A Scalable I/O Architecture for Wide I/O DRAM

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

***Abstract*—A 4 Gb DRAM architecture utilizing a scalable number of data pins is proposed. The architecture does not impact chip size and does not require additional metal layers. The 4 Gb DRAM measure 68.88 mm^2 and achieves an array efficiency of 59.9%. This was accomplished by using a split bank, edge I/O interface, central row, and central column structures. The architecture coincides with the chip size and array efficiency measurements predicted by the ITRS for a 40 nm 2012 production DRAM architecture.**

I. Introduction

Wide I/O DRAM products refer to the use of a large number of data signals. Recent publications report up to 512 I/O signals [1]. Wide I/O DRAM are used in mobile devices that require low power, high bandwidth, and high capacity [2]. These devices utilize stacked die with through silicon vias (TSV). The impact of using a wide I/O data interface carries with it an increase in die size and power consumption. This document details a scalable wide I/O DRAM architecture that does not impact die size. A paper-based study was used to develop a 4 Gb DRAM architecture that can utilize a scalable I/O interface. This work exposes the challenges of creating a wide I/O memory architecture and proposes several architectural changes as solutions.

II. Builing The Memory Array

The 4 Gb DRAM architecture was developed from the bottom up (starting at the memory cell) with die size and cost as the major figures of merit. The 6F^2 DRAM unit cell, published in [3], was used as the fundamental building block. 6F^2 refers to the area of one memory cell with respect to the minimum feature size (F). This study built the 4 Gb DRAM in a 40 nm process node. This gives a minimum feature size of 40 nm and a unit cell area of 0.0096 µm^2.

The unit cell was arrayed horizontally and vertically to create a continuous memory block. The memory array contains 512 bitlines and 512 wordlines. Fig. 1 shows the 256 kb memory block. In this memory block, each crossing of a wordline and bitline contains one memory element. Additional wordlines and bitlines were added for redundancy and dummy edge elements.

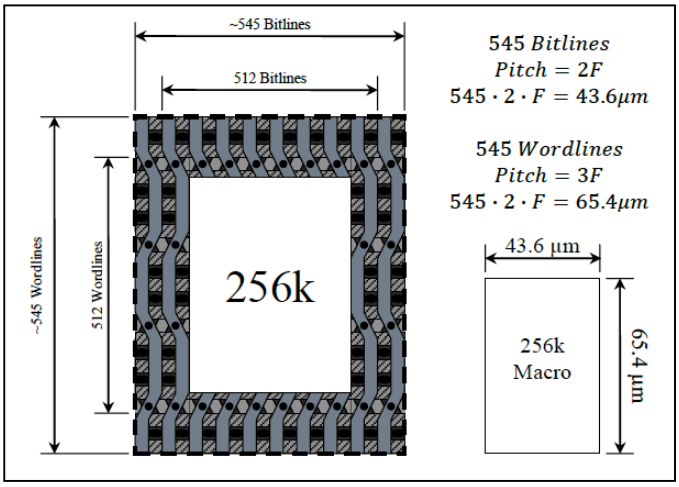

Figure 1. Creation of a 256 kb array

Periphery circuitry is required to access the memory bits contained in the 256 kb memory array. This study assumes bitline sense amplifiers (BLSA) and wordline drivers are used in the periphery. The space allocate for the bitline and wordline circuitry is 100 F [4]. This increases the 256 kb memory array size to 47.6 µm × 69.4 µm. 2048 instances of the 256 kb memory array were used to create the 512 Mb memory bank.

It is critical to configure the 2048 instances in a manner that allows for a low column cycle time. Fig. 2 shows the size of several 512 Mb bank structures. Structure D, in Fig. 2, will have the lowest column cycle time due to the smaller distance the column data has to travel. Structure D uses 128 256 kb arrays in the column-direction and 16 256 kb structures in the row direction. This gives a 512 Mb structure with 64 M-columns and 8 M-rows.

The sizes of the 512 Mb structures in Fig. 2 are derived from expanding the size of the 256 kb memory structure. The approach taken to create the memory structures allows for accurate size estimates of the 512 Mb bank.

