Survey of Non-Volatile Memories: a taxonomy

Lizandro Oliveira[1], Lisandro Silva[1] and Mauricio Pilla[1]

***Abstract*— O consumo de energia é tão importante quanto o desempenho em sistemas embarcados alimentados por bateria, pois cada vez mais estes sistemas precisam processar computação intensiva com um baixo consumo energético. Devido à alta contribuição do acesso à memória no consumo total de energia de sistemas embarcados, a arquitetura de memória influencia fortemente os objetivos dos projetos dos dispositivos embarcados. Novas técnicas são propostas devido aos problemas enfrentados com o avanço da tecnologia, como por exemplo, a memória tradicional baseada em SRAM (*Static Random Access Memory*) *on-chip* tornou-se um gargalo em consumo energético para o projeto de sistemas embarcados, devido principalmente ao seu alto *leakage*. As tecnologias emergentes de memórias não voláteis (*NVM, Non-Volatile Memories*) são soluções candidatas para os futuros sistemas de memória, pois elas possuem algumas vantagens sobre as memórias SRAMs (*Static Random-Access Memory*) e DRAMs (*Dynamic Random-Access Memory*) tradicionais, como por exemplo, um menor *leakage*, uma maior densidade e não volatilidade.**

I. INTRODUCÃO

Segundo Hennessy [1], em um sistema de computador moderno o gargalo dominante na obtenção de alto desempenho e eficiência energética é a distância tecnológica entre o desempenho do processador e a memória tradicional. Esta distância torna-se mais significativa em sistemas embarcados, pois o sistema de memória é um dos principais fatores de desempenho e consumo energético, especialmente nos sistemas embarcados que utilizam bateria [2]. O projeto de sistemas embarcados apresenta muitas restrições e requisitos rígidos. De um modo geral a descrição dos requisitos funcionais não é suficiente para o projeto de um sistema embarcado, devendo ser considerados também requisitos não-funcionais, tais como desempenho, custo, consumo de energia, tamanho físico e peso [3]. Estes requisitos não-funcionais extras ocasionam limitações nas decisões do projeto, provocando preocupações no desempenho e no consumo energético.

Os requisitos necessários para os sistemas embarcados têm motivado a investigação de técnicas de otimização. Na literatura são propostos diversos tipos de otimizações em memória, desde técnicas de otimização em *software*, em *hardware* ou técnicas mistas as quais utilizam tanto otimizações de *hardware* quanto de *software* [2]. Muitas dessas técnicas utilizam NVMs como uma forma de otimização, visto que essas memórias emergentes possuem algumas vantagens em relação as memórias tradicionais DRAM e SRAM. Embora o mercado de memórias emergentes ainda é menor que das memórias tradicionais existe a previsão que este mercado crescerá até 2021, atingindo taxas em torno de 110% ao ano, quando estas novas tecnologias serão utilizadas em vários produtos [4]. Alguns exemplos de memórias emergentes são a PCM/PCRAM (*Phase Change Memory Random-Access Memory*), a MRAM (*Magnetoresistive Random-Access Memory*), a STT-MRAM/STT-RAM (*Spin-Transfer Torque Magnetic Random-Access Memory*), a RRAM/ReRAM (*Resistive Random-Access Memory*), a FRAM/FeRAM (*Ferro-magnetic Random-Access Memory*) e a DWM (*Domain Wall Memory*)

Este artigo está organizado da seguinte maneira. Na Seção II alguns fundamentos sobre memórias NVM's serão apresentados. Na Seção III serão revisados e discutidos trabalhos empregando NVM's em diferentes níveis da hierarquia de memória, como caches, SPM's (*Scratchpad Memory*) e memórias principais, enquanto que as conclusões são discutidas na Seção IV.

II. VISÃO GERAL - MEMÓRIAS

O desempenho total do sistema de memória e o consumo de energia são severamente relacionados com o tempo de acesso à memória e o consumo de energia média, o que faz com que a arquitetura de uma memória seja uma grande preocupação em projetos de sistemas embarcado. Nesta Seção são apresentadas as características das NVM's.

