3-D WiRED: A Novel WIDE I/O DRAM With Energy-Efficient 3-D Bank Organization

Ishan Thakkar and Sudeep Pasricha
Colorado State University

Editor's notes:
WIDE I/O DRAM is a promising 3-D memory architecture for low-power/high-performance computing. This paper proposes a new WIDE I/O DRAM architecture to reduce access latency and energy consumption at the same time, which shows the possibility of further optimization of the WIDE I/O DRAM architecture and the impact of TSV usage in the memory architecture on the performance and energy consumption.
—Sung Kyu Lim, Georgia Institute of Technology

To fulfill the increasing demand of high-bandwidth and low-power memory solutions for embedded system-on-chips (SoCs) in demand today, the Joint Electron Device Engineering Council (JEDEC) proposed a new standard for wide input/output (I/O) DRAMs [1]. WIDE I/O DRAM follows the low-power design methodology of LPDDR2 to achieve lower static power and I/O power dissipation compared to DDR3. In contrast to LPDDR2, the wider I/O of wide I/O DRAM increases peak bandwidth, while its low-capacitance through-silicon-via (TSV)-based interconnects further reduce I/O power.

However, the first prototype of WIDE I/O DRAM demonstrated by Samsung [2] does not take full advantage of TSV-based die stacking as it strongly complies with a conventional 2-D DRAM bank organization having a small bank count and a long intrabank memory access path. Due to this, the power consumption of 3-D-stacked WIDE I/O DRAM core significantly increases with increasing memory capacity, which makes it highly energy inefficient for the green and low-power embedded computing solutions of the future. Moreover, the performance and random access bandwidth of WIDE I/O DRAM are also limited by its conventional 2-D bank organization. These limitations make it imperative to reinvent the bank organization of WIDE I/O DRAM, to improve performance with greater energy efficiency, while limiting area/cost overheads.

Zhang et al. [3] reorganized the WIDE I/O DRAM core in their proposed 3-D SWIFT architecture, which employs a large number of small banks to enable greater bank-level parallelism. But, they over-optimistically eliminate the power constraint and set the two-bank activation window (tTAW) to be zero in order to increase memory performance manifold, which is not practical. Several other 3-D-stacked DRAM architectures have been proposed in recent years [4]–[6] that aim to improve DRAM bank organization to achieve greater performance and throughput for DRAM cores. Unfortunately, these proposals do not efficiently enhance the DRAM core, as they either optimize only a few elements of the

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.
Digital Object Identifier 10.1109/MDAT.2015.2440411
Date of publication: 2 June 2015; date of current version: 30 June 2015.

bank organization or they entirely reinvent it with significant area and cost overheads. Such 3-D-stacked DRAMs require new efficient bank organization, to realize their performance potential while keeping area/cost overheads low.

This paper presents a novel WIDE I/O DRAM architecture called 3-D WiRED, with an enhanced DRAM core to enable low latency and energy-efficient memory access. Through detailed time-energy analysis of a WIDE I/O DRAM prototype [2], we have identified the need to reduce the capacitance of bitlines (BLs), memory bus (M_bus), and global data path to reduce random access latency, read/write energy, and activation-precharge (ActPre) energy. We reorganize DRAM banks and utilize a TSV-based internal M_bus to achieve reduced capacitance of the data access path with increased area efficiency. Our experimental analysis of 3-D WiRED demonstrates improved average latency and energy efficiency over state-of-the-art WIDE I/O and 3-D SWIFT DRAM architectures. Our key contributions with 3-D WiRED are summarized as follows.

- We present detailed breakdowns of timing and energy for the prototype WIDE I/O DRAM reported in [2], through which we identify the key components of DRAM organization that most significantly affect overall latency and energy of the DRAM subsystem.
- We model and study two variants of the state-of-the-art WIDE I/O and 3-D SWIFT DRAMs, to derive an optimum combination of critical enhancements to make in WIDE I/O bank organization that would achieve combined benefits in performance, energy efficiency, cost, and area.
- We employ large-aspect-ratio subarrays to reduce BL capacitance, as proposed in [7], but with better area efficiency. Reduced BL capacitance reduces row cycle time and ActPre energy which relaxes the power constraint and increases bank-level parallelism.
- We reorganize 3-D WiRED banks by using a TSV-based internal M_bus to eliminate global wordlines (GWLs), columnlines, and datalines, which reduces access time and read/write energy.
- We experimentally compare and contrast our 3-D WiRED architecture with state-of-the-art WIDE I/O and 3-D SWIFT DRAM architectures, for several PARSEC benchmarks.

Background and related work

In this section, we identify the key elements of WIDE I/O DRAM organization that need to be improved in order to enhance DRAM energy efficiency.

As discussed earlier, it is imperative to reinvent the bank organization of WIDE I/O DRAM to overcome the shortcomings of its conventional 2-D bank structure. To address this issue, Zhang et al. [3] proposed 3-D SWIFT, which increases bank-level parallelism by eliminating the tTAW power constraint and employing a large number of small banks. They divide each bank of 3-D SWIFT into 16 smaller banks and assume that doing so would eliminate the power constraint ($tTAW = 0$). In general, the tTAW power constraint allows only two bank activates in a rolling window of tTAW time. So, relaxing the power constraint would obviously increase bank-level parallelism by allowing more bank activates in a given time window. But, eliminating the power constraint would allow an infinite number of activates in a given time window, which is impractical and unreal. Thus, more accurate estimation of the power constraint is required to realistically evaluate the performance benefits achieved from the increased bank-level parallelism of 3-D SWIFT.

