Uma Análise Experimental da Utilização de Diferentes Tecnologias de Armazenamento em um SGBD Relacional

Francisco D. B. S. Praciano[1], **J. Filipe L. de Sousa**[1], **Javam C. Machado**[1]

[1]Laboratório de Sistemas e Bancos de Dados (LSBD)
DC/UFC – CEP 60440-900 – Fortaleza – CE – Brasil

{daniel.praciano,filipe.lobo,javam.machado}@lsbd.ufc.br

Abstract. *Traditional Database Management Systems (DBMS) are built on the premise that data is stored on hard disks drives (HDD). Recently, alternatives to HDDs have emerged, such as solid state drives (SSD), non-volatile memories (NVM) and new main memories (DRAM). Different characteristics of these devices may impact the performance of DBMSs. In this work, we propose to analyze a DBMS that stores its data in four different ways, in HDD, SSD NVM, DRAM and in a hybrid way, using the three devices together. To do this, we use a TPC-C workload and discuss the reasons that give rise to the results obtained for each type of storage.*

Resumo. *Os Sistemas Gerenciadores de Banco de Dados (SGBD) tradicionais são construídos com a premissa de que os dados estão armazenados em discos rígidos (HDD). Recentemente, surgiram várias alternativas aos HDDs, tais como as unidades de estado sólido (SSD), as memórias não voláteis (NVM) e as novas memórias principais (DRAM). As diferentes características desses dispositivos podem impactar no desempenho dos SGBDs. Neste trabalho, nos propomos a analisar um SGBD que armazena seus dados de quatro formas distintas, em HDD, SSD NVM, DRAM e de forma híbrida, utilizando os três dispositivos em conjunto. Para isso, usamos a carga de trabalho TPC-C e discutimos os motivos que dão origem aos resultados obtidos para cada tipo de armazenamento.*

1. Introdução

Dada a natureza dos Sistemas Gerenciadores de Banco de Dados (SGBDs), o desenvolvimento de novas tecnologias de armazenamento juntamente com a evolução tecnológica dos dispositivos trazem benefícios tais como o aumento na vazão de transações efetuadas. É muito comum que esses sistemas tenham sido desenvolvidos considerando as características dos dispositivos disponíveis, dificultando o acompanhamento da evolução tecnológica e a capacidade de tirar proveito das melhorias trazidas pelos novos dispositivos de armazenamento. Pensando nisso, esse trabalho tem como objetivo realizar uma avaliação experimental de um SGBD relacional a fim de pontuar e avaliar o impacto que a substituição da tecnologia de armazenamento subjacente pode causar nesses sistemas. Para tanto, nos propomos a observar a performance do PostgreSQL com os diferentes dispositivos de armazenamento, HDD, SSD NVM e DRAM. A investigação apresentada nesse trabalho busca discutir esse impacto por meio dos seguintes questionamentos: 1) Qual o impacto das tecnologias de armazenamento na vazão de um SGBD tradicional? 2) Dada as características dos dados e da carga de trabalho, é possível construir um modelo híbrido de armazenamento? 3) Qual seria a relação custo-benefício desse armazenamento híbrido?

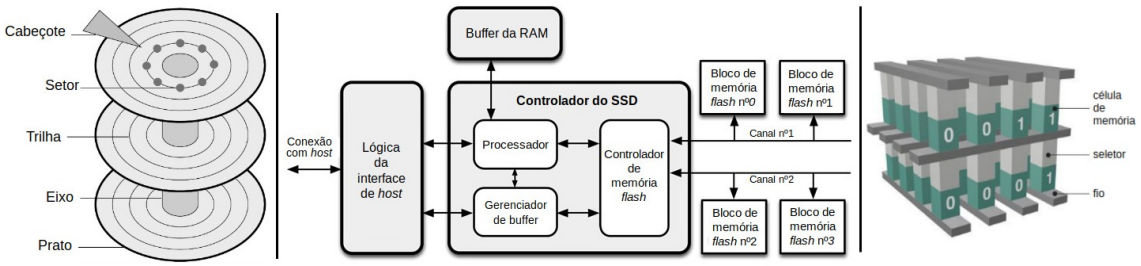

Figura 1. Esquema de funcionamento do HDD, do SSD e do NVM 3D XPoint, respectivamente. Modificado de [Zukowski 2009, Leo Kelion 2015].