A. STT-RAM

Segundo Smullen[5], a STT-RAM utiliza uma junção de túnel magnético (MTJ: *Magnetic Tunnel Junction*) como armazenamento de memória. Uma célula STT-RAM é formada por um transistor de acesso que é ligado a um elemento de memória implementado usando um MTJ, que contém duas camadas ferromagnéticas separadas por uma camada isoladora de óxido. A direção da magnetização de uma camada ferromagnética é fixa enquanto que a outra camada ferromagnética pode ser alterada pela passagem de uma corrente. A resistência do MTJ é determinada pela direção de magnetização relativa dessas duas camadas. Se as duas camadas têm direções diferentes, a resistência do MTJ é alta e vice-versa. Usando essa propriedade, um valor binário é armazenado em uma célula STT-RAM. Para ler o valor armazenado, uma pequena tensão é aplicada entre os terminais MTJ. A corrente que flui através do dispositivo é detectada, e o estado de magnetização é determinado como um resultado.

Embora a STT-RAM possui densidade menor do que a PCM e a RRAM, e maior latência e energia para a operação

*This work was not supported by any organization

[1]H. Kwakernaak is with Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands h.kwakernaak at papercept.net

[2]P. Misra is with the Department of Electrical Engineering, Wright State University, Dayton, OH 45435, USA p.misra at ieee.org

de escrita do que a SRAM, essa memória foi amplamente utilizada para a concepção de caches devido a seu alto *endurance*. O *endurance* é o número de ciclos de gravações que podem ser aplicados a um bloco de memória flash, antes que a mídia de armazenamento torne-se inconfiável. No entanto, apesar de um valor de endurance de 10^{15} foi estimado, o melhor resultado do teste de endurance até agora é inferior a $4x10^{12}$ [6]. Outra vantagem da STT-RAM é que a sua não-volatilidade pode ser negociada para melhorar sua energia de gravação e latência.

Em Smullen [5], os autores modificam o tempo de retenção encolhendo a área planar do MTJ, enquanto que o trabalho de [7] consegue isto diminuindo a espessura da camada livre e baixando a saturação da magnetização que reduz a barreira térmica do MTJ. Como exemplo, [7] mostra que para a frequência de 2 GHz, os valores de latência de gravação de uma STT-RAM de 4 MB para períodos de retenção de 10 anos, 1 segundo e 10 milissegundos são respectivamente 22, 12 e 6 ciclos. Assim, com base na característica da aplicação e no nível da hierarquia da cache, um designer pode escolher um valor apropriado de período de retenção.

B. FeRAM

Resumo das Tecnologias de Memórias.....

C. RRAM

Uma RRAM com comutação unipolar usa um dielétrico isolador [8]. Quando é aplicada uma tensão suficientemente elevada, é formado um filamento ou um percurso condutor no dielétrico isolador. Depois disso, aplicando tensões adequadas, o filamento pode ser ajustado para *set* (o que leva a uma baixa resistência) ou *reset* (o que leva a uma alta resistência).

Comparado a SRAM, uma cache RRAM tem alta densidade, latência de leitura comparável e possui um valor muito inferior de gasto energético referente ao *leakage*. No entanto, as limitações da RRAM são a seu baixo endurance, cerca de 10^{12} [9], alta latência e consumo energético para as operações de escrita. Por exemplo, uma cache típica de 4 MB de RRAM tem uma latência de leitura entre 6 e 8 nanosegundos e uma latência de escrita entre 20 e 30 nanosegundos [10].