The performance and energy efficiency achievable by 3-D SWIFT and WIDE I/O DRAMs can be further improved by enhancing the critical structures inside the banks of these DRAMs. To identify the critical structures out of all the structures of WIDE I/O DRAM organization, it is important to quantify the contributions of all the structures to the overall latency and energy of the memory subsystem. Therefore, we model a 4-stacked version of WIDE I/O DRAM subsystem reported in [2] by adapting the source code of CACTI-3DD [8] for 50-nm technology and $8F^2$ DRAM cell layout. The WIDE I/O DRAM has four independent channels, which are identical in design and operation. More details on the simulation parameters are given in Table 1 and in 3-D WiRED area, energy, and timing analysis. We present the detailed breakdowns of the overall latency and energy of one channel of this wide I/O DRAM model (referred to as wide IO×1) in Figure 1. Figure 1 also presents the detailed energy and delay breakdowns for two variants of the state-of-the-art 3-D SWIFT DRAM architecture and our 3-D WiRED architecture, which are discussed in the 3-D WiRED DRAM architecture section with more details.

As shown in Figure 1a, about 90.12% of the total ActPre energy for wide IO×1 is consumed in the

Table 1 Timing, energy, and area values for various DRAM architectures.

	Wide-I/O		3D-SWIFT		3D-WiRED	
	ARx1	ARx4	ARx1	ARx4		
Frequency (MHz)	800	800	800	800	800	
Page size (Kb)	8192	8192	8192	8192	8192	
Bandwidth (Gbps)	21.3	20	34.8	54.3	56	
# bank/rank	4	4	128	128	128	
Background Leakage Power Breakdown (mW)						
RAS globallines+bus leakage (mW)	1.1968	0.5865	1.1968	0.5865	0.4471	
CAS globallines+bus leakage (mW)	0.0346	0.0346	0.0346	0.0346	0.0346	
RAS peripherals leakage (mW)	0.4220	0.21	0.4220	0.21	0.1457	
CAS peripherals leakage (mW)	0.0135	0.0135	0.0135	0.0135	0.0135	
TSV + driver leakage (mW)	0.0013	0.0013	0.1289	0.1289	1.3574	
Clock power (mW)	6.92	6.92	6.92	6.92	6.92	
PHY leakeage (mW)	1	1	1	1	1	
I/O termination and bias (mW)	2.23	2.23	2.23	2.23	2.23	
Total background leakage power (mW)	*11.8181*	*10.9959*	*11.9469*	*11.1246*	*12.152*	
Refresh Power Breakdown (mW)						
Local wordline (mW)	0.26	0.21	0.31	0.26	0.07	
Row decoder (mW)	0.32	0.27	0.34	0.30	0.07	
Global wordline (mW)	0.74	0.35	0.88	0.44	0.00	
TSV (mW)	1.34	1.28	2.19	2.17	0.85	
Sense amplifier (mW)	6.32	6.04	7.53	7.46	8.33	
RAS bus (mW)	7.49	9.00	1.71	1.90	3.02	
Bitlines (mW)	150.26	76.21	179.07	94.04	105.08	
Total refresh power (mW)	*166.73*	*93.38*	*192.03*	*106.56*	*117.43*	
DRAM Energy and Timing Parameters						
Activation-Precharge energy per access (nJ)	1.813	1.062	1.752	0.982	0.969	
Read/Write energy per access (nJ)	3.959	5.216	2.083	2.724	2.432	
tRCD (ns)	12.5	14.1	11.1	12.4	10.2	
tCAS (ns)	13.5	15.4	7.8	8.5	8.1	
tRAS (ns)	32.2	33.7	25.6	25.6	23.02	
tRC (ns)	41.8	43.7	35.2	35.6	33.2	
tTAW (ns)	50	30	48	27	27	
tRFC (ns)	348	364	292	295	264	
DRAM Die Area Breakdown (mm^2)						
Total die area (mm^2)	35.1232	37.6418	52.3761	55.4899	48.05	
Area efficiency (%)	53.82 %	50.22 %	36.09 %	34.06 %	39.34% + 8.79 mm^2	
Row pre-decoder + decoder area (mm^2)	0.122	0.122	0.1216	0.1216	0.4876	Logic & I/O die area for 3D-WiRED
Column pre-decoder + decoder area (mm^2)	0.0003	0.0007	0.0028	0.0028	0.0111	
Address bus area (mm^2)	0.1208	0.1352	0.8064	0.8448	1.1211	
Data bus area (mm^2)	2.2052	2.4716	17.792	18.88	6.2390	
Center stripe area (mm^2)	1.3516	1.6988	0.6784	0.7680	0.9246	
DRAM cell area (mm^2)	18.9024	18.9024	18.9024	18.9024	18.9024	
Local wordline driver area (mm^2)	8.3858	4.1908	8.3858	4.1908	4.1908	
Sense amplifier area (mm^2)	4.0192	10.1010	4.0192	10.1010	10.1010	
Data driver area (mm^2)	0.0164	0.0164	0.5248	0.5248	2.0992	
TSV area (mm^2)	0.0031	0.0031	1.137	1.137	13.6396	
TSV Properties						
Load capacitance per address TSV (fF)			0.167fF		0.112fF	
Load component per address TSV per layer			One input of the decoder circuit		Inputs of two WL drivers	
Delay per TSV (ns)	0.2546	0.2546	0.2546	0.2546	0.2483	
Energy per TSV (pJ)	0.2337	0.2337	0.2337	0.2337	0.2326	
# TSV tiers	3	3	4	4	4	
# Row address TSVs per rank	15	15	40	40	131072	
# Column address TSVs per rank	13	13	52	52	32768	
# Data TSVs per rank	128	128	16384	16384	16384	