2. Tecnologias de Armazenamento

Os discos rígidos (HDD) são os dispositivos de armazenamento mais utilizados pelos SGBDs relacionais. Esses sistemas adotam várias premissas considerando que o armazenamento dos dados seria feito em um HDD, por exemplo a adoção de uma interface de armazenamento orientada a blocos. Nos últimos anos, novos dispositivos de armazenamento têm sido propostos, como as unidades de estado sólido (SSD) tanto baseadas em memórias *flash* quanto nas novas tecnologias de memórias não-voláteis (NVM).

A Figura 1 mostra os componentes da arquitetura desses dispositivos em que a ausência de partes mecânicas nos SSDs, como eixo de rotação ou cabeçote de leitura, fornece armazenamento persistente com taxas de acesso mais altas do que dos atuais HDDs, enquanto que seu custo por *byte* armazenado tem consistentemente diminuído a cada ano [Shah et al. 2008]. O SSD NVM utilizado nesse trabalho, Intel Optane, é construído baseado na tecnologia *3D XPoint* com uma estrutura em três dimensões (3D) e que possibilita o acesso direto às células de armazenamento. Essa organização, mostrada na Figura 1, traz melhorias tanto no desempenho quanto na densidade das células quando comparado aos SSDs baseados na tecnologia *flash*.

A evolução das tecnologias de DRAMs tornou viável a construção de SGBDs totalmente voltados para a memória principal. Esse tipo de sistema potencializa a maneira de se organizar os dados armazenados de tal modo a aproveitar o endereçamento por *bytes* fornecido por esse tipo de dispositivos. Nesse trabalho, o SGBD utilizado é o PostgreSQL que implementa uma interface orientada a blocos no seu módulo de armazenamento. Por esse motivo, vamos utilizar o sistema de arquivos *ramfs* que permite o endereçamento por blocos da memória principal para realizar a avaliação nesse tipo de dispositivo em paridade com os outros dois que são orientados a blocos.

3. Trabalhos Relacionados

Vários trabalhos na literatura se propuseram a realizar estudos sobre o impacto de armazenar dados em tecnologias distintas nos sistemas de bancos de dados. [Xu et al. 2015] apresentou uma caracterização juntamente com uma análise experimental detalhada do desempenho dos SSDs que fazem uso do padrão NVMe. Para tanto, foram utilizados o Cassandra, o MongoDB e o MySQL. O nosso trabalho também aborda o desempenho de sistema relacional através de uma avaliação experimental, mas diferencia-se quer seja no fato de utilizar um dispositivo diferente, quer seja na maneira de avaliação.

Em um outro trabalho, [Brayner and Monteiro 2016] investigaram o desempenho

de três SGBDs comerciais usando como dispositivo de armazenamento tanto HDD quanto SSD argumentando que, no futuro, esses sistemas deveriam ser conscientes do *hardware* subjacente. A avaliação foi feita com a utilização das cargas de trabalhos TPC-H e TPC-E, diferente da nossa escolha pela carga TPC-C. Além disso, sua avaliação fez uso de dispositivos SSDs de interface SATA e de uso pessoal, diferentemente da nossa estratégia que avalia o armazenamento em um SSD 3D XPoint de alto desempenho.

Um estudo experimental dos SSDs NVMe foi realizado em [Son et al. 2016] com o uso de *microbenchmarks* e também com o TPC-C, sendo o MySQL usado para gerenciar os dados. Os autores mostraram os resultados obtidos quando várias configurações de E/S tanto do sistema operacional quanto do SGBD são usadas. Além da diferença do SGBD escolhido, experimentamos outras configurações utilizando tecnologia diferente.

4. Avaliação do Impacto

Este trabalho investiga, por meio da execução de cargas transacionais, o desempenho medido por vazão e latência de um SGBD relacional quando tecnologias distintas de armazenamento são utilizadas. Para cargas transacionais, o *benchmark* TPC-C é considerado padrão, assim optamos por fazer uso da implementação OLTP-Bench [Difallah et al. 2013].

