Storage Class Memory: Avanços e Limitações em Direção às Memórias Universais

Pedro Ferro Laks[1], **Emilio Francesquini**[1]

[1]Centro de Matemática, Computação e Cognição
Universidade Federal do ABC (UFABC)
Santo André – SP – Brasil

pedro.laks@aluno.ufabc.edu.br, e.francesquini@ufabc.edu.br

Resumo. *Recentemente vem chamando a atenção o avanço nas pesquisas de novas tecnologias de memória chamadas de memórias emergentes, memórias persistentes ou Storage Class Memory, (SCM). Nesse texto discutimos as limitações e dificuldades para a criação de uma memória universal que substituiria toda a hierarquia de memórias por uma única memória. As SCM se aproximam das características necessárias para a criação de uma memória universal, mas ainda não as satisfazem completamente. Assim, as SCM podem pragmaticamente serem vistas como uma adição à hierarquia de memória atual, substituindo ou trabalhando em conjunto com as tecnologias SSD, DRAM e SRAM. Neste texto, exploramos o estado atual das SCM e discutimos sua evolução, com o objetivo final de reduzir a latência de memória para a CPU.*

1. Introdução

Os sistemas de Computação de Alto Desempenho (HPCs) enfrentam, atualmente, um significativo gargalo relacionado ao armazenamento e à necessidade contínua de transferência de dados entre a *Dynamic Random Access Memory* (DRAM) e o armazenamento secundário. Esse processo é caracterizado pela sua lentidão, resultando na subutilização da *Central Processing Unit* (CPU) pela falta de recebimento de dados. Nos últimos anos, houve uma crescente adoção dos *Solid State Drives* (SSD), porém, esses dispositivos não resolvem os desafios associados ao armazenamento interno, como a limitada velocidade e durabilidade quando comparados à memória DRAM. Embora os SSDs superem os *Hard Disk Drives* (HDD) em termos de latência de leitura e escrita, comumente atingindo até $200\times$ na leitura e $10\times$ na escrita, eles ainda permanecem mais lentos que as memórias DRAM, com uma diferença aproximada de 5×10^2 na leitura e 10^4 na escrita [6]. Adicionalmente, os SSDs apresentam uma baixa durabilidade, geralmente suportando cerca de 10^5 ciclos de escrita, um valor na ordem de $\times 10^{10}$ menor que os HDDs e as DRAM, que possuem durabilidade de 10^{15} [6]. Ao longo da evolução natural da tecnologia, foram sendo pesquisadas memórias mais velozes em comparação às tecnologias previamente estabelecidas. O que se notou é que a maioria dessas tecnologias de RAM eram não-voláteis assim como os SSDs, diferentemente das DRAM que são voláteis. Então, por conta desse atributo, ressurgiu a ideia de uma memória universal que unificam a memória de trabalho com a memória secundária, ou até mesmo todas as memórias do sistema [8, 11] que teria como grande vantagem, não precisar fazer a lenta troca de dados entre os armazenamentos. Por conta das limitações da DRAM e do SSD e HDD, essas tecnologias começaram a chamar ainda mais atenção ultimamente, prometendo uma grande evolução. Vale lembrar que a ideia de memória universal já existe desde o primórdios da computação, através da memória de núcleo magnético [7]. O uso de memórias universais seria particularmente

útil para sistemas de HPC, que necessitam de dezenas de HDDs de 16Tb, pois teríamos o grande armazenamento de um SSD ou HDD mas com a velocidade da memória RAM. Outro fator é que a não volatilidade de memórias universais consumiria menos energia e seria mais seguro quanto a perda de dados.

Existem outras memórias muito próximas de universais chamadas de "memórias não voláteis" também chamadas de "memórias emergentes","SCM (Storage Class Memories)", ou "Memórias persistentes", esses são termos sem consenso na literatura e por isso serão explicados na Seção 2 um pouco melhor essas definições. E por fim, existem as memórias híbridas, que buscam de certo modo, simular uma memória universal, utilizando o melhor das memórias antigas e memórias emergentes. O Objetivo dessa revisão é definir as SCM e as memórias universais, demonstrar a diferença entre elas, além de explicar os problemas da ideia de memória universal e porque ela deve ser abandonada. O outro objetivo apresentar alternativas de memórias e novas arquiteturas de memória que visam reduzir o gargalo de latência entre CPU e memória.

