ZettaRAM: A Power-Scalable DRAM Alternative through Charge-Voltage Decoupling

Ravi K. Venkatesan, *Student Member*, *IEEE*, Ahmed S. Al-Zawawi,
Krishnan Sivasubramanian, *Student Member*, *IEEE*, and Eric Rotenberg, *Member*, *IEEE*

Abstract—ZettaRAM™ is a nascent memory technology with roots in molecular electronics. It uses a conventional DRAM architecture except that the conventional capacitor is replaced with a new molecular capacitor. The molecular capacitor has a discrete threshold voltage, above which all molecules are charged and below which all molecules are discharged. Thus, while voltage still controls charging/discharging, the fixed charge deposited on the molecular capacitor is voltage-independent. Charge-voltage decoupling makes it possible to lower voltage from one memory generation to the next while still maintaining the minimum critical charge for reliable operation, whereas DRAM voltage scaling is constrained by charge. Voltage can be scaled inexpensively and reliably by engineering new, more favorable molecules. We analyze how three key molecule parameters influence voltage and then evaluate 23 molecules in the literature. Matching DRAM density and speed, the best molecule yields 61 percent energy savings. While the fixed charge is voltage-independent, speed is voltage-dependent. Thus, voltage is padded for competitive latency. We propose dynamically modulating the padding based on criticality of memory requests, further extending ZettaRAM's energy advantage with negligible system slowdown. Architectural management extends the best molecule's energy savings to 77 percent and extracts energy savings from six otherwise uncompetitive molecules.

Index Terms—DRAM, dynamic voltage scaling, low-power memory, molecular electronics, molecular memory, memory technology.

1 Introduction

ZETTARAM ™ is a new memory technology under development by ZettaCore as a potential replacement for conventional DRAM [26]. ZettaCore's strategy is to initially leverage the large investment in silicon fabs to attain competitive memories within a few years. Accordingly, these new memories are based on conventional DRAM architectures—address decoder, wordline, access transistor, bitline, sense amp, etc. The key innovation is replacing the conventional capacitor in each DRAM cell with a new type of capacitor, which had its genesis in a DARPA-sponsored molecular electronics project [20]. Although one goal of that project was to eventually deploy individual charge-storage molecules as 1-bit memory elements and integrate them with other molecular-scale electronics, ZettaCore currently exploits many charge-storage molecules in aggregate to create a molecular capacitor and replacement for the DRAM capacitor.

The aggregate molecular capacitor retains key advantages of the underlying nanotechnology from which it is derived:

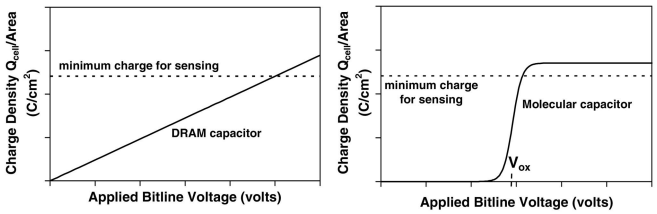
1. *Cost-effective density scaling:* Self-assembly is the process by which the thousands of molecules that make up a molecular capacitor automatically arrange themselves into a single, uniform, dense monolayer. Self-assembly and high charge density of the monolayer reduce or eliminate the need for an elaborate three-dimensional capacitor structure that is required in conventional DRAM to achieve sufficient charge. In DRAM, reducing the cross-sectional area of a cell requires a correspondingly taller structure (stacked capacitor or deep trench capacitor) to maintain enough charge for sensing a "1" within the smaller cross-sectional area. The planar molecular capacitor provides a less expensive means for scaling density.
2. *Precise control of molecules' attributes:* Engineering and synthesizing molecules is precise, predictable/repeatable, and can be done in inexpensive laboratories, whereas tuning bulk properties of semiconductors is expensive and harder to control. Molecular engineering provides precise control over characteristics of molecules, such as the speed with which electrons can be added/removed (affecting the speeds of reading and writing), the voltage at which electrons can be added/removed (affecting read and write power consumption), and monolayer density (affecting charge density and, thus, overall memory density). This provides flexibility in the selection of performance, power consumption, and density.

In this paper, we delve into the circuit-level behavior of the molecular capacitor, uncovering two unique properties of the molecular capacitor and developing the opportunities that they present for extreme power scaling.

- *R.K. Venkatesan is with Intel Technology India Private Limited, Mobile Platforms Architecture Division, Mobility Group, #23-56P, Davarabeesanahalli, Outer Ring Road, Varthur Hobli, Bangalore-560037, India. E-mail: ravi.k.venkatesan@intel.com.*
- *A.S. Al-Zawawi and E. Rotenberg are with the Department of Electrical and Computer Engineering, Center for Embedded Systems Research, North Carolina State University, Partners Building I, Suite 2300, Campus Box 7256, Raleigh, NC 27695-7256. E-mail: aalzawa@ncsu.edu, ericro@ece.ncsu.edu.*
- *K. Sivasubramanian is with The Vanguard Group-IT Division, 2525 Water Ridge Pkwy, 2 Water Ridge #C320, Charlotte, NC 28217. E-mail: krishnan_siva@vanguard.com.*

Manuscript received 2 Oct. 2005; revised 16 Mar. 2006; accepted 7 July 2006; published online 20 Dec. 2006.
For information on obtaining reprints of this article, please send e-mail to: tc@computer.org, and reference IEEECS Log Number TCSI-0340-1005.

0018-9340/07/$25.00 © 2007 IEEE Published by the IEEE Computer Society



(a) (b)

Fig. 1. Charge versus voltage for (a) a conventional capacitor and (b) a molecular capacitor.

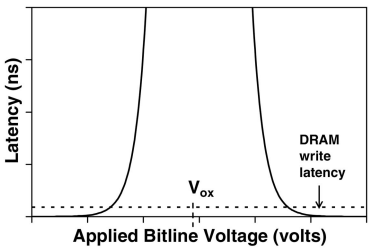


Fig. 2. Intrinsic latency of charging and discharging molecules versus voltage.

1.1 Scalable Power Enabled by Charge-Voltage Decoupling

The amount of charge deposited on the molecular capacitor is fixed—since there is a discrete number of molecules—and independent of the applied voltage. In contrast, the amount of charge deposited on a conventional capacitor depends linearly on the applied voltage. This distinction is illustrated in Fig. 1, which shows charge density (charge per unit area) as a function of voltage for (a) a conventional capacitor and (b) a molecular capacitor. The conventional capacitor exhibits $Q = CV$, where Q is charge, C is capacitance, and V is voltage. The molecular capacitor exhibits a threshold voltage V_{ox} (oxidation potential), above which all of the molecules are charged and below which all of the molecules are discharged. Thus, while voltage still controls charging and discharging of the molecular capacitor via a threshold, the fixed charge itself does not depend on the voltage. We call this property *charge-voltage decoupling* [27].

Conventional DRAM faces a major power scaling challenge in the long term because its charge and voltage are coupled. Because $Q = CV$, there is a minimum write voltage, below which not enough charge is deposited on the conventional capacitor for the sense amplifier to reliably detect a "1" during a later read operation. The minimum charge for reliable sensing is shown with the dashed horizontal line superimposed on the graph in Fig. 1a. The minimum write voltage corresponds to where this line intersects the conventional capacitor curve. The problem is that both the minimum charge and the cell capacitance have remained nearly constant from one memory generation to the next and this trend is projected to continue, making it very difficult to continue lowering voltage in future generations of DRAM [13], [15], [24].

Conversely, scaling the voltage of ZettaRAM is viable because of the charge-voltage decoupling of the molecular capacitor. Charge-voltage decoupling enables voltage to be lowered while still maintaining the minimum critical charge for reliable operation.

Voltage can be scaled inexpensively by engineering more favorable molecules. Key properties of the molecules can be tuned through the choice of molecular "groups" and "linkers," such as the oxidation potential (V_{ox}), electron transfer rate (k^0), and surface concentration (charge density). A key contribution of this paper is providing analyses and insights into how these three parameters influence the operating voltage either directly or indirectly. Another key contribution is demonstrating the long-term power scalability of ZettaRAM by evaluating 23 molecules, synthesized and characterized by ZettaCore. Matching DRAM density and speed, the best molecule yields 61 percent energy savings.

1.2 More Scalable Power by Optimizing the Speed-Voltage Trade-Off

While the molecular capacitor's fixed charge is independent of applied voltage, the speed of charging/discharging the molecules depends on the difference between the applied voltage and the threshold voltage (V_{ox}). We refer to this second property as the *speed-voltage trade-off* or *speed-energy trade-off*. Charging/discharging the molecules becomes exponentially slower the closer the applied voltage is to V_{ox}. This is illustrated in Fig. 2, which shows the intrinsic latency of charging/discharging the molecules as a function of voltage. Superimposed on this graph is the DRAM write latency (dashed line). The overall latency of ZettaRAM is determined by either the latency of charging/discharging the molecules or the latency of the conventional peripheral circuitry used to access the molecular capacitor, whichever is slower. Accordingly, from Fig. 2, ZettaRAM has the same latency as DRAM if the applied voltage is sufficiently "padded" with respect to V_{ox} that the intrinsic latency of the molecules is not the bottleneck.

Padding voltage to achieve competitive latency does not negate the power scaling benefits of charge-voltage decoupling. Nonetheless, if we could reduce the padding without sacrificing performance, we could more fully capitalize on charge-voltage decoupling and thereby extend ZettaRAM's power scaling advantage over DRAM.

We propose architectural techniques to intelligently manage the speed-voltage trade-off. Specifically, the padding is dynamically modulated based on the criticality of memory requests, minimizing energy with negligible system slowdown [1].

1.2.1 Hybrid Write Policy

Requests to DRAM are usually serviced from a row buffer —an entire row (page) of the memory bank held in the row buffer. The row buffer contains the most recently accessed memory page. When a memory request (initiated by the L2 cache) misses in the row buffer, the current open page is closed (write operation) before opening a new page (read operation) to service the request. Bitline voltage swings are caused by both write and read operations. In ZettaRAM, the read operation can be performed only at a fixed voltage, as explained in detail in Section 2.1. However, the write operation can be performed at either a high voltage (fast, but high energy), favoring performance, or a low voltage (slow, but low energy), favoring energy savings.

