Análise de desempenho dos serviços de armazenamento da Nuvem Computacional para execução de *checkpoint*

Charles B. Rodamilans[1], **Edson Borin**[1]

[1] Laboratório Multidisciplinar de Computação de Alto Desempenho (LMCAD)
Instituto de Computação (IC) – Universidade Estadual de Campinas (Unicamp)
Campinas – SP – Brasil

{charles.rodamilans,edson.borin}@ic.unicamp.br

Abstract. *Cloud computing has been used to execute high-performance computing applications due to its potential to reduce costs. The instability of low-cost instances - Spots - can be mitigated with checkpointing, and the performance of the storage service becomes crucial to avoid increasing the total runtime excessively. We compared the performance of four Cloud Computing storages services on AWS (EBS, EFS, FSx for Lustre e S3) for checkpoint persistence and observed that: (1) two of the four storage services showed scalability; (2) the storage service can increase the write performance by up to 727%.*

Resumo. *A Nuvem Computacional tem sido utilizada para executar as aplicações de computação de alto desempenho devido ao seu potencial para reduzir custos. A instabilidade das instâncias com custo reduzido - Spots - podem ser contornadas com mecanismos de checkpointing e o desempenho do serviço armazenamento se torna crucial para evitar o aumento demasiado no tempo total da execução. Foi comparado o desempenho de quatro serviços de armazenamento da Nuvem Computacional da AWS (EBS, EFS, FSx for Lustre e S3) para a persistência de checkpoints e observou-se que: (1) dois dos quatro serviços de armazenamento apresentaram escalabilidade; (2) o serviço de armazenamento pode aumentar o desempenho de escrita em até 727%.*

1. Introdução

A Nuvem Computacional vem sendo utilizada para Computação de Alto Desempenho devido à possibilidade de se obter um menor custo de criação e manutenção dos recursos quando comparada a um ambiente local (*on-premise*) e também por permitir a criação de máquinas virtuais com a configuração (de *software* ou de *hardware*) adequada para a aplicação [Netto et al. 2018]. Neste contexto, uma forma de reduzir os custos da utilização da Nuvem Computacional é utilizar as instâncias de custo reduzido - Spots.

O provedor de Nuvem oferece as instâncias que não estão sendo usadas - Spots - por um preço menor e solicita a sua recuperação quando necessita, fornecendo ao usuário um período de tempo (de 30 a 120 segundos) antes de finalizá-la. Para contornar esta indisponibilidade, pode-se utilizar mecanismos de *checkpointing*, principalmente para as aplicações de alto desempenho que podem demorar dias para serem executadas.

O *checkpoint* é um mecanismo de tolerância a falhas que salva o estado da aplicação e permite restaurar o estado salvo. Isto possibilita que pouco processamento realizado pela aplicação possa ser perdido quando houver a solicitação de retomada da instância por parte do provedor. O *checkpoint* precisa ser persistido para ser posteriormente recuperado e seu desempenho está diretamente relacionado ao desempenho do armazenamento.

A Nuvem Computacional oferece diferentes serviços de armazenamento com diferentes desempenhos, que influenciam diretamente no desempenho do *checkpoint* e, consequentemente, no tempo total de execução da aplicação. O armazenamento em disco

local, normalmente, é a opção mais rápida para a persistência dos arquivos de *checkpoint*. Porém, nem todos os tipos de instâncias da Nuvem Computacional possuem disco local sendo necessário utilizar os armazenamentos que são acessados pela rede.

Neste trabalho comparamos o desempenho de quatro serviços de armazenamento da Nuvem Computacional para a realização do *checkpoint* e apresentamos as seguintes contribuições: (1) dois dos quatro serviços de armazenamento apresentaram escalabilidade horizontal; (2) o serviço de armazenamento pode aumentar o desempenho de escrita em até 727% para a realização do *checkpoint* ; (3) o armazenamento que apresenta a maior vazão esperada tem a variação de eficiência (vazão obtida por razão esperada) de escrita entre 65% a 72%.

