Metodologia para Avaliação de Equipamentos de Computação de Alto Desempenho: Um Estudo de Caso para Bioinformática

Mariza Ferro[1], **Marisa F. Nicolas**[2], **Antonio R. Mury**[1], **Bruno Schulze**[1]

[1]Computação Científica Distribuída (ComCiDis)
Laboratório Nacional de Computação Científica (LNCC)

[2]Laboratório de Bioinformática (LabInfo)
Laboratório Nacional de Computação Científica (LNCC)
Petrópolis – RJ – Brasil
{mariza,marisa,aroberto,schulze}@lncc.br

Abstract. *High Performance Computing is essential to boost scientific progress in many areas of science and to solve a number of complex scientific applications. These applications have different characteristics that require distinct computational resources too. However, there isn't any formal methodology to assist researchers in the proper definition of such equipment. In this work we propose a systematic methodology which allows the evaluation, acquisition and maintenance of high performance computing systems meeting the requirements of scientific applications. For validation of the methodology is proposed a case study for bioinformatics.*

Resumo. *Um grande número de grupos de pesquisa dependem de recursos computacionais de alto desempenho para viabilizar a execução de suas aplicações científicas. Porém, as características dessas aplicações podem ser muito diferentes, exigindo recursos computacionais muito distintos. Entretanto, não existe uma metodologia formal para auxiliar pesquisadores a determinar a infraestrutura computacional adequada ao seu conjunto de aplicações. Este trabalho propõe uma metodologia que possibilite a avaliação, a aquisição e a manutenção de sistemas de computação de alto desempenho, que cumpram os requisitos das aplicações científicas e para validação da metodologia é proposto um estudo de caso para bioinformática.*

1. Introdução

Diferentes áreas de pesquisa dependem de recursos computacionais de alto desempenho (HPC) para viabilizar novas descobertas e alavancar seu progresso científico. Entre essas áreas está a bioinformática, na qual novas estratégias de sequenciamento de DNA de alto desempenho permitem sequenciar e analisar milhares de fragmentos de nucleotídeos em paralelo. Devido a estes avanços são gerados cada vez mais dados, em menor tempo e custo, tornando os projetos de genômica e transcritômica um grande desafio para as análises de bioinformática. Com isso, manter o sucesso da bioinformática no futuro depende do uso dos recursos de HPC. Porém, adquirir, manter e realmente alcançar alto desempenho na execução dos conjuntos de aplicações científicas não é tarefa trivial. Para a obtenção do melhor desempenho na execução das aplicações não basta adotar estratégias simplistas, tal como adquirir a infraestrutura de maneira empírica, baseando-se na classificação do Top500 [1] ou apenas maximizando o número de nós. Esse tipo de

[1]Classificação dos 500 supercomputadores mais poderosos do mundo - http://www.top500.org/

estratégia, em alguns casos, pode até ser prejudicial ao desempenho de certas aplicações pois, enquanto algumas aplicações funcionam bem com qualquer tipo de arquitetura, outras operam melhor com uma configuração particular. Isso porque as aplicações científicas têm diferentes requisitos computacionais, os quais dependem da definição adequada dos sistemas de HPC envolvidos para se obter eficácia e eficiência na resolução dos seus problemas. A proposta deste trabalho é que, para a obtenção do melhor desempenho dos recursos disponíveis, um dos principais requisitos é entender as características das aplicações científicas, seu relacionamento com o tipo de arquiteturas computacionais sobre o qual serão executados e ser capaz de mensurar o impacto deste relacionamento, tornando assim possível entender quais características das arquiteturas podem comprometer o desempenho. Para isso é proposta uma metodologia de avaliação para auxiliar profissionais e pesquisadores, mesmo leigos em sistemas de HPC, a realizarem a aquisição e a manutenção dos seus sistemas, obtendo melhor desempenho, menor custo e principalmente, que cumpram os requisitos operacionais das suas aplicações científicas.

