are touted as key figures of merit, but cost is not considered. The inclusion of multiple advanced packaging techniques (C4 bumps, through silicon vias, die to die vias, and a nascent heat sink technology) increases the cost to produce this type of 3DIC. This dissertation proposes a stacking technique that uses heavily verified technologies to achieve the same types of bandwidth and power metrics with out substantially impacting cost.

Low Cost Stacking Approach

The memory module developed in this dissertation describes a low cost 3DIC stacking technology that overcomes the limitations of TSV, while benefiting from the advantages of 3DIC. The stacking technology was first introduced in a 1993 patent issued to IBM [18]. The invention describes stacking multiple die at an angle and using variable types of connections to connect the substrate to the die. Figure 3.6 gives a visual depiction of this angled stacking technique.

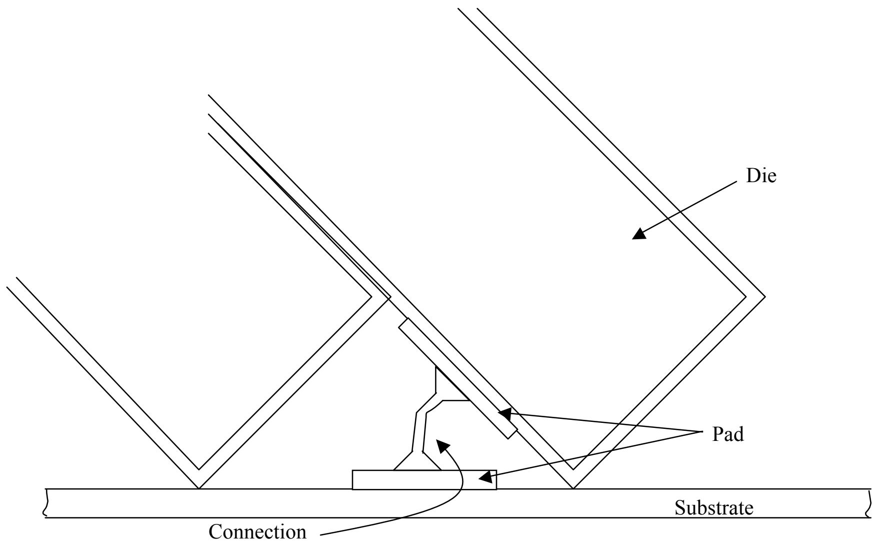


Figure 3.6 – Stepped Device Stacking

The invention describes a variety of ways to connect the angled die to the substrate. The die can be attached with a wire bond (as depicted in Figure 3.6) or with a solder ball. Alternative inventions described attaching the angled die to the substrate by applying solder to both the substrate and die landing sites, placing the two together so that their solder touches, and reflowing the stack to get electrical connection. Figure 3.7 shows an alternative connection method detailed by engineers at MCNC [19].

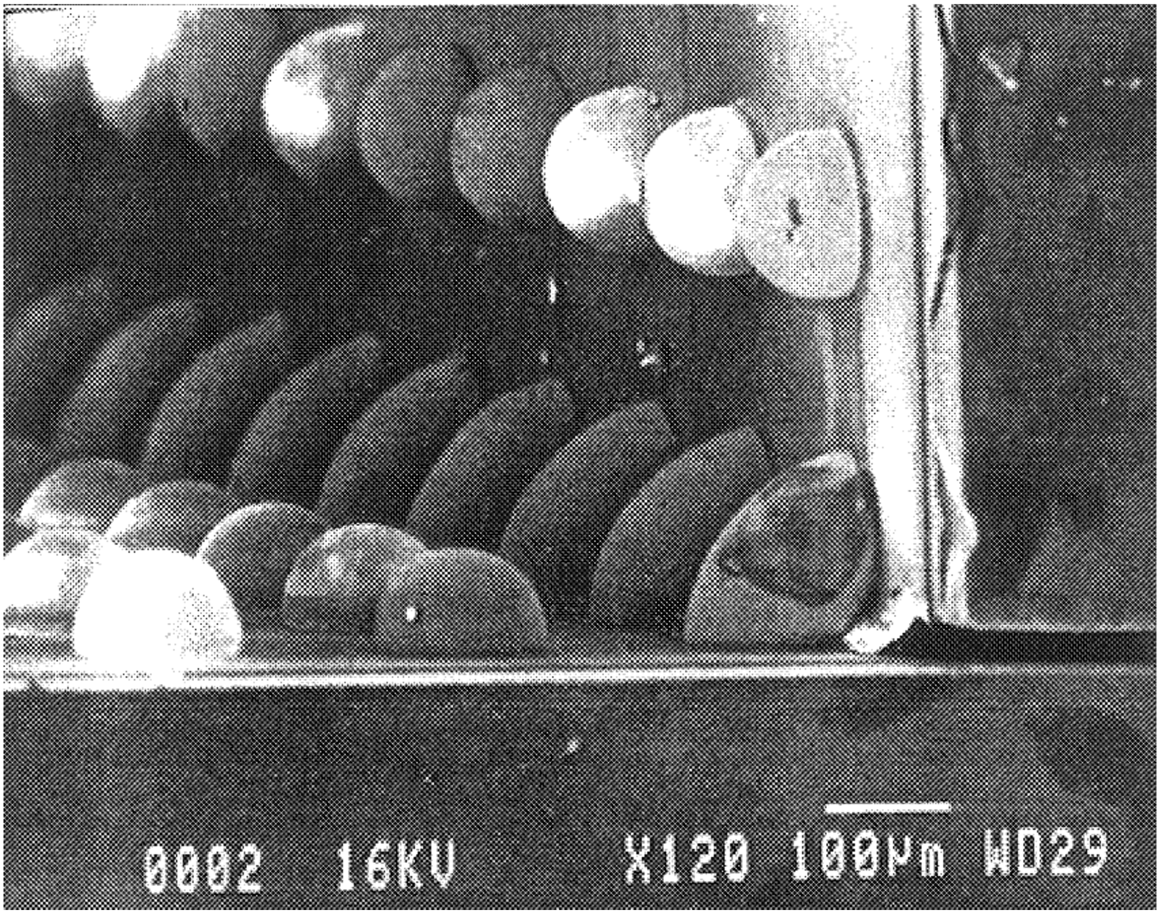


Figure 3.7 – Alternative Method of Angle Bonding the Die to the Substrate

These stacking techniques allow multiple die to be stacked together in a 3DIC configuration. The major advantages of angled bonding technology are cost, power, and bandwidth. The cost of the high capacity memory stack can be greatly reduced by using low cost packaging techniques. The stacking technology places the memory die into a small area allowing for a reduction in the trace lengths. The trace length reduction also reduces the parasitic load that is driven by the I/O of the memory die. This has a benefit of reducing the power consumed by the memory die. The memory chips are placed very close together and the termination can be shared between the die. A configuration where every fourth memory die's termination is active can be used and would result in substantial power reduction. The reduction in trace length also has the potential to increase the bandwidth of the memory die due to the reduction in parasitic capacitance, resistance, and inductance.

Nano-Module

Leveraging all of the benefits of this technology requires the use of an active silicon substrate. Placing registers/buffers directly below the memory die allows for the elimination of the termination resistors on the memory die. Termination is only necessary at the I/O of the substrate.

Redistribution Layer

The redistribution layer (RDL) is used by first routing the center I/O pads of the memory chip to its edge as performed by Bertin, Perlman, and Shanken [2]. This can be done with a single, very low cost, redistribution layer. When multiple layers of RDL are used, it is possible to substantially reduce parasitics on the connections, resulting in a higher bandwidth and lower power connection. The voltage drop due to resistances of the metal traces and edge pad spacing are the major challenges. Figure 3.8 shows a mock up of the redistribution layer used to route the center pads to the edge of a typical DRAM die.

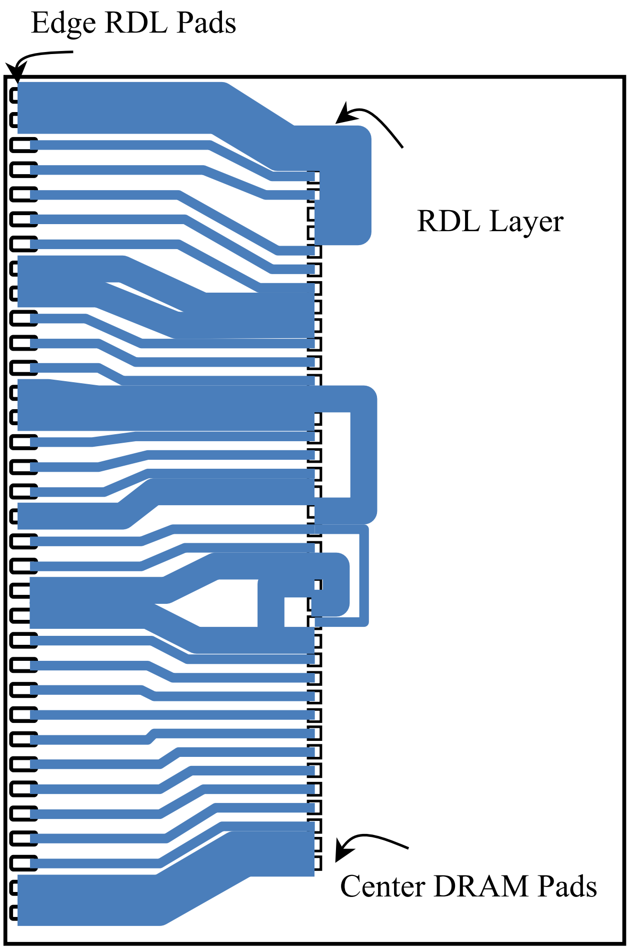


Figure 3.8 – Redistribution Layer Used to Route Pads to the Edge of the Die

A major concern when routing the power signals to the edge of the die is the voltage drop. The voltage drop is due to the high current that can pass through the power and ground routing. Current DRAM die are getting smaller, which limits the space available to route the signals. RDL layers that are used on silicon die can achieve a minimum of 10 µm of spacing and 10 µm of width for the metal tracks. Using two RDL layers to route the signals is preferable but requires expertise at the RDL processing facility. Non-uniformity on the DRAM substrate makes processing the RDL layers difficult. Currently, DRAM die have a higher number of internal pads than external pads, meaning that some power and ground buses need to be consolidated.

The DRAM die size places a limit to the number of I/O pads that can be placed on its edge. A typical DRAM die will have an edge width of approximately 8 mm and a typical pitch of soldered connections is 200 µm. This places a limit of 40 I/O pads that can sit on the edge of the DRAM die. If the solder connection pitch is reduced to 133 µm, it is possible to fit 60 I/O pads on the edge of the DRAM die. This is sufficient for DDR2 memory parts, but DDR3 memory die are using 78 external connections to the memory die (mainly due to an increase in the number of power and ground balls).

A solution to this problem is to use a material surrounding the silicon die that acts as a fan out for the RDL routing. This is used in some wafer scale packaging (WSP) or very fine ball grid array (VFBGA) packages. Infineon demonstrated their fan out technology for their WSP technology [20]. Figure 3.9 shows a cross section view of a die that has been surrounded by a mold compound that creates additional space for I/O routing using RDL layers. The mold compound extends the size of the chip to allow for a large area to place an RDL layer. The bottom of Figure 3.9 contains the solder balls used to pass the signals from the chip to a substrate.

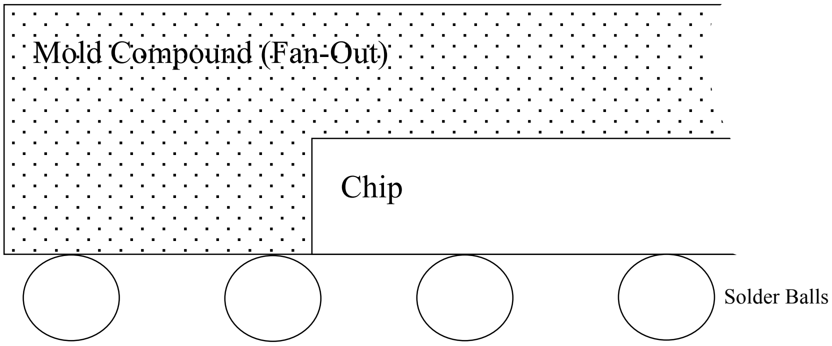


Figure 3.9 – Fan Out Used in Wafer Scale Package

Using RDL layers on the DRAM die allow the signals to be transported from the center of the memory chip to the edge of the memory chip. This solution is applicable for a single row of pads on the DRAM or a double pad row DRAM architecture. Once the signals are routed to the edge of the die, it is possible to mount the chips at an angle onto the substrate.

Substrate Design

The substrate design can be either purely passive (redistribution layer(s) only) or active (substrate contains circuitry). Both types of substrates require mounting sites for the connection to the DRAM die and a redistribution network to connect the DRAM die to the edge wire bonding sites. Figure 3.10 shows a mock up of a typical substrate design.

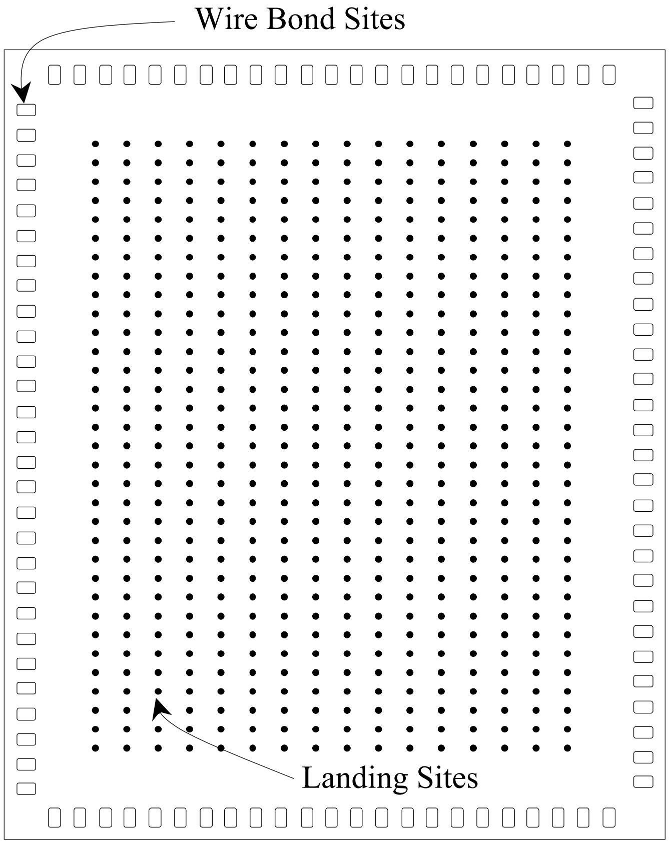


Figure 3.10 – Substrate Mock-Up Showing Wire Bond and Connection Sites

Assuming a 200 μm die to substrate connection pitch and 78 connections on each memory die, it can be shown that the total length of a single landing site in Figure 3.8 is

15.5 mm (15.4 mm plus a connection diameter of 100 µm). This will require the use of a fan out material used on the DRAM die.

Module Size Calculations

The spacing of each wire bond connection row is determined by viewing Figure 3.11.

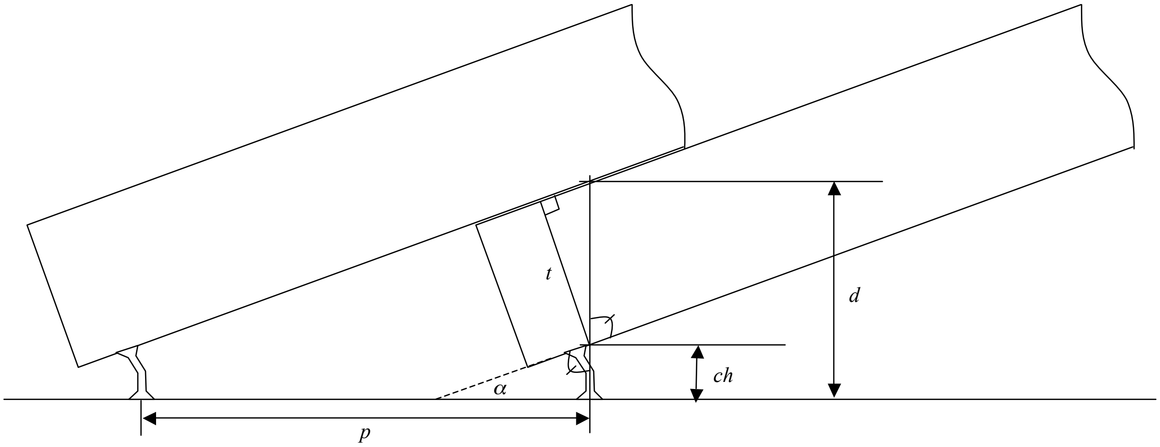


Figure 3.11 – Geometry Used to Determine Die to Substrate Connection Column Pitch

It is found that the pitch of the wire bond connection column (p) is required to be $t/\sin\alpha$, where α is the angle of attachment and t is the thickness of the die and any adhesion. The height of the unmolded package is determined by viewing Figure 3.12.

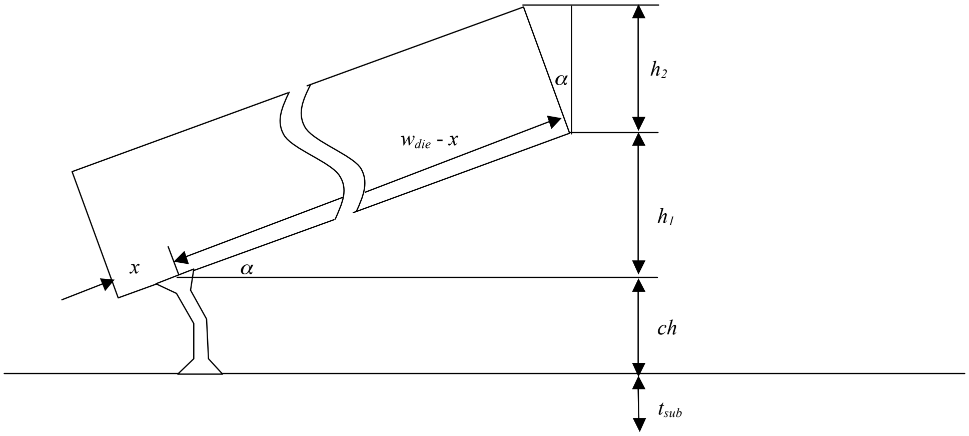


Figure 3.12 – Geometry Used to Determine the Height of the Package

Using the geometry of Figure 3.12, the height of the package can be determined. This geometric analysis can be taken further to determine the exact dimensions of the unmolded package. The following equations summarize the dimensions of the unmolded package. $t_{con\ pictch}$ in the following equations refers to the pitch of the connectors used to connect the die edge pads to the substrate.

$$height = t_{sub\ thickness} + t_{connection} + \frac{t_{die\ thickness}}{\cos\alpha} + \left(t_{die\ width} - t_{die\ pad\ to\ edge}\right)\sin\alpha$$

$$width = \left(no.signals - 1\right)t_{con\ pitch} + t_{con\ diameter} + 2\left(t_{die\ to\ pad\ edge} + t_{wb_1} + t_{wb_2}\right)$$

$$length = 2\left(t_{wb_1} + t_{wb_2}\right) + \sin\alpha\cdot t_{die\ thickness} + \frac{\left(\#\ die - 1\right)t_{die\ thickness}}{\sin\alpha} + \cos\alpha\cdot t_{die\ width}$$

The formulas above use t_{wb1} to represent the spacing between the center of the pad on the memory die and the silicon die, and t_{wb2} as the distance from the center of the pad on the memory die to the edge of the substrate. We can use these equations to determine the size of several configurations. We will assume the following parameters going forward: t_{sub} = 500 µm, t_{die} = 200 µm, attach angle = 20º, $d_{connection}$ = 100 µm, pad to edge spacing of 200 µm, and die size of 8 mm x 10 mm. Using these parameters, a 32-die package would be 3.48 mm by 16.7 mm by 26.5 mm, and an 8-die package would measure 3.48 mm by 16.7 mm by 12.5 mm.

When the memory stack is configured into a dual rank memory module, it is possible to route all signals out with two layers of RDL. Active circuitry can be placed under the memory chips to buffer the signals to and from the memory chips. The substrate can be manufactured in an older technology, which reduces the cost of manufacturing the substrate. The routing channels can be created with RDL layers or with the top layers of the active silicon substrate. This allows for varying possibilities that include passive components (capacitors and inductors) to be created with the top levels of metal. It is even possible to transfer some of the functionality of the memory controller onto the substrate. Once the substrate is defined, it is possible to mount the memory stack on the substrate as seen in the cross section in Figure 3.13.

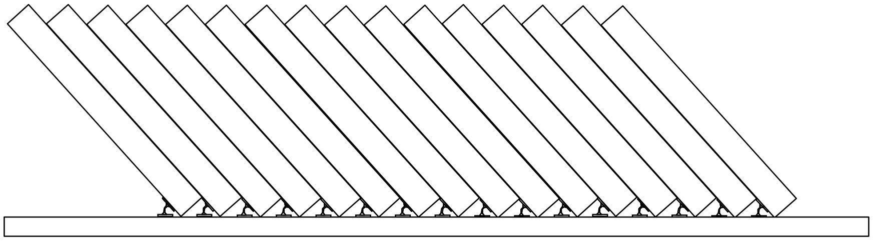


Figure 3.13 – Cross Section View of the Memory Stack

The memory stack can be connected to the microprocessor in varying ways. It is possible to place the memory stack into the same package as the CPU allowing for an extreme reduction in the memory channel power and an increase in the bandwidth of the main memory. Novel interconnect technologies can be used to connect the substrate to the microprocessor. An example of a novel interconnect is a wireless capacitive-coupled interconnect that allows for an order of magnitude increase in the density of interconnects between two die.

It is possible to configure the memory stack as a stand-alone memory module that is placed in an assembly package, much like the microprocessor. This module can be placed in a socket on the motherboard or permanently attached to the motherboard. Depending on the replacement policy, it might be better to use the socket approach. The socket approach allows for the possibility of replacing a defective module, upgrading existing modules, or allowing for scalability of the motherboard's memory.

The placement of the module affects the thermal performance of the nano-module. Heat generated by the memory die must escape the package that contains the memory die. This is true if the stack is placed within an MCM with the processor or as a stand-alone module. A complete thermal solution is outside of the scope of this dissertation but the following discussion is provided for completeness.

Thermal Options

Standard techniques used to remove heat off of the memory module are not applicable to the proposed nano-module due to the novel stacking technique. Figure 2.12 shows that standard DRAM modules place the memory die in a planar manner. This allows for the use of a heat spreader to be placed over the memory module. Thermal paste is placed on the molded die allowing for a high thermal conductivity path between the mold compound and the heat spreader. Heat passes through the thermal paste and collects on the heat spreader. Typically, air is passed over the memory modules allowing the heat to be removed from the system. Without the heat spreader, the nano-module requires heat to be removed from the top or bottom of the module as seen in Figure 3.14.

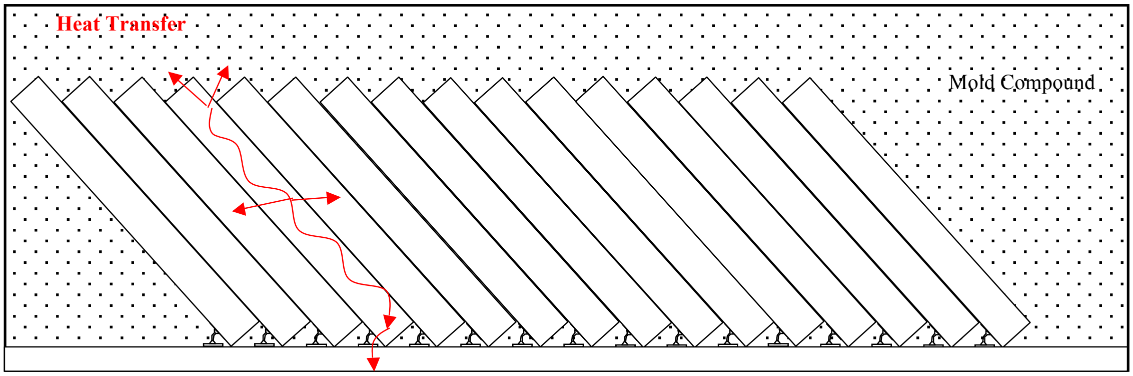


Figure 3.14 – Heat Transfer of the Nano-Module

When power is consumed on the memory die, heat is produced. The heat generated from a die follows the path of the greatest thermal conductivity. The thermal conductivity of materials is described in units of Watts per meter Kelvin (W/mK) and specifies the material's ability to conduct and transfer heat. Silicon and metal have a thermal conductivity greater than 100 W/mK, while the mold compound can have a thermal conductivity less than one W/mK.

