On The Energy-Efficiency of Byte-Addressable Non-Volatile Memory


Published in:
IEEE Computer Architecture Letters

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

General rights
Copyright for the publications made accessible via the Queen’s University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen’s institutional repository that provides access to Queen’s research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person’s rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
On The Energy-Efficiency of Byte-Addressable Non-Volatile Memory

Hans Vandierendonck, Senior Member, IEEE
Queen’s University Belfast
E-mail: h.vandierendonck@qub.ac.uk

Ahmad Hassan
SAP Belfast
E-mail: ahmad.hassan@sap.com

Dimitrios S. Nikolopoulos, Senior Member, IEEE
Queen’s University Belfast
E-mail: d.nikolopoulos@qub.ac.uk

Abstract—Non-Volatile Memory (NVM) technology holds promise to replace SRAM and DRAM at various levels of the memory hierarchy. The interest in NVM is motivated by the difficulty faced in scaling DRAM beyond 22 nm and, long-term, lower cost per bit. While offering higher density and negligible static power (leakage and refresh), NVM suffers increased latency and energy per memory access. This paper develops energy and performance models of memory systems and applies them to understand the energy-efficiency of replacing or complementing DRAM with NVM. Our analysis focuses on the application of NVM in main memory. We demonstrate that NVM such as STT-RAM and RRAM is energy-efficient for memory sizes commonly employed in servers and high-end workstations, but PCM is not. Furthermore, the model is well suited to quickly evaluate the impact of changes to the model parameters, which may be achieved through optimization of the memory architecture, and to determine the key parameters that impact system-level energy and performance.

1 INTRODUCTION

One of the advantages that byte-addressable non-volatile memories (NVM) bring to system design is their close-to-zero leakage power and absence of refresh power. These two types of loss will be jointly referred to as static power. Compared to current state-of-the-art DRAM technology, however, NVM incurs a higher to substantially higher dynamic energy as well as increased latency. This trend holds across most of the NVM contenders, and is valid for at least phase-change memory (PCM) and resistive memory (RRAM), while spin transfer torque memory (STT-RAM) may be faster than DRAM [3].

This paper explores the static vs. dynamic energy trade-off of byte-addressable NVM when used as a DRAM replacement or complement. Our approach uses first-order analytical performance and power models in order to make a number of observations on this trade-off which are generally not expressed in the mainstream computer architecture literature.

2 METHODOLOGY

We aim to model memory technologies using a small set of parameters that characterise access latency, access energy and background energy. While memory has many distinct parameters, such as precharge energy, array and row buffer read/write energy and latency, etc., we summarise those parameters in 3 key values:

- Average access latency ($L$, [CPU cycles]), taking into account statistics for read/write distribution and open page hit rate.
- Dynamic energy ($E_d$, [J]), the energy consumed on average when accessing memory, again representing the common case access.
- Static energy ($P_s$, [W/GB]), the background power consumed per GB of memory capacity. Background power is assumed constant over time and includes the refresh energy for DRAM.

2.1 Performance and Energy Models

We assume a first order processor performance model that distinguishes between a baseline cycles per instruction ($CPI_0$), which is increased by a constant amount for every memory access made. We assume that the processor executes in total $N$ instructions, and requires $M$ memory accesses per instruction. The performance model is thus:

$$T(L) = \frac{N}{\phi} (CPI_0 + ML) \quad (1)$$

Here, $\phi$ is the CPU clock frequency. The average memory access latency $L$ comprises demand fetches and cache line writebacks. We assume that 67% of memory accesses are reads in main memory, following industry practice [10]. Our model furthermore assumes that processes are latency-bound, a situation that appears to be true for modern server workloads [7]. Moreover, we assume parallelism in the memory system is increased to offset technological restrictions of NVM.

Our energy model includes dynamic energy ($E_{d,mem}$) of the memory, static power per GB of memory ($P_{s,mem}$) and processor power consumption ($P_{cpu}$).

$$E_{mem} = E_{d,mem} NM + (P_{s,mem} S + P_{cpu}) T(L) \quad (2)$$

The parameter $S$ represents the main memory size. Processor power consumption is modelled as a constant, the value of which depends on the processor, the program and on the ability of the memory to supply data to the processor. We are thus assuming that any performance degradation caused by NVM will impact CPU energy through an increase in execution time ($T(L)$) but not through a change in $P_{cpu}$.