Two types of memory requests are initiated by the L2 cache, fetch block and writeback block. Several factors converge nicely to direct focus on L2 writebacks: 1) They account for 80 percent of row buffer misses, thus most of the energy savings potential, and 2) they do not directly stall

the processor and thereby offer scheduling flexibility for tolerating extended molecule latency. On the other hand, L2 fetch requests typically stall the processor even with out-of-order execution because the instruction scheduling window is not large enough to accommodate the high memory round-trip latency. Accordingly, we propose a hybrid policy in which slow writes (low energy) are applied to noncritical writebacks that miss in the row buffer and fast writes (high energy) to critical fetches that miss in the row buffer. Applying slow writes to writebacks taps most of the energy savings potential and applying fast writes to fetches ensures little performance degradation.

As one example, we consider the additional energy savings yielded by the hybrid write policy for one of the first synthesized porphyrin molecules. This molecule has $V_{ox} = 0.73$ V. We show that, if fast and slow writes are done at 1.2 V and 1.0 V, respectively, then the hybrid write policy yields 34 percent bitline energy savings (out of a possible 41 percent with uniformly slow writes) with only a 10 percent increase in execution time (as opposed to 81 percent with uniformly slow writes). Thus, the hybrid policy combines the performance of uniformly fast writes with the energy savings of uniformly slow writes. The residual 10 percent performance degradation still exists because, although deferred writebacks do not directly stall the processor, they may fill up the request queues in the memory controller, eventually stalling critical fetches.

1.2.2 Combining Hybrid Write Policy with Eager Writeback

One approach to avoiding any queue-full stalls is to increase the request queue size. The residual performance degradation is reduced to less than 1 percent when the request queues are increased from 4 to 64 entries. However, enlarging the queues increases system cost (each entry contains an entire cache block) and complexity (fetches that bypass queued writebacks must first search the queue for read-after-write hazards).

To avoid the cost and complexity of larger queues, as an alternative approach, we propose employing the eager writeback policy in the L2 cache [14] to evenly spread out writeback requests, potentially eliminating queue-full stalls. In the eager writeback policy, a writeback is issued as soon as a dirty block becomes the LRU block in its set, instead of waiting for the block to be evicted. Issuing the writeback early from the L2 cache compensates for delaying it in the memory controller. Similarly to enlarging the queues, the eager writeback policy reduces the residual performance degradation to less than 1 percent. This is achieved without enlarging the request queues with respect to the baseline system (four entries).

Hybrid fast/slow writes coupled with L2 cache eager writebacks extend the energy savings of ZettaRAM by an additional 25-40 percent (depending on the molecule type) with negligible performance loss and the least complexity in the memory controller. The total energy savings of the best molecule increases from 61 percent with uniformly fast writes to 77 percent with architectural management. Moreover, architectural management extracts energy savings from six otherwise uncompetitive molecules.

1.3 Paper Organization

The rest of this paper is organized as follows: Section 2 provides background on the molecular capacitor, including basic read/write operation, our novel SPICE device model, and our novel derivation of charge density as a function of write voltage. Section 3 presents sample SPICE results for DRAM and ZettaRAM (for a sample molecule). Section 4 describes our framework for system-level experiments. Section 5 presents system-level simulation results for the baseline DRAM system. Section 6 studies the impact of key molecular attributes on operating voltage and then presents system-level results for 23 different molecules. Section 7 presents architectural management of ZettaRAM's speed-voltage trade-off to lower operating voltage even further. Related work is discussed in Section 8. Finally, Section 9 summarizes the paper.

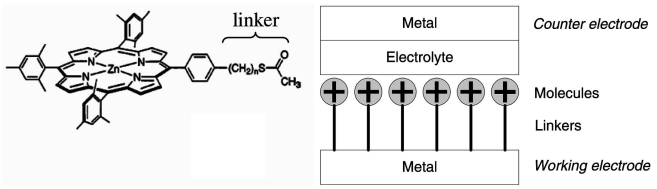


Fig. 3. (a) Individual porphyrin molecule. (b) ZettaRAM molecular capacitor.

2 ZettaRAM Molecular Capacitor

2.1 Molecule Description and Reading/Writing the Molecular Capacitor

A ZettaRAM memory cell is formed by replacing the DRAM capacitor with a molecular capacitor, composed of a self-assembled monolayer of charge-storage molecules (e.g., porphyrin molecules) sandwiched between two electrodes. An individual porphyrin molecule is shown in Fig. 3a and the ZettaRAM molecular capacitor is shown in Fig. 3b. As shown in Fig. 3b, the molecules are attached to the lower metal plate, or *working electrode*, via attachment groups called linkers. A linker is shown in detail in Fig. 3a (its length can be customized). The second electrode, or *counter electrode*, is interfaced to the molecules via an electrolyte.

A molecule can be positively charged by removing an electron, referred to as *oxidation*. Oxidation corresponds to writing a "1." An electron can be added back to the positively charged molecule to return it to the uncharged state, referred to as *reduction*. Reduction corresponds to writing a "0." The molecules are oxidized when the voltage applied across the molecules is greater than the oxidation potential (V_{ox}). In Fig. 3b, this is achieved by applying a voltage greater than V_{ox} on the working electrode relative to the counter electrode, causing electrons to tunnel from the molecules to the working electrode across the linkers. The molecules are reduced when the voltage applied across the molecules is less than V_{ox}. This is achieved by applying a voltage less than V_{ox} on the working electrode relative to the counter electrode, causing electrons to tunnel from the working electrode to the molecules across the linkers.

A more accurate explanation is that oxidation and reduction are always taking place simultaneously since any chemical reaction is a combination of forward and reverse reactions. Equilibrium is reached, at which point the rates of the forward and reverse reactions are equal. Although the

rates are balanced at equilibrium, the molecules have a strong tendency toward either the oxidized state or the reduced state, depending on the applied voltage (above or below the oxidation potential, respectively). The Butler-Volmer equation in the next subsection expresses the nonequilibrium and equilibrium behavior.

Like reading conventional DRAM, reading ZettaRAM is destructive. To read the state of the molecules in a molecular capacitor, they are discharged (if they are initially charged). This is achieved by reducing them, i.e., the bitline is precharged to a voltage below the oxidation potential. The state of the molecules is sensed by detecting the presence (or absence) of a small voltage change on the bitline as the molecules are discharged (unless neutral), which is procedurally similar to sensing in conventional DRAMs.

An idiosyncrasy of the molecular capacitor with regard to reading is that the bitline needs to be precharged to a specific voltage below the oxidation potential, called the *open circuit potential* (OCP) [21]. The molecular capacitor is an electrochemical cell in which the redox species is the porphyrin molecule. The OCP is a well-known artifact of electrochemical cells. Reading at the OCP prevents discharging of the "double-layer capacitance," which would otherwise drown out discharging of the molecules themselves.

Technological problems that have to be considered when integrating the molecular capacitor in a standard CMOS process include potential degradation of molecules during high-temperature processing steps and potential defects due to nonuniform layers of porphyrin molecules on silicon. These issues have been successfully addressed [19], [20], [21].

2.2 SPICE Model of Molecular Capacitor

The oxidation/reduction reactions are shown below, where A is the porphyrin molecule [19].

$$A \leftrightarrow A^{+} + e^{-}. \tag{1}$$

In nonequilibrium (charging or discharging), the net rate of oxidation or reduction—i.e., the net current—is exponentially dependent on the difference between the applied voltage and the oxidation potential. This current is expressed by the Butler-Volmer kinetic model [2], shown below, which forms the basis of our SPICE model.

$$I = F \cdot k^{0} \cdot \left([A] \cdot e^{(1-\alpha)\left(\frac{F}{RT}\right)(V-V_{ox})} - [A^{+}] \cdot e^{-\alpha\left(\frac{F}{RT}\right)(V-V_{ox})} \right). \tag{2}$$

The parameters above are as follows:

$\mathrm{k}^0 = \text{standard rate constant}, \alpha = \text{transfer coefficient},$

$\mathrm{F} = \text{Faraday constant}, \mathrm{R} = \text{gas constant}, \mathrm{T} = \text{temperature},$

$\mathrm{V} = \text{applied voltage}, \mathrm{V_{ox}} = \text{oxidation potential},$

$[\mathrm{A}] = \text{concentration of nonoxidized molecules}$

(in moles per unit area),

and $[\mathrm{A}^+] = \text{concentration of oxidized molecules}$.

The transient current I determines the *intrinsic* speed of reading and writing the molecules. Of course, when we integrate a SPICE model of the molecular capacitor into a complete memory circuit, the overall speed will be determined by several interacting components. That is, like any SPICE device model (e.g., transistor, resistor, capacitor, etc.), when the device model of the molecular capacitor is integrated into a larger circuit, the SPICE simulator correctly solves for currents and voltages at all nodes, accurately reflecting the interaction between the molecular capacitor and the rest of the circuit.

2.3 Highly Nonlinear Capacitance: Charge-Voltage Decoupling

The oxidation/reduction reactions shown in (1) eventually reach an equilibrium. The net current is zero at this equilibrium. We can derive the amount of charge ($\mathrm{Q_{cell}} = [\mathrm{A}^+]$) at equilibrium as a function of the write voltage by substituting $\mathrm{I} = 0$ in the Butler-Volmer equation (2). (This gives us the effective capacitance of the molecular capacitor since capacitance expresses Q as a function of V.) Doing so yields the following $\mathrm{Q_{cell}(V)}$:

$$Q_{cell}\ (V) = [A]_0 \cdot \left[\frac{1}{1 + e^{-\frac{F}{RT}(V-V_{ox})}} \right]. \tag{3}$$

$[\mathrm{A}]_0$ is the total molecule concentration, equal to the sum of [A] and $[\mathrm{A}^+]$. Equation (3) is the basis for the unusual charge density graph shown earlier in Fig. 1b.

The exponential term in the denominator becomes negligible as V is increased slightly above $\mathrm{V_{ox}}$. Thus, just above $\mathrm{V_{ox}}$, the molecular capacitor is nearly fully charged. Conversely, the exponential term in the denominator grows large as V is decreased slightly below $\mathrm{V_{ox}}$. Thus, just below $\mathrm{V_{ox}}$, the molecular capacitor is nearly fully discharged. In other words, while voltage still controls charging and discharging, depending on whether the applied voltage is above or below the threshold, $\mathrm{V_{ox}}$, the *amount* of charge is independent of the applied voltage, as shown in Fig. 1b.

