Boise State University
ScholarWorks

Electrical and Computer Engineering Faculty Publications and Presentations

Department of Electrical and Computer Engineering

4-16-2010

Main Memory with Proximity Communication: A Wide I/O DRAM Architecture

Qawi Harvard
Boise State University

R. Jacob Baker
Boise State University

Robert Drost
VLSI Research Group

©2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. DOI: 10.1109/WMED.2010.5453754

Main Memory with Proximity Communication

A Wide I/O DRAM Architecture

Qawi Harvard and R. Jacob Baker
Department of Electrical and Computer Engineering
Boise State University
Boise, ID, U.S.A.

Robert Drost
VLSI Research Group
Sun Microsystems Laboratory
Menlo Park, CA, U.S.A.

***Abstract* — The bandwidth and power consumption of dynamic random access memory (DRAM), used as the main memory of a computer system, impacts computer execution rates. DRAM manufacturers focus on density increases, due to the innate price per bit decline of main memory, while processor manufacturers continually focus on boosting performance. This leads to a performance gap between the two technologies. Proximity communication promises to increase the off/on chip bandwidth of DRAM products while reducing the power consumption of the main memory system. The design of a memory system employing 4 Gb DRAM chips with a 64-bit wide communication bus using proximity communication is proposed. Technological roadblocks are analyzed and novel solutions are proposed. The proposed 4 Gb DRAM architecture can reduce the power consumption of a main memory system by 50% while increasing the bandwidth by 100%. The 4 Gb chip architecture measures 68.88 mm^2 and has an array efficiency of 59.9%. The estimates are comparable to 2012 International Technology Roadmap for Semiconductors' (ITRS) estimates of 74 mm^2 and 56%, respectively.**

Keywords – DRAM, proximity communication, chip-to-chip, server memory, main memory, bandwidth, power consumption.

I. INTRODUCTION

The performance gap between the computer's processor and its main memory has been growing over the past two decades [1]. Density and die size are the figures of merit for main memory manufacturers. Increasing these performance measurements places a physical limit on the latency of the main memory array due to the parasitics [2]. The limitations keep memory latency scaling at roughly 7%, while processor performance has been scaling at roughly 50%. This performance differential is termed the "memory gap", and refers to the growing performance disparity between the processor core and its main memory.

Processor manufacturers have made several architecture changes that enable computer performance to scale with Moore's Law (double the performance every two years). Multiple cores, increased cache levels, multiple threads, and speculative accessing, have made memory stalls almost transparent to the computer user [3]. Main memory manufacturers increase their density per unit area by developing longer bitlines, longer wordlines, decreased unit cell size, and feature size scaling [4]. Main memory manufacturers alleviate bandwidth limitations by using DRAM pre-fetch. Unfortunately, the pre-fetch architectures did not begin taking hold until 2000 [5]. This places memory bandwidth scaling decades behind processor bandwidth scaling.

Proximity communication is an input/output (I/O) technology that uses capacitors to electrically connect two chips [6]. The off/on chip communication technique has the ability to substantially increase the memory bandwidth and not impact the power consumption [7]. This work develops a memory architecture that utilizes proximity communication to substantially increase bandwidth, while reducing power consumption. This is achieved by allowing a single DRAM chip to provide a full cache line of memory (64 Bytes).

II. PROXIMITY COMMUNICATION

Capacitive coupled proximity communication is a chip-to-chip interface technology that uses the top level of metal on an integrated circuit to form the parallel plates of a capacitor. Two chips are placed face to face and their top level of metal is allowed to come within close proximity (1 µm – 20 µm) of each other without touching. This arrangement creates a parallel plate capacitor.

A. Advantages

The advantages of proximity communication allow for a significant reduction of parasitics in the transmission channel, which increases bandwidth and lowers power relative to other chip-to-chip interconnects. Fig. 1 depicts a cross sectional view of two chips using proximity communication as the I/O interface.

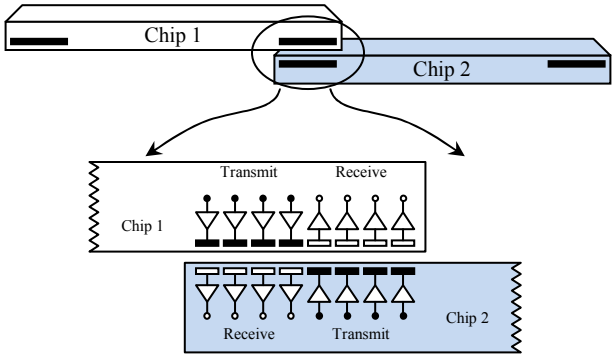

Figure 1. Cross section view of placing two chips face-to-face and within close proximity of each other [6].

