Tamanhos de Requisições de E/S de Aplicações HPC em um Supercomputador

Gessica Azevedo[1], Jean Luca Bez[1,3], Pablo Pavan[1], Francieli Boito[2], Philippe Navaux[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]LaBRI, Université de Bordeaux, Inria, CNRS, Bordeaux-INP – Bordeaux, France

[3]Barcelona Supercomputing Center (BSC) – Barcelona, Spain

{gessica.azevedo,jean.bez, pablo.pavan, navaux}@inf.ufrgs.br

Abstract. *This study seeks to identify the most common I/O request sizes used by HPC applications in supercomputers. We use data from an entire year of characterization with the Darshan tool, in the Intrepid Blue Gene/P, to identify access patterns and request sizes. Therefore, we contribute so that new optimization techniques can be evaluated considering these sizes that are similar to those found in these environments.*

Resumo. *Este estudo busca identificar tamanhos de requisições de E/S mais comuns utilizados por aplicações HPC em supercomputadores. Utilizamos dados de um ano inteiro de caracterização com a ferramenta Darshan, no Intrepid Blue Gene/P, para identificar os padrões de acesso e tamanhos de requisições. Assim, contribuímos para que novas técnicas de otimização possam ser avaliadas considerando estes tamanhos semelhantes ao encontrado nestes ambientes.*

1. Introdução

As aplicações científicas que executam em sistemas de computação de alto desempenho (HPC) demandam alta capacidade de armazenamento e eficiência no acesso aos dados. Os sistemas de arquivos paralelos (PFS) atuam fornecendo uma abstração do sistema de armazenamento para estas aplicações, que lidam com grande quantidade de dados, fazendo com que uma centena de nós de computação precisem acessar um sistema de armazenamento compartilhado de forma simultânea. No entanto, dependendo da forma como as aplicações fazem suas requisições de E/S, i.e., seu padrão de acesso, o desempenho pode ser prejudicado. Um exemplo disso é o uso de requisições muito pequenas, que não compensa o custo de acesso ao sistema de armazenamento remoto [Boito et al. 2018].

Para aplicar técnicas de otimização nestas aplicações, primeiro precisamos entender o comportamento de E/S da aplicação. Ferramentas como Darshan [Carns et al. 2009], desenvolvido no Argonne Leadership Computing Facility (ALCF), fornecem uma caracterização criando perfis das aplicações. Com o objetivo de identificar os tamanhos de requisições mais comuns, observados nos padrões de acessos das aplicações em um supercomputador, este estudo analisou os dados do Darshan no Intrepid Blue Gene/P. Os resultados aqui apresentados contribuem para que novas técnicas de

otimização possam ser testadas, utilizando tamanhos de requisições semelhantes aos de um ambiente de grande escala.

O restante do artigo está organizado da seguinte forma. Informações sobre os dados utilizados neste estudo e a metodologia empregada são detalhados na Seção 2. Os resultados e análises são apresentados na Seção 3. A Seção 4 discute trabalhos relacionados. Finalmente, na Seção 5, concluímos este artigo e discutimos trabalhos futuros.

2. Metodologia

Entre 2010 e 2013, o ALCF coletou dados de execução de uma variedade de aplicações científicas usando a ferramenta de caracterização de E/S chamada Darshan, no Intrepid Blue Gene/P, número 23 na lista Top500 de novembro de 2011. Darshan intercepta o fluxo de funções de E/S e registra uma coleção de estatísticas para cada arquivo que é aberto [Carns. 2013]. As informações coletadas permitem identificar tamanhos e padrões de acesso, operações e tempo gasto em operações de E/S.

Essas informações foram usadas em trabalho anterior [Pavan et al. 2019], o qual foi extraído dados relevantes para o estudo. O foco neste estudo foi em dados de 2012 gerado pela versão Darshan 2.0, resultando em 91.603 *jobs*. A taxa de cobertura da aplicação variou entre 20% e 80% por semana [Carns. 2013], pois, o Darshan apenas instrumentou aplicações que chamaram com sucesso `MPI_Init()` e `MPI_Finalize()`, porém, isso não impede que as aplicações utilizem POSIX, pois ambas estas funções são utilizadas apenas para agrupar informações sobre a execução da aplicação. Agrupamos as informações por mês e padrão de acesso, buscando identificar o tamanho de requisição mais comum para cada padrão e seu comportamento ao longo do ano. Para isso observamos os valores mínimos, medianos, máximos e os quartis.

