Short-Stack: Pushing Back the Pin Bandwidth Wall with FinFET-based eDRAM In-Package Last Level Cache

He Xiao, Wen Yueh, Saibal Mukhopadhyay, *Member, IEEE,* and Sudhakar Yalamanchili, *Fellow, IEEE*
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332
{hxiao, wyueh3}@gatech.edu, {saibal.mukhopadhyay, sudha.yalamanchili}@ece.gatech.edu

Abstract—The slow growth of the number of pins per package coupled with increasing device densities is leading to decreasing off-chip memory bandwidth per core which in turn leads to reductions in system level performance. In this work we present a 2-tier stacked IC structure, referred to as the *Short-Stack*, to push back this pin bandwidth wall. The Short-Stack consists of a processor die of multiple cores in the bottom tier, and a multi-banked last level cache (LLC) die including the memory controllers and network-on-chip (NoC) in the top tier, face-to-face bonded with the processor tier. We characterize the timing delay, power consumption, and density of an eDRAM implementation of the LLC, and consider a range of LLC designs using SRAM and eDRAM. Using a full-system, cycle-level simulator, we conduct a quantitative analysis of a 16-core Short-Stack executing SPLASH-2 benchmarks. We present performance results in comparison to i) a traditional 2D processor implementation and ii) a multi-tier 3D package with stacked DRAM memory in terms of performance, power and energy efficiency

Index Terms—Pin Bandwidth, Short-Stack, 3D IC, FinFET eDRAM.

1 Introduction

As computing systems move to extreme scale, the number of cores integrated into the package will dramatically increase exerting pressure on the pin bandwidth between the on-chip cores and off-chip memory system. According to a recent study, supply pins will take up a large proportion of the total number of pins in a package [1], indicating a decreasing available pin bandwidth per core and affecting energy and time performance especially for memory intensive applications. To address this pin bandwidth problem, we propose an effective approach to reduce the off-chip bandwidth requirements by deploying 3D IC technology and optimizing the cache system using FinFET based eDRAM. The novel microarchitecture, referred to as the Short-Stack, stacks the processor tier and the LLC tier together into a single package with face-to-face bonding. With the abundant LLC capacity offered by eDRAM, large inter-tier bandwith and low-latency communication can overcome the performance loss due to limited pin bandwidth. It is also demonstrated to be competitive with 3D architectures using stacked DRAM (i.e. Micron's Hybrid Memory Cube) in terms of manufacturing complexity and TSV reliability.

In this paper, we focus on a homogenous multicore architecture as our target microarchitecture that uses sixteen. x86 out-of-order cores. Each core has a private L1 cache and shares the LLC with other cores. The model is constructed with a cycle-based full system simulator Manifold [2], which adopts the eDRAM characterizations from our HSpice simulation. The simulation also deploys the Energy Introspector library [3] that provides multi-physics modeling capability, e.g., thermal, energy, and reliability. We compare the performance of the Short-Stack with a traditional planar processor design and a 3D architecture with DRAM stacks [4] in terms of performance, power consumption and energy efficiency. The results show that the Short-Stack has considerable promise to deal with evolving pin bandwidth constraints.

2 Pin Bandwidth Problem

The total pin count doubles approximately every six years on average across difference ranges of processor design [1]. Meanwhile, the supply pins required for the package grows as the square root of the supply current to maintain constant resistive loss per supply pin, indicating an increasing pressure on signal pins available for the package.

Possible solutions to address this problem include either reducing the current demands of the package (i.e. integrating on-chip voltage regulators) or reducing the off-chip memory bandwidth requirement. In this work, we present a two-tier 3D microarchitecture structure called the Short-Stack to minimize the demand for pin bandwidth with a FinFET-based eDRAM configured as the last level cache (LLC). The benefits of deploying the Short-Stack structure come from two sources. First, replacing the SRAM with eDRAM cells in the LLC will roughly increase the cache capacity by 2X with little performance loss [5]. For a typical LLC design optimized for bandwidth, increasing the cache capacity and associativity will reduce off-chip data requests. Second, using the 3D structure, the bandwidth and latency between the core and LLC tier decreases, which will improve the overall system performance (measured in instructions/cycle or IPC) and shorten the performance gap between planar and stacked DRAM architectures.

