Energy Optimization in 3D MPSoCs with Wide-I/O DRAM Using Temperature Variation Aware Bank-wise Refresh

Mohammadsadegh Sadri*, Matthias Jung†, Christian Weis†, Norbert Wehn† and Luca Benini*‡
*DEI, University of Bologna, Italy. Email: {mohammadsadegh.sadr2,luca.benini}@unibo.it
†Microelectronic Systems Design Research Group, TU Kaiserslautern, Germany. Email: {jungma,weis,wehn}@eit.uni-kl.de
‡Department of Information Technology and Electrical Engineering, ETHZ, Switzerland. Email: lbenini@iis.ee.ethz.ch

***Abstract*—Heterogeneous 3D integrated systems with Wide-I/O DRAMs are a promising solution to squeeze more functionality and storage bits into an ever decreasing volume. Unfortunately, with 3D stacking, the challenges of high power densities and thermal dissipation are exacerbated. We improve DRAM refresh power by considering the lateral and vertical temperature variations in the 3D structure and adapting the per-DRAM-bank refresh period accordingly. In order to provide proof of our concepts we develop an advanced virtual platform which models the performance, power, and thermal behavior of a 3D-integrated MPSoC with Wide-I/O DRAMs in detail. On this platform we run the Android OS with real-world benchmarks to quantify the advantages of our ideas. We show improvements of 16% in DRAM refresh power due to temperature variation aware bank-wise refresh. Furthermore, two solutions are investigated to speedup system simulations: (1) Adaptive tuning of sampling intervals based on the estimated chip thermal profile, which results in speedups of $2X$. (2) Hardware acceleration of thermal simulations using the Maxeler engine, which shows possible speedups of $12X$.**

I. Introduction

Energy and thermal dissipation are limiting the efficiency of today's green computing solutions. More than 40% of the system energy in existing platforms is consumed by DRAMs [1]. 3D packaging of systems starts to break down the memory and bandwidth walls. However, this comes at the price of increased power density and less horizontal heat removal capability of the thinned dies. The thermal issues of 3D ICs cannot be solved by tweaking the technology and circuits alone. In fact, a 3D stacked SoC aggravates the thermal crisis and forces enhancements in the architecture and memory organization. Early detection of architectural shortcomings and thermal hazards is crucial to the design of sub-20-nm 3D chips. For instance, current microprocessor architectures are inefficient for running datacenter workloads mainly because of the mismatch between the workload characteristics and the organization of the memory subsystem. Consequently, the detailed analysis of the memory subsystem is very important as it unveils possible bottlenecks and issues which impact the system energy and efficiency.

Therefore we perform an in-depth study on performance, timing, power and leakage of Wide-I/O DRAMs and track its key parameters with temperature change. Through careful evaluation of temperature distributions in a 3D IC with Wide-I/O DRAMs, we propose architectural enhancements for the DRAM subsystem which improve energy consumption. We consider the lateral and vertical variation in temperature of the 3D DRAM dies and refresh each of the DRAM banks at a separate rate according to its own temperature. In order to assess and quantify the advantages of our proposed ideas, we build a suitable virtual infrastructure which considers all key characteristics of a 3D MPSoC with Wide-I/O DRAMs in detail. Our virtual platform uses *Transaction Level Models* (TLM), since they are well suited for fast system-level simulation and exploration of designs. The analysis of the memory subsystem requires a timing accurate behavior of the CPU cores. The *gem5* [2] architecture level full-system simulator is selected for this purpose, since it models system operations in detail and generates realistic traces of memory accesses, which can be replayed very fast inside the TLM environment.

978-3-9815370-2-4/DATE14/ © 2014 EDAA

To better understand the contributions of this work, we first illustrate the developed TLM environment (Section III), and the DRAM (Section III-A), CPU (Section III-B) and thermal (Section III-C) models. Then we discuss the temperature variation aware management of refresh rates for the 3D Wide-I/O DRAM (Section III-D). The adaptive sampling method to improve simulation speed is described in Section III-E.