Os 22 padrões de acesso observados ao longo do ano podem ser classificados por operações de escrita/leitura em arquivo único ou compartilhado, escrita/leitura sequencial (o `offset` não precisa ser adjacente e sim, maior que o anterior) e consecutiva (onde o próximo `offset` a ser acessado é imediatamente adjacente ao anterior) e por interfaces, sendo elas POSIX e MPI-IO. Utilizando de estimativas de fases de E/S, obtidas no mesmo trabalho anterior, destes 22 padrões, selecionamos os seis padrões que foram observados durante um período maior entre as aplicações para realizar nossa análise, estes são apresentados na Tabela 1.

3. Resultados e Análises

A Figura 1 ilustra o tamanho médio das requisições para os seis padrões de acesso. O eixo x representa os dias e o eixo y o tamanho médio (em KB). Agrupamos os resultados em meses e pelo padrão de acesso, indicado pelas letras de A-F. A descrição destes padrões está presente na Tabela 1.

Tabela 1. Tamanhos de requisições dos padrões de acesso

	Padrão de Acesso	Min. (Bytes)	Q1 (Bytes)	Mediana (KB)	Média (KB)	Q3 (KB)	Max. (MB)
A	POSIX, Write, Sequential Unique-file	1	99	15,7	3120,3	135	128
B	POSIX, Write, Consecutive Unique-file	1	4	0,003	0,019	0,003	16
C	POSIX, Read, Shared-file	1	160	0,27	93,6	64	256
D	POSIX, Read, Consecutive Unique-file	1	4096	4	12,6	4	16
E	POSIX, Write, Unique-file	1	2184	35,1	580,3	71,8	256
F	MPI-IO, Write, Shared-file	4	4194304	4096	6168,3	8192	122

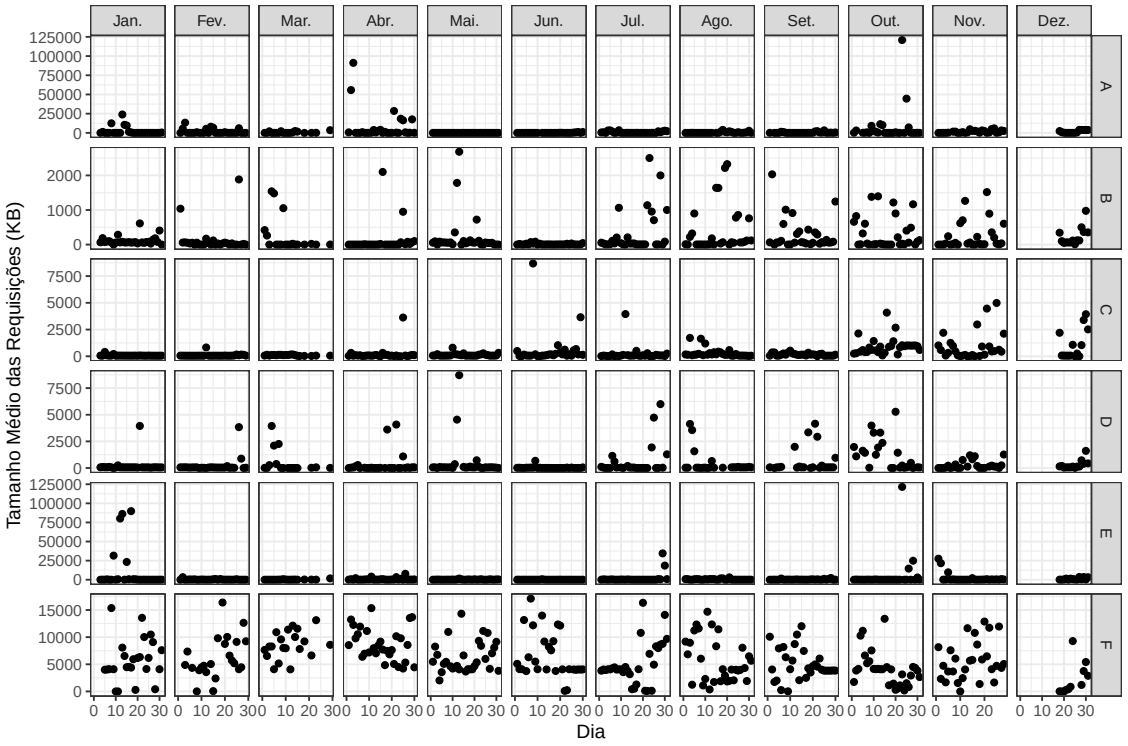

Figura 1. Tamanho médio das requisições (KB) ao decorrer dos dias de um ano. O eixo y é diferente para cada linha.

