CACTI-IO: CACTI With Off-Chip Power-Area-Timing Models

Norman P. Jouppi[3], Andrew B. Kahng[1,2], Naveen Muralimanohar[3] and Vaishnav Srinivas[1]
UC San Diego [1]ECE and [2]CSE Departments, La Jolla, CA. {abk, vaishnav}@ucsd.edu
[3]HP Labs, Palo Alto, CA. {norm.jouppi, naveen.muralimanohar}@hp.com

***Abstract*—We describe CACTI-IO, an extension to CACTI [4] that includes power, area and timing models for the IO and PHY of the off-chip memory interface for various server and mobile configurations. CACTI-IO enables design space exploration of the off-chip IO along with the DRAM and cache parameters. We describe the models added and three case studies that use CACTI-IO to study the tradeoffs between memory capacity, bandwidth and power.**

The case studies show that CACTI-IO helps (i) provide IO power numbers that can be fed into a system simulator for accurate power calculations, (ii) optimize off-chip configurations including the bus width, number of ranks, memory data width and off-chip bus frequency, especially for novel buffer-based topologies, and (iii) enable architects to quickly explore new interconnect technologies, including 3-D interconnect. We find that buffers on board and 3-D technologies offer an attractive design space involving power, bandwidth and capacity when appropriate interconnect parameters are deployed.

***Keywords:* CACTI, DRAM, IO, memory interface, power and timing models.**

I. Introduction

The interface to the DRAM, including the PHY, I/O circuit (IO) and interconnect, is becoming increasingly important for the performance and power of the memory subsystem [15], [16], [17], [25], [31], [37]. As capacities scale faster than memory densities [7], there is an ever-increasing need to support a larger number of memory dies, especially for high-end server systems [29], often raising cooling costs. Mobile systems can afford to use multi-chip package (MCP) or stacked-die point-to-point memory configurations; by contrast, servers have traditionally relied on a dual-inline memory module (DIMM) to support larger capacities. With modern server memory sizes exceeding 1 TB, the contribution of memory power can reach 30-57% of total server power [37], with a sizable fraction (up to 50% in some systems) coming from the off-chip interconnect. The memory interface incurs performance bottlenecks due to challenges with interface bandwidth and latency. The bandwidth of the interface is limited by (i) the data rate, owing to the DRAM interface timing closure, signal integrity over the interconnect, and limitations of source-synchronous signaling [3], [41], and (ii) the width of the bus, which is often limited by size and the cost of package pins.

CACTI [4] is an analytical memory modeling tool which can calculate delay, power, area and cycle time for various memory technologies. For a given set of input parameters, the tool performs a detailed design space exploration across different array organizations and on-chip interconnects, and outputs a design that meets the input constraints. CACTI-D [19] is an extension of CACTI that models the on-chip portion of the DRAM (Dynamic Random Access Memory).

In this paper we describe CACTI-IO, an extension to CACTI, illustrated in Figure 1. CACTI-IO allows the user to describe the configuration(s) of interest, including the capacity and organization of the memory dies, target bandwidth, and interconnect parameters. CACTI-IO includes analytical models for the interface power, including suitable lookup tables for some of the analog components in the PHY. It also includes voltage and timing uncertainty models that help relate parameters that affect power and timing. Voltage and timing budgets are traditionally used by interface designers to begin building components of the interface [1], [3], [34], [42] and budget the eye diagram between the DRAM, interconnect, and the controller as shown in Figure 2. The *Eye Mask* represents the portion of the eye budgeted for the *Rx* (receiver). The setup/hold slacks and noise margins represent the budgets for the interconnect and the *Tx* (transmitter).

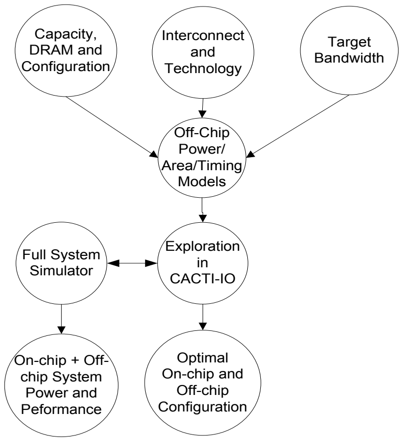

Capacity, DRAM and Configuration
Interconnect and Technology
Target Bandwidth
Off-Chip Power/ Area/Timing Models
Full System Simulator
Exploration in CACTI-IO
On-chip + Off-chip System Power and Peformance
Optimal On-chip and Off-chip Configuration

Fig. 1. CACTI-IO: Off-chip modeling and exploration within CACTI.

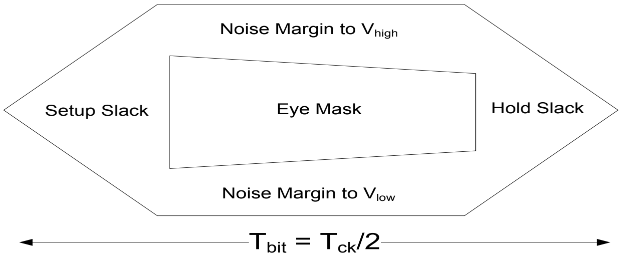


Fig. 2. Memory interface eye diagram for voltage and noise budgets.

Final optimization of the IO circuit, off-chip configuration and signaling parameters requires detailed design of circuits along with SPICE analysis, including detailed signal integrity and power integrity analyses; this can take months for a new design [3]. CACTI-IO is not a substitute for detailed analyses, but rather serves as a quick estimate for the system architect to enable the right tradeoffs between the large number of non-trivial IO and off-chip parameters. Up-front identification of the off-chip design space at an architectural level is crucial for driving next-generation memory interface design.

The main objectives for the CACTI-IO tool are as follows.

IEEE/ACM International Conference on Computer-Aided Design (ICCAD) 2012, November 5-8, 2012, San Jose, California, USA

(1) Obtain IO power numbers for different topologies and modes of operation that can be fed into a full-system simulator. The tradeoffs between performance, power and capacity in the memory subsystem are non-trivial [14], [19], but previous studies often do not explore alternatives for the memory interface to a standard DDR3 configuration. Furthermore, most simulators, including McPAT [18] and DRAMSIM [28], do not model the interface power and timing, and have no visibility into the details of the PHY and IO. CACTI-IO provides IO power numbers for Read, Write, Idle (only clock active) and Sleep modes that can easily be integrated into a system simulator. This enables architects to see the most significant on-chip and off-chip sources of power across modes.