2.2 Comparative Energy Model

The comparative energy model contrasts the energy consumption when using a memory technology $T$ versus when using technology $U$. We vary $T$ and $U$ over DRAM, PCM, STT-RAM and RRAM in this paper.
Assuming main memory is comprised either entirely of technology $T$ or entirely of technology $U$, the energy difference $\Delta E = E_T - E_U$ is equal to:

$$\Delta E = \frac{N}{\phi}(\phi \Delta E_d M + CPI_0 \Delta P_d S + \Delta E_s M S + P_{cpu} M \Delta L)$$

(3)

where $\Delta E_d = E_{d,T} - E_{d,U}$, the difference in dynamic energy per memory access, $\Delta P_d = P_{s,T} - P_{s,U}$, the difference in static energy per gigabyte, and $\Delta E_s = P_{s,T} L_T - P_{s,U} L_U$, the difference in static energy attributable to the difference in memory access time.

We use the energy equation to characterize the workloads for which either technology $T$ or $U$ consumes less energy. We characterize workloads by two key properties: $M$, the memory access rate of the workload, and $S$, the memory footprint of the workload. Each (hypothetical) workload is depicted in a 2-dimensional space by assigning it to the point in the space whose $X$-coordinate corresponds to the value of $S$ and its $Y$-coordinate corresponds to the value of $M$. Then, for each coordinate, we evaluate $\Delta E$. For some workloads (values of $M$ and $S$), $\Delta E$ is positive, while for other workloads it is negative.

We identify the boundary between workloads that favour one of the technologies $T$ and $U$ by solving the equation $\Delta E = 0$. There exist three non-trivial types of solutions to this equation, as depicted in Figure 1. The governing situation depends on the values of the energy and latency parameters.

Interestingly, in all three scenarios the boundary $\Delta E = 0$ is governed by its asymptotic behavior for infinitely large $M$ and $S$. We calculate these asymptotes as:

$$M^\infty = -CPI_0 \Delta P_d / \Delta E_s$$

$$S^\infty = -(\phi \Delta E_d + P_{cpu} \Delta L) / \Delta E_s$$

(5)

The signs of the differences of the model parameters determine the signs of $M^\infty$ and $S^\infty$, which correlates with the three scenarios indicated in Figure 1. A fourth scenario is possible if all differences carry the same sign, in which case one memory technology trivially outperforms the other.

The scenarios will occur under different conditions. The first scenario (Figure 1(a)) occurs when $\Delta P_d$ and $\Delta E_s$ have the same sign (as $M^\infty < 0$), i.e., the static energy consumed by one technology is less than the other’s and increases in access delay in the first technology do not offset the static energy gain. Moreover, the sign of $S^\infty$ is positive, implying that the technology with less static energy has higher dynamic energy and delay. This is the common scenario when comparing byte-addressable non-volatile memories against DRAM.

The second and third scenarios (Figure 1(b), (c)) occur when the reduction in static energy is offset by increased access latency ($\Delta P_d$ and $\Delta E_s$ have different signs). Moreover, the increase in latency may be coupled with increase in memory access energy (Figure 1(b)), or not (Figure 1(c)).

The presence of the $S^\infty$ threshold in scenario I may seem counterintuitive: when increasing the memory access rate while keeping memory size constant, shouldn’t $U$ consume more energy than $T$ at some point? This is not necessarily the case, as increasing the memory access rate $M$ increases the execution time (Equation 1), and thus also increases static energy. The model predicts that for particular values of dynamic and static energy, there is a situation where static energy grows faster with increasing memory access rate than dynamic energy does.

Alternatively, one may restructure Equation 3:

$$\Delta E = \frac{N}{\phi}(\Delta E_s M (S - S^\infty) + CPI_0 \Delta P_d S + P_{cpu} \Delta L)$$

(6)

Thus, the sign of $M$ in $\Delta E$ depends on whether $S > S^\infty$.

### 2.2 Energy-Delay Product

Similar observations can be made for the energy-delay product (EDP) as for energy, although now $\Delta EDP$ is a quadratic function of $M$:

$$\Delta EDP = \frac{N^2}{\phi^2} (\Delta P_d S CPI_0 l_d^2 + (\phi \Delta E_d + 2 \Delta E_s S + 2 P_{cpu} \Delta L) CPI_0 M + (\phi \Delta EDP_d + EDP_S S + P_{cpu} \Delta L^2) M^2)$$

Here, $\Delta E^2 = L^2_d - L^2_{nvm}$, $\Delta EDP_d = E_d, L_d^{dram} - E_d, L_d^{nvm}$ and $\Delta EDP_S = P_s, L_d^{dram} - P_s, L_d^{nvm}$. We find four scenarios that are qualitatively the same as for energy, with thresholds:

$$M^\infty = -CPI_0(\Delta E_d \pm \sqrt{P_{s,dram} P_{s,nvm} \Delta L}) / \Delta EDP_S$$

$$S^\infty = -(\phi \Delta EDP_d + P_{cpu} \Delta L^2) / \Delta EDP_S$$

$M^\infty$ potentially has two values. Only positive values matter.

