Reliability-Performance Tradeoffs between 2.5D and 3D-Stacked DRAM Processors

Syed Minhaj Hassan, William J. Song, Saibal Mukhopadhyay, and Sudhakar Yalamanchili
School of Electrical and Computer Engineering,
Georgia Institute of Technology,
Atlanta, GA 30332

***Abstract*—Three-dimensional DRAM stacking has emerged as a vehicle for scaling system densities and performance improvement. The two design choices for interfacing to processors are - i) a separate core die connected to the DRAM stack via a silicon interposer (2.5D), and ii) DRAM die stacked on top of the core die (3D). These alternatives have different performance, power, and reliability behaviors. Specifically, 3D designs realize higher performance but operate at higher temperatures and thus exhibit lower lifetime. On the other hand, 2.5D designs provide lower bandwidth between the core die and the DRAM stack, but exhibit significantly longer lifetime due to less thermally-induced degradation. This paper explores this tradeoff between reliability and performance of 3D and 2.5D stacked memory systems. Our results indicate that, in general, lower voltage and frequency operations with 3D stacked systems may achieve balanced reliability-performance tradeoff.**

***Index Terms*—3D Stacked DRAM, Die Stacking Technology, Lifetime, Refresh Rate, Reliability, Silicon Interposer Technology, System Performance, Temperature**

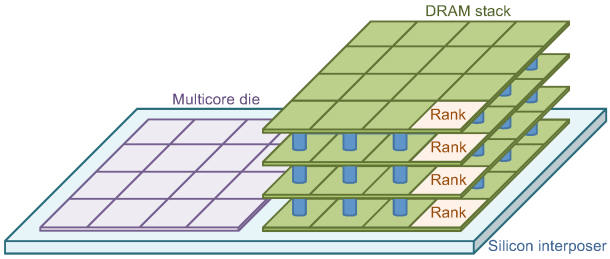
I. Introduction

Performance and energy overheads of data movement in large-scale computing systems motivate a transition towards data-centric computing designs known as *near-data processing* or *processing near memory*. Three-dimensional DRAM stack components, such as *high-bandwidth memory* (HBM) or *hybrid memory cube* (HMC), have emerged as vehicles for scaling system densities and performance due to i) increased inter-tier bandwidth, ii) reduced inter-tier latencies, and iii) the ability to integrate dies from different process technologies as a means of customization and hence performance improvement. There can be two possible ways of integrating DRAM stacks with a multicore die; 1) a separate core die with one or more DRAM stacks connected via a silicon interposer and 2) stacking DRAM dies on top of the core die. The former method is generally referred to as a *2.5D system*, and the latter as a *3D system* as shown in Figure 1a and 1b, respectively.

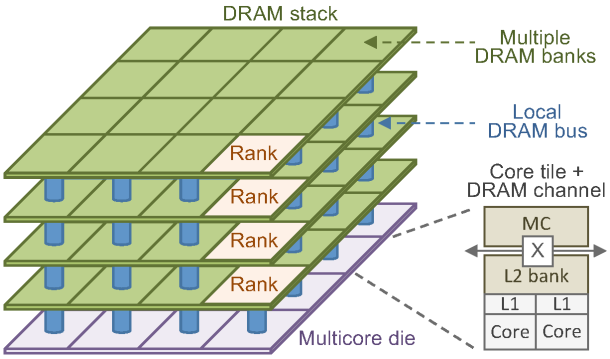
These two stacked DRAM implementations have distinct advantages and drawbacks. The 3D design provides greater memory bandwidth, but DRAM dies experience higher operating temperatures induced by the underlying multicore die and application characteristics. Higher temperature requires more frequent DRAM refresh operations (or shorter duty cycles until all DRAM rows are refreshed), which have a negative impact on performance because of reduction in available memory bandwidth. The 2.5D design provides lower memory bandwidth than the 3D design, although still higher than feasible with conventional off-chip memories.

More importantly, such distinct thermal footprints cause different reliability behaviors. Although the 3D-integrated DRAM stack provides superior performance to the 2.5D implementation, its higher operating temperature diminishes DRAM reliability. Therefore, *reliability-performance tradeoff* is an imminent challenge for stacked DRAM processors, and this paper explores this tradeoff between 2.5D and 3D implementations, encompassing the analysis of 3D microarchitectures, applications, and DRAM reliability.

This paper makes the following contributions.



(a) 2.5D: DRAM stack integrated with a core die on a silicon interposer.



