Improving the Latency-Area Tradeoffs for DRAM Design with Coarse-Grained Monolithic 3D (M3D) Integration

Chao-Hsuan Huang, Ishan G Thakkar
Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY, U.S.A.
{chu276, igthakkar}@uky.edu

***Abstract*— Over the years, the DRAM latency has not scaled proportionally with its density due to the cost-centric mindset of the DRAM industry. Prior work has shown that this shortcoming can be overcome by reducing the critical length of DRAM access path. However, doing so decreases DRAM area-efficiency, exacerbating the latency-area tradeoffs for DRAM design. In this paper, we show that reorganizing DRAM cell-arrays using the emerging monolithic 3D (M3D) integration technology can improve these fundamental latency-area tradeoffs. Based on our evaluation results for PARSEC benchmarks, our designed M3D DRAM cell-array organizations can yield up to 9.56% less latency and up to 21.21% less energy-delay product (EDP), with up to 14% less DRAM die area, compared to the conventional 2D DDR4 DRAM.**

***Index Terms*—DRAM, Monolithic 3D Integration, Bitlines, Sense Amplifiers, DRAM Access Latency**

I. Introduction

Over the years since the emergence of DRAM, various manufacturers have deliberately sacrificed the access latency benefits of the continuing DRAM process scaling, to achieve greater cell density (i.e., more DRAM cells per unit die area) and lower cost-per-bit for DRAM, by sharing area-hungry DRAM access peripherals (e.g., sense amplifiers (SAs)) with increasingly large number DRAM cells [1]. Consequently, most DRAM designs today have very long internal critical access path, corresponding to having many DRAM cells interconnected through a long wire called a *bitline* [1]. This design choice has slowed down the DRAM latency scaling, which in turn has exacerbated the "Memory Wall" problem by widening the performance gap between the processor and DRAM subsystems even further.

To alleviate the "Memory Wall" problem, which is crucial for meeting the performance demands of the modern data-driven computing applications, an efficient solution has been to use short-bitline DRAM architectures (e.g., [2], [3], [4]). However, these architectures require more SAs for a given die capacity, increasing the die area, and thus, reducing the die's cell density and cost-per-bit. As a result of this inherent area-latency tradeoffs in short-bitline DRAMs, the industry has relegated them to specialized applications only such as high-end networking systems (e.g., [5]) that can tolerate a very high cost for a very low latency. *For more widespread adoption of the short-bitline DRAM architectures, the per-die cell density for such DRAM architectures needs to be increased, for which improving the fundamental latency-area tradeoffs for DRAM design is of paramount importance.*

To improve the latency-area tradeoffs for DRAM design, and consequently improve the per-die cell density for DRAM, we propose to use the emerging monolithic 3D (M3D) integration technology [6]. In this paper, we show for the first time that reorganizing the traditional 1T1C (1-transistor 1-capacitor) DRAM die (we consider DDR4 DRAM [7]) at the subarray-level granularity with the M3D technology can mitigate the inherent latency-area tradeoffs for DRAM design, in spite of suffering from performance degradation related to the M3D fabrication process [8]. Our idea is to partition the sense-amplifiers and other peripherals on a different M3D tier from the tier with DRAM cell-arrays. We present two different M3D DDR4 DRAM designs, both with improved cell density (die area) and access latency, compared to the baseline 2D DDR4 DRAM of the same capacity.

Our key contributions in this paper are summarized below.

- To relax the latency-area tradeoffs for DRAMs, we reorganize the cell-array of the commodity 2D DDR4 DRAM [7] using the coarse-grained M3D integration technology;
- We present the subarray-level bank layouts as well as the latency, area, and energy analysis (based on SPICE and other circuit-level simulations) for our designed M3D DDR4 DRAMs;
- We evaluate our designed M3D DDR4 DRAM architectures using Gem5 [9] based full-system simulations with PARSEC benchmarks [10], and compare their performance and energy-efficiency with the conventional 2D DDR4 DRAM.