2. Aplicações Científicas e Avaliação de Desempenho

Um grande número de aplicações científicas têm sido implementadas para execução em equipamentos computacionais de alto desempenho. Na bioinformática não é diferente e suas aplicações possuem requisitos para sua execução distintos da maioria das aplicações de outros domínios científicos. Existem algumas propriedades dominantes nestas aplicações, e é necessário entendê-las para determinar quais os recursos necessários para executá-las de maneira eficiente. A exigência de recursos tais como, tempo de CPU, quantidade de memória, largura de banda, entre outros, são informações importantes, pois permitem identificar quais aspectos da configuração são limitantes de desempenho para uma aplicação. Apesar desse tipo de informação estar disponível para os usuários, as medidas apresentadas são baseadas em análises distantes das necessidades dos usuários, os quais desconhecem se o equipamento de HPC em uso, ou que se pretende adquirir, é a melhor opção para maximizar a eficiência das suas aplicações. A maneira tradicional para se realizar as avaliações de desempenho desses equipamentos é por meio do uso de *benchmarks*. Porém, a maioria das suites de *benchmarks* comparam uma arquiteturas com a outra sob um único foco; nesse caso o que os *benchmarks* fazem são entregar um número relativo ao desempenho que unicamente caracteriza um sistema/arquitetura em relação a outro. Essa avaliação é feita sem levar em consideração requisitos das aplicações e o papel da arquitetura para o desempenho da mesma. Por exemplo, o Linpack é o *benchmark* mais amplamente utilizado e também é métrica para classificar sistemas de HPC no Top500. Entretanto, segundo [Heroux and Dongarra 2013] "*o Linpack está cada vez menos confiável como a única métrica de desempenho para uma coleção cada vez maior de importantes aplicações cientificas e de engenharia*". O Linpack favorece arquiteturas com altas taxas de processamento em ponto-flutuante. Porém, segundo [Albayraktaroglu et al. 2005], para grande parte das aplicações de bioinformática a importância das operações de ponto flutuante (uso intensivo de CPU) são relativamente baixas em relação a outras operações, tais como comparações de strings e buscas em disco. Além disso, ao contrário do que se imagina, que quanto mais recursos melhor, adotar essa estratégia simplista muitas vezes pode até ser prejudicial ao desempenho, como demonstrado no trabalho de [Cypher et al. 1993] onde foram estudados os comportamentos de oito aplicações científicas com características bastante heterogêneas. Neste trabalho é proposto o uso de classes de aplicações para esse tipo de avaliação, onde cada

classe representa um padrão comum de requisito por parte da aplicação, em termos de processamento e comunicação. Assim, é possível conhecer os fatores limitantes de desempenho para um determinado grupo de aplicações. Com base nisso, está em desenvolvimento uma metodologia sistemática para avaliação de sistemas de HPC, com foco no bom desempenho das aplicações científicas. O desenvolvimento da metodologia proposta neste trabalho, tem como referência a Análise Operacional (AO) e é brevemente apresentada a seguir.

3. Metodologia Proposta e Estudo de Caso

No momento da aquisição de um novo sistema de HPC, a tomada de decisão sobre qual é o mais adequado para atender a um determinado grupo de aplicações é uma decisão que exige conhecimento sobre o hardware, as aplicações e a interação entre eles. A complexidade do hardware somada ao crescente desenvolvimento tecnológico, e particularmente dos sistemas de HPC para os quais novas arquiteturas são lançadas a cada seis meses, torna essa tomada de decisão de alta complexidade. Nesse caso, contar apenas com a experiências ou com o uso de *benchmarks*, não é adequado. Para isso está em desenvolvimento uma metodologia que auxilie pesquisadores e técnicos a determinarem a infraestrutura computacional adequada, com foco no seu conjunto de aplicações científicas. Essa metodologia usa como referência a AO, a qual consiste num conjunto de procedimentos necessários para fornecer subsídios, que possam auxiliar no processo de tomada de decisões quanto à obtenção, ao emprego e às modificações de um sistema avaliado. Por meio da AO é possível estimar a capacidade do sistema de cumprir efetivamente a função para o qual foi adquirido [Wagner et al. 1999]. Com a aplicação da metodologia será possível identificar e categorizar as aplicações em classes e entender seus comportamentos em diferentes arquiteturas (*multi-core* e *manycore*), mapear cada classe de aplicação para aplicações representativas (*benchmarks*, *kernels*, etc), criando assim um conjunto de testes que medem os parâmetros realmente relevantes ao conjunto de aplicações do usuário; ao contrário do simples uso das medidas obtidas pelos *benchmarcks* utilizados pelos fabricantes, os quais apenas fornecem os picos teóricos do equipamento, sem levar em consideração as aplicações para o qual será utilizado. Além disso, definir Medidas de Eficácia Operacional (MEO) por meio da qual são definidos os parâmetros significativos para cada classe de aplicação e seus valores de referência para atender às necessidades do usuário; por exemplo, linguagem de programação pode ser um parâmetro relevante, caso o usuário não queira migrar seus códigos de aplicações para se adequar a uma arquitetura específica pois, muitas arquiteturas tem baixo desempenho se otimizações e migração de linguagem de programação não são realizadas. Com a execução dos testes é possível avaliar o desempenho do conjunto de aplicações em um determinado equipamento e permite ao pesquisador solicitar aos fornecedores dos sistemas de HPC, que executem o seu conjunto de testes e requisitar quais parâmetros devem ser medidos durante sua execução. Com isso, o usuário consegue auditar os dados coletados, e não ficar mais "refém" dos testes executados pelos fabricantes. Outra contribuição obtida com o uso da metologia é avaliar a adequabilidade operacional do sistema, não apenas durante a sua aquisição, mas para a manutenção periódica desses sistemas, evitando sua degradação, com o consequente aumento da vida útil da infraestrutura já existente. Além disso, os resultados obtidos com o uso da metologia fornecerão subsídios para os grupos de pesquisa junto as agências de fomento no momento da aquisição de um novo equipamento. Um estudo de caso para avaliação da metodologia proposta está sendo realizado junto ao projeto