(b) 3D: DRAM stack on top of a multicore die.

Figure 1. Illustration of integrating DRAM stacks with core dies. The baseline core tiles consists of 2 out-of-order cores, generating significant heat, along with a bank of shared L2 cache.

- Analyzes and compares the reliability-performance tradeoffs between multicore systems with 3D and 2.5D stacked memory.
- Characterizes the relationship between DRAM refresh operations and memory system organization.
- Identifies correlation between frequency, performance, temperature, and reliability and uses it to quantify the reliability-performance tradeoff.
- Empirically determines that feasible applications with 3D memories are memory-intensive applications, running at low operating frequencies.

II. Experiment Methods

Our baseline 3D and 2.5D systems are composed of a 4x4 tiled multicore architecture with 32 out-of-order cores as shown in Figure 1b. Each core is associated with a private L1 cache. Each tile is coupled to a local L2 cache bank. All L2 banks collectively operate as a shared L2 cache. For 3D stacks, the core die also contains a 4x4 tori network with 16 memory controllers (MCs). Each MC controls a Micron style DRAM vault consisting of four vertical sub-layers of DRAM communicating with the MC through a 64-bit wide shared through silicon via (TSV) bus and with each sub-die acting as a single rank. For the 2.5D system, the base

978-1-4673-9157-5/16/$31.00 ©2016 IEEE

logic layer consists of 4 MCs connected through 4 parallel memory channels. The four MCs each control a quarter of a die, with multiple dies sharing a channel similar to the 3D system.

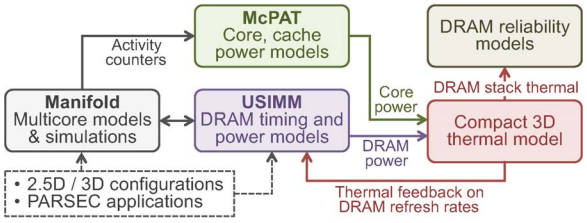


Figure 2. Simulation flow of evaluating reliability, thermal, performance tradeoffs between 2.5D and 3D-stacked DRAM processors.

Figure 2 depicts the simulation flow used in our study. We used the Manifold full system microarchitecture simulator [9] for performance simulations that boots a Linux kernel and executes multi-threaded PARSEC and SPLASH applications [1]. Architectural activity counters from Manifold simulations were supplied to McPAT [3] to estimate core area and power. We adopted USIMM [2] to simulate 3D-stacked DRAM and estimate the power of DRAM banks. Performance and power of 2.5D and 3D implementations with PARSEC benchmarks were measured by varying the clock frequency of cores between 800MHz and 4GHz. Collected power traces were input to a compact 3D thermal model [10], which first converts the floorplan power to thermal grid power and calculates the temperature based on power density and thermal coupling of each grid. The calculated steady-state temperatures are then used to calculate refresh rates of different dies (ranks) and channels which are fed back to USIMM to calculate the updated performance results.

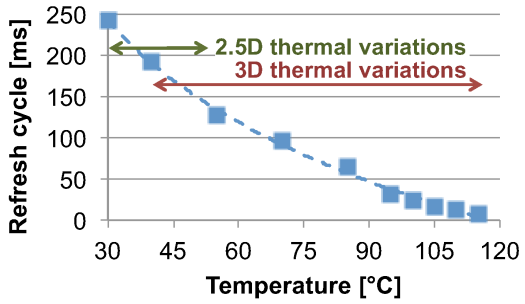


Figure 3. Impact of operating temperatures on DRAM refresh rates [6], and the extent of thermal variations in 2.5D and 3D processors. Higher temperature means slower DRAM retention time and significantly higher refresh rate.

III. Performance and Thermal Correlations in 2.5D vs 3D-Stacked Memory Systems

The baseline multicore processor die dissipates power increasing its temperature. In our 3D system configuration, this die is at the bottom of the stack while the heat sink and the fans are connected to the top with multiple DRAM layers in between. This decreases the heat removal capability of the heat sinks increasing the temperature of the chips. Furthermore, since the cores, the caches and the DRAMs are stacked, the overall power of the system is dissipated in a smaller volume increasing power density and hence has a higher overall temperature as compared to conventional 2D or 2.5D systems. The logic layer in the 2.5D stack, on the other hand, has relatively lower power density (per volume), thus the operating temperature of 2.5D stack is smaller as well. Thermal coupling from the baseline core die is also low as long as the distance between the two dies is greater than 10mm [11].