TSV Resistance	TSV Length	TSV Pitch	TSV Diameter	TSV Capacitance	Area per TSV
734.42 mOhm	8 µm	4 µm	2 µm	154.57 fF	12.56 $µm^2$

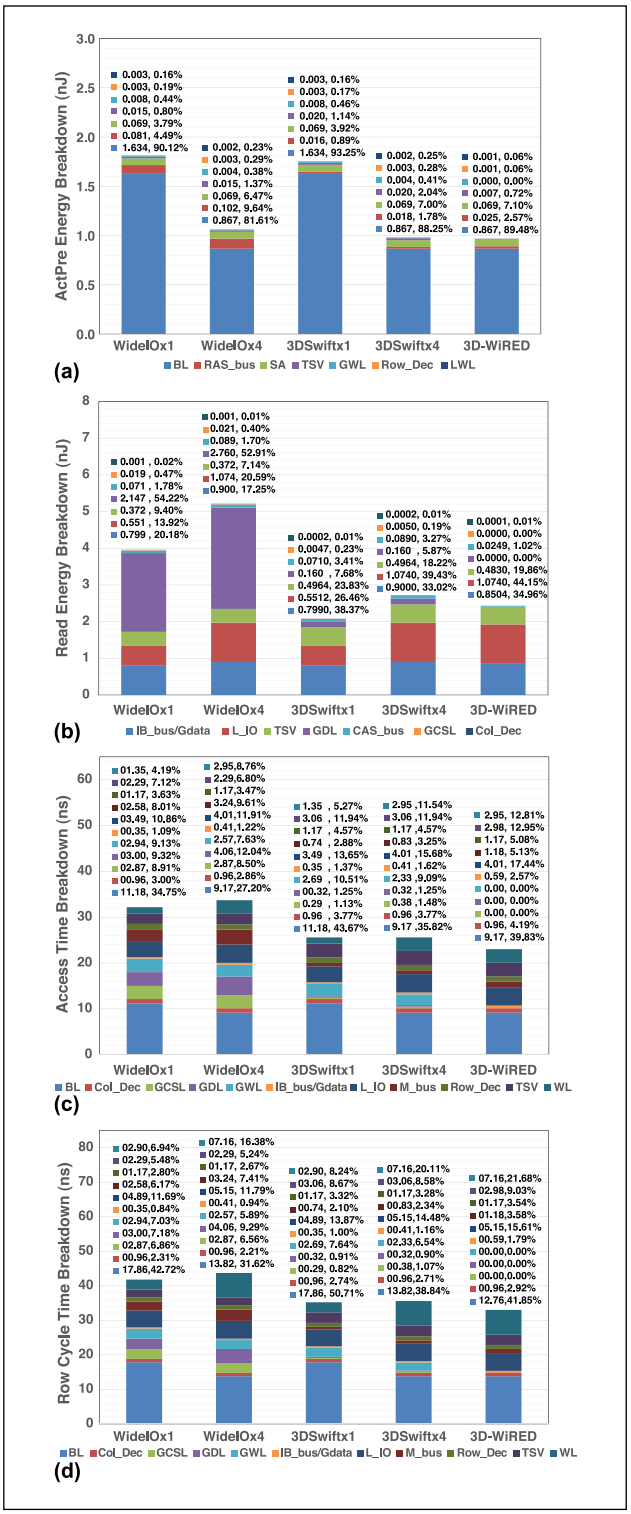

Figure 1. Energy and timing for various DRAM architectures: (a) ActPre energy breakdown; (b) read/write energy breakdown; (c) access-time breakdown; and (d) row-cycle time breakdown. (IB_bus is replaced by Gdata in case of 3-D WiRED.)

BLs. Therefore, it is critical to reduce BL capacitance (which in turn will reduce BL energy) if the total ActPre energy is to be minimized. Figure 1b shows that global datalines (GDL), intrabank data bus (IB_bus), and local I/O (L_IO) are critical components for read energy. From the timing results in Figure 1c and d, it is evident that BLs, GDLs, global column select lines (GCSLs), and GWLs are critical components for access time (Figure 1c), and wordlines (WLs), BLs, GDLs, GCSLs, GWLs, and M_bus are critical components for row cycle time (Figure 1d). M_bus delay is sum of row access bus (RAS_bus) and column access bus (CAS_bus) delays. While the decoder (Row_Dec/Col_Dec) also contributes notably to the timings, further reducing decoder delay is very difficult, and, therefore, we ignore this component during our bank optimization.

In summary, the BLs, M_bus, and global lines (GDLs, GWLs, and GCSLs) are the most critical components that most significantly affect the overall latency and energy of the memory subsystem. So, it is important to reduce the energy and delay of these critical components if the overall latency and energy consumption of the memory subsystem is to be minimized.