3 EDRAM Cell Modeling

In advanced process technologies, SRAM becomes increasingly difficult to design due to read-write access contention, and eDRAM is emerging as a rising alternative to the mainstream 6T SRAM design. The eDRAM cells are more compact than SRAM due to fewer transistors in the cell design; reducing require cell area by nearly 50%. The associated leakage power is also reduced due to a reduction in the number of devices. The absent of access contention between read and write accesses also improves the voltage margin in eDRAM cells.

In eDRAM parametric analysis, a high performance gaincell EDRAM cell is considered. A two-transistor (2T) NFET design is modeled for the analysis. In order to model equivalent speed of the SRAM, the NFET design is used in Figure 1. A similar 3T eDRAM has been implemented 65 nm process [6], yet

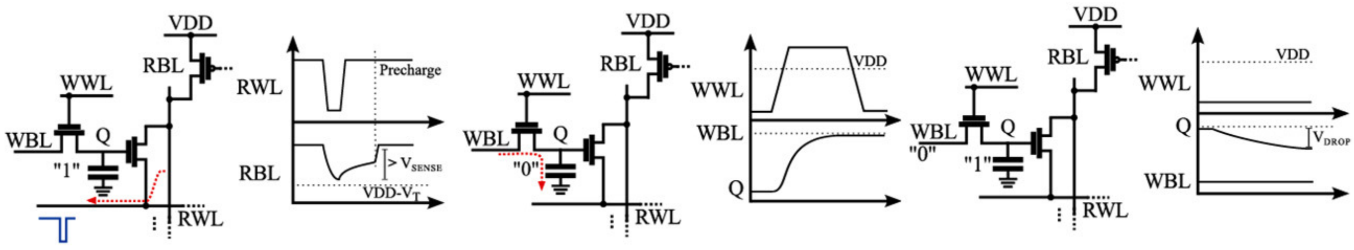


Fig. 1. The simulation methodology for thermal and supply cross-talk aware eDRAM analysis. The methodology co-simulates supply and thermal grids with process variation aware eDRAM analysis.

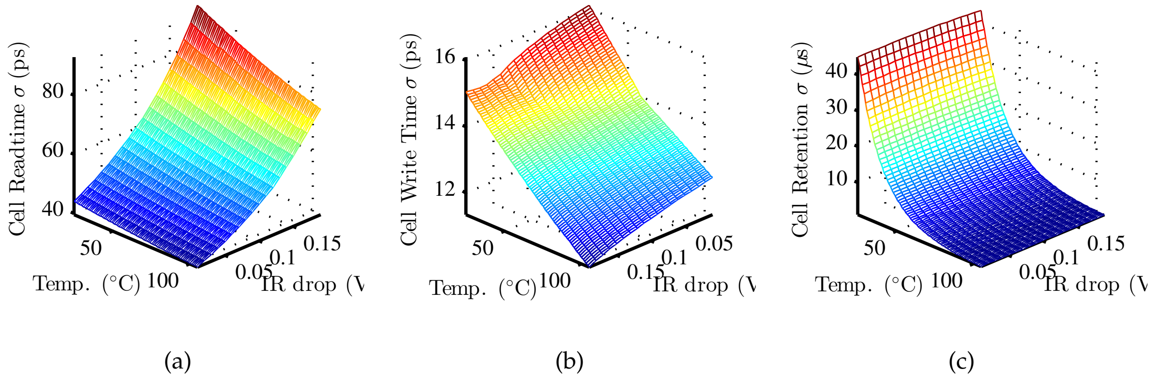


Fig. 2. The temperature sensitive delay for FinFET eDRAM: (a) read time, (b) write time, and (c) cell retention simulation.