The retention time characteristics of various DRAM cells vary widely. The commonly used refresh rate of 64ms is a conservative value designed for the weak cells to operate at high temperatures. Some recent works have pointed towards decreasing the refresh rate for lower temperatures [13]. The self refresh current values as described in data sheets of modern DDRs [12], indicate lower refresh rates at lower temperatures [14]. We extrapolated temperature vs refresh rate curve using these sources as shown in Figure 3. The temperature of the DRAM stack in the 2.5D design is primarily induced by memory activities, whereas the temperature in the 3D stack is dominated by the temperature of underlying multicore die and experiences much wider range of operating temperatures. This raises both performance and reliability concerns in the 3D design.

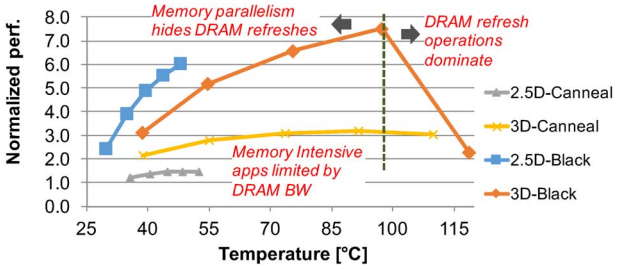


Figure 4. Impact of operating temperatures on the performance of 2.5D and 3D-stacked processors. 3D-stacked DRAM's performance is not significantly penalized by higher operating temperatures. Note: Black represents compute-intensive applications and canneal represents memory-intensive applications.

Figure 4 shows the impact of DRAM temperatures (and corresponding refresh rates) on the performance of 2.5D and 3D-stacked processors. Different temperatures in 3D and 2.5D designs are obtained by varying the frequency of the cores. Here, we assume that the power budget allows both designs to operate at higher frequencies. One representative memory-intensive and one compute-intensive application is used. The DRAM bandwidth of memory-intensive application saturates quickly with little effect on performance when increasing the frequency of the cores. Although higher bandwidth in 3D memories results in better performance as compared to 2.5D memories, the improvement is mitigated by the increased refresh rate at higher temperature reducing memory bandwidth availability. The performance of compute-intensive applications, on the other hand, increases with core frequency. In this case again, 3D-stacked DRAM provides superior performance despite its higher operating temperature. Even though higher operating temperature in the 3D design requires more frequent DRAM refresh operations than the 2.5D design, channel- and bank-level parallelism in 3D effectively hides DRAM refresh operations by parallelizing refreshes with normal read/write accesses. Thus, higher operating temperatures have relatively minor impact on overall performance. Eventually at very high temperatures, the performance gets dominated by DRAM refreshes and the application becomes memory bound.

IV. Frequency, Reliability and Thermal Correlations

In the previous section, we established that 3D DRAM delivers higher performance than the 2.5D design for both compute- and memory-intensive applications, even at a higher operating temperature. However, increased temperature in 3D may affect system

reliability significantly. To understand how DRAM reliability is dependent on its operating temperature, we adopted a DRAM lifetime reliability model from Micron [4], where mean-time-before-failures (MTBF) is proportional to temperature and voltage acceleration factors, AF_T and AF_V,

$$MTBF \propto AF_T \times AF_V = e^{\frac{E_a}{k}\left(\frac{1}{T_o}-\frac{1}{T_s}\right)} \times e^{\beta(V_s-V_o)} \quad (1)$$

where E_a is the activation energy, and k is the Boltzmann's constant. T_o and T_s denote operation and stress temperatures, respectively. β is a process-dependent constant, and V_o and V_s are operation and stress voltages.

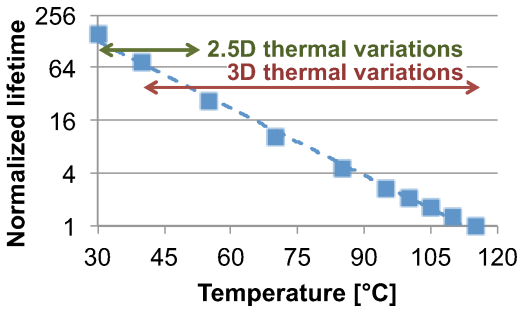


Figure 5. Impact of temperatures on DRAM lifetime reliability, and the extent of thermal variations in 2.5D and 3D processors. Lifetime reduces exponentially with temperature.

Figure 5 shows the correlation between operating temperature and the resulting DRAM lifetime. It can be seen that DRAM lifetime degrades very quickly with increased operating temperature.