2. Fundamentação Teórica e Estado Atual das Memórias

Muitos artigos como Rudan et al [11] comentam sobre memórias universais e SCM, porém não possuem enfoque nelas. Outros artigos como Baldi et al [9], até focam em memórias universais, porém diferente desse, acabam possuindo um certo viés, por terem como base artigos de memórias específicas. Desse modo, esse artigo priorizou uma revisão sistemática sobre o tema de memórias universais e SCM e focando em artigos que citavam mais de uma memória, mais gerais. Nesse processo, foram utilizadas as palavras chaves "emerging memories AND universal memories AND memory AND dram AND flash AND ns AND endurance", nas seguintes bases de dados: Acm, SpringerLink, ScienceDirect, IEEE e Scopus. Com isso foram coletados 45 artigos, dos quais 19 mencionam memórias universais, e uma parcela deles será mencionada no texto[1].

As memórias universais são uma tecnologia que prometia muito nos anos 2000, juntamente com as storage class memory, elas deveriam ser a tecnologia do futuro. No entanto, isso não aconteceu exatamente, e devido a certas limitações, chegamos a apenas uma fração dos objetivos previstos por essas tecnologias [11].

2.1. Definições de Universal Memory e Storage Class Memories

Tanto o texto do Thomasian e Alexander [14] quanto o texto do Dimitrakis, Panagiotis [1] consideram que os artigos do R.Freitas e W.Wilcke [12], e o artigo do G.Burr et al[4] começaram a definir o que seriam as *Storage Class* Memory. No texto [4] a *Storage Class Memory* é considerada muito mais como uma evolução dos HDDs e SSDs do que propriamente uma memória universal capaz de substituir também a DRAM e *Static Random Acess Memory* (SRAM). De um modo geral, o artigo propõe uma memória com baixo custo por bit, alta densidade, alta performance, boa retenção de dados e alta durabilidade de escrita. Sendo como objetivo dessa memória, ir além da memória flash e atender a demanda de futuros servidores. A ideia das SCM no geral é ficar com uma latência entre os HDDs/SSDs e a DRAM, ser uma camada adicional na hierarquia de memória.

O texto [12] apesar de também citar as SCM para substituir os HDDs/SSDs ele já é um pouco mais otimista e considera que é bem provável que as SCM fiquem num meio do caminho entre essas tecnologias e a DRAM, podendo dividir funções de memória

[1] A metodologia completa pode ser acessada na seguinte referência: Metodologia da Revisão Sistemática

de trabalho. Sendo assim o artigo põe como requisitos: não volatilidade, ausência de partes móveis, latência entre dezenas e centenas de ns, e baixo custo. Ambos os artigos consideram que se a SCM evoluir muito, ela pode virar uma memória universal, com o artigo [4] citando diretamente as memórias universais e o artigo [12] apenas dizem que elas podem substituir do HDD até a DRAM.

Ou seja, de forma resumida, as SCM podem ocupar diversos lugares da hierarquia de memória, mas normalmente ficam entre o armazenamento primário e secundário, e todas as tecnologias que visam chegar ao ponto de memória universal, são SCM. A memória universal, por outro lado, precisa substituir quase que por inteira essa hierarquia, mas pelo menos a memória primária e secundária. [11]

2.2. Tecnologias Atuais e Limitações

Toda essa ideia das memórias universais ganhou força nos anos 2000 em que havia uma grande dificuldade para escalar tanto a Flash, como a DRAM e por isso surgiam novas tecnologias tentando substituir as tecnologias até então vigentes. O grande problema das memórias universais é que as características de retenção de dados, latência de leitura e escrita, janela de tensão de leitura e escrita, durabilidade de escrita consumo de energia e densidade estão inter-relacionadas [11]. Ou seja, quando se tenta focar em uma dessas características, outra acaba sendo prejudicada.

Sendo mais específico, normalmente baixa energia de escrita leva a uma baixa retenção de dados, baixa tensão de leitura gera uma leitura lenta, [9], a durabilidade de ciclos de escrita sempre vem com o custo de uma menor faixa de tensão de escrita, um maior tempo de escrita leva a maior retenção de dados e menor durabilidade de ciclos de leitura e escrita. As principais tecnologias potenciais para memórias universais são *Spin Transfer Torque Magnetic RAM* (STT-MRAM), *Phase Change Memory* (PCM) e *Resistive RAM* (ReRAM) [10] e todas elas apresentam uma ou mais dessas características limitantes mencionadas acima.