3 Sample SPICE Results

In this section, SPICE simulations are performed to determine 1) operating voltages and 2) read/write latencies for both DRAM and ZettaRAM. Note that different molecule types exhibit different operating voltages (we target the same latencies for all molecule types). For the purposes of this section, we present SPICE simulations for only one molecule type (molecule #9 in Table 5). Molecule 9 yields similar system performance and energy as conventional DRAM without our architectural management technique. In Sections 6 and 7, we consider 23 different molecules and architectural management.

We use 0.18μ technology and assume a 10:1 ratio between bitline capacitance and cell capacitance [10]. The sense amps are designed accordingly and are based on designs in the DRAM literature [12].

3.1 SPICE Models

Fig. 4a shows the SPICE model of the DRAM architecture, including bitline, wordline, access transistor, conventional capacitor, and sense amplifier.

Fig. 4b shows the SPICE device model of the molecular capacitor. The voltage-controlled current source implements (2). The current depends on three variables: [A], $[\mathrm{A}^+]$, and V.

Fig. 4c shows the SPICE model of the ZettaRAM architecture, including bitline, wordline, access transistor, molecular capacitor, and sense amplifier. The only difference between the DRAM and ZettaRAM SPICE models is the type of capacitor used inside the cell (conventional versus molecular, respectively).

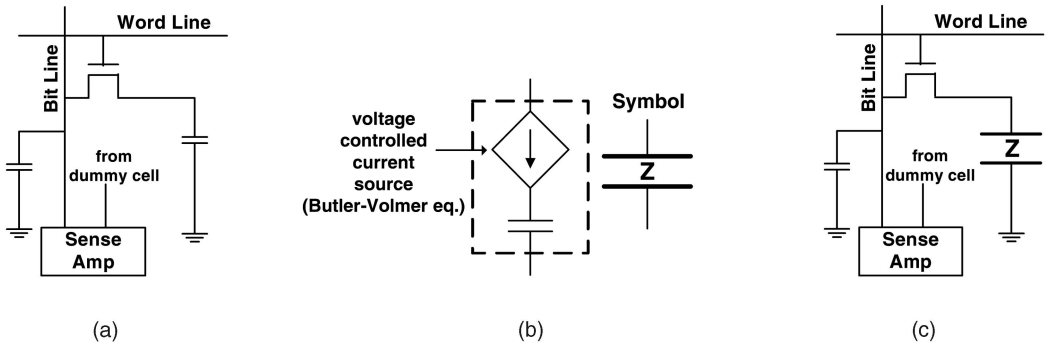


Fig. 4. (a) DRAM circuit. (b) SPICE device model of molecular capacitor. (c) ZettaRAM circuit.

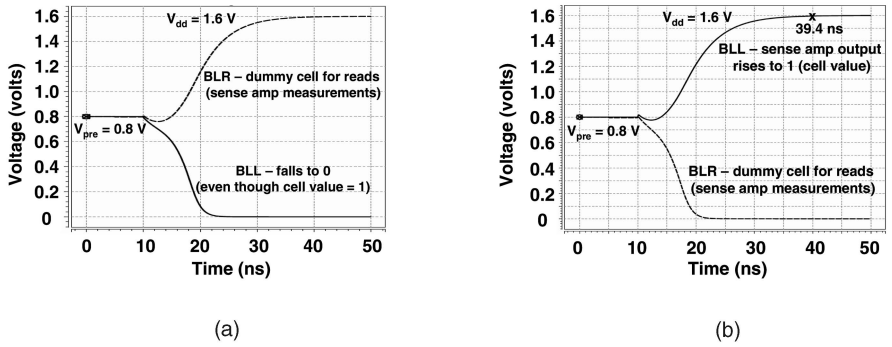


Fig. 5. (a) Writing DRAM capacitor below 1.25 V causes the subsequent read operation to fail. (b) Read latency of DRAM is 29 ns.

3.2 DRAM SPICE Results

The linear relationship between charge and voltage in a conventional capacitor places a lower bound on the DRAM write voltage for writing a "1." Below this voltage, the charge deposited on the capacitor is not enough for the sense amplifier to reliably sense a "1" during a later read operation. We determine this lower bound experimentally and call this write voltage $V_{d\_write\_1}$. Searching in increments of 0.05 V, we determined $V_{d\_write\_1} = 1.25$ V. The graph in Fig. 5a shows that writing the DRAM capacitor at 1.2 V causes sensing to fail during a later read operation since there is not enough charge on the capacitor.

Next, we determine the read and write latencies of DRAM. SPICE produces a read latency of 29 ns, as shown in Fig. 5b. SPICE produces a write latency of 8.6 ns for $V_{d\_write\_1} = 1.25$ V (a graph is not included here due to space constraints).

3.3 ZettaRAM SPICE Results

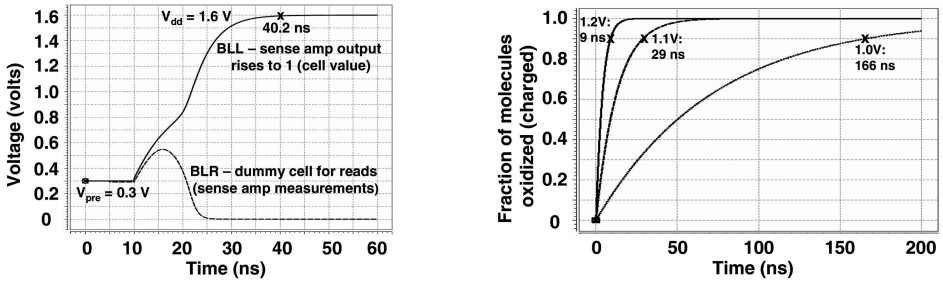
In the previous subsection, we showed that the conventional capacitor of DRAM is not sufficiently charged below 1.25 V from the standpoint of correct sensing during a later read operation. On the other hand, writing the molecular capacitor at a voltage as low as 1.0 V (and possibly lower) results in correct sensing during a later read operation, as shown in Fig. 6a.

Next, we determine the write latencies of ZettaRAM as a function of the ZettaRAM write voltage, $V_{z\_write\_1}$. In the first experiment, we use DRAM's minimum write voltage, $V_{d\_write\_1} = 1.25$ V. The ZettaRAM write latency at this voltage is 8.2 ns (a graph is not included here due to space constraints), similar to the DRAM write latency (8.6 ns) reported in the previous subsection. This means that, for $V_{z\_write\_1} = V_{d\_write\_1}$, the conventional peripheral circuitry used to access the molecular capacitor limits the speed, not the intrinsic speed of the molecules.

The ZettaRAM molecular capacitor can be reliably written below 1.25 V, although the intrinsic speed of the molecules begins to limit the overall write speed at lower voltages. The SPICE results in Fig. 6b show exponentially increasing write latency with decreasing write voltage: 9 ns at 1.2 V, 29 ns at 1.1 V, and 166 ns at 1.0 V.

Reading is competitive with conventional DRAM because the read voltage (OCP, Section 2.1) is typically sufficiently lower than V_{ox} such that the molecule current is much faster than the peripheral sensing apparatus and, thus, does not limit the speed of reading. Thus, the read latency of ZettaRAM is dictated by the peripheral sensing circuit, common to both DRAM and ZettaRAM. This is confirmed by SPICE simulations. The SPICE result in Fig. 6a shows that the latency of reading ZettaRAM is 30 ns, similar to the read latency of DRAM (29 ns) measured in the previous subsection. Reading is procedurally similar for the conventional and molecular capacitors—it is based on sensing a small change in charge on the precharged bitline.

Reading the molecular capacitor is tantamount to writing "0" since the read voltage is below V_{ox}, fully discharging the molecular capacitor. So far, we only discussed multiple write voltages for writing a "1." For writing a "0," we consider only a single write voltage equal to the read voltage. Incidentally, this is a fast write voltage. Bitline operations always alternate between reading (open page) and writing (close page), so keeping the write "0" voltage the same as the read voltage eliminates many bitline transitions altogether. Considering slower write "0" voltages between the read voltage and V_{ox} will only increase the number of bitline transitions, thus increasing energy



(a) (b)

Fig. 6. (a) Writing molecular capacitor as low as 1.0 V subsequently results in correct sensing. (b) ZettaRAM write latency (90 percent of molecules oxidized) for three applied voltages.

consumption. This will become clearer in Section 3.5, where we summarize bitline transitions for DRAM and ZettaRAM.

3.4 Retention Time Comparison of DRAM and ZettaRAM

The retention times of the two technologies are comparable because leakage is an artifact of the access transistor and the initial stored charge is the same. This is confirmed by the SPICE results shown in Fig. 7. For example, at 40 ms, the conventional capacitor and molecular capacitor retain 32 percent and 51 percent of the initial charge, respectively. The molecular capacitor demonstrates an improved decay curve at the beginning. The retention times of both memories can be improved by applying a negative substrate bias, reducing the leakage current of the access transistor. What we want to demonstrate here is the comparable retention times.

ZettaRAM exhibits an unconventional linear retention time curve. The molecular capacitor's working electrode, which is connected to the access transistor, decays sharply to V_{ox} and stays close to this voltage level throughout the decay process. In fact, this is due to the nonlinear charge-voltage characteristic described in Section 2.3 and depicted in Fig. 1b (charge drops from fully charged to fully discharged over a narrow voltage interval). Since voltage is nearly constant throughout the decay process, leakage current is nearly constant throughout as well. (Leakage current varies with the voltage at the drain of the access transistor.) Since leakage is nearly constant, the charge decays linearly with time. We confirmed that the slope corresponds to the leakage current of the access transistor.

3.5 ZettaRAM and DRAM: Comparison Summary

Table 1 and Table 2 summarize the SPICE results for DRAM and ZettaRAM (molecule #9 only). Table 1 shows read/write latencies. Table 2 shows operating voltages and implied bitline voltage transitions that depend on consecutive memory operations.

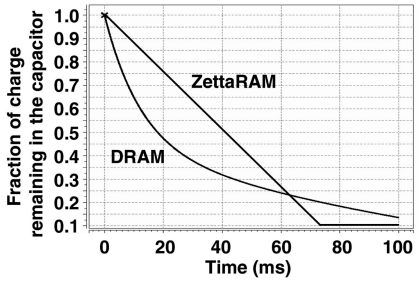


Fig. 7. Retention times.

