3D-Wiz: A Novel High Bandwidth, Optically Interfaced 3D DRAM Architecture with Reduced Random Access Time

Ishan G Thakkar, Sudeep Pasricha
Department of Electrical and Computer Engineering
Colorado State University, Fort Collins, CO, U.S.A.
{ishan.thakkar, sudeep}@colostate.edu

Abstract — **This paper introduces 3D-Wiz, which is a high bandwidth, low latency, optically interfaced 3D DRAM architecture with fine grained data organization and activation. 3D-Wiz integrates sub-bank level 3D partitioning of the data array to enable fine-grained activation and greater memory parallelism. A novel method of routing the internal memory bus using TSVs and fan-out buffers enables 3D-Wiz to use smaller dimension subarrays without significant area overhead. This in turn reduces the random access latency and activation-precharge energy. 3D-Wiz demonstrates access latency of 19.5ns and row cycle time of 25ns. It yields per access activation energy and precharge energy of 0.78nJ and 0.62nJ respectively with 42.5% area efficiency. 3D-Wiz yields the best latency and energy consumption values per access among other well-known 3D DRAM architectures. Experimental results with PARSEC benchmarks indicate that 3D-Wiz achieves 38.8% improvement in performance, 81.1% reduction in power consumption, and 77.1% reduction in energy-delay product (EDP) on average over 3D DRAM architectures from prior work.**

I. INTRODUCTION

In recent years, DRAM latency has not improved as rapidly as DRAM capacity and bandwidth with shrinking technology, owing to the well-known "memory-wall" problem [1]. Continued process technology scaling has enabled commodity memory system solutions to exploit smaller and faster transistors to improve memory capacity and bandwidth. However, the traditional means of improving memory access performance by increasing clock frequency is no longer practical due to increasingly stringent power constraints that limit further frequency scaling. Thus, performance improvements in memory systems today are primarily reliant on latency tolerance techniques such as multi-level caches, row prefetching, burst mode access, memory scheduling [20] and memory parallelism [2][8]. But the performance improvements obtained through these techniques are not expected to scale well for high performance computing systems of the future [18]. Moreover, preserving the minimum standard capacitance of a DRAM cell is becoming increasingly challenging with shrinking feature size [3]. These trends are forcing designers to reinvent DRAM architectures so as to overcome the hurdles in DRAM performance scaling.

Since the emergence of 3D integration technology, 3D-stacked DRAM has been a promising option to alleviate many of the physical limitations of commodity DRAMs. In recent years, several 3D DRAM architectures have been proposed [4][5][6][7], that have exploited fine-grained partitioning and activation to notably improve memory parallelism. Such 3D-stacked DRAMs require new methods for efficient address and data path routing and 3D cell organization, to realize their potential and achieve improved memory parallelism and latency.

In this paper, we propose 3D-Wiz, a new 3D DRAM architecture with a photonic interface that improves access latency and energy consumption characteristics over prior efforts. Our key contributions can be summarized as follows:

- **Fine-grained 3D organization of DRAM data array:** Bank-level parallelism in a commodity DRAM data array is limited by the total number of banks the data array is partitioned into. A typical DRAM data array is partitioned into 4 to 16 banks. These banks are very large in capacity, ranging from 32Mb to 256Mb, and hence they are few in number. To enable higher memory parallelism in 3D-Wiz, a large bank is further divided into multiple smaller banks called sub-banks. All the subarrays in a sub-bank are addressed in parallel and work in lock-step to serve an entire cacheline. Each sub-bank is many-fold smaller than a commodity bank and acts independently, which greatly increases memory parallelism and reduces energy.
- **All TSV, 3D upright routing of internal memory bus**: In 3D-stacked DRAM, the address and data buses are routed to each die layer in the vertical direction using TSVs, and after reaching a die they are routed to the edges of individual banks along 2D routing paths. The pre-decoded address lines are routed to individual subarrays along the global word lines and global column select lines. At a subarray level, the global address lines are further decoded to drive local word lines and local column select lines. Similarly, data buses reach individual subarrays as local data lines, which are driven by global data lines. These 2D routing paths and global peripheral circuits (decoders, repeaters and drivers of global lines) incur latency, energy and area overhead at each die. We found in our study that by intelligently using TSVs to route data and address lines to individual subarrays, the need to use 2D routing paths and global peripheral circuits at each die can be eliminated, which enables the use of smaller sized subbanks and subarrays. This significantly minimizes overall area, latency and energy consumption.
- **Reduced random access time:** The DRAM access time is a function of the sum of row to column command delay (t_RCD) and column command to data out delay (t_CAS). For a subarray of interest, the timing constraints t_RCD and t_CAS are proportional to the capacitive loading of word lines and bit-lines respectively, which are increasing functions of the number of columns and number of rows respectively [8]. We reduced the access time of 3D-Wiz by carefully reducing the number of rows and columns in a subarray so as not to harm the DRAM cell area efficiency compared to the area efficiency of other 3D-stacked DRAM architectures.
- **High bandwidth photonic interface:** We observed that the conventional electrical interface of the DDRx family and the differential lane based interface of Micron's HMC [7] cannot support very high bandwidths. So in 3D-Wiz, we propose using a higher bandwidth, dense wavelength division multiplexed (DWDM) based photonic interface.

978-1-4799-6492-5/14/$31.00 ©2014 IEEE

II. Background and Motivation

In this section, we briefly discuss the state-of-the-art data array organizations for 3D-stacked DRAMs, and identify the fundamental elements that contribute significantly to the overall latency, energy consumption, and die area.