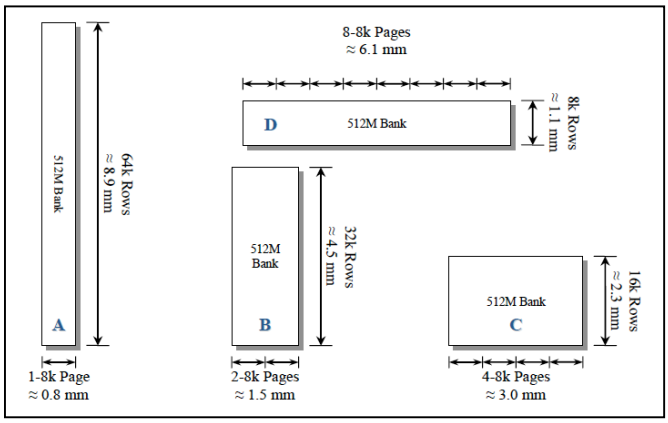

Figure 2. Creation of a 512 Mb bank

Typical DRAM devices operate with an 8 kb page size. The page size refers to the number of bitlines that are accessed when a wordline fires. The bank structure used in this work requires the use of three additional bits to be sent across the bank, with the master wordline signal, to each wordline driver. This allows the page size to be decoded when a wordline is fired. A power saving technique, which lends itself to the reduction in locality associated with multi-core processors [5], is to reduce the page size. This architecture provides an avenue for sending additional page decode bits to reduce the page size below 8 kb.

The 4 Gb DRAM architecture, seen in Fig. 3, uses the 512 Mb bank structure along with a centralized row, centralized column, and edge aligned I/O interface. The centralized row structure was allocated 300 µm. The centralized column structure was allocated 200 µm, and the edge aligned I/O interface was allocated 600 µm. The global row and column circuitry contain periphery circuitry needed to access the memory array (examples include fuses, timers, decode logic, and control circuitry). The edge aligned I/O contain pads, a datapath, voltage generators, and other periphery circuitry.

Using an edge aligned I/O interface, allows for the reduction of the number of global column structures. If the edge aligned I/O were placed in the center of the DRAM die, two global column structures would be required.

The centralized column and row circuitry allows for a reduction in the number of global circuits required. Typically, global circuitry is placed in the center of each memory bank.

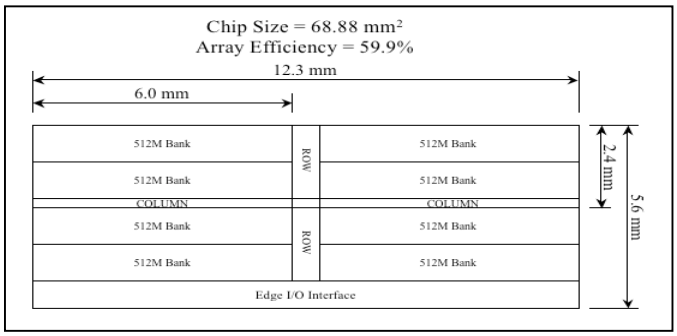

Figure 3. 4 Gb Scalable I/O DRAM Architecture

Our proposed bank structure (with a smaller global column length, see Fig. 2) allows for a centralized column structure to be used. The centralized structures lead to a reduction in the overall chip size, and an increase in the array efficiency over other architectures.

III. Accessing the Array

The wide I/O architecture increases the number of memory bits that can be accessed in parallel. As more global routing channels are used, the die size must increase to allow space for the metal tracks or an additional layer of metal is required. Each of these solutions carries with it a cost premium. Our solution reduces cost by not employing either of these solutions.

A new bank structure, discussed in Section II, and a half-bank structure are used to ensure a low cost manufacturing technology. The half-bank access scheme reduces the number of global I/O tracks required. Fig. 4 shows the implementation of a half-bank access. Each of the 512 Mb memory banks have been divided into two 256 Mb half-banks. The half-banks reside on either side of the global row. During an access, a wordline is fired in both half-banks. The 8192 page is divided between the half-banks (4096 bits are fired in each half bank). This allows half of the I/O signals to be accessed in each half-bank.