A conventional DRAM device has 512 WLs and 512 BLs per subarray. Some recent work on DRAM architectures [4], [7] reduces the energy and delay of BLs by reducing the number of WLs per DRAM subarray. However, the reduction in WLs per subarray is done without proportionally increasing the number of BLs per subarray, which significantly increases area overhead of WL drivers and sense amps for a given memory capacity, harming the area efficiency of the DRAM die. Instead, in 3-D WiRED, we reduce the number of WLs per subarray by a factor of two, while increasing the number of BLs per subarray by the same factor yielding better area efficiency compared to DRAM architectures from prior work [4], [7].

A few 3-D DRAM architectures [4]–[6] reduce the length and capacitance of the M_bus and global lines to benefit overall DRAM latency and energy. Loh et al. [5] and Micron's hybrid memory cube (HMC) [6] employ TSVs to partition the rank across multiple layers in order to reduce the length of interbank M_bus and global lines, which results in reduced latency and energy consumption. Thakkar and Pasricha [4] improve upon these DRAM architectures and propose the 3-D-Wiz 3-D DRAM

architecture that eliminates global lines by employing aggressive vertical routing of intrabank buses using TSVs and fanout buffers. However, the use of fanout buffers and TSVs at subarray-level granularity requires costly changes in local peripheral circuits. Also, use of fanout buffers incurs significant delay, area, and cost overhead. Moreover, these existing 3-D DRAM architectures enhance only a single critical component or a random combination of critical components in a DRAM subsystem, which is suboptimal. Thus, the 3-D DRAM architectures from prior work fail to achieve combined benefits in performance, energy efficiency, area, and cost, which is vital for emerging high-performance and low-power embedded memory solutions. By considering an optimal combination of critical components to be enhanced in the DRAM subsystem, our 3-D WiRED architecture is able to overcome the drawback of state-of-the-art DRAM architectures, as discussed next.

Three-dimensional WiRED DRAM architecture

This section describes our proposed 3-D WiRED architecture in detail. First, the bank organization of the existing WIDE I/O and 3-D SWIFT DRAM architectures is discussed in the Bank organization of wide I/O and 3-D-swift DRAMs section. The Bank organization of 3-D WiRED DRAM section describes the bank organization of our 3-D WiRED DRAM architecture. Finally, the area, energy, and timing analysis of 3-D WiRED DRAM and the variants of WIDE I/O and 3-D SWIFT DRAMs are presented in the 3-D WiRED area, energy, and timing analysis section.

Bank organization of WIDE I/O and 3-D SWIFT DRAMs

In this section, we describe two variants of the wide I/O and 3-D SWIFT DRAM architectures each.

We consider the 4-stacked WIDE I/O DRAM architecture mentioned in the previous section as the baseline model. This baseline WIDE I/O DRAM architecture consists of a stack of four 4-GB DRAM dies. This 16-GB stack of four DRAM dies constitutes four independent channels of 4-GB size each. Figure 2a shows a schematic of a WIDE I/O DRAM channel. The channel consists of four ranks with each rank having four identical banks. A rank is defined as a portion of memory from one memory die, which logically corresponds to a single channel within the memory stack. Figure 2a shows the WIDE I/O bank structure for two variants identified as AR×1 and AR×4. Each subarray of variant AR×1 of WIDE I/O DRAM has 512 WLs and 512 BLs, which corresponds to a subarray aspect ratio (AR) of one. A 16×64 array of these 512×512 subarrays makes up one bank of variant AR×1. On the other hand, each subarray of variant AR×4 of WIDE I/O DRAM has 256 WLs and 1024 BLs, which corresponds to a subarray AR of four. An 8×128 array of these 256×1024 subarrays makes up one bank of variant AR×4. Each bank of both variants of WIDE I/O DRAM has 32 768 rows and 64 columns with columnwidth of 128 bits. As shown in Figure 2a, global peripherals (row/column decoders, global lines, and their drivers) occupy a notable amount of area within a WIDE I/O rank.

Figure 2b shows a schematic of the 3-D SWIFT DRAM channel [3]. As shown in the figure, the channel consists of four ranks, each of which is partitioned across four DRAM dies. The decoder and control logic of 3-D SWIFT DRAM channel are relocated on a separate logic die, which decreases total DRAM die area resulting in better area efficiency. As discussed in [3], 3-D SWIFT divides each rank in 128 identical banks, which increases bank-level parallelism provided the power constraint is significantly reduced. Figure 2b depicts the 3-D SWIFT bank structure for two variants: AR×1 and AR×4. A 16×2 array of 512×512 subarrays makes up one bank of variant AR×1. On the other hand, an 8×4 array of 256×1024 subarrays makes up one bank of variant AR×4. Each bank of both variants of 3-D SWIFT DRAM has 1024 rows and 64 columns with a columnwidth of 128 bits.

The variant AR×1 of 3-D SWIFT has shorter row address bus (RAS_bus), shorter column address bus (CAS_bus), shorter global lines (GCSLs and GDLs), and larger bank count compared to the variant AR×1 of WIDE I/O. The variants AR×4 of both architectures have shorter BLs compared to the variants AR×1. The detailed analysis of how these architectural variants behave in terms of area, latency, and energy is presented in the 3-D WiRED area, energy, and timing analysis section.