Different organizations of data array differ in the way banks and ranks are stacked and partitioned across the 3D die-stack. Accordingly, the 3D organizations of data array are referred to as coarse-grained or fine-grained rank or bank-level partitioning [9]. Loh [4] was the first to highlight the potential performance benefits achievable by altering the data array organization for conventional 3D-stacked DRAMs. Previous approaches (prior to [4]) did not fully exploit 3D-stacking technology, as the individual structures inside DRAMs were still 2D. Loh [4] proposed splitting a rank across multiple layers instead of laying out a rank on a single layer, and showed up to 1.75× DRAM performance improvement. However, the area overhead due to increased number of TSVs per rank was not considered. Woo et al. [5] proposed the SMART-3D DRAM architecture that used a vertical L2-fetch and write-back network made up of a large array of TSVs to hide the latency behind large data transfers. Kang et al. [6] proposed extending the commodity DDR3 architecture to 3D-DDR3 and showed benefits in performance due to the reduced length and latency of a TSV-based memory bus. Micron's hybrid memory cube (HMC) exploits fine-grained rank-level partitioning to further reduce the access latency and increase memory parallelism [7]. All of these approaches [4]-[7] exploit the bandwidth from vertical TSVs to reduce access latency. But, they miss out on the potential of 3D data array organization and TSVs to more aggressively improve performance and reduce activation-precharge energy in DRAM data arrays.

Chen et al. [9] discussed the pros and cons of coarse-grained and fine-grained rank-level partitioning for 3D DRAMs with respect to latency, energy consumption, and area efficiency. For the coarse-grained rank-level cell organization, the inter-bank communication within one rank occurs via 2D routing paths. Chen et al. [9] show that the activation energy and latency of the fine-grained design are reduced by 48.5% and 46.9% respectively compared to the coarse-grained design, because of the reduced bank size and optimized data path routing in the fine-grained design. The total die area for the fine-grained design reduces by 35.9% in spite of a 3.7% TSV area overhead. Also, the latency of the internal memory bus is reduced by 62.8% for the fine-grained design. The reason behind the improved results for the fine-grained model is that the model utilizes the potential bandwidth of TSVs for inter-bank transfers, which relaxes the need for 2D routing paths for such transfers and alleviates related overheads. From these results, it can be concluded that many of the limitations that arise with DRAM scaling can be overcome by intelligently organizing the data array. Moreover, by carefully choosing the bank size and an appropriate scheme for TSV based routing of address and data paths, DRAM performance and energy consumption may be improved even further.

The organization of the DRAM data array is only one of the limiting factors that affect DRAM performance and energy. The other more fundamental limiting factor is RC loading of the memory access path, which is a function of the number of rows and columns in the accessed subarray. This fact encourages designers to reduce the number of rows and columns in a subarray. But, such a reduction can significantly increase the area overhead, harming overall area efficiency, if it is done without due deliberation. *3D-Wiz exploits the benefits of intelligent array organization and aggressive vertical routing of address and data paths to counter-balance this overhead, which enables the use of smaller dimension subarrays.*

Furthermore, the concurrency in state-of-the-art DRAM modules is best exploited only inside the banks due to limitations arising from bus contention outside the banks. The contention is a result of internal memory bus and IO bus sharing among all the banks due to pin-bandwidth limitations of the traditional DRAM interface. Therefore, increasing concurrency within a memory system has limited benefits unless this pin-bandwidth limitation is alleviated. *3D-Wiz exploits intelligent layout of TSV buses and uses a high speed photonic interface to address this issue.*

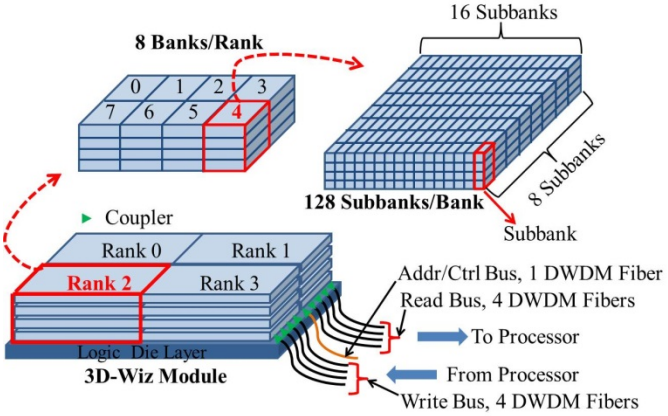

Fig. 1: Schematic of 3D-Wiz architecture and constituent elements.

III. 3D-Wiz Architecture: Overview

In this section, an example DRAM design is described to demonstrate the 3D-Wiz architecture. As shown in Fig. 1, one 3D-Wiz module of 8 Gb capacity (50nm) consists of a stack of 4 dies, which is divided into 4 identical ranks. Each rank consists of 8 identical banks. The stack of 4 DRAM dies is further stacked onto one logic die. All of the global control logic (except subarray-level control) of DRAM is integrated on the logic die. The logic die also contains all the opto-electrical circuits required to support a bidirectional, 256-bit wide, DWDM data bus, which acts as an interface between the processor side memory controller and 3D-Wiz module. The following subsections describe the 3D-Wiz architecture in more detail.

A. Fine-grained Data Array Organization: Subbank-level Stacking and Partitioning

To enable fine-grained activation, a bank of 256Mb size is divided into 128 smaller subbanks of size 2Mb each. Each subbank works independently and is capable of serving an entire cache line. In other words, each subbank is functionality wise similar to a bank, except that it is smaller in size. Therefore, partitioning of a rank (of 2 Gb size) into 8 identical banks is amounts to a partitioning of the rank into 1024 identical subbanks. Each subbank is folded across 4 die layers and consists of a total of 32 subarrays, 8 subarrays of which are on one layer. Fig. 2 shows the layout of a subbank on one die layer. This fine-grained array organization of 3D-wiz increases memory parallelism manifold, as each subbank can work independently. In commodity DRAMs, the tFAW power constraint significantly reduces bank-level parallelism in a rank [19]. However, the smaller size of a subbank in 3D-Wiz eliminates the tFAW constraint. This theoretically allows any number of subbanks per rank to be activated in a given period of time. But, due to the photonic bus interface contention, the memory controller cannot

send ACT/READ/WRITE command to more than one subbanks concurrently. However, the requests to different subbanks can be pipelined through the photonic bus, because the internal memory bus made of TSVs is not shared among all the subbanks (as will become clear in the following subsection). As a result, ideally, all the subbanks in a rank can be activated in a pipline.