Para a realização dos experimentos, foi utilizada uma máquina com HDD, SSD NVMe e DRAM conforme a Tabela 1, um processador Intel Xeon E5-2609 v3 1.9GHz, 15M Cache, 6 núcleos, com a versão do kernel GNU/Linux 4.15.0. Além disso, o PostgreSQL (versão 11.1) foi escolhido como SGBD relacional. Para obter cada resultado apresentado nas próximas subseções foram realizados 10 execuções independentes em cada um dos dispositivos do respectivo experimento e, então, a média da métrica de interesse foi calculada. Na avaliação de cada dispositivo, nos asseguramos de eliminar E/Ss nos outros dispositivos. Por fim, destaca-se que o cache utilizado pelo PostgreSQL é composto de duas partes: uma própria e uma outra fornecida pelo sistema operacional (SO). Como essa última pode impactar no resultado obtido, por exemplo melhorando o desempenho do HDD, modificamos o nosso *setup* de forma a assegurar que o cache do SO não afete o desempenho dos dispositivos.

4.1. O Impacto no Desempenho

Antes de avaliarmos o PostgreSQL com os dispositivos mencionados, inicialmente fizemos uso de duas ferramentas, *dd* e *fio*, que estão presentes nos sistemas Linux a fim de pontuarmos as diferenças de desempenho entre os dispositivos sem ainda utilizar um sistema complexo como um SGBD. A variação da taxa de transferência entre os dispositivos é mostrada na Figura 2 quando fazemos uso da ferramenta *dd* que permite a leitura ou escrita de arquivos em um dispositivo específico. Esta Figura mostra a taxa de transferência alcançada para ler e escrever um arquivo de 1GB em cada um dos dispositivos. É possível observar que, tomando como referência a taxa de transferência do HDD, o SSD NVM tem taxa 5 vezes maior, enquanto que para a DRAM a mesma taxa é 15 vezes maior.

Da mesma forma, usamos o *fio* a fim de obter mais detalhes sobre as possíveis maneiras de realizar operações de E/S: leitura aleatória (LA), escrita aleatória (EA), leitura sequencial (LS) e escrita sequencial (ES). Na Figura 3, é apresentada a vazão (número de operações de entrada e saída por segundo - IOPS) dessas operações em cada um dos dispositivos. Chama a atenção a pequena vazão alcançada pelo HDD quanto às operações realizadas de maneira aleatória. Isso é esperado devido às características do HDD

Característica	HDD 7.2K RPM	SSD NVM	DRAM
Fabricante	Dell	Intel	Smart
Capacidade (GB)	1000	375	24
Latência (ms)	$8,3$	$0,01$	$5 \cdot 10^{-7}$
Largura de Banda (GB/s)	1,2	2,4	17
Preço/bit (R$/GB)	1,99	20,27	43,75

Tabela 1. Características dos dispositivos de armazenamento.

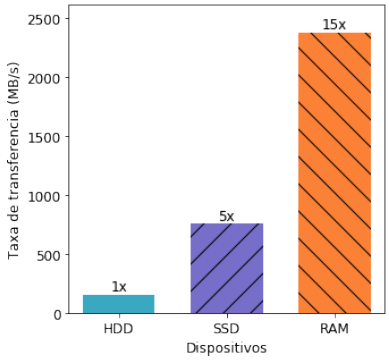

Figura 2. Taxa de transferência utilizando *dd*.

que fazem com que as partes mecânicas tenham um grande impacto no desempenho. Por outro lado, ressalta-se o ganho de desempenho desse dispositivo quando as operações são sequenciais. Essa diferença entre operações aleatórias e sequenciais não ocorre nos outros dispositivos. Por fim, a constante evolução dos dispositivos secundários permite que a vazão alcançada por eles se aproxime daquela alcançada pelas memórias principais.

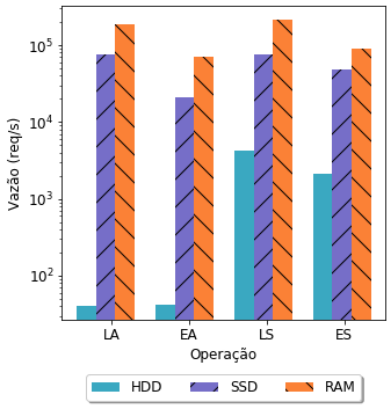

Figura 3. Vazão (IOPS) dos dispositivos obtidos por meio da ferramenta *fio*.