The heat generated on a memory die will pass through adjacent memory die, wire connections at the bottom of the die, silicon substrate, and mold compound. The rate of heat transfer is lowest at the mold compound. This is due to the difference in the thermal conductivity between the materials. If air is passed over the top and bottom of the module, some of the heat will be removed. The large path through the silicon die to the wired connections at the bottom of the die will lead to a large temperature gradient.

The temperature gradient will result in a wide variety of performance across the memory die and creates hot spots. The temperature gradient causes a performance loss. The hot spots can degrade the life of the memory die due to high temperature stresses. Heat must be transferred from the memory die to the outside of the package in order to prevent these affects. Figure 3.15 shows a solution for heat removal on the nano-module.

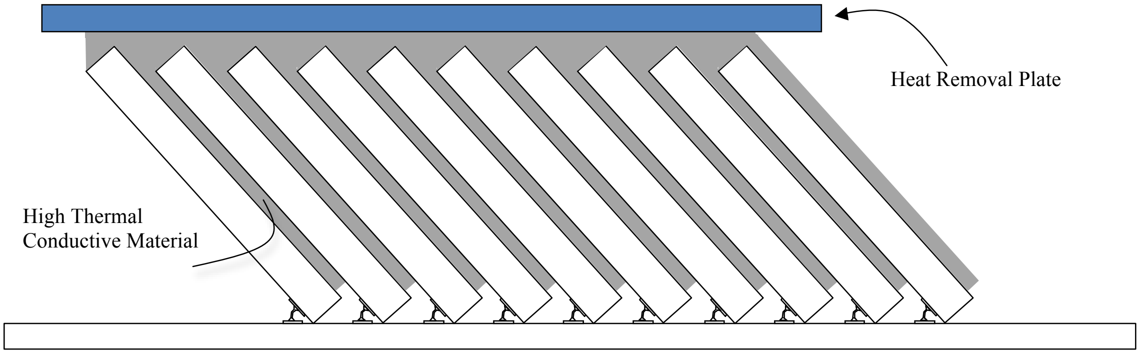


Figure 3.15 – Proposed Heat Removal Solution

Figure 3.15 shows the use of a high thermally conductive material placed between the memory die and extending to the top of the nano-module. A heat removal plate, with high thermal conductivity, is placed at the top of the nano-module. The heat removal plate is used to transfer the heat out of the nano-module. Once the heat is transferred to the heat removal plate, a variety of techniques can be used to remove the heat. These include passive and active heat sinks.

CHAPTER FOUR – DRAM ARCHITECTURE

Increasing the bandwidth of the main memory can be accomplished by increasing the number of data signals used by the memory chip. Increasing the number of data signals using conventional interconnect technologies increases power consumption. Wireless interconnects are being used in 3DIC configurations due to their low power and high bandwidth configurations. Leveraging these advantages in DRAM requires several changes to the memory chip. This chapter develops a DRAM architecture that is created for low power and high bandwidth applications. The architecture is well suited for wireless interconnect technologies.

4 Gb DRAM Architecture

Capacitive-coupled interconnects offer the largest I/O count over other interconnect technologies. This DRAM architecture utilizes up to 64 data signals due to the low power interconnect technology. Using a 4 Gb DRAM architecture (described in [21]) as a starting point, it is possible to create a wide I/O interface that integrates a wireless interconnect channel. Figure 4.1 shows a block diagram of the 4 Gb DRAM architecture.

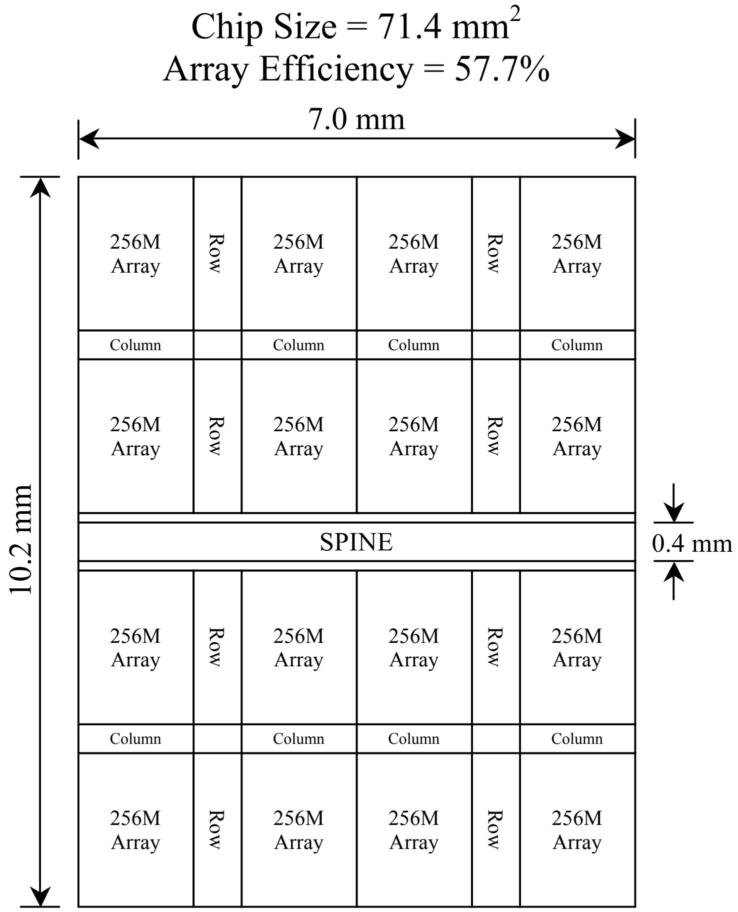


Figure 4.1 – A 4 Gb DRAM Architecture

Capacitive-coupled interconnects work when two die are placed face to face and a capacitor is formed between the top level of metal on each die. This approach works well when the I/O pads are placed on the edge of the chips. The first step in developing the new DRAM architecture was to move the pads to the edge of the die. The second step was to change the structure of the memory banks and periphery circuitry to maximize array efficiency while keeping column and row cycle times constant. Making these changes resulted in the discovery of several challenges.

Edge Aligned I/O Pads

Moving the pads to the edge would require address, command, and data signals to be buffered into the memory array. The time difference experienced from routing the data signals to the closest and furthest memory arrays would create a significant design challenge and increase the cycle time specifications of the memory. The addition of buffers would also increase the power consumption of the memory die. Due to these challenges, an alternative solution was needed.

A centralized row and column structure can be used to reduce the die size. The centralized column structure requires signals to be driven 5 mm to the edge of the memory array. Comparing this to the previous column path, which drove data signals 2.5 mm, it is apparent that the column path frequency will decrease. However, the use of a wide I/O interface will allow for more bits to be read in parallel. This, in effect, can normalize the column throughput. With this approach, column path throughput can be normalized, or increased, at the expense of access latency. The centralized row structure follows the same logic as the centralized column structure.

Main memory used for desktop and server applications operate with eight internal banks. This allows a single row to be active in each bank simultaneously, allowing the row cycle time to be circumvented. The eight-bank architecture allows for access scheduling to open word lines in each bank. Due to this, the row cycle is much greater than the column cycle time. The row cycle time is dominated by the time it takes to drive the large parasitic resistance and capacitance of the word line. This allows the typical access time to be a function of the column cycle time and not the word line cycle time. The large row cycle time and access scheduling allows for the design of a central row structure to service the memory banks.

Figure 4.2 shows the representation of eight internal banks in the 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 parasitic resistance and capacitance.

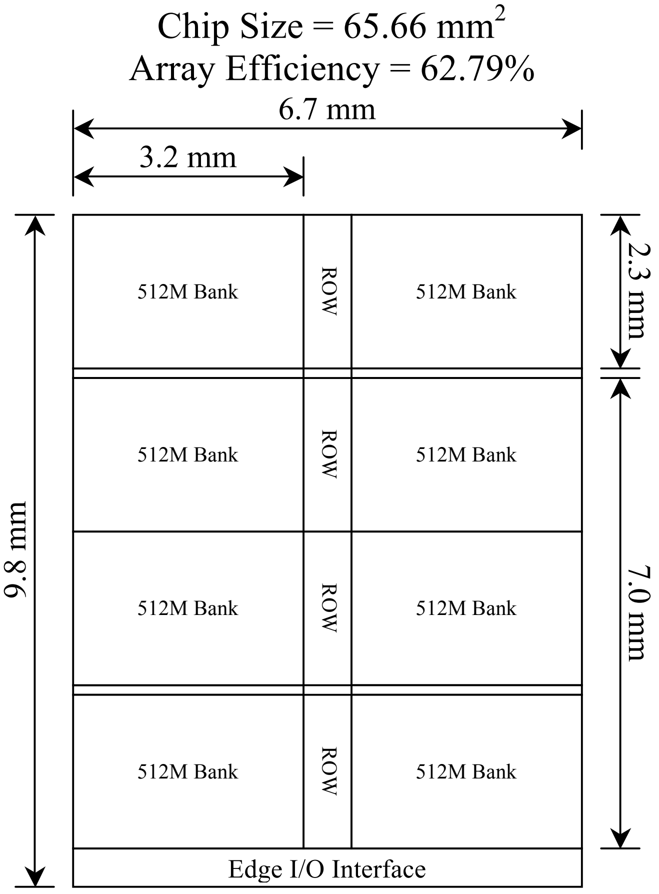


Figure 4.2 – DRAM Architecture with Edge I/O

The implemented architecture decisions allowed for the integration of a capacitive-coupled interconnect residing at the edge of the memory chip. Changing to a centralized column and row circuitry allowed for an 8.0% decrease in chip size and an 8.8% increase in array efficiency. The die size savings can be used for an increase in the number of die per wafer (cost), or additional support circuitry can be added to the die while keeping the die size constant. Establishing the centralized structures early in the architecture definition allowed for enough research and development time to implement the architecture.

The initial wide I/O architecture allowed for an immediate sense of the challenges associated with integrating capacitive-coupled interconnects on a DRAM chip. The initial choices were made with respect to chip size and array efficiency, while the challenges were in trying to buffer the signals into the DRAM chip. Alleviating some of the initial challenges required the communication channel be placed on the side of the DRAM chip rather than the bottom. Placing the communication channel on the side of the DRAM chip requires an additional review of the 512 Mb bank structure.

Bank Structure

When a word line is activated in a memory bank, a page of data is latched into the bit line sense amplifier. The number of bit line sense amplifiers activated is referred to as a page. Current DRAM devices utilize an 8k page. The power needed to charge 8k bit line capacitance sets the majority of the power consumption of the DRAM chip. While 8k bit line memory bits are read during a single access, only 64 bits are sent off-chip, which leads to a very low energy efficiency per bit accessed. 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. The page size of the DRAM can also be reduced because the temporal and spatial locality is decreasing as we move towards multi-core/multi-thread applications.

The processors memory controller requests data to and from the main memory in large parallel data accesses. The amount of data accessed is termed a cache line and comprises 64 bytes of data (or 512 bits of data). Multiple independent threads operating on a processor access data independently. The probability that the data accessed in one thread will reside on a word line opened by another thread is very low. The use of a memory access scheduler and multiple banks has allowed this technology to continue scaling.

The memory access scheduler is used to store outstanding data requests to and from main memory. The requests are staggered in a way to ensure temporal and spatial locality. As the number of independent threads increases, the hardware required to sustain proper temporal and spatial locality to an open row will begin to have diminishing returns. The 8k-page size is becoming irrelevant and a reduction in the page size is a logical next step. Reducing the page size has a significant impact on the memory power consumption and the hardware required to reduce the page size by two is minimal. The bank structure used in the modified 4 Gb DRAM architecture realizes the same 8k page size, with the realization that a reduction in page size requires minimal effort.

There are multiple ways to create a 512 Mb bank structure. The most efficient bank structure that works well with edge aligned interconnect circuitry was developed. For this reason, we choose bank structure D in Figure 4.3.

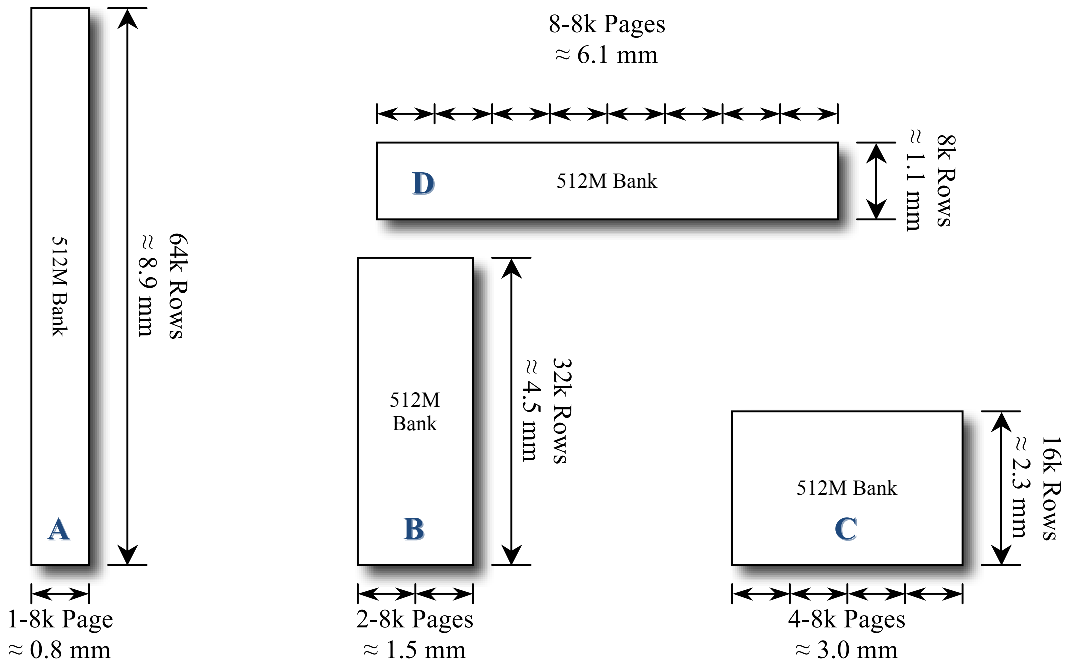


Figure 4.3 – Variable Page Size in Bank Structure

The 8k-page size per memory bank sets the number of possible structures a 512 Mb DRAM array can have. Each bank structure in Figure 4.3 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 512 Mb memory array.

Side Mount 4 Gb 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 wide I/O DRAM architecture. Placing the interconnect 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 4.3. 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 capacitive-coupled enabled architectures.

Going with a centralized global column structure can further reduce the chip size of the initial side mount architecture. The increase in global column and global row parasitic resistance and capacitance is a trade-off to reducing the chip size and increasing the array efficiency. Improving upon the initial 4 Gb side mount architecture is achieved when bank structure D (see Figure 4.3) is used. The DRAM architecture in Figure 4.4 is used as the basis for discussing further challenges that impact the incorporation of capacitive-coupled interconnects into DRAM architectures.

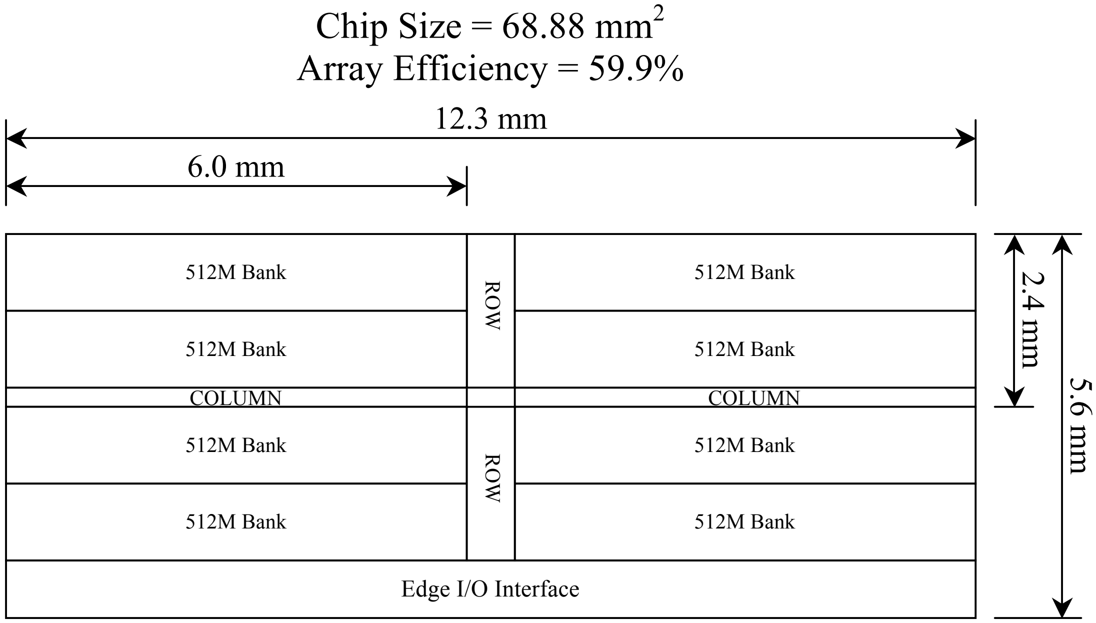


Figure 4.4 – Final 4 Gb DRAM Architecture

The DRAM architecture in Figure 4.4 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% [22]. The side-mount DRAM architecture in Figure 4.4 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 2 Gb [23]. The 4 Gb DRAM architecture depicted in Figure 4.4 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 (see Figure 4.1) will greatly reduce the manufacturing cost associated with a wide I/O DRAM architecture.

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.

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 (with low resistivity) and therefore increases the bandwidth of the global I/O routing. The lowest level of metal is used for the global word line circuitry because the extra delay of driving the larger parasitic resistance and capacitance has a smaller effect on the total row access latency, which is dominated by the word line parasitic resistance and capacitance. The current pitch of the top layer of metal is approximately four times the minimum feature size. In a 40 nm process, this pitch is 160 nm. These challenges can be overcome by dividing the bank structure, changing the datapath design.

Dividing the Bank Structure

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 4.5 shows the concept of separating the 8k-page between two half-banks.

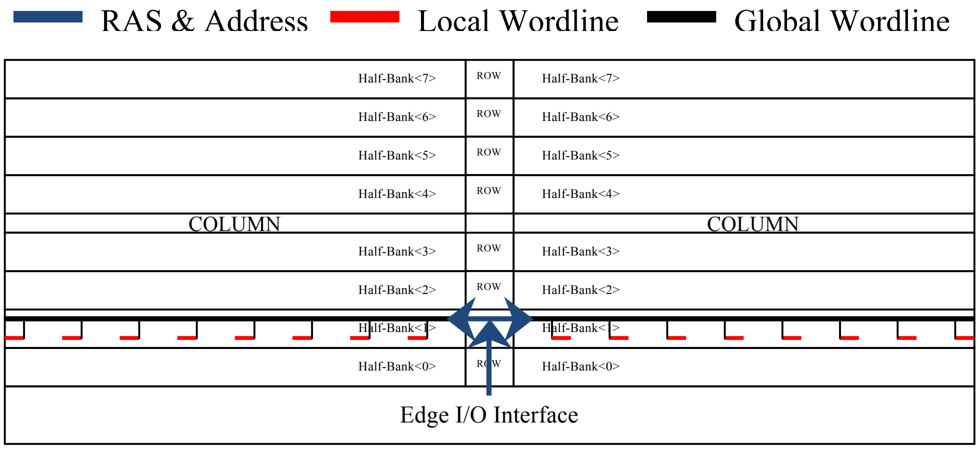


Figure 4.5 – Division of the Page

The architecture developed in the previous section has a half-bank width of approximately 6 mm. 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 6 mm 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 160 nm, 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 by 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 capacitive-coupled wide I/O DRAM architecture.

Datapath Design

The 256 Mb half-banks utilize 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. Distributing the 8k-page into two half-banks requires a 4k-page size per half-bank. This means that out of the 64k bit lines only 4k bit line sense amplifiers fire. Firing only 4k bit lines versus the total 64k bit lines requires a 16:1 ratio for the half-bank bit lines. This is accomplished by sending four additional bits with the master word line to the local word line drivers. The four bits are decoded at the local word line driver and perform the 16:1 page decode. Figure 4.6 depicts how the array macros are decoded so that only 4k of the 64k bit line sense amplifiers are fired during a word line activation.

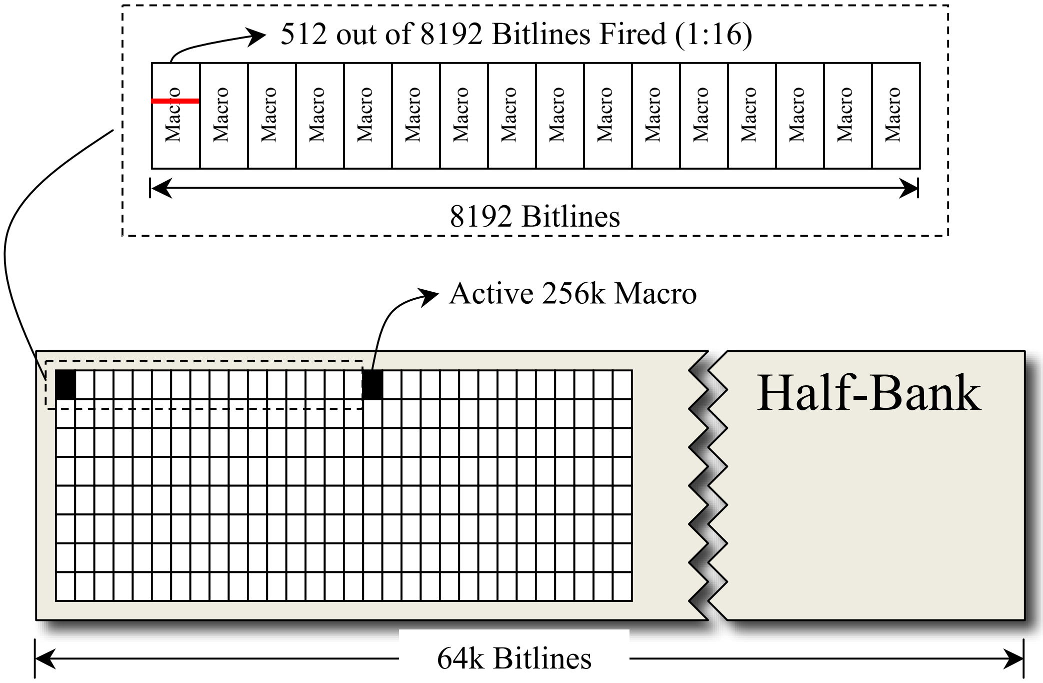


Figure 4.6 – Demonstrating 16:1 Decoding

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 standard 4 Gb DRAM architecture allocated 100F (F denotes the minimum feature size, in this case 40 nm) space for the bit line sense amplifier region [19]. The local I/O signals are differential, requiring 64 data lines per bit line sense amplifier. The space allocated for the bit line sense amplifiers was approximated as 100F. We can reduce this challenge by segmenting the bit line sense amplifiers to above and below the 256 kb memory array. The 32 I/O signals required by the bit line sense amplifier is still a challenge considering only 4 local I/O signals are being used in current DRAM architectures.

In the case of the architecture that utilizes 64 data pins, Figure 4.7 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 word line drivers. Each of the 32 local I/O signals found in a bit line sense amplifier must span this 800 µm width to keep a 23.4 µm pitch on the global I/O.

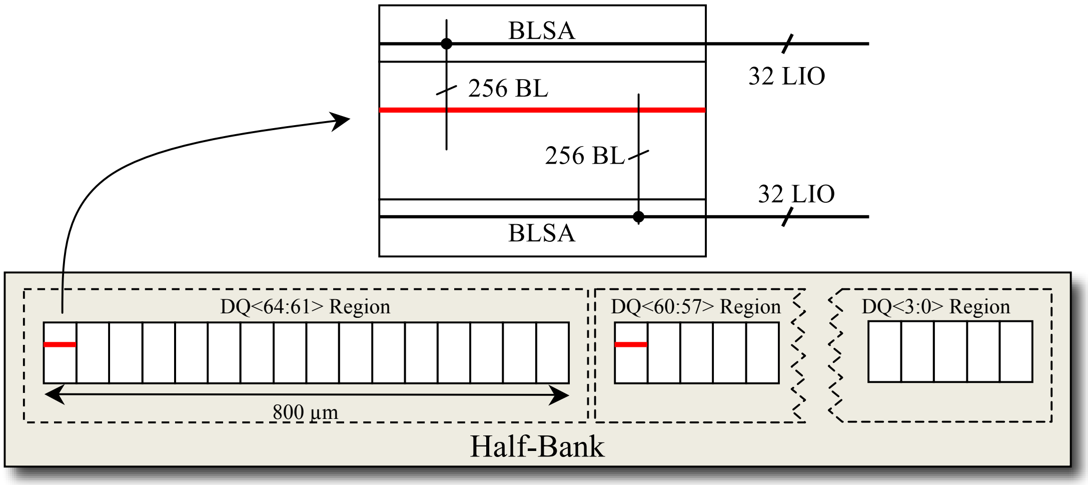


Figure 4.7 – Local I/O Routing

Typically, the local I/O signals are routed in the tungsten metal 0 layer due to no bit lines being used in the bit line 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 capacitive-coupled interconnects.