Bitline energy, which can constitute up to 96 percent of overall energy in DRAM [11], depends on the applied voltage and magnitude of the voltage change ($E_{BL} = C_{BL} \cdot V_{BL} \cdot \Delta V_{BL}$, where E_{BL} is the bitline energy, C_{BL} is the bitline capacitance, V_{BL} is the bitline voltage, and ΔV_{BL} is the bitline voltage change). The actual magnitude of bitline transitions depends on the nature of consecutive operations which cause a voltage change on the bitline— read, write 0, and write 1. We now analyze these individual bitline transitions in depth for both DRAM and ZettaRAM.

Table 2 shows operating voltages and bitline voltage transitions for DRAM and ZettaRAM (molecule #9). Because L2 cache requests are always serviced from the page held in the row buffer, bitline operations always alternate between reading (open page) and writing (close page). This yields only four valid transitions: read followed by write-0, read followed by write-1, write-0 followed by read, and write-1 followed by read. The first row in the table shows the percentage breakdown of these four transitions. One benchmark from the SPEC2K benchmark suite (*mcf*) is shown. The other benchmarks show similar breakdowns. The second row shows the DRAM voltage differential for each transition, using the voltages derived in Section 3.2. Table entries for positive voltage transitions are highlighted, which we use in the energy accounting. Although the previous SPICE experiments used $V_{DD} = 1.6\,V$ due to our available technology files (and a corresponding read precharge voltage of 0.8 V), for energy accounting, we use $V_{DD} = V_{d\_write\_1}$. This adjustment minimizes DRAM energy by applying a lower voltage differential for the higher percentage write-0 $\rightarrow$ read transitions.

The third row shows ZettaRAM voltage differentials, using fast writes, for molecule #9 ($V_{z\_write\_1\_fast} = 1.2$ V). The fourth row shows ZettaRAM voltage differentials, using slow writes, for molecule #9 ($V_{z\_write\_1\_slow} = 1.0$ V). Because the write-0 and read voltages are the same (Section 3.3), two of the transitions incur no voltage change.

Recall that fast writes match the latency of DRAM writes and slow writes increase latency for further energy savings. In Section 7, we propose a hybrid policy, wherein fast writes are used to service critical memory requests and slow writes are used to service noncritical memory requests.

The lack of any write-0 $\rightarrow$ read transitions gives ZettaRAM a substantial energy advantage over conventional DRAM. Conceivably, the same strategy of unifying the read potential and the write-0 potential may be applicable

TABLE 1
Read/Write Latencies for DRAM and ZettaRAM (Molecule #9)

Characteristic	DRAM	ZettaRAM (molecule #9)
Precharge time (write an entire row)	9 ns	Function of applied voltage [9 ns @ 1.2V – 166 ns @ 1V]
Row access time (read an entire row)	29 ns	30 ns

TABLE 2
Operating Voltages and Bitline Voltage Transitions for DRAM and ZettaRAM (Molecule #9)

	Bitline Transition			
	read → write 0	read → write 1	Write 0 → read	write 1 → read
% of all transitions, benchmark = *mcf*	28.46%	21.48%	28.48%	21.58%
Conventional DRAM ΔV	$-(V_{DD}/2)$ = -0.625	$+(V_{d\_write\_1}-V_{DD}/2)$ = 0.625	$+(V_{DD}/2)$ = 0.625	$-(V_{d\_write\_1}-V_{DD}/2)$ = -0.625
Fast ZettaRAM ΔV (#9) ($V_{z\_write\_1\_fast}$ = 1.2 V)	0	$+(V_{z\_write\_1\_fast}-V_{ocp})$ = 0.9	0	$-(V_{z\_write\_1\_fast}-V_{ocp})$ = -0.9
Slow ZettaRAM ΔV (#9) ($V_{z\_write\_1\_slow}$ = 1.0 V)	0	$+(V_{z\_write\_1\_slow}-V_{ocp})$ = 0.7	0	$-(V_{z\_write\_1\_slow}-V_{ocp})$ = -0.7
Baseline DRAM ΔV ($V_{read} = V_{d\_write\_0} = V_{ocp}$ = 0.3V)	0	$+(V_{d\_write\_1}-V_{read})$ = 0.95	0	$-(V_{d\_write\_1}-V_{read})$ = -0.95

in future DRAMs. To level the playing field, we enhance the DRAM by lowering the read potential from $V_{DD}/2$ and raising the write-0 voltage from 0 V, both to V_{ocp} (of molecule #9 in this particular case). (Like ZettaRAM, the enhanced DRAM sense amplifier senses logic "0" via the absence of a bitline shift.) This enhanced DRAM is the baseline for all architectural experiments performed in the subsequent sections. Voltage differentials for this baseline DRAM are shown in the last row of Table 2.

4 Experimental Framework

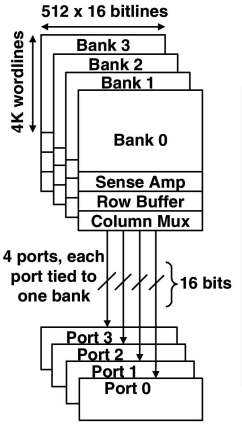
4.1 Memory Simulator: Modeling Timing

The memory simulator models the internal state and operation (timing and functionality) of ZettaRAM. The interleaved ZettaRAM memory system, shown in Fig. 8, is based on the synchronous DRAM (SDRAM) architecture [16].

The ZettaRAM memory system has four independent ports, with each port tied to a bank. The memory controller maps physical addresses to memory addresses (bank id, row id, and column id) and schedules pending memory requests. The memory controller maintains a separate queue of pending memory requests for each bank. There are two types of memory requests initiated by the L2 cache, fetch block and writeback block.

Memory access reordering is used by default. Fetch requests circumvent queued writeback requests unless there is an address match. Where indicated, we also investigate configurations with memory access reordering disabled.

A ZettaRAM *page* is a row in memory that is read into the row buffer to service memory requests. The memory controller can use one of two different policies to manage pages—open page policy and close page policy. In the close page policy, a page is "closed" after servicing the memory request, i.e., the page is immediately written back into its memory array. In the open page policy, a page is left "open" after reading the page into the row buffer, i.e., the data is held in the row buffer (cached). By keeping the page open, subsequent accesses to the same page do not incur the penalty of opening the page. However, if there is a request to a different page in the same bank, the open page policy suffers the penalty of closing the current page before opening the new page, thus sometimes increasing the wait time of fetch and writeback requests. Nonetheless, we find that the open page policy significantly outperforms the close page policy, so we consider only open page policy in our simulations.



ZettaRAM configuration

Chip configuration	8 Mb x 16
# chips	4
# banks per chip	4
# bits per column per chip	16
Data bus width	64 bits
Row Addressing	4K (A0 – A11)
Column Addressing	512 (A0 – A8)
Bank Addressing	4 (BA0 – BA1)
Row Access Time - RAS	30 ns
Col. Access Time - CAS	16 ns
Precharge Time - PRE	Variable (voltage dependent)

Fig. 8. Interleaved ZettaRAM memory system.

4.2 Memory Simulator: Modeling Energy

Bitline energy, i.e., energy required to charge the bitline when opening or closing a page, can constitute up to 96 percent of the total memory system energy [11]. Thus, in our experiments, we measure bitline energy consumption in main memory. We track the voltage states of all bitlines in order to measure the energy required to charge the bitlines for a particular memory operation.

Assuming a single voltage supply (V_{DD}), the energy to charge a bitline is

$$E_{bitline} = C_{BL} \cdot V_{DD} \cdot (\Delta V_{BL}) = C_{BL} \cdot V_{DD} \cdot (V_{z\_write\_1} - V_{ocp}).$$

Thus, dynamically adjusting the write-1 voltage yields linear energy scaling. If we use a dedicated voltage supply for charging the bitline ($V_{z\_write\_1}$), then

TABLE 3
Processor Configuration

Microarchitecture	4-issue OOO superscalar, 7-stage pipeline
Frequency	1 GHz
Reorder Buffer	128 entries
Issue queue, LSQ	64 entries
Function units	4, universal
Branch predictor	gshare, 2^{16} entries
L1 I & D caches (split)	8 KB, 4-way, 64 B line size
L2 cache (unified)	256 KB, 8-way, 128 B line size, writeback
Hit latencies	L1: 2 ns, L2: 10 ns
MSHRs	L1: 32, L2: 8
Bus	400 MHz 64-bit

$$E_{bitline} = C_{BL} \cdot V_{z\_write\_1} \cdot (V_{z\_write\_1} - V_{ocp}).$$

Now, dynamically adjusting the write-1 voltage yields quadratic energy scaling. In this paper, we assume dual voltage supplies for the dual write voltages ($V_{z\_write\_1\_fast}$ and $V_{z\_write\_1\_slow}$). The supplies can be implemented using high-efficiency DC-DC converters [4]. Dual voltages were implemented in drowsy caches and selected in one to two cycles via a MUX [7], a technique we borrow.

The analytical model $C_{BL} \cdot V_{DD} \cdot (\Delta V_{BL})$ is derived by integrating power across the voltage supply $(V_{DD} \times I)$, which yields the overall energy consumed, as opposed to integrating power across only the bitline capacitor $(V_{BL} \times I)$. The analytical model was compared against SPICE simulations and they match exactly.

4.3 Cycle-Level Simulator

Our memory simulator is integrated with a custom detailed cycle-level processor simulator. The SimpleScalar ISA (PISA) [3] and compiler (gcc-based) are used. The processor configuration is given in Table 3. The cache and bus configurations are based on the Pentium 4 processor [9]. The L1 instruction and data caches each allow up to 32 outstanding misses. The L2 cache allows up to eight outstanding fetch requests at a time. Increasing the number of L2 MSHRs beyond eight provided only minor performance benefits. The maximum number of outstanding L2 writeback requests is only limited by the buffering in the memory controller.

4.4 Benchmarks

We use eight different integer benchmarks from the SPEC2000 benchmark suite with reference inputs. We used SimPoint to determine the appropriate starting simulation point for each benchmark [23]. One hundred million instructions are then simulated from this simulation point. The SimPoints chosen for each benchmark are shown in Table 4. Table 4 also shows the rates of L1 and L2 cache misses (per 1,000 instructions) and L2 writebacks (per 1,000 instructions) to main memory for each benchmark.