Current DDR3 DRAM, with eight data pins operating at a burst length of eight, use only 64 bits of the opened page. This proposed architecture assumes a scalable I/O interface that can utilize 512 data signals (or more) of the open page. Increasing the number of data signals read from an open page increases the energy efficiency of the DRAM. When 512 bits are accessed in parallel, each half-bank supplies 256 bits. With a 4× metal pitch, the global I/O routing consumes less than 1 % of the total metal. This amount of metal usage allows the number of global I/O signals to be scalable beyond 512 bits.

Traditional high-capacity desktop DRAM uses a triple metal process [6]. The proposed architecture can be manufactured with a two metal process. The reduction in the number of processing steps required to manufacture this architecture contributes to its low cost. The half-bank structure alleviates global routing congestion, but retrieving large amounts of data from the memory cell creates congestion at the local I/O routing level.

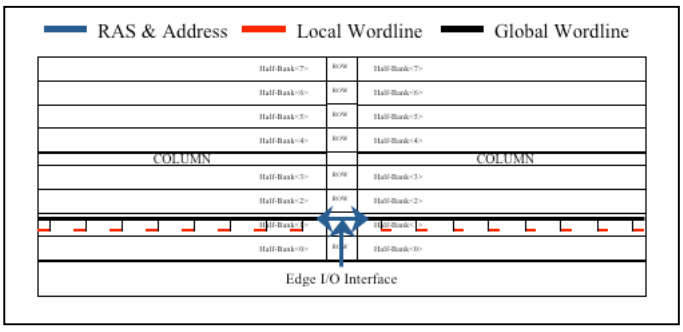

Figure 4. Half-Bank access scheme

A. *Local I/O Routing*

The major challenge associated with accessing a large number of data bits from the memory array can be found in the local I/O routing channel. As discussed in Section II, the 512 Mb bank will have 128 256 kb memory blocks. A 4096 bit half-page will require a wordline to fire in eight of the 128 memory blocks (each block having 512 bitlines). Keeping the 25.6 µm pitch of the global I/O routes, requires the memory blocks to be evenly distributed across the half-bank.

If the wide I/O architecture is operated with 64 data pins, operating with a burst length of eight, then the physical data mapping would appear as in Fig. 5. The local I/O (LIO) routing channels are located on the top and bottom of the 256 kb memory blocks. In this architecture 64 LIO channels are allocated. Using 64 LIO channels requires a large area for the metal tracks. For these reasons LIO routing is a major challenge of this architecture. The use of inactive bitlines in the neighboring memory blocks is proposed as one solution [7].

As the number of data signals is increased, this architecture becomes modular. The full column path is repeated at each DQ region. A SLICE architecture [7] can be used to simply the full chip verification process by enabling each SLICE to be verified independently.

The 64 LIO lines are differential signals. A total of 32-memory bit are transferred to the differential LIO lines from the accessed 256 kb memory block. The 32 memory bits are mapped to four DQ signals, as seen in Fig. 5. These 32 bits satisfy the burst length of eight for each DQ. Using this mapping, the number of DQ signals and burst length are scalable.

Each DQ region spans 800 µm ($1/16^{th}$ of length of the 512 Mb memory bank). The parasitic resistance and capacitance of the LIO line is another challenge for this architecture. However, this limitation is obviated when more global I/O metal tracks are used. If the global I/O tracks are distributed closer to the accessed 256 kb memory block, the distance of each LIO can be decreased.

Fig. 6 illustrates the connection of the LIO lines to the global I/O lines. The global I/O signals are routed to the global column structure located in the center of the array. The global column circuitry contains helper flip-flops to generated full logic levels, and buffers used to buffer the global I/O signal to the edge aligned I/O interface.