II. Related Work

Two integrated performance, power and thermal modeling infrastructures for evaluation of energy and thermal management policies are presented in [3] and [4]. Both of the tool-sets mainly integrate a group of simulation software (*gem5*, *DRAMSim2*, *Hotspot* and etc.) to perform performance, power and thermal modeling. The advantages of our infrastructure over these tool-sets are the following:

1) We use TLM models to integrate different functional units of MPSoC together. The TLM environment provides us a high level of flexibility in creating various hardware structures and performing simulations.
2) Our infrastructure is tailored to the simulation of 3D-integrated Wide-I/O DRAM memory systems and tracks the timing and power of Wide-I/O DRAMs with the changes in temperature. This is contrary to other papers which rely on common DRAM models (e.g. DDR3) to estimate the behavior of a 3D stacked DRAM component.
3) Prior publications use adapted versions of 2D thermal simulators to perform 3D thermal estimations. Unlike, we use the 3D-ICE [5] thermal simulator which is inherently designed for 3D chips.
4) To mimic the workloads executed by today's mobile phones, we run Android OS and a set of real-world benchmarks on our multi-core ARM MPSoC. This is opposed to papers, which run benchmarks in the bare-metal or at most Linux environment.

A 3D MPSoC with Wide-I/O DRAM is presented in [6]. The 3D-IC features thermal sensors which can be used for online monitoring of temperature and tuning thermal models as well. The floorplan developed in this paper for the thermal model is indeed inspired by this work. An exploration of 3D DRAM architecture which results in a highly efficient Wide-I/O DRAM subsystem is presented in [7]. The work however, focuses on architectural issues only and does not consider the effect of temperature variation on the performance metrics of the Wide-I/O DRAM.

A well-known and often used power model for DRAM is provided by Micron [8]. This model has certain limitations. First, Micron uses the minimal timing constraints from the

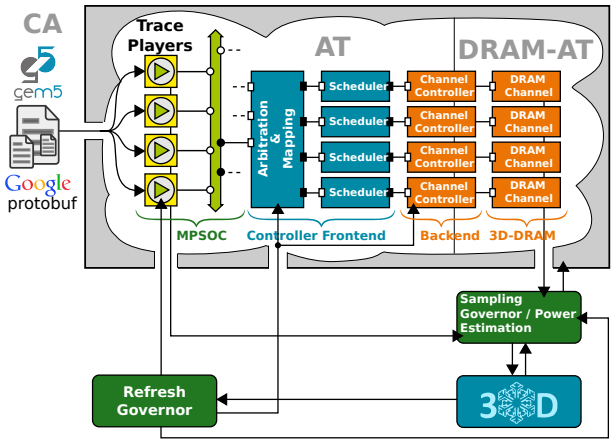

Fig. 1: Virtual Platform with Thermal Control Loop

data sheet specifications instead of the actual timings. Second, Micron assumes that the controller uses a close-page policy. An improved version of this power model was presented in [9], which uses actual timings from transactions. We use an enhanced version of this power model in our design [10]. The refresh rate of a DRAM device depends on its leakiest cells. However, the number of low retention time cells is relatively small compared to the total number of cells in a DRAM. A method for reducing the total refresh rate by grouping the DRAM rows into different retention time bins and applying different refresh rates on them is presented in [11]. We extend these ideas by studying the relationship between refresh rate of a DRAM bank and its temperature. We show that due to the lateral and vertical temperature variations in 3D MPSoCs, it is not necessary to refresh all banks of a DRAM channel at a similar rate.

III. The ESL Virtual Infrastructure

The developed infrastructure which is shown in Figure 1 contains five major parts:

- The *gem5* Environment: it models the operation of ARM CPU cores. *gem5* is configured to capture complete DRAM access traces (after L2 cache). Further descriptions come in Section III-B.
- The TLM Environment: it models the entire MPSoC. It contains the TLM models for the DRAM memory system (Section III-A) as well as *gem5* trace players which resemble the CPU cores in the TLM environment.
- Power models: receive the performance statistics of CPU cores and DRAMs and estimate their power. Sections III-B and III-A discuss these in more detail.
- Thermal model: calculates the thermal profile of the chip (Section III-C).
- Governors: are two different units to govern the sampling interval of simulation; t_{sim} (Section III-E) and the refresh rate of DRAM (Section III-D).