(2) Enable co-optimization of off-chip and on-chip power and performance, especially for new off-chip topologies. Historically, off-chip parameters (i.e., signaling properties and circuit parameters) have been limited to standardized configurations including DIMMs, with operating voltage, frequency, data rates and IO parameters strictly governed by standards. A major drawback and design limiter – especially when operating at high frequencies – in this simplistic design context is the number of DIMMs that can be connected to a channel. This often limits memory capacity, creating a *memory wall*. Recent large enterprise servers and multicore processors instead use one or more intermediate buffers to expand capacity and alleviate signal integrity issues. Such a design still adheres to DRAM standards but has more flexibility with respect to the interconnect architecture that connects memory and compute modules, including serial interfaces between the buffer and the CPU. While current and future memory system capacity and performance greatly depend on various IO choices, to date there is no systematic way to identify the optimal off-chip topology that meets a specific design goal, including capacity and bandwidth. ***CACTI-IO provides a way for architects to systematically optimize IO choices in conjunction with the rest of the memory architecture.*** Below, we illustrate how CACTI-IO can help optimize a number of off-chip parameters – number of ranks (fanout on the data bus), memory data width, bus frequency, supply voltage, address bus fanout and bus width, – for given capacity and bandwidth requirements. CACTI-IO can also be used to evaluate the number of buffers needed in complex, high-end memory configurations, along with their associated overheads.

(3) Enable exploration of emerging memory technologies. With the advent of new interconnect and memory technologies, including 3-D TSS (through-silicon stacking) based interconnect being proposed for DRAM as well as new memory technologies such as MRAM (magnetic RAM) and PCRAM (phase-change RAM) [36], architects are exploring novel memory architectures involving special off-chip caches and write buffers to filter writes or reduce write overhead. Note that most emerging alternatives to DRAM suffer from high write energy or low write endurance. The use of additional buffers plays a critical role in such off-chip caches, and there is a need to explore the changing on-chip and off-chip design space. When designing new off-chip configurations, many new tradeoffs arise based on the choice of off-chip interconnect, termination type, number of fanouts, operating frequency and interface type (serial vs. parallel). CACTI-IO provides flexible baseline IO models that can be easily tailored to new technologies and used to explore tradeoffs at a system level.

In summary, the key contributions of this paper are:

- models for power, area and timing of the IO, PHY and interconnect for server and mobile configurations;
- CACTI-IO, an extension to CACTI that includes these models, thus enabling tradeoffs within the memory subsystem; and
- three industry-driven case studies that use CACTI-IO to optimize parameters of the off-chip topology, including the number of ranks, memory data width and address bus fanout.

In the remainder of this paper, Section II describes the interface models, including those for power, voltage margins, timing margins and area. Section III presents CACTI-IO using three case studies, showing a summary of the power and timing as well as optimal off-chip configurations. Section IV summarizes our conclusions.

II. IO, PHY and Interconnect Models

Power and timing models for interconnect and terminations have been well documented and validated over the years [1], [2], [6]. Complete details of the IO, PHY and interconnect models included in CACTI-IO and their validation are beyond the scope of this paper, but in this section we briefly summarize key details. Complete details about the models and their validation can be found in the CACTI-IO technical report [5].

As shown in [5], the models have been validated against SPICE simulations and measurements for several configurations used in the case studies below. These models scale with off-chip interconnect technology and on-die process technology. Our goal here is to show the framework of the baseline models, which can then be adapted and validated for any customized configuration needed, including new interconnect technologies. Further discussion of the portability to different technologies can be found in [5].

A. Power Models

Power is calculated for three different modes: Active (peak activity, Read and Write), Idle (no data activity, but clock is enabled and terminations are on), and Sleep (clock and terminations are disabled, in addition to no data activity). Based on the duty cycle spent among the Active (both Read and Write), Idle and Sleep modes, CACTI-IO projects the total IO power consumed. Our models include the following.

(1) Dynamic IO Power. The switching power at the load capacitances is described in Equation (1), where N_{pins} is the number of signal pins; D_c is the duty cycle of activity; α is the activity factor for the signal switching (number of 0 to 1 transitions per clock period, i.e. $\alpha = 1$ for a clock signal); i denotes various nodes along the interconnect, with possibly different swings in a terminated or low-swing scheme; C_{Total_i} is the capacitance at node i; V_{sw_i} is the swing of the signal at node i; V_{dd} is the supply voltage; and f is the frequency of operation.

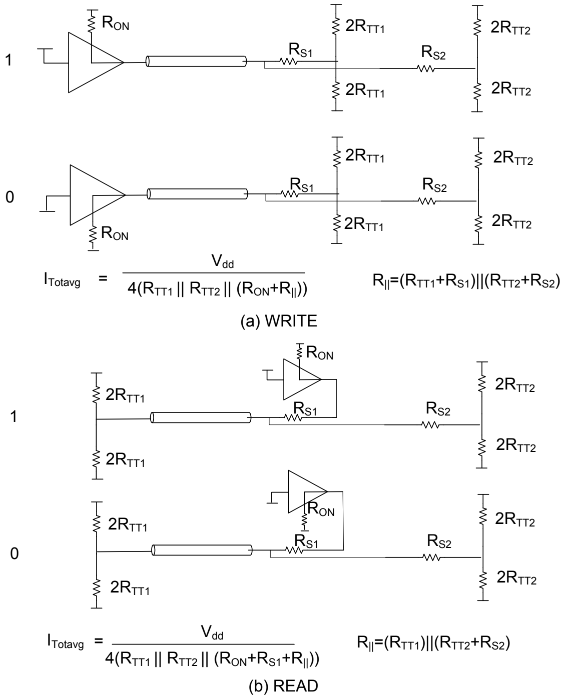
$$P_{dyn} = N_{pins} D_c \alpha (\sum_i C_{Total_i} V_{sw_i}) V_{dd} f \qquad (1)$$

(2) Interconnect power. The power dissipated on the interconnect ($P_{dyn\_interconnect}$) is given by Equation (2). The energy/bit consumed on the interconnect ($E_{bit}^{interconnect}$) is described in Equation (3), where Z_0 is the characteristic impedance of the line, t_L is the flight time (time taken for the signal to traverse the line length) and t_b is the bit period. For high-end servers, generally $2t_L > t_b$ since the interconnect is long, while for mobile configurations, generally $2t_L < t_b$. For an FR-4 based interconnect used on printed circuit boards, t_L is approximately 180 ps/inch. The interconnect is generally modeled as a transmission line (unlike an on-die RC network [2]) when $t_L > t_r/3$, where t_r is the rise time of the signal.

$$P_{dyn\_interconnect} = N_{pins} D_c \alpha E_{bit}^{interconnect} f \qquad (2)$$

$$E_{bit}^{interconnect} = \begin{cases} \frac{t_L V_{sw} V_{dd}}{Z_0} & \text{if } 2t_L \leq t_b \\ \frac{t_b V_{sw} V_{dd}}{Z_0} & \text{if } 2t_L > t_b \end{cases} \qquad (3)$$