B. Subbank Floorplan and 3D Signal Routing for Internal Memory Bus

Fig. 2 shows the layout of subbanks and constituent subarrays on one die layer, along with the TSV bus layout. As mentioned earlier, a single subbank consists of 32 subarrays. All 32 constituent subarrays of a subbank are addressed in parallel, and they work in lockstep to serve a cache line. Each subarray has 4 data lines, which allows a subbank to serve a total of 128 data bits in one data burst. Therefore, 3D-wiz requires a burst length of 4 to serve a 64B cache line. As shown in the figure, all 8 subarrays belonging to the on-die part of a subbank are lined up along the Y-direction, with number of subarrays along the X-direction being one. In other words, a subbank is 8 subarrays long, one subarray wide and 4 die layers high. In a bank, a total of 8 subbanks are arranged in the Y-direction to form one column of subbanks. A total of 16 subbank-columns are arranged in the X-direction, which is also the number of subbanks in a subbank-row. Two adjacent subbank-columns share one TSV bus, which is laid out in the Y-direction along the length of a subbank-column. As shown in the figure, each TSV bus is logically partitioned into 8 sections, where each section is shared between two subbanks in adjacent columns and is responsible for routing decoded data and address lines to the shared subbanks.

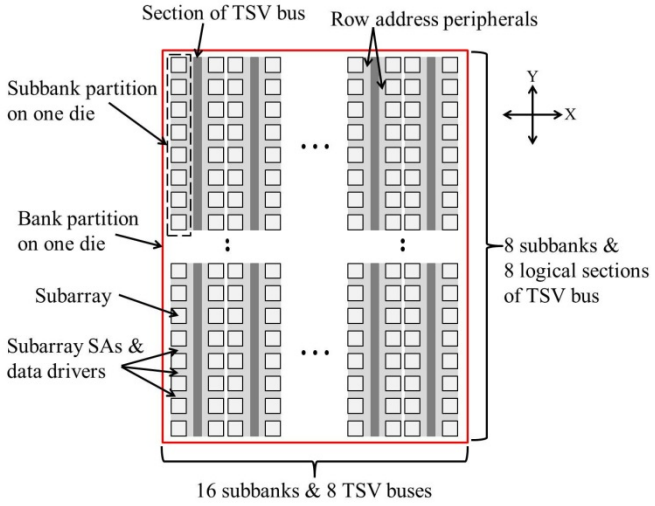

Fig. 2: Schematic floorplan of a bank partition on one die.

3D-Wiz benefits in performance, area, and latency from the aforementioned arrangement of subbanks and TSV bus. The pre-decoder and decoder circuits for row and column address lines are located on the logic die. The decoded address lines are routed vertically via the TSV bus, which feeds the local word lines and column select lines of subarrays in a subbank. This arrangement eliminates the need for using on-die 2D routing, global word lines, global column select lines, and circuits connecting global lines with local lines. Elimination of on-die 2D routing of data and address paths to individual subbanks and the placement of the decoder circuits on the logic die saves significant die area, which helps in counter-balancing the overhead of 32 TSV buses.

Only one section (out of 8 total sections) of one TSV bus is responsible for feeding the address lines to all 32 parallel addressed subarrays of a subbank. Intuitively, this should require 32 sets of decoded row address and column address TSVs per TSV bus section. But such an organization would significantly increase the TSV bus area, jeopardizing all of the benefits of vertical routing using TSVs. Alternatively, as all the subarrays in a subbank are addressed in parallel, we propose the use of only one set of address TSVs. In this case, each address TSV, when it reaches a die layer, is used to feed in one 1:8 fan out buffer after a repeater stage. Using one 1:8 fan out buffer per address TSV on each die layer enables 3D-Wiz to drive all 32 subarrays of a subbank using just one set of address TSVs. The area overhead incurred due to fan out buffers is insignificant compared to the benefits, as discussed in the next paragraph.

Fig. 3 shows the on-die layouts of a TSV bus section and shared subbanks for two different cases. Fig. 3(a) shows the layout for the case when a dedicated set of address TSVs is used for each subarray of a subbank. Fig. 3(b) depicts the layout for the case when a single set of address TSVs are used to drive all 32 subarrays of a subbank. In the former case, a dedicated address TSV and a dedicated control TSV are required per wordline of each subarray, with total number of row address TSVs per subbank equal to 2 × number of wordlines in a subarray × number of subarrays in a subbank. The control TSV is used to select between the shared subbanks. In the latter case, an address TSV and a control TSV are required per wordline of one subarray, with a significantly reduced total number of row address TSVs per subbank equal to 2 × number of wordlines in a subarray. Thus using one 1:8 fan out buffer per address TSV notably reduces the TSV bus area for the latter case.