Considering the above descriptions, Figure 2 shows the detailed procedure for one simulation step. In the following, we

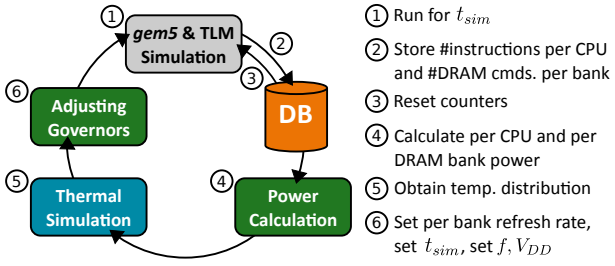

Fig. 2: Flow of one Co-Simulation step

TABLE I: REFRESH RATES OF DRAM BANKS VS. TEMPERATURE

Temp (°C)	35	45	65	85	87.5	88.75	90
Refresh Rate (ms)	192	128	96	64	56	52	48

Temp (°C)	91.25	92.5	95	100	105	bigger than 105
Ref Rate (ms)	44	40	32	24	16	8

describe the important parts of the infrastructure in detail.

A. DRAM Model

The 3D-DRAM memory subsystem used in our platform, consists of a controller frontend, a channel controller and a Wide I/O DRAM model. It is shown in Figure 1. The standard TLM AT protocol is extended with a DRAM specific protocol (DRAM-AT) [12] to provide very fast simulation speeds with high timing accuracy.

We improve the DRAM power model of [9] to create our DRAM power model for the TLM platform [10]. For this work we extended the model of the DRAM controller to support in addition to the auto-refresh command `REFA`, separate refreshes per bank or groups of banks: `REFB`. The DRAM power model accepts per-bank activity statistics and generates per-bank power values at each simulation step. This functionality is later exploited by the refresh governor of the MPSoC to assign different refresh rates to the different banks of a DRAM channel according to their temperature. Table I represents the relationship between temperature and refresh rates of a DRAM bank (50 nm technology). The refresh rates in this table are calculated based on [13]. Table I is built taking into account that commercial DRAM products exploit temperature sensors with high level of accuracy (± 1 °C) around the calibrated junction temperature range (90 °C).

For a detailed description of the DRAM power model and its validation methodology and results, please refer to [9] [7].

Our MPSoC contains 4 channels of Wide-I/O DRAM. Each channel consists of 8 banks and spans 4 DRAM dies. Thus each DRAM die, contains 2 banks per DRAM channel, see Figure 3 (Banks 0 and 1 belong to Mem Die 1, banks 2 and 3 belong to Mem Die 2, etc.). Figure 4 (a) shows the placement of DRAM banks in each DRAM die. The density of each DRAM die is 2 Gbit. The physical dimensions of DRAM banks and the silicon die are calculated based on 50 nm technology [14]. Section III-C describes the physical properties of the chip in more detail.

B. CPU Model

Our CPU model consists of two distinct parts: the *gem5* simulator and TLM *gem5* trace players. Overall, we first run our benchmarks on the *gem5* simulator and record detailed traces of DRAM accesses (after L2 cache). Inside the TLM environment, we build a complete virtual MPSoC which contains the *gem5* trace players to represent the CPU cores and other necessary glue logic. We then re-play the recorded traces inside the TLM environment to perform different evaluations on temperature, power and energy consumption of the MPSoC. The timings of the traces will be dynamically adjusted at the time of play-back inside the TLM environment according to the timings and latencies introduced by the DRAM models.

gem5 is configured for the ARM ISA and it is run in detailed, out-of-order mode. We run Android 2.3 Gingerbread on this platform and use three real-world well-known benchmarks to stress the CPUs and the memory. Our selected benchmarks are: *AndEBench*, *0xBench*, and *SmartBench*.

C. Thermal Model

We build our thermal model using the 3D-ICE [5] thermal simulator which supports definition of layers with non-homogeneous materials in the chip stack. To create a complete realistic thermal model, real-world numbers are used to define

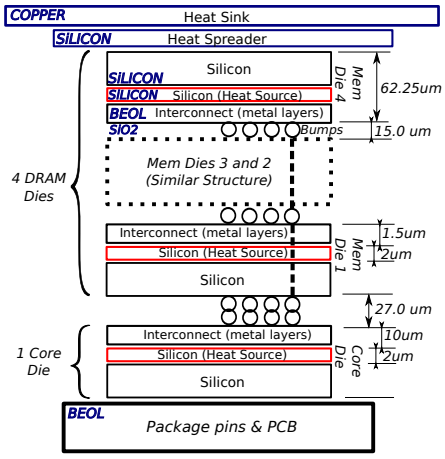

Fig. 3: 3D stack-up of the virtual MPSoC

the key dimensions of the 3D structure [14] [15]. Figure 3 shows selected values for 3D stackup. The thermal floorplan of each silicon die which is indicative of the size and coordinates of on-die units is shown in Figure 4. As Figure 4 shows each silicon die is 8.6×7.4 mm^2. For thermal simulation, the chip is divided into equal sized cubes of 0.3 mm^3. Each transient thermal simulation is done with a duration of t_{sim}. The spatial resolution of the thermal model is fixed during the entire simulation.