(3) Termination Power. The IO termination power is provided for various termination options, including unterminated (as used in LPDDR2 and Wide-IO), center-tap (as used in DDR3), VDDQ (as in DDR4) and differential terminations (as used in M-XDR). The voltage swing set by the terminations is fed into the dynamic power equation described above in Equation (1). Terminations are used to improve signal integrity and achieve higher speeds, and the values depend on the interconnect length as well as the frequency or timing requirements. Terminations on the DQ (data) bus typically use an ODT (on-die termination) scheme, while those on the CA (command-address) bus use a fly-by termination scheme to the multiple loads. Figures 3 and 4 show the DDR3 DQ and CA



1
0
R_{ON}
$2R_{TT1}$
$2R_{TT2}$
R_{S1}
R_{S2}
$I_{Totavg} = \frac{V_{dd}}{4(R_{TT1} \| R_{TT2} \| (R_{ON}+R_{\|}))}$
$R_{\|}=(R_{TT1}+R_{S1})\|(R_{TT2}+R_{S2})$
(a) WRITE
$I_{Totavg} = \frac{V_{dd}}{4(R_{TT1} \| R_{TT2} \| (R_{ON}+R_{S1}+R_{\|}))}$
$R_{\|}=(R_{TT1})\|(R_{TT2}+R_{S2})$
(b) READ

Fig. 3. DDR3 DQ dual rank termination.

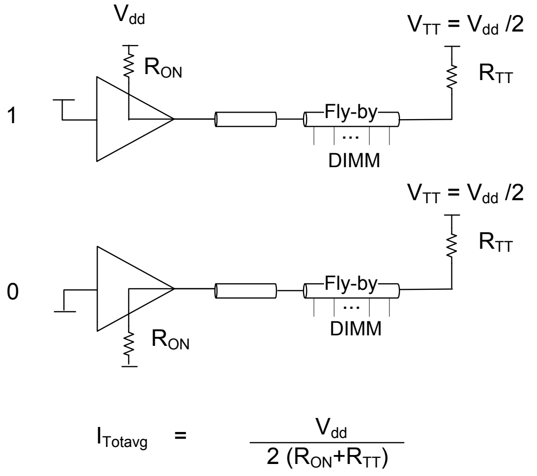


Fig. 4. DDR3 CA termination.

termination schemes along with the static current consumed by them as used in [20].

(4) PHY Power. The PHY includes analog and digital components used to retime the IO signals on the interface. A wide range of implementations exist for the PHY [15], [16], [17], [25], [26], [27], that vary in power and are fine-tuned to specific design requirements. Currently, the user can change the inputs for the PHY power based on a specific implementation. In the future, we propose to extend this framework to explore PHY optimization for system target metrics using analytical models for some of the building blocks in the PHY, rather than a lookup table. Tables I and II respectively show the active dynamic power per bit and static power of building blocks in an example PHY implementation for a x128 3-D configuration. The building blocks are representative of typical PHY components [15], [16], [17], [25], [26], [27]. The power breakdown is shown for three data rates (0.5, 1 and 2 Gbps) for the 3-D configuration to enable exploration of interface options in Section III. Table III shows exemplary values of dynamic and static power for a DDR3-1600 PHY; these are used for the case studies in Section III below. At lower data rates, certain components are not required, indicated by N/A in Tables I and II.

TABLE I
PHY ACTIVE DYNAMIC POWER /BIT FOR 3-D CONFIGURATIONS.

Building Block	Dynamic Power (mW/Gbps)		
	500 Mbps	1 Gbps	2 Gbps
Datapath	0.1	0.2	0.5
Phase Rotator	N/A	0.1	0.2
Clock Tree	0.1	0.2	0.4
Duty Cycle Correction	N/A	N/A	0.05
Deskewing	N/A	N/A	0.05
PLL	N/A	N/A	0.05

TABLE II
PHY STATIC POWER FOR A X128 3-D CONFIGURATION.

Building Block	Static Power (mW)		
	500 Mbps	1 Gbps	2 Gbps
Phase Rotator	N/A	1	10
PLL	N/A	N/A	10

TABLE III
PHY DYNAMIC /BIT AND STATIC POWER FOR A X64 DDR3-1600.

Building Block	Dynamic Power (mW/Gbps)	Static Power (mW)
Datapath	0.5	0
Phase Rotator	0.2	10
Clock Tree	0.8	0
Rx	0.2	20
Duty Cycle Correction	0.05	0
Deskewing	0.05	0
Write/Read Leveling	0.05	0
PLL	0.1	10

B. Voltage and Timing Margins

The minimum achievable clock period T_{ck} depends on the voltage and timing budgets (i.e., eye diagram and/or BER (bit error rate) compliance). Traditionally, the memory interface budgets have been based on the worst-case analysis approach shown in Figure 2, where the budgets are divided between the DRAM, the interconnect and the controller chip or SOC. With increasing speeds there is a need for a statistical analysis approach similar to serial links [35], [39] during detailed design analysis. However, for architectural exploration and relative tradeoffs, we continue to use worst-case budgets in our initial framework, with the option of accounting for optimism or pessimism based on measurements or prior correlation between the two approaches. This correlation factor also helps address different BER requirements for server DIMM modules that include error correction (ECC) schemes [3], [29], [32].

(1) Timing budgets. The key interface timing equations are based on DRAM AC timing parameters in the JEDEC specification [21], [22]. There are nuances to the system timing based on the controller design and clocking architecture, but most rely on measuring setup and hold slacks to ensure positive margins.

It is interesting to note that while the DQ bus is DDR in almost all DRAMs today, the CA bus is mostly SDR (single data rate), except for LPDDR2 and LPDDR3 where the CA bus is DDR [21], [22]. In addition, the CA bus provides an option for 2T (two clock-cycles) and 3T (three clock-cycles) timing to relax the requirements when heavily loaded. This is done since the CA bus is typically shared across all memories in the DIMM.

The jitter on the interface is the true limiter of the timing budget, and optimizing the interface for low jitter is the key challenge. The common sources of jitter include Tx jitter, ISI (inter-symbol interference), crosstalk, SSO (simultaneously switching outputs), supply noise and Rx jitter [3].

Jitter can be estimated from various deterministic (DJ_i) and random (RJ_i) sources as follows [3]:

$$T_{jitter} = \sum_i DJ_i + \sqrt{\sum_i RJ_i^2} \qquad (4)$$

$$T_{jitter}(\mathbb{F}_0) = T_{jitter\_avg} + \sum_i (T_{jitter}(F_i = F_{i0}) - T_{jitter\_avg}) \qquad (5)$$

Here, factor F_i is a parameter that affects T_{jitter} [3]. $\mathbb{F}_0$ is the value of a set of factors $F_i = F_{i0}$ for which we calculate the jitter, $T_{jitter}(\mathbb{F}_0)$, as an estimate assuming there is no interaction between the factors F_i [3]. This is done efficiently by running a Design of Experiments (DOE) for a set of orthogonal array experiments as defined by the Taguchi method [3], [24]. T_{jitter_avg} represents the average jitter from all the experiments in the orthogonal array, while $T_{jitter}(F_i = F_{i0})$ represents the average jitter from all experiments where $F_i = F_{i0}$. For cases where F_{i0} is not part of the orthogonal array, a piecewise linear approximation is employed.