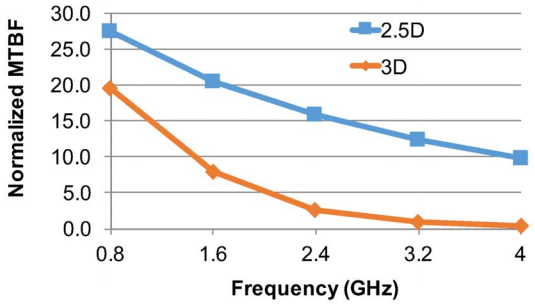
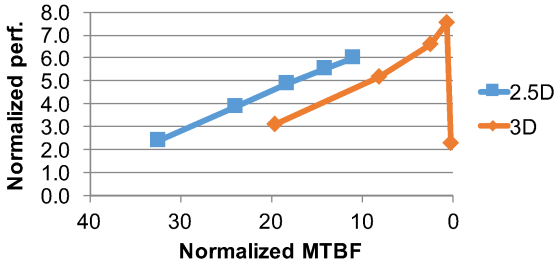


Figure 6. 2.5D and 3D lifetime at different operating frequencies. 3D lifetime degrades rapidly with frequency.

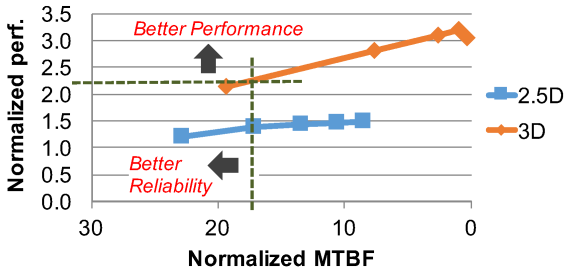
Using the thermal-reliability relation discussed above, figure 6 shows the impact of increasing core frequency on DRAM lifetime reliability. The result is an average of multiple PARSEC/SPLASH benchmarks. Since the estimated operating temperatures of 3D stacks are greater than 2.5D stacks, 3D DRAMs exhibits much lower lifetime. Furthermore, 3D lifetime degrades quickly with increased frequency, which suggests that 3D systems should be operated at much lower frequency than their 2.5D counterparts.

V. Reliability-Performance Tradeoff

From the previous two sections, it can be concluded that 3D designs can increase frequencies for performance gains but reliability concerns prohibits sustained operation at such higher temperatures. This section explores the reliability performance tradeoff. Figure 7a and 7b plot the performance and reliability of 2.5D and 3D-stacked processors for representative compute- and memory-intensive applications. For compute-intensive applications, 2.5D designs are



(a) Compute-intensive application (Blackscholes)



(b) Memory-intensive application (Canneal)

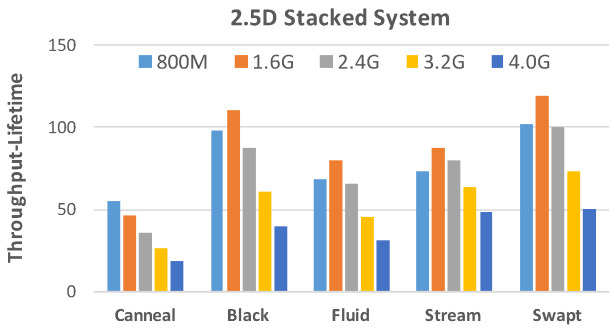
Figure 7. Performance and reliability tradeoff between 2.5D and 3D-stacked DRAM processors. Compute-intensive applications has better tradeoff with 2.5D while memory-intensive applications prefers 3D memory systems.

better. They not only have better reliability characteristics but also have higher performance than 3D at the same reliability design point. Performance can further be improved by operating at higher frequencies. For memory-intensive applications, on the other hand, 3D stacked designs are better based on their superior performance. Reducing frequency to improve lifetime reliability still achieves better performance relative to comparable 2.5D design, even at higher frequencies.