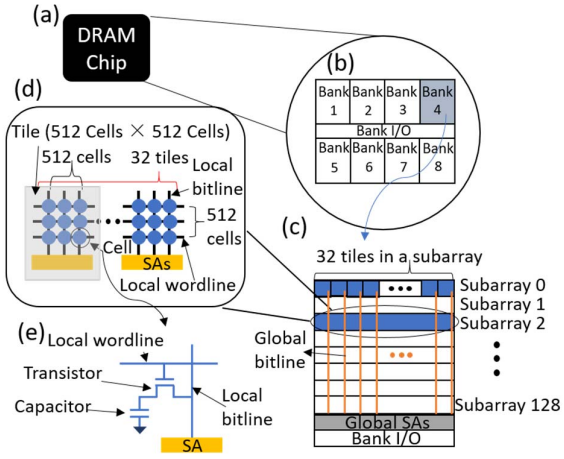


Fig. 1: Schematic structures of (a), (b) a DRAM chip, (c) a DRAM bank, (d) a DRAM subarray, and (e) DRAM cell. SAs: Sense Amplifiers.

II. Background on DRAM Structure and Operation

A. DRAM Chip Structure, Operation, and Timing Constraints

A DRAM chip typically employs a hierarchical cell-array organization, which is briefly illustrated in Fig. 1. A cell is the smallest unit in the hierarchy, and the critical path for accessing a cell includes a local bitline, a local SA, a global bitline, a global SA, and bank I/O (Fig. 1). Fig. 2 illustrates three DRAM operation

2576-6996/20/$31.00 ©2020 IEEE
DOI 10.1109/ICCD50377.2020.00076

phases (activation, data I/O, and precharging), along with their related DRAM timing parameters and activities during these phases that occur in various DRAM structures, such as DRAM array, peripherals, command bus, and data bus. The definitions of various DRAM timing parameters and the DRAM structures that dominate the latency contributions for respective timing parameters are listed in Table 1. *From Table 1, lengths of local and global bitlines are major contributors to all critical access latency parameters*.

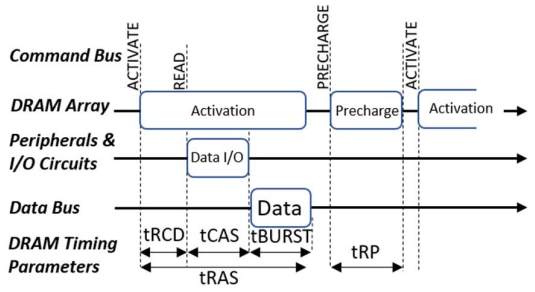


Fig. 2: Three phases of DRAM operation and related timing parameters.

TABLE 1. VARIOUS DRAM TIMING PARAMETERS.

Timing Parameters	Descriptions	DRAM Structure that Mainly Contributes to the Delay
tRCD	Row to Column Command Delay	Local Bitline
tCAS	Column Access Strobe Latency	Global Bitline, I/O
tBURST	Data Burst Duration	Interface
tRAS	Row Access Strobe	Local and Global Bitline, I/O, Interface
tRP	Precharge Delay	Local Bitline