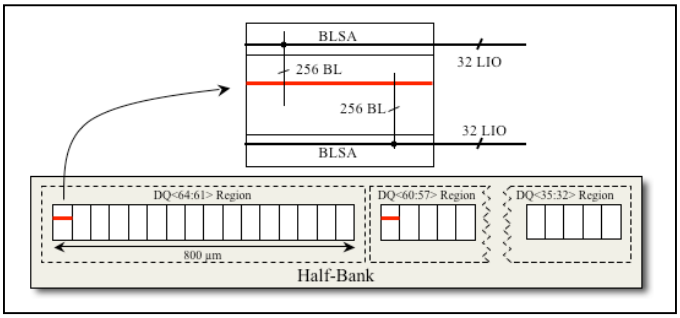

Figure 5. Local I/O routing in a half-bank

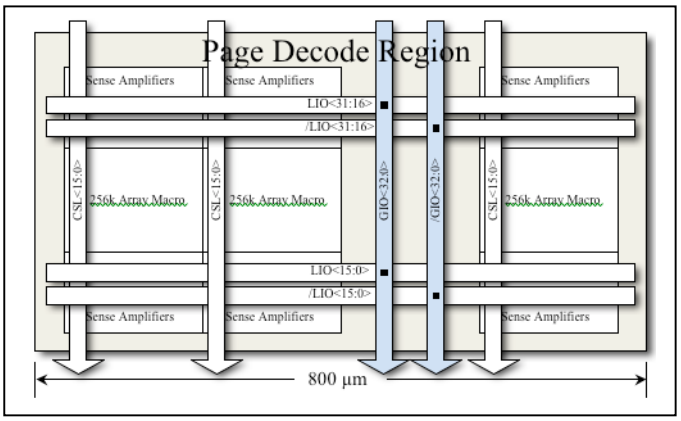

Figure 6. Illustrating the LIO to GIO connection

IV. Energy Efficency and Bandwidth

The wide I/O architecture has increased energy efficiency and bandwidth compared to traditional DRAM architectures. Typically, when a DRAM die is activated a full page (8192 bits) is accessed. Of these 8192 bits, only 64 data bits are supplied externally through the eight I/O pads found in traditional architectures. This creates a poor usage of the energy required to access 8192 bits. Energy efficiency is increased when more bits are used from the accessed page. The wide I/O architecture enables this energy efficiency. When a single die supplies 512 bits, it is only necessary to access one die to supply 64 Bytes of data. Fig. 7 compares the relative energy consumed between the proposed scalable I/O interface and a typical DDR3 desktop memory die. Fig. 7 assumes the proposed architectures are configured to use a burst length of eight, and as the number of data signals is increased fewer die need to be accessed to supply the same amount of data.

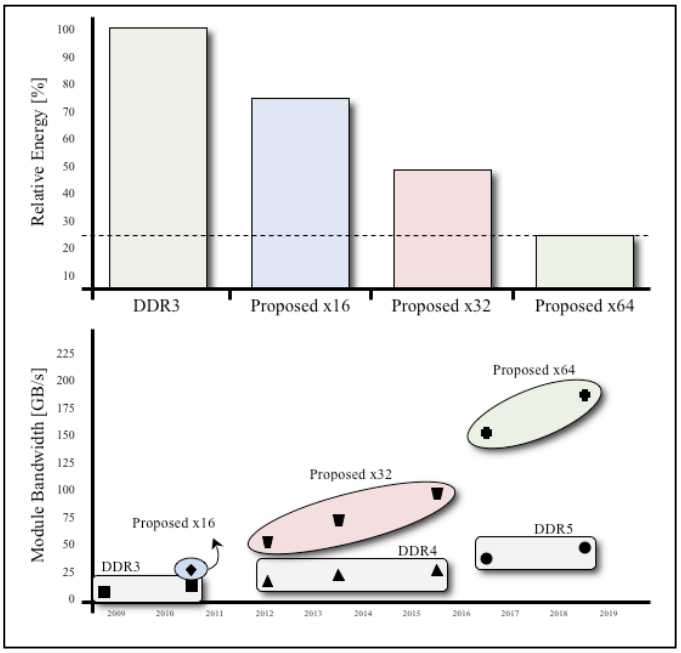

Figure 7. Energy and bandwidth improvements

The module bandwidth increase, seen in Fig. 7, is enabled by allowing each scalable I/O DRAM to supply data to the memory channel in parallel. In this way, a single die can provide several module configurations were developed to enable the larger number of data signals [8].