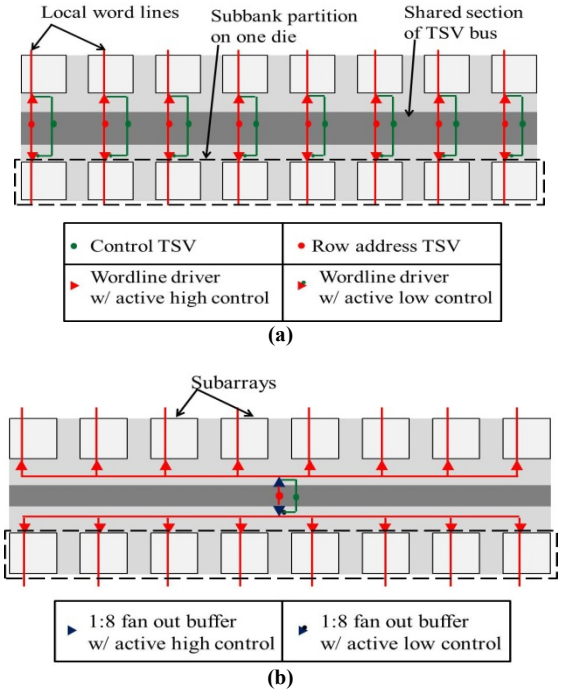

Fig. 3: Schematic layout of the partition of one TSV bus section on one die. (a) when a dedicated set of address TSVs is used for each subarray of a subbank. (b) when a single set of address TSVs is used to drive all 32 subarrays of a subbank using 1:8 fan out buffers.

C. Random Access Time

Every subarray in well-known 3D DRAM designs has 512 rows and 512 columns, whereas, a subarray in 3D-Wiz DRAM has 256 rows and 256 columns. While designing the 3D-Wiz

architecture, the number of rows and number of columns in a subarray were set to 256, because this value gave the anticipated benefits without incurring a significant area overhead (see Section IV). Consequently, the reduction in capacitive loading of wires because of smaller subarray size and elimination of global lines yields, for 3D-Wiz, the access time and the row cycle time of 19.5ns and 25ns respectively, which is significantly lower than in other architectures, as discussed in more detail in Section IV.

D. High Bandwidth Photonic Interface

In this subsection, we first discuss the maximum theoretical bandwidth achievable with 3D-Wiz and justify the use of a high speed and high bandwidth photonic interface, before describing the functionality of the logic die that supports the interface.

1) Bandwidth Analysis

As discussed earlier, ideally 1024 subbanks can be activated in pipeline per rank of the 3D-Wiz DRAM, which makes it possible to serve one request (64B) per cycle assuming at least 25ns (tRC) time between two consecutive requests to the same subbank and assuming 50 stages deep pipeline (25ns (tRC) × 2GHz (clock rate) = 50). The logic die of the 3D-Wiz contains one 128-bit wide and 4-entry deep register bank per TSV bus section to facilitate data buffering for pipeline. The value of the number of entries in the data buffer per TSV bus section is assumed to be 4.

This organization enables each 3D-Wiz module to achieve a peak data rate of 512x4 bits/cycle (512bits/cycle/rank), if the memory controller is designed to send requests to all four ranks in parallel. This is equal to 512GBps peak bandwidth at 2GHz clock rate. A conventional electrical bus based interface of the DDRx family cannot support such a high bandwidth due to pin-bandwidth limitations. To address this issue, the Hybrid Memory Cube (HMC) from Micron proposes the use of high speed serial links at the interface [7]. The high speed link in HMC consists of several differential lanes as the fundamental building blocks. Each differential lane is claimed to achieve the maximum data transfer rate of 10Gbps. Thus, 3D-Wiz would require about 410 such differential lanes to achieve a 512GBps data transfer rate across the interface. This is a prohibitively large number of lanes that would cause serious packaging issues due to pin-limitations.

Alternatively, several recent works on DRAM architectures have proposed dense wavelength division multiplexed (DWDM) photonic interfaces to achieve high bandwidth [10], [11]. A DWDM optical fiber can carry approximately 64 wavelengths, which creates 64 channels on a single fiber [21]. A typical DWDM fiber link consists of several components such as ring modulators, photonic waveguides, SerDes (serialization/deserialization) components, and photo detectors, which work together in sync to achieve high speed and high bandwidth data transfers [21][22]. The resulting bandwidth for a DWDM fiber link depends on the cumulative bandwidth of individual components. In recent years, innovations in Si-photonics technology have enabled on-chip Si waveguides, ring modulators, and ring detectors to function at 20Gbps data rate [12]. Also, advancements in CMOS technology have enabled a single lane SerDes to operate at 25Gbps data rate [13]. Thus, cumulatively, single wavelength channels within fiber links can today operate at 20Gbps data rate, which corresponds to a 160GBps bandwidth per DWDM fiber. Consequently, the bandwidth requirement of 3D-Wiz can be fulfilled by just 4 such DWDM fibers (~640GBps), which is easily achievable well within the pin-constraints. Therefore, we propose utilizing such a high bandwidth photonic interface for 3D-Wiz.

The photonic interface in 3D-Wiz is comprised of two links (as shown in Fig. 1): one for reads and the other for writes. Each link consists of 4 unidirectional DWDM fibers. Each DWDM fiber in read and write links supports 64 wavelengths, making the read and write bus to be 256-bits wide each. The photonic interface also uses one additional link consisting of a single DWDM fiber for transmitting address and control signals. The address/control fiber supports a total of 40 wavelengths, with 32 wavelengths for addresses and 8 wavelengths for control signaling.

2) Design and Functioning of Logic-die

Fig. 4 shows the functional block diagram of the logic die. The logic die is functionally divided into two parts: photonic and electrical. The photonic part of the logic die does all of the optical to electrical (O-to-E) and electrical to optical (E-to-O) conversions; and optically interfaces with the memory controller on the processor chip. As shown in the figure, the photonic part of the logic die consists of on-chip Si waveguides (Si WGs), photonic couplers, and photonic stops. The electrical part of the logic die consists of address/control decoder logic, TSV drivers, read/write buffers, and the control and power delivery network. It interfaces with the DRAM cell array through TSV buses. The functionality and role of all of these components on the logic die is discussed next.