B. Latency-Area Tradeoffs for 2D DRAMs

From Table 1, having shorter local bitlines is the fundamental approach for reducing tRCD, tCAS, tRP, and close-page access latency (tRCD+tCAS+tBURST). However, from [1], reducing the length of local bitlines comes at the cost of exacerbated latency-area tradeoffs. To evaluate these latency-area tradeoffs for different local bitline lengths, we evaluated the die area, tRCD and close-page access latencies for *iso-capacity* DDR4 [7] bank organizations with 512, 256, 128, 64, 32 cells per local bitline (respectively referred to as DDR4-512, DDR4-256, DDR4-128, DDR4-64, and DDR4-32), using our CACTI [12] and SPICE [13] based DDR4 models discussed in Section IV. The results of our evaluation are plotted in Fig. 3. Fig. 3 also plots results for our proposed M3D organizations, which will be discussed in the next section. From Fig. 3, as we move from DDR4-512 to DDR4-32, tRCD does not reduce beyond DDR4-128 without drastic (>2×) increase in die area. This is because, as move from DDR4-512 to DDR4-32, a greater number of subarrays and SA stripes are required in a DRAM bank of unchanged capacity due to the shortened local bitlines, which increases the total DRAM die area. After DDR4-128, the reduction in tRCD due to the reduction in local bitline length becomes negligible, but the increase in DRAM die are still remains significant. On the other hand, as we move from DDR4-512 to DDR4-32, the close-page access latency starts increasing significantly from DDR4-128. This is because, due to the increasing number of required subarrays, the length of global bitlines increases, contributing more significantly to the tCAS and close page access latencies. Thus, contrary to the observation made for tRCD, *shorter local bitlines yield longer close-page access latencies, which can result in longer average memory access latency*.

It is clear from these findings that shortening local bitlines does not help unless the global bitlines can also be shortened in concurrence, without incurring any extra die area cost. Intuitively, global bitlines can be shortened reducing the bank size and increasing the bank count per DRAM die. However, doing so cannot come without significant decrease in the per-die cell density of DRAM. Therefore, to address this shortcoming, we take a promising alternative approach of reorganizing DRAM banks using the coarse-grained monolithic 3D integration (M3D) technology, as discussed next.

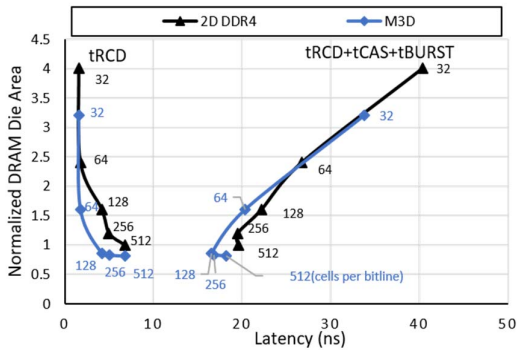


Fig. 3: Normalized DRAM die area versus tRCD and close-page access latency (tRCD + tCAS + tBURST) for various local bitline lengths (cells per local bitline) for conventional 2D and M3D-enhanced DDR4 DRAM.

III. REORGANIZING DRAMS WITH M3D INTEGRATION

A. Monolithic 3D Integration Technology

M3D technology enables sequential processing and integration of multiple tiers (mostly up to two tiers) of logic circuits on the same die. To vertically connect various components located on different M3D tiers, the M3D-integrated chips utilize monolithic inter-tier vias (MIVs) that are several orders of magnitude smaller in physical dimensions (~50nm×100nm) than TSVs (~1-3μm×10-30μm) [14]. Moreover, an MIV has 10Ω resistance and 0.2fF capacitance. *This enables vertical routing of connections using MIVs with nanoscale contact pitch and negligible overheads of parasitic loading*. More details on the M3D integration technology can be found in [8]. *The disadvantage of M3D integration is* that, due to the sequential integration process, the resistance of the required tungsten interconnects on the bottom M3D tier can increase by up to 2×, and the transistor performance on the second/top tier can degrade by 10~20% [8]. We mitigate this tier degradation issue by employing an established workaround from [8] to make the best use of the M3D technology for designing better performing DRAM organizations.