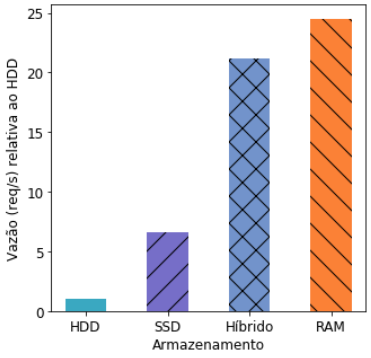

Figura 4. Vazão (req/s) em cada um dos dispositivos ao executar o TPC-C.

Visando responder ao primeiro ponto levantado no início desse trabalho, experimentamos a carga de trabalho padrão do TPC-C (i.e., porcentagem 45-43-4-4-4 de cada uma das transações) com tempo de *warm-up* de 5 segundos e fator de escala igual a 50 juntamente com a configuração padrão do PostgreSQL cujo resultado está apresentado nas Figuras 4 e 5. A Figura 4 mostra que a vazão do dispositivo de armazenamento subjacente pode acelerar em até 24 vezes o processamento de transações pelo PostgreSQL. Somente a troca do HDD pelo SSD já traz uma melhora de 7 vezes na vazão dos SGBDs. Conclui-se que a troca do dispositivo, sem mudanças na configuração do SBGD, já traz um ganho para o desempenho desses sistemas ao lidar com cargas de trabalho transacionais. Esse efeito também é visualizado na queda da latência das transações, conforme a Figura 5. Esta melhora se dá porque as transações do TPC-C realizam operações aleatórias de entrada e saída, característica marcante das cargas transacionais. Consequentemente, o desempenho é impactado pelos motivos apresentados acima ao utilizar o *fio*.

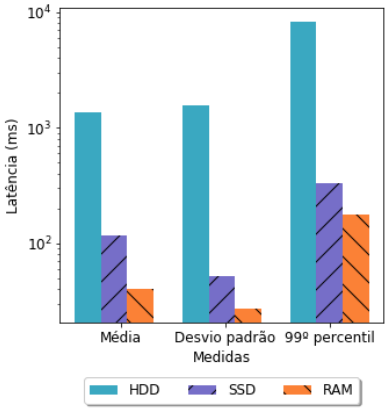

Figura 5. Medidas da latência (ms) em cada um dos dispositivos ao executar o TPC-C.

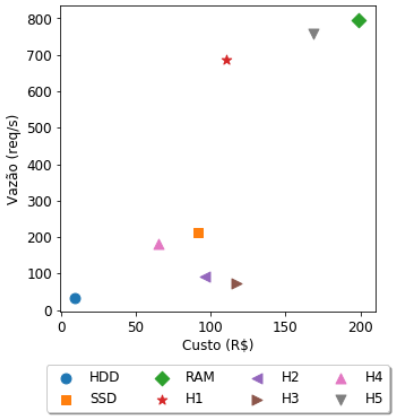

Figura 6. Relação custo-benefício do HDD, SSD e das formas híbridas.

4.2. Armazenamento Híbrido

O TPC-C é constituído por 9 relações e 5 tipos de transações. No armazenamento híbrido, os três dispositivos são usados conjuntamente de modo a realizar a divisão do banco de dados entre eles. Para tanto, utilizamos duas maneiras para decidir em qual dispositivo cada relação deveria ser armazenada, sendo que ambas seguiram uma distribuição uniforme, isto é, cada um dos dispositivos recebeu 3 relações. Enquanto que na primeira maneira é utilizado o tamanho das relações como fator de decisão, na segunda é utilizada a quantidade de operações de leitura e escrita que são realizadas nas relações. Cinco configurações (H1 - H5) foram geradas, sendo que H2 e H5 utilizaram a primeira forma de alocação e as restantes utilizaram a segunda. Como pode ser visto na Figura 4, a vazão alcançada pela configuração H1, a qual obteve o melhor custo-benefício, apresentou uma melhora em torno de 20 vezes quando comparado com a vazão do HDD. Nessa configuração, as relações TPC-C *Warehouse*, *District* e *Customer* foram alocadas na RAM, enquanto que as *Order-Line*, *Order* e *Stock* no SSD e, por fim, as *History*, *Item* e *New-Order* no HDD. Observa-se portanto que é possível construir um modelo híbrido de armazenamento de forma a considerar a carga de trabalho e ainda assim obter um bom desempenho.