Utilizing a large number of I/O pads requires the use of low power receiver designs. Mobile wide I/O architectures utilize TSV connections that have smaller parasitics compared to wire bond connections. This allows the receivers to consume lower power. A low cost alternative to TSV connections is capacitive-coupled interconnects. This wireless technology achieves the same chip-to-chip interconnect pitch as TSV (on the order of 10 μm), cost less, has higher bandwidth, and achieves lower power [9].

The edge aligned I/O interface is suited for capacitive coupling between die. Traditionally, mobile DRAM multi-chip packages contain two-DRAM die on top of the processor. Two edge aligned I/O interfaces, utilizing capacitive coupled interconnects, can be placed face down on top of the processor. This allows for a low cost alternative for a mobile wide I/O DRAM architecture.

V. SUMMARY

A 4 Gb DRAM architecture that uses a scalable I/O interface was developed in this study. A $6F^2$ memory cell was used to construct the memory array and to get reasonable die size estimates. The 4 Gb DRAM architecture used a 40 nm minimum feature size, measured 68.88 mm^2, and achieved an array efficiency of 59.9 %. The architecture can be manufactured in a two metal process. The die size and metal process allows this architecture to be considered low cost.

The architecture utilizes a distributed page and half-bank memory structure to distribute the large number of data signals across the memory array. Global and local routing challenges were exposed and several novel solutions were presented. The scalable I/O interface allows the more data bits to be read from a single memory die. This increases the energy per bit efficiency compared to traditional architectures. This is due to using more of the bits of an open page. The scalable I/O interface can also be used in a high bandwidth application. The large number of data signals read from the array can be presented to the memory channel. This allows the memory bandwidth to be increased. Fig. 6 depicts these improvements.

REFERENCES

[1] T. Sekiguchi, K. Ono, A. Kotabe, Y. Yanagawa, "1-Tbyte/s 1-Gbit DRAM Architecture Using 3-D Interconnect for High-Throughput Computing," *Solid-State Circuits, IEEE Journal of,* vol. 46, no. 4, pp. 828-837, April 2011

[2] J. Kim, et al., "A 1.2V 12.8 GB/s 2Gb mobile Wide-I/O DRAM with 4x128 I/Os using TSV-based stacking," *Solid-State Circuits Conference, IEEE International,* pp. 496-498, 20-24 Feb. 2011

[3] F. Fishburn, et al., "A 78nm $6F^2$ DRAM technology for multigigabit densities," *VLSI Technology, Symposium on*, pp. 28-29, 15-17 June 2004

[4] B. Keeth, R.J. Baker, B. Johnson, F. Lin, *DRAM Circuit Design: Fundamental and High-Speed Topics, Second Edition,* Wiley-IEEE, 2008

[5] A. Jung, J. Leverich, R.S. Schreiber, N.P. Jouppi, "Multicore DIMM: an Energy Efficient Memory Module with Independently Controlled DRAMs," *Computer Architecture Letters,* vol. 8, no. 1, pp. 5-8, Jan. 2009

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

[7] Q. Harvard, "Wide I/O DRAM Architecture Utilizing Proximity Communication," Boise State University Theses and Dissertations, Paper 72, 2009

[8] Q. Harvard, "Low-Power, High-Bandwidth and Ultra-Small Memory Module Design," Boise State University Theses and Dissertations, [submitted] 2011.

[9] Q. Harvard, R.J. Baker, "A 4.0 Gbps 15 fJ/bit Receiver Design for Capacitive-Coupled Wireless Interconnects," *Custom Integrated Circuits Conference (CICC), 2011*, [submitted] Sept. 2011