Bank organization of 3-D WiRED DRAM

As evident from the discussion in the previous section, the variant AR×4 of the 3-D SWIFT architecture combines the benefits of shorter address bus, shorter BLs, shorter global lines, and larger bank count. Our 3-D WiRED architecture improves upon

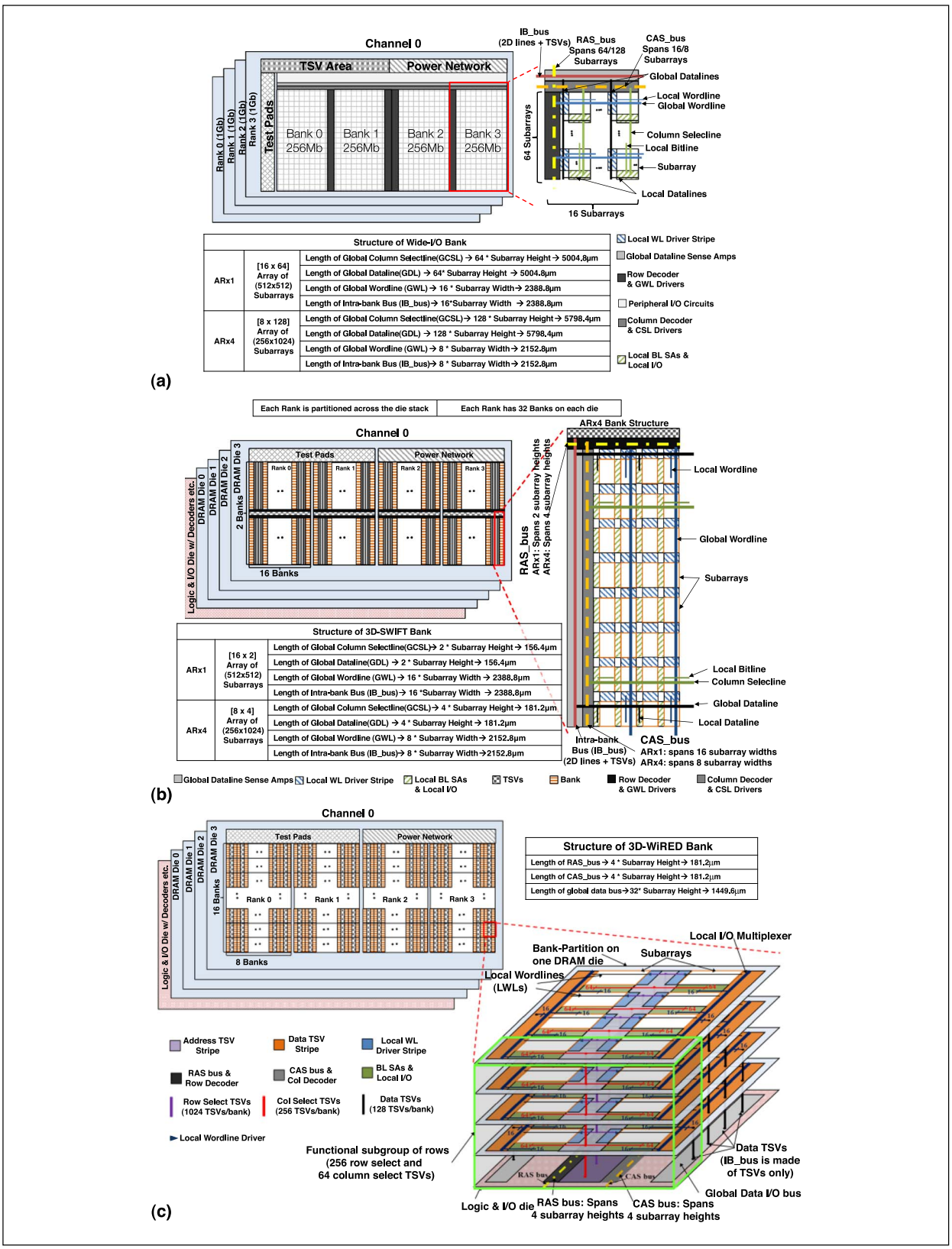

Figure 2. (a) Schematic layout of WIDE I/O DRAM channel. (b) Schematic layout of 3-D SWIFT DRAM channel. (c) Schematic layout of 3-D WiRED DRAM channel.

the variant AR×4 of 3-D SWIFT by eliminating the global lines and reducing the length of the intrabank data bus. Figure 2c shows a schematic of the 3-D WiRED DRAM channel. As shown in the figure, the intrabank data bus (IB_bus), which comprises 2-D data lines (spanning multiple subarrays) and TSVs in 3-D SWIFT and WIDE I/O DRAMs, is implemented with only TSVs in our 3-D WiRED DRAM. It is evident that the length of the TSV-based IB-bus in 3-D WiRED is 32 μm (as one TSV is 8 μm long and the IB_bus traverses through four DRAM layers), which is about 69× shorter than the length of IB_bus (~2.2 mm) in 3-D SWIFT DRAM (see Figure 2b). However, in addition to IB_bus, 3-D WiRED requires a global data bus that connects the global I/O of the DRAM module to the IB_bus. As shown in Figure 2c, each 3-D WiRED rank has a total of 128 banks, each of which is partitioned across four DRAM dies. The bank partition on each die consists of a 2×4 array of 256×1024 subarrays. Each bank in 3-D WiRED has 1024 rows (1024 row select TSVs) and each row has 64 columns with a columnwidth of 128 bits. The rows of a bank are divided into four functional subgroups with 256 rows per subgroup. Each subgroup has eight subarrays, and it is partitioned across four DRAM dies with two subarrays on each die. Each subgroup is functionally independent. For independent operation, each subgroup needs a dedicated set of 64 column select TSVs requiring a total of 256 (4×256) column select TSVs per bank. Unlike 3-D SWIFT, each row of a 3-D WiRED DRAM bank is folded across four DRAM dies. The use of TSVs at subarray level granularity enables direct and efficient connection of the address buses (RAS_bus and CAS_bus) to the local peripherals of individual subarrays, but also significantly increases TSV area overhead for the 3-D WiRED DRAM. However, area benefits obtained due to elimination of global lines and related peripherals along with the vertical routing of IB_bus counterbalances the TSV area overhead, resulting in comparable area efficiency. The results of the area, energy, and timing analysis of 3-D WiRED are presented in the following section.

