The State of ZettaRAM

(Invited Paper)

Eric Rotenberg
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC USA
ericro@ece.ncsu.edu, http://www.tinker.ncsu.edu/ericro

Ravi K. Venkatesan
Mobile Platforms Architecture Division
Intel Technology India Private Limited
Bangalore, India
ravi.k.venkatesan@intel.com

***Abstract*—ZettaRAM is a nascent memory technology with roots in molecular electronics. ZettaRAM patents and papers are distilled and consolidated into a unified discussion. Various embodiments and key novel properties are discussed with a bias toward computer architecture and system design implications. Embodiments include transistor-free crossbar arrays and two hybrid molecule/silicon implementations, a Flash-like cell and a 1T-1C DRAM cell. Key properties of the core technology include (1) flexibility and precision through molecular engineering, (2) self-assembly, (3) scalability through charge-voltage decoupling, (4) speed/energy tradeoff, (5) multiple discrete states, and (6) mixed molecules. Implications include inexpensive fabrication of high performance memory (by all metrics), practical mixed logic/DRAM, 3D memory, exceeding DRAM power scaling limits, intelligent power management, efficient multi-bit storage, memory hierarchies cohabiting the same space, and multiple virtual products in one physical product. Thus, molecular memory has qualities of a disruptive technology. Computer architects and system designers should play a central role in charting its use.**

I. Introduction

Computer architectures are heavily influenced by parameters imposed by memory technologies. Memory hierarchies, virtual memory, prefetching, multithreading, and large-window processors are some well-known examples of architectural innovations influenced by memory constraints.

This paper surveys ZettaRAM™, a nascent memory technology based on molecular electronics. From patents and papers, we distill a number of embodiments and key properties of the core technology. This consolidation lays the foundation for researchers to explore how ZettaRAM may influence computer architectures based on favorable properties and shifting trade-offs.

II. Brief History

The core technology underlying ZettaRAM had its genesis in the DARPA molecular electronics program circa 1999. A 2x2 crossbar array was demonstrated by the originating scientists at the University of California – Riverside (electrochemistry side) and North Carolina State University (molecule synthesis side) [7][15][17]. Later, electrical and computer engineering researchers from North Carolina State University joined, where the key need was scaling from manual fabrication in the lab to wafer fabrication [18]. Other embodiments besides crossbars evolved, to leverage industry's investment and know-how in semiconductor fabrication. This evolution provided the basis for a startup company, ZettaCore™, founded by the originating scientists [25].

III. Core Technology

The core technology consists of a self-assembled monolayer of charge-storage molecules sandwiched between two electrodes [16], as shown in Figure 1(a). The class of molecules is called porphyrins, an example of which is shown in Figure 1(b). Another class of molecules, called ferrocenes, is also used. Within these classes, many different molecules can be synthesized to precisely "engineer" desired values for molecular attributes.

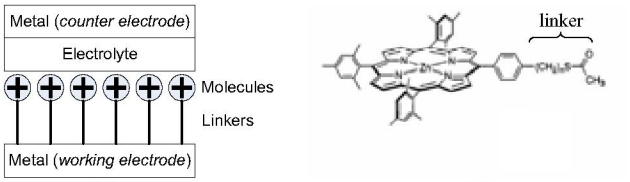

Figure 1. (a) ZettaRAM Molecular Capacitor. (b) Individual porphyrin molecule and linker.

As shown in Figure 1(a), the charge-storage molecules are attached to one of the electrodes, the working electrode, via attachment groups called linkers. Notice, in Figure 1(b), the length of the linker can be selected by choosing the desired number of CH_2 groups in series. The length of the linker provides another dimension for engineering attributes.

The second electrode, the counter electrode, is interfaced to the molecules via an electrolyte.

Neither the electrolyte nor linkers are electron conductors. Thus, when the molecules are charged, the charge is isolated and thereby retained.

The molecules are positively charged via oxidation: one electron is removed from each molecule when a positive voltage greater than the oxidation potential, V_{ox}, is applied to the working electrode relative to the counter electrode. Although the linker is non-conductive, the applied voltage causes an electron to *tunnel* from the molecule to the working electrode, through the linker layer.

Conversely, molecules are discharged via reduction: one electron is returned to each molecule if the voltage applied to the working electrode relative to the counter electrode is less than V_{ox}. In this case, electrons tunnel from the working electrode to the molecules, through the linker layer.

