

Comparador de Memórias Emergentes: Uma Ferramenta para a Análise de Novas Tecnologias

Pedro Ferro Laks¹, Emilio Francesquini¹

¹Centro 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

Abstract. *Recently, the advancement of research into new memory technologies, known as emerging memories, persistent memories, or Storage Class Memory (SCM), has been gaining attention. In this text, we discuss a comparison tool for these new memory devices. This tool features an interface divided into two columns to display information about each memory type. The data were gathered through a systematic review and can receive internal contributions. Our goal is to provide a reference material on which technologies currently deserve greater attention and where they should be positioned within the memory hierarchy.*

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 uma ferramenta de comparação desses novos dispositivos de memória. Essa ferramenta tem uma interface que se divide em 2 colunas para as informações de cada memória. Os dados foram obtidos via revisão sistemática e podem receber contribuições internos. Esperamos fornecer um material de referência sobre quais tecnologias atualmente merecem uma maior atenção e onde elas devem se situar na hierarquia de memória.*

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 memória dinâmica de acesso aleatório (DRAM) e o armazenamento secundário. Este processo é caracterizado por sua lentidão, resultando na subutilização da Unidade Central de Processamento (CPU) devido à falta de recebimento de dados em tempo hábil. Nos últimos anos, houve uma crescente adoção dos discos de estado sólido (SSD); entretanto, esses dispositivos não resolvem os desafios associados ao armazenamento interno, como a velocidade limitada e a menor durabilidade em comparação com a memória DRAM. Embora os SSDs superem os discos rígidos (HDD) em termos de latência de leitura e escrita, comumente atingindo até 200 vezes mais na leitura e 10 vezes mais na escrita, eles ainda permanecem mais lentos que as memórias DRAM, com uma diferença aproximada de 500 vezes na leitura e 10^4 vezes na escrita [3]. Além disso, os SSDs não têm conseguido reduzir significativamente a latência de leitura e escrita ao longo dos anos. Adicionalmente, os SSDs apresentam baixa durabilidade, geralmente suportando cerca de 10^5 ciclos de escrita, um valor na ordem de 10^{10} vezes menor que os HDDs e as DRAM, que possuem durabilidade de 10^{15} ciclos. Outro problema é que a

própria DRAM apresenta limitações, não conseguindo evoluir na mesma velocidade de antes, o que tem exacerbado o gargalo de memória [3][2]. A SRAM enfrenta o mesmo problema de escalabilidade que a DRAM, mesmo com os avanços nos processos de litografia utilizados em sua fabricação. À medida que a miniaturização dos componentes continua a avançar, a densidade dos transistores aumenta, mas isso também intensifica os desafios relacionados à dissipação de calor, consumo de energia e estabilidade do circuito. Essas limitações tecnológicas impedem a SRAM de acompanhar a demanda crescente por maior capacidade e desempenho nas aplicações de computação que exigem um alto desempenho [4]. Ao longo da evolução tecnológica, foram exploradas novas tecnologias de memória visando maior desempenho e capacidade de armazenamento. Essas novas tecnologias emergentes, alternativas à DRAM, são majoritariamente memórias não voláteis, o que significa que não necessitam de uma fonte de energia constante para manter seu conteúdo. Devido a essa característica, elas foram denominadas memórias emergentes. As memórias emergentes têm o potencial de substituir as memórias tradicionais ou atuar como uma camada adicional na hierarquia de memória. Um possível uso dessas tecnologias seria eliminar ou reduzir a troca de dados entre o armazenamento primário e secundário, entregando os dados à CPU de maneira mais rápida [3][2].

Para uma melhor compreensão do panorama dessas novas tecnologias, propõe-se o desenvolvimento de uma ferramenta do tipo comparador. Essa ferramenta está em formato de site e estruturada em duas colunas, permitindo a comparação simultânea de dados. Por meio dessa plataforma, os usuários poderão obter uma visão mais clara sobre o posicionamento de diferentes tipos de memória em relação a suas concorrentes emergentes e às tecnologias atualmente vigentes.

2. Metodologia e Revisão Sistemática

Todos os dados utilizados na ferramenta foram obtidos a partir de uma revisão sistemática refinada, baseada na análise do nosso próprio artigo [1]. O refinamento consistiu na extração dos dados operacionais das memórias presentes em cada artigo revisado.

As memórias SOT-MRAM, STT-MRAM, PCM, ReRAM, FeRAM, FeFET foram as mais citadas e por isso foram analisadas. Como base de comparação, as memórias vigentes DRAM, SRAM, Nand-flash, Nor-flash também foram selecionadas. Os dados coletados dessas memórias foram: latência de escrita, latência de leitura, energia de escrita, energia de leitura, durabilidade em ciclos de escritas, e tamanho de célula em F².