The wide I/O DRAM architecture will access one half page of memory per half-bank. Firing a word line performs the task of accessing the half page. The additional page decode bits can be used to select one of 16 256 kb memory macros. The other 15 unselected memory macros will have their bit lines equilibrated to a fixed voltage. It is possible to use the unused bit lines to route local I/O signals through the unused 256 kb memory macros. Figure 4.8 shows how the page decode region can be mapped through unused memory macros and how this can reduce the local I/O routing challenge.

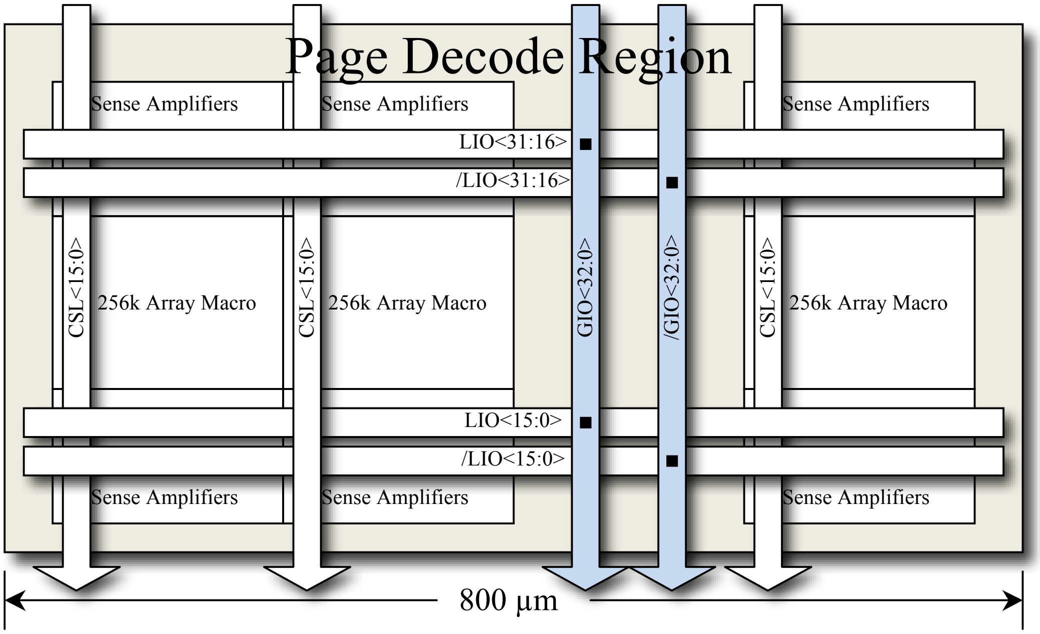


Figure 4.8 – Local and Global I/O Routing

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. Figure 4.9 shows how the wide I/O architecture developed in this chapter can be sliced into many identical slices.

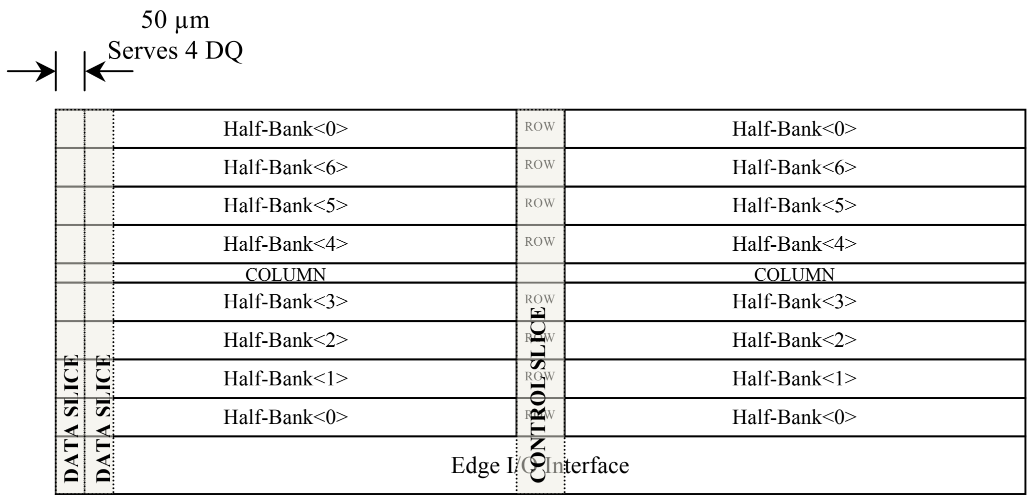


Figure 4.9 – Demonstration of SLICE Architecture

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 capacitive-coupled scaling, the slice can incorporate more input data pins and increase the serialization circuitry in the data path. Figure 4.10 shows the implementation of a data and control slice depicting power routing, block placement, block size, I/O signals, and control signal routing.

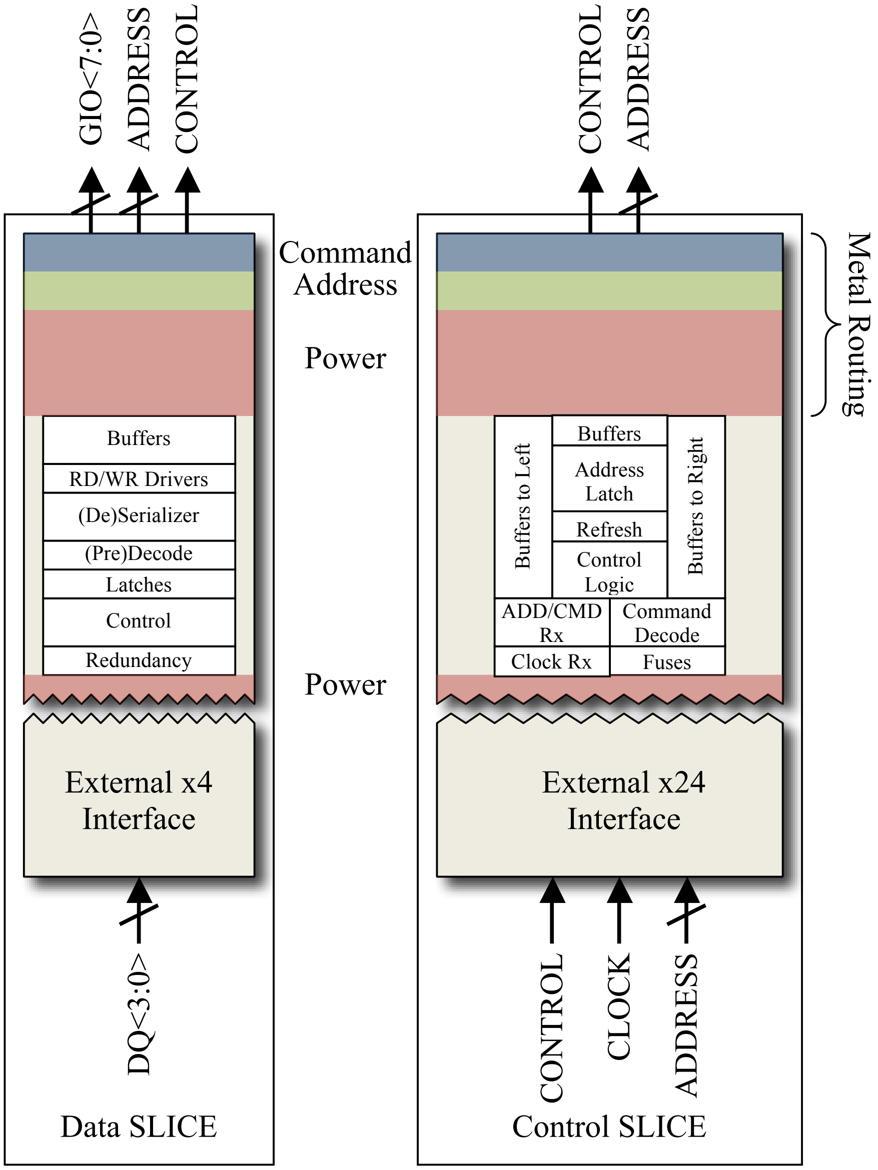


Figure 4.10 – Data and Control SLICE

Summary

Developing a wide I/O DRAM architecture that is suitable for capacitive-coupled interconnects requires the communication channel to be moved to the side of the DRAM chip. This enables a capacitive-coupled wide I/O 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 capacitive coupling with 32 data pins. The developed DRAM architecture placed a page size specification of 8k that 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 [21]. The wide I/O architecture discussed here 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.

The wide I/O DRAM architecture utilizing capacitive-coupled interconnects enables several technological advantages over existing DRAM architectures. Figure 4.11 compares the energy per bit and bandwidth of conventional DRAM modules that use eight data pins per DRAM chip with the wide I/O architectures discussed here. Fixing the page size and increasing the I/O count through the capacitive-coupled interconnects wide I/O DRAM architecture allows for an energy efficient DRAM architecture.

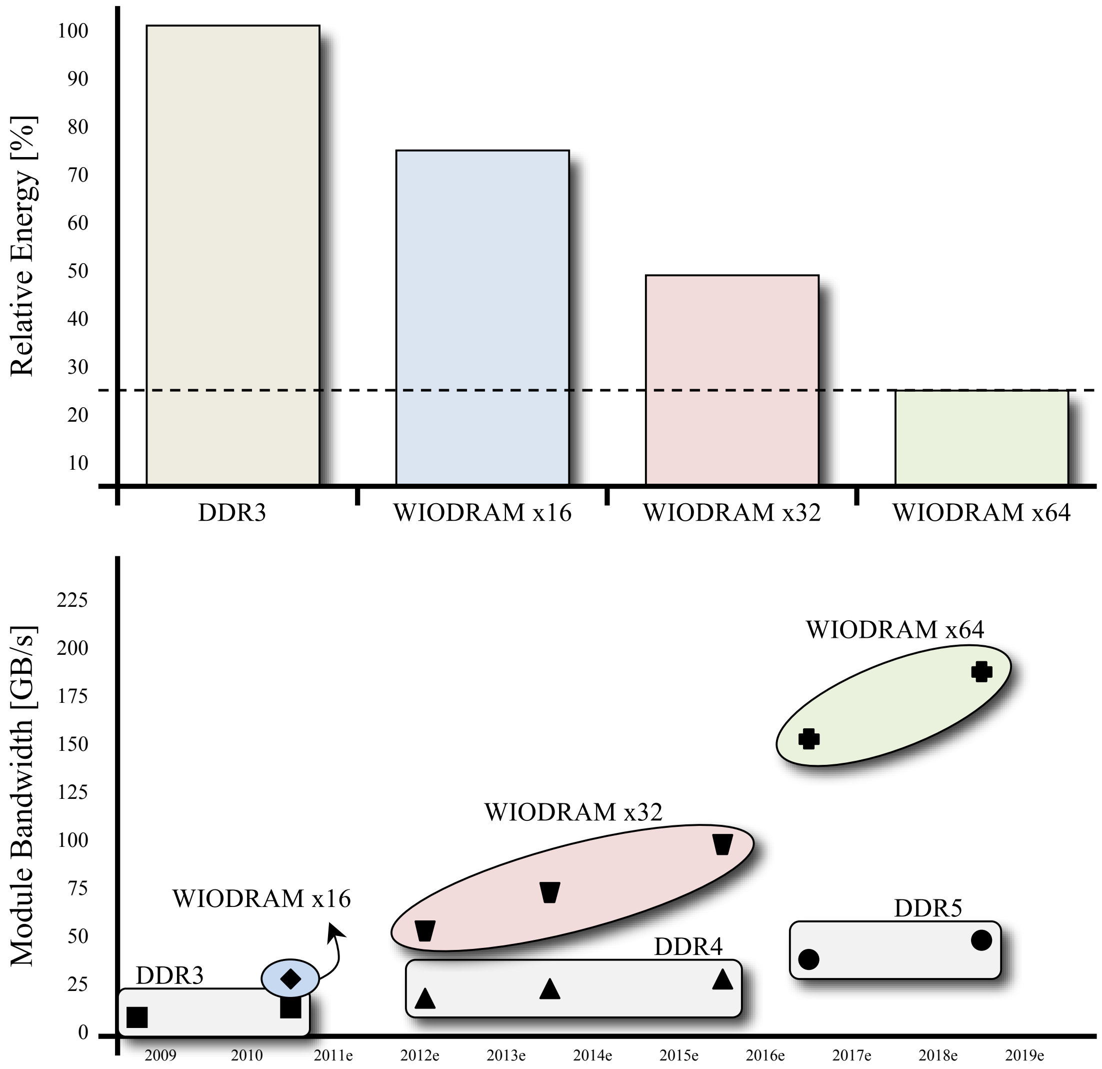


Figure 4.11 – Energy per Bit and Chip Bandwidth Estimates

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 4.11 also shows the energy efficiency advantage of using capacitive-coupled interconnects DRAM compared to conventional DRAM architectures.

Although it is possible to only access one capacitive-coupled interconnect 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 data bus width. The projected bandwidth trends shown in Figure 4.11 clearly depict the advantage of using capacitive-coupled DRAM over current and future DRAM technologies.

CHAPTER FIVE – HIGH BANDWIDTH INTERCONNECT

Using 64 data pins in a DRAM is impractical due to the large power consumed by conventional DRAM I/O. AC coupled interconnects provide a path to a low power alternative.

AC coupled interconnects have been used in a variety of ways over the years. Electrocardiography used coupling connects to determine the heart beat of individuals because the beating of the heart induces small voltage changes on the skin. AC coupled interconnects are also used for serial-deserial (SERDES) applications to remove the DC signal as data is transmitted between two different voltage domains. Automobiles use AC coupling to prevent damage to electronics if the battery is connected improperly. It is the inherent attribute of AC coupling that it removes the DC signal, which is exploited by automobile and SERDES designers. AC coupling was first introduced in semiconductor memories via low cost interconnect used in multi-chip modules (MCM).

AC Coupled Interconnects

AC coupled interconnects occurs when two IC chips are placed close together. The top level of metal is used to create an active component (capacitor or inductor). The AC component is used to transmit data between the two chips. Figure 5.1 shows a diagram of an AC coupled interconnect between two chips [24].

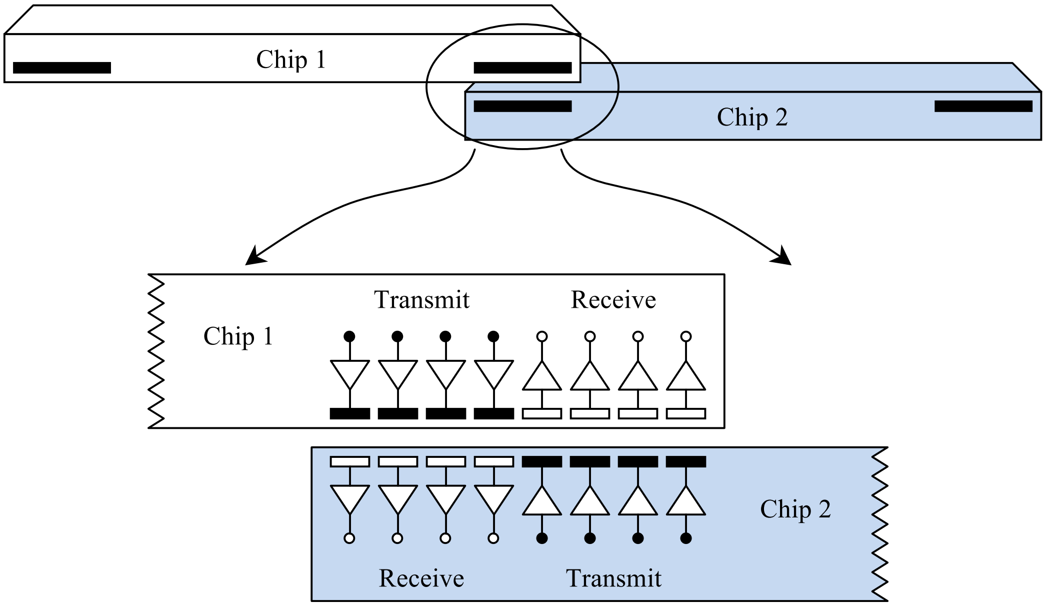


Figure 5.1 – AC-Coupled Interconnect

As seen in Figure 5.1, the top level of metal on each die, or a redistribution layer, is used to form a capacitor once the die are brought together. In the same way, an inductor can be formed on each chip and the signals can be transmitted through a type of transformer. The advantages of AC coupled interconnect are an increased bandwidth, low power, and low cost.

The increased bandwidth comes from several key innovations of the technology. Since the passivation layer is still present over the pads, the ESD structures can be removed or made smaller because there is not a physical connection. The pad sizes can be reduced because a suitable capacitance value can be obtained with a smaller pad size. The use of the bonding wire found in typical off-chip interconnections sets the dimensions of the larger pad sizes. The reduction of the ESD structures also reduces the parasitic capacitances present on the pad. This means that the signals can be transmitted at a higher rate because of the reduced capacitive loading.

The removal of a physical connection to an external voltage also allows for the use of standard threshold voltage devices on the input receiver. A standard practice is to use high threshold voltage devices as the input devices for all signals coming from an external voltage, such as the I/O pins. This causes receiver designers to increase the size of the input devices to overcome the reduction in transconductance gain found when using high threshold devices. The increase in device size requires the use of higher biasing currents (higher power) to fully switch the larger devices capacitances.

The lower capacitance of the chip-to-chip interconnect also reduces the power consumed by the transmitter circuits. When the transmitter and receiver circuits are designed properly, the high power consumed by the termination resistors can also be removed. These configurations allow AC coupled interconnects to consume much less power.

Cost is reduced by the removal of the relatively high cost wire-bonding steps. Further, since the die will not be physically connected together, the cost to remove a defective part in a multi-chip module (MCM) can be reduced. Typically, if a single die in an MCM is found to be defective and the goal is to replace the defective part, all of the solder connections must be removed. A defective die is then removed and a working die is placed in its place, then the MCM is reconnected. The cost associated with these high force and high temperature operations can be removed when die are simply glued together.

Contributions of Salzman and Knight

Capacitive-coupled interconnects were introduced to IC components in 1994 with a paper by Salzman and Knight [25]. Their work described the use of capacitive-coupled interconnects to reduce the cost of testing bare silicon die before assembling them into a MCM. They used fuzz buttons to power the device under test. Figure 5.2 shows a diagram of their testing configuration [24].

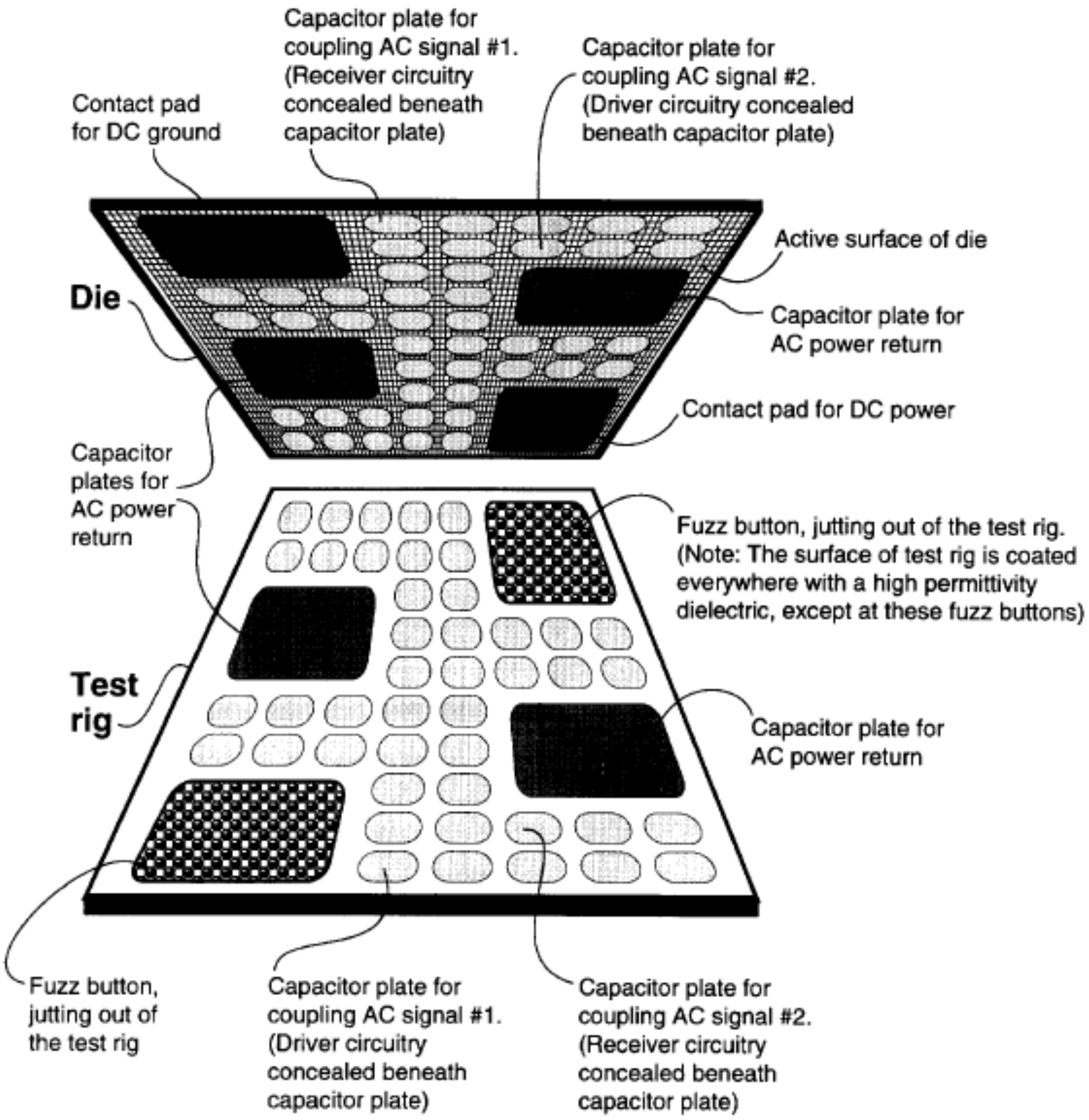

Capacitor plate for coupling AC signal #1. (Receiver circuitry concealed beneath capacitor plate)
Capacitor plate for coupling AC signal #2. (Driver circuitry concealed beneath capacitor plate)
Contact pad for DC ground
Active surface of die
Die
Capacitor plate for AC power return
Contact pad for DC power
Capacitor plates for AC power return
Fuzz button, jutting out of the test rig. (Note: The surface of test rig is coated everywhere with a high permittivity dielectric, except at these fuzz buttons)
Test rig
Capacitor plate for AC power return
Fuzz button, jutting out of the test rig
Capacitor plate for coupling AC signal #1. (Driver circuitry concealed beneath capacitor plate)
Capacitor plate for coupling AC signal #2. (Receiver circuitry concealed beneath capacitor plate)

Figure 5.2 – Testing Known Good Die with Capacitive-Coupled Interconnects

Knight and Salzman began introducing new ways of using capacitive-coupled interconnects in MCM configurations. It was found that an MCM could be built that utilized capacitive-coupled interconnects. The advantages of the technology were shown to be a higher bandwidth and lower cost associated with building an MCM.

When capacitive coupling is used, the transmitted signal becomes a function of *dV/dt* (rate of change) rather than the exact voltage being transmitted. This allows the same signal to be transmitted regardless of the operating voltage. This has significant implications when die using multiple power domains are used. It becomes possible to transmit between two die with differing voltage domains without the use of complicated level shifting circuitry.

The higher bandwidth also became apparent when realizing that the size of the parallel plates can be made smaller than conventional pads used with wire bonding. The area of the parallel plates has to be approximately 280 times the distance between the plates to achieve a 10 fF capacitor. Given a 1 µm separation, the pad size would have to be 16 µm by 16 µm. Given a separation of 10 µm, the pad size would have to be 53 µm by 53 µm. These values are considerably less than the 100 µm by 100µm size of traditional I/O pads. The values above result in an inherent bandwidth increase, due to sheer size, of 4000% and 400%, respectively. As the pad size is reduced, more pads can occupy the same area allowing for an increase in bandwidth.