1-4244-0391-X/06/$20.00 ©2006 IEEE

Charged molecules (+1 state) correspond to a one stored in the device. Likewise, discharged molecules (0 or neutral state) correspond to a zero stored in the device. As discussed later, more complex molecules have been developed that support more than just two charge states (0, +1, +2, and higher) [1][2][4][6][10], providing multi-bit storage.

The electrolyte not only interfaces the counter electrode to the molecules, it also provides charge shielding. The electrolyte consists of positive and negative ions that form aligned dipoles when the molecules are positively charged. Electric field lines emanating from the positively charged molecules are sinked by both the working electrode on one side and the electrolyte's negative ions on the other side. Without the electrolyte, an unsustainable high electric field would form across the very short linkers.

Read, write, and storage operations are performed as follows. For all operations, the counter electrode is gounded.

- Write one: A voltage higher than V_{ox} is applied to the working electrode. This charges the molecules.
- Write zero: A voltage lower than V_{ox} is applied to the working electrode. This may be ground or the OCP (see read operation below). This discharges the molecules.
- Read: Reading is functionally equivalent to writing a zero. If the molecules are already charged (one), then applying a voltage below V_{ox} causes them to discharge. The discharge current is detected by a current sense amplifier [19], signaling a one. If the molecules are already discharged, then there is no discharge current, signaling a zero. Regardless of the state of the molecules, a separate and unwanted discharge current is produced by the electrolyte unless the applied voltage is equal to the open circuit potential (OCP), a known artifact of electrochemical cells [17]. Thus, reading is performed at the OCP, which may be different from ground but still less than V_{ox}. Notice that reading is destructive, so the value that is read must be immediately written back, similar to conventional DRAM.
- Storage mode (no change to the state of the molecules): The working electrode is electrically isolated via a switch, so that neither charging nor discharging occurs (no current). While the core technology is volatile, intrinsic retention times on the order of tens to hundreds of seconds have been measured in the lab [15]. However, in the 1T-1C DRAM-like embodiment introduced later, leakage current in the conventional access transistor (an imperfect switch) may limit retention times to only seconds as in conventional DRAM [24].

IV. Three Embodiments

A. Crossbar

The first memory architecture to be proposed was a simple crossbar with no semiconductor components (no transistors, diodes, etc.) [7][15]. A crossbar consists of two wire planes, as shown in Figure 2. Parallel wires run east-west in one plane and north-south in the other plane. Thus, the wires are equivalent to wordlines and bitlines (respectively) in conventional memory architectures. A continuous, non-patterned, self-assembled monolayer (SAM) of molecules (along with corresponding linkers and electrolyte) is sandwiched between the two wire planes.

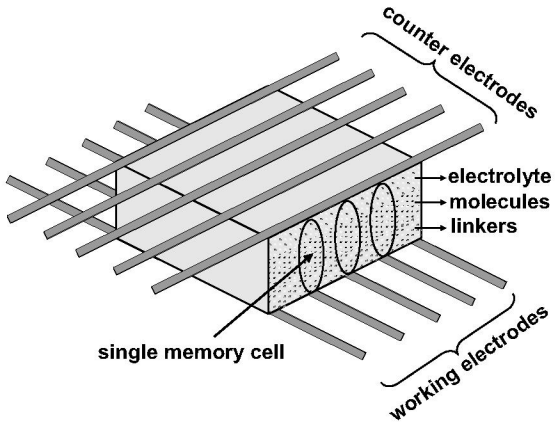

Figure 2. Crossbar embodiment.

The molecule layer does not need to be explicitly patterned to create a 2D array of memory cells. The key idea is that only those groups of molecules situated between intersecting east-west and north-south wires form the memory cells. Thus, the top wire plane provides counter electrodes and the bottom wire plane provides working electrodes. Other molecules not between intersecting wires are inert, as they have no electrodes.

A single memory cell may consist of hundreds or thousands of molecules, depending on molecule concentration (which can be engineered) and wire pitch. A single memory cell is selected for reading or writing, by asserting its corresponding counter electrode wire and working electrode wire. Read and write operations, as controlled by the two electrodes, were explained in Section III. Multiple cells in a whole row can be written with different logic values at the same time, by grounding their shared counter electrode and applying desired voltages (greater than or less than V_{ox}) on their separate working electrodes.