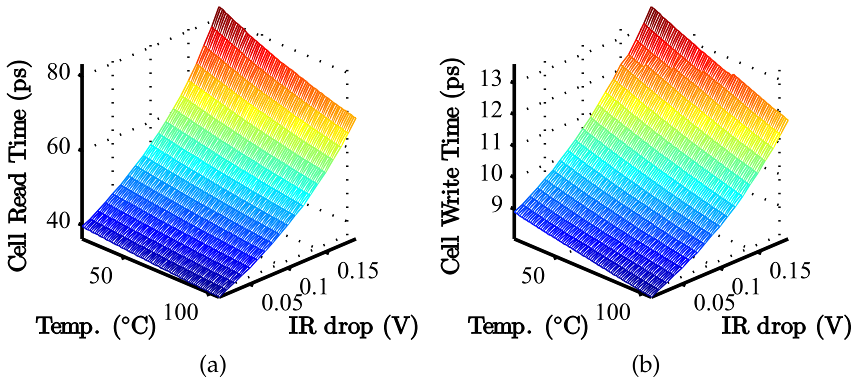


Fig. 3. The delay on FinFET SRAM: (a) read time, (b) write time simulation.

related design unlike the proposed 2T NFET EDRAM requires additional read transistor stack.

During the read operation, a sneak current path exists through all the unselected cells storing Q=1 in the columns. he sneak current path terminates as soon as the RWL is disabled. A differential sense-amplifier (with a constant reference) is used per column to sense the small RBL drop during the short pulse. Outputs of the SAs are multiplexed (column decoder) to access one entry at a time. Other than the cell access, the critical path of a EDRAM sub-bank is very similar to the 6T SRAM system.

The write operation uses single NFET. Writing Q=0 can be facilitated without modification. For Q=1, using a voltage of V_{mem} less than the peripheral voltage (V_{logic}=V_{mem} + V_{T}) guarantees a robust write. The eDRAM cell needs constantly refresh because of the non-regenerative charge inside the cell. The cell retention time is measured by the time of fully charged cell discharged to $100mV + VDD/2$.

The simulation result of FinFET SRAM and eDRAM are shown in Figure 2 and Figure 3. The read time and write time are affected more by the supply droop than the temperature. The thermal response of the FinFET devices in simulation shows interesting trends. Compared to SRAM, the read time is the same order of magnitude of the SRAM system with the same array configuration. This suggests that for a read dominant system the performance impact of eDRAM is very low. The write time is intrinsically worse than that of the SRAM design. However, in both SRAM and eDRAM cases the delay for random is still masked by the read time and hence not a significant impact to the system. The retention time impacts the overall eDRAM availability, power, and bandwidth. At room temperature corner, the retention time in FinFET is significant. Yet with a design of an advanced package for the eDRAM system, we can extend the system bandwidth significantly over a traditional 6T SRAM LLC.

4 The Short-Stack Structure and Model

The Short-Stack, depicted in Figure 4, consists of a processor die and an LLC die. Both dies are stacked in a 3D 2-tier structure using 16nm technology. In the simulation model used in this paper, the bottom tier implements 16, x86 Nehalem-like out-of-order cores, each with a private 16KB L1 instruction cache and 32KB data cache. Each core has 5 components: FE (pipeline frontend and L1 instruction cache), SCH (Out-of-Order scheduler), INT (integer unit), FPU (float-point unit) and DL1 (L1 data cache). The top tier is the proposed eDRAM LLC partitioned into 16 banks. Each cache bank contains 2MB capacity, and the content is shared among processors. The memory controllers are also integrated at the corners of the top tier. When there is an off-chip memory request, it will be passed down through the bottom tier via TSVs by the on-chip memory controllers.