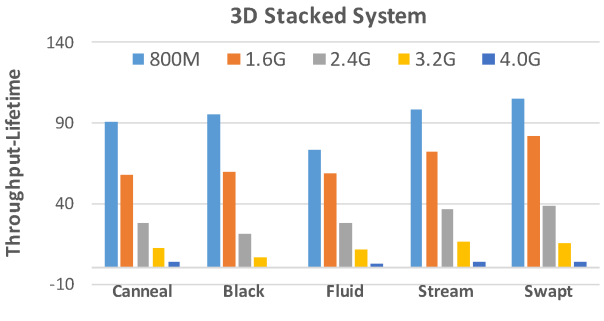
A. Throughput-Lifetime Product

Song et al. [7] suggested using a metric of *throughput-lifetime product* (TLP) to quantify reliability-performance tradeoffs. Figure 8a and 8b illustrate the TLP of various applications at different frequencies for 2.5D and 3D stacked memory systems, respectively. In general, TLP of 3D design is higher at low frequency, specially in the case of memory-intensive applications. However it decreases rapidly with increasing frequency due to lower lifetime reliability. On the other hand, the variation in TLP in a 2.5D stacked system is much lower, due to a much smaller variation in temperature with varying core frequencies. In fact, for most applications (except canneal), small increase in frequency increases the overall TLP, showing 2.5D stacks can be operated at higher frequencies. To conclude, the 3D design is suited for cases of memory-intensive applications operating at low frequency. For all other cases, the 2.5D system achieves better performance-reliability tradeoff.

Finally, figure 9 plots the TLP vs. lifetime of 2.5D and 3D-stacked processors average across all applications. The graph indicates that the optimal reliability-performance tradeoff is achieved by low-frequency operations of 3D design that yields high throughput with low operating temperatures. However, the difference between the optimal TLP points of 2.5D and 3D designs is not significant. Choice of 2.5D or 3D design depends on performance or relia-



(a) TLP with 2.5D Memory



(b) TLP with 3D Memory

Figure 8. Throughput-lifetime product of various applications with various frequencies. 2.5D stacked systems exhibit much lower variation in TLP with changing frequency as compared to their 3D counterpart.

bility requirements. If a processor is performance-constrained (or necessitates high throughput), 3D-stacked DRAM provides greater throughput than the 2.5D design. However, rigorous reliability requirements may favor the 2.5D implementation, where better reliability is traded with lower throughput.

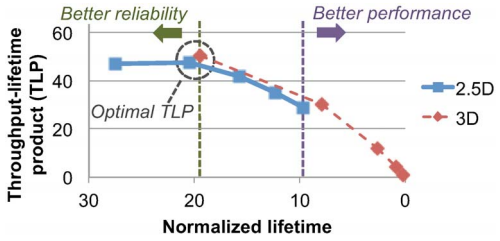


Figure 9. Throughput-lifetime product evaluation of 2.5D and 3D-stacked DRAM processors. Optimal reliability-performance is achieved by 3D systems operating at lower frequencies. However, the difference between 2.5D and 3D stacks is small.

VI. Summarizing the Observations and Future Work

The tradeoff between performance and reliability between 2.5D and 3D-stacked DRAM processors is an important design question. This paper analyzes this trade-off and identifies the type of applications and operating range where 2.5D vs. 3D stacked systems should be operated. The following summarizes key insights obtained from the analysis:

- *High-temperature operations, thus more frequent DRAM refreshes in 3D-stacked DRAM processors, have minor impact on overall performance because of increased channel- and bank-level parallelism available in such systems, allowing refreshes to be performed inherently in parallel with memory accesses.*
- *3D-stacked DRAM provides better performance than 2.5D for most applications, but it exhibits poorer lifetime due to higher operating temperature. This effect is significant in the case of compute-intensive applications.*
- *2.5D design is favored in the presence of rigorous reliability requirements.*
- *Compute-intensive applications have better performance-reliability trade-offs with 2.5D designs, even at higher frequencies, whereas memory-intensive applications favor 3D stacked systems operating at lower frequencies.*
- *Low voltage and frequency operation with 3D stacked systems may achieve balanced reliability-performance tradeoff.*

We conclude by pointing out the fact that this analysis used a tiled 3D floorplan with all cores active at the same time. This means both the core and the memory dies have a relatively uniform power density across all the tiles. This reduces the variations in the thermal profile of the overall system. In systems where the core floorplan is non-uniform (e.g., systems with large cache space or with asymmetric and heterogeneous systems), there will be a higher thermal variation across various parts/channels of the 3D memory. The DRAM channels above the hotter region will have higher refresh rate while the DRAM channels above the cooler regions will require fewer refreshes. It is expected that in such a scenario, an application will simultaneously suffer high bandwidth loss from some DRAM channels while low bandwidth loss from other channels.