Because there are no semiconductor components (transistors, diodes, etc.), the crossbar implementation is simple to fabricate and dense. No devices need to be patterned. Moreover, cells do not need to be lithographically patterned and etched because groups of molecules sandwiched between two wires implicitly form memory cells. Because there are no other components, density is only limited by the wire pitch. While conventional flash and DRAM technologies may also be able to approach this ideal density, "hiding" extra components in the third dimension is complicated and expensive.

B. MoleFET

The second memory architecture is modeled after conventional Flash memory [3][12][13][14].

A conventional Flash cell consists of a special *field effect transistor* (FET), as shown in Figure 3(a). The special FET

gets its non-volatile memory capability from the "floating gate" embedded in the gate oxide. By applying a high drain voltage, electrons can be tunneled from the floating gate through the thin oxide barrier to the silicon channel. This leaves the floating gate positively charged indefinitely or until explicitly erased by reverse tunneling. The Flash cell is read by turning the transistor on, i.e., apply a positive gate-to-source voltage V_{gs} and positive drain-to-source voltage V_{ds} (V_{ds} is lower than the programming voltage described above, thus the charge state of the floating gate is not disturbed). The charge state of the floating gate modulates the drain current. A positively charged floating gate decreases the perceived threshold voltage (turn-on voltage) of the transistor, inducing a larger drain current when the transistor is turned on, compared to a neutral floating gate. Thus, a larger than nominal drain current signals a one and a nominal drain current signals a zero. The read operation is non-destructive.

The MoleFET embodiment is similar to a Flash FET, except that the floating gate is replaced with the linkers, molecules, and electrolyte, as shown in Figure 3(b). Otherwise MoleFET operation is essentially the same. Positively charged vs. neutral molecules correspond to a positively charged vs. neutral floating gate. The charge state of the molecules modulates the drain current for distinguishing between a one and a zero, when reading the MoleFET.

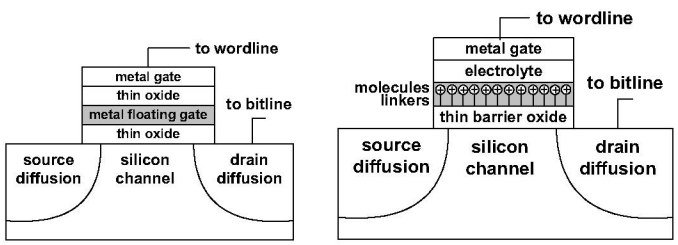

Figure 3. (a) Conventional Flash cell. (b) MoleFET.

While the linkers alone may provide a sufficient tunneling barrier (as explained in Section III), the thin barrier oxide is needed for non-destructive reads. As with the Flash FET, the MoleFET is read by applying a positive gate voltage to turn the transistor on. Since the gate is the counter electrode, the working electrode (silicon channel) is negatively biased with respect to the counter electrode, thus applying a voltage less than V_{ox} to the molecules. Without the thin barrier oxide, this bias arrangement reverse tunnels electrons from the silicon channel through the linkers to the molecules, discharging the molecules if they are initially positively charged. The issue here is that the core technology does not provide "hysteresis": charging and discharging occurs just above and below a single voltage, V_{ox}. The thin barrier oxide provides hysteresis [3][13]: the working electrode (with respect to counter electrode) must be some delta lower than V_{ox} to discharge the molecules.

As we will discuss in Section V, the discrete charge of molecules in the MoleFET provides a level of precision that is difficult to achieve with non-discrete charge of the floating gate in conventional Flash. This is especially advantageous for implementing multiple bits per cell more efficiently than conventional multi-bit Flash.

C. DRAM Derivative

The third memory architecture is modeled after conventional DRAM [8][9]. Compared to the crossbar and MoleFET, this embodiment seems to diverge the least from conventional memory. Because it is so close to conventional DRAM, this embodiment seems most promising in the near term.

A conventional 1T-1C DRAM cell consists of one transistor (1T) and one capacitor (1C). A 1T-1C DRAM cell is shown in Figure 4(a). The transistor controls access to the capacitor, which stores the bit.

ZettaRAM uses the same 1T-1C cell, except the conventional capacitor is replaced with a molecular capacitor. The molecular capacitor is the core device described earlier in Section III and depicted in Figure 1(a). Instead of attaching the linkers to a metal working electrode, the linkers can be attached directly to the drain diffusion of the transistor, as shown in Figure 4(b). The drain is the working electrode.

