Avaliação de desempenho do OLAM com o PVFS2

Francieli Zanon Boito[1], Rodrigo Kassick[1], Laércio Pilla[1], Philippe O. A. Navaux[1], Claudio Schepke[1], Nicolas Maillard[1], Carla Osthoff[2], Pablo Grunmann[2], Pedro Dias[2], Jairo Panetta[3]

[1]Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brasil

[2]Laboratório Nacional de Computação Científica (LNCC), Petrópolis, RJ, Brasil

[3]Instituto Nacional de Pesquisas Espaciais (INPE), Cachoeira Paulista, SP, Brasil

1. Introdução

Modelos de previsão de tempo e clima são grandes usuários de computação de alto desempenho. Essas aplicações paralelas manipulam grandes quantidades de dados, acessados e gerados por todos os nós envolvidos na computação. Assim, o desempenho das operações de entrada e saída é crucial para a viabilidade dessas execuções. Esse artigo apresenta uma avaliação do modelo OLAM com o sistema de arquivos paralelo PVFS2.

2. *Ocean-Land-Atmosphere Model* e *Parallel Virtual File System*

O *Ocean-Land-Atmosphere Model* (OLAM) [Walko and Avissar 2008] é um modelo criado na *Duke University* em FORTRAN 90 e paralelizado com MPI. O globo é representado por uma malha de icosaedros, e a execução do programa se dá iterativamente: após uma fase de inicialização, em que alguns arquivos com dados de entrada são lidos por todos os processos envolvidos, há uma sequência de *timesteps*. Cada *timestep* evolui o estado da atmosfera em uma quantidade configurável de tempo. A cada número predeterminado de avanços, cada processo cria e escreve em um *history file*. Quanto mais processos executando, menores serão os arquivos gerados, mas maior será o seu número.

O *Parallel Virtual File System* (PVFS2) [Carns et al. 2000] é um sistema de arquivos paralelo de alto desempenho de grande visibilidade mantido por diversas universidades e laboratórios nacionais norte-americanos. O papel de servidor é desempenhado por diversos servidores de dados e de metadados, que podem ser executados na mesma máquina.

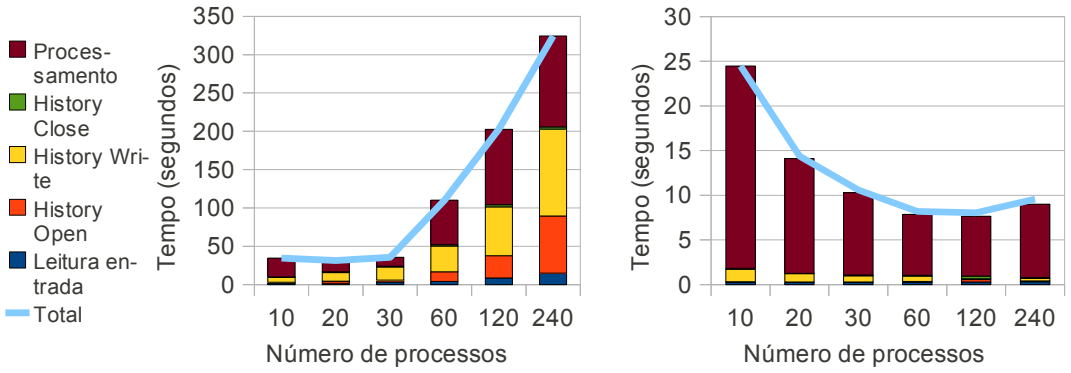
3. Resultados

A versão 3.3 do OLAM foi instrumentada para fornecer os tempos gastos em cada operação de entrada e saída. A versão 2.8.2 do PVFS2 foi obtida da página do projeto[1], e o sistema foi configurado com 4 máquinas atuando como servidores de dados e de metadados e 30 máquinas clientes acessando o sistema de arquivos através do módulo para *kernel* do Linux. Os testes foram executados no *cluster* Griffon do Grid5000[2] (Nancy), cujos nós são equipados com dois processadores Intel Xeon *quad-core* e interconectados por Gigabit Ethernet.

O gráfico da Figura 1 mostra os tempos obtidos usando o PVFS2 como meio de armazenamento e os discos locais dos nós envolvidos. O uso do sistema de arquivos

[1]http://www.pvfs.org/

[2]http://www.grid5000.fr/



(a) PVFS2 (b) Discos locais

Figura 1. Tempos de execução do OLAM com até 240 processos MPI em 30 máquinas clientes usando o PVFS2 e os discos locais aos nós

distribuído se mostrou até 36 vezes pior do que o uso do armazenamento local, não apresentando *speedup*. Quanto maior o número de processos, maior ficou o tempo necessário para efetuar as operações de entrada e saída, fenômeno não observado nos discos locais. Isso ocorre porque se sobrecarregou o sistema de arquivos, especialmente o gerenciamento de metadados.

Além disso, aumentando o paralelismo, o tempo gasto em processamento acaba aumentando, mesmo enquanto diminui a carga a ser executada por cada processo. Isso ocorre possivelmente porque o sistema de arquivos coloca em *buffers* as pequenas requisições, concorrendo pelo uso da rede e da memória com as etapas de processamento e com as trocas de mensagens entre os processos MPI.

4. Conclusão e Trabalhos Futuros

Esse artigo apresentou um estudo de desempenho do OLAM com o sistema de arquivos PVFS2. Em um cenário de 30 nós, o uso do sistema de arquivos se mostrou até 36 vezes pior que os discos locais aos nós. No entanto, é importante observar que o uso de armazenamento local impõe a tarefa de copiar os dados de entrada para todos os nós antes da execução e de obter todos os resultados desses nós ao final. Além disso, a possibilidade de escolher essa abordagem está condicionada à disponibilidade e presença de armazenamento nas máquinas do *cluster* utilizado.

Para a continuidade do trabalho, planeja-se o estudo de outras otimizações que possam beneficiar o desempenho de E/S e saída do OLAM, como escalonamento de operações de entrada e saída.

Referências

Carns, P. H., Walter B. Ligon, I. I. I., Ross, R. B., and Thakur, R. (2000). Pvfs: a parallel file system for linux clusters. In *Proceedings of the 4th conference on 4th Annual Linux Showcase and Conference (ALS'00)*, pages 28–28, Berkeley, CA, USA. USENIX Association.

Walko, R. and Avissar, R. (2008). The Ocean–Land–Atmosphere Model (OLAM). Part I: Shallow-Water Tests. *Monthly Weather Review*, 136:4033–4044.