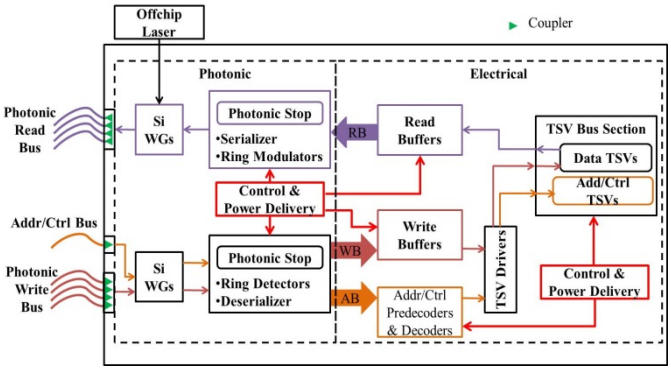

Fig. 4: Functional block diagram of the logic die in 3D-Wiz

For each transaction, firstly the incoming bit stream (originating at the processor) from the photonic address/control bus is coupled to the on-chip Si waveguides by couplers. These photonic couplers are primarily responsible for realizing a low loss coupling between on-chip Si waveguides and off-chip DWDM fibers. Next, the bit stream is passed through a photonic stop, where the constituent ring detectors convert the photonic bit stream into an electrical bit stream. The serialized electrical bit stream is then de-serialized before it electrically drives the address/control bus (AB). The signals of this AB are decoded before they are used to drive an appropriate section of one of the 32 TSV buses. The decoded address/control signals activate an appropriate subbank to serve the subsequent read or write request. For a write request, the data is written to the requested subbank from the write buffer through the TSV bus section.

It is important to provide a write buffer at the interface of the electrical write bus (WB) and TSV bus, because the rate at which data is written on the electrical WB (for a 10GHz photonic clock rate) is greater than the rate at which it can be written to the subbank (for a 2GHz memory clock rate). Similarly, read buffers are provided at the interface of the electrical read bus (RB) and TSV bus to facilitate synchronization of data transfer between two different clock rates. One 128-bit wide and 4-entry deep buffer is allocated to each section of the TSV bus. Lastly, the control and power delivery unit of the logic die controls the on die circuits and delivers power to the DRAM module.

IV. 3D-WIZ AREA, TIMING AND ENERGY ANALYSIS

The area, timing and energy analysis for the 3D-Wiz architecture was performed by modifying the code for CACTI-3DD [9]. A similar analysis was conducted for other well-known 3D DRAM architectures such as the 3D stacked photonic DRAM (3DSPDRAM) [11], 3D DRAM from Samsung (3DSams) [6], and the hybrid memory cube (HMC) from Micron [7]. The results of the analysis for these architectures were compared with the results for our 3D-Wiz and with commodity DDR3 DRAM architectures. The models of the aforementioned 3D DRAM architectures were implemented in CACTI-3DD to the best of our knowledge using the technology parameters for the 50nm node. The 3D DRAM architectures from prior work were chosen so as to provide a broad coverage of the full spectrum of 3D DRAM designs. For instance, the 3D DRAM architecture from Samsung realizes coarse-grained rank level stacking [9], whereas 3DSPDRAM implements single subarray access for a coarse-grained rank-level 3D data array organization [9]. On the other hand, the HMC architecture from Micron is designed to exploit fine-grained rank level partitioning [9].

Table 1: Timing parameters for different DRAM designs

	t_AC (ns)	t_RAS (ns)	t_RP (ns)	t_RC (ns)
3D-Wiz	19.5	20.4	4.7	25.1
3D-Wiz 512	21.7	25.3	6.3	31.6
HMC	22.5	27.4	10.6	38
3DSPDRAM	26.3	28.3	12	40.3
3D-Sams	24.7	40.5	21.3	61.8
DDR3	36.4	32.2	33.2	69.5

The energy, timing and area analysis for 3D-Wiz with two different subarray sizes was done to select the optimum subarray size entailing the best results without significant area overhead. The results for the following two subarray sizes are given in this section: 1) Subarray with 512 columns and 512 rows (512×512), and 2) subarray with 256 columns and 256 rows (256×256). The 3D-Wiz architecture with 512×512 subarray dimensions is referred to as '3D-Wiz 512' in the rest of the paper, and the 3D-Wiz architecture with 256×256 subarray dimensions is referred to as '3D-Wiz'.

Table 2: Per access energy values for different DRAM designs

	Activation Energy (nJ)	Precharge Energy (nJ)	Refresh Power (mW)
3D-Wiz	0.78	0.62	0.26
3D-Wiz 512	1.27	1.08	0.31
HMC	1.23	1.13	0.20
3DSPDRAM	1.08	1.01	0.20
3D-Sams	1.81	1.68	0.77
DDR3	7.45	7.30	1.67

Table 1 lists the different timing parameters obtained from CACTI-3DD based models for the DRAM designs considered in our analysis. 3D-Wiz demonstrates access latency of 19.5ns and row cycle time of 25ns, which are better than other architectures on average by about 27% and 52% respectively. The HMC entails the second best results for access latency and row cycle time.

Similarly, Table 2 gives the per access values of activation-precharge energy for all the various DRAM designs. It also lists refresh power. 3D-Wiz yields per access activation energy and precharge energy of 0.78nJ and 0.62nJ respectively, which demonstrates significant improvement over other DRAM architectures. 3D-Wiz has the best results for energy because of the reduction in subarray dimensions and also because the size of a subbank in 3D-Wiz is smaller than the bank size in other architectures. The smaller bank size (or fine-grained activation) also benefits HMC to yield low activation-precharge energy compared to 3D-Sams. In spite of a similar bank size and data array organization, 3DSPDRAM has lower energy than 3D-Sams, because 3DSPDRAM implements a single subarray access (SSA) [11]. In a SSA implementation, an entire cache line is served by a single subarray [11], which reduces the granularity of activation for 3DSPDRAM compared to 3D-Sams, reducing the activation energy for 3DSPDRAM.