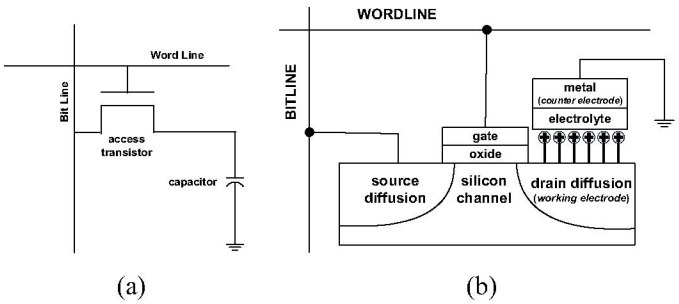

Figure 4. (a) 1T-1C DRAM cell. (b) DRAM with molecular capacitor.

Replacing the conventional capacitor with the molecular capacitor yields a more scalable form of DRAM, as we discuss in Section V.

V. Key Properties

A. Flexibility and Precision

Hundreds of different porphyrin and ferrocene molecules have been synthesized and are being characterized by ZettaCore. Synthetic chemistry offers significant flexibility in customizing molecular attributes, through the design of organic molecules and attachment groups.

While semiconductors have also demonstrated significant flexibility, sophisticated "recipes" and expensive facilities are needed to carefully control doping levels, bake times, etch times, etc.

The complexity of semiconductors arises from the fact that key attributes such as threshold voltage depend on bulk properties (e.g., dopant concentration). In contrast, intrinsic chemical properties of molecules, such as oxidation potential, yield precision with low cost and complexity.

Molecule selection provides precise control over characteristics such as the electron transfer rate (affecting the speeds of reading and writing), the threshold voltage V_{ox} (affecting read and write power consumption), and monolayer density (affecting charge density and thus overall memory density).

B. Self-Assembly

Self-assembly is a process by which molecules automatically arrange themselves into a single, uniform, dense monolayer. The quality of the monolayer may be affected by imperfections in the substrate to which the linkers attach, whether metal, oxide, silicon, or some other surface. Nonetheless, the autonomous and parallel nature of self-assembly is efficient from a fabrication standpoint.

Self-assembly leads us to reconsider possibilities that are impractical with conventional memory technology.

- *Mixed logic/DRAM chips (DRAM embodiment).* Traditionally, DRAM and logic processes are very different. This is due in part to the processing steps needed to fabricate stacked capacitors in DRAM. Self-assembled monolayers of molecules yield significant charge density without having to construct unwieldy stacked capacitors. There is still the problem of high leakage in logic processes that reduce DRAM retention times. Perhaps the thin barrier oxide used in the MoleFET can be reapplied to the molecular capacitor of the DRAM embodiment, yielding a longer-retention molecular capacitor more resistant to the high leakage of logic processes. The oxide will slow reading and writing. Oxide thickness should be carefully selected to balance the retention-time/speed tradeoff.

- *3D stacking (crossbar embodiment).* The molecules can self-assemble in arbitrary places as long as there is a compatible attachment surface. This suggests a convenient path toward 3D memory stacking. The crossbar implementation is most suited to this, as there are no semiconductor devices. Given that multiple metal layers are common, it is conceivable that multiple crossbars can be stacked vertically.

C. Charge-Voltage Decoupling

Venkatesan et al. [22][23] identify charge-voltage decoupling in ZettaRAM as a means for extending the limits of DRAM voltage (power) scaling.

In conventional DRAM, charge is constrained by the equation Q=CV (Q is charge, C is capacitance, V is write voltage). This constraint makes it difficult to simultaneously scale DRAM density and voltage from one generation to the next. Sufficient Q is needed for reliable sensing. Reducing the 2D area of the capacitor (higher density) is already difficult: it requires building a correspondingly taller capacitor to keep C the same, hence, Q the same. If we also want to lower the operating voltage, C must actually be increased to compensate. Yet, C is predicted to remain relatively constant [20], limiting voltage scaling (hence power scaling) in future generations of DRAM.

Replacing the conventional capacitor with the molecular capacitor offers a reprieve. As explained in Section III, the molecules are fully charged/discharged when the applied voltage is slightly above/below the threshold voltage V_{ox}. Note that the *amount* of fixed charge does not depend on the voltage. Voltage only controls whether the molecules are charged or discharged. Decoupling the amount of charge from voltage means that the operating voltage can be reduced arbitrarily from one memory generation to the next, while keeping Q constant for reliable sensing. Molecular engineering has already demonstrated a wide range of threshold voltages. For example, Venkatesan identified one molecule that yields similar performance and charge density as the current generation of DRAM, with an operating voltage of only 0.65V, compared to 1.25V for conventional DRAM [23]. Thus, the ZettaRAM form of DRAM may help extend the roadmap of this important memory technology.