D. PCM

A PCM usa um material de mudança de fase chamado GST, que é uma liga de germânio, antimônio e telúrio. Quando a liga é aquecida a uma temperatura muito alta e rapidamente resfriada, transita em uma substância amorfa com alta resistência elétrica (*reset*). Por outro lado, quando a liga é aquecida a uma temperatura entre a cristalização e o ponto de fusão e resfriada lentamente, cristaliza até um estado físico com menor resistência (*set*). Para a operação *set*, quando o GST é aquecido a uma temperatura entre a temperatura de cristalização ($\sim 300\,^{\circ}$C) e a temperatura de fusão ($\sim 600\,^{\circ}$C) durante um período de tempo, GST transforma-se no estado cristalino que corresponde a um Lógica '1'. Para a operação *reset*, quando o GST é aquecido acima do ponto de fusão e extinguido rapidamente, GST transforma-se no estado amorfo que corresponde ao '0' lógico. Esta propriedade física é usada para armazenar um valor binário em uma célula PCM. A estrutura básica de uma célula da PCM consiste em um transistor NMOS e um dispositivo de mudança de fase.

Os dois desafios mais graves no uso de PCM para projetar caches *on-chip* são sua resistência limitada de escrita e alta latência de gravação. Uma vez que o tráfego de escrita para uma cache é muito mais pesado do que para uma memória principal e a *endurance* da PCM é apenas perto de 10^8 escritas [11] e [12], para diversas aplicações, uma cache que utiliza PCM pode falhar em menos de uma hora. Uma típica cache de 4 MB utilizando PCM possui uma latência de leitura entre 15 e 20 nanosegundos e uma latência de escrita entre 150 e 170 nanosegundos [10] e [12].

E. DWM

A DWM funciona controlando a parede do domínio (DW: *Domain Wall*) em nanofios ferromagnéticos [13]. O fio ferromagnético pode ter múltiplos domínios que são separados por paredes de domínio. Esses domínios podem ser individualmente programados para armazenar um único bit (na forma de uma direção de magnetização) e assim, o DWM pode armazenar múltiplos bits por célula de memória.

Logicamente, uma macrocélula DWM aparece como uma fita, que armazena múltiplos bits e pode ser deslocada em qualquer direção. O desafio na utilização de DWM é que o tempo consumido no acesso a um bit depende da sua localização em relação à porta de acesso, o que leva a uma latência de acesso não uniforme e torna o desempenho dependente do número de operações de deslocamento necessárias por acesso. Comparado com outras NVMs, DWM é mais recente, e ainda está em fase de pesquisa e protótipos.

F. Comparativo entre as Memórias

Em um sistema embarcado genérico, o sistema de memória principal pode estar contido parcialmente dentro do chip *on-chip* e fora do chip *off-chip* [2]. As memórias *on-chip* e *off-chip* influenciam o desempenho e o consumo de energia em sistemas embarcados. No modelo on-chip as SPM's podem ser utilizadas juntamente com memórias caches, como memórias de alta velocidade. O tipo tradicional de memória empregada nas SPMs e caches são as memórias SRAMs.

Com os avanços da tecnologia CMOS, com a fabricação de transistores cada vez menores, o *leakage* da SRAM cresceu consideravelmente, resultando em uma parte significativa do consumo energético total em diversos chips de semicondutores [14]. Comparando com a memória tradicional SRAM, as memórias emergentes STT-RAM e PCRAM proporcionam um leakage menor e uma maior densidade. Comparando-as entre si, STT-RAM possui uma menor latência de acesso e energia dinâmica, enquanto PCRAM possui maior densidade.

Segundo Banakar [15] uma cache utilizando memória SRAM normalmente possui um consumo energético entre

25% a 45% do consumo total do chip. Os autores utilizaram uma SPM com o objetivo de reduzir o consumo de energia e de área, os resultados apresentaram uma redução de 40% no consumo energético e uma redução em média de 46% na área. Assim, a memória SPM, tem sido amplamente adotada em muitos sistemas embarcados devido a sua menor área e menor consumo energético. Além disso, a scratchpad pode proporcionar muitas vezes uma melhor previsibilidade e sincronização em dispositivos de tempo real, por ser uma memória gerenciada por software /citeli2005memory.