The analysis can further be extended by running applications with different characteristics simultaneously. In such cases, the overall temperature will be higher due to compute-intensive applications while the memory bandwidth requirement will also be higher due to the memory-intensive applications, thus incurring both reliability and performance penalties simultaneously. We plan to extend the current work to these domains in future.

Acknowledgments

This research was supported in part by the National Science Foundation under grant CNS 0855110, Sandia National Laboratories, and the Defense Advanced Research Projects Agency (DARPA) contract HR0011-14-1-0002. We also acknowledge the detailed and constructive comments of the reviewers.

References

[1] C. Bienia, S. Kumar, J. Singh, and K. Li, "The PARSEC benchmark suite: characterization and architectural implications," *PACT*, *Int. Conf. Parallel Archit. Compil. Tech.*, pp. 72-81, Oct. 2008.

[2] N. Chatterjee, R. Balasubramonian, M. Shevgoor, S. Pugsley, A. Udipi, A. Shafiee, K. Sudan, M. Awasthi, and Z. Chishti, "USIMM: the Utah SImulated Memory Module," *JWAC*, *JILP Workshop Comput. Archit. Competit.*, Feb. 2012.

[3] S. Li, J. Ahn, R. Strong, J. Brockman, D. Tullsen, and N. Jouppi, "McPAT: Integrated power, area, timing modeling framework for multicore architectures," *MICRO*, *Int. Symp. Microarchit.*, pp. 469-480, Dec. 2009.

[4] Micron Technology, "Uprating semiconductors for high-temperature applications," Micron Technical Note TN-00-18, pp. 1-14, 2004.

[5] M. Radulovic, D. Zivanovic, D. Ruiz, B. Supinski, S. McKee, P. Radojkovic, and E. Ayguade, "Another trip to the wall: how much will stacked DRAM benefit HPC?" *MemSys*, *Int. Symp. Memory Syst.*, Oct. 2015.

[6] M. Sadri, M. Jung, C. Weis, N. Wehn, and L. Benini, "Energy optimization in 3D MPSoCs with wide-I/O DRAM using temperature variation aware bank-wise refresh," *DATE*, *Design, Autom. Test Europe Conf. Exhibit.*, pp. 1-4, Mar. 2014.

[7] W. Song, S. Mukhopadhyay, and S. Yalamanchili, "Managing performance-reliability tradeoffs in multicore processors," *IRPS*, *IEEE Int. Reliability Physics Symp.*, pp. 3C.1.1-3C.1.7 Apr. 2015.

[8] Z. Wan, Y. Kim, and Y. Joshi, "Compact modeling of 3D-stacked die inter-tier microfluidic cooling under non-uniform heat flux," *ASME*, *Int. Mech. Eng. Congr. Exhib.*, pp. 911-917, Sep. 2012.

[9] W. Wang, J. Beu, R. Bheda, T. Conte, Z. Dong, C. Kersey, M. Rasquinha, G. Riley, W. Song, H. Xiao, P. Xu, and S. Yalamanchili, "Manifold: a parallel simulation framework for multicore systems," *ISPASS*, *IEEE Int. Symp. on Perform. Anal. Syst. Softw.*, pp. 106-115, Mar. 2014.

[10] Z. Wan, Y. J. Kim, and Y. K. Joshi, "Compact modeling of 3d stacked die inter-tier microfluidic cooling under non-uniform heat flux," in *ASME 2012 International Mechanical Engineering Congress and Exposition*. American Society of Mechanical Engineers, 2012, pp. 911–917.

[11] L. Zheng, Y. Zhang, and M. Bakir, "A silicon interposer platform utilizing microfluidic cooling for high-performance computing systems," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 5, no. 10, pp. 1379–1386, Oct 2015.

[12] *MT41J128M8BY-18E*, Micron Technology. Available: http://download.micron.com/pdf/datasheets/dram/ddr3/1Gb_20DDR320_SDRAM.pdf

[13] M. Sadri *et al.*, "Energy optimization in 3d mpsocs with wide-i/o dram using temperature variation aware bank-wise refresh," in *Proceedings of the conference on Design, Automation & Test in Europe*. European Design and Automation Association, 2014, p. 281.

[14] Matthias Jung, der Zulian, Deepak M. Mathew, Matthias Herrmann, Christian Brugger, Christian Weis, and Norbert Wehn, "Omitting Refresh: A Case Study for Commodity and Wide I/O DRAMs," in *Proceedings of the 2015 International Symposium on Memory Systems (MEMSYS '15)*. ACM, New York, NY, USA, 85-91.