### 3 Comparison of DRAM vs. NVM

We compare DRAM and NVM using the methodology outlined above. Model parameter values for non-volatile memory...
thresholds are larger for EDP than they are for energy. This is
Plugging in values for the model parameters shows that the thresholds
for 5 different MPKI values between 2 and 125. By varying the sizes of PCM and DRAM for a given MPKI configuration, we
for 50% MPKI. Thus, the main conclusion that a threshold exists is valid. The thresholds are quite loose for large values. One may define the value $S^\infty$ where $\Delta E = 0$ and $MPKI = 100$, which is a practically meaningful bound. For PCM and the energy equation, $S^\infty$ is 220 GB.

3.2 Energy-Delay Product
Plugging in values for the model parameters shows that the $S^\infty$ thresholds are larger for EDP than they are for energy. This is expected, as NVM increases delay, which must be made up by larger savings in static power.

The results (Table 1) show that PCM is not quite as promising a technology as STT-RAM and RRAM: A memory size of 1.34 TB is required to make PCM outperform DRAM on EDP. However, technologies such as STT-RAM and RRAM can outperform DRAM for memory sizes of 130.9 GB and 75.1 GB, respectively.

### 3.3 NVM Latency Tolerance
The thresholds $M^\infty$ and $S^\infty$ depend on the memory access latencies $L_{dram}$ and $L_{nvm}$ through $\Delta L$ and $\Delta E_s$. We can rewrite the thresholds in function of $L_{dram}$ and $\Delta L$ (avoiding the use of $L_{nvm}$), e.g., for the energy equation:

$$
S^\infty = -(\delta E_d + P_{cpu} \Delta L)/(\Delta P_s L_{dram} + P_s \Delta L)
$$

Note that $P_{cpu}$ is assumed zero, so $S^\infty$ increases in absolute value as $\Delta L$ grows. When $\Delta L$ becomes too large, $S^\infty$ becomes negative and the scenario changes (Figure 1). For a technology like PCM, the contribution of $\Delta L$ to $S^\infty$ is much larger than the contribution of $\Delta E_s$. Consequently, minimising $\Delta L$ at the expense of $\Delta E_d$ increases the applicability of PCM. If, however, NVM is optimised in isolation (i.e., $P_{cpu}$ is assumed zero), then $S^\infty$ is minimized by minimizing $\Delta E_s$. This assumption corresponds to optimizing a single memory characteristic in isolation. Such a solution would be sub-optimal at the system-level, where minimizing $\Delta L$ is more important than minimizing $\Delta E_d$.

### 4 Hybrid Memory Systems
A hybrid main memory consists of $S_{dram}$GB of DRAM memory and $S_{nvm}$ GB of non-volatile memory. A fraction $\mu$ of the memory accesses are directed to NVM, while a fraction $1 - \mu$ is directed to DRAM [12]. We assume reads and writes are distributed evenly across the memory types. We make no particular assumptions on whether memory traffic is distributed between DRAM and NVM by hardware or software.

We define energy per memory access $E_{h,hyb} = (1 - \mu) E_{dram} + \mu E_{nvm}$ and average memory access latency $L_{hyb} = (1 - \mu) L_{dram} + \mu L_{nvm}$. Static energy consumption is calculated as the weighed sum of static energy consumption in both DRAM and NVM.

The energy equation for hybrid memory is given as:

$$
E_{hyb} = E_{d,hyb} M N + (P_{s,dram} S_{dram} + P_{s,nvm} S_{nvm} + P_{cpu}) T(L_{hyb})
$$