Ao contrário de memórias baseadas em carga, como DRAM, NVMs armazenam seus dados através da mudança do estado físico. Uma vez que uma operação de escrita para NVM envolve a mudança de seu estado físico, uma operação de gravação para NVM possui maior latência e consumo energético do que uma operação de leitura, levando a assimetria leitura-escrita [16]. Da mesma forma, a latência de escrita e o consumo energético da transição da lógica 1 para 0 é maior do que a de 0 para 1, levando a assimetria de escrita de 0/1 [17]. As NVMs também permitem o armazenamento de múltiplos bits de dados em uma única célula de memória. Isto é referido como armazenamento de células de múltiplos níveis (MLC: *Multi-Level Cell*) e conduz a um aumento significativo na densidade de armazenamento suportado por estas memórias.

Segundo Wu [18] e Sun [19] a memória STT-RAM é mais adequada para memórias de último nível, enquanto que a PCRAM é promissora como uma alternativa à DRAM na memória principal [20]. Embora a STT-RAM possui densidade menor do que a PCM e a RRAM, e maior latência e energia para a operação de escrita do que a SRAM, essa memória foi amplamente utilizada para a concepção de caches devido a seu alto endurance. Enquanto que a PCM é adequada para a memória principal ou hierarquias inferiores de cache, por exemplo, cache L3 ou mesmo cache L4 [18], onde a sua latência elevada pode ser tolerada e a alta densidade pode ser utilizada para evitar acessos fora do chip.

Embora STT-RAM apresente muitas características atraentes, como baixo leakage e alta densidade, esse tipo de memória possui alguns problemas. Ao contrário da SRAM, na qual operações de leitura e escrita consomem o mesmo tempo e energia, uma operação de gravação na STT-RAM requer muito mais tempo de latência e maior energia do que uma operação de leitura. Além disso, a latência e a energia de operações de escrita convencionais em uma STT-RAM são várias vezes maiores do que os da SRAM para um mesmo tamanho de memória.

Novos modelos de STT-RAM foram propostos para diminuir os problemas envolvidos com as operações de escrita. Segundo Amiri [21] e Tadisina [22], as PMTJ (*Perpendicular Magnetic Tunnel Junction*) foram desenvolvidas para alcançar uma baixa corrente de comutação, mantendo uma alta estabilidade térmica para as STT-RAM. Segundo Fujita [14] o modelo de memória p-STT-RAM (*Perpendicular Spin-Transfer Torque Magnetic Random-Access Memory*) possui uma maior probabilidade para substituir a SRAM do que a STT-MRAM, pois a p-STT-RAM apresenta uma maior velocidade de acesso e endurance. Já em Xu [23], os autores conseguiram diminuir significativamente os problemas de gravação em STT-RAM SPM, graças a um cuidadoso processo de cootimização entre os dispositivos da arquitetura.

A propriedade comum a todas 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 [24].

A tabela ...

TABLE I
CLASSIFICAÇÃO DE MEMÓRIAS NÃO-VOLÁTEIS

Classificação	**Referências**
Tecnologia de memória	Listar papers
Memórias híbridas	Listar
Cache	Listar
SPM	Listar
Memória principal	Listar
Economia de energia	Listar
Melhoria de desempenho	Listar

III. TRABALHOS RELACIONADOS COM NVM'S

O emprego de NVM's tem sido investigado em diferentes níveis da hierarquia de mémoria por diversos trabalhos. Nos trabalhos de Mittal [25] e [24], os autores revisam diversas propostas de otimizações em caches e memórias principais, entretanto não abordam nenhuma utilização em SPMs. As SPM's empregando NVM's são investigadas em [26] e [27], comparando-as com tecnologias tradicionais de memória.