(2) Voltage Budgets. A voltage budget can be developed for voltage margins as follows [1], based on a worst-case analysis, where V_N is the voltage noise, K_N is the proportionality coefficient for the proportional noise sources (that are proportional to the signal swing V_{sw}), V_{NI} is the noise due to independent noise sources and V_M is the voltage margin. Crosstalk, ISI and SSO are typical proportional noise sources [1], while the Rx-offset, sensitivity and independent supply noise are typical independent noise sources.

$$V_N = K_N \cdot V_{sw} + V_{NI} \qquad (6)$$

$$K_N = K_{xtalk} + K_{ISI} + K_{SSO} \qquad (7)$$

$$V_{NI} = V_{Rx-offset} + V_{Rx-sens} + V_{supply} \qquad (8)$$

$$V_M = \frac{V_{sw}}{2} - V_N \qquad (9)$$

A DOE analysis for the voltage noise coefficient, K_N, can be performed in a similar manner as described above for T_{jitter}.

C. Area Models

The area of the IO is modeled as shown below in Equation (10), where N_{IO} is the number of signals, f is the frequency, and R_{ON} and R_{TT1} are the impedance of the IO driver and the on-die termination circuit respectively as shown in Figure 3. A_0, k_0, k_1, k_2 and k_3 are constants for a given technology and design.

$$\begin{aligned} Area_{IO} = N_{IO} \cdot \left(A_0 + \frac{k_0}{min(R_{ON}, 2 \cdot R_{TT1})}\right) + \\ N_{IO} \cdot \left(\frac{1}{R_{ON}}\right) \cdot (k_1 * f + k_2 * f^2 + k_3 * f^3) \end{aligned} \qquad (10)$$

The area of the last stage of the driver is proportional to $1/R_{ON}$ or the drive current, and the fanout in the IO for the predriver stages is proportional to f, the frequency of the interface, to reflect the proportional edge rates needed based on the frequency. In the event that the on-die termination $(2 \cdot R_{TT1})$ is smaller than R_{ON}, the driver size is determined by $1/(2 \cdot R_{TT1})$. A_0 is the fixed area of the rest of the IO, which includes ESD protection.

III. CACTI-IO

CACTI-IO is an extended version of CACTI [4] that includes the models described in Section II above. CACTI-IO allows for a quick search of optimal IO configuration parameters that help optimize power and performance of the IO along with the DRAM and cache subsystem.

CACTI has analytical models for all the basic building blocks of a memory [19]: decoder, sense-amplifier, crossbar, on-chip wires, DRAM/SRAM cell and latch. We extend it to include the off-chip models presented in this paper. This requires modifying CACTI's global on-chip interconnect to include buffers at the PHY and drivers at the bank edge to connect to the IO circuit. Since all calculations are based on the ITRS [38] technology parameters, the energy and delay values calculated by CACTI are guaranteed to be mutually consistent. When a user inputs memory parameters and energy/delay constraints into CACTI, the tool performs an exhaustive design space exploration involving different array sizes, degrees of multiplexing, and interconnect choices to identify an optimal configuration. This exhaustive search now also considers off-chip power, area and timing by searching for optimal number of ranks, memory data width (x4, x8, x16 or x32 DRAMs), off-chip bus frequency and bus width.

We present three case studies: (1) high-capacity DDR3 based server configurations in Section III.B; (2) 3-D memory configurations for high-bandwidth systems in Section III.C; and (3) BOOM (Buffered Output On Module), a novel LPDDRx based configuration for servers [10] in Section III.D. All comparisons in the case studies are shown for one channel of the memory controller.

The IO power shown in the case studies is the peak power during activity, except in Section III.D for the BOOM case study, where we show how CACTI-IO can project the total system power as a sum of both IO and DRAM power and provide quick design-space exploration of both off-chip and on-chip components together. The case studies show the variety of options the IO models provide, as well as the achievable range of capacities and power efficiencies, making for interesting tradeoffs for the architect.

To further highlight the utility of CACTI-IO, we study two tradeoffs in more detail for the BOOM designs: in Section III.E we discuss optimal fanout of the data bus, and in Section III.F we discuss the optimal fanout of the address bus.

A. Simulation Methodology

For studies of the high-capacity DDR3 configurations and 3-D configurations, we run the CACTI-IO models stand-alone to provide IO power comparisons described in Sections III.B and III.C below. For the BOOM cases, we use a multi-core simulator [11] built on top of PIN [12] to provide the activity factor and idle-time information for multi-programmed workload mixes from SPLASH2 [13]. While different benchmarks will yield different results, we expect that overall trends for IO and DRAM power will remain stable. We model a 16-core processor with two memory controllers. Each controller has a dedicated memory channel and each channel has four ranks. Number of reads, writes, activates, idle cycles, and power down cycles from this simulation are fed into CACTI-IO to evaluate the DRAM as well as IO energy averaged over the SPLASH2 benchmarks for the different BOOM configurations described in Section III.D.

B. High-capacity DDR3 Configurations

We compare several configurations shown in Table IV for a x64 DDR3 memory channel; they all use a DIMM. RDIMM refers to a Registered DIMM, where the command and address signals are buffered to allow for increased capacity. A Load Reduced DIMM (LRDIMM) [30] has a buffer for both address and data signals,

TABLE IV

CASE STUDY 1: SUMMARY OF POWER FOR DIFFERENT SERVER CONFIGURATIONS USING X4 4GBIT DRAMS.

Configuration	Capacity (GB)	No. of DQ loads	BW (GB/s)	P_{IO} (W)	$P_{CPU-Buf}$ (W)	P_{PHY} (W)	Efficiency (GBps/W)	Efficiency·GB (GB·GBps/W)
2 RDIMMs dual rank	32	4	12.8	4.7	0.55	0.6	2.19	70.1
3 RDIMMs dual rank	48	6	12.8	6.2	0.55	0.8	1.70	81.6
3 LRDIMMs dual rank	48	2	12.8	4.86	3.2	0.8	1.44	69.12
3 LRDIMMs quad rank w/ 2-die stack	**96**	2x2d	12.8	5.1	3.2	0.8	1.41	135.4
BoB w/ 2 channels 2 dual rank RDIMMs	64	4	25.6	10.8	0.34	1.2	2.07	132.5

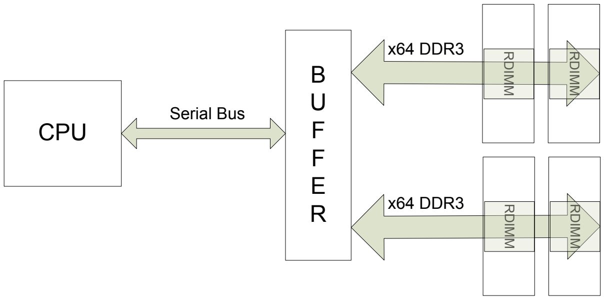


Fig. 5. BoB (Buffer-on-Board) [9].

allowing further increase in capacity at the cost of some data latency due to the buffering. The last configuration listed uses a Buffer-on-Board (BoB) from Intel [9] shown in Figure 5. In this configuration, the buffer is not integrated into the DIMM, but is rather a stand-alone chip on the board. The buffer drives two RDIMMs and has two channels (4 RDIMMs in all). While the interface between the RDIMM or LRDIMM and the CPU remains a DDR3 bus, the interface between the BoB and CPU is a proprietary serial interface [9].

