**RESUMOS DOS PAPERS**

A Figura 1 exibe a capacidade, a velocidade dos acessos e o *endurance* de diversos modelos de memórias. Nesta figura nota-se modelos de memórias com acessos rápidos como SRAM, STT-RAM, p-STT-RAM e DRAM, com acessos moderados como PCM e ReRAM e com acessos lentos, como por exemplo, discos rígidos e memória flash. Observa-se que o modelo p-STT-RAM possui velocidade de acesso a memória superior ao modelo STT-RAM.

|  |
| --- |
|  |
| Figura 1 – Capacidade e velocidade de acesso de diversas memórias.  Fonte: FUJITA et al., 2015. |

O trabalho de Mittal, Vetter e Li (2015), apresenta uma avaliação comparativa das propriedades das diferentes tecnologias de memórias (Tabela 1). Os valores apresentados podem não ser satisfatórios, entretanto com a pesquisa contínua, os valores das propriedades podem ser melhorados significativamente.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tabela 1 – Características das diferentes tecnologias de memórias | | | | | | |
|  | SRAM | EDRAM | STT-RAM | RRAM | PCM | DWM |
| Tamanho da Célula (F2) | 120 - 200 | 60 – 100 | 6 - 50 | 4 - 10 | 4 - 12 | > 2 |
| Endurance | 1016 | 1016 | 4 x 1012 | 1011 | 108 - 109 | 1016 |
| Velocidade (L/E) | Muito Veloz | Veloz | Veloz/Lento | Veloz/Lento | Lento/Muito Lento | Veloz/Lento |
| Leakage | Alto | Médio | Baixo | Baixo | Baixo | Baixo |
| Energia Dinâmica (L/E) | Baixo | Médio | Baixo/Alto | Baixo/Alto | Médio/Alto | Baixo/Alto |
| Período de retenção | N/A | 30 - 100 µs | N/A (Quando Relaxado) | N/A | N/A | N/A |
| Fonte: MITTAL, VETTER e LI, 2015. | | | | | | |

O trabalho de Mittal e Vetter (2016) também discute propriedades aproximadas para diferentes tecnologias de memória (Tabela 2). Pode-se notar que estes valores devem ser considerados simbólicos, pois, futuras pesquisas podem levar a mudanças nestes parâmetros. Na Tabela 2, a granularidade de acesso refere-se à quantidade mínima de dados que são lidos / escritos em cada acesso e a *endurance* refere-se ao número de gravações que o bloco de memória pode suportar antes de se tornar não confiável. A propriedade comum a todos as NVMs é que sua latência / energia de escrita é significativamente maior do que a latência / energia de leitura. Além disso, em condições normais, as NVMs podem reter dados durante vários anos sem a necessidade de qualquer energia de *standby*.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tabela 2 – Propriedades aproximadas em nível do dispositivo das tecnologias de memórias | | | | | | | |
|  | Tamanho da Célula | Granularidade de Acesso | Latência de Leitura | Latência de Escrita | Latência de Apagar | Endurance | Energia de Espera |
| HDD | N/A | 512 B | 5 ms | 5 ms | N/A | > 1015 | 1W |
| SLC Flash | 4 – 6 F2 | 4 KB | 25 µs | 500 µs | 2 ms | 104- 105 | 0 |
| DRAM | 6 – 10 F2 | 64 B | 50 ns | 50 ns | N/A | > 1015 | Energia de atualização |
| PCM | 4 – 12 F2 | 64 B | 50 ns | 500 ns | N/A | 108- 109 | 0 |
| STT-RAM | 6 – 50 F2 | 64 B | 10 ns | 50 ns | N/A | > 1015 | 0 |
| ReRAM | 4 – 10 F2 | 64 B | 10 ns | 50 ns | N/A | 1011 | 0 |
| Fonte: MITTAL e VETTER, 2016. | | | | | | | |

**Otimizações realizadas em *Caches***

No trabalho de Mittal, Vetter e Li (2015), os autores apresentam diversas pesquisas sobre abordagens arquiteturais propostas para a concepção de sistemas de memória, especificamente *caches* utilizando memórias com tecnologias emergentes. Neste trabalho de revisão, os autores apresentam uma classificação dessas tecnologias e abordagens arquiteturais baseadas em suas características principais e conseguem realçar as semelhanças e diferenças dessas abordagens.

Em Komalan et al. (2015), os autores utilizaram uma memória não volátil para substituir as memórias SRAM em *caches* no nível L1. No artigo, os autores substituíram a cache de dados L1 SRAM por uma memória STT-MRAM e exploraram a penalização de leitura da cache L1 de dados com o emprego da NVM em um processador ARM. O impacto da substituição da memória SRAM por uma STT-MRAM foi avaliado quanto ao desempenho de *benchmarks*. Primeiramente, observou-se uma sobrecarga de 54% nas penalidades de desempenho dos *benchmarks* analisados, associadas ao emprego da memória STT-MRAM quando comparada com a SRAM. Estas penalidades se devem a latência de leitura de uma STT-MRAM ser bem superior se comparada com a latência de uma SRAM. Após foram realizadas algumas modificações na arquitetura da cache de dados empregada, que resultaram em uma redução destas penalidades, através da utilização de um VWB (*Very Wide Buffer*) entre a STT-MRAM e o processador. A Figura 2 ilustra a utilização do VWB.