2. Materiais e métodos

Os experimentos foram executados na Nuvem Computacional da *Amazon Web Services* (AWS), no período de novembro a dezembro de 2019, na região us-east-1b. Para obtenção dos tempos do *checkpoint*, cada experimento foi executado 3 vezes. Os resultados no gráfico apontam a média com intervalo de confiança de 95%. Utilizamos o *benchmark* NAS/NPB LU (NPB 3.3.1, Classe D) e realizamos os *checkpoints* com o auxílio da ferramenta *Distributed MultiThreaded Checkpointing* (DMTCP 2.5.2), sendo que o coordenador foi executado em uma das máquinas de processamento.

Utilizamos a instância do tipo c5n.18xlarge (não tem disco local). Esta instância possui 72 CPUs virtuais (vCPUs), é baseada no processador Intel Xeon Platinum 3,0 GHz, com 192 GiB de memória RAM, e possui a largura de banda de rede de 100 Gbps (com *Elastic Network Adapter* (ENA)) e de EBS de 14 Gbps. O sistema Operacional utilizado foi o Ubuntu 16.04.5 LTS. Para cada máquina virtual (MV), foram utilizados 64 vCPUs (ao invés de 72 vCPUs). Os experimentos foram realizados com 1, 2, 4 e 8 MVs; 64, 128, 256, 512 vCPUS; e um processo MPI por vCPU.

Os serviços de armazenamento da AWS utilizados foram *Elastic Block Store* (EBS), *Elastic File System* (EFS), *FSx for Lustre* e *Simple Storage Service* (S3). Os serviços EBS e EFS possuem dois modos de vazão (*throughput*): *bursting*, que possui picos de desempenho durante um período de tempo; e provisionado, que oferece desempenho constante a um custo maior. Utilizou-se o modo *bursting* e monitorou-se os créditos do *bursting* (métrica `BurstCreditBalance` do serviço de monitoramento AWS CloudWatch) para garantir o uso do pico de desempenho durante os experimentos.

O EBS é um serviço de armazenamento em bloco. Foi empregado o de Finalidade Geral (SSD gp2). Cada volume foi criado com o tamanho 100 GiB e fornece a vazão máxima de 128 MiB/s por volume. Foi criado 1 volume EBS para cada MV com sistema de arquivos ext4. O EFS é um serviço de sistema de arquivos que opera com o protocolo *Network File System* versão 4 (NFSv4). O modo de desempenho utilizado foi de uso geral (*General Purpose*) e o modo de vazão foi o de *bursting*. No tamanho de 100 GiB, o modo *bursting* fornece a vazão de 100 MiB/s por até 72 minutos por dia. O tamanho do sistema de arquivos do EFS é dimensionado conforme a necessidade (tamanho dinâmico).

O FSx for Lustre é um serviço de sistema de arquivo paralelo e distribuído baseado no Lustre. Foi utilizado o Lustre versão 2.10.6 e um sistema de arquivos de tamanho 1.200 GiB (tamanho mínimo oferecido pelo provedor de Nuvem). Isto significa que foi criado apenas um disco (ou *Object Storage Target* - OST do Lustre) e é oferecido a vazão máxima de 240 MiB/s. O S3 é um serviço de armazenamento de objetos. Foi utilizado o S3 de Propósito Geral (*S3 Standard*) e para montar o S3 como sistema de arquivos foi utilizado o programa S3FS-Fuse versão 1.79. O S3 possui o tamanho dinâmico e não foi encontrado na documentação da AWS o desempenho esperado para este serviço.

3. Resultados e discussões

Os tamanhos dos arquivos gerados pela realização do *checkpoint* com o DMTCP e EBS obtidos foram utilizados como base para calcular o tamanho do *checkpoint* por processo e para todas as MVs (Tabela 1). O tamanho do arquivo de *checkpoint* por processo (e por MV) diminui com o aumento do número de processos devido à diminuição da memória necessária por cada processo do NPB LU (também observado em [Cao et al. 2014]).