D. Speed/Energy Tradeoff

Although the amount of charge does not depend on voltage, the intrinsic speed of the molecules increases exponentially with the difference between the applied voltage and the threshold voltage V_{ox}. Charging/discharging is very fast if the difference is just a few tenths of a volt. The conventional CMOS peripheral circuitry used to access the molecular memory is the bottleneck in this regime. However, the molecule latency increases dramatically as applied voltage approaches V_{ox}.

Thus, for good performance, the operating voltage must be "padded" with respect to V_{ox}. For example, the 0.65V operating voltage reported in the previous section includes some padding.

There is an opportunity to further reduce the power consumption of ZettaRAM, if good performance can be assured without exclusively relying on padding voltage. Venkatesan et al. exploit architectural insights to intelligently manage the novel speed/energy tradeoff [22]. They propose a dynamic voltage scaling approach, in which a nominal write voltage is used for critical requests that noticeably affect system performance, and a lower voltage is used for non-critical requests. In their study, they find that only 20% of memory requests are critical. Their hybrid fast/slow write policy is able to achieve most of the energy savings of uniformly slow writes with the performance of uniformly fast writes.

Intelligent management of ZettaRAM hints at the possibilities that may exist for innovation through a combination of computer architecture and technology.

E. Multiple Discrete States

In addition to molecules with two charge states (neutral and +1) for storing a single bit, molecules have been developed with more than two charge states for storing multiple bits [1][2][4][6][10].

Multi-bit storage has been implemented in some commercial Flash memories. For example, two bits are

implemented by charging the floating gate to one of four charge levels, corresponding to the 00, 01, 10, and 11 states. The difference between successive charge levels causes a sufficient difference in drain current to reliably and quickly determine the state. One way to vary the charge level of the floating gate, and thereby program a 00, 01, 10, or 11 into the cell, is to vary the programming time. Since charge is a continuous quantity in this context, there is bound to be some variability which may be addressed via careful design, conservative noise margins, and the like.

The discrete states of molecules provide a more efficient multi-bit storage solution. The drain current varies discretely with the charge state, with no intrinsic variability from the molecular memory itself. Moreover, this property can be exploited in all three embodiments, whereas multi-bit storage has traditionally been designated for Flash memory.

F. Admixtures

We are not limited to using only one type of molecule in a chip. Different molecules can be mixed [21]. While hybrid technologies are commonplace, nanotechnology offers a new twist in that different molecules can be mixed in the *same physical space* either as admixtures or by stacking molecules. Thus, molecules with different speeds, different retention times, different threshold voltages, and different concentrations can co-exist in the same macro space. This opens up new possibilities to computer architects, as it adds a new dimension to implementing memory hierarchies. Consider that a fast, low-voltage, short-retention primary storage can cohabit the same space as a slow, high-voltage, long-retention secondary storage. This unusual hierarchy presents new challenges and opportunities for optimizing data "placement" for power and performance. Admixtures enable different business models, such as shipping a product with multiple molecules but configuring it to use only one molecule or the other, thus implementing many virtual products in one physical product.

VI. Summary

A key contribution of this work is consolidating, distilling, and interpreting high-level knowledge about ZettaCore's molecular memory technology and three embodiments of this technology. Key properties are also consolidated into a unified discussion. In some cases, the unified framework reveals new insights, ruminations, and implications that have not been explicitly published elsewhere. Overall, ZettaRAM has qualities of a disruptive technology. It also has many properties that can significantly influence computer architectures and system design. Thus, computer architects and designers have a central role to play in charting the use of this new technology. Ultimate prospects depend on the commercialization efforts currently underway.

References

[1] C. Clausen et al. Synthesis of Thiol-Derivatized Porphyrin Dimers and Trimers for Studies of Architectural Effects on Multibit Information Storage. J. Org. Chem., 65, 7363-7370.

[2] C. Clausen et al. Investigation of Tightly Coupled Porphyrin Arrays Comprised of Identical Monomers for Multibit Information Storage. J. Org. Chem., 65, 7371-7378.

[3] S. Gowda et al. Hybrid Silicon/Molecular FETs: A Study of the Interaction of Redox-Active Molecules With Silicon MOSFETs. IEEE Transactions on Nanotechnology, 5(3):258-264, May 2006.