All configurations shown in Table IV use x4 4Gb memory devices. We study the interface to the DRAM as the bottleneck in the system, and the timing on the interface between the buffer and the host CPU is assumed not to be the limiting factor in the study. The table lists the power consumed due to the IO on the DRAM interface (P_{IO}), the PHYs (P_{PHY}), and the IO on the interface between the CPU and the buffer ($P_{CPU-Buf}$). All configurations are assumed to operate at 800 MHz (DDR3-1600) and 1.5 V. As can be seen from the table, the LRDIMM offers a 50% increase in capacity (96 GB for a x64 channel) compared to the 3-RDIMM for a 17% decrease in efficiency. The product of capacity and efficiency is the highest for LRDIMM, at 135.4 GB·GBps/W. The BoB configuration offers a 30% increase in capacity and a 2X bandwidth improvement over the 3-RDIMM with 23% better power efficiency. Its product of capacity and efficiency is 132.5 GB·GBps/W.

This case study highlights the ability of CACTI-IO to calculate IO power numbers for various configurations under consideration, and search for an optimal solution based on either total capacity (3-LRDIMM with 2-die stack), or efficiency (2-RDIMM), or perhaps a very good balance between the two (BoB). The BoB design presents a novel means of increasing capacity using a buffer on the board, while maintaining efficiency and low pin-count using a serial bus to the CPU with 2X the bandwidth (25.6 GB/s).

C. 3-D Stacking Using Wide-IO

In our second case study, we evaluate different 3-D stacking configurations to maximize bandwidth. The configurations chosen include a 3-D TSS (Through Silicon Stack) 4-die 4Gb stacked DRAM with 4x128 channels [33], an 8-die stack with 4x128 channels, and narrower buses (4x64 and 4x32 as opposed to 4x128) with same bandwidth, all of which connect to the CPU directly, exposing the die stack to the external pin loading. We also include the Hybrid Memory Cube (HMC) proposed by Micron [8], wherein the memory controller is included along with the DRAM stack, and connected by a 16x128 interconnect (a serial interface is proposed to connect the HMC to the CPU) [8]. All configurations are assumed to operate at 1.2V [40]. The data-rate on the interface is limited by the relaxed DRAM timing parameters and data-rates proposed for Wide-IO [40], although CACTI-IO predicts some changes from the proposed data-rates based on the jitter sensitivity to loading and R_{ON}.

Table V shows the results for these configurations calculated by CACTI-IO. As can be seen, the power efficiency varies by around 2X, with the HMC showing the highest efficiency (56 GBps/W), and a 3-D stack using a 4x32 bus showing the lowest efficiency (27 GBps/W). A peak bandwidth of 176 GB/s for 16x128 channels is achieved for the HMC with a 4-die stack, a 4.76X improvement over the standard 3-D TSS stack in an external connection using 4x128 channels. The isolation provided by the HMC to the CPU allows the bus to operate faster without the additional external loading.

The 4x64 and 4x32 cases shown in Table V represent narrower buses that achieve the same bandwidth. The PHY power (taken from Tables I and II) goes up considerably for the x32 case since the complexity increases at 1066 MHz; this leads to the poorest efficiency. CACTI-IO can furthermore predict V_{ddmin} based on the voltage noise parameters as described in Equations (6) - (9). The V_{ddmin} and the scaled efficiency at V_{ddmin} are shown in Table V. CACTI-IO predicts that the HMC can further scale down to 0.85V and improve its efficiency to 100 GBps/W.

This case study highlights the ability of CACTI-IO to calculate IO power and timing for a new interconnect technology such as 3-D, including the novel Hybrid Memory Cube. The baseline models included in CACTI-IO can be configured for DDR3 based signaling as well as for 3-D interconnect. We see that CACTI-IO is able to identify the solution with the highest bandwidth and efficiency (HMC) and also predict how much the efficiency would be affected when going from 4x128 to 4x32 due to PHY power increase for the higher data rates. CACTI-IO is also able to predict V_{ddmin} for the given frequency and loading, providing a 1.8X improvement in power efficiency for the HMC.

D. BOOM: LPDDRx for Servers

BOOM (Buffered Output On Module) architecture [10] from Hewlett-Packard relies on a buffer chip on the board that connects to lower-speed and lower-power LPDDRx memories. To match the channel bandwidth, BOOM uses a wider DIMM-internal bus (from the buffer to the DRAMs) as shown in Figure 6. Further, BOOM has the option of grouping multiple physical ranks into a single logical rank [10]. BOOM allows for commodity LPDDRx DRAMs with lower power, but achieves high bandwidth and capacity through wider buses. As servers become more sensitive to memory subsystem

TABLE V
CASE STUDY 2: SUMMARY OF POWER FOR DIFFERENT 3-D CONFIGURATIONS.

Configuration	Capacity (GB)	BW$_{MAX}$ (GB/s)	IO Power (W)	PHY Power (W)	Efficiency (GBps/W)	V$_{ddmin}$ (V)	Efficiency @ V$_{ddmin}$ (GBps/W)
3-D 4-die	2	37	0.9	0.06	38	1	54
3-D 8-die	4	37	1.2	0.06	30	1.2	30
4x64	2	34	0.74	0.14	38	1.2	38
4x32	2	34	0.84	0.44	27	1.2	27
HMC	2	**176**	2.96	0.29	56	0.85	100

TABLE VI
CASE STUDY 3: SUMMARY OF POWER FOR DIFFERENT BOOM CONFIGURATIONS.