Para o padrão A, observa-se que o tamanho médio se mantém abaixo dos 1 KB, mas há exceções que podem chegar até 122 MB. O padrão B também tem uma tendência para o tamanho médio até os 1 KB, porém, foi detectada uma maior variação entre os tamanhos, principalmente na segunda metade do ano. Algumas excederam 1 MB, mas a maioria permaneceu no intervalo entre 1 KB e 1 MB. No padrão C também houve uma concentração dos tamanhos na faixa de até 1 KB, apresentando um aumento de variação nos últimos 3 meses do ano. Essas entre 1 KB e 4 MB (com exceção de um valor discrepante que excede os 7 MB). Este comportamento também se repete para os padrões D e E, com apenas a diferença de que para o padrão E, as variações ficam em torno de 24 MB, com alguns casos excedendo 73 MB e um máximo de 122 MB. O padrão F é o mais distinto, pois, registrou maior variabilidade nos tamanhos durante o ano.

Em geral, podemos concluir que destes seis padrões utilizados como amostra, cinco deles se mantém com a tendência do tamanho médio das requisições entre 1 KB e 1 MB. O único padrão que se destaca por sua alta variabilidade para os tamanhos médios é o padrão F, correspondente ao MPI-IO, Write Shared-file, e o motivo para esta flexibilidade deste padrão são as otimizações (como *data sieving* e *collective buffering*) que esta interface disponibiliza.

4. Trabalhos Relacionados

Carns et al. analisou o comportamento de 66 aplicações no supercomputador Interprid (ALCF) [Carns et al. 2011] durante dois meses de 2010. Os autores demonstraram que o tamanho de requisições de escrita mais recorrente era entre 100 KiB (*kibibytes*) e 1 MiB

(*mebibytes*), enquanto para operações de leitura era entre 100 *bytes* e 1 KiB. Percebeu-se que poucas aplicações influenciavam os tamanhos de acesso observados, e se essas aplicações fossem desconsideradas na análise, o tamanho de acesso mais frequente mudaria para 100 KiB e 1 MiB, para ambas as operações.

Apesar de o trabalho anterior também ter utilizado dados sobre a carga de E/S do Interprid, o estudo utilizou um conjunto de dados menor. Nosso estudo investiga o comportamento de E/S ao decorrer de um ano inteiro (2012) e considera o tamanho observado para os diferentes padrões de acesso e não somente pela interface ou operação.

5. Conclusão e Trabalhos Futuros

Este estudo utilizou dados de um ano inteiro de caracterização com o Darshan no supercomputador Intrepid Blue Gene/P. Foi possível determinar os tamanhos de requisições de E/S mais comuns considerando os diferentes padrões de acesso observados.

Para avaliar novas técnicas de otimização (inicialmente utilizando *benchmarks*), é necessário utilizar parâmetros próximos da realidade para que a validação seja consistente. Desta forma, identificar os tamanhos de requisições mais comuns contribui para que os testes mantenham sua confiabilidade, já que estes parâmetros estarão de acordo com a realidade das aplicações HPC. Como trabalhos futuros, pretendemos expandir a análise para os demais padrões observados na máquina ao longo do ano.

Agradecimentos

A pesquisa recebeu financiamento do PIBIC CNPq-UFRGS, da CAPES, concessão N. 001, do CNPq e do projeto Petrobras, concessão N. 2016 / 00133-9. Esta pesquisa utilizou recursos da *Argonne Leadership Computing Facility* no Laboratório Nacional de Argonne, que é apoiado pelo *Office of Science of the U.S. Department of Energy* sob contrato DE-AC02-06CH11357.

Referências

Boito, F. Z., Inacio, E. C., Bez, J., Navaux, P. O. A., Dantas, M. A. R., and Denneulin, Y. (2018). A Checkpoint of Research on Parallel I/O for High-Performance Computing. *In 2018 ACM Computing Surveys (ACM)*.

Carns., P. (2013). ALCF I/O Data Repository. Technical report, Argonne Leadership Computing Facility.

Carns, P., Harms, K., Allcock, W., Bacon, C., Lang, S., Latham, R., and Ross, R. (2011). Understanding and improving computational science storage access through continuous characterization. *In 2011 IEEE 27th Symposium on Mass Storage Systems and Technologies*, page 7(3):8:1–8:26.

Carns, P., Latham, R., Ross, R., K. Iskra, S. L., and Riley., K. (2009). 24/7 Characterization of petascale I/O workloads. *In 2009 IEEE International Conference on Cluster Computing and Workshops*, pages 1–10.

Pavan, P. J., Bez, J., Serpa, M. S., Boito, F. Z., and Navaux, P. O. A. (2019). An Unsupervised Learning Approach for I/O Behavior Characterization. *In 2019 31st International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*.