[4] D. Gryko et al. Studies Related to the Design and Synthesis of a Molecular Octal Counter. J. Mater. Chem.Vol. 11, pp. 1162-1180, 2001.

[5] D. T. Gryko et al. Synthesis of "Porphyrin-Linker-Thiol" Molecules with Diverse Linkers for Studies of Molecular-Based Information Storage. J. Org. Chem., 65, 7345-7355, 2000.

[6] D. T. Gryko et al. Synthesis of Thiol-Derivatized Ferrocene-Porphyrins for Studies of Multibit Information Storage. J. Org. Chem., 65, 7356-7362, 2000.

[7] D. T. Gryko et al. High Density Non-volatile Memory Device Incorporating Thiol-derivatized Porphyrins.US Patent #6,208,553, 2001.

[8] W. G. Kuhr, A. R. Gallo. Molecular Memory Devices and Methods. US Patent Publication # 20060092687, May 2006.

[9] W. G. Kuhr, A. R. Gallo. Molecular Memory Arrays and Devices. US Patent Publication # 20050162895, July 2005.

[10] J. Li et al. Synthesis of Thiol-Derivatized Europium Porphyrinic Triple-Decker Sandwich Complexes for Multibit Molecular Information Storage. J. Org. Chem., 65, 7379-7390, 2000.

[11] Z. Liu et al. Molecular Memories That Survive Silicon Device Processing and Real-World Operation. Science.302:1543-1545, 2003.

[12] G. Mathur, S. Gowda, V. Misra. Threshold Voltage-Assisted Reduction of Molecules in Hybrid Silicon/Molecular Memory Devices. 5th IEEE Conference on Nanotechnology, Vol. 1, pp. 442 – 445, July 2005.

[13] G. Mathur et al. Hybrid CMOS/Molecular Memories using Redox-Active Self-assembled Monolayers. 3rd IEEE Conference on Nanotechnology, Vol. 1, pp. 307 – 310, Aug. 2003.

[14] V. Misra et al. Method and System for Molecular Charge Storage Field Effect Transistor. US Patent #6,674,121, Jan. 2004.

[15] K. M. Roth et al. Comparison of Electron-Transfer and Charge-Retention Characteristics of Porphyrin-Containing Self-Assembled Monolayers Designed for Molecular Information Storage. J. Phys. Chem. B, 106, 8639-8648, 2002.

[16] K. M. Roth et al. Molecular Approach Toward Information Storage Based on the Redox Properties of Porphyrins in Self-Assembled Monolayers. J. Vac. Sci. Technology B, 18, 2359–2364, 2000.

[17] K. M. Roth et al. Characterization of Charge Storage in Redox-Active Self-Assembled Monolayers. Langmuir, 18, 4030–4040, 2002.

[18] Q. Li et al. Capacitance and Conductance Characterization of Ferrocene-containing Self-assembled Monolayers on Silicon Surfaces for Memory Applications. Appl. Phys. Lett. vol. 81, pp. 1494-1496, 2002.

[19] Y. Nishida and W. Liu. An Interpolating Sense Circuit for Molecular Memory. IEEE Custom Integrated Circuit Conference, pp.103-106, May 2002.

[20] J. A. Mandelman et al. Challenges and Future Directions for the Scaling of Dynamic Random-Access Memory (DRAM). IBM J. Res. and Dev. Vol. 46, No. 2/3, Mar/May 2002.

[21] E. Rotenberg, J. Lindsey. Variable-Persistence Molecular Memory Devices and Methods of Operation Thereof. US Patent #6,944,047, Sep. 2005.

[22] R. K. Venkatesan, A. S. Al-Zawawi, and E. Rotenberg. Tapping ZettaRAM™ for Low-Power Memory Systems. HPCA-11, pp. 83-94, Feb. 2005.

[23] R. K. Venkatesan, A. S. Al-Zawawi, K. Siva, and E. Rotenberg. ZettaRAM™: A Power-Scalable DRAM Alternative through Charge-Voltage Decoupling. IEEE Transactions on Computers, Special Issue – Nano Systems and Computing. In press.

[24] R. K. Venkatesan, S. Herr, and E. Rotenberg. Retention-Aware Placement in DRAM (RAPID): Software Methods for Quasi-Non-Volatile DRAM. HPCA-12, pp. 157-167, Feb. 2006.

[25] ZettaCore - http://www.zettacore.com