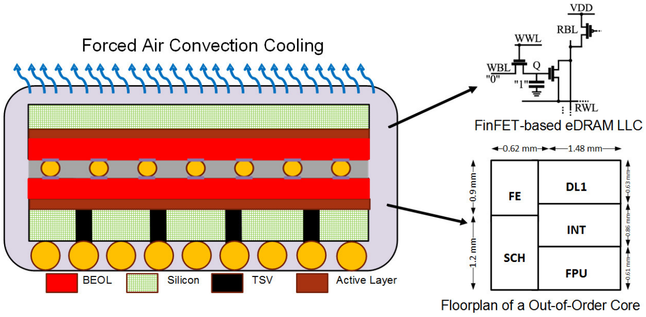


Fig. 4. The Short-Stack Structure with FinFET-based eDRAM LLC

Each tier of the Short-Stack contains 3 layers: BEOL layer, active layer and silicon base layer. The BEOL layer obtained by lift-off process is used for bonding and routing with a thickness of 25 μm in our simulation model. The device layout lies in the active layer, from where the heat is generated. The thickness of the active layer is 10 μm. The silicon base layer represents the silicon substrate, and has a thickness of 25 μm. TSVs are embedded in the silicon base layer of the processor die to establish the off-chip memory communication. The Short-Stack is placed on a bismaleimide triazine (BT) substrate through a silicon interposer. The BT substrate is attached to the printed circuit board using solder ball array. Forced air convection is assumed at the top of the chip stack with a heat transfer coefficient of 100 $W/m^2\ ^\circ C$.

5 Results and Analysis

5.1 Simulation Framework

The simulator is built based on the Manifold cycle-level full system simulation infrastructure, which executes application binaries selected from the SPLASH-2 [7] benchmark on the Linux guest OS, listed in Table 1. The timing model interacts with the Energy Introspector library to power consumption and thermal distribution. McPAT [8] and 3D-ICE [9] are used as the power and thermal models respectively. Applications are fastforwarded to the region of interest and executed until completion, and system states are sampled every 10μm.

We evaluate 4 system configurations with a fixed silicon area. The baseline 2D configuration places the cores and LLC in the same planar floorplan. The SS-sram and SS-edram are both Short-Stack configurations with LLC implementation of SRAM and eDRAM respectively. The 3D system model stacks the main memory on top of the microarchitecture, which serves as the upper-bound of system performance. The LLC of the 3D system uses eDRAM instead of SRAM due to leakage concerns at high temperature [10].

TABLE 1
SPLASH-2 benchmark characterization

application	uops	flops	memR	memW
barnes	2437M	11.9%	20.3%	15.6%
fmm	2624M	33.9%	18.1%	3.1%
lu-nc	415.9M	18.7%	21.1%	9.7%
radiosity	2891M	-	17.6%	10%
radix	325.8M	-	23.7%	13.8%
raytrace	719.6M	-	25.2%	9.6%
water-ns	675.1M	21.3%	17.6%	7.7%
ocean-c	665.4M	26.7%	21.6%	4.9%

5.2 System Performance

Figure 5.a shows the system throughput comparison of the 4 configurations. The 3D architecture has the highest performance gain because of the high memory bandwidth and low access delay, improving throughput by 43.4% on average compared to the baseline 2D configuration. ss-SRAM and ss-eDRAM also show an average MIPS improvement of 25.5% and 25.8% respectively. The increased LLC capacity of the ss-eDRAM compensates for the access loss due to eDRAM cell retention (i.e. cell refresh). For example, the throughput gain of ss-eDRAM is 27.6% compared to 23.6% in ss-SRAM when executing the radix application, indicating an improved LLC hit rate in ss-eDRAM as a result of the larger cache capacity and associativity.