que abrange a utilização de métodos de bioinformática, modelagem molecular e transcritômica para a detecção e priorização de desenvolvimento de novos fármacos para o controle de doenças causadas pelas bactérias *Klebsiella pneumoniae* (infecções hospitalares) e *Mycobacterium tuberculosis* (tuberculose). O presente projeto possui relevância para as populações de ambos os países participes, Brasil e Argentina, uma vez que todos enfrentam, na prática clínica, surtos de infecções hospitalares causados por estirpes de *K. pneumoniae* além de sermos, também, países onde a tuberculose, causada pelo bacilo *M. tuberculosis*, ainda representa uma ameaça à saúde pública a despeito das políticas públicas de vacinação implementadas. Assim, para o desenvolvimento do projeto, a definição adequada dos recursos de HPC é de grande relevância, permitindo acelerar a obtenção dos resultados. O estudo de caso da metodologia proposta envolve o *workflow* científico para o sequenciamento do transcritôma da bactéria *K. pneumoniae*. Todas as aplicações científicas envolvidas nesse *workflow* estão sendo identificadas e classificadas, bem como a infraestrutura computacional envolvida no projeto. Ao final deste estudo MEOs para esse grupo de aplicações científicas serão definidas, a arquitetura atual será avaliada quanto a sua adequabilidade operacional e se necessário uma nova arquitetura será proposta. Além disso, a metodologia será avaliada quanto a sua eficiência em avaliar o sistema e a sua facilidade de uso pelos pesquisadores.

4. Considerações Finais

A principal contribuição deste trabalho é o de desenvolver uma metodologia para aquisição e manutenção de sistemas de HPC e a criação de um conjunto de métricas para avaliação do desempenho das aplicações científicas nessas arquiteturas de HPC. Esse tipo de abordagem, baseada nos requisitos das aplicações científicas, ainda não é utilizada no Brasil, e a percepção da sua importância ainda hoje está restrita a um pequeno grupo de laboratórios, que no entanto, são aqueles com o maior desenvolvimento tecnológico e maior capacidade computacional existente. Assim, o desenvolvimento de uma metodologia e suas métricas, será de grande contribuição aos pesquisadores e usuários de sistemas de HPC.

Referências

Albayraktaroglu, K., Jaleel, A., Wu, X., Franklin, M., Jacob, B., Tseng, C.-W., and Yeung, D. (2005). Biobench: A benchmark suite of bioinformatics applications. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software, 2005*, ISPASS '05, pages 2–9, Washington, DC, USA. IEEE Computer Society.

Cypher, R., Ho, A., Konstantinidou, S., and Messina, P. (1993). Architectural requirements of parallel scientific applications with explicit communication. *SIGARCH Comput. Archit. News*, 21(2):2–13.

Heroux, M. A. and Dongarra, J. (2013). Toward a New Metric for Ranking High Performance Computing Systems. *SAND2013 - 4744*.

Wagner, D., Mylander, W., and Sanders, T. (1999). *Naval Operations Analysis*. Naval Institute Press, Annapolis, Maryland, USA, 3rd edition.