4.3. Relação Custo-Benefício

Mesmo que a troca direta do HDD pela RAM traga melhoria significativa no desempenho do PostgreSQL ao se executar uma carga transacional, não se pode desconsiderar o custo adjacente relativo a essa troca. Muito embora as tecnologias de memória principal tenham evoluído constantemente ao longo dos anos e, dessa forma, melhorado o custo-benefício desse tipo de tecnologia de armazenamento, ainda existe uma diferença relevante quando se compara o custo dessas tecnologias. Por exemplo, no caso dos dispositivos utilizados nesse trabalho, essa diferença está em torno de 40 reais por cada GB. Com o advento das novas tecnologias de NVM, como a 3D XPoint aqui utilizada, essa lacuna diminui, mas ainda é considerável. A diferença entre o custo por GB do HDD para o SSD utilizado é de aproximadamente 18 reais. A Figura 6 mostra a relação entre o custo de manter o banco de dados e a vazão obtida, quer seja utilizando um dos dispositivos, quer seja realizando as várias formas híbridas descritas anteriormente. Observe as configurações híbridas H1 e H4. Com um custo 2 vezes menor e mantendo uma vazão bem próxima

do SSD, a configuração H4 apresenta uma melhor relação custo-benefício em relação ao SSD, podendo dessa forma ser uma alternativa viável. A configuração H1 obteve o maior ganho na relação custo-benefício, visto que essa apresentou um desempenho similar ao da RAM com um custo menor. Pegando como base essa configuração, temos que a relação custo-benefício do armazenamento híbrido seria de 6, enquanto que a RAM é de 4.

5. Conclusões e Trabalhos Futuros

Este trabalho apresenta uma análise experimental da influência que as diferentes tecnologias de armazenamento podem trazer para o desempenho de um SGBD relacional, PostgreSQL, quando são utilizadas como o hardware subjacente. Baseado nos resultados experimentais, discutimos os três questionamentos levantados nesse trabalho. Constatou-se que a rápida evolução das tecnologias de armazenamento traz uma melhoria significativa para os SGBDs. Em particular, pontuamos que a troca direta de HDD para DRAM causou um impacto positivo de 24 vezes na vazão de uma carga transacional. Não obstante, o custo dos dispositivos pode ser proibitivo. Assim, mostramos que é possível construir um armazenamento híbrido de tal forma a melhorar o custo-benefício em até 2 vezes em relação à RAM. Futuramente, pretendemos aprofundar o estudo com o intuito de propor uma solução autônoma para escolher a melhor organização dos dados no formato híbrido.

Agradecimentos

Esta pesquisa foi parcialmente apoiada pela CAPES (processo #1782887) e LSBD/UFC. Agradecemos ao Ítalo Cavalcante de Abreu pelo auxílio dado neste trabalho.

Referências

Brayner, A. and Monteiro, J. M. (2016). Hardware-aware database systems: A new era for database technology is coming - vision paper. In *31º Simpósio Brasileiro de Banco de Dados, SBBD 2016, Salvador, Bahia, Brasil, October 4-7, 2016.*, pages 187–192.

Difallah, D. E., Pavlo, A., Curino, C., and Cudre-Mauroux, P. (2013). Oltp-bench: An extensible testbed for benchmarking relational databases. *Proc. VLDB Endow.*, 7(4):277–288.

Leo Kelion, BBC, I. M. (2015). 3d xpoint technology. `https://www.bbc.com/news/technology-33675734`. Accessed: 2019-07-15.

Shah, M. A., Harizopoulos, S., Wiener, J. L., and Graefe, G. (2008). Fast scans and joins using flash drives. In *4th Workshop on Data Management on New Hardware, DaMoN 2008, Vancouver, BC, Canada, June 13, 2008*, pages 17–24.

Son, Y., Kang, H., Han, H., and Yeom, H. Y. (2016). An empirical evaluation and analysis of the performance of NVM express solid state drive. *Cluster Computing*, 19(3):1541–1553.

Xu, Q., Siyamwala, H., Ghosh, M., Suri, T., Awasthi, M., Guz, Z., Shayesteh, A., and Balakrishnan, V. (2015). Performance analysis of nvme ssds and their implication on real world databases. In *Proceedings of the 8th ACM International Systems and Storage Conference, SYSTOR 2015, Haifa, Israel, May 26-28, 2015*, pages 6:1–6:11.

Zukowski, M. (2009). Balancing vectorized query execution with bandwidth-optimized storage. *Journal of Computational Physics - J COMPUT PHYS*.