978-1-4244-6575-0/10/$26.00 ©2010 IEEE

The removal of off chip wires allows for a fixed impedance to be delivered to the transmission channel. The passivation over the metal pads is not opened, as in wire-bonded applications, which allows the electrostatic discharge (ESD) protection circuitry to be removed, and for this reason the on-die resistive termination is superfluous [8].

The increase in I/O density is another advantage of proximity communication. The capacitance of parallel plate capacitors is at least 10 pF/mm^2 (with a 1 µm gap). 400 I/O channels per mm^2 is possible when each communication channel uses 25 fF. The configuration creates a research avenue for scaling the transmission channel below 25 fF.

Placing multiple die into a single package requires complicated wire-bonding technologies used for chip-to-chip interconnects [9]. Proximity communication allows chips to be simply glued in place, which increases the ease of testability. System in package (SiP) configurations can be tested, and defective chips easily replaced while using proximity communication.

B. *Challenges*

Chip misalignment is a major challenge associated with the development of proximity communication. Researchers at Sun Microsystems were able to develop a novel solution to this problem [10]. Through the development of electronic sensors, which could be incorporated into the same silicon substrate as transmit and receive circuits, it was possible to determine the misalignment of the two chips. Electrical steering circuits were developed that allowed the transmit data to be driven to multiple receiver pads to electrically realign the transmission channel.

III. DRAM TRENDS

Incorporated proximity communication into a DRAM architecture without understanding the DRAM market will result in a product that does not meet the need of current memory and computer systems.

A. *Effect of Price Decline and Scaling*

The performance differential between microprocessors and the main memory system is referred to as the memory gap and is often misunderstood. The memory gap measures the microprocessor's instructions per second and the main memory's access latency. The confusion occurs when you blindly relate these two figures of merit. DRAM manufacturers focus the majority of their innovations on the process technology that allows for an increase in density. The reason for this is due to the innate price per bit decline of DRAM.

DRAM manufacturers are forced to focus on density scaling over access latency, or I/O bandwidth, due to the historic 36% price per bit decline. Putting this into perspective, if two gigabits of memory chip costs $2.00 today, then four gigabits would cost $1.64 in two years. Density scaling in main memory chips follows Moore's Law. The doubling of the number of transistors in main memory chips every two years (or √2 every year) is used to increase the density of the die.

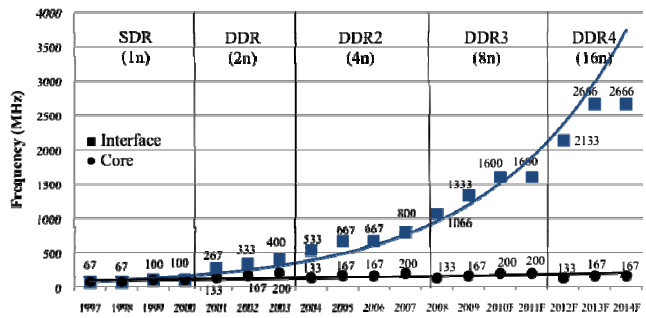

Figure 2. Array pre-fetch, of two and higher, has allowed off chip bandwidth to increase at a rate of 26% per year since 2000, while the core frequency does not scale [5].

Microprocessors use the extra transistors to increase the number of instructions that can be completed each second. Main memory manufacturers have the ability to arbitrarily set the latency and bandwidth of the memory chip. These chips sacrifice power and die size, which creates their inability to compete in the main memory market, which requires large densities [11]. Instead, these products find their place in varying applications that do not require large density.

Reducing the minimum feature size of the components in the memory array achieves the required density scaling. The reduction in feature size increases the parasitics associated with the array and places a physical limit to the bandwidth. The parasitics have placed a limit on the column bandwidth of current memory chips to 133 MHz – 200 MHz. Each generation of main memory starts with a column access of 133 MHz, and then transitions to 167 MHz, and then to 200 MHz to achieve a generational approach to chip bandwidth. Fig. 2 shows this bandwidth trend in main memory chips.

Array pre-fetch allows the DRAM to sustain a larger off-chip bandwidth. Array pre-fetch refers to accessing all bits of the latency at once, and serializing the data in the data path. Main memory chips have been operating with four, eight, or sixteen data pins over the past three generations (DDR, DDR2, DDR3), with eight data pins being most common. The maximum bandwidth that can be achieved with DDR3 chips is 12.8 Gbps (8 data pins, pre-fetch of 8, at 200 MHz). Increasing chip speeds above 12.8 Gbps requires an increase in pre-fetch (to 16) or an increase in data pins, due to the column access limit.

B. *Memory Channel Bandwidth Limitation*

Current computer systems use a 64-bit wide data bus to communicate between the main memory and the microprocessor. Series stub terminated logic connections are used in computer system memory channels due to the ease of memory upgrades. The series stub terminated logic refers to terminating electrical signals at each memory module with a resistive pull up device that prevents transmission line reflections from interfering with transmitted data on the shared memory channel.

