A Comparison of Energy Efficient Cache Design Architectures and Cache Scheme Techniques throughout the Past Decade

Krystal Folkes
Department of Electrical and Computer Engineering
University of Central Florida
Orlando, FL 32816-2362

Abstract— Cache designs with the capability of higher than standard performance for geo-technological and other elevated performance requirements are the focal point for data management demands at this time. Between volatile and non-volatile multilevel caches, there are data retrieval opportunities and methods for faster, smoother, and more proficient delivery. In short, a cache is a component designed to store data in a faster pace than the receiving of said data from the main memory. Caches vary by types, including but not limited to fully-associative, direct-mapped, and set-associative.

A few methods I have personally come across, orchestrated to increase cache productivity begin with alleviation to the service cost caused by critical loads from Last Level Cache. Removing SRAM-based L2 will give way to decreasing read service duration and improving data retention time. The alternative suggested by one abstract is the imputation Spin-Transfer Torque Random Access Memory. By causing miss ratio to decline exponentially and sustaining consistent cache requests, this proposal, in my professional opinion, is just shy of brilliant. In addition to that, similarly to the concept of Lithium batteries elongating life in Apple products, another abstract yielded the idea of a “hybrid architecture” designed to implement as specific write buffer with overall minimal dimensions for stamina.

Both Improving prefetching in relations with eDRAM LLC (as spoken of in at least two other abstracts) appear to have complimentary potential for strong kinetic action. To “propose a more accurate and application-relevant vulnerability estimation approach to conventional LLC SEU analysis methods” in conjunction with “efficacy of...systems, namely, request prioritization and hybrid local-global prefetch control, “ is a collaboration. What kind of collaboration? one that can precisely improve accuracy and accurately improve precision.

I. INTRODUCTION

Today, multilevel cache designs are now necessary for fulfilling the growing demand needed for performance and energy efficiency supply. Multilevel cache has two categories of memory types: volatile memory and non-volatile memory. Volatile memory is defined as data that is eventually lost when the memory is not consistently powered, and non–volatile memory is described as having the capabilities to retrieve data when power to the memory is cycled on and off. [11] In particular, a volatile multilevel cache designs include SRAM (Static Random-Access Memory) and eDRAM (Embedded Dynamic Random-Access Memory) is commonly used in comparison to the non-volatile STT-RAM (Spin Transfer Torque RAM) and PRAM (Phase Change RAM).

Cache is a hardware or software component that stores data, such that requests for data are provided faster than retrieval from the main memory. Cache has various types of cache block placements, direct-mapped, fully-associative, and set-associative, for storing data. Cache block placements are generally either takes advantage of principle of temporal or spatial locality. The cache block placements have different associativity. Associativity: design approach providing flexibility by which a block of memory can be associated with a corresponding line in cache. [11] Direct mapped cache has no associativity and each block is mapped to only one cache line. Fully associative cache has unbounded associativity meaning each block can be stored in any cache line. Set associative cache in bounded associativity referred to as k-way set associative (k lines are possible to store each block). [11] Figure 1 is to reference general cache design and bit encodings for the block placements.

Keywords— SRAM, eDRAM, STT-RAM, PRAM, energy consumption, associativity
Cache hit ratio (READ hit rate) is the portion of data found in the cache without referencing the memory. Subsequently, when there is a cache READ miss, the data is not found in cache and needs to be allocated from the memory to cache. The miss rate (READ miss rate) is the [number of address cache misses / number of address referenced]*100%. Similarly, hit ratio (READ hit rate) is the [number of address cache hits / number of address referenced]*100%. Thus, having other cache coherence policies of a WRITE hit and miss involving a DRITY copy of data [1].

The paper will further discuss ten multilevel cache design architectures and compare cache configurations over the past decade.

II. LITERATURE REVIEW
As technology advancement continues, SRAM is facing challenges such as leakage power and scalability. Recent research in STT-RAM has higher storage density and negligible leakage power [8]. In particular, a PRAM cache is advantageous as it has 16 times higher capacity than a SRAM cache with the same die area as well as consumes significantly less leakage energy [7].

From 2016 to 2010 research shows that replacing SRAM with STT-RAM, in the past deployment was in favor of large SRAM. Thus looking for alternative solutions eDRAM was creates with an on-chip memory hierarchy. Though the solution is not perfect. The eDRAM technology suffers from high dynamic energy consumption due to mandatory periodic refresh required to keep the stored value in the valid state. [4] The critical charge of eDRAM is dependent on storage and bit-line capacitance, write voltage, and the minimum voltage difference required by the sense amplifier thus can cause soft errors in life of the cell. [2]

Other studies showed the efficient STT-RAM write management algorithms read and write policies are used to reduce the number of STT-RAM read Write lines in a cache. [10]

III. DATA ANALYSIS

Fig. 2. <Energy vs Year >

IV. CONCLUSION

Discovering that two, what seem to be binary, categories of memory type, non-volatile and volatile illustrated the different scenarios that reality and the imagination can devise about powered caches. Problems can arise in consistently powered caches such as overload or elongated read durations and likewise caches with ‘breaks’ may have lower performance or tolerance rates for hard work. Some lines filter into both categories and there are strong as well as weak aspects of both, but the most important part as I noticed in [11] is that there is always opportunity for innovation.

REFERENCES


## TABLE I. <LEVEL CACHE CONFIGURATION COMPARISON>

<table>
<thead>
<tr>
<th>Parameters for the below techniques, Year</th>
<th>Processor</th>
<th>Level 1 (L1) for Instruction (I) or Data (D)</th>
<th>Level 2 (L2)</th>
<th>Level 3 (L3) or Last Level Cache (LLC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-----------------</td>
<td>-------</td>
<td>----------</td>
<td>------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Khoshavi [2]</td>
<td>8</td>
<td>3GHz</td>
<td>32KB</td>
<td>8-way</td>
</tr>
<tr>
<td>Sun [6]</td>
<td>4</td>
<td>2GHz</td>
<td>32KB</td>
<td>4-way</td>
</tr>
<tr>
<td>Chen [3]</td>
<td>4</td>
<td>3.3GHz</td>
<td>32KB</td>
<td>8-way</td>
</tr>
<tr>
<td>Khoshavi [4]</td>
<td>8</td>
<td>3GHz</td>
<td>32KB</td>
<td>8-way</td>
</tr>
<tr>
<td>Crawford [1]</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Lin [5]</td>
<td>2</td>
<td>800MHz</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Syu [10]</td>
<td>4</td>
<td>2GHz</td>
<td>32KB/64KB</td>
<td>4-way</td>
</tr>
<tr>
<td>Li [8]</td>
<td>1</td>
<td>N/A</td>
<td>32/8/2KB</td>
<td>4/1/3 way</td>
</tr>
<tr>
<td>Joo [7]</td>
<td>1</td>
<td>2GHz</td>
<td>32KB</td>
<td>Direct mapped</td>
</tr>
</tbody>
</table>

*CL* = Cache line
Calculation for “# of CL” columns:
Manually compute the number of cache lines given the capacity value as listed in capacity column, assuming the cache line size is always 64 Bytes

Protocol column = {Write Back (WB), Write Through (WT), MESI, MOESI, Not Available (N/A)}