A. SPM's

Em Wang [26], os autores investigaram primeiramente a substituição de uma memória SPM empregando SRAM por uma SPM baseada em STT-RAM. Posteriormente, os autores também avaliaram uma SPM híbrida (SRAM+STT-RAM), onde os dados mais escritos são alocados na SRAM e os dados mais lidos são alocados na STT-RAM. Na abordagem híbrida, foram utilizadas diferentes áreas (proporções) de SRAM e STT-RAM. Para a realização do trabalho, utilizaram uma plataforma de simulação construída sobre a ferramenta SIMICS [28] juntamente com o GEMS. Para a abordagem híbrida os autores fixaram a área de uma SPM de 64KB SRAM como linha base para realizarem a comparação de diversas proporções. 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 [26] a memória STT-RAM pode ser projetada cerca de quatro vezes mais densa que a memória SRAM, estas duas SPMs de densidades diferentes possuem uma área total semelhante de silício, devido a variação na área requerida por cada tecnologia. 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 [26], 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 a alta densidade, esta é apontada como uma memória promissora para modelos híbridos de memória ou para a substituição da SRAM.

No trabalho de Hu [27], uma SPM tradicional baseada em SRAM também foi comparada com uma SPM híbrida, porém empregando como memória emergente uma PCM. 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. 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.

Os trabalhos citados, [27] e [26], utilizaram o NVsim [10] 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 benchmarks selecionados foram retirados do MiBench [29]. Os 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.

B. Caches

Trabalhos caches

C. Memória Principal

Trabalhos MP.

IV. CONCLUSÕES

A conclusion secti

REFERENCES

[1] John L. Hennessy and David A. Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2011.

[2] W. Wolf and M. Kandemir. Memory system optimization of embedded software. *Proceedings of the IEEE*, 91(1):165–182, 2003.

[3] Marilyn Wolf. *Computers as components: principles of embedded computing system design*. Elsevier, 2012.

[4] Yole Développement. Storage-class memory will be the clear go-to market for emerging non-volatile memory in 2021. [Online]. Available: http://www.yole.fr/Emerging_NVM_Market.aspx#.WYEepojyvDf, 28 July 2016.

[5] C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan. Relaxing non-volatility for fast and energy-efficient stt-ram caches. In *High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium on*, pages 50–61. IEEE, 2011.

[6] Y. Huai. Spin-transfer torque mram (stt-mram): Challenges and prospects. *AAPPS bulletin*, 18(6):33–40, 2008.

[7] A. Jog, A. K. Mishra, C. Xu, Y. Xie, V. Narayanan, R. Iyer, and C. R. Das. Cache revive: architecting volatile stt-ram caches for enhanced performance in cmps. In *Proceedings of the 49th Annual Design Automation Conference*, pages 243–252. ACM, 2012.

[8] H. Li and Y. Chen. An overview of non-volatile memory technology and the implication for tools and architectures. In *Design, Automation & Test in Europe Conference & Exhibition, 2009. DATE'09.*, pages 731–736. IEEE, 2009.

[9] Y.-B. Kim, S. R. Lee, D. Lee, C. B. Lee, M. Chang, J. H. Hur, M.-J. Lee, G.-S. Park, C. J. Kim, U.-I. Chung, et al. Bi-layered rram with unlimited endurance and extremely uniform switching. In *VLSI Technology (VLSIT), 2011 Symposium on*, pages 52–53. IEEE, 2011.

[10] X. Dong, C. Xu, N. Jouppi, and Y. Xie. Nvsim: A circuit-level performance, energy, and area model for emerging non-volatile memory. In *Emerging Memory Technologies*, pages 15–50. Springer, 2014.

[11] J. Wang, X. Dong, Y. Xie, and N. P. Jouppi. i2wap: Improving non-volatile cache lifetime by reducing inter-and intra-set write variations. In *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*, pages 234–245. IEEE, 2013.

[12] Y. Joo, D. Niu, X. Dong, G. Sun, N. Chang, and Y. Xie. Energy-and endurance-aware design of phase change memory caches. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2010*, pages 136–141. IEEE, 2010.

[13] R. Venkatesan, V. Kozhikkottu, C. Augustine, A. Raychowdhury, K. Roy, and A. Raghunathan. Tapecache: a high density, energy efficient cache based on domain wall memory. In *Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design*, pages 185–190. ACM, 2012.