Prior Art

Since the inception of capacitive-coupled interconnects, there have been several research programs describing different applications for the new interconnect technology. Franzon, along with Salzman and Knight, described the first application past an MCM [26]. They first determined that switch fabrics were a good application for capacitive-coupled interconnects. The receiver design, discussed in the paper, describes the use of pulse signaling at the input of the receiver. Figure 5.3 shows the discussed receiver design [24].

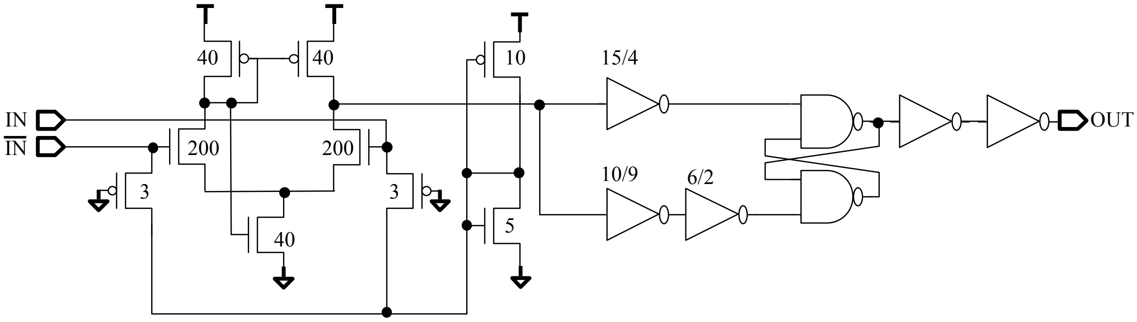


Figure 5.3 – Initial Receiver Design for Capacitive-Coupled Interconnects

The receiver uses a self-biased differential amplifier to provide slight amplification and a level shift of the input signals. It is reported that this design could be used with differential signaling or single-ended signaling. The shorted inverter biases the input nodes through the always-on PMOS devices. The output of the differential amplifier is fed to two asymmetric inverters. The outputs of the inverters are sent to a set-reset latch. The use of the set-reset latch allows for the transmission of long strings of ones or zeros. The design works for pulsed signaling, but consumes higher power with the self-biased differential amplifier.

Additional publications by Wilson et al. and Salzman et al. describe the benefits and challenges of using capacitive-coupled interconnects [27, 28]. A major published challenge is the delivery of power connections to the die. Franzon alleviated the issue using silicon trenches filled with solder balls. Franzon's receiver designs culminated with the use to two different receivers as seen in Figure 5.4 [25, 26].

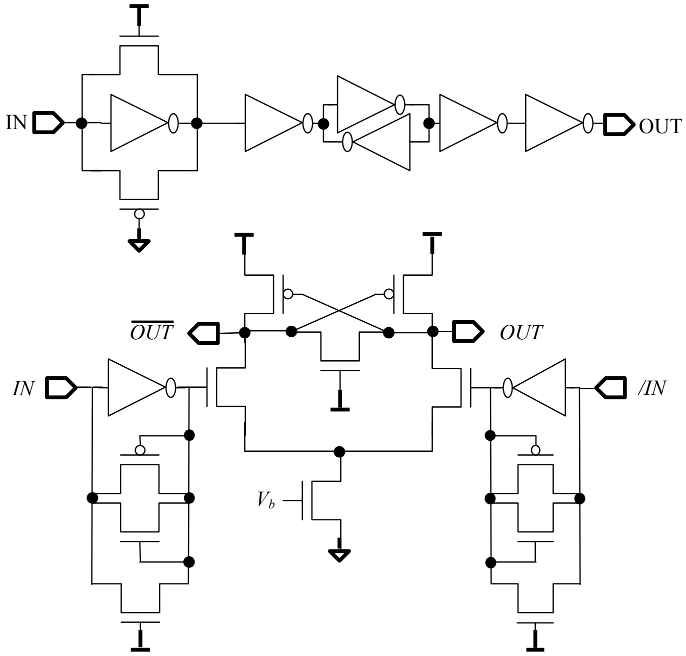


Figure 5.4 – Two Different AC Coupled Receiver Designs

The first schematic shows the use of an inverter that is biased in its high gain region by the shorted input inverter. The negative feedback NMOS and PMOS devices help create a DC bias and allow the output of the inverter connected to the input to amplify the input signal. The second inverter is clamped at its high gain region to amplify the input signal. The latch is used to latch the input signal. The biasing inverter is improved in the second receiver by adding the additional transistors connected to the inputs. These devices are used to short the input inverter to create a DC bias and to clamp the signals from going too low or high (causing inter symbol interference). The latch is used to convert the return-to-zero (RTZ) signal to a non-return-to-zero (NRZ) signal. The two designs suffer from excess power consumption due to the first and second stage of the receivers being biased in their high gain region and consume excess power.

Proposed Receiver Design

A major advantage of capacitive-coupled interconnect is their high I/O count due to smaller pad sizes. The size of the receivers and transmitters must be small in order to accommodate the high I/O count. The designs described thus far can only be used in devices that can consume large amounts of power. This limits the applications to a space of designs that does not need to use capacitive-coupled interconnects. Memory components could use the extra bandwidth and low power I/O but must use I/O circuitry that consumes less energy than 1 pJ/bit at 1 GHz (1 mA/GHz/bit). This can be accomplished by removing the need to use two stages of the design that consume large amounts of power because they are biased in their high gain region. A better design for capacitive interconnects would consider the circuit seen in Figure 5.5.

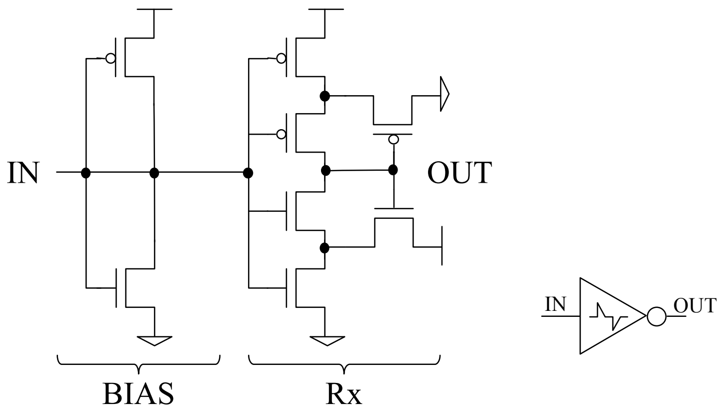


Figure 5.5 – Low Power Capacitive-Coupled Receiver

The bias inverter is designed to bias the Schmitt trigger in its high gain region; but unlike the previous designs, the Schmitt trigger latches to a fixed state and does not consume excess power due to its input bias. The Schmitt trigger behaves as a latch for RTZ signaling (pulse signaling), making it ideal for capacitive-coupled interconnects. The RTZ latching takes effect when the input of the Schmitt trigger is set to be in-between the switching points of the Schmitt trigger. Figure 5.6 shows the transient response of the low power receiver design. In this demonstration, the input signal transitions are differentiated across the coupling capacitor and the differentiated signal appears on node B. The Schmitt trigger latches the RTZ signal once it switches through its high/low switching regions. The output of the Schmitt trigger is a NRZ signal that successfully recovers the transmitted signal.

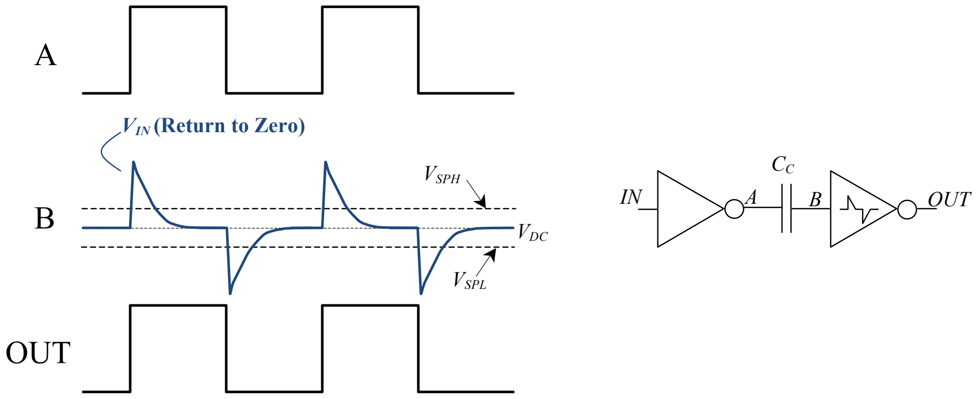


Figure 5.6 – Transient Response of the Low Power Receiver Design

This design effectively receives DC (static 1 or static 0) signals. The Schmitt trigger behaves as a latch on an edge transition and holds the value of a logic high or logic low until the input signal transitions to a different state. In this way, the capacitive-coupled interconnect behaves as a DC connection for logic pulses.

The advantage of this topology is that it does not require two stages that consume large amounts of power. The output of the Schmitt trigger is stable at logic high or low, substantially reducing the power of the circuit. The resistive discharge path through the bias circuit must have a time constant less than ½ of the bit rate to prevent inter-symbol interference (ISI). ISI refers to an attribute of a transmitted signal where adjacent transmitted signals interfere with each causing distortion in the signal

When designed properly, the bias inverter will consume a small amount of static power. The bias inverter can be designed to consume very little power (< 1 µA) but this induces ISI into the input signal, because the DC bias is not stable before the next symbol. The ISI can be tolerated in some applications where low power and low bandwidth are specified.

The power consumed by this circuit is a function of the number of gates tied to the output of the Schmitt trigger. As technology shrinks, the value of this capacitance also shrinks. This is due to the physical area of the gate capacitance reducing. This is the desired power scaling direction for the receiver design. The bias circuit power consumption also shrinks as technology evolves.

The reduction in capacitance attached to the input signal allows for a corresponding increase in the bias resistance for the same time constant. A practical implementation of this receiver can be used with feedback that changes the hysteresis and biasing resistor depending on the sensed coupling capacitor.

Providing proof of concept of this receiver design is performed in two different technologies. A test chip was fabricated in 0.5 µm technology to give proof of concept. An additional test chip was fabricated in a 65 nm technology to give proof of scalability.

0.5 µm CMOS Design (Proof of Concept)

The initial silicon test chip used ON Semiconductors 0.5 µm provided by MOSIS to demonstrate the capabilities of the receiver design. On chip polysilicon capacitors were used to create the coupling capacitor. Attempting to communicate between two chips utilizing capacitive-coupling with unproven transmitters, receivers, capacitors, and alignment techniques posed a significant risk. For this reason, on chip capacitors were used to verify the novel receiver design. Figure 5.7 shows the schematic of a test structure used to characterize the design.

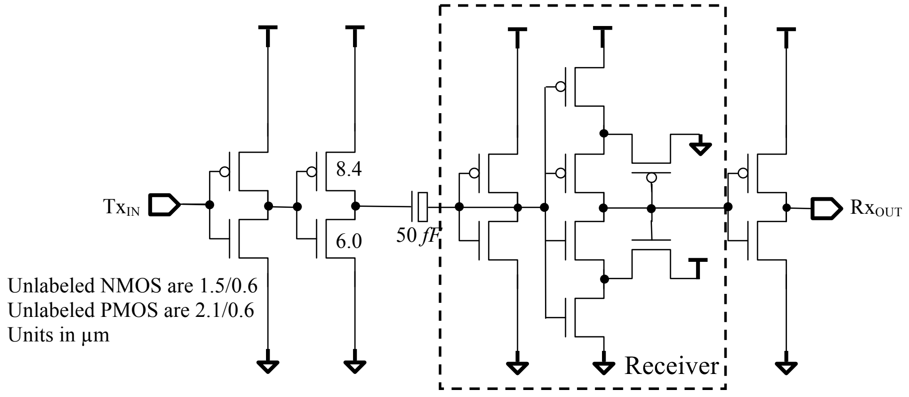


Figure 5.7 – Test Structure of the 0.5 µm Receiver Design

The test structure comprises a transmitter that is simply an inverter sized properly to drive the 50 fF coupling capacitor. The bias circuit uses minimum length devices (L = 0.6 µm) with PMOS width of 2.1 µm and NMOS width of 1.5 µm. The ISI is reduced by designing the biasing leg to have a resistance that creates a time constant with the coupling capacitor of approximately ½ of the bit rate.

When the resistance of the bias devices is larger, the current consumption of the biasing leg can be substantially reduced. The reduced power is achieved by reducing the timing margin for latching the input signal in the data path. DRAM chips use an 8-bit burst with a write pre-amble. The worst-case ISI will result when comparing two different input waveforms. The first is when an input pattern of seven symbols of one type (1 or 0) and the last symbol being the alternative versus its opposite pattern. If these patterns can be tolerated, it is best to use the low power approach.

The Schmitt trigger uses the same width and length for all of the MOSFETs. The Schmitt trigger has a high switching point of 3.2 V and a low switching point of 1.8 V. The bias circuit provides a 2.5 V bias to the Schmitt trigger, which is biased at half of the power supply (5 V). The DC current consumed by the receiver is 240 µA, 120 µA by the DC bias circuit, and 120 µA by the hysteresis devices in the Schmitt trigger.

The design was created to operate at 5 V and 200 Mbps. The design consumes ~ 8 pJ/bit total power (transmitter and receiver). The DC current passing through the bias circuit and hysteresis devices consumes 75% of the energy, while the switching components consumes the other 25%. Figure 5.8 shows the simulation results of a 100 MHz signal being transmitted through a 100 fF capacitor. The top pane of Figure 5.8 shows the input signal, the middle pane shows the input signal after the coupling capacitor, and the bottom pane shows the receivers output signal.

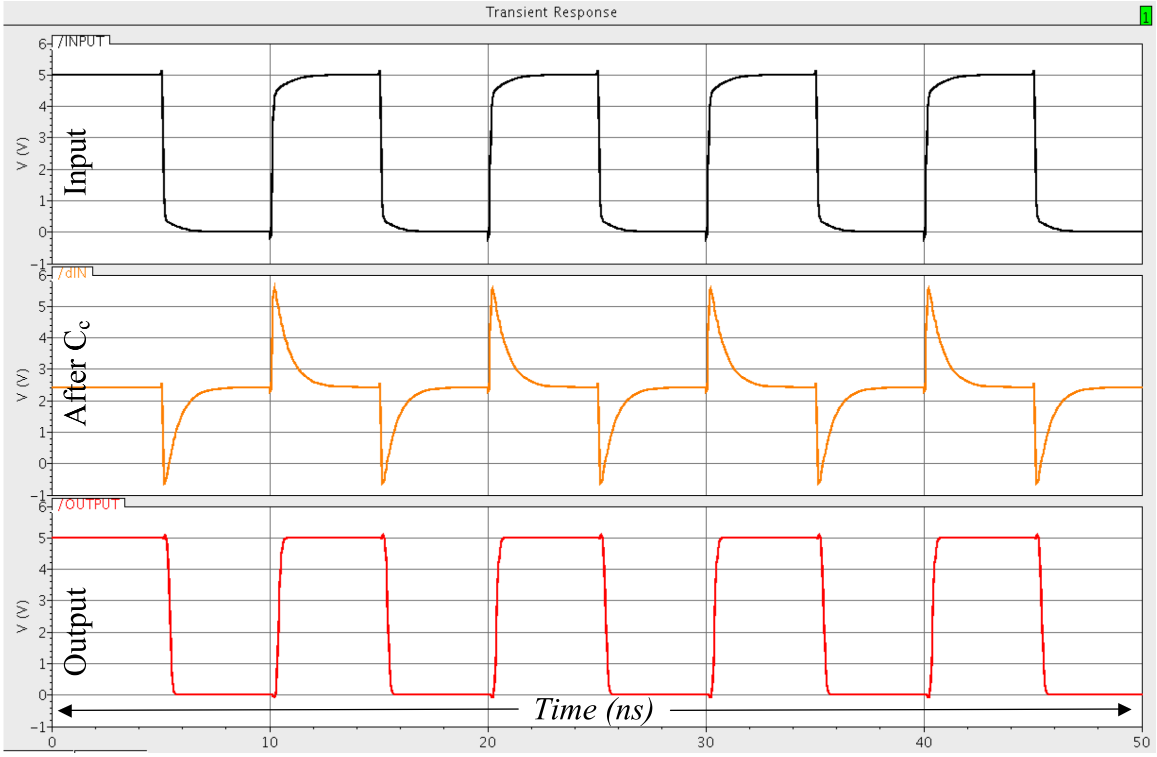


Figure 5.8 – Simulation Results of the Receiver

The coupling capacitors were created on the test chip using poly-to-poly capacitors. These capacitors have a large parasitic substrate capacitor approximately equal to 2/3 of the capacitor size. Our 50 fF capacitors had a substrate capacitor equal to

40 fF according to the parasitic extraction of the layout. This capacitor value would reduce the signal being transmitted through the capacitor due to the capacitor divider presented to the coupling capacitor. The connections to the coupling capacitor could be reduced (place the parasitic capacitance at the transmitter side), but this would inject unwanted substrate noise through the bottom plate of the capacitor. For this reason, the bottom plate was connected to the input receiver in an attempt to reduce substrate noise injection.

2.3.1 Chip Layout

The test chip contained test structures using 50 fF, 100 fF, and 250 fF capacitors. Multiple receivers were designed that had varying hysteresis and device sizes. Figure 5.9 shows the layout of a typical test structure containing a 50 fF coupling capacitor.

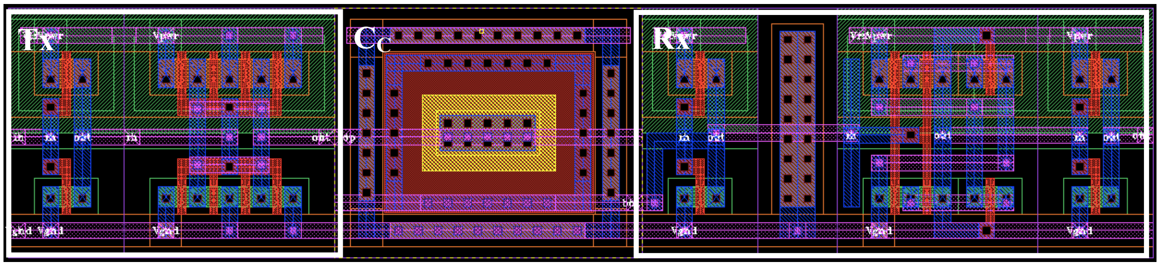


Figure 5.9 – Layout of the 50 fF Test Structure

The size of the receiver is 32.55 μm x 18.9 μm (an area of 615.195 μm^2). The layout shows that the voltage supply of each element was isolated from the others so that the power consumption could be measured independently during testing. The test chip included varying test structures that vary the coupling capacitor and switching characteristics of the receiver design. Table 5.1 lists the test structures contained in the test chip.

Table 5.1 – List of test structures on the 0.5 µm test chip

Structure	Type	Coupling Capacitor (fF)	Switch Device Width (µm)	Hysteresis Device Length (µm)
A	DC	-	2.1/1.5	0.6
B	DC	-	2.1/1.5	1.2
C	DC	-	4.2/3.0	1.2
D & E	Transient	250	2.1/1.5	0.6
F & G	Transient	100	2.1/1.5	0.6
H & I	Transient	100	2.1/1.5	1.2
J & K	Transient	100	4.2/3.0	1.2
L & M	Transient	50	2.1/1.5	0.6
N & O	Transient	50	2.1/1.5	1.2

The test chip measured 1.5 mm × 1.5 mm and contained the 15 test structures listed in Table 5.1. Decoupling capacitors were used internally between *VDD* and *GND* to filter out high frequency switching noise. Figure 5.10 shows a picture of the full chip layout.

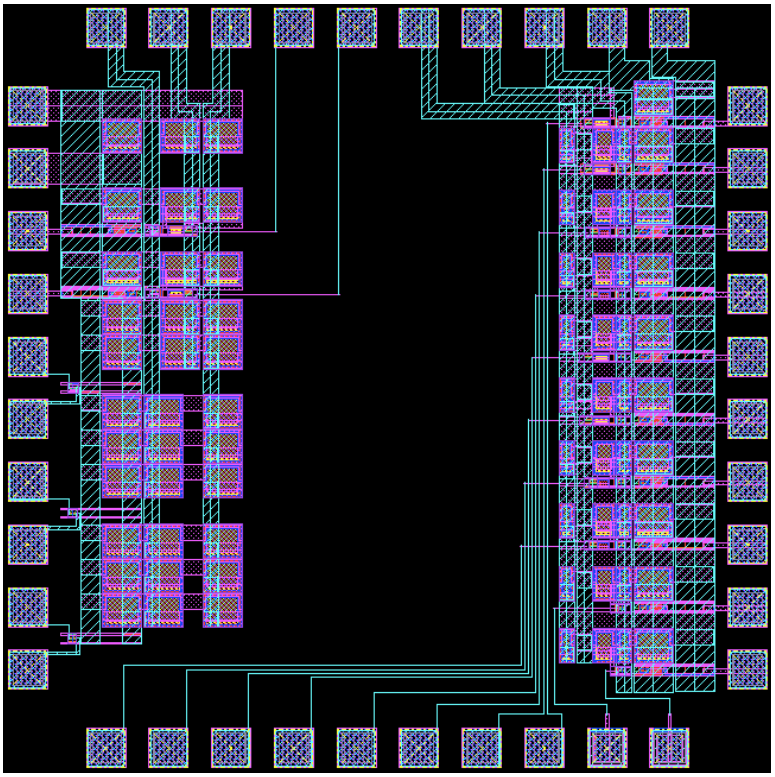


Figure 5.10 – Layout of the 0.5 µm Test Chip

Experimental Results

The silicon chip was packaged to a dual inline package (DIP) 40 package at MOSIS. Figure 5.11 shows the micrograph of the 0.5 μm test chip after wire bonding to the DIP 40 package.