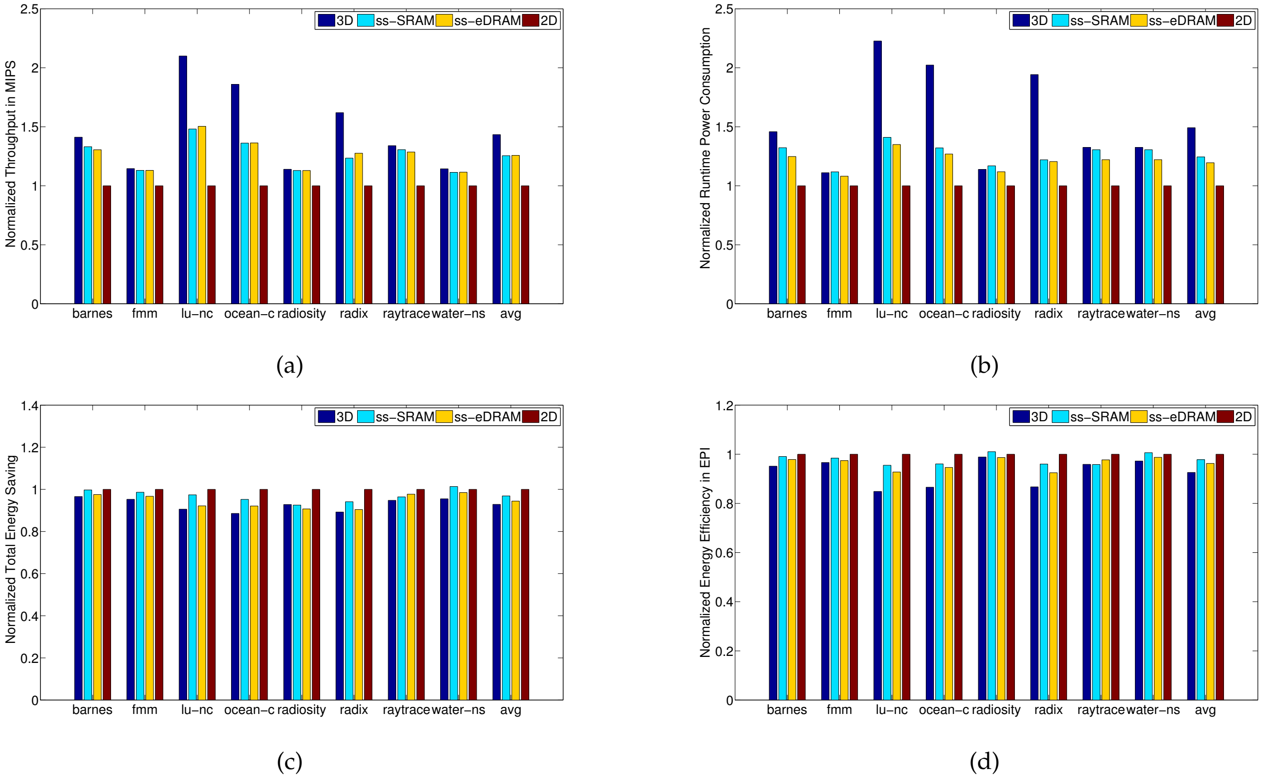
5.3 Power and Energy Consumption

The power consumption is proportional to the system performance, as shown in Figure 5.b. The power increase is significant among memory bound applications such as lu-nc, ocean-c and radix. These applications are accelerated by at least 25% using the Short-Stack structure. Meanwhile, the power consumption of the computational bounded application increases by around 10% in Short-Stack. The average power increase of ss-SRAM and ss-eDRAM are 24.4% and 19.5% respectively. The ss-eDRAM has overall 5% less power consumption than ss-SRAM as the leakage power of the eDRAM LLC is approximately 55% less compared to the SRAM implementations.

Although the average power increases in both 3D and Short-Stack, the total energy consumed by the entire system is reduced, as depicted in Figure 5.c. The total execution time of each application is largely reduced, and unnecessary energy (i.e. LLC leakage power) is saved due to shorter execution time. The 3D architecture achieves the highest energy saving of 7.4% as the execution time reduces to 70%. ss-SRAM and ss-eDRAM get a energy saving of 2.9% and 5.7% respectively. It is noticed that the energy is increased by 1.3% in ss-SRAM when executing water-ns. Water-ns runs at higher temperature due to its high IPC, and the increase of the leakage power offsets the benefits of faster execution.

5.4 Energy Efficiency

The energy efficiency is calculated based EPI, which measures the average energy to execute a single instruction. Figure 5.d provides a comparison among the 4 configurations. The system EPI is reduced both in 3D and Short-Stack, as the LLC in both systems is energy efficient as a result of improved throughput and reduced energy. The 3D structure reduces the average EPI by 7.4%, while the Short-Stack reduces the EPI by 2% and 3.7% respectively in ss-SRAM and ss-eDRAM.