5 Baseline DRAM Energy and Performance

Fig. 9 shows (a) bitline energy consumption and (b) execution times for DRAM operating at 1.25 V. The request queue size for each bank is fixed at four entries. Memory access reordering is used in the baseline unless otherwise indicated. Since 1.25 V is the lowest reliable write voltage for DRAM, we use this system as our baseline and all ZettaRAM performance and energy measurements are normalized with respect to this baseline.

6 Exploring Molecular Engineering for Long-Term Power Scalability

This section describes how molecular attributes can be engineered to lower the ZettaRAM write voltage from one generation to the next. Synthetic chemists can precisely tune key properties of the molecules through the choice of molecular "groups" and "linkers," such as the oxidation potential, electron transfer rate, and surface concentration (charge density). Many molecules have been synthesized and characterized by ZettaCore. Among these, we identify molecules that yield ZettaRAMs comparable to or better than DRAM in all respects—density (as determined by *cell area*), performance (as determined by *write latency*), and power (as determined by *voltage*).

To compare these ZettaRAMs with DRAM, we fix two of the variables, *cell area* and *write latency*, in order to focus on the third variable, *voltage*. Since cell area is fixed, we consider only those molecules with charge density greater than or equal to that of DRAM so that the minimum amount of charge for reliable sensing is available. For fast writes, we target the same write latency as DRAM. Targeting faster intrinsic molecular speeds is of no use because, as mentioned before, the conventional peripheral circuitry used to access the molecular capacitor limits write latency anyway.

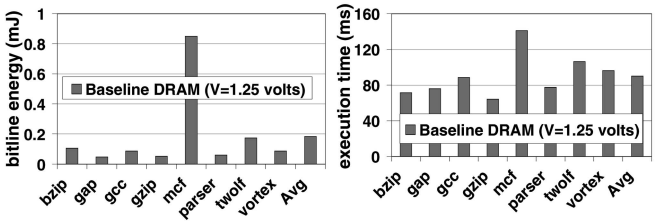
Having pinned down the cell area and write latency, voltage is the only unknown variable. In this situation, voltage is influenced by three key molecular attributes—oxidation potential, electron transfer rate, and surface concentration. We now give insight into how these three molecular attributes affect the write voltage.

1. *Oxidation potential* (V_{ox}): Write latency is determined by the current, which in turn is exponentially dependent on the difference between the applied voltage and the oxidation potential, as seen in (2) of Section 2.2. Since write latency, hence current, is

TABLE 4
SPEC2K Benchmarks

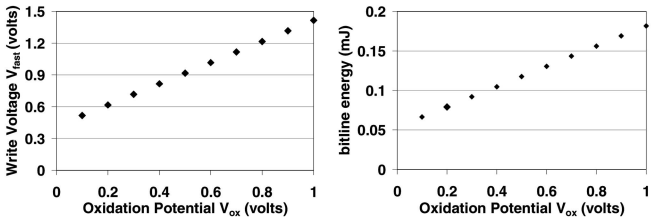
Benchmark	SimPoint (billions of instr.)	L1 misses*	L2 misses*	writebacks*	writebacks that close page*
bzip	1	84.8	13.3	4.6	2.8
gap	209.5	87.8	4.2	1.8	1.2
gcc	11	98.8	9.6	3.13	2.4
gzip	48.7	97.0	4.7	1.91	1.5
mcf	31.7	208.6	80.3	31.84	23.8
parser	1.7	58.9	5.4	2.12	1.5
twolf	3.2	110.5	22.8	7.61	4.9
vortex	5.8	81.2	7.5	2.9	2.4

* *per 1,000 instructions*



(a) (b)

Fig. 9. (a) Bitline energy consumption and (b) execution times for DRAM operating at 1.25 V.



(a) (b)

Fig. 10. Effect of oxidation potential $\mathrm{V_{ox}}$ on (a) write voltage and (b) bitline energy.

fixed, the difference $\mathrm{V} - \mathrm{V_{ox}}$ must be fixed. Thus, a decrease in oxidation potential results in an equal decrease in the write voltage.

2. *Electron transfer rate constant* (k^0): The current also depends on the electron transfer rate constant. A higher rate constant implies that the same current can be generated at a lower write voltage. Therefore, an increase in rate constant results in a decrease in write voltage. However, the relationship is nonlinear because the rate constant is a coefficient of the Butler-Volmer exponential term, whereas voltage is in the exponent.
3. *Surface concentration*: Since cell area is fixed, a higher surface concentration yields more molecules in the cell. However, the molecular capacitor needs to be charged only to the minimum charge that is required by the sense amplifiers for reliable sensing, i.e., the target number of charged molecules is fixed. Therefore, a higher surface concentration implies a correspondingly smaller fraction of the total molecules needs to be charged. Due to the nature of the Butler-Volmer equation, a smaller fraction can be charged faster than a larger fraction, even if the absolute number of charged molecules is the same in both cases (analogous to the radioactive half-life principle—the fraction is what matters). Since we want to fix write latency, we can now offset the higher speed of charging a smaller fraction by lowering the write voltage (slowing it back down to the target write latency). To sum up, we can exploit a higher surface concentration of molecules to lower the write voltage, thus saving energy. However, as with rate constant, the relationship is nonlinear because concentration is a coefficient of the Butler-Volmer exponential term, whereas voltage is in the exponent. In our analyses, $\mathrm{Q_0}$ refers to the total number of molecules present in the molecular capacitor.

The relationship between write voltage and the three molecular parameters—$\mathrm{V_{ox}}$, $\mathrm{k^0}$, and $\mathrm{Q_0}$—is expressed indirectly by (4), which shows the intrinsic molecule write latency $\mathrm{t_{mol\_write}}$ as a function of the molecular parameters $\mathrm{V_{ox}}$, $\mathrm{k^0}$, and $\mathrm{Q_0}$. We fix $\mathrm{t_{mol\_write}}$ and then numerically solve for write voltage, given parameters $\mathrm{V_{ox}}$, $\mathrm{k^0}$, and $\mathrm{Q_0}$ for a particular molecule.

$$t_{mol\_write} = \frac{1}{(k_O + k_R)} \cdot \ln\left\{\frac{k_O}{k_O - (k_O + k_R)\cdot(Q_f/Q_0)}\right\} \quad (4)$$

$$k_O = k^0 \cdot e^{(1-\alpha)\left(\frac{F}{RT}\right)(V - V_{ox})} \qquad k_R = k^0 \cdot e^{-\alpha\left(\frac{F}{RT}\right)(V - V_{ox})}.$$

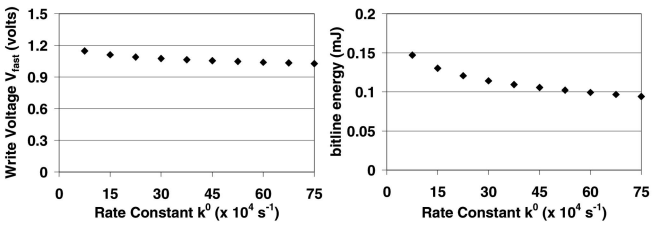
The parameters above are as follows: $\mathrm{k_O} =$ oxidation rate constant; $\mathrm{k_R} =$ reduction rate constant; $\mathrm{Q_f}$ = final charge in the molecular capacitor determined by the minimum charge required by the sense amplifiers for reliable sensing; $\mathrm{Q_0}$ = total molecules in the molecular capacitor, equal to the surface concentration of the molecules multiplied by the fixed cell area; $\mathrm{k^0} =$ standard rate constant; $\alpha =$ transfer coefficient; $\mathrm{F} =$ Faraday constant; $\mathrm{R} =$ gas constant; $\mathrm{T} =$ temperature; $\mathrm{V} =$ applied voltage; $\mathrm{V_{ox}} =$ oxidation potential.

In Section 6.1, we study the sensitivity of write voltage and energy to changes in each of the three molecular attributes—$\mathrm{V_{ox}}$, $\mathrm{k^0}$, and $\mathrm{Q_0}$. In Section 6.2, we apply the above analysis technique to determine the fast and slow write voltages for 23 molecules, based on their attributes obtained from the literature. Section 6.3 presents energy consumption of ZettaRAMs with different molecules and employing uniformly fast writes (corresponding results for hybrid fast/slow writes are presented in Section 7.4).

6.1 Sensitivity of Voltage and Energy to Molecular Attributes—$\mathrm{V_{ox}}$, $\mathrm{k^0}$, and $\mathrm{Q_0}$

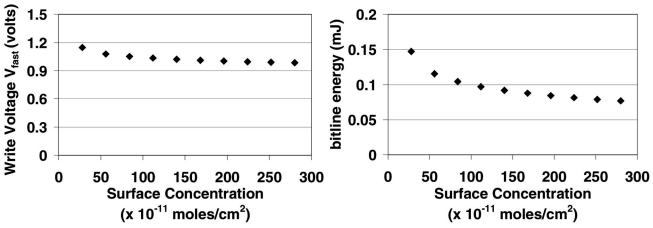
First, we study the effect of increasing the oxidation potential on write voltage and bitline energy. In order to study the effect of oxidation potential in isolation, we fix surface concentration and rate constant to be the same as that of the porphyrin molecule used in our SPICE experiments (molecule #9: $28 \times 10^{-11}\,\mathrm{moles/cm^2}$ and $7.5 \times 10^4\,\mathrm{s^{-1}}$, respectively). Fig. 10 shows the effect of increasing $\mathrm{V_{ox}}$ on (a) write voltage and (b) bitline energy. We observe that write voltage changes by the same magnitude as $\mathrm{V_{ox}}$, for example, decreasing $\mathrm{V_{ox}}$ by 0.9 V results in a 0.9 V decrease in write voltage. Thus, the write voltage is highly sensitive to changes in oxidation potential. Interestingly, the relationship between oxidation potential and bitline energy is also linear, as shown in Fig. 10b, even though bitline energy is proportional to both the write voltage $\mathrm{V_{BL}}$ and the voltage swing $\Delta\mathrm{V_{BL}}$ ($\mathrm{E_{BL}} = \mathrm{C_{BL}} \cdot \mathrm{V_{BL}} \cdot \Delta\mathrm{V_{BL}}$). This is because the write voltage changes by the same amount as the oxidation potential and, as a result, $\Delta\mathrm{V_{BL}}$ remains constant. Nonetheless, from Fig. 10, lowering $\mathrm{V_{ox}}$ by a factor of 10X (from 1.0 V to 0.1 V) yields a 64 percent reduction in both write voltage and energy consumption.