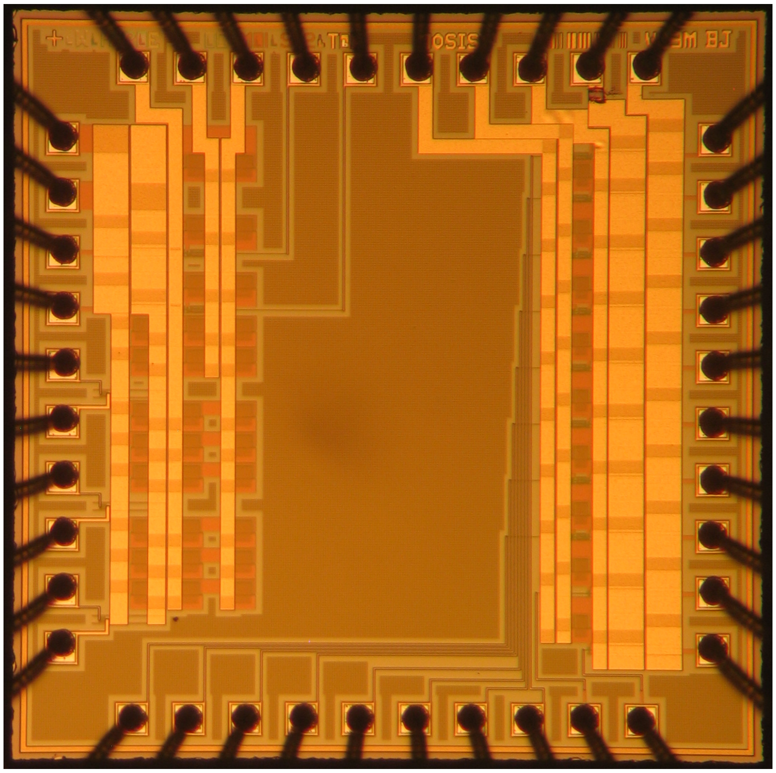
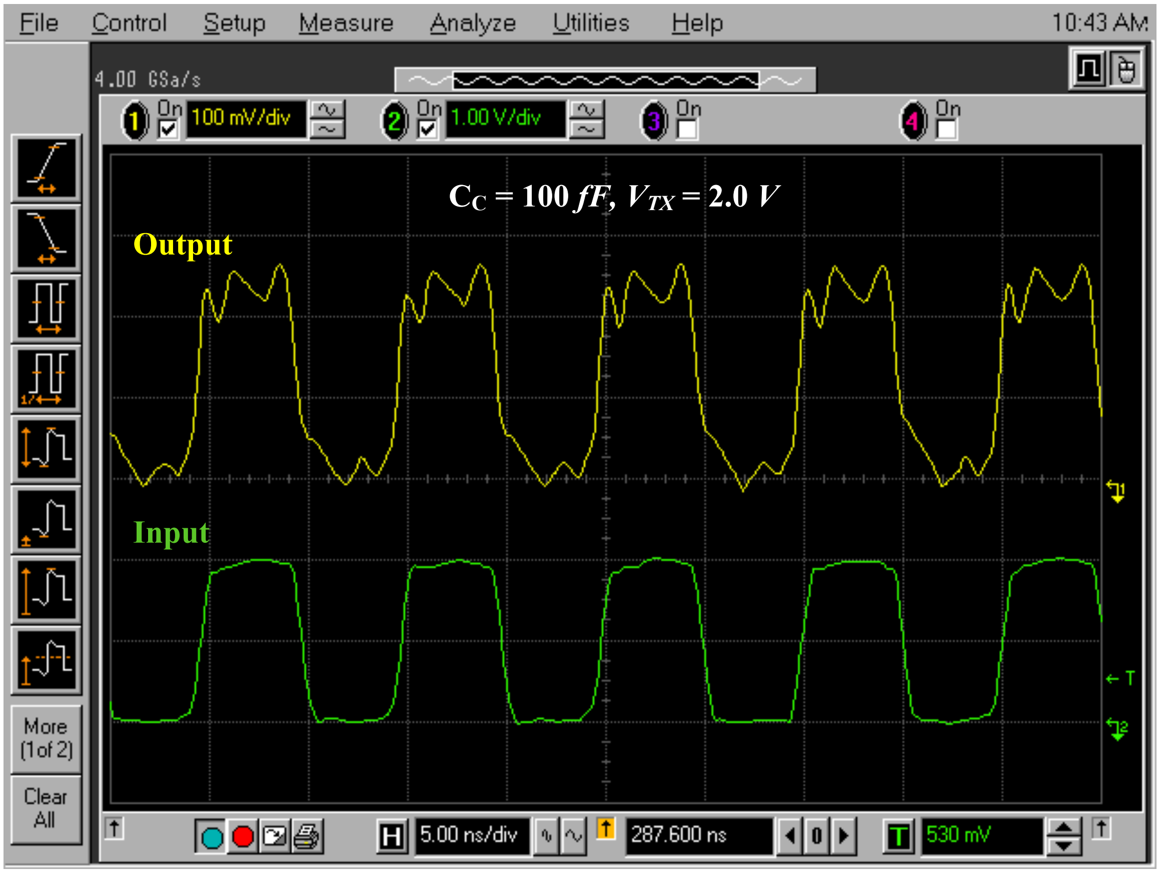


Figure 5.11 – Microphotograph of 0.5 μm Test Chip

The test results gathered from silicon matched simulation results. The design transmitted and received data reliability using transmitter voltages ranging from 2.0 V – 6.0 V and a receiver voltage of 5.0 V. This gives proof that the design can work between chips with different power supply values. Using 2.0 V as the transmitter voltage shows the design can operate with less than the simulated 8 pJ/bit of energy at 200 Mbps. Experimental results showed that the design worked down to a transmitted voltage of only 1.3 V, allowing for a reduction of energy from 8 pJ/bit to approximately 3 pJ/bit at

200 Mbps. The test set up used a simple breadboard to connect the packaged chip to the test equipment.

Figure 5.12 shows the results of the test chip running at 200 Mbps. The output signal was driven off chip through a resistive divider to isolate the large parasitic capacitance of the bonding wire, test board, and scope probe. The simulated output voltage swing was 200 mV, which is verified with the waveforms in Figure 5.12.



4.00 GSa/s

1 On ☑ 100 mV/div

2 On ☑ 1.00 V/div

3 On □

4 On □

H 5.00 ns/div

287.600 ns

T 530 mV

Figure 5.12 – Test Results of Using 100 fF Coupling Capacitor

The results shown in Figure 5.12 used a coupling capacitor of 100 fF to successfully transmit and receive data at 200 Mbps. A DC power supply of 5 V was used to power the receivers and the rest of the test chip. A pulse generation circuit was used to generate the input square wave and used a 2.0 V peak-to-peak transmitter voltage (V_{TX}).

The ringing of the output signal in Figure 5.12 was due to the inductance of the breadboard. Figure 5.13 shows the test set up used to generate the waveforms.

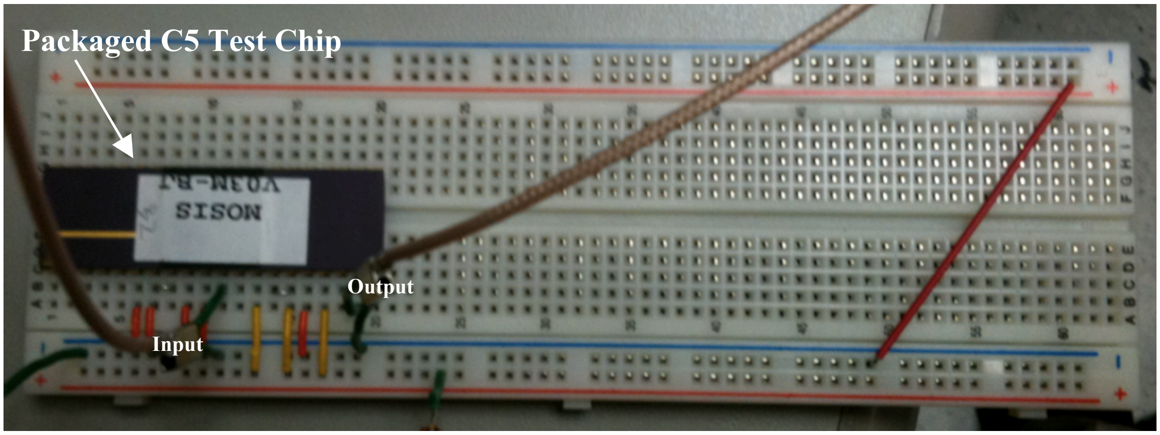


Figure 5.13 – Test Setup Used to Gather Data

The test setup was fairly straightforward. Once power measurements were taken, all of the chips power supplies were shorted together. The input signals were applied directly to the test chip. The output signal was fed directly to the oscillators input. Figure 5.14 shows the equipment used to generate the waveforms.

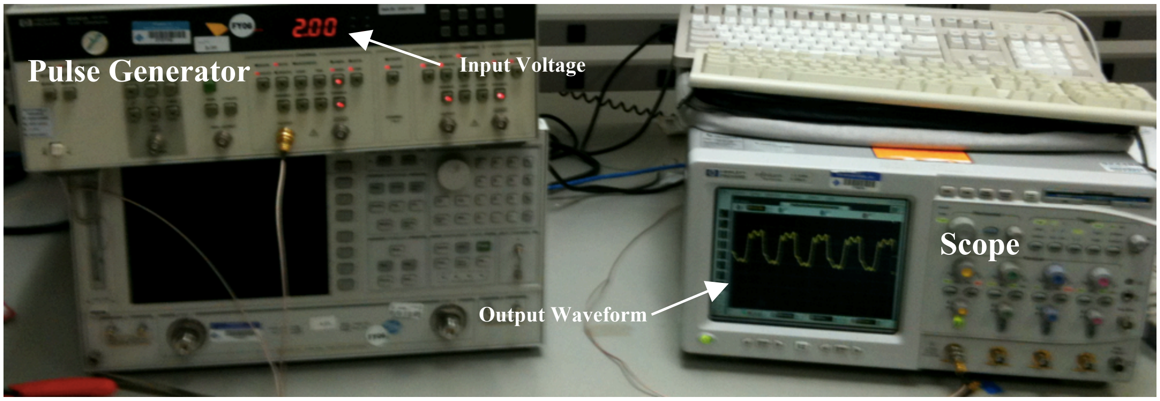


Figure 5.14 – Equipment Used to Gather Data

The 0.5 μm design gave proof of concept on a low power and robust design at 200 Mbps. The next design was fabricated in a 65 nm process to show the ability of this topology to scale with technology scaling.

65 nm CMOS Design (Proof of Scalability)

The 65 nm design used the same topology for the transmitter and receiver designs as the 0.5 µm design. The 65 nm design used inter-digitated metal-to-metal capacitors for the coupling capacitor and was optimized for higher frequency operation. The coupling capacitor was changed to remove the large parasitic substrate capacitance of the poly-to-poly capacitor. This, along with the smaller device sizes, allowed the coupling capacitors to be scaled down to 15 fF.

Increasing main memory bandwidth by using a wide I/O architecture requires the use of receivers that consume less than 100 fJ/bit at data rates above 1 Gbps. The 65 nm design used a 1.2 V power supply and operated at 4 Gbps. A phase locked loop (PLL) and pseudo random binary sequence generator (PBRS) was used to generate an onboard 2 GHz signal and an output data eye at 2 Gbps. Simulation results show the receiver design consumed less than 15 fJ/bit of energy at 4 Gbps. This is two orders of magnitude reduction in energy consumption of typical DRAM I/O circuitry that consumes 1 pJ/bit at 2 Gbps – 4 Gbps. Figure 5.15 shows the schematic of the 65 nm receiver design.

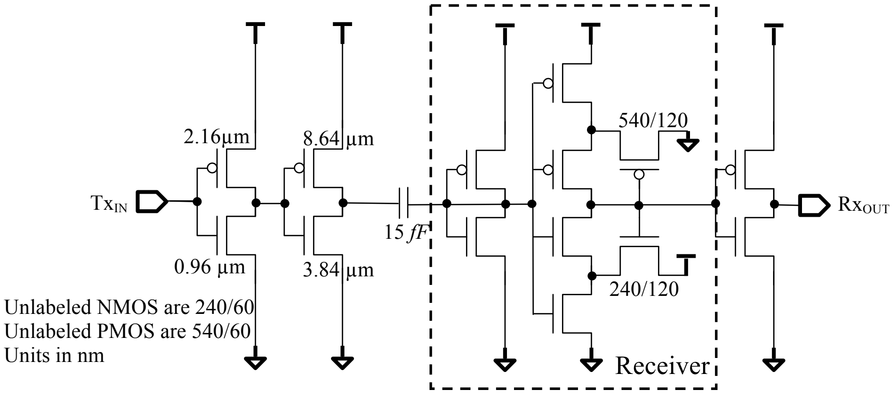


Figure 5.15 – 65 nm Receiver Design

The DC current consumed by the biasing circuit and the hysteresis devices of the Schmitt trigger was less than 20 µA. At a power supply of 1.2 V and a bit rate of 4.0

Gbps (1 bit every 250 ps), the static energy is 6 fJ per bit. The dynamic energy is a function of the capacitance connected to the output of the Schmitt trigger and the clock frequency. Each gate and junction connection loads the output of the Schmitt trigger by 2 fF. Assuming 10 connections to the output of the Schmitt trigger gives a total capacitance of 20 fF. The dynamic power of the Schmitt trigger is calculated as $P = CV^2f$. This equation gives a dynamic power of approximately 60 µW consumed every 250 ps or an energy per bit of approximately 15 fJ/bit at 4 Gbps. These values match the simulation results. The layout of the receiver design measures 3.73 µm by 4.72 µm (area of 17.64 µm^2) and is shown in Figure 5.16.

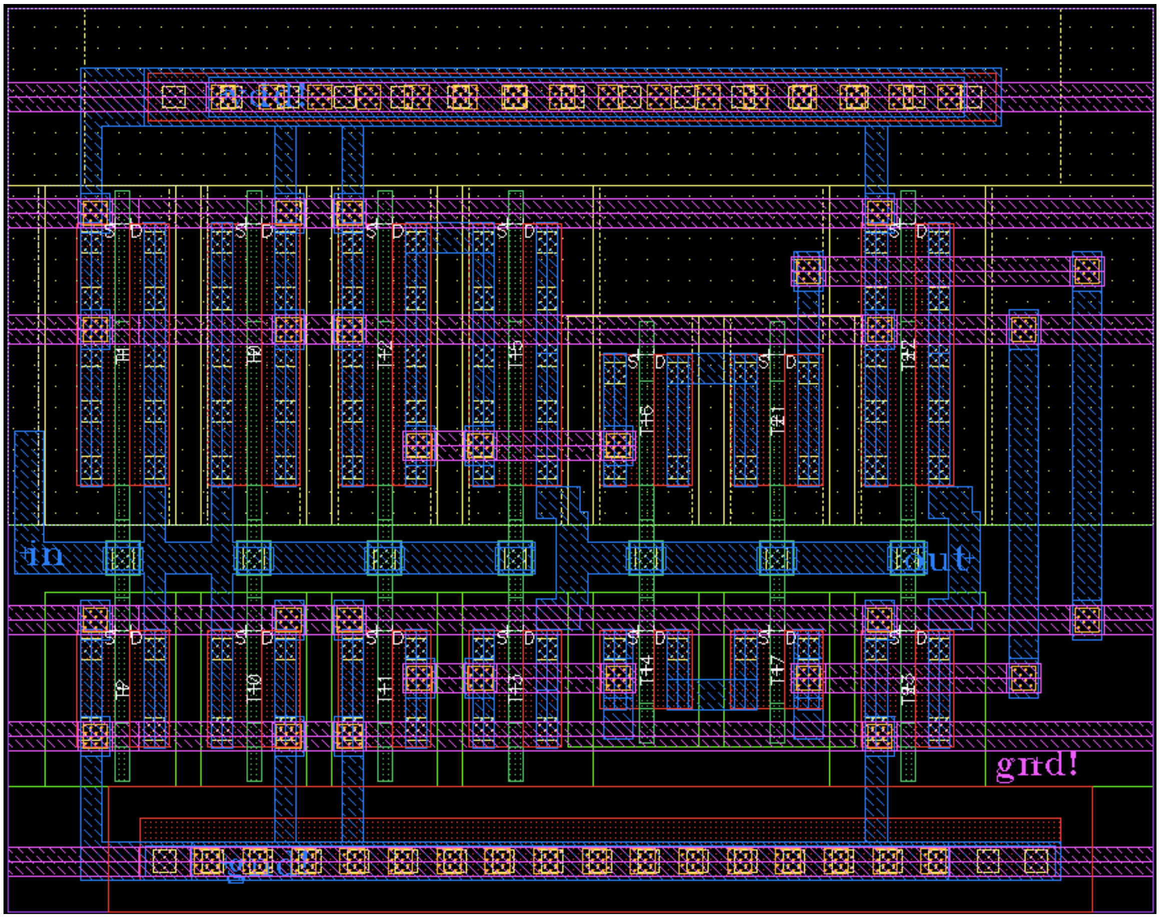


Figure 5.16 – Layout of the 65 nm Receiver Design

The receiver design used minimum length devices (L = 60 nm) for all devices. The output of the Schmitt trigger was fed to a minimum sized inverter (Wp = 540 nm, Wn = 240 nm) to minimize the capacitive load. Figure 5.17 shows the simulation results using a receiver optimally designed for a 15 fF coupling capacitor. The top pane of Figure 5.17 is the input signal, the middle pane is the input signal after the coupling capacitor, and the bottom pane is the output of the receiver.

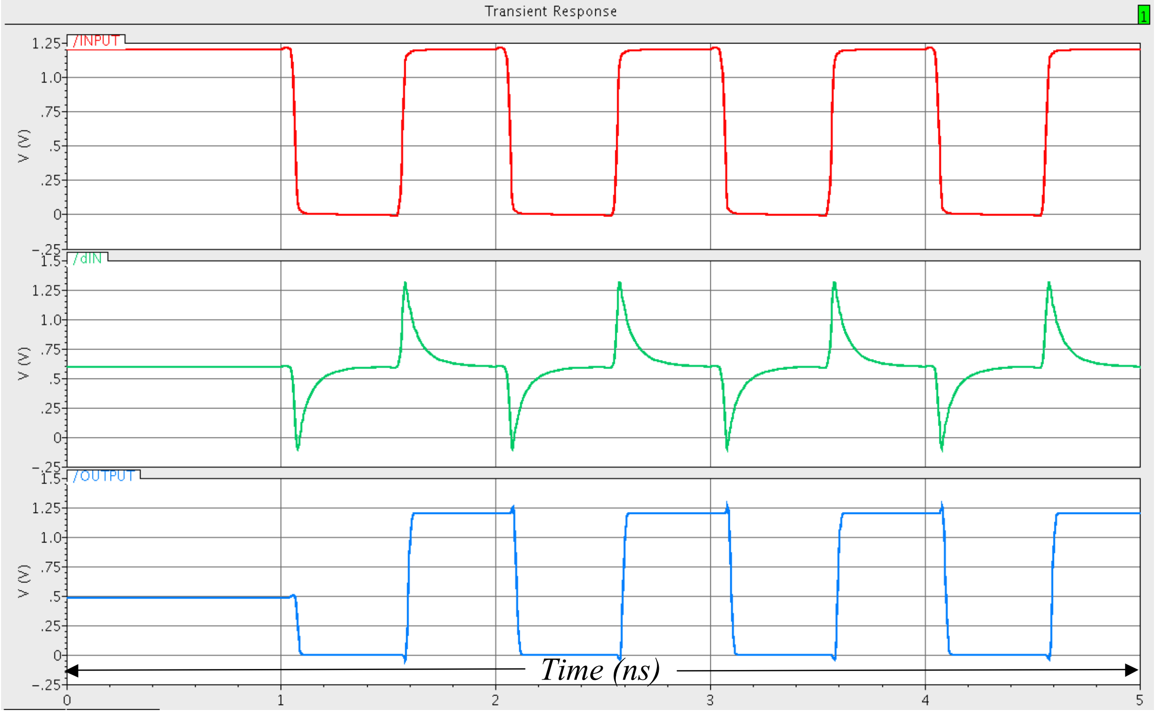


Figure 5.17 – Simulation Results of the Receiver

The simulation results show the receiver parking in an unknown state before input edges are present. This would not be the case during normal power up on a silicon die due to the latch flipping to a known state when the power supplied is slowly ramped up to its final value. In the case of simulations, when the power supply begins at its final value, the receiver requires one edge to successfully latch into a known state. The receiver design was fabricated in IBM's 65 nm low power process to prove the operation of the low power receiver.

Substantial periphery circuitry was required to test the high-speed operation of the 65 nm design. The major component of the peripheral circuitry is the phase locked loop (PLL).

Phase Locked Loop

The PLL was needed to generate a high frequency signal that could be transmitted through the on-chip capacitive-coupled interface in the full-chip simulations. The PLL was manufactured with the receiver design in the unlikely case that a high frequency signal generator was not available. The PLL was required to synthesize a 6.25 MHz input clock into a 2.0 GHz output clock. The design needed to have a 400 MHz lock range, over damped response (no jitter accumulation), and produce a relatively low jitter output clock. The design also needed to be uncomplicated, robust, and stable. For these reasons, a textbook design using a phase frequency detector with charge pump output was used [29]. The block diagram of the PLL with each element's gain (denoted by *A*) is seen in Figure 5.18.

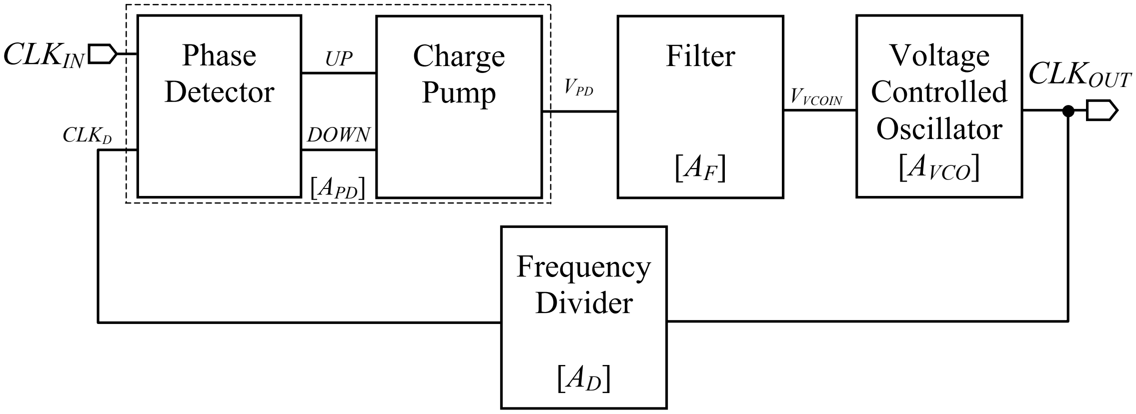


Figure 5.18 – Block Diagram of PLL

The PLL contains a phase frequency detector (PFD), charge pump (CP), filter, voltage controlled oscillator (VCO), and divider circuit. The blocks work together to output a clock that is phase locked (coincident edges) with the input clock, and to generate an output clock that has a larger frequency than the input signal. The block diagram can be analyzed as a negative feedback system with closed loop gain equal to:

$$A_{CL} = \frac{A_{OL}}{1 + A_{OL}\beta}$$

Where β is the feedback factor (in our case A_D) and A_{OL} is the open loop gain (in our case $A_{PD}\cdot A_F\cdot A_{VCO}$). The feedback of the PLL allows for precise control of the system and must be analyzed to ensure that the feedback signal does not add to the input signal. If the feedback signal is added to the input signal, the system becomes unstable and will not operate properly. The following section details the design of the over damped second order PLL design.

Phase Frequency Detector

The purpose of the PFD is to measure the difference in phase of the input clock and the divided down feedback clock. The PFD outputs a signal that is a representation of the phase difference of the two input signals. The circuit works by first detecting which rising edge occurred first. If the input clock edge rises before the feedback clock, the *UP* signal goes high and stays high until the rising edge of the feedback clock arises. If both edges occur at the same time, then both outputs (*UP* and *DOWN*) remain low. Figure 5.19 shows the CMOS implementation of the PFD [27].

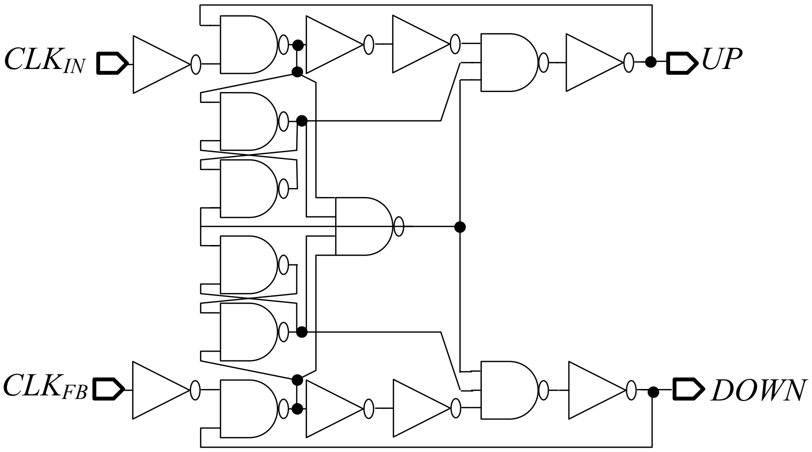


Figure 5.19 – CMOS Implementation of the PFD

Figure 5.20 shows the simulation results of the CMOS implementation of the PFD. In the simulation, the *data* signal refers to the input clock (purple) and the *dclock* signal refers to the feedback clock (red). The simulation begins with the rising edge of *data* arriving before the rising edge of *dclock*, which sets the *UP* output high (yellow). When the rising edge of *dclock* arrives at the PFD, the *UP* signal is set low. The simulation also shows the case of *dclock* arriving first, which sets the output *DOWN* high (blue).

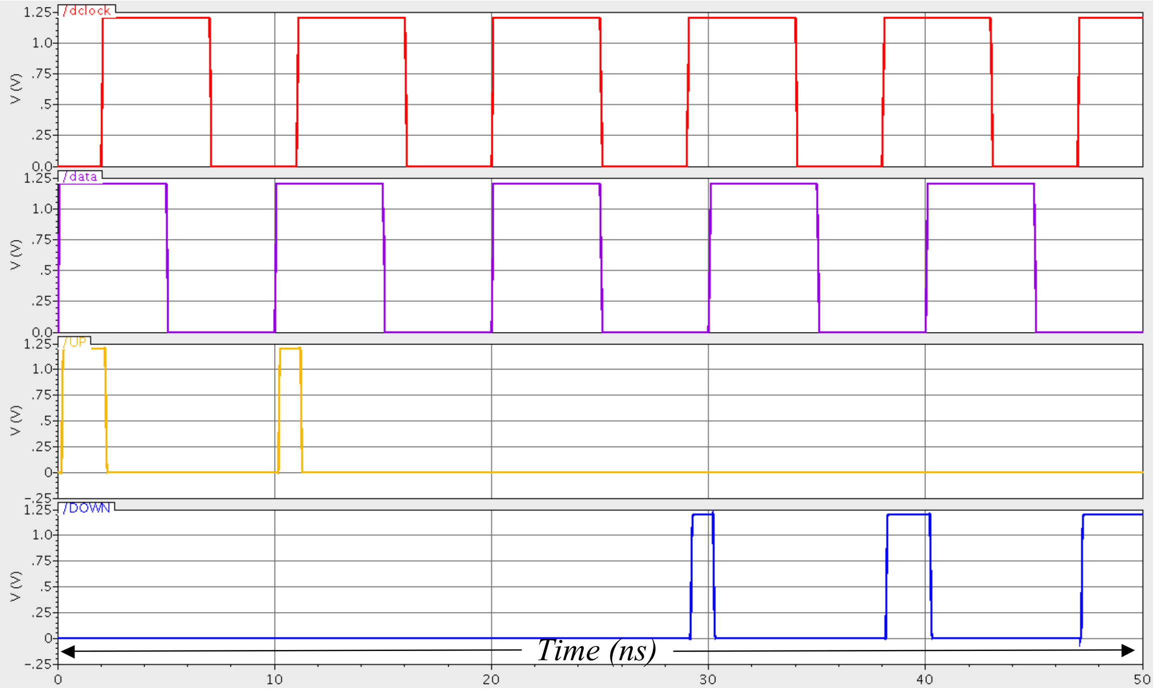


Figure 5.20 – Simulation Results of the PFD

The PFD is used in conjunction with a charge pump and loop filter to set the input voltage of the VCO. The *UP* and *DOWN* signals are combined and translated into a current (I_{PUMP}) that drives the loop filter. When the *UP* signal is high, the charge pump supplies the full I_{PUMP} to the loop filter. When the *DOWN* signal is high, the charge pump removes the full I_{PUMP} from the loop filter. Figure 5.21 demonstrates this concept by plotting the transfer characteristics of the charge pump.