Tabela 1. Tamanhos dos *checkpoints* (obtidos com EBS).

MVs	Processos	*Ckpt* Por Processo (GiB)	*Ckpt* Por MV (GiB)	*Ckpt* Todas MVs (GiB)
1	64	0,19	12,11	12,11
2	128	0,11	6,92	13,84
4	256	0,06	4,09	16,36
8	512	0,04	2,72	21,79

O EBS obteve o melhor tempo de execução, com variação de 29% a 727%, sobre os demais serviços de armazenamento (Tabela 2), com a exceção do experimento de 1 MV, em que foi pior 37% comparado ao FSx for Lustre. A boa escalabilidade apresentada (Figura 1) ocorre devido ao número de volumes alocados ser o mesmo número de MVs. Isto evita a concorrência das múltiplas MVs tentando acessar o mesmo volume e a vazão esperada para cada MV permanece a mesma (128 MiB/s) (Tabela 3).

Apesar do S3 possuir o pior desempenho para a maioria dos experimentos, este armazenamento apresentou uma boa escalabilidade e obteve o segundo melhor resultado para 8 MVs. A escalabilidade apresentada pelo S3 (com S3FS-FUSE) está relacionada ao desempenho que o serviço promove com o aumento da quantidade de instâncias (consequentemente, o aumento da largura de banda de rede para acessar o S3) e da quantidade de conexões HTTP. Dessa forma, ao aumentarmos o número de MVs e processos, obtemos um aumento da largura de banda da rede e da quantidade de conexões, respectivamente. Outro fator que contribui para a escalabilidade é diminuição dos tamanhos dos arquivos de *checkpoint* com o aumento do número de processos (Tabela 1), com uma redução de 475% do tamanho do arquivo a ser persistido, quando comparado a 1 MV. A vazão para cada MV variou de 35,3 a 20,5 MiB/s (Tabela 3) e a vazão total (Num. MVs * Vazão Obt. para 1 MV), aumenta com a quantidade de MVs, variando de 35,3 MiB/s a 160,4 MiB/s. Não foi encontrado o desempenho esperado do S3 na documentação da AWS.

A pior escalabilidade apresentada foi a do EFS . Isto ocorre porque o EFS forneceu a mesma vazão de 100 MiB/s para todos os experimentos, devido a Qualidade de Serviço (*Quality of Service* - QoS) imposta pela AWS que restringe a vazão da rede de acordo com o tamanho do sistema de arquivo utilizado. Dessa forma, com o aumento do número de MVs e de processos, aumentou-se a concorrência pela utilização da rede e a vazão esperada diminuiu de acordo com o número de MVs, variando de 100 MiB/s com 1 MV a 12,5 para cada uma das 8 MVs (Tabela 3). Observa-se que, para a maioria dos experimentos, a vazão obtida ficou próxima da esperada, com variação de 91% a 99%.

FSx for Lustre também não apresentou uma boa escalabilidade pois foi utilizada a configuração mínima de tamanho permitida pela AWS (1,2 TiB) e, com isso, somente um disco foi alocado para o sistema de arquivo Lustre (1 OST). Isto aumentou a concorrência dos processos pelo disco e, possivelmente, tornou-se o gargalo. Devido ao FSx for Lustre possuir um único disco, o aumento do número de MVs também implica na divisão da vazão pelo número de MVs, variando de 240 MiB/s para 1 MV a 30 MiB/s para cada uma das 8 MVs (Tabela 3). Observa-se que a eficiência do FSx for Lustre foi pior quando comparada com o EFS; isto ocorre, possivelmente, devido à concorrência dos processos escrevendo em partes diferentes de um único disco do FSx for Lustre enquanto no EFS a AWS garante a vazão de rede com diversos discos em *backend* para persistir os dados.