B. Design of Monolithic 3D (M3D) DRAMs

We reorganize DDR4 DRAM [7] with M3D technology. In our designed M3D DDR4 variants, to avoid performance degradation on M3D tiers, we place the SAs and other peripherals (e.g., write drivers, precharge units, SA I/O, local wordline drivers, address decoders) on the bottom tier, and the DRAM cell-arrays (including the DRAM interconnects such as bitlines and wordlines) on the top tier. Fig. 4 shows schematics of DDR4-512, M3D-512 and M3D-128 organizations. Moreover, we evaluated tRCD, close-page access latency, and die area for these and other M3D organizations (M3D-512 to M3D-32) to derive the latency-area tradeoffs for M3D designs, shown in Fig. 3. For M3D-512 (Fig. 4(b)), placing SAs and peripherals underneath the DRAM tiles shortens global bitline length L_{GBL} per subarray by 234F (117F for SAs + 90F for precharge units + 27F for write drivers), yielding total L_{GBL} to be 132,969F for M3D-512, compared to L_{GBL} of 162,687F for DDR4-

512 (Fig. 4(a)). As a result of reduced L_{GBL}, M3D-512 achieves reduced tCAS of 8.9ns, compared to tCAS of 10.3ns for DDR4-512. Moreover, we evaluate that the area of a 128Mb M3D-512 bank is 3.2mm^2, which is significantly less than the 3.9mm^2 area of a 128Mb DDR4-512 bank. Along the same lines, M3D-128 reaches the pinnacle of the benefits of M3D integration (Fig. 4(c)), for which L_{GBL} of 142,569F and L_{LBL} of 256F are achieved (Fig. 4(c)). These values of L_{GBL} and L_{LBL} are ~1.14× and 4× less respectively than the L_{GBL} and L_{LBL} values for DDR4-512. Moreover, we evaluate that the area of a 128Mb M3D-128 bank is 3.4mm^2, which is only 0.2mm^2 less than the 3.2mm^2 area of a 128Mb M3D-512 bank. Due to these benefits, the tRCD and close-page access latency curves for the M3D organizations are closer to the origin than the curves for the DDR4 organizations (Fig. 3), which indicates that the M3D organizations relax the fundamental latency-area tradeoffs for DRAM design. *These results corroborate the excellent capabilities of the M3D technology in mitigating the fundamental latency-area tradeoffs for DRAMs, to achieve simultaneous benefits in DRAM access latency and per-die cell density.*

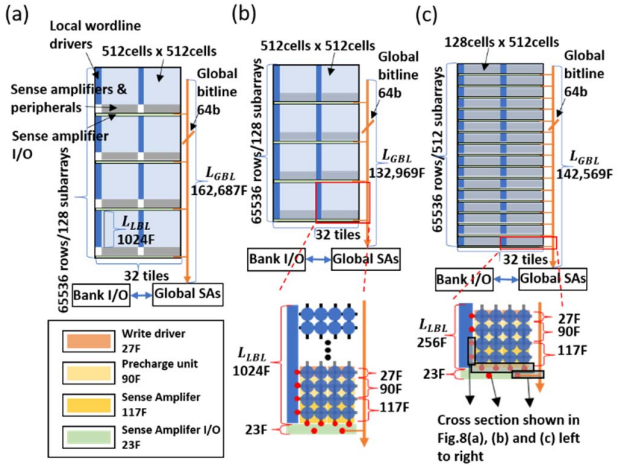


Fig. 4: Illustration of three example bank organizations of the folded-bitline DRAM; (a) 512 cells per local bitline 2D DDR4 DRAM (DDR4-512), (b) 512 cells per local bitline M3D DRAM (M3D-512), and (c) 128 cells per local bitline M3D DRAM (M3D-128). Although the local/global address decoders are not shown, they are placed on the bottom tier.