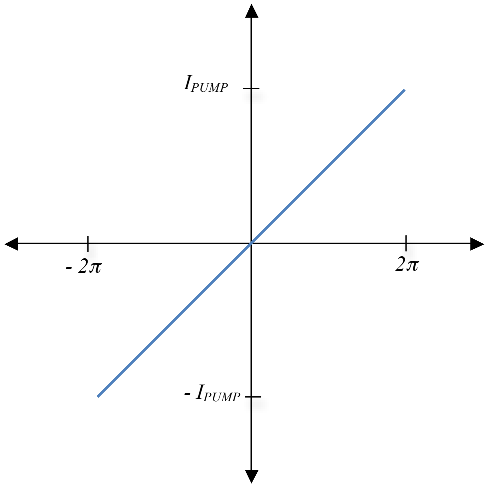


Figure 5.21 – Transfer Characteristic of the PFD and CP

Figure 5.21 shows that the gain of the PFD is:

$$A_{PD} = \frac{I_{PUMP}}{2\pi}$$

The pump current was designed to be 10 µA and the simulation results give a pump current of 12 µA. This sets A_{PD} = 1.91 µA/radian. The value of I_{PUMP} is set by the charge pump design. The charge pump combines the output signals of the PFD and translates them into a current.

Charge Pump and Filter

The charge pump used in the PLL is responsible for converting the PFD output signals (*UP* and *DOWN*) into a current. When *UP* is asserted, the charge pump will add current to the filter and when *DOWN* is asserted, the charge pump will remove current from the filter. This action allows for precise control of the input voltage to the VCO. The charge pump used in this design can be seen in Figure 5.22 [27].

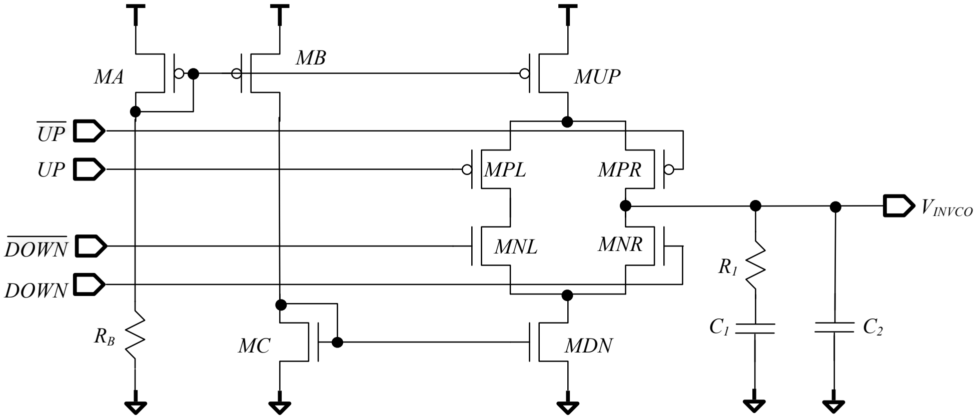


Figure 5.22 – Implementation of the Charge Pump and Filter

The charge pump works by first developing a bias current (I_{PUMP}) with resistor R_B and the gate-drain connected PMOS device *MA*. The current is mirrored over to *MB* and used to create an *NMOS* current mirror. When designed correctly, devices *MUP* and *MDN* will source/sink I_{PUMP}. The currents in *MUP* and *MDN* are steered through the charge pump depending on the state of the input signals. The addition of transistors *MPL* and *MNL* are used to keep devices *MDN* and *MUP* biased correctly when both inputs are low.

Figure 5.22 also depicts the filter design used in the PLL. Capacitor C_2 is used to filter out voltage jumps due to I_{PUMP} and R_1 and is approximately 10 times smaller than C_1. The filter's resistor and capacitor are used to stabilize the loop and to filter the input signal of the VCO. The gain of the filter is obtained by realizing that the filter has a current input and voltage out and that C_2 can be neglected for all practical purposes.

$$A_F = \frac{1 + sR_1C_1}{sC_1}$$

The values for the passive components in the filter are $R_1 = 50\ k\Omega$, $C_1 = 12\ pF$, and $C_2 = 1\ pF$. These values were selected to give the PLL system an over damped response. The filter provides an over damped voltage to the input of the VCO where it is used to control the output frequency of the VCO.

Voltage Controlled Oscillator and Divider

The VCO design used a five stage current starved oscillator for the VCO design. Figure 5.23 shows a schematic of the VCO [27].

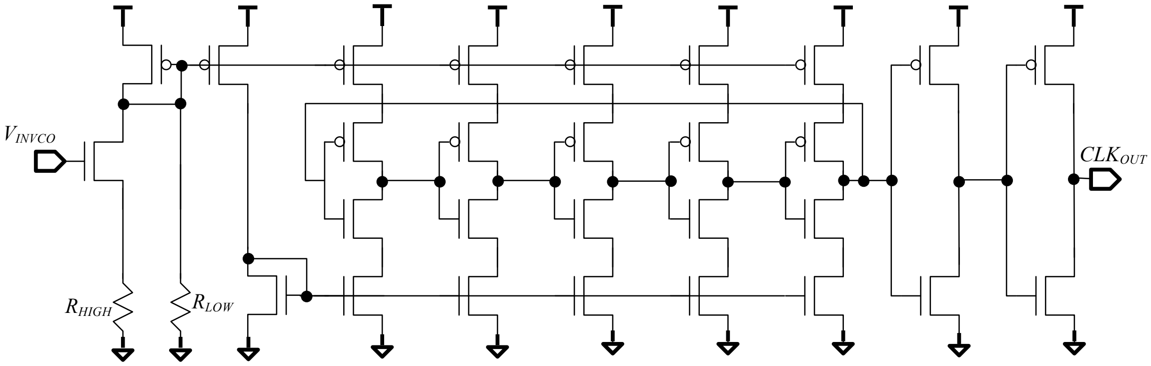


Figure 5.23 – Schematic of the Voltage-Controlled Oscillator

The input voltage is transformed into a current and sets the current supplied to the current starved oscillator. The R_{HIGH} is used to linearize the voltage to current transformation and to set the maximum output frequency. The R_{LOW} resistor is used to set the minimum output frequency of the oscillator (when the input voltage is less than the threshold voltage of the input transistor). The VCO was designed to have a center frequency of 2 GHz with an input voltage of *VDD/2* (0.6 V) and a frequency range of ± 200 MHz. The gain of the VCO is found by analyzing the equation below.

$$A_{VCO} = 2\pi \cdot \frac{f_{MAX} - f_{MIN}}{V_{MAX} - V_{MIN}}$$

Figure 5.24 shows the transfer characteristics of the VCO using simulation data taken at a typical design corner.

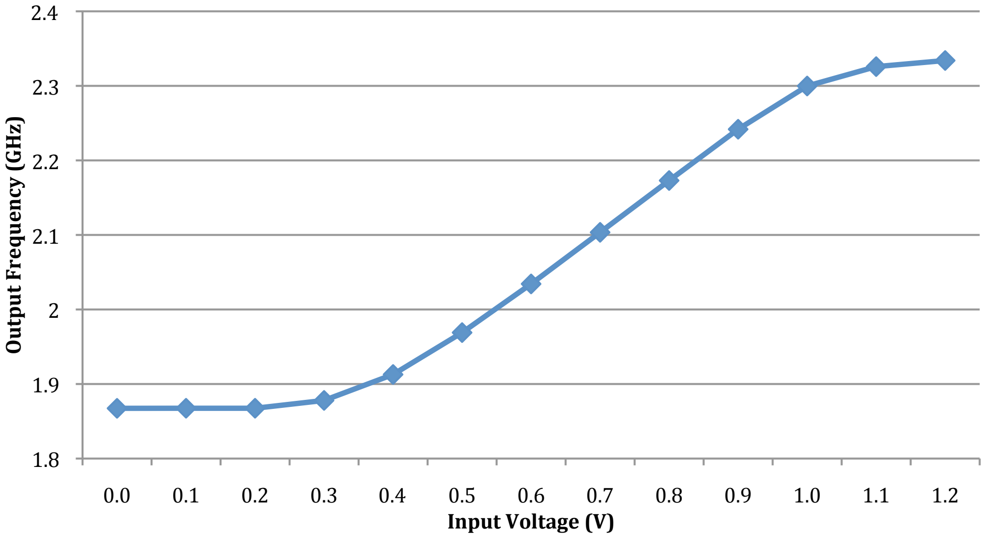


Figure 5.24 – Transfer Function of the VCO

The simulation results show a design that is not centered around the nominal output clock frequency of 2.0 GHz. This is not a problem for the design of the PLL that uses the PFD with charge pump output. The gain of the VCO can be determined by analyzing the transfer function and its value is $A_{VCO} = 3.59 \times 10^9$ radians/V·s. The gain of the VCO is valid in the linear region of the transfer curve seen in Figure 5.24.

The output of the VCO is a frequency and the input to the PFD is a phase. This requires the integration of the VCO output frequency. The phase of the output clock is determined with the following equation.

$$\phi_{CLK_{OUT}} = V_{INVCO} \cdot \frac{A_{VCO}}{s}$$

The output clock frequency is divided down using toggle flip-flops to match the frequency of the input clock. The input clock frequency is 6.25 MHz and the output clock frequency is 2 GHz. This sets the feedback factor of the PLL to 32. The phase of the feedback clock can be determined using the following equation.

$$\phi_{CLK_D} = \frac{1}{N} \cdot \phi_{CLK_{OUT}} = \beta \cdot \phi_{CLK_{OUT}}$$

Loop Characteristics

The open loop gain of the PLL is the product of the gains for each block.

$$A_{OL} = A_{PD} A_F A_{VCO}$$

Using this in the closed loop equation, remembering that the output frequency of the VCO is integrated to get phase, gives the following transfer function.

$$H(s) = \frac{\phi_{CLK_{IN}}}{\phi_{CLK_{OUT}}} = \frac{A_{PD} A_F \frac{A_{VCO}}{s}}{1 + \beta A_{PD} A_F \frac{A_{VCO}}{s}} = \frac{A_{PD} A_F A_{VCO}}{s + \beta A_{PD} A_F A_{VCO}}$$

Remembering that:

$$A_F = \frac{1 + sR_1C_1}{sC_1}$$

$$H(s) = \frac{\phi_{CLK_{IN}}}{\phi_{CLK_{OUT}}} = \frac{A_{PD} A_{VCO} \left(\frac{1+sR_1C_1}{C_1}\right)}{s^2 + s\left(\frac{A_{PD} A_{VCO} R_1}{N}\right) + \left(\frac{A_{PD} A_{VCO}}{NC_1}\right)}$$

The transfer function shows that the PLL design is a second order system and may oscillate. The PLL was designed as an over damped system, which ensures an exponential settling rather than oscillations. The transfer function gives the following second order parameters for the natural frequency and damping factor.

$$\omega_n = \sqrt{\frac{A_{PD} A_{VCO}}{NC_1}}$$

$$\varsigma = \frac{\omega_n}{2} R_1 C_1$$

Using the component values and gains calculated previously, the design has a natural frequency of 4.2×10^9 radians/V·s and a damping factor of 1.27. Figure 5.25 shows the simulation results of the PLL achieving phase lock.

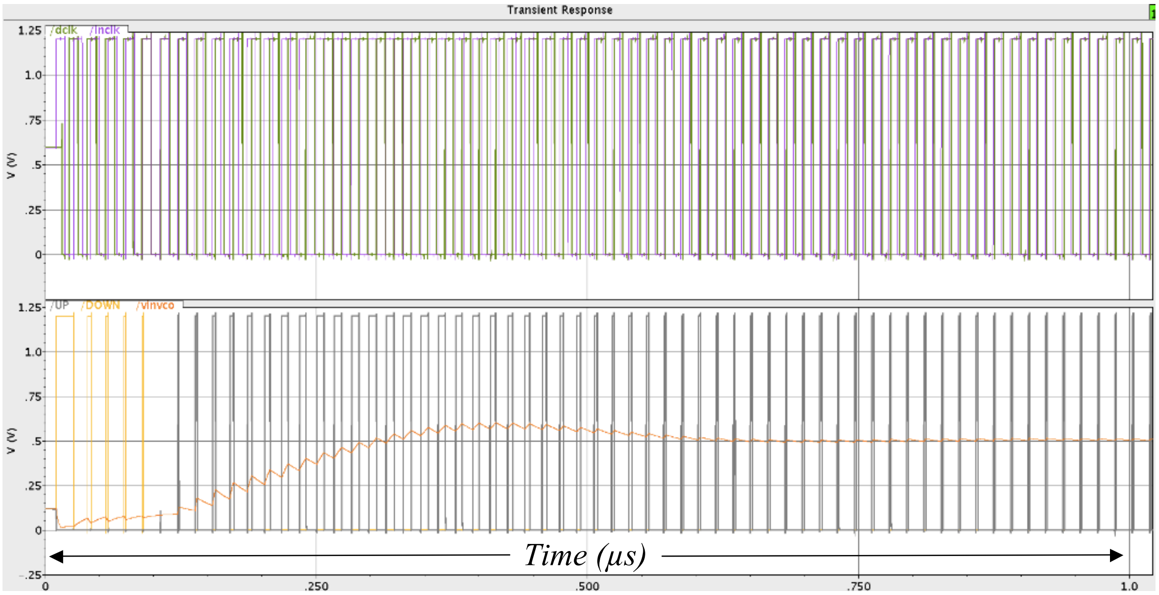


Figure 5.25 – Transient Simulation Showing the PLL Locking

The two signals in the top plane of Figure 5.25 are the input clock (*inclk*) and the feedback clock (*dclock*). The bottom pane of Figure 5.25 shows the PFD control signals (*UP* and *DOWN*) along with the input voltage to the VCO. The two clock signals begin the simulation with an unknown phase relationship. The PFD detects the phase relationship of the two signals and asserts the respective control signal. This causes the charge pump to set the input voltage to the VCO. As the simulation progresses, the ripple on the input to the VCO decreases to zero and the loop achieves lock. Figure 5.26 shows a zoomed plot of the simulation results.

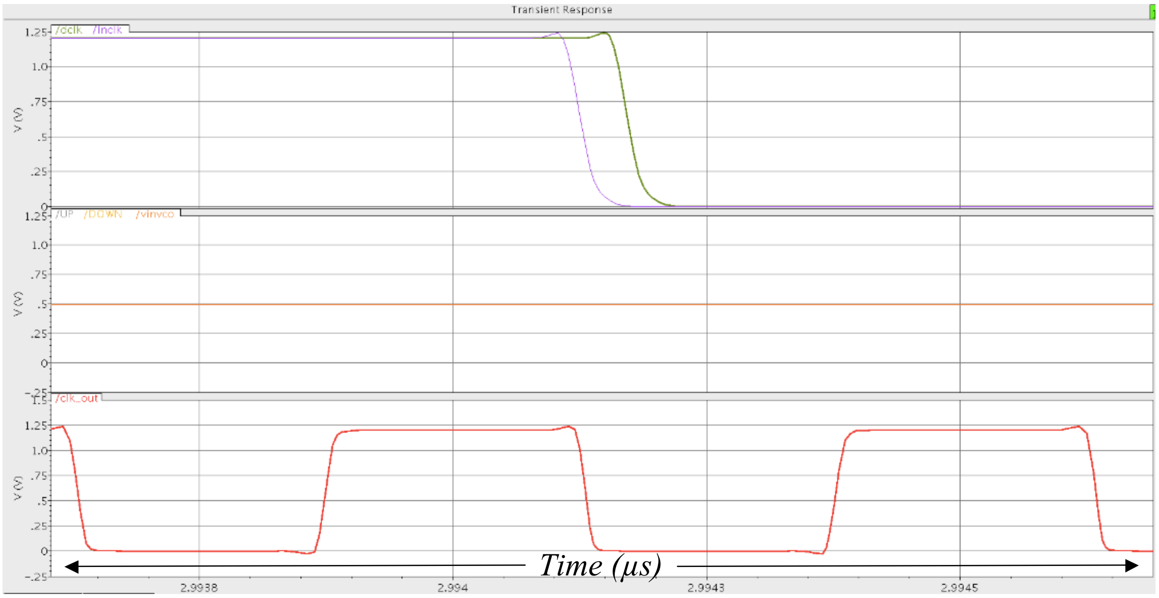


Figure 5.26 – Zoomed Plot of the PLL at Lock

The top pane of Figure 5.26 contains a zoomed plot of the input clock and the feedback clock. The middle pane of Figure 5.26 contains the input voltage of the VCO, and the bottom pane contains the output clock. The simulation results show lock at approximately 3 μs (the *UP* and *DOWN* signals are no longer asserting). The phase difference of the two input signals to the PFD is less than 30 ps and the input to the VCO is steady at 0.5 V. The output jitter of the PLL was characterized at a typical corner (25 °C, *VDD* = 1.2 V, no *VDD* noise, typical process) to analyze the stability and settling of the PLL and was measured at ± 5 ps.

The PLL design used digital and analog components operating at relatively high frequencies. Supply noise, signal-to-signal coupling, and an unbalanced layout could have a large impact on the output jitter of the PLL. For these reasons, the layout of the PLL was critical. The final layout of the PLL can be seen in Figure 5.27. The VCO was placed away from the other circuitry and surrounded by substrate/well ties and decoupling capacitors.

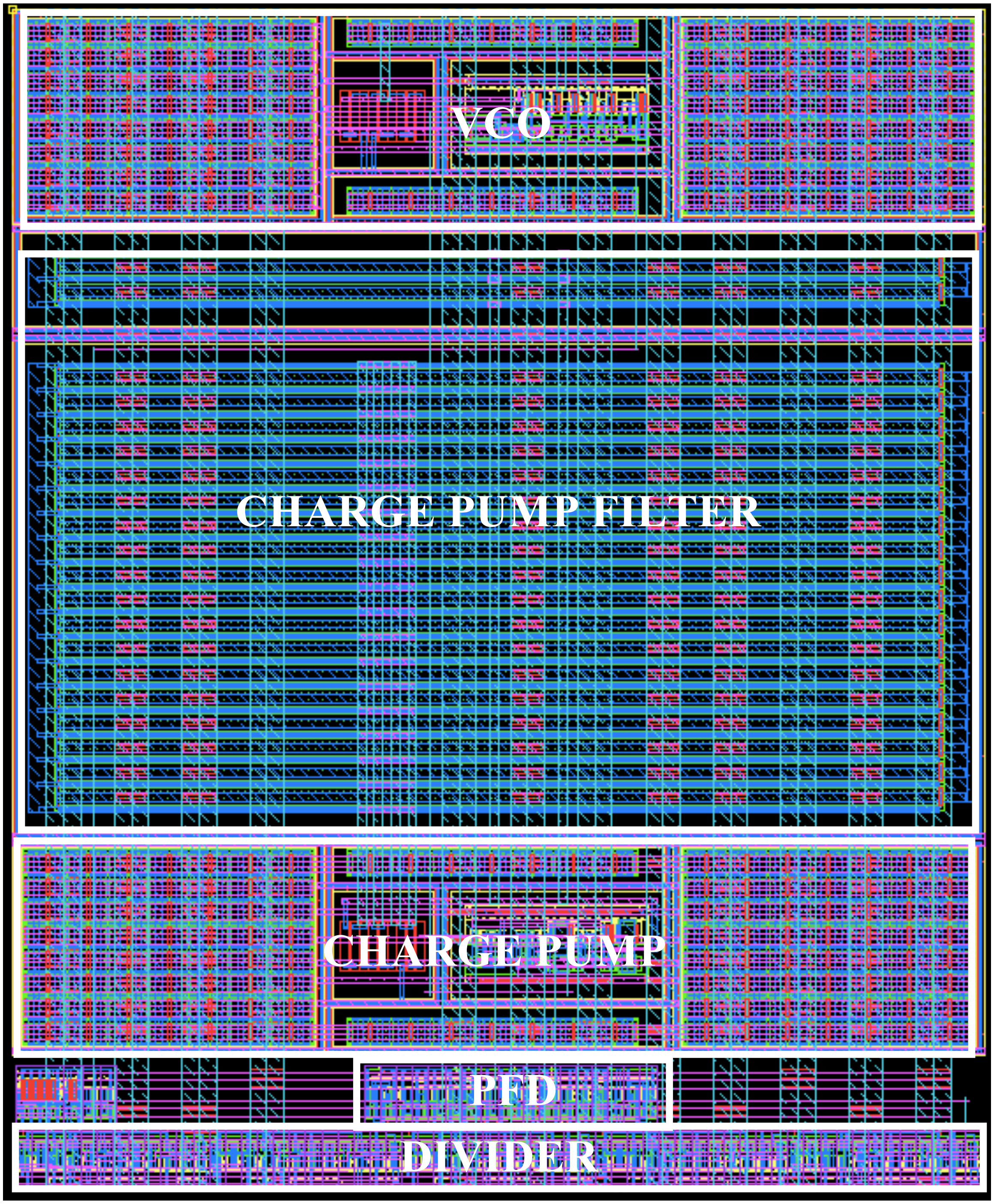


Figure 5.27 – Layout of the 2.0 GHz PLL

The majority of the layout is consumed by the filter capacitor for the charge pump. Decoupling capacitors are used to create a symmetric layout for the charge pump and VCO, along with decoupling noise on the power rails.

Additional Periphery Circuitry

Using the output of the PLL allows the receiver design to be tested with a deterministic 2.0 GHz clock signal. This will mimic a 4.0 Gbps data signal, but input signals are not always periodic and can be influenced by ISI. Testing the effects of ISI required the use of a pseudo random binary sequence (PBRS) generator. The design used a linear feedback shift register implemented in a 6 ⊗ 9 configuration to achieve a pseudo random sequence of 511 symbols at a maximum output rate of 2 Gbps [30]. Figure 5.28 shows the schematic of the PBRS generator.

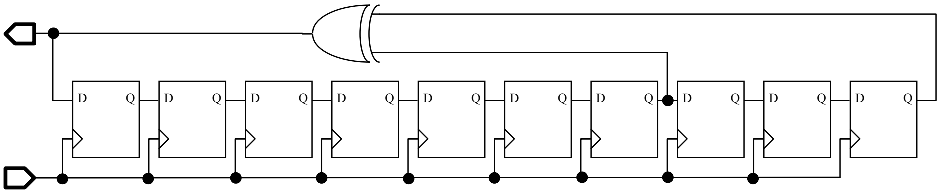


Figure 5.28 – Schematic of the PRBS Generator

The output of the PRBS generator was fed to multiple transmitters that coupled the signal across the capacitive interface to the low power receiver. This allows for an ISI analysis against various receiver designs and coupling capacitors. Simulation results show 10 ps of jitter passed to the PRBS from the PLL. Due to the input jitter and PRBS circuitry, the output of the PRBS showed a data eye closure of 50 ps (10 % of the data eye). This reduction in the data eye is passed to the input receiver. Figure 5.29 shows the simulated data eye taken at the output of the receiver.