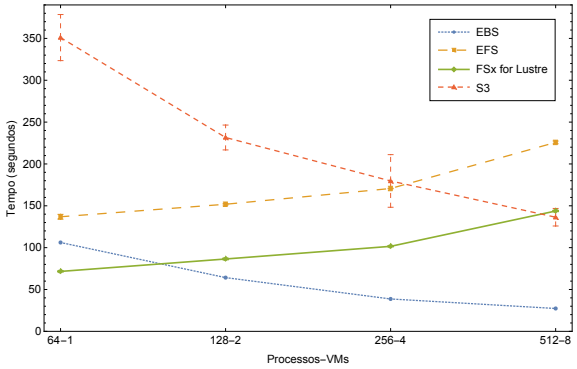

Figura 1. Escalabilidade dos armazenamentos para 1 *checkpoint*

Tabela 2. Razão de desempenho com o EBS para 1 *checkpoint*

Proc.	EBS	EFS	FSx Lustre	S3
64	1,00	1,29	0,67	3,31
128	1,00	2,37	1,35	3,61
256	1,00	4,41	2,63	4,64
512	1,00	8,27	5,27	4,99

Tabela 3. Vazão de escrita por MV - NPB LU.D. Valores de Obt. e Esp. em MiB/s.

MVs	EBS			EFS			FSx for Lustre			S3	
	Obt.	Esp.	Razão	Obt.	Esp.	Razão	Obt.	Esp.	Razão	Obt.	Esp.[a]
1	116,8	128,0	0,91	90,6	100,0	0,91	173,2	240,0	0,72	35,3	-
2	110,4	128,0	0,86	46,7	50,0	0,93	82,0	120,0	0,68	30,6	-
4	108,2	128,0	0,85	24,5	25,0	0,98	41,2	60,0	0,69	23,3	-
8	102,2	128,0	0,80	12,4	12,5	0,99	19,4	30,0	0,65	20,5	-

[a] Não encontrado na documentação da AWS.

4. Conclusão

Este trabalho realizou a análise de 4 serviços de armazenamento na Nuvem Computacional para a realização de *checkpoint* com instâncias de custo reduzido - Spot. Os resultados demonstraram que dois dos 4 serviços de armazenamento são escalados horizontalmente: (a) EBS, devido do aumento da vazão obtida com o incremento da quantidade de volumes EBS (1 volume para cada MV); (b) S3, devido o aumento a quantidade de MVs melhorar a largura de banda total da rede e também o incremento do número de processos adicionar novas conexões com o sistema S3.

Os resultados também ratificaram que o EBS possui os melhores tempos de execução - variação de 29% a 727% - para a maioria dos experimentos realizados. O serviço Lustre não apresentou a vazão esperada, possivelmente, devido à concorrência ocasionada pelos processos de *checkpoint* para utilização de um único disco (OST) e o EFS apresentou a pior escalabilidade devido à vazão esperada (restringida pelo QoS da AWS) ser a menor e por diminuir com o aumento de MVs. Pretende-se, como trabalho futuro, otimizar o desempenho dos serviços de armazenamento aumento-se a quantidade de discos (OST) (FSx for Lustre) e criar um sistema de arquivo para cada MV (FSx for Lustre e EFS).

Agradecimentos

Os autores agradecem a FAPESP (CCES 13/08293-7), o CNPq (140653/2017-1) e a Petrobras pelo apoio financeiro e o Laboratório Multidisciplinar de Computação de Alto Desempenho do Instituto de Computação da Unicamp pelo suporte computacional.

Referências

Cao, J., Kerr, G., Arya, K., and Cooperman, G. (2014). Transparent checkpoint-restart over infiniband. In *Proceedings of the 23rd international symposium on High-performance parallel and distributed computing*, pages 13–24.

Netto, M. A., Calheiros, R. N., Rodrigues, E. R., Cunha, R. L., and Buyya, R. (2018). Hpc cloud for scientific and business applications: Taxonomy, vision, and research challenges. *ACM Computing Surveys (CSUR)*, 51(1):1–29.