Implementation Overheads for M3D DRAM Organizations: To implement our proposed M3D DRAM organizations, we route the connections of the SAs and other peripherals on the bottom tier to the DRAM interconnects on the top tier using MIVs and tier-specific metal-via stack. Fig. 5 illustrates the MIV-based vertical interconnects' cross-sections. Evidently, each vertical connection includes one M1-M5 metal-via stack and an MIV. We extract the parasitic resistance and capacitance values for the vertical interconnects from [15] to be 0.23fF and 20Ω for the worst-case scenario (i.e., highest parasitic loading) shown in Fig. 5(c). In addition, our M3D organizations also suffer from the performance degradation of the DRAM cell access transistors placed on the top tier. We evaluate this degradation in terms of I_{ON}-I_{OFF} characteristics using the methods from [8]. We incorporate the vertical interconnects' parasitic values and the degraded access transistors' I_{ON}-I_{OFF} characteristics in our LTSpice model from [16], to evaluate their impact on various DRAM latency parameters such as tRCD and tRP. Fig. 6(a) shows the results of our LTSpice simulations for tRCD parameter extraction for the DDR4-512, M3D-512, and M3D-128 organizations. As discussed earlier, both DDR4-512 and M3D-512 have the same value of 1024F for L_{LBL}. From Fig. 6(a), even with the addition of parasitic overheads of vertical interconnects and performance degradation of the access transistor, tRCD latency for M3D-512 hardly changes significantly compared to the tRCD latency for DDR4-512. *From these findings, we can conclude that M3D integration incurs negligible overhead for our proposed M3D DRAM organizations.*

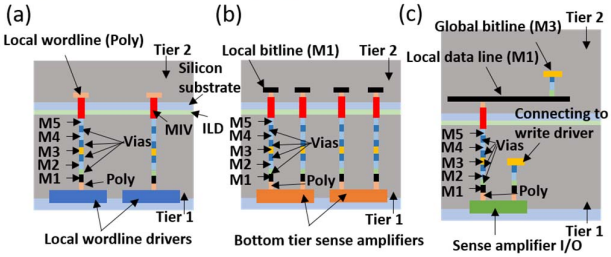


Fig. 5: Illustration of the vertical interconnects' cross-sections between (a) local wordline drivers and local wordlines, (b) sense amplifiers (SAs) and local bitlines, and (c) sense amplifier (SA) I/O and local data line & global bitline, for our proposed M3D DRAMs. ILD: Inter Layer Die electric; MIV: Monolithic Inter-tier Vias. Although the local/global address decoders are not shown, they are placed on the bottom tier.

TABLE 2. MODELING PARAMETERS FOR VARIOUS DDR4 AND M3D DRAM ORGANIZATIONS.

	2D DDR4	M3D 512	M3D 128
Ranks	1	1	1
Banks	8	8	8
Page Size	16kb	16kb	16kb
Cells per Bitline	512	512	128
Timing Parameters (ns)			
tRCD	6.77	6.78	4.2
tCAS	10.29	8.96	9.82
tRP	9.58	9.6	4.04
tRC	26.64	25.34	18.05
tFAW	35.8	35.3	14.4
tREFI	7800	7800	7800
Per Access Energy Values (nJ)			
Activation Energy	0.59	0.58	0.24
Read Energy	1.1	0.94	1.05
Write Energy	1.1	0.94	1.05
Refresh Energy	35.22	32.51	23.23
Area Analysis			
Subarray (mm^2)	0.031	0.025	0.007
Bank (mm^2)	3.926	3.209	3.42
#MIVs per Bank	0	5,243,008	14,680,576
MIV Area per Bank (mm^2)	0	0.01	0.029
Subarray Height	1281F	1047F	279F
Local Bitline Length	1024F	1024F	256F
Local Bitline Resistance	20000Ω	20010Ω	5010Ω
Local Bitline Capacitance	72fF	72.2fF	18.2fF