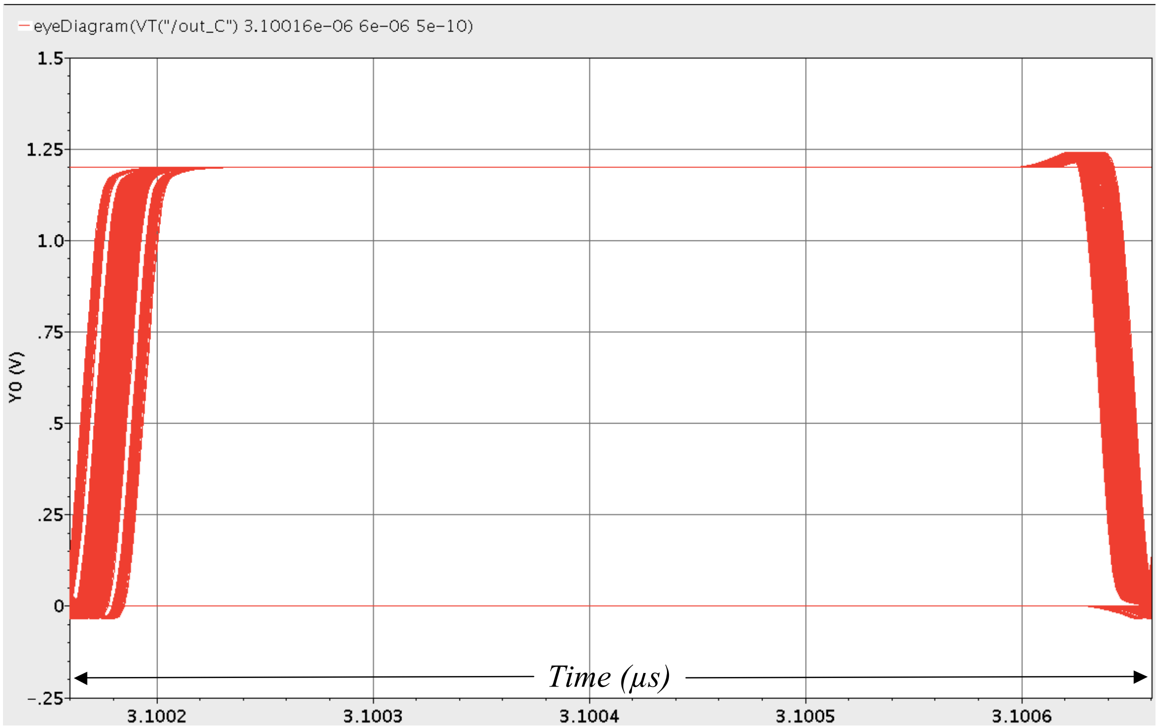


Figure 5.29 – Simulated Data Eye

The simulated data eye shows that the receiver does not contribute to the eye closure caused by the PLL and PRBS generator. The output eye closure is approximately 50 ps, the same as the input data eye closure. Adequate test structures were created to verify the simulation results.

Test Structures

Along with the data eye test circuits, several test structures were included in the 65 nm test chip. Table 5.2 lists the test structures.

Table 5.2 – List of Test Structures on the 65 nm Test Chip

Structure	Type	Coupling Capacitor (fF)
1	Data Eye	20, 40, 80, 100
2	Data Eye	1, 5, 10, 15, 20
3	Cap Sweep	20, 40, 80, 100
4	PLL Cap Sweep	20, 40, 80, 100
5	PLL Only	-

Test structure one contains a PLL, PRBS generator, multiple copies of the transmitter, and multiple copies of the receiver design. The receiver contained in test structure one was designed for a nominal 50 fF capacitor. Varying capacitor values were used with the receivers in test structure one to analyze the effects of changes in the value of the coupling capacitor. Test structure two is similar to test structure one, except the receiver design is optimized for a 15 fF coupling capacitor. Test structure three is similar to test structures one and two but does not contain a PLL and PRBS generator. Test structure three allows the receiver to be tested with a variety of digital input signals.

Test structure four includes the PLL and multiple coupling capacitor values. The PRBS generator is left out to determine the effects of transmitting a clock like signal across a capacitive interface, and to test the receivers ability to successfully receive 4 Gbps. Test structure five contains only the PLL for characterization.

All test structures were placed on a 2 mm by 2 mm chip fabricated in an IBM 65 nm low power process at MOSIS. The chip layout can be seen in Figure 5.30.

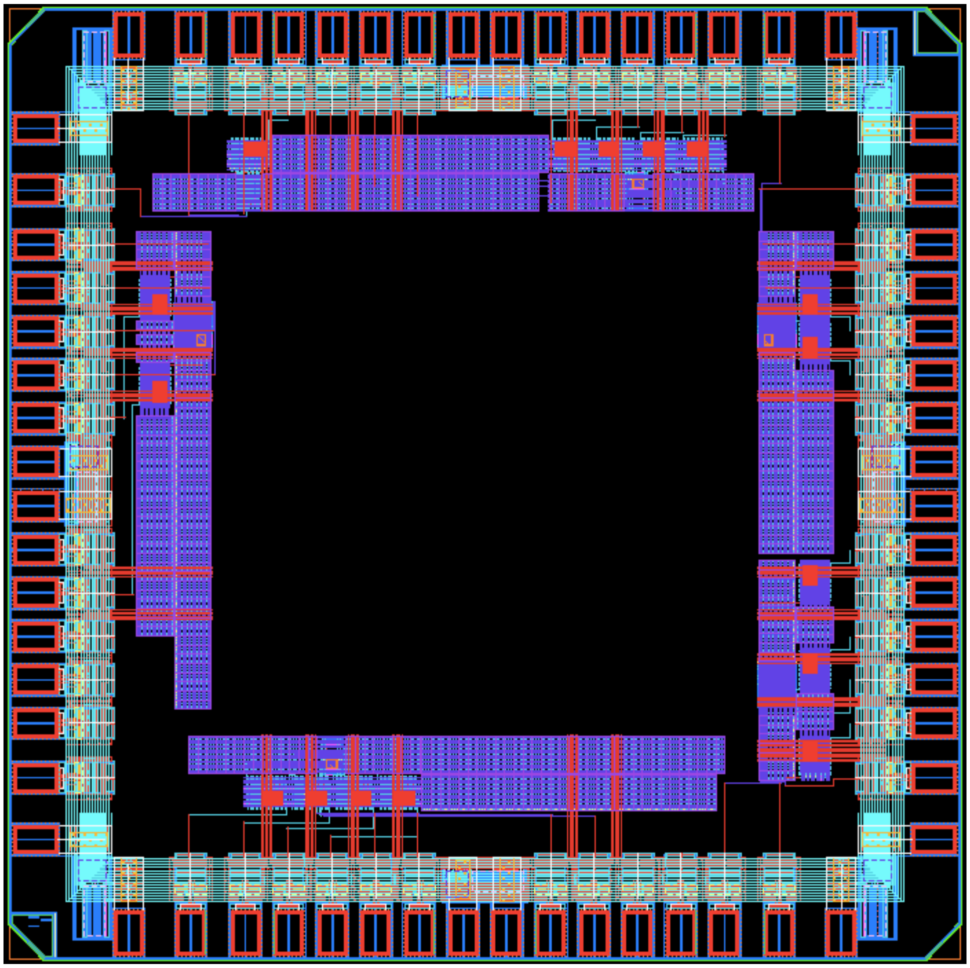


Figure 5.30 – Chip Layout of the 65 nm Test Chip

The high frequency output signals required the use of an alternative packaging solution, compared to the 0.5 μm test chip. For this reason, the test chip was wire bonded to a printed circuit board (PCB). The next section summarizes the design of the PCB used to test the 65 nm design.

Printed Circuit Board

The wire bonding capabilities at Boise State University resulted in only critical signals being bonded to the PCB. This was due to the required pitch of the wire bonds. The redundancy of test structures allowed for a full characterization of the design. Figure 5.31 shows the schematic of the PCB design.

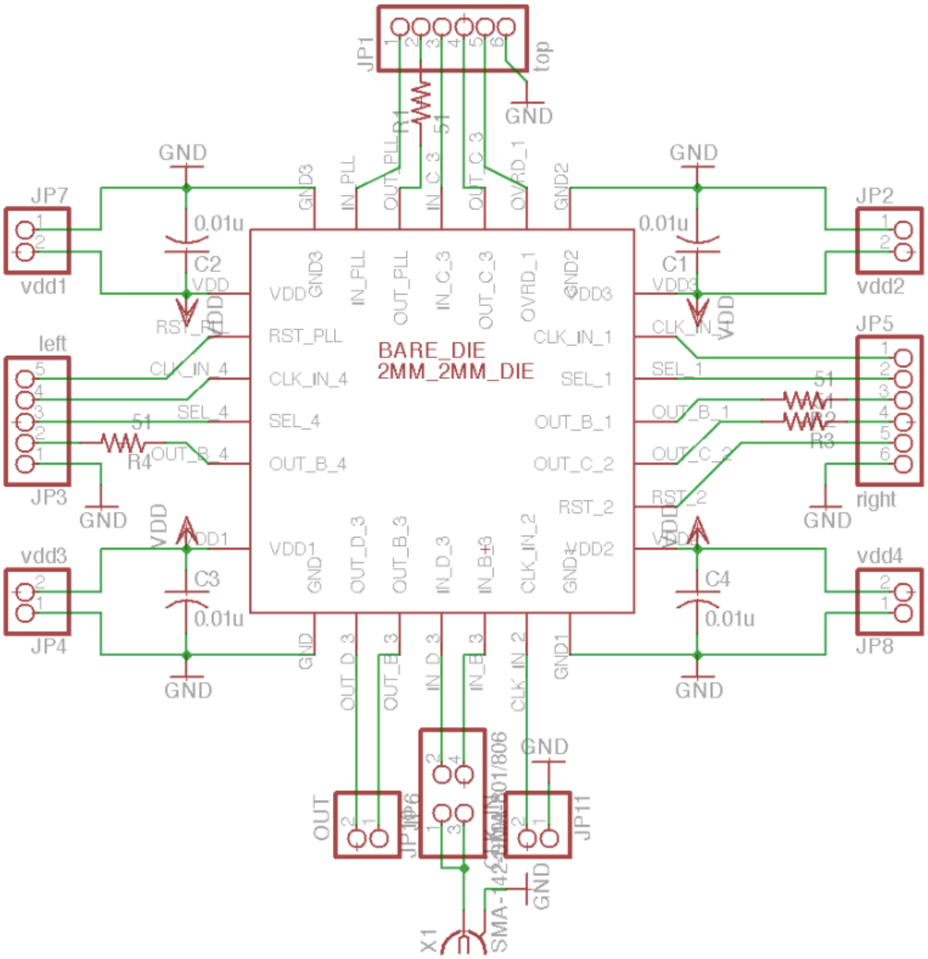


Figure 5.31 – PCB Schematic

The sub-miniature version A (SMA) coaxial connector is used to deliver a high frequency signal to test structure three because it does not use the PLL. All other connections use a standard pillar to connect to the test equipment. The PCB board was designed in a two-layer PCB board process and measured 2” by 2”. Figure 5.32 shows the layout of the PCB.

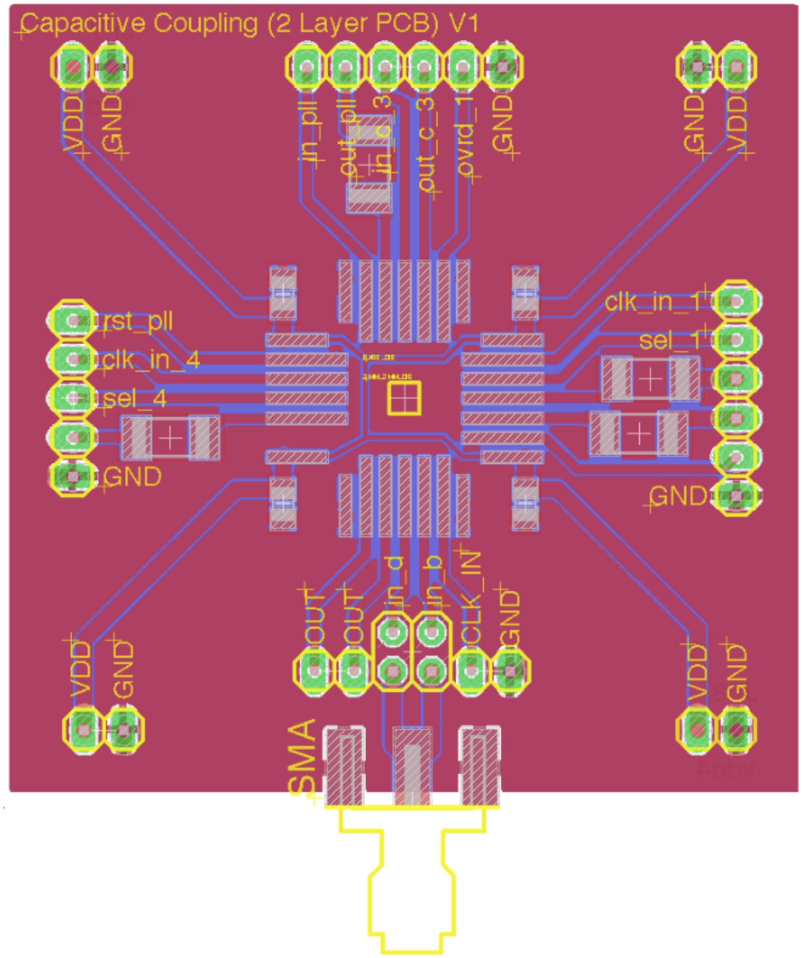


Figure 5.32 – PCB Layout

The PCB board was successfully manufactured and components were soldered to the board. Figure 5.33 shows the test board populated with the passive components. 50 Ω resistors were placed in series with high-speed output signals to reduce the active peaking caused by the combination of package inductance and capacitance. 10 nF capacitors were placed near the die as coupling capacitors to mitigate power supply noise issues.

Attempts were made to wire bond the test chip to the PCB board without success. The wire bonds were unable to adhere to the large PCB wire traces. It may be possible to eliminate this issue if the PCB wire bond sites are made smaller.

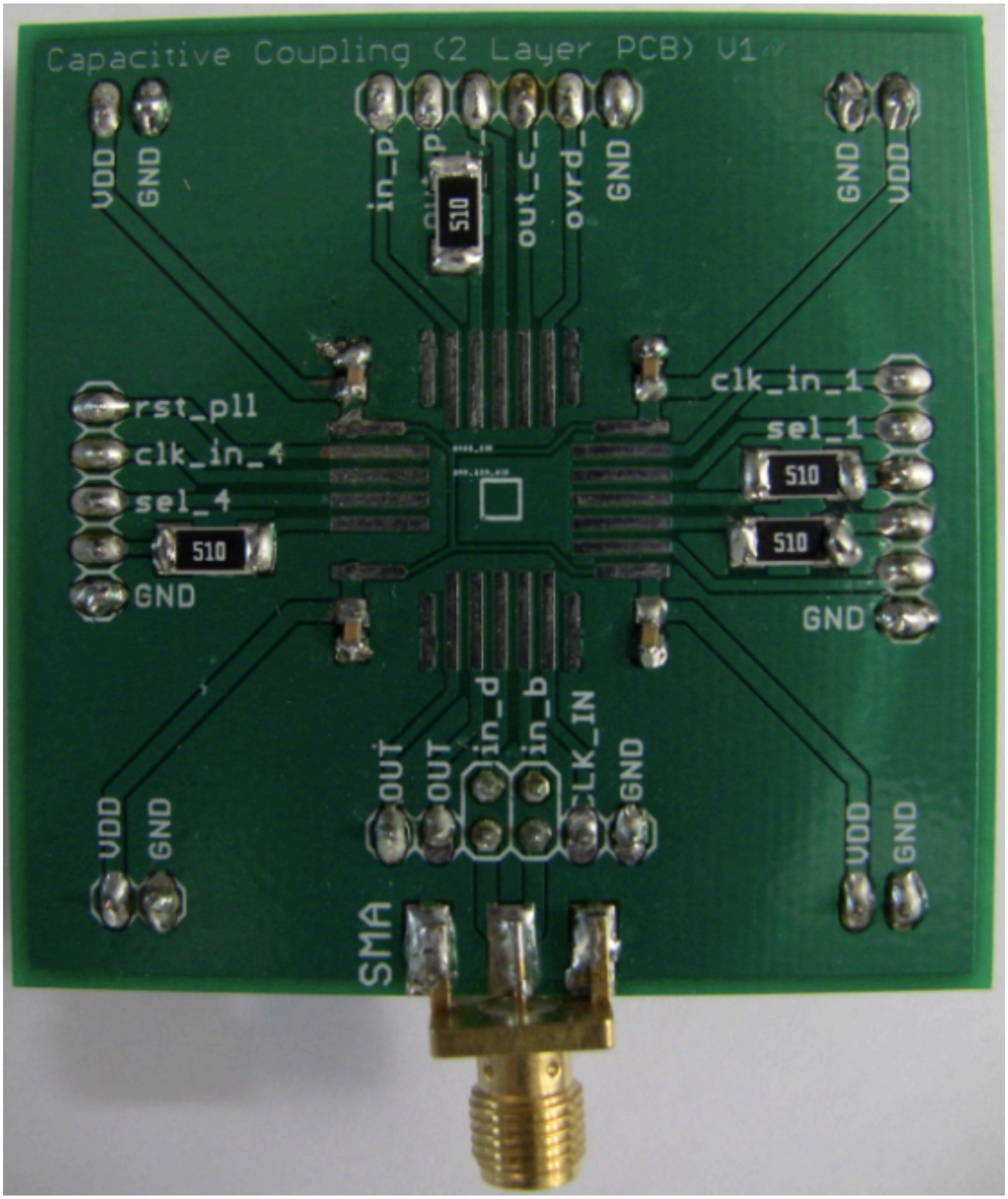


Figure 5.33 – Picture of the Populated Test Board

Experimental Results

The 65 nm test chip was placed in DIP24 packages for initial functionality checks. Figure 5.34 shows the test chip. The only discernable structures obtained by viewing Figure 5.34 is the power and ground pads. The supply pads contain the top level of metal perturbing into the die, making them visible. The rest of the chip is covered with top-level metal fill, which inhibits visibility of the silicon structures.

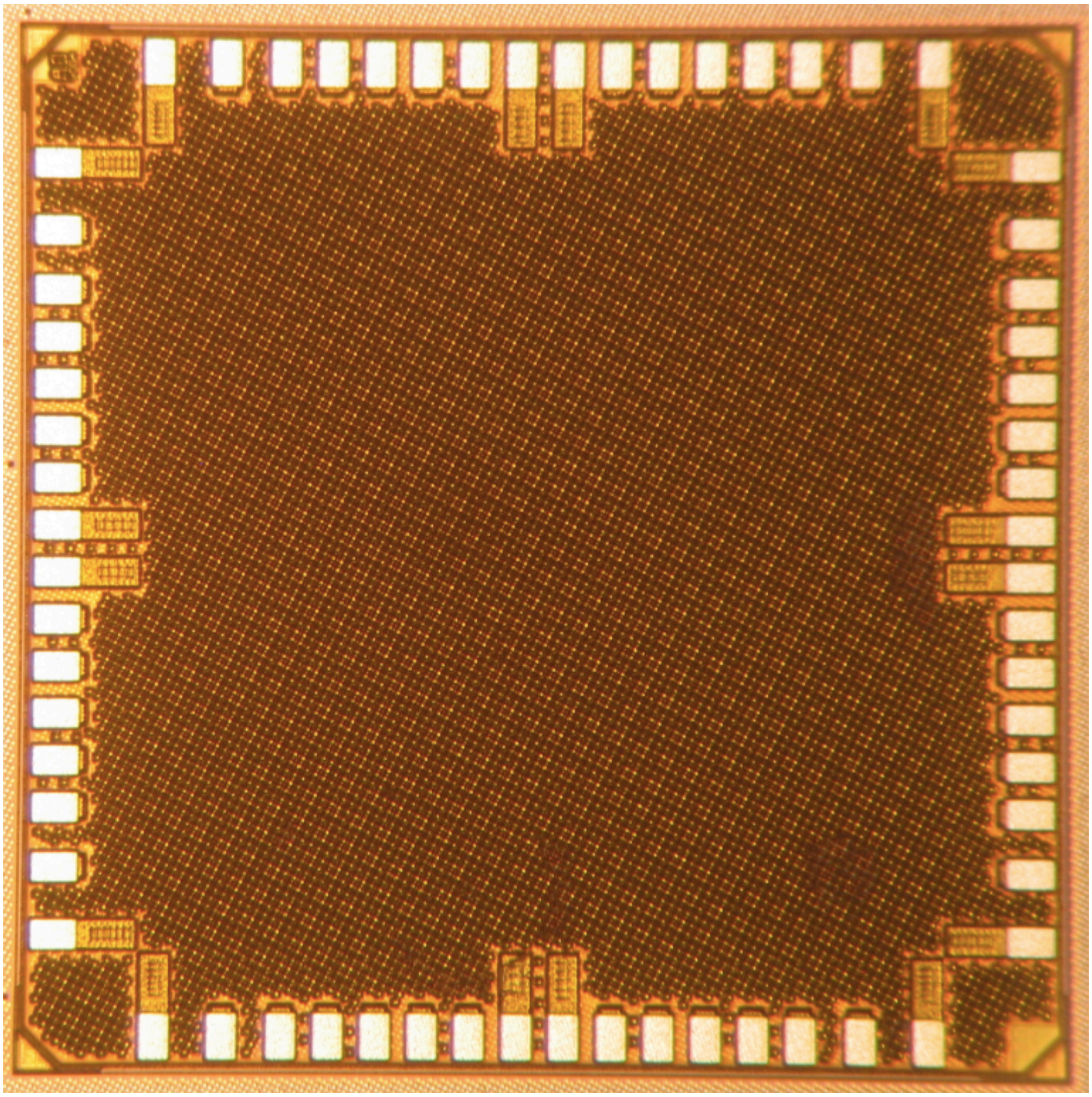


Figure 5.34 – Microphotograph of the 65 nm Test Chip

Figure 5.35 shows the scope output taken at 10 MHz. The top trace (yellow) is the input signal presented to the die, while the bottom trace (green) is the output of the test chip. These results confirm the successful transmission and reception of signals across a 50 fF coupling capacitor in 65 nm CMOS technology.

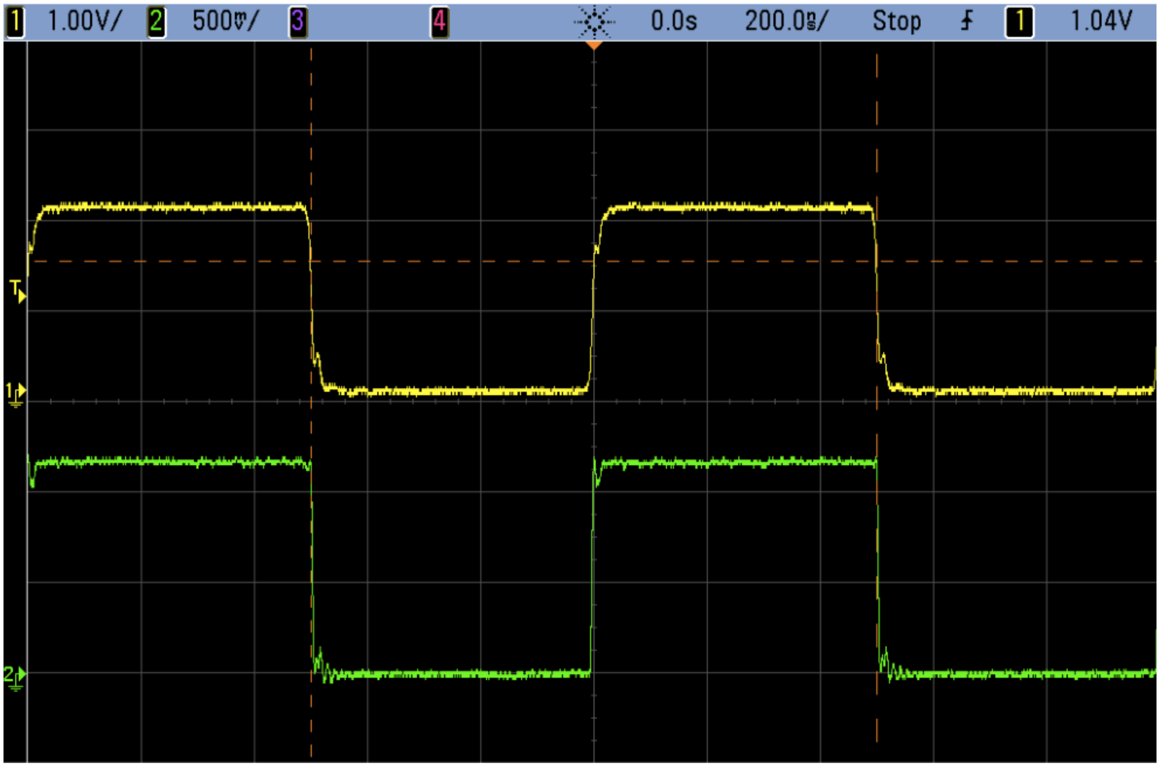


Figure 5.35 – Input and Output Signals Taken from the 65 nm Test Chip

High frequency results (> 200 Mbps) required the use of “dead bug” test setup. Figure 5.36 shows the setup used to test the die at high frequencies. The dead bug configuration is used to remove parasitic inductances found in typical source and probe cables.