Next, we study the effect of rate constant $\mathrm{k^0}$ on the write voltage and bitline energy. Here, we fix surface concentration and oxidation potential to be the same as that of the porphyrin molecule used in our SPICE experiments (molecule #9: $28 \times 10^{-11}\,\mathrm{moles/cm^2}$ and 0.73 V, respectively). Fig. 11 shows the effect of increasing rate constant $\mathrm{k^0}$



(a) (b)

Fig. 11. Effect of rate constant k^0 on (a) write voltage and (b) bitline energy.



(a) (b)

Fig. 12. Effect of surface concentration on (a) write voltage and (b) bitline energy.

on (a) write voltage and (b) bitline energy. We observe that increasing the rate constant by a factor of 10X (from 7.5 to $75 \times 10^4\ s^{-1}$) yields only a 10 percent decrease in write voltage. However, this corresponds to a more substantial decrease in bitline energy of 36 percent, due to its quadratic dependence on write voltage.

Finally, we study the effect of surface concentration on the write voltage and bitline energy. Here, we fix rate constant and oxidation potential to be the same as that of the porphyrin molecule used in our SPICE experiments (molecule #9: $7.5 \times 10^4\ s^{-1}$ and 0.73 V, respectively). Fig. 12 shows the effect of increasing the surface concentration on (a) write voltage and (b) bitline energy. Increasing surface concentration by a factor of 10X (from 28 to $280 \times 10^{-11}\ moles/cm^2$) yields only a 14 percent decrease in write voltage, which corresponds to a more substantial decrease in bitline energy of 48 percent.

From the above analysis, among the three molecular attributes, oxidation potential has the largest effect on write voltage. Nonetheless, all three molecular attributes have a large effect on bitline energy. Bitline energy only depends linearly on the write voltage, when oxidation potential is varied. Consequently, a 64 percent decrease in write voltage results in a 64 percent decrease in energy consumption. On the other hand, bitline energy depends on the square of the write voltage when the rate constant and surface concentration are varied. As a result, a 10-14 percent decrease in write voltage is magnified to a 36-48 percent decrease in energy consumption.

6.2 Analysis of Molecules

The selected molecules and their attributes are shown in Table 5 [19], [22], [25]. Also shown are the fast and slow write voltages, calculated using the methodology derived earlier in this section. The fast write voltage is used to determine energy with uniformly fast writes. Both the fast and slow write voltages are used to determine energy with our hybrid write policy (dynamic voltage modulation), described in Section 7.

6.3 Energy Savings with Uniformly Fast Writes

Fig. 13 shows ZettaRAM energy consumption normalized to that of DRAM for each molecule type. All write operations are performed at a voltage, V_{fast}, such that the ZettaRAM has the same performance as DRAM. In other

TABLE 5
Molecules with Comparable or Better Charge Density than DRAM

Molecules	Surface Concentration (x 10^{-11} mol/cm^2)	Rate Constant k^0 (x 10^4 s^{-1})	Oxidation Potential V_{ox} (V)	V_{fast} (V)	V_{slow} (V)
(1) Triple decker O2 2/3+	96	9.7	0.3	0.65	0.45
(2) Triple decker O2 3/4+	420	4.5	0.48	0.79	0.59
(3) TD-Tpd(TD-2/3+)	35	10	0.31	0.73	0.53
(4) Triple decker O2 1/2+	25	16	0.24	0.73	0.53
(5) TD-Tpd(TD-3/4+)	96	8.7	0.77	1.13	0.93
(6) TD-Phenylethynylphenyl Linker state 3	27	7	0.52	1	0.8
(7) TD-Phenylethynylphenyl Linker state 3	27	2.2	0.39	0.93	0.73
(8) TD-Tpd(TD-2/3+)	35	8.5	0.76	1.19	0.99
(9) PM-1	28	7.5	0.73	1.2	1
(10) TD-Tpd(TD-3/4+)	96	8.8	1.23	1.59	1.39
(11) ZnP-Tpd(ZnP-1/2+)	25	10.7	0.68	1.19	0.99
(12) Triple decker T1 3/4+	96	8	1.44	1.8	1.6
(13) ZnP-Tpd(ZnP-1/2+)	25	5.5	0.71	1.25	1.05
(14) Triple decker D1 3/4+	64	0.9	0.9	1.4	1.2
(15) Triple decker M1 3/4+	35	7.9	1.23	1.66	1.46
(16) Triple decker M2 2/3+	27	5.8	0.99	1.48	1.28
(17) Fc-ZnP-Tpd (ZnP-1/2+)	25	9.8	1.13	1.64	1.44
(18) TD-Phenylethynylphenyl Linker state 4	35	5.4	1.48	1.93	1.73
(19) Fc-ZnP-Tpd (ZnP-1/2+)	25	4.5	1.03	1.58	1.38
(20) TD-Phenylethynylphenyl Linker state 4	64	0.9	1.3	1.8	1.6
(21) Triple decker T1 2/3+	27	5.4	1.46	1.95	1.75
(22) Triple decker M1 2/3+	25	7.3	1.34	1.87	1.67
(23) Triple decker D1 2/3+	27	4.5	1.68	2.18	1.98

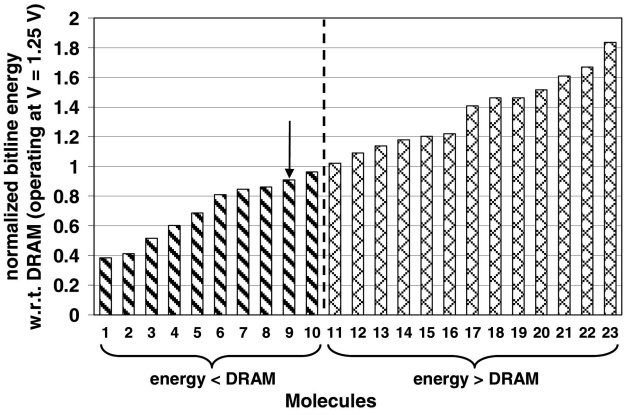


Fig. 13. ZettaRAM with uniformly fast writes (V_{fast}): normalized energy for the 23 molecules.

words, at this voltage, the peripheral circuitry is the overall performance limiter, not the molecules. V_{fast} is different for each molecule and is given in Table 5. Among the 23 molecules, 13 operate with higher energy consumption than DRAM. An unfavorable combination of oxidation potential, rate constant, and surface concentration yields a higher write voltage. Ten molecules operate with lower energy consumption than DRAM. This is because a favorable combination of molecular attributes yields a lower write voltage, resulting in significantly lower energy consumption than DRAM. The best molecule yields 61 percent energy savings, with uniformly fast writes. The porphyrin molecule used in the earlier SPICE experiments is highlighted via the arrow in Fig. 13. Eight different molecules yield lower energy than it.

7 Intelligent Management of ZettaRAM

Recall that the ZettaRAM voltage is padded to achieve competitive performance. Thus, there is room to lower voltage even further if some other means can be found to maintain good performance.

In this section, we describe our hybrid write policy, which dynamically modulates the write voltage based on the criticality of memory requests, thereby maximizing energy savings without degrading overall system performance. We first demonstrate the hybrid write policy and other architectural techniques using molecule #9. The energy consumption of ZettaRAM using molecule #9 and uniformly fast writes is close to the energy consumption of DRAM, a convenient scenario for specifically highlighting the impact of hybrid fast/slow writes. This in-depth evaluation of one molecule is then followed by summarized results for all 23 molecules.

The results presented in the graphs in this section are averaged over the eight SPEC2K benchmarks and normalized to the baseline DRAM, unless stated otherwise.

7.1 Trade-Off between Bitline Energy and System Performance

We first quantify the trade-off between bitline energy and system performance as the ZettaRAM write voltage is changed. Fig. 14 shows (a) bitline energy consumption and (b) execution times, for ZettaRAM operating at fixed write voltages of 1.0 V through 1.25 V in 0.05 volt increments. At 1.25 V and 1.2 V, the execution times for ZettaRAM and the baseline DRAM are equal because the molecules are fast enough above 1.2 V such that the write latency is dictated by the peripheral circuitry. However, at lower voltages, overall write latency is determined by the intrinsic speed of the molecules, degrading system performance.

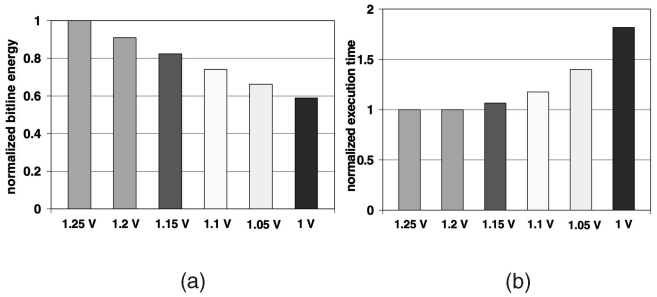


Fig. 14. ZettaRAM (molecule #9) with various write voltages. (a) Bitline energy with respect to DRAM. (b) Execution time with respect to DRAM.

From Fig. 14a, lowering the write voltage from 1.25 V to 1.0 V reduces bitline energy by 41 percent. However, as expected, execution time increases by 81 percent, as shown in Fig. 14b. This is because write latency increases exponentially with decreasing write voltage. Thus, writes done at 1.2 V favor performance, whereas writes done at 1.0 V favor energy savings.

7.2 Hybrid Write Policy

All L2 cache requests to memory are serviced from the row buffer. If the requested page is not in the row buffer, the current page is closed (written back to array) and the requested page is opened (read from array). Closing a page corresponds to a write operation and can be done with either a fast write or slow write, favoring either performance or energy, respectively. Opening a page corresponds to a read operation and is always performed at the OCP. Thus, our hybrid write policy applies to closing a page.

Two types of memory requests are initiated by the L2 cache, fetch block and writeback block. The graph in Fig. 15a shows that 71-82 percent (79 percent on average) of all closed pages are closed because of writebacks that miss in the row buffer. Only 18-29 percent (21 percent on average) of all closed pages are due to fetches that miss in the row buffer. Writebacks exhibit significantly lower locality than fetches, with respect to the row buffer. Fig. 15b shows that fetches miss only 10-20 percent of the time, whereas writebacks miss 60-82 percent of the time (71 percent on average). Because writebacks cause most of the closed pages, they constitute most of the energy savings potential. Therefore, we can tap most of the energy savings potential of ZettaRAM by focusing on writebacks. Moreover, writebacks are not timing critical because they do not stall the processor directly, thereby offering scheduling flexibility to tolerate slow writes. In contrast, fetch requests are timing critical because they stall the processor directly.