The cost of a DRAM module depends on the area efficiency of the DRAM die. Therefore, area efficiency is a very important parameter to consider while designing a new DRAM architecture. Area efficiency of a DRAM die is defined as the percentage of the total die area which corresponds to the DRAM cell area. The total DRAM die area includes the area covered by peripherals, memory bus and TSVs in addition to the area covered by DRAM cells. A DRAM architecture with a high area efficiency is thus able to store more useful information compared to a lower area efficiency DRAM architecture, within the same die area.

Table 3: Area efficiency, TSV area overhead for different DRAM designs

	Area Efficiency (%)	TSV area overhead (mm^2)
3D-Wiz	42.5	3054×10^{-3}
3D-Wiz 512	38.4	7134×10^{-3}
HMC	54.5	44.7×10^{-3}
3DSPDRAM	42.6	0.8×10^{-3}
3D-Sams	43.7	1.7×10^{-3}
DDR3	46.7	-

Table 3 shows the results for area efficiency and TSV area overhead of different DRAM designs. All TSVs in this study were modeled based on ITRS projections for intermediate interconnect level TSVs [14]. The area overhead of a fan out buffer was calculated to be $5.2\mu m^2$ which makes the total area of fan out buffers to be $6mm^2$ (about 6.3% of the total die area). It is evident from the results that despite the significant increase in the TSV area overhead for 3D-Wiz, its area efficiency is comparable to most of the architectures listed in the table. This is because the TSV area overhead is counter-balanced by the area benefits obtained by optimized interconnects and relocation of decoder logic on the logic die. However, the comparable area efficiency of 3D-Wiz does not translate into great cost benefits in this case, as the cost of logic die has to be added in the total cost of the DRAM module. Nonetheless, the significant improvements in access time and energy consumption make 3D-Wiz a promising architecture candidate for future 3D DRAMs. The next section presents more detailed experimental comparisons, for real application runs.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

We performed trace-driven simulation analysis to compare 3D-Wiz with other DRAM architectures. Memory access traces for the PARSEC benchmark suite [17] were extracted from detailed cycle-accurate simulations using gem5 [16]. We considered twelve different applications from the PARSEC suite: Blackscholes (A), Bodytrack (B), Canneal (C), Dedup (D), Facesim (E), Ferret (F), Fluidanimate (G), Freqmine (H), Streamcluster (I), Swaptions (J), Vips (K), x264 (L). We ran each PARSEC benchmark for a "warm-up" period of 100 million instructions and captured memory access traces from the subsequent 50 million instructions extracted. These memory traces were then provided as inputs to the DRAM simulator

DRAMSim2 [15], which we modified heavily to model 3D-Wiz and other DRAM architectures.

Table 4: Modeling parameters for the photonic interface

Read laser power [11]	1.22 mW
Write laser power [11]	0.49 mW
Addr/Ctrl laser power [11]	0.24 mW
Ring detector energy [11]	44 fJ/bit
Thermal tuning power [11]	50µW/ring
Ring modulator energy [11]	47 fJ/bit
Modulator/detector delay [12]	0.05 ns
SerDes delay [13]	0.04 ns

Table 5: DRAMSim2 simulation configurations

3D-Wiz	8Gb module, 4 ranks, 1024 subbanks/rank, Stacked die count: 4 128-bit wide TSV data bus, Burst length 4
HMC	8Gb module, 16 banks/rank (vault), Stacked die count: 4 256-bit wide TSV data bus, Burst length 2
3DSPDRAM	8Gb module, 8 banks/rank, Stacked die count: 8 16-bit wide TSV data bus, Burst length 32, Single subarray access (SSA)
3DSAMS	8Gb module, 4 ranks, 8 banks/rank, Stacked die count: 4 32-bit wide TSV data bus, Burst length 16
DDR3	8Gb module, 1 rank, 8 banks/rank, 64-bit wide JEDEC data bus, Burst length 16

Table 4 gives the parameters used for modeling the photonic interface. We carefully modeled the differential lane based memory interface of HMC and the photonic interface of 3D-Wiz to the best of our knowledge. Table 5 shows the memory configurations used in DRAMSim2 for the comparison across different DRAM architectures. A FCFS scheduling scheme and a closed page policy were used for all simulations. Performance, total power consumption, and energy-delay product values for the memory subsystem were obtained from DRAMSim2. Performance was calculated as the inverse of average access latency. The results of simulations with synthetic and PARSEC benchmarks are discussed in the following subsections.

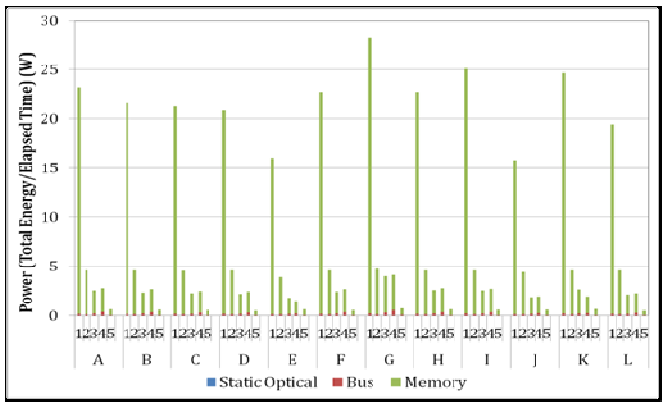

Fig. 5: Power consumption values for PARSEC benchmarks. (1: DDR3, 2: 3DSPDRAM, 3: 3DSAMS, 4: HMC, 5: 3D-Wiz).