Three-dimensional WiRED area, energy, and timing analysis

We performed area, timing, and energy analysis by adapting the code for CACTI-3DD [8] to model the 3-D WiRED architecture. A similar analysis was conducted for AR×1 and AR×4 variants of the WIDE I/O and 3-D SWIFT DRAM architectures. The models of these DRAM architectures were implemented in CACTI-3DD using the technology parameters for the 50-nm node. All TSVs in this study were modeled based on International Technology Roadmap for Semiconductors (ITRS) projections for intermediate interconnect level TSVs [9]. The core-to-memory interface for wide I/O, 3-D SWIFT and 3-D WiRED DRAM architectures is based on the JEDEC standardized WIDE I/O interface protocol [1]. The results of this analysis are given in Table 1, along with TSV parameters and breakdowns of refresh power, leakage power, and DRAM die area. We also extracted detailed breakdowns of the per-access energy and latency of the AR×1 and AR×4 variants of the WIDE I/O and 3-D-SWIFT DRAM architectures, and 3-D WiRED architecture, which are given in Figure 1.

It is evident from Figure 1a that 3-D WiRED has a smaller value of ActPre energy compared to other DRAMs, because of its shorter RAS_bus, shorter BLs, and elimination of GWLs. Similarly, Figure 1b shows that 3-D WiRED has a smaller value of read energy compared to 3-D Swift×4, wide IO×1 and wide IO×4 DRAMs, because of elimination of global lines (GDLs and GCSLs). Due to the combined effects of shorter GDLs and shorter local I/O lines, 3-D Swift×1 has smaller read energy than 3-D WiRED. In spite of having a very large number of address TSVs, 3-D WiRED DRAM (as shown in Figure 1a and b and in Table 1 for the refresh power breakdown) does not incur very high energy overhead of TSVs. This is because only one pair of address TSVs (one column address TSV plus one row address TSV) is activated during a DRAM read/write operation, and, therefore, only one pair of address TSVs contribute to the per-access values of ActPre energy and read energy. Moreover, as explained in [8], the coupling capacitance of TSV does not significantly increase with increasing TSV density, keeping the total capacitance of TSV (intrinsic capacitance plus coupling capacitance) unchanged with increase in TSV density. This means that TSV delay, which is a function of TSV capacitance, remains unchanged with increase in TSV density. Hence, the delay of TSVs for 3-D WiRED is the same as the TSV delay for other DRAM architectures. Besides, Tezzaron's high-density TSV technology enables integration of millions of TSVs in a single 3-D-stacked DRAM module [4], [13], which