D. Refresh Governor

At each simulation step the refresh governor receives the estimated temperatures and defines the refresh rates of DRAMs based on their maximum temperature (Table I).

We perform statistical analysis on the temperature profile of our 3D MPSoC, and we measure lateral and vertical variation in temperature in the 3D structure. For instance, with AndEBench, when all 8 CPU cores are running at 1.4 GHz, an averaged vertical temperature variation of 5.6 °C can be seen across 4 DRAM dies. In the first DRAM die, averaged lateral difference in temperature between two adjacent DRAM banks of a same channel is 3.3 °C. As Table I shows when the averaged DRAM die temperature is > 85 °C, the mentioned lateral and vertical temperature variations cause significant differences in the required refresh rate of each DRAM bank.

Due to these observations, we implemented the following key idea: instead of defining the refresh rate based on the maximum temperature seen across the entire channel and refreshing all DRAM banks at the same rate, we select the *refresh rate of each bank separately* based on its own maximum temperature. As described in III-A we have extended our DRAM subsystem model to support handling of separate per-bank refresh commands. As we will show in section IV-A this increases the overall refresh period (makes refreshes happen less frequently) and improves the energy consumption.

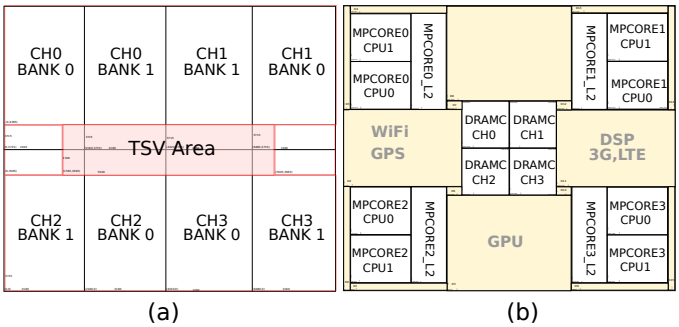

Fig. 4: Geometries used for thermal model: (a) DRAM die (b) Core die.

TABLE II: t_{sim} VS. MAXIMUM CHIP TEMPERATURE

Maximum Temperature	40	60	70	80	90	100
Sampling Interval (ms)	32	16	8	4	2	1

TABLE III: BANK-WISE REFRESH VS NORMAL METHOD. THERMAL CONTROLLER OFF, CPU CORES @1.25 GHZ, ALL VALUES ARE AVERAGED.

Benchmark	AndEBench		0xBench		SmartBench	
Bank-wise Refresh	Off	On	Off	On	Off	On
Refresh Period (ms)	26.27	30.57	29.06	37.18	33.81	43.68
REF Power (W)	0.0297	0.0262	0.0221	0.0177	0.0181	0.0149
DRAM Power (W)	0.0842	0.0820	0.160	0.1586	0.1247	0.1232
IPC	0.5176	0.5236	0.5436	0.5515	0.4690	0.4733

E. Sampling Governor

The t_{sim} parameter affects the speed of the simulation significantly. It determines how frequently the thermal profile of the chip should be estimated and governor routines should be invoked. Basically, a high resolution sampling in time is only needed when the temperature values for silicon dies are near critical[1] regions. For each sampling interval, if T_{max} is the maximum temperature of the entire 3D structure, and T_{TH} is the defined thermal hazard threshold, t_{sim} will be set to its smallest value when T_{max} approaches T_{TH}. t_{sim} grows gradually as the distance between T_{max} and T_{TH} increases.

Without loss of generality, we select $min(t_{sim})$ equal to $t_{sim}^{min} = 1$ ms. This sampling period is small enough to show the effectiveness of our closed loop management ideas completely. Table II shows selected sampling intervals for each temperature range. For $T < 40$ °C, t_{sim} of 32.0 ms and for $80 < T_{max} < 90$ °C, t_{sim} of 2.0 ms are chosen. For temperature values above 90.0 °C, t_{sim} is 1 ms.

IV. Experimental Results

We conduct a set of experiments to demonstrate the advantages of the previously described contributions. First, the temperature variation aware bank-wise refresh is presented. Then, the speed-up gained by adaptive sampling is quantified.