Configuration	Capacity (GB)	No. of DQ loads	BW (GB/s)	P$_{IO}$ (W)	P$_{CPU-Buf}$ (W)	P$_{PHY}$ (W)	Efficiency (GBps/W)
x8 BOOM-N2-D-800	16	4	12.8	4.96	3.52	0.8	1.38
x8 BOOM-N4-L-400	32	4	12.8	2.51	3.52	0.4	2.0
x8 BOOM-N4-L-400 with serial bus to host	32	4	12.8	2.51	0.34	0.4	3.94

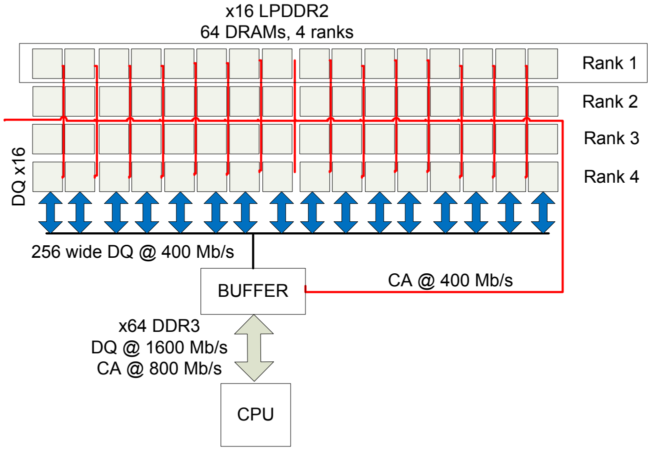


Fig. 6. BOOM-N4-L-400 configuration with x16 devices [10].

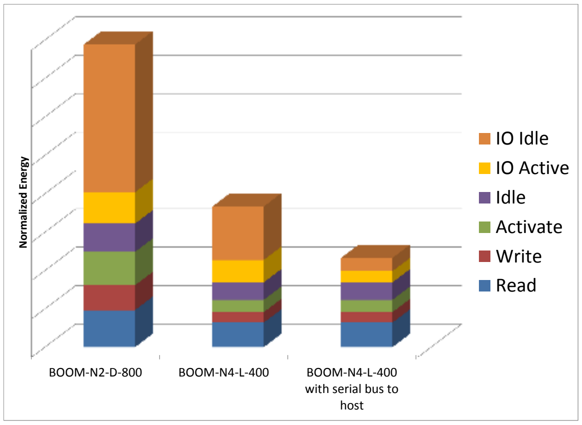


Fig. 7. Normalized system (DRAM+IO) energy for BOOM configurations.

power, BOOM provides a valuable means for use of mobile DRAM to achieve better power efficiency while still meeting server performance requirements.

Table VI summarizes the IO peak power for three BOOM configurations [10]. The power is shown per memory channel (equivalent of a x64 DDR3 channel). A BOOM configuration is denoted as BOOM-Nn-X-Y, where n is a ratio of the wider internal bus to the channel's x64 bus, X is DRAM type (D for DDR3 and L for LPDDR2) and Y is DRAM data rate (typically 1600/n Mb/s). All BOOM configurations shown use x8 memories.

Table VI clearly shows a 2X improvement in IO power (P_{IO}) from buffer to DRAM using LPDDRx memories to achieve the same bandwidth when we compare BOOM-N2-D-800 (using DDR3 DRAM) and BOOM-N4-L-400 (using LPDDR2 DRAM).

Additionally, BOOM offers the advantage of using a custom interface between the CPU host and the buffer chip. Instead of a standard x64 DDR3 interface, a serial bus similar to the BoB [9] case in Section III.B above can be used. This further improves the total efficiency by 2X, achieving a nearly 2.85X improvement in total power efficiency over a DDR3-based design.

To highlight the ability of CACTI-IO to provide combined DRAM and IO power, we compare the three BOOM configurations with respect to normalized energy, shown in Figure 7.

The simulation methodology used to obtain the normalized energy is described in Section III.A. The total energy is broken down into the DRAM core power (Read, Write, Activate, Precharge), the IO Active power (Read and Write) and the IO Idle power (mainly due to terminations and the active clock).

We make the following observations.

- The IO power is a significant portion of the combined power (DRAM+IO): 59% for the DDR3-based (BOOM-N2-D-800) configuration and 54% for the LPDDR2-based configuration (BOOM-N4-L-400). When using a serial bus from the buffer to the host, the IO power for BOOM-N4-L-400 reduces to 27% of the total power.
- The IO Idle power is a very significant contributor. The BOOM-N4-L-400 design reduces the IO Idle power by using LPDDR2 unterminated signaling, but since the BOOM configuration still relies on a DDR3 type bus from the buffer to the host as shown in Figure 6, the IO Idle power for the whole channel is still significant.
- Once the DRAM core becomes efficient, IO becomes a major contributor to the total power. Replacing DDR3 memories with LPDDR2 alone is not as efficient as further reducing the IO Idle power using a serial bus instead of a DDR3 style bus to the host. The BOOM-N4-L-400 design with a serial host provides a 3.4X energy savings (DRAM+IO) over the BOOM-N2-D-800 design.

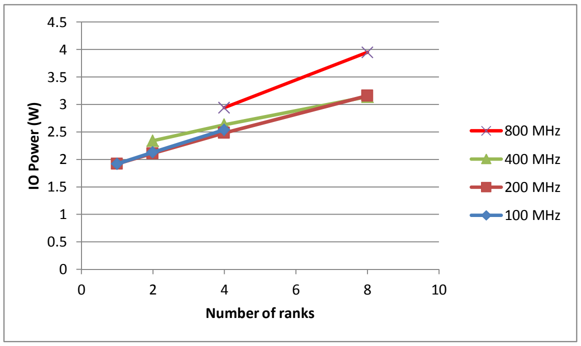


Fig. 8. IO power vs. number of ranks for BOOM-LPDDR2.

While Table VI only compares the IO Active power, Figure 6 also accounts for IO Idle power and projects total energy based on active and idle times. While the serial bus only provides a 2.85X savings in IO Active power, it provides an 11X savings in IO Idle power when compared to the BOOM-N2-D-800 design.

- The number of power-down cycles is around 15% of the total cycles. More aggressive power-down will help reduce the IO Idle power. Supply scaling is also an option at lower frequencies in the case of BOOM-N4-L-400.

This case study highlights CACTI-IO's ability to provide IO power numbers to a system simulator, which can then provide valuable insight into total system power. Only combining the IO and DRAM power brings out the right tradeoffs needed to further improve efficiency. The study also highlights how CACTI-IO can be used to optimize a buffer-based topology such as BOOM, where IO choices including bus frequency and width can make a 2.85X difference in IO Active power and nearly 11X difference in IO Idle power.

E. Optimizing Fanout for the Data Bus

We now illustrate how one can calculate the optimal number of physical ranks in a BOOM type configuration to minimize IO power for a fixed capacity and bandwidth (BW). The number of physical ranks represents the fanout on the data bus. For this example, we assume that the memory density per DRAM die is fixed.