[14] S. Fujita, H. Noguchi, K. Ikegami, S. Takeda, K. Nomura, and K. Abe. Technology trends and near-future applications of embedded stt-mram. In *Memory Workshop (IMW), 2015 IEEE International*, pages 1–5. IEEE, 2015.

[15] R. Banakar, S. Steinke, B.-S. Lee, M. Balakrishnan, and P. Marwedel. Scratchpad memory: design alternative for cache on-chip memory in embedded systems. In *Proceedings of the tenth international symposium on Hardware/software codesign*, pages 73–78. ACM, 2002.

[16] R. Bishnoi, F. Oboril, M. Ebrahimi, and M. B. Tahoori. Avoiding unnecessary write operations in stt-mram for low power implementation. In *Quality Electronic Design (ISQED), 2014 15th International Symposium on*, pages 548–553. IEEE, 2014.

[17] R. Bishnoi, M. Ebrahimi, F. Oboril, and M. B. Tahoori. Asynchronous asymmetrical write termination (aawt) for a low power stt-mram. In *Proceedings of the conference on Design, Automation & Test in Europe*, page 180. European Design and Automation Association, 2014.

[18] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie. Hybrid cache architecture with disparate memory technologies. In *ACM SIGARCH computer architecture news*, volume 37, pages 34–45. ACM, 2009.

[19] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen. A novel architecture of the 3d stacked mram l2 cache for cmps. In *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on*, pages 239–249. IEEE, 2009.

[20] B. C Lee, E. Ipek, O. Mutlu, and D. Burger. Architecting phase change memory as a scalable dram alternative. In *ACM SIGARCH Computer Architecture News*, volume 37, pages 2–13. ACM, 2009.

[21] P. K. Amiri, Z. M. Zeng, J. Langer, H. Zhao, G. Rowlands, Y.-J. Chen, I. N. Krivorotov, J.-P. Wang, H. W. Jiang, J. A. Katine, et al. Switching current reduction using perpendicular anisotropy in cofeb–mgo magnetic tunnel junctions. *Applied Physics Letters*, 98 (11):112507, 2011.

[22] Z. R. Tadisina, A. Natarajarathinam, B. D. Clark, A. L. Highsmith, T. Mewes, S. Gupta, E. Chen, and S. Wang. Perpendicular magnetic tunnel junctions using co-based multilayers. *Journal of Applied Physics*, 107(9):09C703, 2010.

[23] C. Xu, D. Niu, X. Zhu, S. H Kang, M. Nowak, and Y. Xie. Device-architecture co-optimization of stt-ram based memory for low power embedded systems. In *Proceedings of the International Conference on Computer-Aided Design*, pages 463–470. IEEE Press, 2011.

[24] S. Mittal and J. S. Vetter. A survey of software techniques for using non-volatile memories for storage and main memory systems. *IEEE Transactions on Parallel and Distributed Systems*, 27(5):1537–1550, 2016.

[25] S. Mittal, J. S. Vetter, and D. Li. A survey of architectural approaches for managing embedded dram and non-volatile on-chip caches. *IEEE Transactions on Parallel and Distributed Systems*, 26(6):1524–1537, 2015.

[26] P. Wang, G. Sun, T. Wang, Y. Xie, and J. Cong. Designing scratchpad memory architecture with emerging stt-ram memory technologies. In

Circuits and Systems (ISCAS), 2013 IEEE International Symposium on, pages 1244–1247. IEEE, 2013.

[27] J. Hu, C. J. Xue, Q. Zhuge, W.-C. Tseng, and E. H.-M. Sha. Data allocation optimization for hybrid scratch pad memory with sram and nonvolatile memory. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 21(6):1094–1102, 2013.

[28] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hallberg, J. Hogberg, F. Larsson, A. Moestedt, and B. Werner. Simics: A full system simulation platform. *Computer*, 35(2):50–58, 2002.

[29] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown. Mibench: A free, commercially representative embedded benchmark suite. In *Workload Characterization, 2001. WWC-4. 2001 IEEE International Workshop on*, pages 3–14. IEEE, 2001.