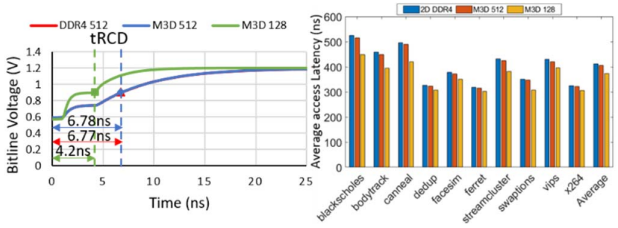
IV. AREA, TIMING, AND ENERGY ANALYSIS

We modeled various DRAM organizations for 22nm technology node using CACTI [12]. Each DRAM cell consumes $6F^2$ area, while the height and pitch of a SA are 117F and 6F respectively. We evaluate the lengths of local and global bitlines also using CACTI based models of various DDR4 and M3D organizations. For M3D organizations, we hide the area consumed by the SAs and other peripherals, to come up with bank and DRAM die area. We extract energy values from CACTI based models as well. Moreover, to evaluate various DRAM latency parameters and close-page access latency, we use the sense amplifier with DRAM subarray bitline model from [16] in LTspice [13]. The model from

[16] is for 45nm, so we scale it to 22nm following the standard scaling guidelines for wires and interconnects in CMOS technologies. Our extracted modeling parameters are listed in Table 2 for various DDR4 and M3D DRAMs.

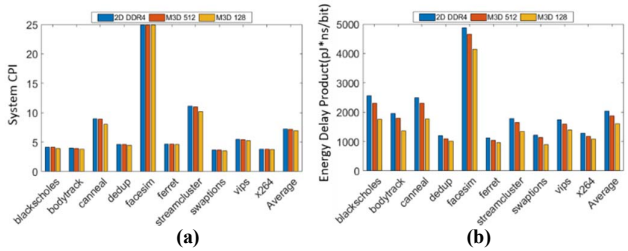
V. Simulation Setup and Results

We performed trace-driven simulations using NVmain [17] to compare the power and energy-delay product values for our considered DRAM organizations. We consider the iso-area organizations DDR4-512, M3D-512, and M3D-128 for system-level comparison. We also perform full-system simulations in Gem5 [9], to evaluate cycles per instruction (CPI) and average latency results. We used the PARSEC benchmarks [10] for the analysis, the trace files were extracted from detailed cycle-accurate simulations using GEM5 [9]. The configuration of GEM5 for both trace-driven and full-system simulations is shown in [11] and [18]. We considered 10 different applications form the PARSEC suite: Blackscholes, Bodytrack, Canneal, Dedup, Facesim, Ferret, Streamcluster, Swaptions, Vips, and X264. For the trace-driven simulations, we ran each PARSEC benchmark for a "warm up" period of one billion instructions and captured memory access traces form the subsequent one billion instructions extracted. For the full-system simulations, we run PARSEC benchmarks in their critical regions of interest (ROIs) in Gem5. We use parameters from Table 2 to model the DDR4-512, M3D-512, and M3D-128 organizations in Gem5 and NVMain. Fig. 7(a) shows system-level cycle per instruction (CPI) values for our considered DRAM organizations across PARSEC benchmarks. Compared to the baseline DDR4-512, M3D-512 and M3D-128 organizations yield about 0.54% and 3.74% lower system CPI respectively. Similarly, Fig. 6(b) shows average access latency values. Compared to the baseline DDR4-512, M3D-512 and M3D-128 organizations yield about 1.65% and 9.56% less average latency respectively. Shorter tRC time and shorter close-page access latencies for the M3D-512 and M3D-128 organizations result in lower CPI and average latency values for them.



(a) (b)

Fig. 6: (a) Results of LTspice simulations for tRCD extraction for the DDR4-512, M3D-512 and M3D-128 organizations, and (b) average access latency results for the DDR4-512 (blue), M3D-512 (red), and M3D-128 (yellow) organizations across PARSEC benchmarks.



(a) (b)

Fig. 7: (a) System cycles per instruction (CPI), and (b) energy-delay product (EDP) results for the DDR4-512 (blue), M3D-512 (red), and M3D-128 (yellow) organizations across PARSEC benchmarks.

Fig. 7(b) shows energy-delay product (EDP) values. EDP indicates how balanced different designs are in terms of energy consumption and delay. We calculate EDP by multiplying energy per bit (pJ/bit) with average access latency (ns), while energy per bit is total power divided by throughput (bit/s). The results show that M3D-512 and M3D-128 respectively have 7.49% and 21.21% lower EDP than the baseline DDR4-512.