A. Temperature Variation Aware Bank-wise Refresh

We execute two sets of simulations; In the first set the bank-wise refresh is disabled. For each DRAM channel, the refresh governor finds the maximum temperature of the channel. Then suitable refresh period will be selected (Table I) and used for all DRAM banks of the channel. In the second set, the bank-wise refresh is active (Section III-D). We perform the test while the thermal controller is off and all CPU cores are playing their own respective traces at 1.25 GHz. DRAMs are running at 200

[1] e.g. temperatures in which the refresh interval of the DRAM should change frequently, or defined thresholds for the thermal controller units.

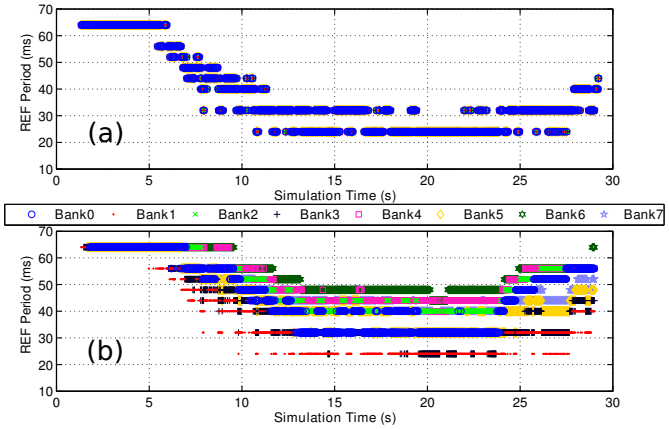

Fig. 5: Refresh periods of Channel 3, (a) Bank-wise off (b) Bank-wise on (SmartBench).

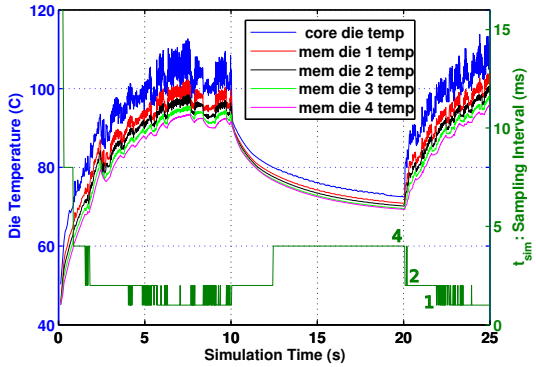

Fig. 6: Change in sampling interval relative to temperature

MHz and ambient temperature (T_{amb}) is fixed at 45 °C.

Figure 5 shows the refresh periods for 8 banks of the first DRAM channel. When the bank-wise refresh is off, refresh periods are equal for all 8 banks. When bank-wise refresh is active, the banks located on the lower dies have smaller refresh periods compared to colder banks at higher dies. For example, at $t = 20$ s all of the banks are refreshed with a period of 24 ms when the bank-wise refresh is off. However, when the bank-wise refresh is active, only the bank 0 and bank 3 are refreshed at this period and other banks are refreshed at higher periods. As we see in Table III the bank-wise refresh results in an average improvement of 24% in refresh rate, and 16.4% in averaged refresh power. In near future, half of the DRAM power will be related to refresh [11]. Thus, the proposed idea can significantly improve the total energy consumption.

B. Adaptive Sampling

Figure 6 shows the dependency of t_{sim} to the die temperature for a test in which we turn off and on all CPU cores at $t = 10$ s and $t = 20$ s, respectively. The x axis is time, the left y axis is temperature and the right y axis is t_{sim}.

We quantify the speedup gained by adaptive sampling and ensure that the simulation accuracy remains at acceptable levels. We perform two tests: first, we execute the simulation task with the smallest fixed $t_{sim} = 1$ ms. Then, we re-run the same simulation and allow t_{sim} to change adaptively. The thermal governor is off during these tests and CPU cores are running *AndEBench* benchmark at 1.25 GHz. For each test we record the total execution time of the simulation and complete history of temperatures. The accuracy of the proposed adaptive method is acceptable, if it reports similar *critical* temperature values at similar points in time as the fixed method. We refer to the fixed trace containing temperatures of the core die and for each point at time with temperature $T > T_{crit}$ we look-up the temperature of the identical point in adaptive method and calculate the difference. The execution time of the simulation is 1.46×10^4 and 6.6×10^3 seconds for fixed and adaptive methods respectively. This is a speedup of $2.21X$. Averaged t_{sim} for adaptive method is 1.99 ms. For $T_{crit} = 90$ °C, the maximum temperature difference between identical points in adaptive and fixed traces is 0.87 °C.