Esses dados são de extrema importância, pois determinam em qual lugar da hierarquia de memória, cada tecnologia pode entrar. As latências determinam o quão rápido a memória consegue entregar as informações para a CPU. A energia, significa o consumo energético por bit de cada memória. A durabilidade significa o número de ciclos de escrita que cada memória consegue fazer até falhar. O tamanho de célula em F² é importante, pois determina a densidade da memória e desse modo, determina a possibilidade de memórias com grande capacidade de armazenamento.

Os dados coletados de cada artigo, foram colocados em planilhas que foram analisadas através de códigos Python. Nessa planilha também há o ano de cada memória, no entanto, quando não foi possível encontrar o ano dela foi utilizado o ano do artigo.

3. Ferramenta de Comparação

Na ferramenta de comparação, os dados coletados foram convertidos em gráficos. Para fazê-los, foi utilizado o Python por meio do Google Colab ¹. Como o objetivo era que o gráfico fosse interativo, foi utilizado a extensão do Vega Lite ². Através dele, é possível fazer gráficos de densidade kernel, no qual revela a probabilidade de cada valor ocorrer. Por exemplo, um dado com densidade kernel 0,2 é um valor 2 vezes mais provável do que um de densidade 0,1. Com o Vega-Lite, foram implementados dois mecanismos de interação para o usuário. O primeiro permite selecionar um intervalo de datas diretamente no gráfico superior, clicando e arrastando o cursor sobre o período desejado. Essa seleção funciona como um filtro e atualiza automaticamente o gráfico de densidade abaixo, exibindo apenas os dados correspondentes ao intervalo de datas escolhido. O segundo mecanismo é um seletor de largura de banda de densidade kernel, no qual o usuário pode fazer o ajuste fino da densidade de probabilidade, ou seja, o quão preciso ou aproximado são os dados. Aqui abaixo estão alguns exemplos de gráficos obtidos:

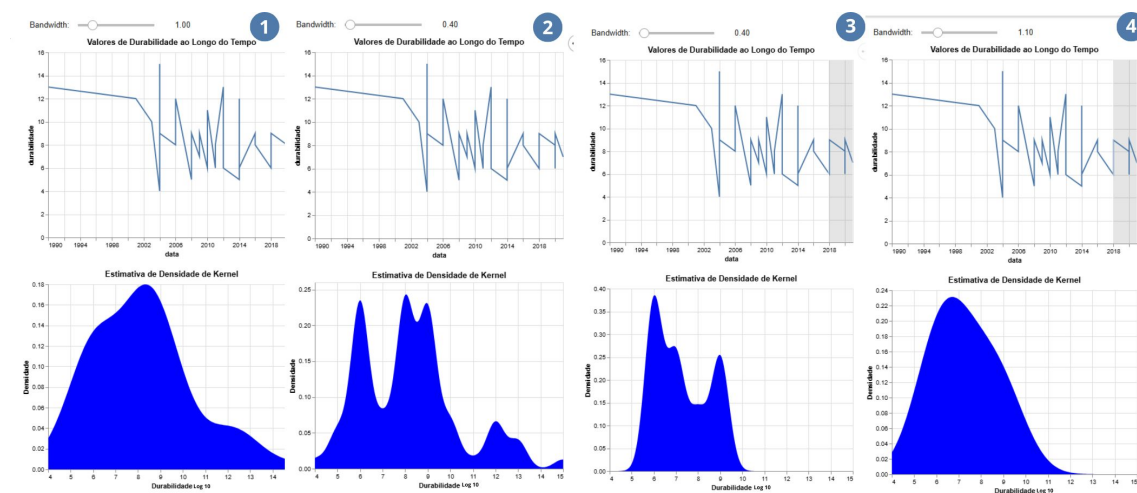


Figura 1. Gráfico 1: largura de banda 1 e todos os anos, Gráfico 2: largura de banda 0,4 e todos os anos, Gráfico 3: largura de banda 0,4 e data 2018 adiante, Gráfico 4: largura de banda 1 e data 2018 adiante

Da figura 1 para a 2 é possível analisar o impacto da alteração da largura de banda na densidade kernel. Um valor interessante de se notar é a durabilidade de 10^7 nos gráficos de 2018 adiante. Esse valor se destaca significativamente com uma largura de banda de 1,0. No entanto, ao reduzi-la para 0,4 e, conseqüentemente, suavizar menos a distribuição, observa-se que 10^7 é menos frequente que 10^6 . Ainda assim, ele se posiciona entre 10^8 e 10^9 , que são os valores mais recorrentes, com predominância de 10^6 . Dessa forma, 10^7 pode ser considerado um valor altamente provável dentro desse intervalo.