B. Results for PARSEC Benchmarks

This subsection presents the performance, power, and energy-delay product values for all of the 3D DRAM designs shown in Table 5 obtained for PARSEC benchmarks. Fig. 5 shows power consumption values for the various DRAM architectures across the PARSEC benchmarks. The figure gives the total power which is a sum of static optical power, dynamic memory access power (labeled as 'Memory') and power consumed by the interface bus (labeled as 'Bus') that interfaces the DRAM module with the processor. It can be observed that 3D-Wiz consumes about 81.1% less power on average over all the other 3D DRAM architectures. More specifically, 3D-Wiz consumes about 97.2%, 87%, 75.3% and 75.4% less power on average over DDR3, 3DSPDRAM, 3DSAMS and HMC respectively. The reason for the lower power consumption in 3D-Wiz is smaller values of per access activation-precharge energy, the effect of which cumulates over time to minimize average power consumption.

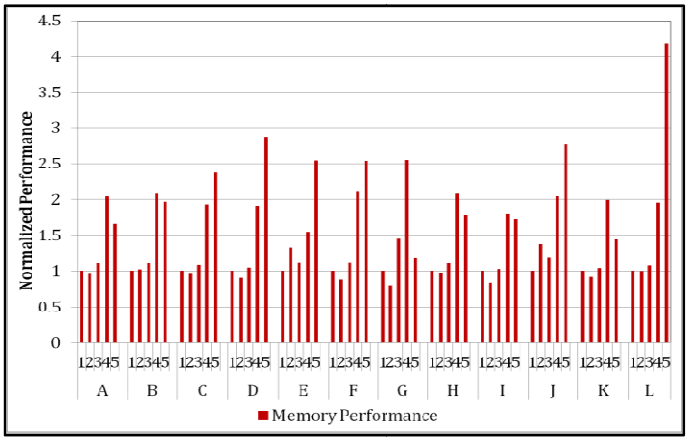

Fig. 6: Performance values for PARSEC benchmarks. (1: DDR3, 2: 3DSPDRAM, 3: 3DSAMS, 4: HMC and 5: 3D-Wiz).

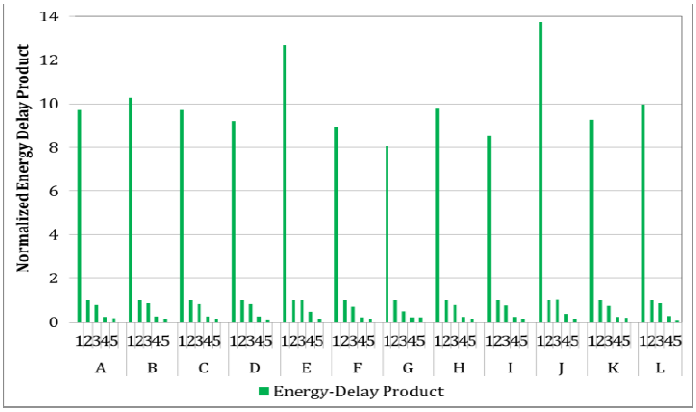

Fig. 7: Energy-delay product values for PARSEC benchmarks. (1: DDR3, 2: 3DSPDRAM, 3: 3DSAMS, 4: HMC and 5: 3D-Wiz).

Fig. 6 shows performance values for different DRAM designs normalized to the performance of DDR3, across the PARSEC benchmarks. 3D-Wiz demonstrates about 38.8% greater performance on average over all the other 3D DRAM architectures. More specifically, 3D-Wiz demonstrates about 50.5%, 53%, and 43.4% greater performance values on average over DDR3, 3DSPDRAM and 3DSAMS respectively. For some applications, the performance of HMC is slightly better than 3D-Wiz, while in other applications, the performance of 3D-Wiz dominates that for HMC. On average, 3D-Wiz yields about 1.2% greater performance over HMC. As discussed in Section II and in [9], the fine-grained rank-level 3D partitioning of the data array in HMC better utilizes potential TSV bandwidth compared to the coarse-grained rank-level partitioning used in 3DSPDRAM and 3DSAMS. Due to this reason, HMC has an edge over 3DSPDRAM and 3DSAMS, which translates into a performance edge for HMC over 3DSPDRAM and 3DSAMS. We have used *ch:rank:row:col:bank* address mapping scheme for HMC, whereas for 3D-Wiz and all other DRAM architectures we have

used *ch:rank:bank:col:row* address mapping scheme. The *ch:rank:bank:col:row* scheme results in a larger queueing latency in 3D-Wiz and other DRAMs compared to the *ch:rank:row:col:bank* scheme in HMC for benchmarks A, B, H, I and K, as this benchmarks have coarse/medium granularity parallelism. Therefore, HMC has better performance than 3D-Wiz and other DRAMs for these benchmarks. The reason behind the greater performance of 3D-Wiz over other DRAM designs is the reduced RC loading of access path in 3D-Wiz, which is a result of the smaller subarrays and elimination of global lines.

Fig. 7 shows the energy-delay product (EDP) values (calculated in mJ.µs) for different DRAM designs normalized to the EDP value of 3DSPDRAM. 3D-Wiz yields about 77.1% less EDP value on average over all the other 3D DRAM architectures. More specifically, 3D-Wiz yields about 98.4%, 84.5%, 79.5% and 39.2% less EDP values on average over DDR3, 3DSPDRAM, 3DSAMS and HMC respectively. These improvements in EDP for 3D-Wiz follow directly from the power and performance improvements that were discussed earlier.