|  |
| --- |
|  |
| Figura 2 – Modificação da organização D-cache L1 SRAM com uma D-cache L1 STT-MRAM, utilizando o VWB.  Fonte: KOMALAN et al., 2015. |

Após a realização da modificação arquitetural, Komalan et al. (2015), ainda propuseram e realizaram otimizações de software, ou seja, transformações e otimizações de código, exploraram o paralelismo através de vetorizações, alinhamentos de loop, e reduções do número de decisões de desvios. Todas estas otimizações foram automaticamente realizadas e conseguiram reduzir ainda mais as penalidades de leitura.

Para avaliar a eficácia das modificações propostas, os autores implementaram o projeto via o simulador GEM5 (BINKERT, et al., 2011) e utilizaram um conjunto de *benchmark* do PolyBench (POUCHET e YUKI, 2011). Através de experimentos, verificaram a influência do tamanho do VWB, na redução das penalidades. Os experimentos apontaram que ao aumentar o tamanho do VWB, reduz-se as penalidades. Entretanto, aspectos de custo e energia associados ao aumento do VWB também devem ser considerados, assim como o roteamento do *layout*, pois o *layout* do projeto torna-se mais complexo a medida em que o VWB é aumentado. Assim, os autores indicaram que o tamanho ideal para o VWB é cerca de 2KBits.

Ainda no trabalho de Komalan et al. (2015), são realizadas comparações entre a solução arquitetural, com outros trabalhos relacionados. A solução arquitetural oferece melhor desempenho (menor porcentagem de penalidades), devido as otimizações de software e modificações em hardware. Porém, quando as mesmas transformações são realizadas em uma SRAM, a solução com SRAM apresentou um desempenho melhor em relação a solução com a STT-MRAM, cerca de 8%. Entretanto, vale destacar que essa comparação não leva em conta as vantagens óbvias oferecidas pela cache com NVM (STT-MRAM), como uma área e gasto energético menores e não volatilidade. Assim, conclui-se que, pelo trabalho de Komalan et al. (2015), as explorações realizadas juntamente com as transformações e otimizações de código podem reduzir a penalidade de desempenho introduzido pela utilização da NVM que inicialmente era de 54% para níveis toleráveis, cerca de 8%, mesmo para os piores casos testados.

**Otimizações realizadas em *Scratchpads***

Uma otimização muito utilizada em sistemas embarcados é o uso de memórias *scratchpads*, pois apresentam vantagens quando comparadas com as memórias *caches,* como discutido anteriormente*.* Pesquisam vem sendo realizadas nas quais são avaliados o emprego de tecnologias emergentes em memórias *scratchpads*, como emWang et al. (2013) eHu et al. (2013).

Em Wang et al. (2013), os autores focaram em substituir uma memória *scratchpad* (SPM) empregando SRAM por uma *scratchpad* empregando STT-RAM. A memória SST-RAM foi escolhida dentre outras NVMs por apresentar vantagens em alguns aspectos, como por exemplo, melhores resultados para as operações de leitura e escrita (latência e consumo energético), para determinados tamanhos de memórias. Os resultados de ambas versões da SPM*,* com SST-RAM e com SRAM, foram comparados, os critérios observados na comparação foram o consumo energético e desempenho para operações de escrita e leitura juntamente com a área e o *leakage* de cada tipo de memória empregada. Posteriormente, os autores também avaliaram uma SPM híbrida, a qual emprega ambas as tecnologias, SRAM e STT-RAM, onde os dados mais escritos são alocados na SRAM e os dados mais lidos são alocados na STT-RAM.

Para a realização do trabalho, os autores utilizaram uma plataforma de simulação construída sobre a ferramenta SIMICS juntamente com o GEMS. Os *benchmarks* selecionados para o trabalho foram retirados do *MiBench* e a tecnologia empregada foi de 45nm (nanômetros). Para estimar as latências e o consumo de energia para operações de leitura / escrita para um determinado tamanho de memória não volátil foi utilizado o simulador NVsim (DONG, et al. 2012). Com a ferramenta NVsim, três tipos diferentes de otimizações, latência otimizada, produto da energia-*delay* (EDP) otimizado, e energia otimizado (WANG, et al., 2013) foram extraídos.

Como mencionado anteriormente, os autores exploraram duas abordagens diferentes para *scratchpad*, a primeira abordagem é uma substituição direta de uma memória SRAM por uma memória STT-RAM, a segunda abordagem é um *design* híbrido (SRAM e STT-RAM) para a memória *scratchpad*. Na abordagem híbrida, foram utilizadas diferentes áreas (proporções) de SRAM e STT-RAM.

Na primeira abordagem, a substituição direta de uma memória SRAM por uma memória STT-RAM, no quesito desempenho (IPC, Instruções Por Ciclo), em média a opção de otimização de energia resultou em uma degradação de cerca de 8%, enquanto que a latência otimizada e o produto energia-*delay* melhoraram cerca de 24% e 10%, respectivamente. Ainda na substituição direta, no quesito consumo energético, em média a latência otimizada sofreu uma degradação, cerca de 20%, enquanto que o produto energia-*delay* e a energia otimizada melhoraram cerca de 18% e 31%, respectivamente.