The resistive termination network, along with module loading, places a bandwidth limit on the memory channel. Server applications require a substantially larger density (or number of main memory modules) then personal computers.

Registered, fully buffered, and load reduced DIMMs were developed for server applications to increase the number of DIMMs per memory channel. These innovations have a cost and power premium associated with them.

IV. X64 DRAM ARCHITECTURE

A 4 Gb DRAM architecture utilizing proximity communication was developed that is realizable with existing technology and meets 2012 ITRS predictions [12]. Challenges associated with incorporating proximity communication into DRAM were characterized and several innovations were developed that alleviated these challenges. A novel global I/O routing structure was discussed that promises to increase the number of data signals that can be read and written to a memory array. The slice architecture was developed to increase the modularity of memory systems.

A. Moving the Pads

Moving the communication channel to the edge of the DRAM chip creates several interesting challenges when performing an architectural feasibility study. The bank structure used in this research alleviates the initial challenges. Once the communication channel is moved to the edge of the die additional circuitry is required to buffer the signals into the memory chip. Limiting the number of rows per bank creates a "short" bank that reduces global data and command signals, eliminating the need for additional buffers.

The inexpensive process technology of DRAM chips utilizes 2 – 3 layers of metal above the memory capacitor. This places an intrinsic limit to the number of global I/O tracks over each bank. Due to this, the half-bank structure used in this proposal has 64k columns and 8k rows. This half-bank structure must decode the 64k columns into eight 8k pages. A by 64 DRAM chip operating with a pre-fetch of eight requires 512 bits to be accessed at once. Accessing 512 bits from one bank requires the use of a half-bank to reduce the total metal usage. Each half-bank supplies 256 bits of data. This allows the global I/O track to be spread across the chip, limiting metal usage for the global I/O bus. The challenges of buffering the signals into the array and limited routing channels are circumvented by using the proposed bank and segmented page structures. Fig. 3 shows the block diagram of the 4 Gb DRAM die. The half-bank structure can be thought of as dividing each bank horizontally, and firing a wordline in each half-bank.

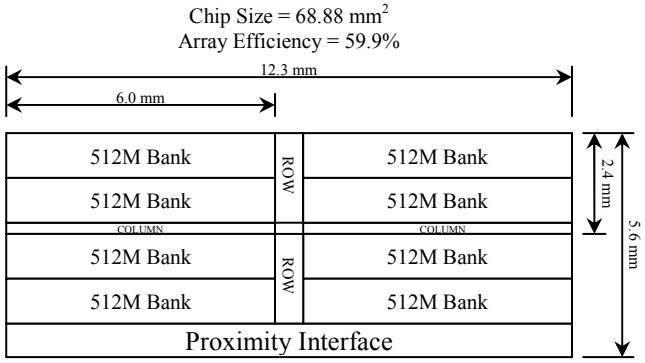

Figure 3. A 4 Gb DRAM architecture incorporating proximity communication and centralized row and column circuitry.

B. Local I/O Routing

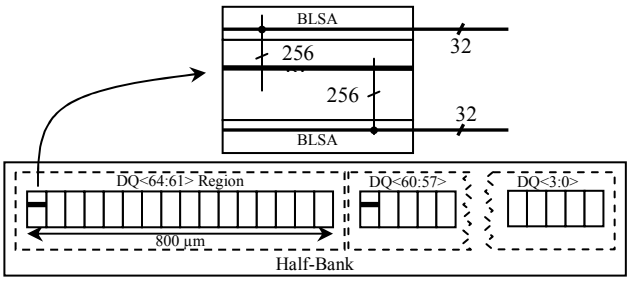

Figure 4. Space and data mapping of the local input/output routing within a half-bank.

The by 16 and by 32 proximity configurations will not require any significant innovation, but the by 64 configuration will require additional innovation for local I/O routing. The large number of global I/O tracks (256 per half-bank) requires 32 data signals from each 256 kb memory array. Moving 32 data signals from the bitline sense amplifiers to the global I/O track is a major challenge due to the limited routing space above the bitline sense amplifiers. Increasing the page size will alleviate this challenge but will also increase the power consumption. Instead, these signals can be routed to the top and bottom of each 256 kb memory segment, as seen in Fig. 4. An additional avenue for architectural research consists of routing the data signals through adjacent inactive bitlines (above and below).

C. New Global I/O Routing

As mentioned above the memory array operates at a maximum frequency of 200 MHz due to the parasitics of the memory array. The global I/O route does not share the parasitics of the memory array and can operate at a higher frequency. Insertion muxes, and additional latches can be used to keep the global I/O bus fully occupied with data. A column path protocol can be developed that allows for multiple banks to be accessed and data stored in the local I/O channels. Busy, ready, and data insertion requests can be used to allow the global I/O routing to operate at a higher frequency, while the memory array remains operating at frequencies below 200 MHz.