VI. Conclusions

This paper introduced 3D-Wiz, a novel high bandwidth and low latency 3D DRAM architecture. 3D-Wiz integrates sub-bank level 3D partitioning of the data array to enable fine-grained activation and greater memory parallelism than other 3D DRAM architectures. The 3D vertical routing of the internal memory bus using TSVs and fan-out buffers enable 3D-Wiz to use smaller dimension subarrays without significant area overhead. This in turn reduces the random access latency and activation-precharge energy. Consequently, 3D-Wiz yields on average 81.1%, 38.8% and 77.1% improvements in power consumption, performance and energy-delay product (EDP) respectively over other 3D DRAM architectures.

The significant improvements demonstrated by 3D-Wiz position it as a promising architecture for future DRAMs. The performance of the 3D-Wiz memory system can be further improved by using intelligent scheduling schemes and novel memory controller designs so that the greater parallelism of 3D-Wiz architecture can be better exploited. Furthermore, the queuing latency of the read and write buffers can be improved by choosing the appropriate size of the buffers. The intelligent selection of buffer size can also help curb the leakage power of the buffers. Moreover, the capacity of the 3D-Wiz DRAM module can be greatly scaled by using intelligent arbitration techniques for the photonic bus. Thus, with potential opportunities for further improvements, the 3D-Wiz architecture can become an even more promising solution for future DRAMs.

ACKNOWLEDGMENTS

This research is supported by grants from SRC, NSF (CCF-1252500, CCF-1302693), and AFOSR (FA9550-13-1-0110).

REFERENCES

[1] P. Kogge and K. Bergman, "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems," 2008.

[2] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas and K. Yelick, "A Case for Intelligent RAM," in *IEEE Micro*, 1997.

[3] J. Mukundan, H. Hunter, K.-h. Kim, J. Stuecheli and J. F. Martinez, "Understanding and Mitigating Refresh Overheads in High-Density DDR4 DRAM Systems," in *ISCA'13*, 2013.

[4] G. H. Loh, "3D-Stacked Memory Architectures for Multi-core Processors," in *ISCA'35*, 2008.

[5] D. H. Woo, N. H. Seong, D. L. Lewia and H. S. Lee, "An Optimized 3D-Stacked DRAM Architecture by Exploiting Excessive, High-Density TSV Bandwidth," in *HPCA'16*, Jan 2010.

[6] U. Kang, H.-J. Chung, H. Seongmoo, S.-H. Ahn, H. Lee, S.-H. Cha, J. Ahn, D. Kwon, J. H. Kim, J.-W. Lee, H.-S. Joo, W.-S. Kim, H.-K. Kim, E.-M. Lee, S.-R. Kim, K.-H. Ma, D.-H. Jang, N.-S. Kim, M.-S. Choi, S.-J. Oh, J.-B. Lee, T.-K. Jung, J.-H. Yoo and C. Kim, "8 Gb 3D DDR3 DRAM Using Through-Silicon-Via Technology," in *JSSC*, 2010.

[7] J. T. Pawlowski, "Hybrid Memory Cude (HMC)," in *in proceedings of Hot Chips 23*, 2011.

[8] Y. H. Son, O. Seongil, R. Yuhwan, H. W. Lee and J. H. Ahn, "Reducing Memory Access Latency with Asymmetric DRAM Bank Organizations," in *ISCA'13*.

[9] K. Chen, S. Li, N. Muralimanohar, J. H. Ahn, J. B. Brockman and N. P. Jouppi, "CACTI-3DD: Architecture-level Modeling for 3D Die-stacked DRAM Main Memory," in *DATE*, 2012.

[10] A. Hadke, T. Benavides, S. J. Ben Yoo, R. Amirtharajah and V. Akella, "OCDIMM: Scaling the DRAM Memory Wall Using WDM based Optical Interconnects," in *HOTI*, 2008.

[11] A. N. Udipi, N. Muralimanohar, R. Ralsubramonian and P. N. Jouppi, "Combining Memory and a controller with Photonics through 3D-Stacking to Enable Scalable and Energy-Efficient Systems," in *ISCA'11*, 2011.

[12] K. Bergman, L. P. Carloni, A. Biberman, J. Chan and G. Hendry, Photonic Network-On-Chip Design, Springer, May 2013.

[13] C. Zhong, C. Liu and F. Zhong, "25Gbps SerDes," IEEE HSSG, Orlando FL, 2007.

[14] Semiconductor Industries Association, "International Technology Roadmap for Semiconductors".

[15] P. Rosenfeld, E. Cooper-Balis and B. Jacob, "DRAMSim2: A Cycle Accurate Memory System Simulator," *IEEE COMPUTER ARCHITECTURE LETTERS,* 2011.

[16] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill and D. A. Wood, "The gem5 Simulator," in *Computer Architecture News*, May 2011.

[17] C. Bienia, S. Kumar, J. P. Singh and K. Li, "The PARSEC Benchmark Suit: Characterization and Architectural Implications," in *PACT*, Oct 2008.

[18] S. Beamer et al., "Re-Architecting DRAM Memory Systems with Monolithically Integrated Silicon Photonics," in *ISCA'10*, June 2010.

[19] B. Jacob, S. W. NG and D. T. Wang, Memory Systems: Cache, DRAM, Disk, Morgan Kaufmann, 2007.

[20] T. Pimpalkhute and S. Pasricha, "NoC Scheduling for Improved Application-Aware and Memory-Aware Transfers in Multi-Core Systems," in *IEEE International Conference on VLSI Design (VLSID)*, Jan 2014.

[21] S. Bahirat and S. Pasricha, "OPAL: A Multi-Layer Hybrid Photonic NoC for 3D ICs," in *IEEE/ACM Asia & South Pacific Design Automation Conference (ASPDAC 2011)*, Yokohama, Japan, Jan 2011.

[22] S. Pasricha and N. Dutt, "Trends in Emerging On-Chip Interconnect Technologies," *IPSJ Transactions on System LSI Design Methodology,* vol. 1, 2008.