If N_R is the number of ranks, W_B the bus-width, W_M the memory data-width and f the data rate, then [7]:

$$N_R \cdot (W_B/W_M) = Capacity \tag{11}$$

$$W_B \cdot 2f = BW \tag{12}$$

Figure 8 shows the IO power as we vary the number of ranks to meet a capacity of 64 DRAMs and a bandwidth of 12.8 GB/s for an LPDDR2 bus. The IO power varies for different bus frequencies f, as the width of the bus and the memory data-widths vary to meet the conditions in Equations (11-12). The memory data-width is chosen to be x4, x8, x16 or x32 for the LPDDRx memories. The number of ranks is 1, 2, 4 or 8. The bus-width is x64, x128, x256 or x512, and the bus frequency is 800 MHz, 400 MHz, 200 MHz or 100 MHz.

As can be seen from Figure 8, the wider and slower LPDDR2 bus provides the lowest power. A 512-wide bus using x8 memories in a single-rank configuration running at 100 MHz consumes the lowest power at 1.92 W, while a 64-wide bus using x8 memories in an eight-rank configuration running at 800 MHz consumes the highest power at 3.94 W. Also to be noted are the diminishing returns of scaling down to a lower speed once the bus is scaled to 200 MHz, owing to high-impedance terminations. This frequency at which termination is no longer needed depends on the interconnect length and the loading, which change based on the topology and technology as determined by the jitter DOE analysis.

One of the downsides to having a wider and slower bus is the cost of area on the die, package and board. CACTI-IO predicts the impact on on-die area as we scale frequency and bus-width to keep the bandwidth constant. Shown in Figure 9 is the IO area vs. frequency for low fanouts (1 or 2 ranks) in 28nm technology, such that total bandwidth is kept constant. Also shown is the R_{out} that is used in Equation (10) to calculate the area. Wider buses result in a net increase in area even though they operate at lower frequencies. In a buffer chip this may be acceptable as there is less premium on area than on a CPU or DRAM die. Since there is almost a 2X increase in area going from the 200 MHz to the 100 MHz solution, while there is hardly any difference in power, it may be prudent to choose the 200 MHz solution. The optimal solution would then be N_R = 1, W_B = 256, W_M = 4 and f = 200MHz. This example highlights CACTI-IO's ability to optimize the number of ranks based on IO power and any user-provided IO area, thus helping to optimize the IO configuration for a buffer-based design.

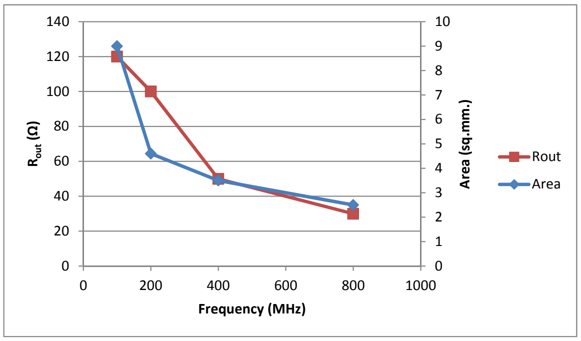


Fig. 9. Area vs. frequency for a constant-bandwidth BOOM-LPDDR2.

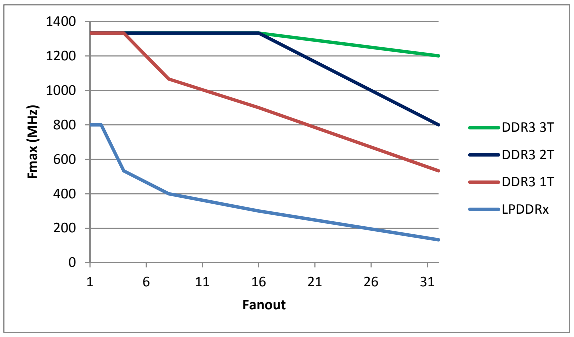


Fig. 10. Fanout vs. Fmax for a typical DDR3 CA bus.

F. Optimizing Fanout for the Address Bus

As we increase capacity, the address bus incurs a penalty as all memories on the channel share a common address bus. The LPDDR2 and LPDDR3 standards [23] offer address buses at DDR speeds, with no option for 2T (2 clock-cycle) timing [22]. This idiosyncrasy in the DRAM specification is not easily exposed to architects, but CACTI-IO allows for verified configurations to be systematically provided to architects.

To calculate the maximum achievable speed for a fly-by topology as shown in Figure 4, we need to define the sensitivity of the jitter on the CA (command-address) bus to the fanout of the bus as shown in Equation (5). Figure 10 shows the maximum achievable clock frequency on the CA bus for DDR3 and LPDDR2/3 as a function of the fanout for a representative channel. For DDR3, the 2T and 3T timing options allow for relaxed timing on the CA bus [21].

Given the limitation for the LPDDR2 address fanout owing to the DDR speed requirement, multiple address buses may be needed to achieve higher capacities. For instance, based on the example in

Figure 10, with a fanout of 16 we would need two LPDDR2 CA buses to support 400 MHz, while a single CA bus on DDR3 could support 1066 MHz with 2T timing.

With a buffer-based design, it is possible to have multiple address buses for a given channel between the buffer chip and the DRAMs. This would provide a means to limit the fanout on the address bus. Architects can optimize the design for a given address speed with optimal latency and burst requirements, including sub-ranking [10]. Understanding the limitations of the address bus allows architects to plan to overcome or minimize its impact on system performance.

IV. Summary

We have presented CACTI-IO, a version of CACTI that models the off-chip memory interface for server and mobile configurations. Its models include off-chip power and IO area, as well as voltage and timing margins that help define the maximum achievable bandwidth. Our framework permits quick design space exploration with the rest of the memory subsystem and provides a systematic way for architects to explore the off-chip design space. It also exposes DRAM signaling standards and their idiosyncrasies to architects, while still providing an easily-extensible framework for customization of off-chip topologies and technologies.

Using CACTI-IO, we have also illustrated the tradeoffs between capacity, bandwidth, area and power of the memory interface through three industry-driven case studies. These clearly show the ability of CACTI-IO to calculate IO power for various configurations, including DIMMs, 3-D interconnect, and buffer-based designs such as BoB and BOOM. CACTI-IO helps determine the lowest-power off-chip configuration (bus width, memory data width, number of physical ranks, address bus fanout, minimum supply voltage, and bus frequency) for given capacity and bandwidth requirements. We have demonstrated how this impacts the data and address buses differently by studying the optimal fanout for each in a BOOM design.

Furthermore, we have highlighted the capability of CACTI-IO to combine IO and DRAM power, which shows the significant contribution of IO power to the total (DRAM+IO) memory power (up to 59% in some cases). We have observed the relative importance of IO Idle power by using CACTI-IO and a system simulator together to calculate system energy in various modes (Read, Write, Activate, Precharge, Idle). A combination of a wider and slower bus to the DRAM and a faster serial bus to the CPU provides the lowest IO Idle power.

We plan to make CACTI-IO publicly available online. We expect that the new capabilities provided by this tool will enable improved understanding of memory interface issues, allowing architects to evaluate customized off-chip buffer-based designs as well as new interconnect technologies for impact on system power and performance.