D. Modular Architecture

Main memory DRAM chips use a large number of repeated structures and symmetry. The proposed modular architecture speeds up design verification. Each modular architecture contains all circuitry required for one data pin to read and write. Combining many of these modular structures together will create the entire chip. A data, command, and clock modular architecture was developed during this research.

The first advantage of this architecture is that the time required for chip verification can be reduced significantly. Due to the sheer number of transistors on a modern DRAM chip, simulating an extracted netlist can take several weeks to complete. Using smaller modular blocks to fully verify the data, command, and clock paths within the chip will reduce the time required to perform validation on the extracted netlist

because each block is self contained. The second advantage of this modular structure is that varying densities of memory chips can be easily constructed for varying applications. DRAM chips utilizing both proximity communication and this modular structure can simply be glued directly over their application with the correct density and I/O count. This has the possibility of revolutionizing the way chips access off-chip data. Instead of driving data requests away from the central circuitry of an integrated circuit to the memory channel, it is possible to simply send signals up towards the memory chip. This approach provides the exact memory requirement at the exact place it is required, reducing the access latency considerably.

V. SUMMARY

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

A distributed page and bank structure was developed to enable the possibility of using proximity communication with 32 data pins. The architecture utilized the standard main memory page size specification of 8k, which allows the array power consumption to remain competitive with current and future DRAM architectures.

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

The wide I/O DRAM architecture utilizing proximity communication enables several technological advantages over existing DRAM architectures. Fixing the page size and increasing the I/O count through the wide I/O DRAM architecture allows for an energy efficient DRAM architecture. Fig. 5 shows the relative energy per bit estimates for DRAM chips utilizing proximity communication.

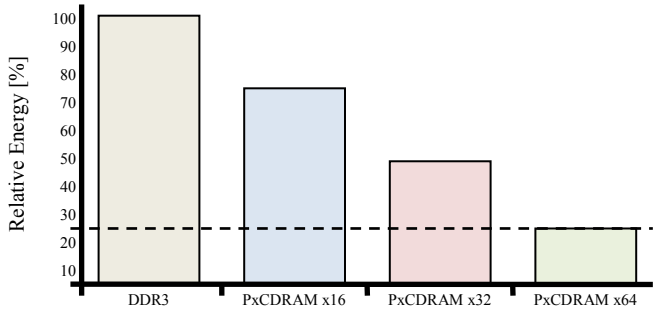

Figure 5. Energy per bit comparison.

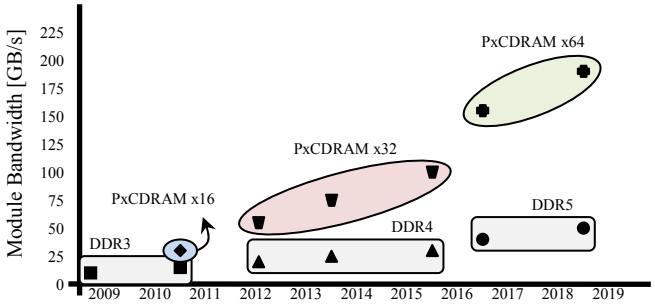

Figure 6. Module bandwidth comparison of current and future main memory compared to a main memory chips using proximity communication.

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.

Although it is possible to only access one proximity communication DRAM chip to supply the full 64 bytes of data to the memory controller, it is also possible to increase the amount of data accessed by increasing the memory channel width. The projected bandwidth trend shown in Fig. 6 clearly shows the advantage of using proximity communication DRAM over current and future DRAM technologies.

REFERENCES

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

[2] D. Rhosen, "The evolution of DDR," VIA Technology Forum, 2005.

[3] D. Patterson, "Latency lags bandwidth," Communications of the ACM, vol. 47, Issue 10, pp. 71-75, October 2004.

[4] D. Klein, "The future of memory and sorage: closing the gap," Microsoft WinHEC 2007, May 2007.

[5] Rambus, "Challenges and solutions for future main memory," http://www.rambus.com/assets/documents/products/future_main_memory_whitepaper.pdf, May 2009.

[6] 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.

[7] Q. Harvard, "Wide I/O DRAM architecture utilizing proximity communiation," Master's thesis, Boise State University, December 2009.

[8] D. Salzman, T. Knight, "Capacitively coupled multichip modules," Multichip Module Conference Proceedings, pp. 487-494, April 1994.

[9] K. Kilbuck, "Main memory technology direction," Microsoft WinHEC 2007, May 2007.

[10] R. Drost, R. Ho, R. Hopkins, I. Sutherland, "Electronic alignment for proximity communication," IEEE International Solid State Circuits Conference, vol. 1, pp. 144-145, February 2004.

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

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