2.3. Tecnologias que se aproximam de Memórias Universais

Apesar de todas as limitações citadas na subseção anterior, as SCM ainda possuem o seu lugar na arquitetura de memória, e se aproximam das memórias universais ao atuar tanto como armazenamento primário como secundário. Analisando os dados de cada memória, é notável que existem SCM mais adequadas para substituir a DRAM e o cache, como STT-MRAM e *Spin Orbit Torque RAM* (SOT-MRAM) e outras mais adequadas para substituir o SSD, como a ReRAM e PCM [13, 3, 11]. Apesar de não haver atualmente uma arquitetura que misture essas diversas SCM, também não existem grandes limitações para que isso ocorra num futuro breve, apenas é necessário que essas memórias evoluam em latência e durabilidade de ciclos de escrita, latência de leitura, retenção de dados, para poder substituir as memórias SSD, DRAM e SRAM.

Como as SCM ainda estão em desenvolvimento, utilizá-las emparelhadas com outras memórias tradicionais, pode ser uma solução adequada. Um exemplo desse emparelhamento são as memórias híbridas, em que se utiliza tanto a SCM como a DRAM ou SRAM em conjunto. Ela pode ser de dois jeitos, um utilizando a memória RAM como cache da SCM e outra compartilhando o mesmo espaço de endereçamento, operando de modo paralelo, como também visto na memória Intel Optane [5]. A ideia é se aproveitar da maior densidade das SCM, aliada a performance da DRAM. Quando se trata de uma arquitetura híbrida, alguns problemas devem ser abordados. No geral as

SCM sofrem de uma escrita mais lenta que a leitura, alto consumo de energia (apesar de quando desligadas não consumirem), e durabilidade de ciclos de escrita limitado, [11] com exceção da STT-MRAM [2]. Desse modo é proposto em grande parte da literatura, que as SCM, na utilização como cache, foquem em processar dados de leitura intensa, enquanto a DRAM/SRAM, processem mais os dados da escrita [11].

3. Conclusão

No mercado atual não há nenhuma tecnologia que consiga substituir a memória principal e a memória secundária ao mesmo tempo. Como discutido no artigo, é improvável que isso ocorra num futuro próximo, porém, parte das tecnologias que tinham esse objetivo, ainda devem ser úteis. As SCM, podem substituir as memórias atuais, mesmo que não de forma universal. Outro uso das SCM são as memórias híbridas que também devem ser cada vez mais comuns no futuro. Desse modo, mesmo que o caminho das memórias universais seja improvável, toda essa pesquisa sobre elas deve servir para o mesmo objetivo de reduzir o gargalo entre CPU e armazenamento, via novas memórias.

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo nº 2019/26702-8 pelo apoio a este trabalho.

Referências

[1] P. Dimitrakis. Introduction to nvm devices. *Charge-Trapping Non-Volatile Memories: Volume 1–Basic and Advanced Devices*, pages 1–36, 2015.

[2] X. Dong et al. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions*, 31(7):994–1007, 2012.

[3] R. Gastaldi and G. Campardo. *In Search of the Next Memory: Inside the Circuitry from the Oldest to the Emerging Non-Volatile Memories*. Springer, 2017.

[4] G.Burr et al. Overview of candidate, device technologies for storage-class memory. *IBM Journal of Research and Development*, 52(4.5):449–464, 2008.

[5] Intel Corp. *Intel 64 and IA-32 Architectures Optimization Reference Manual*, 2022.

[6] J.Mittal and S.Mittal. Opportunities for nonvolatile memory systems in extreme-scale high-performance computing. *Computing in Science & Engineering*, 17(2), 2015.

[7] K.Ando et al. Non-volatile memories. *Normally-Off Computing*, pages 27–55, 2017.

[8] C. H. Lam. The universal semiconductor memory. In *2012 IEEE 11th International Conference on Solid-State and Integrated Circuit Technology*, pages 1–5, 2012.

[9] L.Baldi et al. Emerging memories. In *2013 Proceedings of the European Solid-State Device Research Conference (ESSDERC)*, pages 30–36, 2013.

[10] M. Marinella. The future of memory. In *2013 IEEE Aerospace Conference*, 2013.

[11] M.Rudan et al. *Springer Handbook of Semiconductor Devices*. Springer, 2023.

[12] R.Freitas and W.Wilcke. Storage-class memory: The next storage system technology. *IBM Journal of Research and Development*, 52(4.5):439–447, 2008.

[13] S.Yu and P.Chen. Emerging memory technologies: Recent trends and prospects. *IEEE Solid-State Circuits Magazine*, 8(2):43–56, 2016.

[14] A. Thomasian. *Storage Systems: Organization, Performance, Coding, Reliability, and Their Data Processing*. Academic Press, 2021.