(a)

(b)

(c)

(d)

Fig. 5. The comparison of different system configurations running SPLASH-2 on: (b) system throughput in terms of MIPS (c) normalized runtime power consumption (d) normalized energy saving (f) energy efficiency in terms of EPI

The memory bound applications (i.e. lu-nc, ocean-c) achieve over 5% EPI reduction in both Short-Stack configurations, as the performance gain surpasses the power increase executing these applications. Computational bound applications benefit less using Short-Stack due to the limited memory interactions.

When the temperature is high, ss-SRAM will suffer from the degradation of energy efficiency. The leakage power of the SRAM cells is significantly increased when executing radiosity and water-ns, and the EPI is increased by 2.2% and 0.6% compared to the 2D baseline.

6 Conclusion

The pin bandwidth constraints set limits to the progress of future processor design. In this paper, we propose the 2-tier Short-Stack structure to address this problem using FinFET-based eDRAM cells as the system LLC. Short-Stack brings significant performance improvement due to the substantial increase of the LLC bandwidth, and improves over 25% average performance gain compared to the 2D. On the other hand, the larger LLC capacity and cache associativity enabled by the eDRAM implementation reduces the miss rate and thus the demand of off-chip memory bandwidth. Moreover, we model 4 different system configurations using SRAM and eDRAM LLC in detail and study the impact of the Short-Stack. Full system simulation results show that the Short-Stack structure with eDRAM LLC saves 5.6% of energy consumption on average and improves approximately 4% of energy efficiency, suggesting the ss-eDRAM is a viable alternative for future processor designs.

Acknowledgments

This work was supported by the Semiconductor Research Corporation under task 2318.001.

References

[1] Phillip Stanley, Victoria Cabezas, Ronald P. Luijten, *Pinned to the Walls - Impact of Packaging and Application Properties on the Memory and Power Walls*, ISLPED 2011, pp. 51-56, 2011.

[2] J. Wang, J. Beu, R. Behda, T. Conte, Z. Dong, C. Kersey, M. Rasquinha, G. Riely, W. Song, H. Xiao, P. Xu, and S. Yalamanchili, *Manifold: A Parallel Simulation Framework for Multicore Ssytems*, ISPASS 2014, pp. 106-115, 2014.

[3] W. J. Song, S. Mukhopadhyay, S. Yalamanchili, *Energy Introspector: A parallel, composable framework for integrated power-reliability-thermal modeling for multicore architectures*, ISPASS 2014, pp. 145-144, 2014.

[4] He Xiao, Wen Yueh, S. Mukhopadhyay, and S. Yalamanchili, *Multi-physics driven Co-design of 3D Multicore Architectures*, InterPACKIC-NMM 2015, 2015.

[5] Mu-Tien Chang, Paul Rosenfeld, Shih-Lien Lu, and Bruce Jacob, *Technology Comparison for Large Last-Level Caches: Low-Leakage SRAM, Low Write-Energy STT-RAM, and Refresh-Optimized eDRAM*, HPCA 2013, pp. 143-154, 2013.

[6] Youn Sung Park, David Blaauw, Dennis Sylvester, and Zhengya Zhang, *A 1.6-mm2 38-mW 1.5-Gb/s LDPC decoder enabled by refresh-free embedded DRAM*, VLSIC 2012, pp. 114-115, 2012.

[7] C. Bienia, S. Kumar, and K. Li, *PARSEC vs. SPLASH-2: A quantitative comparison of two multithreaded benchmark suites on Chip-Multiprocessors*, IISWC 2008, pp. 47-56, 2008.

[8] Sheng Li, Jung-Ho Ahn, R.D. Strong, J.B. Brockman, D.M. Tullsen, and N.P. Jouppi, *McMAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures*, MICRO 42, pp. 469-480, 2009.

[9] A. Sridhar, A. Vincenzi, M. Ruggiero, Thomas Brunschwiler, and D. Atienza, *3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling*, ICCAD 2010, pp. 463-470, 2010.

[10] H. Xiao, Z. Wan, S. Yalamanchili and Y. Joshi, *Leakage power characterization and minimization in 3D stacked multi-core chips with microfluidic cooling*, SEMI-THERM, pp. 207-212, 2014.