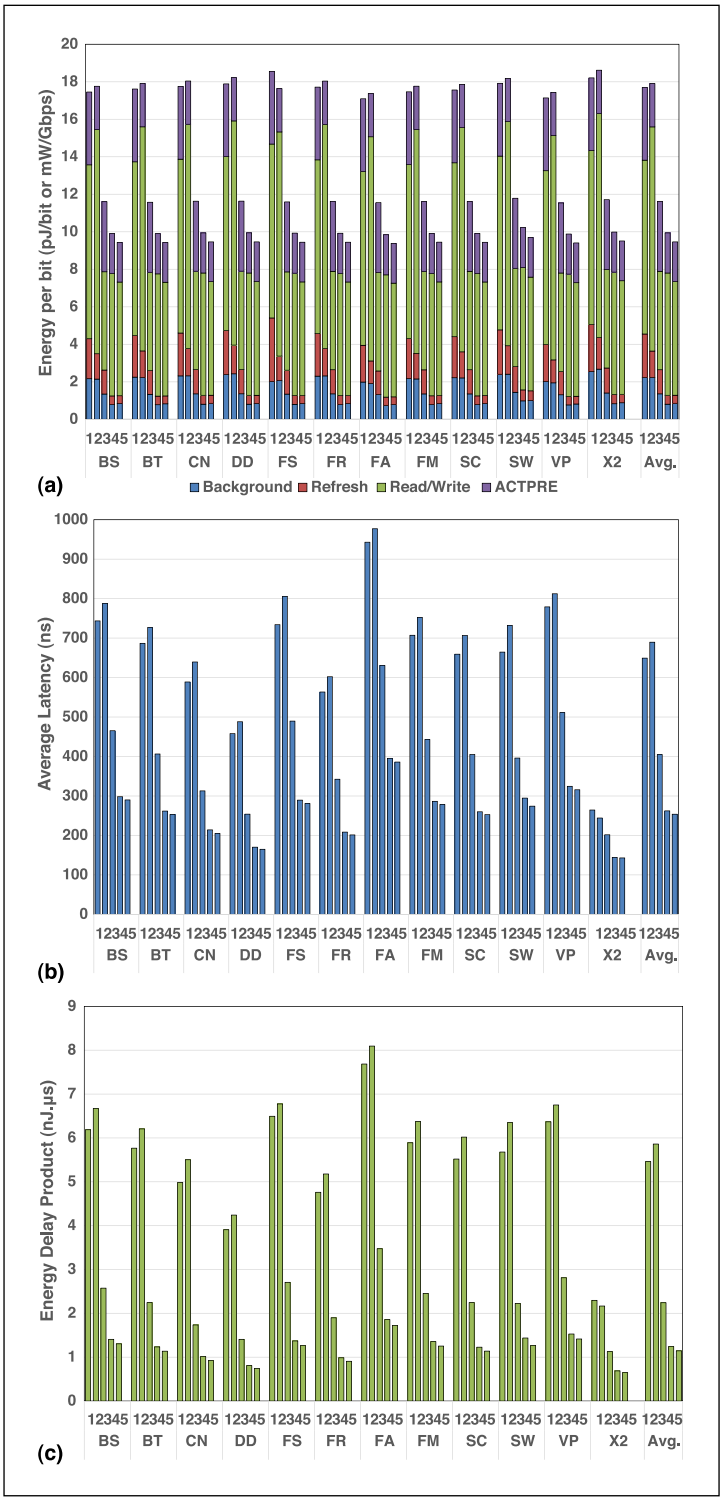

Figure 3. DRAM comparison results for PARSEC benchmarks: (a) energy-per-bit values. (b) average-latency values; (c) energy-delay product values. (1: wideIO-AR×1; 2: wideIO-AR×4; 3: 3-D-SWIFT-AR×1; 4: 3-D-SWIFT-AR×4; 5: 3-D WiRED.)

makes it possible to implement a very large number of TSVs in case of 3-D WiRED DRAM. Finally, due to the combined effect of shorter BLs, shorter IB_bus, and elimination of global lines (GWLs, GCSLs, and GDLs), our proposed 3-D WiRED DRAM architecture achieves smaller values of access time and row cycle time than all the other DRAM architectures (Figure 1c and d).

JEDEC has standardized the tTAW time for WIDE I/O DRAM to be 50 ns. Therefore, we assume the tTAW time for our baseline variant AR×1 of WIDE I/O DRAM to be 50 ns. 3-D WiRED DRAM has about 46% less ActPre energy than the baseline variant AR×1 of WIDE I/O DRAM. This translates into a relaxed tTAW constraint of 27 ns, which is about 54% of the tTAW for the AR×1 variant of WIDE I/O DRAM.

Evaluation results

We performed trace-driven simulation analysis to compare 3-D WiRED with other state-of-the-art DRAM architectures. Memory access traces for the PARSEC benchmark suite [10] were extracted from detailed cycle-accurate simulations using gem5 [11]. We considered 12 different applications from the PARSEC benchmark suite: Blackscholes (BS), Bodytrack (BT), Canneal (CN), Dedup (DD), Facesim (FS), Ferret (FR), Fluidanimate (FA), Freqmine (FM), Streamcluster (SC), Swaptions (SW), Vips (VP), and ×264 (X2). We ran each PARSEC benchmark for a "warm-up" period of one billion instructions and captured memory access traces from the subsequent one billion instructions extracted. These memory traces were then provided as inputs to the DRAM simulator DRAMSim2 [12], which we modified heavily to model 3-D WiRED and the other DRAM architectures.

A rank-based round-robin scheduling scheme and a closed page policy were used for all simulations. Energy-per-bit, average-latency, and energy-delay product values for the memory subsystem were obtained from DRAMSim2. Figure 3a shows energy-per-bit values for the various DRAM architectures across the PARSEC benchmarks. The energy per bit was calculated by dividing the average power (in milliwatts) by the throughput (in gigabits per second). The figure gives the total energy per bit which is a sum of background energy-per-bit, refresh energy-per-bit, ActPre energy-per-bit (ACTPRE), and read/write energy-per-bit values. It can be observed that 3-D WiRED consumes about 33.8% less energy

per bit on average, compared to all the other DRAM architectures. More specifically, 3-D WiRED consumes approximately 46.5%, 47.1%, 18.6%, and 4.8% less energy per bit on average over the wide IO-AR×1, wideIO-AR×4, 3-D-SWIFT-AR×1, and 3-D-SWIFT-AR×4 architectures, respectively. The reason for the lower energy consumption in 3-D WiRED is due to the smaller values of per-access ActPre energy, read/write energy, and relatively high throughput, the effect of which cumulates to minimize energy per bit.

Figure 3b shows average-latency values for the various DRAM architectures across the PARSEC benchmarks. The average latency was calculated by dividing the total latency with a total number of transactions. Despite having the same access time and row cycle time, 3-D-SWIFT-AR×4 yields lower average latency than 3-D-SWIFT-AR×1. This is due to the reduced tTAW time constraint for 3-D-SWIFT-AR×4, which enables increased bank-level parallelism. Similarly, wideIO-AR×4 also has larger access time and smaller tTAW time than wideIO-AR×1, but it does not translate into lower average latency for wide IO-AR×4, because smaller bank count for wide IO-AR×4 does not allow the reduced tTAW constraint to increase bank-level parallelism. It can be observed that 3-D WiRED yields about 49.4% less average latency on average over all the other DRAM architectures. More specifically, 3-D WiRED yields 60.9%, 63.2%, 37.3%, and 3.2% less average latency on average over wideIO-AR×1, wideIO-AR×4, 3-D-SWIFT-AR×1, and 3-D-SWIFT-AR×4, respectively.