C. Hardware Acceleration of Thermal Simulation

Thermal simulation is very computational intensive. Based on our measurements, more than 65% (with adaptive sampling) to 90% (fixed sampling) of the simulation time is dedicated to thermal simulation. We study the feasibility of accelerating the execution of 3D-ICE with the aid of a specialized hardware. Using the OProfile [16] performance analyzer, we first identify the *execution time hotspots* of the 3D-ICE. Our investigations using OProfile show that the tool is running the CBLAS DGEMV (matrix-vector multiplication) function during more than 90% of its execution time. We select the Maxeler [17] hardware acceleration engine (featuring one Xilinx Virtex6-SX475T FPGA) as target platform and adapt the 3D-ICE source code to the Maxeler development flow. This allows us to off-load any computational intensive part of the software to the hardware containing our computational kernels. To obtain a measure of feasible level of speedup by the Maxeler hardware, we focused on hardware acceleration of the key routine: *DGEMV*. We implemented the computational kernel using *maxj* [17] and validated its functionality in practice. A speedup of more than $12X$ is seen for execution of DGEMV in comparison with one core of Intel i7-860 (at 2.8 GHz) while the FPGA is running at 150 MHz and 75% of its hardware multipliers and less than half of its logic slices are consumed.

V. Conclusion

We demonstrated the temperature variation aware bank-wise refresh which improves the overall refresh rate of the system effectively and decreases the power consumption of Wide-I/O DRAMs. To prove our ideas, we presented a virtual infrastructure featuring a TLM environment and detailed power and thermal models for the 3D chip. To speedup simulations, we devised a method to tune the sampling interval of simulation adaptively according to the thermal profile of the chip. We also studied the feasibility of hardware acceleration of the thermal simulation using the Maxeler engine.

Acknowledgment

This work is supported, in parts, by the EU FP7 ERC Project MULTITHERMAN (GA n. 291125).

References

[1] J. Lin, et al. *Thermal Modeling and Management of DRAM Systems*. IEEE Trans. on Computers., 2013.

[2] N. Binkert, et al. *The gem5 Simulator*. SIGARCH Comput. Archit. News, 39, 2011.

[3] A. Rodrigues, et al. *Improvements to the structural simulation toolkit*. In Proc. of SIMUTOOLS 2012.

[4] J. Meng, et al. *Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints*. In Proc. of DAC 2012.

[5] A. Sridhar, et al. *3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling*. In Proc. of ICCAD 2010.

[6] D. Dutoit et al. *A 0.9 pJ/bit, 12.8 GByte/s WideIO memory interface in a 3D-IC NoC-based MPSoC*. In VLSI Technology (VLSIT), 2013 Symposium on, pages C22–C23, 2013.

[7] C. Weis, et al. *Exploration and Optimization of 3-D Integrated DRAM Subsystems*. IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems, 2013.

[8] Micron Technology Inc. *TN-41-01: Calculating Memory System Power for DDR3*. Technical report, 2007.

[9] K. Chandrasekar, et al. *System and Circuit Level Power Modeling of Energy-Efficient 3D-Stacked Wide I/O DRAMs*. In Proc. of DATE 2013.

[10] M. Jung, et al. *Power Modelling of 3D-Stacked Memories with TLM2.0 Based Virtual Platforms*. In Proc. of SNUG 2013.

[11] J. Liu, et al. *RAIDR: Retention-Aware Intelligent DRAM Refresh*. In Proc. of ISCA 2012.

[12] M. Jung, et al. *TLM Modelling of 3D Stacked Wide I/O DRAM Subsystems: A Virtual Platform for Memory Controller Design Space Exploration*. In Proc. of RAPIDO 2013.

[13] Micron Technology Inc. *4Gb: x16, x32 Mobile LPDDR3 SDRAM, http://www.micron.com/products/dram/mobile-lpdram*, July 2013.

[14] 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 Journal of Solid-State Circuits, 47, 2012.

[15] Y. Temiz, et al. *A CMOS-compatible chip-to-chip 3D integration platform*. In Proc. of ECTC 2012, 2012.

[16] J. Levon, et al. *Oprofile: A system profiler for Linux, http://oprofile.sourceforge.net*, 2013.

[17] O. Mencer. *Maximum performance computing for exascale applications*. In Proc. of SAMOS 2012, 2012.