Como a otimização produto energia-*delay* (EDP) obteve os melhores resultados, tanto para desempenho como para consumo energético, na primeira abordagem (substituição direta), foi utilizada a otimização da EDP para as diversas proporções da SPM híbrida (segunda abordagem). Os autores fixaram a área de uma SPM de 64KB SRAM como linha base para realizarem a comparação de diversas proporções do modelo híbrido. Por exemplo, os autores utilizaram a proporção de 1:1 de área para a SPM híbrida, o que significa que a SPM híbrida possui 32KB de SRAM e 128KB de STT-RAM, enquanto o SPM linha de base tem 64 KB de SRAM. Segundo Wang et al. (2013) a memória STT-RAM pode ser projetada cerca de quatro vezes mais densa que a memória SRAM e estas duas SPMs de densidades diferentes possuíram uma área total semelhante de silício, devido a variação na área requerida por cada tecnologia.

Na segunda abordagem, *design* híbrido (SRAM e STT-RAM), na comparação de desempenho para as diversas proporções para a SPM híbrida, a proporção 2:1 (SRAM:STT-RAM) possui os melhores resultados, cerca de 36% em média. Na comparação do consumo energético para as diversas proporções para a SPM híbrida, o modelo ALLSTT-RAM (empregando somente memória STT-RAM), apresentou os melhores resultados, cerca de 29% em média. Na comparação do produto do desempenho e consumo energético para as diversas proporções da SPM híbrida, a proporção 2:1 (SRAM:STT-RAM) apresentou os melhores resultados, cerca de 45% em média.

Pelo trabalho de Wang et al., (2013), percebe-se que, através das explorações realizadas juntamente com as otimizações demonstradas, que a arquitetura SPM híbrida pode superar SPM’s constituídas somente por SRAM ou STT-RAM. Devido as características atraentes da STT-RAM, como o baixo *leakage* e de alta densidade, esta é apontada como uma memória promissora para modelos híbridos de memória ou para a substituição da SRAM.

Em Hu et al. (2013), os autores utilizaram como memória emergente uma PCM. O trabalho teve como o estudo de caso uma SPM híbrida composta de 16KB SRAM e 64KB PCM, a qual foi comparada com uma SPM de 32KB SRAM. Ainda nesse trabalho, os autores exploraram diferentes algoritmos para a alocação dos endereços nos tipos de memórias utilizadas, propondo um novo algoritmo para alocação em memórias híbridas. Os benchmarks empregados pelos autores para a realização do trabalho foram selecionados do *MiBench*. De acordo com os resultados experimentais, com a ajuda do algoritmo proposto pelos autores, o modelo híbrido de arquitetura SPM reduziu o tempo de acesso à memória em 18.17%, a energia dinâmica em 24.29%, e o *leakage* em 37.34% quando comparada com uma SPM contendo SRAM com a mesma área.

Ambos os trabalhos citados, Hu et al. (2013) e Wang et al., (2013), utilizaram o NVsim (DONG, et al. 2012) com tecnologia 45nm, para a obtenção dos parâmetros das memórias PCM, STT-RAM e SRAM, como as latências de acesso e o consumo energético para as operações de escrita e leitura. Os dois trabalhos conseguiram reduzir significativamente o consumo energético e melhorar o desempenho para diversos benchmarks executados em um processador ARM, através do emprego de tecnologias emergentes de memória em SPMs.

REFERÊNCIAS

FUJITA, S. et al. Technology trends and near-future applications of embedded STT-MRAM. **IEEE International Memory Workshop (IMW)**. 2015.

HU, J. et al. Data Allocation Optimization for Hybrid Scratch Pad Memory with SRAM and Nonvolatile Memory. **IEEE Transactions on Very Large Scale Integration Systems** (VLSI), Volume: 21, Issue: 6, p. 1094 – 1102, june, 2013.

KOMALAN, M. P. et al. System level exploration of a STT-MRAM based Level 1 Data-Cache. **Design, Automation & Test in Europe Conference & Exhibition (DATE)**. p. 1311-1316, march, 2015.

MITTAL, S.; VETTER, J. S. A Survey of Software Techniques for Using Non-Volatile Memories for Storage and Main Memory Systems. **IEEE Transactions on Parallel and Distributed Systems**, Volume: 27, Issue: 5, may 1, 2016.

MITTAL, S.; VETTER, J. S.; LI, D. A Survey of Architectural Approaches for Managing Embedded DRAM and Non-Volatile On-Chip Caches. **IEEE Transactions on Parallel and Distributed Systems**, Volume: 26, Issue: 6, june 1, 2015.

WANG, P. et al. Designing Scratchpad Memory Architecture with Emerging STT-RAM Memory Technologies. **Proceedings of IEEE International Symposium on Circuits and Systems** (ISCAS), Beijing, China, may, p. 1244 – 1247, 2013.