### Table 1
Model parameters for non-volatile memory technologies.

<table>
<thead>
<tr>
<th>Memory Technology</th>
<th>DRAM</th>
<th>PCM</th>
<th>STT-RAM</th>
<th>RRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_d [\text{J}]$</td>
<td>1.56E-8</td>
<td>3.66E-7</td>
<td>2.25E-7</td>
<td>2.75E-8</td>
</tr>
<tr>
<td>$P_s [\text{W/GB}]$</td>
<td>3.66E-1</td>
<td>3.66E-3</td>
<td>4.88E-2</td>
<td>7.31E-3</td>
</tr>
<tr>
<td>$L [\text{cy}]$</td>
<td>45.5</td>
<td>165</td>
<td>60.6</td>
<td>61.9</td>
</tr>
<tr>
<td>$\Delta E_d [\text{J/GB}]$</td>
<td>-</td>
<td>-5.34E-7</td>
<td>-2.09E-6</td>
<td>-5.23E-9</td>
</tr>
<tr>
<td>$\Delta P_s [\text{W/GB}]$</td>
<td>-</td>
<td>3.62E-1</td>
<td>3.17E-1</td>
<td>3.35E-1</td>
</tr>
<tr>
<td>$\Delta E_f [\text{J/GB}]$</td>
<td>-</td>
<td>1.60E-1</td>
<td>1.37E+1</td>
<td>1.62E+1</td>
</tr>
<tr>
<td>$\Delta L [\text{cy}]$</td>
<td>-</td>
<td>-1.19E+2</td>
<td>-1.52E+1</td>
<td>-1.65E+1</td>
</tr>
<tr>
<td>$\Delta EDP_s [\text{Ms}]$</td>
<td>-6.02E-5</td>
<td>-1.29E-5</td>
<td>-7.66E-7</td>
<td>-8.25E-9</td>
</tr>
<tr>
<td>$\Delta EDP_f [\text{Ms}]$</td>
<td>6.63E-2</td>
<td>5.76E+2</td>
<td>1.77E+3</td>
<td>3.58E-1</td>
</tr>
</tbody>
</table>

| Energy Comparison |  |
| $S^\infty [\text{GB}]$ | 270.0 | 65.4 | 31.6 |  |
| $S^\infty [\text{GB}]$ | 1340.2 | 65.4 | 31.6 |  |

| EDP Comparison |  |
| $E_{hyb}$ |  |

**Fig. 2.** Characterisation of workload space. The annotation $X > Y$ indicates a region of workloads where memory technology $X$ is more energy-efficient than technology $Y$.
We are interested in the comparison of a DRAM-only memory system versus a hybrid memory system. We define $\Delta E$ as the difference of energy consumed by a DRAM-only memory system of size $S_{\text{DRAM}} + S_{\text{NVM}}$ and average latency $L_{\text{DRAM}}$ and a hybrid memory system with sizes $S_{\text{DRAM}}$ and $S_{\text{NVM}}$ and corresponding latencies $L_{\text{DRAM}}$ and $L_{\text{NVM}}$. The model leads to similar observations as before, i.e., four distinct scenarios are possible, with values for the thresholds given as follows:

$$M^\infty = -\frac{CPI_{\text{dram}} \Delta P_s}{\Delta E_s + \frac{1}{\mu} \Delta P_s L_{\text{dram}}}$$

$$S_n_{\text{NVM}} = -\frac{\phi \Delta E_s + (P_{\text{dram}} S_{\text{dram}} + P_{\text{cpu}}) \Delta L}{\Delta E_s + \frac{1}{\mu} \Delta P_s L_{\text{dram}}}$$

A similar analysis was performed to compare the EDP of a hybrid memory system with a DRAM-only system.

## 5 RELATED WORK

Hybrid memories are a commonly studied approach to address main memory energy consumption. Researchers often assume small memory sizes, e.g., 8 GB or less [6], [14], [15]. As such, dynamic energy consumption is important in these studies. In this paper, we have shown the existence of a threshold memory size $S^\infty$ above which NVM outperforms DRAM in terms of energy and EDP. As static energy dominates for large memory sizes, many of the proposed techniques to reduce NVM energy consumption become mute at scale, or they should be interpreted as a way of lowering the $S^\infty$ threshold.

This work does not optimize the performance of NVM. Performance must be optimised independently, typically through caching. Similarly, a common goal is to modify the NVM memory layout [3], [5], [9]. Our hybrid memory model may be applied to these studies by adjusting the parameters $E_d$, $P_s$ and $L$ to the optimised values. Note that asymmetric read/write memory latencies can be expressed in the parameter $L$ by computing the appropriate average latency. The parameter $\mu$ can model caching. Conversely, the model may be applied to evaluate the benefits that optimisation of the model parameters may bring, select what parameters to optimise, or express constraints on the optimisation (e.g., $S^\infty < 64 \text{ GB}$).

## 6 CONCLUSION

We have presented an analytical model for comparing energy and energy-delay product of non-volatile and DRAM memory technologies. Workloads are characterised by memory size and access frequency. The model shows that byte-addressable NVM technologies are more energy-efficient than DRAM when the memory size exceeds a technology-dependent threshold. The model can be used to quickly evaluate the impact of changes to the memory architecture on energy, such as DRAM caching and write-avoidance schemes, by substituting improved values for the technology parameters. It can also indicate what memory system parameters to optimise.

The presented model applies to latency-sensitive workloads. In future work, bandwidth restrictions and read/write asymmetry will be addressed.

## ACKNOWLEDGMENTS

This work is supported by the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme (FP7/2007-2013), grant agreement no. 327744.

## REFERENCES