Last, Figure 3c shows energy-delay product (EDP) values for the various DRAM architectures across the PARSEC benchmarks. It can be observed that 3-D WiRED yields about 69.1% less EDP on average over all the other DRAM architectures. More specifically, 3-D WiRED yields about 79.1%, 80.5%, 49%, and 7.9% less EDP on average over the wideIO-AR×1, wide IO-AR×4, 3-D-SWIFT-AR×1, and 3-D-SWIFT-AR×4 architectures, respectively. These improvements in EDP for 3-D WiRED follow directly from the energy-per-bit and average-latency improvements that were discussed earlier.

We proposed the novel 3-D WiRED DRAM architecture that yields on average 31.2%, 32.9%, and 52.8% improvements in energy per bit, average latency, and EDP over state-of-the-art wide I/O and 3-D SWIFT DRAM architectures. These promising results indicate that the performance and energy efficiency of contemporary wide I/O DRAM core can be greatly improved by aggressively using TSVs at subarray-level granularity. However, several challenges still need to be overcome to support such a fine-grained 3-D integration. In particular, 3-D-stacking technology suffers from low yield and thermal/noise issues which can affect the productization of these systems. As 3-D-stacking technology matures, architectures such as the one proposed in this work can greatly improve density, performance, and energy efficiency of DRAM cores. ■

Acknowledgment

This work was supported by the Semiconductor Research Corporation (SRC), National Science Foundation (NSF) under Grants CCF-1252500 and CCF-1302693 and by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-13-1-0110.

■ References

[1] *Wide I/O Single Data Rate*, Standard JESD229, JEDEC Solid State Technology Association, 2011.

[2] J. S. Kim et al., "A 1.2 V 12.8 GB/s 2 Gb mobile wide-I/O DRAM with 4 128 I/Os using TSV based stacking," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 107–116, Jan. 2012.

[3] T. Zhang, C. Xu, K. Chen, G. Sun, and Y. Xie, "3D-SWIFT: A high-performance 3D-stacked wide IO DRAM," in *Proc. 24th Ed. Great Lakes Symp. VLSI*, 2014, pp. 51–56.

[4] I. Thakkar and S. Pasricha, "3D-Wiz: A novel high bandwidth, optically interfaced 3D DRAM architecture with reduced random access time," in *Proc. IEEE Int. Conf. Comput. Design*, 2014, DOI: 10.1109/ICCD.2014.6974654.

[5] G. H. Loh, "3D-stacked memory architectures for multi-core processors," in *Proc. Int. Symp. Comput. Architect.*, 2008, pp. 453–464.

[6] J. T. Pawlowski, "Hybrid memory cude (HMC)," in *Proc. Hot Chips*, 2011. [Online]. Available: http://www.hotchips.org/wp-content/uploads/hc_archives/hc23/HC23.18.3-memory-FPGA/HC23.18.320-HybridCube-Pawlowski-Micron.pdf

[7] Y. H. Son, O. Seongil, Y. Ro, J. W. Lee, and J. H. Ahn, "Reducing memory access latency with asymmetric DRAM bank organizations," in *Proc. 40th Annu. Int. Symp. Comput. Architect.*, 2013, pp. 380–391.

[8] K. Chen et al., "CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory," in *Proc. Design Autom. Test Eur.*, 2012, pp. 33–38.

[9] *International Technology Roadmap for Semiconductors*, Semiconductor Industries Association, 2011.

[10] C. Bienia et al., "The PARSEC benchmark suit: Characterization and architectural implications," in *Proc. 17th Int. Conf. Parallel Architect. Compilat.*, 2008, pp. 72–81.

[11] N. Binkert et al., "The gem5 simulator," *Comput. Architect. News*, vol. 39, pp. 1–7, 2011, DOI: 10.1145/2024716.2024718.

[12] P. Rosenfeld, E. Cooper-Balis, and B. Jacob, "DRAMSim2: A cycle accurate memory system simulator," *IEEE Comput. Architect. Lett.*, vol. 10, no. 1, pp. 16–19, Jan.–Jun. 2011.

[13] B. Giridhar et al., "Exploring DRAM organizations for energy-efficient and resilient exascale memories," in *Proc. Int. Conf. High Performance Comput. Netw. Storage Anal.*, 2013, DOI: 10.1145/2503210.2503215.

Ishan Thakkar is currently working toward a PhD in electrical engineering at Colorado State University, Fort Collins, CO, USA. His research interests include 3-D DRAM architectures, nonvolatile memories, and high-speed optical interfaces. Thakkar has an MS in electrical engineering from Colorado State University. He is a Student Member of the IEEE.

Sudeep Pasricha is an Associate Professor at the Electrical and Computer Engineering Department, Colorado State University, Fort Collins, CO, USA. His research interests include embedded, mobile, and high-performance computing. Pasricha has a PhD in computer science from the University of California Irvine, Irvine, CA, USA (2008). He is a Senior Member of the IEEE and the Association for Computing Machinery (ACM).

■ Direct questions and comments about this article to Ishan Thakkar, Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523-1373 USA; ishan.thakkar@colostate.edu.