Accordingly, we propose a hybrid write policy, where fetches and writebacks are handled differently. If a critical fetch request causes a row buffer miss, the current page is closed using a fast write (high energy). If a noncritical writeback request causes a row buffer miss, the current page is closed using a slow write (low energy). Applying slow writes to writebacks taps most of the energy savings potential and applying fast writes to fetches ensures competitive performance. The pie chart in Fig. 16 illustrates the hybrid write policy. Fast and slow writes are done at 1.2 V and 1.0 V, respectively, for this particular molecule type (#9).

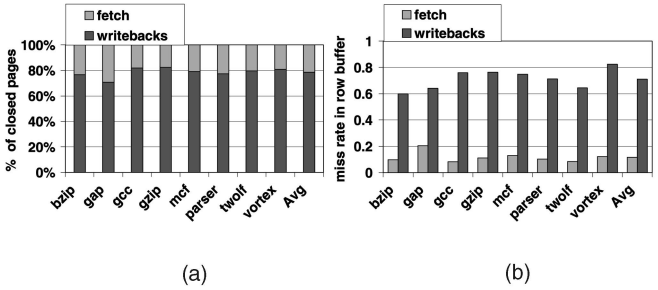


Fig. 15. (a) Percentage of closed pages that are closed due to writebacks versus fetches. (b) Row buffer miss rates for writebacks and fetches.

Fig. 17 shows (a) bitline energy consumption and (b) execution time for ZettaRAM with different write policies. The first bar ("Fast") corresponds to all writes done at 1.2 V. Similarly, the second bar ("Slow") corresponds to all writes done at 1.0 V. These two bars are reproduced from Fig. 14 to facilitate comparisons with the hybrid write policies. The third bar in Fig. 17 shows that the hybrid write policy, as predicted, taps most of the energy savings potential with only a mild system slowdown. The hybrid write policy achieves 34 percent energy savings (third bar in Fig. 17a), out of a possible 41 percent energy savings with uniformly slow writes (second bar in Fig. 17a). Moreover, the hybrid write policy increases execution time by only 10 percent (third bar in Fig. 17b), as opposed to 81 percent with uniformly slow writes (second bar in Fig. 17b). Thus, the hybrid write policy couples the performance of uniformly fast writes with the energy savings of uniformly slow writes.

7.3 Eliminating Residual Slowdown

Although deferred writebacks do not directly stall the processor, they may fill up the request queues in the memory controller, eventually stalling critical fetch requests. This is the cause for the residual 10 percent slowdown.

7.3.1 Hybrid Write Policy Coupled with Large Queues and Access Reordering

Queue pressure can be relieved by enlarging the queues. As expected, the residual slowdown is reduced from 10 percent to less than 1 percent when the queues are enlarged from four to 64 entries, as shown in the fourth bar of Fig. 17b. However, enlarging the queues increases system cost. Each entry contains an entire cache block, thus four 64-entry queues cost 31 KB more than four 4-entry queues. Enlarging the queues also increases complexity and power. Scheduling younger fetches before older writebacks requires searching the entire queue for possible address conflicts.

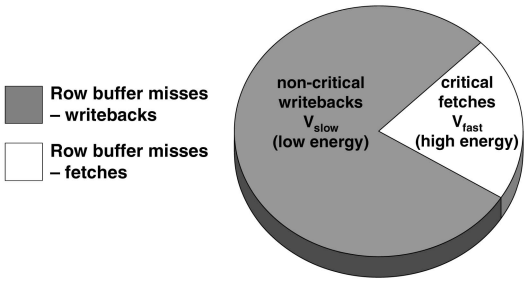


Fig. 16. Hybrid write policy.

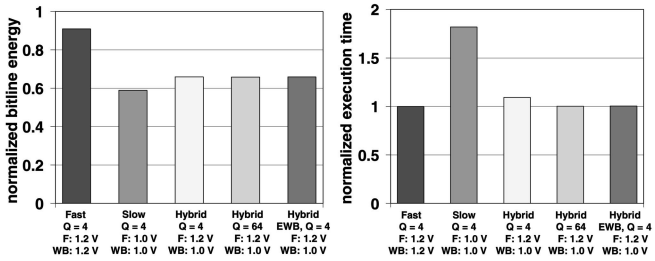


Fig. 17. ZettaRAM (molecule #9) with various write policies. (a) Bitline energy with respect to DRAM. (b) Execution time with respect to DRAM.

7.3.2 Hybrid Write Policy Coupled with L2 Cache Eager Writeback

To avoid the cost, complexity, and power of large queues, we propose employing the eager writeback policy in the L2 cache [14] to evenly spread out writebacks, reducing the frequency of queue stalls. In the eager writeback policy, a writeback is issued as soon as a dirty block becomes the LRU block in its set, instead of waiting for the block to be evicted.

Fig. 18 shows the arrival time (in cycles) of the next request after a writeback request starts closing a page for the hybrid write policy with four queue entries (top graph) and the hybrid write policy with four queue entries coupled with the eager writeback policy in the L2 cache (bottom graph). The measurements are for *mcf* (other benchmarks show similar patterns). With the eager writeback policy, once a writeback request starts closing a page, the next request does not arrive for at least 100 cycles. Without it, a quarter of all next requests arrive between 0 and 100 cycles.

Thus, with eager writebacks, we can probably do well with a small queue in spite of delaying writebacks in the memory controller. *Effectively, issuing the writeback early from the L2 cache compensates for delaying it in the memory controller.*

The fifth bar in Fig. 17 confirms our prediction. For this result, the L2 cache employs the eager writeback policy for both the ZettaRAM and baseline DRAM. We observe that the eager writeback policy improves performance of the

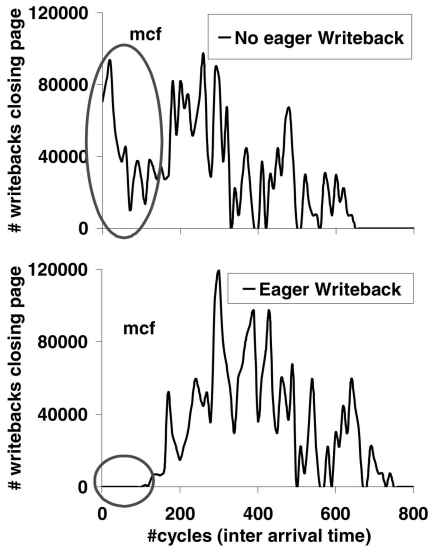


Fig. 18. Arrival time (in cycles) of the next request after a writeback request starts closing a page.

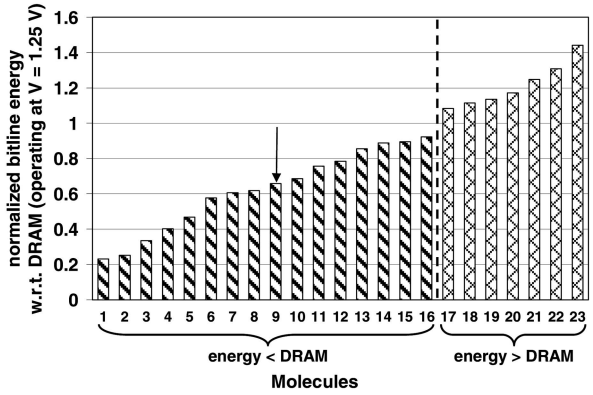


Fig. 19. Normalized bitline energy of ZettaRAM using the hybrid write policy (V_{fast}, V_{slow}) and L2-cache eager writebacks, for 23 different molecules.

baseline DRAM by only 0.6-1.3 percent and bitline energy consumption is unaffected. Fig. 17 shows that the hybrid write policy in conjunction with the L2 cache eager writeback policy yields 34 percent energy savings with negligible slowdown. This is achieved without enlarging the request queues with respect to the baseline system.

7.4 Energy Savings with Hybrid Write Policy for Different Molecules

In this section, the hybrid write policy (coupled with L2-cache eager writebacks) is evaluated for different molecules. The voltages V_{fast} and V_{slow} for each molecule are given in Table 5. Depending on the molecule, the hybrid write policy decreases bitline energy consumption 25-40 percent with respect to uniformly fast writes. Fig. 19 shows energy with respect to the baseline DRAM. The best molecule now yields energy savings of 77 percent (previously 61 percent). Moreover, six of the molecules that yielded higher energy than DRAM with uniformly fast writes now yield lower energy with hybrid fast/slow writes.

8 Related Work

Itoh et al. [13], Mandelman et al. [15], and Teng [24] investigated DRAM scaling trends and concluded that DRAM cell capacitance will remain steady at around 30 fF to 40 fF and will not scale with technology [15], [24]. The charge in the capacitor must be large enough to generate a bitline voltage change that can be reliably sensed, including compensating for various noise sources (radiation, leakage current, and electrical imbalances between pairs of bitlines). While noise induced by radiation decreases with each generation, leakage current and electrical imbalances remain nearly the same from one generation to the next. Therefore, the required charge has not reduced much with each new generation [13], [15].

Itoh et al. quantified energy consumption in main memory and concluded that bitline energy consumption is the main component of the total memory system energy consumption [10], [11].

There has been much work on energy management of DRAM, exploiting multiple low-power modes and multiple banks. Delaluz et al. [5], Fan et al. [6], and Ozturk and Kandemir [18] propose energy management schemes that switch parts of the main memory among four different operating modes (active, standby, nap, power down). We exploit the lower operating voltage of ZettaRAM to reduce power in the active mode. ZettaRAM is based on DRAM architectures. Therefore, all of the above power-saving techniques for transitioning among different power modes are applicable to ZettaRAM.

Galatsis et al. [8] provide an overview of emerging memory technologies, identified by the International Technology Roadmap for Semiconductors (ITRS). These include phase change memory, floating body DRAM, nano floating gate memory, single electron memory, insulator resistance change memory, and molecular memory. Natarajan and Alvandpour [17] discuss the potential of alternative memory technologies such as Ferroelectric RAM, Magnetic RAM, Organic RAM, and Thyristor RAM.