VI. Conclusions

In this paper, we showed how the fundamental latency-area tradeoffs for DRAM can be mitigated by reorganizing DRAM cell-arrays using the emerging monolithic 3D (M3D) integration technology. We evaluated the latency-area tradeoffs for various configurations of 2D DDR4 and M3D DRAMs. Based on our evaluation results for PARSEC benchmarks, we found that our designed M3D DRAM cell-array organizations can yield up to 9.56% less latency and up to 21.21% less energy-delay product (EDP), with up to 14% less DRAM die area, compared to the conventional 2D DDR4 DRAM. These results corroborate the excellent capabilities of the M3D technology in mitigating the fundamental latency-area tradeoffs for DRAMs, to achieve simultaneous benefits in DRAM access latency and per-die cell density.

References

[1] D. Lee *et al.*,"Tiered-latency DRAM: A low latency and low cost DRAM architecture," *HPCA*, 2013.

[2] T. Kimura *et al.*, "64Mb 6.8ns random row access DRAM macro for ASICs," *ISSCC*, 1999.

[3] Micron. RLDRAM 2 and 3 Specifications. http://www.micron.com/products/dram/rldram-memory.Linear Technology Corp., "LTspice IV," http://www.linear.com/LTspice

[4] Y. Sato *et al*. Fast Cycle RAM (FCRAM); a 20-ns random row access, pipe-lined operating DRAM. In Symposium on VLSI Circuits, 1998.

[5] C. Toal *et al*.,"An RLDRAM II Implementation of a 10Gbps Shared Packet Buffer for Network Processing," *AHS*, 2007.

[6] P. Batude *et al.*, "3-D sequential integration: A key enabling technology for heterogeneous co-integration of new function with CMOS,"*IEEE JETCAS, 2012.*

[7] Micron Technology, Inc., "8Gb: x4, x8, x16 DDR4 SDRAM - Micron Technology,Inc.",https://www.micron.com//media/client/global/documents/products/data-sheet/dram/ddr4/8gb_ddr4_sdram.pdf

[8] S. Panth *et al.*," Tier Degradation of Monolithic 3-D ICs: A Power Performance Study at Different Technology Nodes*," IEEE TCAD*, 2017.

[9] N. Binkert *et al*., "The Gem5 simulator," *CAN*, 2011.

[10] C. Bienia *et al.*,"The PARSEC Benchmark suite: Characterization and architectural implications," *PACT*, 2008.

[11] I. Thakkar, *et al*., "3D-ProWiz: An Energy-Efficient and Optically-Interfaced 3D DRAM Architecture with Reduced Data Access Overhead," *TMSCS*, 2015.

[12] R. Balasubramonian *et al* "CACTI 7: New Tools for Interconnect Exploration in Innovative Off-Chip Memories," *TACO*,2017.

[13] Linear Technology Corp., "LTspice IV," http://www.linear.com/LTspice

[14] S. Musavvir *et al*., "Inter-Tier Process Variation-Aware Monolithic 3D NoC Architectures," ArXiv, 2019. arXiv:1906.04293v1.

[15] Y. Lee, P. Morrow and S. K. Lim, "Ultra high density logic designs using transistor-level monolithic 3D integration," *ICCAD*, 2012.

[16] Kevin K. Chang *et al.*, "Understanding Reduced-Voltage Operation in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms," *POMACS*, 2017 .

[17] M. Poremba *et al.*,"NVMain 2.0: A user-friendly memory simulator to model (non-)volatile memory systems," *IEEE LCA*, 2015.

[18] I. Thakkar *et al*., "DyPhase: A Dynamic Phase Change Memory Architecture with Symmetric Write Latency and Restorable Endurance," *IEEE TCAD,* 2018.