Olhando para o caso de datas gerais, a discrepância é ainda mais evidente. Com uma largura de banda de 1,0, 10^7 aparece como um dos principais valores. No entanto, ao reduzir o kernel, observa-se um buraco em 10^7 , indicando que, na realidade, não é um valor tão frequente. A conclusão obtida é semelhante ao caso de 2018 em diante: devido à grande presença de 10^6 , 10^8 e 10^9 , o valor 10^7 atua, na prática, como um valor médio e provável dentro desse intervalo. Por fim, o seletor de data permite o usuário compreender

¹<https://colab.google/>

²<https://vega.github.io/vega-lite/>

como a tecnologia evoluiu ao longo do tempo, e como ela está atualmente.

Algo que também é possível de ser feito com a ferramenta é comparar duas tecnologias diferentes de memória:

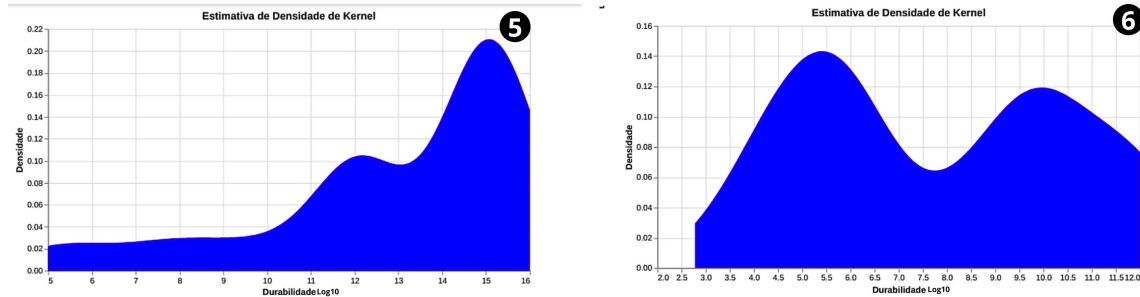


Figura 2. Gráfico: 5 STT-MRAM, Gráfico 6: ReRAM

Nesse exemplo, é possível notar que a STT-MRAM possui, em geral, uma durabilidade significativamente superior à ReRAM. Enquanto a ReRAM apresenta variações, com picos em 10^5 e 10^{10} , a STT-MRAM exibe valores bem concentrados em torno de 10^{15} e 10^{16} . Outro aspecto a ser analisado é a regularidade da curva, que indica o nível de maturidade da tecnologia. Curvas com menor variação tendem a representar tecnologias mais consolidadas, enquanto curvas com maior variabilidade, como no caso da ReRAM, podem indicar tecnologias ainda em desenvolvimento ou até mesmo a presença de duas tecnologias distintas no mesmo gráfico.

A ferramenta de comparação está em estágio de desenvolvimento. Ele foi feito utilizando Css, Html, e Python, mais especificamente flask do Python para a ferramenta funcionar de forma dinâmica.

4. Conclusão

Criamos um mecanismo comparador entre memórias emergentes para compilar dados de cada tecnologia ao longo dos anos, facilitando a compreensão da hierarquia de memória e da evolução tecnológica. Por meio de gráficos de densidade de probabilidade, conseguimos visualizar os valores esperados e comuns de cada característica operacional das memórias emergentes. Essa análise permite identificar quais tecnologias apresentam os melhores atributos para diferentes aspectos de desempenho. Dessa forma, nosso objetivo é fornecer um material de referência abrangente e acessível, auxiliando na comparação e seleção das memórias emergentes mais adequadas para cada aplicação.

Referências

- [1] P. F. Laks and E. Franceschini. Storage class memory: Avanços e limitações em direção às memórias universais. In *(ERAD-SP)*. SBC, 2024.
- [2] M. Si, H.-Y. Cheng, T. Ando, G. Hu, and P. D. Ye. Overview and outlook of emerging non-volatile memories. *Mrs Bulletin*, 46(10):946–958, 2021.
- [3] J. S. Vetter and S. Mittal. Opportunities for nonvolatile memory systems in extreme-scale high-performance computing. *Computing in Science & Engineering*, 2015.
- [4] S. Yu and T.-H. Kim. Semiconductor memory technologies: State-of-the-art and future trends. *Computer*, 57(4):150–154, 2024.