9 Summary

This paper uncovers key properties of ZettaRAM and develops opportunities presented by them for extreme power scaling. ZettaRAM's basis in molecular electronics endows it with charge-voltage decoupling. The molecular capacitor exhibits a threshold voltage, above which all molecules are charged and below which all molecules are neutral. In this way, voltage is only a control mechanism, by way of the threshold, and does not influence the fixed charge itself. This is a powerful concept in the domain of charge-based electronic memories such as DRAM as it enables voltage to be lowered while easily maintaining the critical charge. Without it, DRAM power scaling is severely constrained by the critical charge.

ZettaRAM voltage can be inexpensively lowered by engineering new molecules. This paper contributes the first analyses and insights into how three key molecular attributes influence voltage and energy. In some cases (e.g., concentration), the connection to voltage is subtle. Nonetheless, all three attributes strongly influence energy consumption. This study can guide molecular engineering with a broader, system-level view. Of the 23 molecules evaluated in this paper, the best one yields 61 percent energy savings over DRAM for equal density and performance.

We also study a second property whereby the speed of charging/discharging depends on voltage. Write voltage is padded for competitive latency. Here, there is an opportunity to optimize the speed-voltage trade-off using architectural insights. We develop a hybrid write policy that differentiates between timing critical versus noncritical memory requests, dynamically reducing the write voltage for noncritical requests. This approach successfully combines the energy savings of uniformly slow writes with the high performance of uniformly fast writes.

The static and dynamic voltage scaling techniques enabled by ZettaRAM and developed in this paper may help extend the roadmap of future charge-based electronic memories.

Acknowledgments

The authors thank Ken Mobley and Werner Kuhr for valuable discussions on modeling the molecular capacitor, Rhett Davis for insights on multiple supply voltages, and the anonymous reviewers for their valuable comments. This research was supported by a grant from ZettaCore. This work was performed while Ravi K. Venkatesan and Krishnan Sivasubramanian were with North Carolina State University.

REFERENCES

[1] R.K. Venkatesan, A.S. Al-Zawawi, and E. Rotenberg, "Tapping ZettaRAM for Low-Power Memory Systems," *Proc. 11th Int'l Symp. High-Performance Computer Architecture,* pp. 83-94, Feb. 2005.

[2] A. Bard and L. Faulkner, *Electrochemical Methods: Fundamentals and Applications,* pp. 92-96. John Wiley and Sons, 2001.

[3] D. Burger, T. Austin, and S. Bennett, "Evaluating Future Microprocessors: The Simplescalar Toolset," Technical Report CS-TR-96-1308, Computer Science Dept., Univ. of Wisconsin-Madison, July 1996.

[4] A.P. Chandrakasan and R.W. Brodersen, "Minimizing Power Consumption in Digital CMOS Circuits," *Proc. IEEE,* vol. 83, no. 4, pp. 498-523, Apr. 1995.

[5] V. Delaluz, A. Sivasubramaniam, M. Kandemir, N. Vijaykrishnan, and M.J. Irwin, "Scheduler-Based DRAM Energy Management," *Proc. Design Automation Conf.,* June 2002.

[6] X. Fan, C.S. Ellis, and A.R. Lebeck, "Memory Controller Policies for DRAM Power Management," *Proc. Int'l Symp. Low Power Electronics and Design,* Aug. 2001.

[7] K. Flautner, N.S. Kim, S. Martin, D. Blaauw, and T. Mudge, "Drowsy Caches: Simple Techniques for Reducing Leakage Power," *Proc. Int'l Symp. Computer Architecture,* July 2002.

[8] K. Galatsis, K. Wang, Y. Botros, Y. Yang, Y. Xie, J.E. Stoddart, R.B. Kaner, C. Ozkan, J. Liu, M. Ozkan, C. Zhou, and K.W. Kim, "Emerging Memory Devices," *IEEE Circuits and Devices Magazine,* vol. 22, no. 3, pp. 12-21, May/June 2006.

[9] G. Hinton, D. Sager, M. Upton, D. Boggs, D. Carmean, A. Kyker, and P. Roussel, "The Microarchitecture of the Pentium 4 Processor," *Intel Technology J.,* Q1 2001.

[10] K. Itoh, K. Sasaki, and Y. Nakagome, "Trends in Low-Power RAM Circuit Technologies," *Proc. IEEE,* vol. 83, no. 4, pp. 524-543, Apr. 1995.

[11] K. Itoh, *VLSI Memory Chip Design,* pp. 117-123. Springer Series in Advanced Microelectronics, 2001.

[12] K. Itoh, *VLSI Memory Chip Design,* p. 403. Springer Series in Advanced Microelectronics, 2001.

[13] K. Itoh, Y. Nakagome, S. Kimura, and T. Watanabe, "Limitations and Challenges of Multigigabit DRAM Chip Design," *IEEE J. Solid-State Circuits,* vol. 32, no. 5, pp. 624-634, May 1997.

[14] H.S. Lee, G.S. Tyson, and M.K. Farrens, "Eager Writeback—A Technique for Improving Bandwidth Utilization," *Proc. 33rd Int'l Symp. Microarchitecture,* pp. 11-21, 2000.

[15] J.A. Mandelman, R.H. Dennard, G.B. Bronner, J.K. DeBrosse, R. Divakaruni, Y. Li, and C.J. Radens, "Challenges and Future Directions for the Scaling of Dynamic Random-Access Memory (DRAM)," *IBM J. Research and Development,* vol. 46, nos. 2/3, Mar./May 2002.

[16] Micron SDRAM 8Mx16x4 Part No. MT48LC32M16A2TG-75, 2003.

[17] S. Natarajan and A. Alvandpour, "Emerging Memory Technologies—Mainstream or Hearsay?" *Proc. IEEE Int'l Symp. VLSI Design, Automation, and Test,* pp. 222-228, Apr. 2005.

[18] O. Ozturk and M. Kandemir, "Data Replication in Banked DRAMs for Reducing Energy Consumption," *Proc. Seventh Int'l Symp. Quality Electronic Design (ISQED '06),* Mar. 2006.

[19] K.M. Roth, D.T. Gryko, P.C. Clausen, J. Li, J.S. Lindsey, W.G. Kuhr, and D.F. Bocian, "Comparison of Electron-Transfer and Charge-Retention Characteristics of Porphyrin-Containing Self-Assembled Monolayers Designed for Molecular Information Storage," *J. Physics and Chemistry B,* vol. 106, pp. 8639-8648, 2002.

[20] K.M. Roth, N. Doutha, R.B. Dabke, D.T. Gryko, C. Clausen, J.S. Lindsey, D.F. Bocian, and W.G. Kuhr, "Molecular Approach toward Information Storage Based on the Redox Properties of Porphyrins in Self-Assembled Monolayers," *J. Vacuum Science Technology B,* vol. 18, pp. 2359-2364, 2000.

[21] K.M. Roth, J.S. Lindsey, D.F. Bocian, and W.G. Kuhr, "Characterization of Charge Storage in Redox-Active Self-Assembled Monolayers," *Langmuir,* vol. 18, pp. 4030-4040, 2002.

[22] K. Schweikart, V.L. Malinovskii, A.A. Yasseri, J. Li, A.B. Lysenko, D.F. Bocian, and J.S. Lindsey, "Synthesis and Characterization of BiS (S-acetylthio)-Derivatized Europium Triple-Decker Monomers and Oligomers," *J. Inorganic Chemistry,* vol. 42, pp. 7431-7446, 2003.

[23] T. Sherwood, E. Perelman, G. Hamerly, and B. Calder, "Automatically Characterizing Large Scale Program Behavior," *Proc. 10th Int'l Conf. Architectural Support for Programming Languages and Operating Systems (ASPLOS),* Oct. 2002.

[24] C.W. Teng, "DRAM Technology Trend," *Proc. 1995 Int'l Symp. VLSI Technology, Systems, and Applications,* pp. 295-299, 1995.

[25] L. Wei, K. Padmaja, W.J. Youngblood, A.B. Lysenko, J.S. Lindsey, and D.F. Bocian, "Diverse Redox-Active Molecules Bearing Identical Thiol-Terminated Tripodal Tethers for Studies of Molecular Information Storage," *J. Organic Chemistry,* vol. 69, pp. 1461-1469, 2004.

[26] ZettaCore, http://www.zettacore.com, 2006.

[27] R.K. Venkatesan, "Power-Scalable Memory: Exploiting Typical Charge Retention in DRAM and Charge-Voltage Decoupling in ZettaRAM," PhD thesis, North Carolina State Univ., July 2006.

Ravi K. Venkatesan received the BTech degree from the Indian Institute of Technology (IIT) Madras, India, in 1999 and the MS and PhD degrees in computer engineering from North Carolina State University in 2001 and 2006, respectively. He is a senior platform architecture engineer in the Mobile Platforms Architecture Division at Intel, India. From 2001 to 2002, he worked as a PowerPC development engineer at IBM Microelectronics Division, Research Triangle Park, North Carolina. The subject of his PhD dissertation was developing architectural techniques to achieve power-scalable memory, exploiting conventional memory technologies (DRAM) as well as emerging ones (ZettaRAM). His research interests include the broad areas of high-performance and low-power processor and memory architectures. He is a student member of the IEEE.

Ahmed S. AL-Zawawi is a PhD candidate in the Department of Electrical and Computer Engineering at North Carolina State University. He received the BS (2001) and MS (2002) degrees in computer engineering from North Carolina State University. His research interests include computer architecture, high-performance microarchitecture, and compiler optimization.

Krishnan Sivasubramanian received the bachelor of science degree in computer engineering in 2006 from North Carolina State University. A graduate of the University Honors Program, he is a student member of the IEEE and a member of the National Society of Collegiate Scholars. His research interests include embedded systems, molecular electronics, and underwater robotics. He currently works as an information systems engineer for The Vanguard Group, Inc. in Charlotte, North Carolina.

Eric Rotenberg received the BS degree in electrical engineering (1991) and the MS and PhD degrees in computer sciences (1996, 1999) from the University of Wisconsin-Madison. He is an associate professor of electrical and computer engineering at North Carolina State University. From 1992 to 1994, he participated in the design of IBM's AS/400 computer in Rochester, Minnesota. His research interests are broadly in the area of computer architecture. He currently has research projects on novel high-performance processor architectures, fault-tolerant processor architectures, architectures for real-time embedded systems, and low-power techniques for conventional and nascent memory technologies. He is a member of the IEEE.