The configuration uses a copper ground plane. The DIP24 package is placed on its backside (top of the package) and passive components are soldered to the necessary pins of the package. The top and right hand side of Figure 5.36 show BNA to SMA connectors used to drop coaxial connections down to SMA connectors.

The coaxial cables provide the input to the test board and provide transmission of the output signal to a high frequency scope. The input signal is fed to the package as a sine wave referenced to ground. A resistor divider provides a DC bias to the input pin, while a capacitor is used to block the DC component of the input signal. The output signal is fed to the SMA output connector through a 510 Ω resistor. The input of the scope is set to a 50 Ω termination. This configuration provides an approximate 11:1 ratio that is compensated by setting the gain of the scope to match this attenuation.

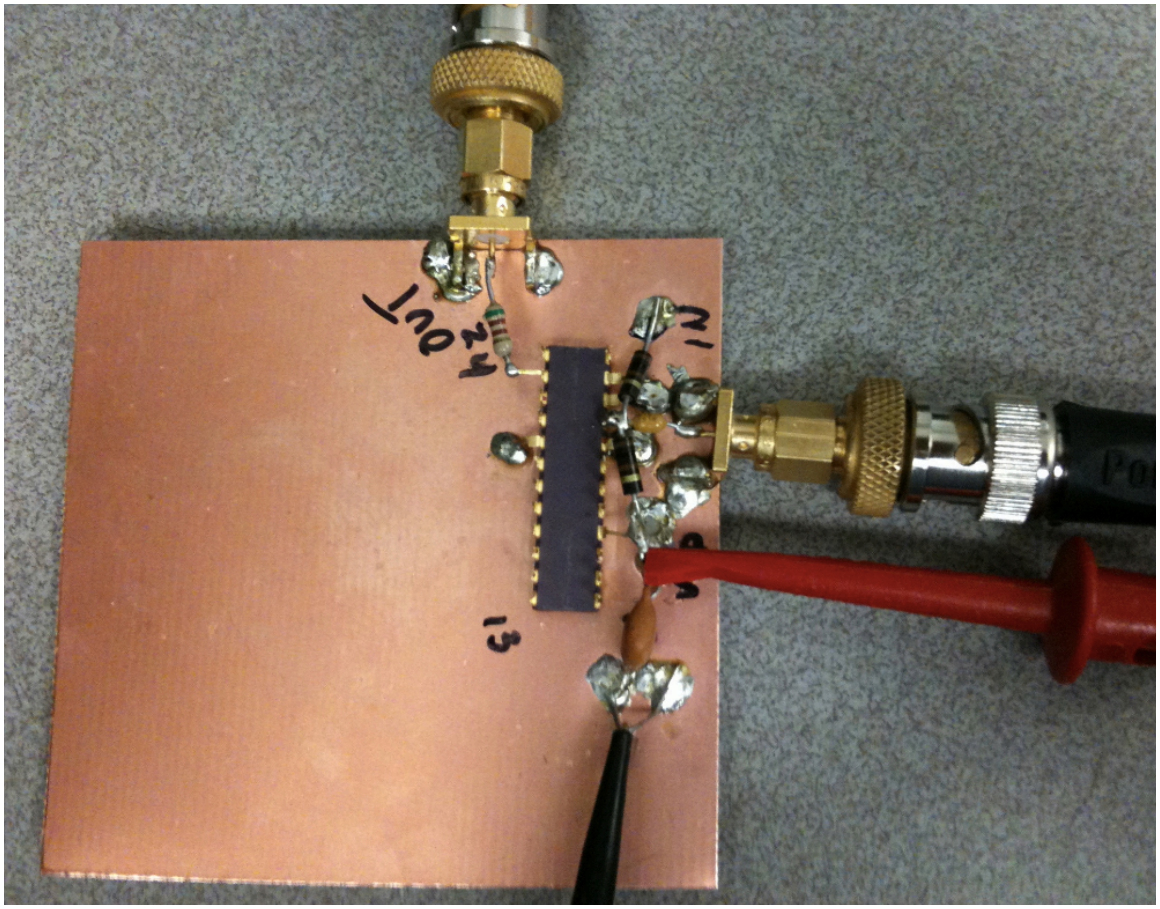


Figure 5.36 – Dead Bug Test Setup

Although the high frequency behavior of the DIP package and passive components degrade the input and output signals, this test setup provided high frequency verification of the 65 nm test chip. Figure 5.37 shows the output of the test chip when receiving a 2 Gbps input signal. The results were taken using a 1 GHz scope. The scope attenuates signals greater than 2 Gbps. This causes a square wave to appear as a sine wave, because only the first harmonic is successfully represented.

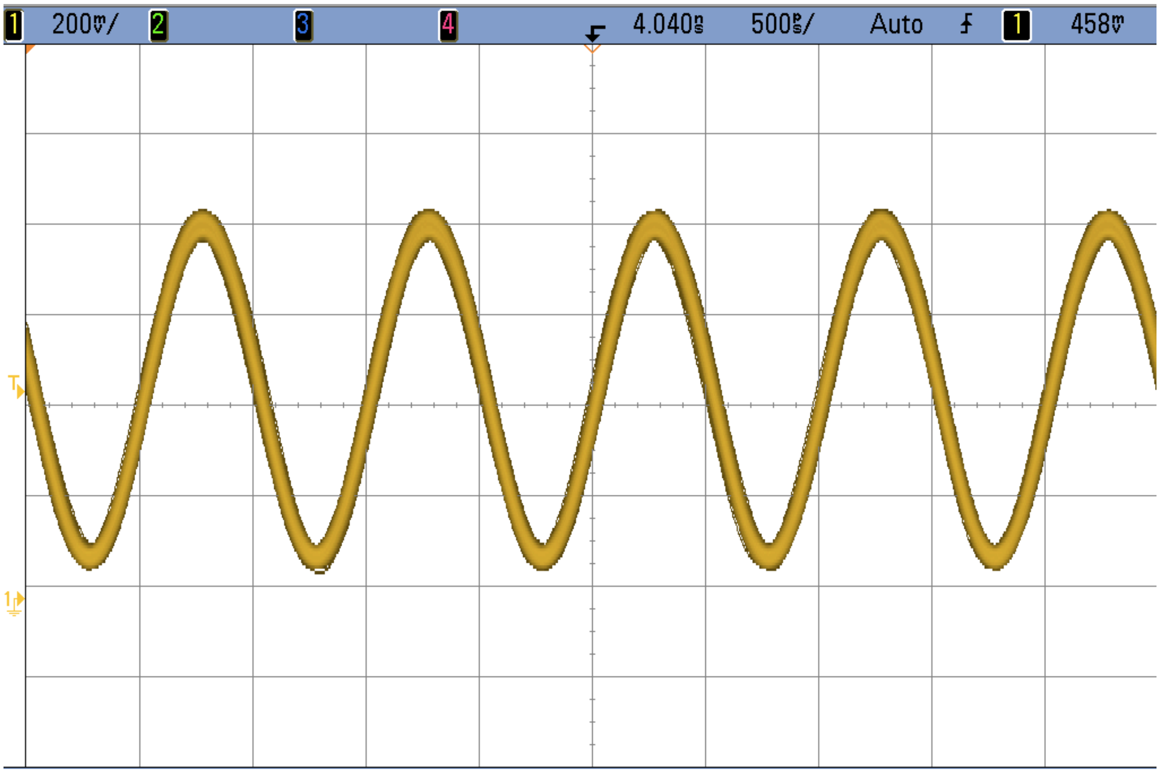


Figure 5.37 – 1 GHz Output of the 65 nm Test Chip

The results obtained in Figure 5.37 are from the 50 fF receiver design operating at 1.2 V. The receiver successfully received a 2 Gbps input signal at 0.9 V, providing the possibility of 25% reduction in energy consumption. The test setup does not attenuate signals that operate below 800 Mbps as severely as signals above 2 Gbps. Figure 5.38 shows the output of the receiver at 800 Mbps.

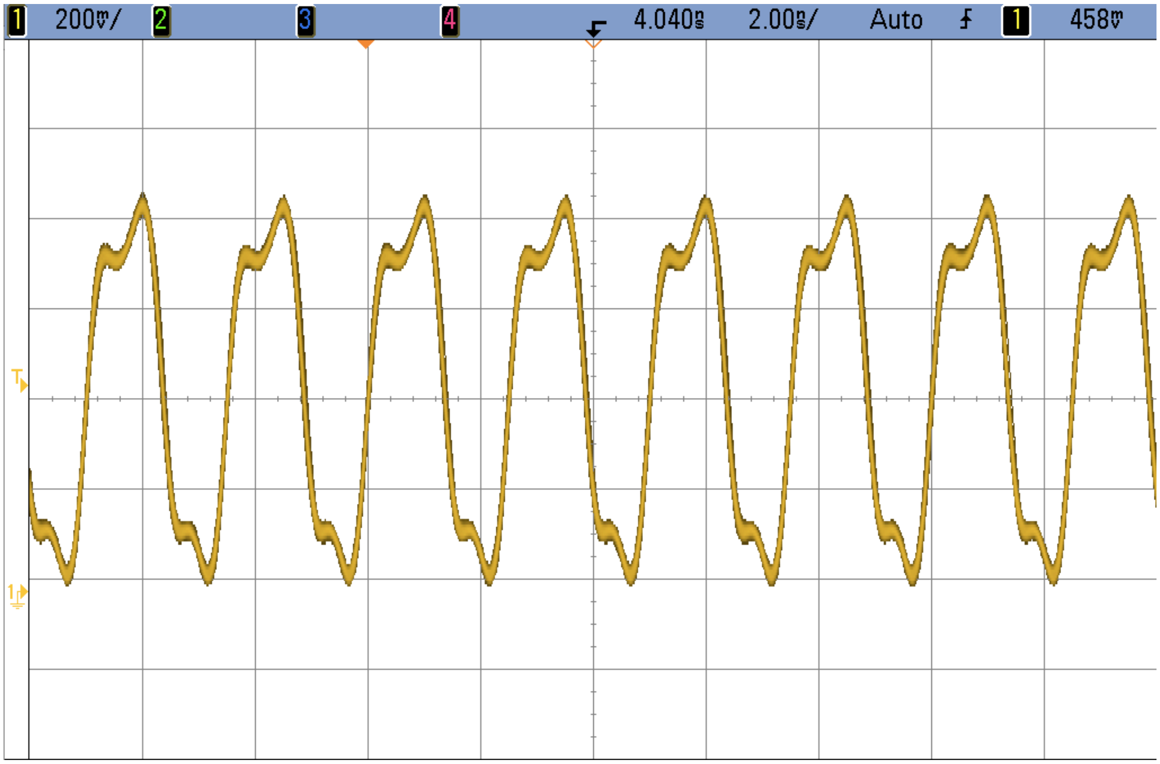


Figure 5.38 – 400 MHz Output Signal of the 65 nm Test Chip

Experimental results show that the receivers can operate reliably with a varying capacitor value. Data was successfully transmitted and received with a minimum coupling capacitance of 50 fF using a 1.2 V power supply for both the transmitter and receiver. The power consumed by the receiver optimally designed for a 50 fF coupling capacitor was less than 20 μW. The reliable bandwidth obtained, on silicon, with the proposed receiver design was 2.0 Gbps.

Simulation data supplemented the experimental results when high frequency measurements were not available and allowed for the development of an aggregate bandwidth metric. The receiver area measured 17.64 μm^2, allowing for an aggregate bandwidth of 226 Tbps/mm^2. The power density obtained when using the maximum aggregate bandwidth is 1.4 mW/mm^2. These values give a total power metric of 162

Tbps/mW/mm^2. These results are summarized and compared against other capacitive-coupled receiver designs.

Table 5.3 – Summary of Results

Work	Process	Supply (V)	Data Rate (Gbps)	Coupling (fF)	Bandwidth (Gbps/mm^2)	Energy (pJ/bit)	Requires CLK
Kanda 2003 [31]	0.35 μm	3.3	1.27	10	2.117×10^3	2.4	Yes
Franzon 2006 [27]	0.18 μm	1.8	3	150	5.55×10^2	5.0	No
Fazzi 2007 [32]	0.13 μm	1.2	1.23	10	1.922×10^4	0.14	Yes
Kim 2009 [33]	0.18 μm	1.8	2	600	6.90×10^2	0.8	Yes
This Work	0.5 μm	5.0	0.2	50	3.25×10^2	8	No
This Work	65 nm	1.2	4	15	2.268×10^5	0.015	No

The advantages of the receiver design, proposed in this dissertation, is demonstrated when viewing Table 5.3. The major figures of merit (bandwidth and energy) are orders of magnitude better than previous designs. The receiver designs presented by Kanda et al., Fazzi et al., and Kim et al. require the use of a clock signal [31 – 33]. The circuitry required to generate and transmit a clock signal increases the amount of power past what was reported. The clock circuitry adds additional power. Increasing the number of data signals to minimize the power contribution of the clock generation and transmission creates additional challenges and overhead to the receiver design.

As the number of transmitted signals increases, the total area of the transmission circuits increases. The clock signal is required to travel the entire area. Random noise, routing offsets, thermal variation, process variation, and power supply variation causes the clock signal to have varying arrival times at each transmission circuit. The varying arrival times are referred to as clock skew. As the clock skew increases, the use clock and data recovery circuitry is required to receive the transmitted data correctly. This additional circuitry often includes a phase locked or delay locked loop. The power, area, and design complexity required to support a receiver design that uses a clock signal reduces the effectiveness of a low power capacitive coupled interface. The receiver design used in this dissertation removes the use of a clock signal, thereby reducing the power, area, and complexity.

Summary

In this chapter, a historic perspective of capacitive-coupled interconnects was presented describing the shortcomings of prior art. A novel receiver design was produced that demonstrated lower power consumption than prior inventions. A test chip was produced in 0.5 µm to give proof of concept. Test results proved that a 200 Mbps signal could be transmitted across a 100 fF capacitor and properly received without error. The dynamic energy consumption of the receiver was 2.4 pJ/bit and the total energy (dynamic and static) consumed at 200 Mbps was 8 pJ/bit at a transmitter voltage of 5 V and 3 pJ/bit at a transmitter voltage of 1.3 V. The test chip was able to successfully transmit a clock signal with a transmission voltage of 1.3 V and receive the signal with a receiver powered with 5.0 V.

A 65 nm test chip was produced to study the high frequency operation and scalability of the new receiver design. The test chip contained a PLL and PRBS generator that provided data eye measurements. The 65 nm test chip successfully coupled across a 50 fF capacitor at 2 Gbps without error and operated correctly down to 0.9 V. The total energy consumed by the receiver was 15 fJ/bit, which is two orders of magnitude less power compared with conventional DRAM receivers. Typical DRAM receivers have an energy consumption in the pJ/bit range while operating at several Gbps. This allows the possibility of a two orders of magnitude increase in the I/O count of modern DRAM chips without impacting power, or a reduction of two orders of magnitude in I/O power consumption when the I/O count is not increased.

CHAPTER SIX – CONCLUSION AND FUTURE WORK

The computer industry is transitioning towards an increase in cloud computing. Mobile devices (laptops, tablets, and smart phones) are able to access the cloud and some devices have access to virtualization software. It is now possible to access high performance computers over the Internet with your mobile device. This shift in the industry will increase the number and performance of server and mobile devices.

This dissertation researched the impact of main memory on the power consumption of both mobile and server platforms. The correlation between memory capacity and bandwidth were directly related to performance and energy consumption. A new memory architecture was proposed that utilized novel techniques to increase the bandwidth and capacity of main memory while substantially reducing the power consumed by the memory.

An advanced packaging technology was proposed that leverages previous innovations to describe a mounting technique that uses slanted memory die and a substrate to develop a low cost alternative to current 3DIC configurations. The proposed memory module measures less than 2 cm^3 and provides the needed capacity increase for server and mobile platforms.

A 4 Gb DRAM wide I/O DRAM architecture was proposed that used up to 64 data pins to achieve high bandwidth. The DRAM architecture was designed to utilize the 64 data pins for high bandwidth, low power, or a combination of the two. This proposed

DRAM architecture consumes 50% less power and achieves a 100% increase in the bandwidth compared to traditional DRAM.

This innovation provides a substantial reduction in the energy per bit of modern DRAM devices. It was proposed that the page size of modern DRAM components be decreased due to the increase in independent CPU threads reducing the temporal and spatial locality of an open page. With a reduced page size, the memory array can reduce its power consumption while requiring less complexity at the memory controller due to scheduling. The wide I/O interface is made possible through innovations provided by capacitive-coupled interconnects.

A low power capacitive-coupled receiver was demonstrated in both 0.5 µm and 65 nm CMOS technologies. The 0.5 µm design provided proof of concept and allowed for an energy consumption of 8 pJ/bit at 200 Mbps. The test chip worked down to a transmitted voltage of 1.3 V. The 65 nm test chip provided the proof of scalability for the new receiver design. The 65 nm receivers consumed less than 15 fJ/bit at 4 Gbps, compared to traditional receiver design that consume energy in the pJ/bit range. This low power value allows for a 64 bit wide DRAM architecture to achieve substantial bandwidth increases while reducing the power consumed.

The nano-module proposed in this dissertation would benefit most by using the DRAM architecture and receiver design developed in this dissertation. The power and bandwidth of main memory, using these techniques, would help solve the power vs. performance issues seen in server and mobile platforms.

Continuing this line of research would require work in ESD. The ESD structures can be made smaller, and even removed, due to the removal of a physical connection.

Removing the physical connection does not guarantee the removal of all ESD events, and further research would need to determine the types of ESD events that would occur during packaging. These ESD events should be applied to a capacitive-coupled interface to determine their affects.

The nano-module can be built and tested to address packaging concerns such as temperature and mechanical stress affects on the 3DIC configuration. There are multitudes of ways to attach the die to the substrate that were not covered in this dissertation but can be found in the references.

BIBLIOGRAPHY

[1] Val, C.; Lemoine, T.; , "3-D interconnection for ultra-dense multichip modules," *Components, Hybrids, and Manufacturing Technology, IEEE Transactions on* , vol.13, no.4, pp.814-821, Dec 1990

[2] Bertin, C.L.; Perlman, D.J.; Shanken, S.N.; , "Evaluation of a three-dimensional memory cube system," *Components, Hybrids, and Manufacturing Technology, IEEE Transactions on* , vol.16, no.8, pp.1006-1011, Dec 1993

[3] Uksong Kang; Hoe-Ju Chung; Seongmoo Heo; Duk-Ha Park; Hoon Lee; Jin Ho Kim; Soon-Hong Ahn; Soo-Ho Cha; Jaesung Ahn; DukMin Kwon; Jae-Wook Lee; Han-Sung Joo; Woo-Seop Kim; Dong Hyeon Jang; Nam Seog Kim; Jung-Hwan Choi; Tae-Gyeong Chung; Jei-Hwan Yoo; Joo Sun Choi; Changhyun Kim; Young-Hyun Jun; , "8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology," *Solid-State Circuits, IEEE Journal of* , vol.45, no.1, pp.111-119, Jan. 2010

[4] Matthias, T.; Kim, B.; Burgstaller, D.; Wimplinger, M.; Lindner, P., "State-of-the-art Thin Wafer Processing," Chip Scale Review, vol. 14, no. 4, pp. 26, July 2010.

[5] U.S. Enviornmental Protection Agency, "Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431," 2007.

[6] L. Minask, B. Ellison, "The Problem of Power Consumption in Servers," Intel Press, 2009, http://www.intel.com/intelpress/articles/rpcs1.htm

[7] D. Patterson, J. Hennessy, Computer Organization and Design, 4th ed., Morgan Kaufmann Publishers, San Francisco, 2009.

[8] Karp, J.; Regitz, W.; Chou, S.; , "A 4096-bit dynamic MOS RAM," *Solid-State Circuits Conference. Digest of Technical Papers. 1972 IEEE International* , vol.XV, no., pp. 10- 11, Feb 1972

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

[10] B. Gervasi, " Time to Rethink DDR4," MEMCON 2010, http://discobolusdesigns.com/personal/20100721a_gervasi_rethinking_ddr4.pdf

[11] Various IBM server datasheets. www.ibm.com

[12] "Power-Efficiency with 2, 4, 6, and 8 Gigabytes of Memory for Intel and AMD Servers," Neal Nelson & Associates, White Paper 2007.

[13] Rambus, "Challenges and Solutions for Future Main Memory," http://www.rambus.com/assests/documents/products/future_main_memory_white paper.pdf, May 2009.

[14] Intel AMB Datasheet, http://www.intel.com/assets/pdf/datasheet/313072.pdf, pg 38.

[15] "*Intel Server Board S5520UR and SS5520URT, Technical Product Specification*" Rev. 1.6, July 2010, Intel Corporation.

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

[17] Black, B., Annavaram, M., Brekelbaum, N., DeVale, J., Jiang, L., Loh, G., McCauley, D., Morrow, P., Nelson, D., Pantuso, D., Reed, P., Rupley, J., Shankar, S., Shen, J., Webb, C., "Die Stacking (3D) Microarchitecture," Symposium on Microarchitecture, 39th Annual IEEE/ACM International, 2006

[18] Cotues, "Stepped Electronic Device Package," U.S. Patent 5,239,447, Aug. 24, 1993.

[19] G. Rinne, P. Deane, "Microelectronic Packaging Using Arched Solder Columns," U.S. Patent 5,963,793, Oct. 5, 1999.

[20] R. Plieninger, "Challenges and New Solutions for High Integration IC Packaging," ESTC, July 2006, http://141.30.122.65/Keynotes/6-Plieninger-ESTC_Keynote_20060907.pdf

[21] Harvard, Q., "Wide I/O Dram Architecture Utilizing Proximity Communication" (2009). *Boise State University Theses and Dissertations.* Paper 72.

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

[23] K. Kilbuck, “Main Memory Technology Direction,” Microsoft WinHEC 2007, May 2007.

[24] 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.

[25] Salzman, D.; Knight, T., Jr., "Capacitive coupling solves the known good die problem," *Multi-Chip Module Conference, 1994. MCMC-94, Proceedings., 1994 IEEE* , vol., no., pp.95-100, 15-17 Mar 1994

[26] Salzman, D.; Knight, T., Jr.; Franzon, P., "Application of capacitive coupling to switch fabrics," Multi-Chip Module Conference, 1995. MCMC-95, Proceedings., 1995 IEEE , vol., no., pp.195-199, 31 Jan-2 Feb 1995

[27] Wilson, J.; Mick, S.; Jian Xu; Lei Luo; Bonafede, S.; Huffman, A.; LaBennett, R.; Franzon, P.D.; , "Fully Integrated AC Coupled Interconnect Using Buried Bumps," Advanced Packaging, IEEE Transactions on , vol.30, no.2, pp.191-199, May 2007

[28] Luo, L.; Wilson, J.M.; Mick, S.E.; Jian Xu; Liang Zhang; Franzon, P.D.; , "3 gb/s AC coupled chip-to-chip communication using a low swing pulse receiver," Solid-State Circuits, IEEE Journal of , vol.41, no.1, pp. 287- 296, Jan. 2006

[29] R. Baker, CMOS: Circuit Design, Layout, and Simulation, Third Edition, Wiley-IEEE, 2010

[30] O. Schwartsglass, “PRBS Work,” The Hebrew University of Jerusalem, VLSI class notes, 2002. http://www.cs.huji.ac.il/course/2002/vlsilab/files/prbs/PRBS.pdf

[31] Kanda, K. Antono, D.D., Ishida, K., Kawaguchi, H., Kuroda, T., Sakurai, T.; “1.27 Gb/s/pin 3 mW/pin Wireless Superconnect (WSC) Interface Scheme,” IEEE Solid-State Circuits Conference, Session 10, Paper 10.7, Feburuary 11th, 2003.

[32] Fazzi, A. Canegallo, R., Ciccarelli, L., Magani, L., Natali, F., Jung, E., Rolandi, P., Guerrieri, R., “3-D Capacitive Interconnections With Mono- and Bi-Directional Capabilities,” Solid-State Circuits, IEEE Journal of, vol. 43, no. 1, pp. 275-284, Jan. 2008

[33] Kim, G., Takamiya, M., Sakurai, T., “A 25-mV-Sensitivity 2-Gb/s Optimum-Logic-Threshold Capacitive-Coupling Receiver fro Wireless Wafer Probing

Systems," Circuits and Systmes, IEEE Transactions on, vol. 56, no. 9, pp. 710-713, Sept. 2009