References

[1] W. Dally and J. Poulton, *Digital Systems Engineering*, Cambridge University Press, 1998.

[2] H. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley, 1990.

[3] D. Oh and C. Yuan. *High-Speed Signaling: Jitter Modeling, Analysis, and Budgeting*, Prentice Hall, 2011.

[4] CACTI. http://www.hpl.hp.com/research/cacti/

[5] N. P. Jouppi, A. B. Kahng, N. Muralimanohar and V. Srinivas, "CACTI-IO Technical Report," *technical report* CS2012-0986, UC San Diego CSE Department, August 2012.

[6] N. Chang, K. Kim and J. Cho, "Bus Encoding for Low-Power High-Performance Memory Systems," *Proc. IEEE DAC*, 2000, pp. 800-805.

[7] A. B. Kahng and V. Srinivas, "Mobile System Considerations for SDRAM Interface Trends," *Proc. ACM/IEEE SLIP Workshop*, 2011, pp. 1-8.

[8] J. Baloria, "Micron Reinvents DRAM Memory: Hybrid Memory Cube," *Proc. IDF Workshop*, Sept. 2011.

[9] Intel's Scalable Memory Buffer. http://tinyurl.com/7xbt27o

[10] D. H. Yoon, J. Chang, N. Muralimanohar and P. Ranganathan, "BOOM: Enabling Mobile Memory Based Low-Power Server DIMMs," *Proc. IEEE ISCA*, 2012, pp 25-36.

[11] McSim. http://cal.snu.ac.kr/mediawiki/index.php/McSim

[12] C.-K. Luk et al., "PIN: Building Customized Program Analysis Tools with Dynamic Instrumentation," *Proc. ACM PLDI*, 2005, pp. 190-200.

[13] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh and A. Gupta, "The SPLASH-2 Programs: Characterization and Methodological Considerations," *Proc. IEEE ISCA*, 1995, pp. 24-36.

[14] H. Zheng and Z. Zhu, "Power and Performance Trade-Offs in Contemporary DRAM System Designs for Multicore Processors," *IEEE Trans. on Computers* 59(8) (2010), pp. 1033-1046.

[15] H. Lee et al., "A 16 Gb/s/Link, 64 GB/s Bidirectional Asymmetric Memory Interface," *IEEE JSSC* 44(4) (2009), pp. 1235-1247.

[16] J. Poulton, R. Palmer, A. M. Fuller, T. Greer, J. Eyles, W. J. Dally and M. Horowitz, "A 14-mW 6.25-Gb/s Transceiver in 90-nm CMOS," *IEEE JSSC* 42(12) (2007), pp. 2745-2757.

[17] F. O'Mahony et al., "A 47x10Gb/s 1.4mW/(Gb/s) Parallel Interface in 45nm CMOS," *Proc. IEEE ISSCC*, 2010, pp. 156-158.

[18] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen and N. P. Jouppi, "McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures," *Proc. IEEE/ACM MICRO*, 2009, pp. 469-480.

[19] S. Thoziyoor, J. Ahn, M. Monchiero, J. B. Brockman and N. P. Jouppi, "A Comprehensive Memory Modeling Tool and its Application to the Design and Analysis of Future Memory Hierarchies," *Proc. IEEE ISCA*, 2008, pp. 51-62.

[20] Micron DRAM System Power Calculators. http://www.micron.com/support/dram/power_calc.html

[21] JEDEC DDR3 Specification JESD79-3B.

[22] JEDEC LPDDR2 Specification JESD209-2C.

[23] JEDEC. http://www.jedec.org

[24] G. Taguchi, *Introduction to Quality Engineering*, 2nd ed., McGraw-Hill, 1996.

[25] R. Palmer, J. Poulton, A. Fuller, J. Chen and J. Zerbe, "Design Considerations for Low-Power High-Performance Mobile Logic and Memory Interfaces," *Proc. IEEE ASSCC*, 2008, pp. 205-208.

[26] J. Ellis, "Overcoming Obstacles for Closing Timing for DDR3-1600 and Beyond," *Denali MemCon*, 2010.

[27] A. Vaidyanath, "Challenges and Solutions for GHz DDR3 Memory Interface Design," *Denali MemCon*, 2010.

[28] D. Wang, B. Ganesh, N. Tuaycharoen, K. Baynes, A. Jaleel and B. Jacob, "DRAMsim: A Memory System Simulator," *ACM SIGARCH Computer Architecture News - special issue* 33(4) (2005), pp. 100-107.

[29] HP Memory Technology Evolution: An Overview of System Memory Technologies. http://tinyurl.com/7mvktcn

[30] http://www.micron.com/products/dram_modules/lrdimm.html

[31] "Challenges and Solutions for Future Main Memory," *Rambus White Paper*, May 2009. http://tinyurl.com/cetetsz

[32] B. Schroeder, E. Pinheiro and W. Weber, "DRAM Errors in the Wild: A Large-Scale Field Study," *Proc. ACM SIGMETRICS*, 2009, pp. 193-204.

[33] J.-S. Kim et al., "A 1.2V 12.8GB/s 2Gb Mobile Wide-I/O DRAM with 4128 I/Os Using TSV-Based Stacking," *Proc. IEEE ISSCC*, 2011, pp. 496-498.

[34] S. Sarkar, A. Brahme and S. Chandar, "Design Margin Methodology for DDR Interface," *Proc. IEEE EPEPS*, 2007, pp. 167-170.

[35] S. Chaudhuri, J. McCall and J. Salmon, "Proposal for BER Based Specifications for DDR4," *Proc. IEEE EPEPS*, 2010, pp. 121-124.

[36] M. Qureshi, V. Srinivasan and J. Rivers, "Scalable High-Performance Main Memory System Using Phase-Change Memory Technology," *Proc. IEEE ISCA*, 2009, pp. 24-33.

[37] HP Power Advisor. http://h18000.www1.hp.com/products/solutions /power/index.html.

[38] *International Technology Roadmap for Semiconductors*, 2011 edition. http://www.itrs.net/

[39] B. K. Casper, M. Haycock and R. Mooney, "An Accurate and Efficient Analysis Method for Multi-Gb/s Chip-to-Chip Signaling Schemes," *Proc. IEEE VLSIC*, 2002, pp. 54-57.

[40] Future-Mobile JEDEC Draft Wide IO Specification.

[41] M. A. Horowitz, C.-K. K. Yang and S. Sidiropoulos, "High-Speed Electrical Signaling: Overview and Limitations," *IEEE Trans. on Advanced Packaging* 31(4) (2008), pp. 722-730.

[42] D. Oh, F. Lambrecht, J. H. Ren, S. Chang, B. Chia, C. Madden and C. Yuan, "Prediction of System Performance Based on Component Jitter and Noise Budgets," *Proc. IEEE EPEPS*, 2007, pp. 33-36.