Exploração de Módulos Paralelo Híbrido de Bioinformática para Ambientes GPU de Supercomputação

Guilherme Freire[1,2], **Kary Ocaña**[2], **Micaella Coelho**[2], **Carla Osthoff**[2]

[1]Faculdade de Educação Tecnológica do Estado do Rio de Janeiro (FAETERJ)
Petrópolis – RJ – Brasil

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

{gfreire, karyann, micaella, osthoff}@lncc.br

Abstract. *Bayesian phylogenetic algorithms are computationally intensive. BEAST Bayesian software coupled to BEAGLE 3 high-performance library had been tested on the hybrid resources of the Santos Dumont supercomputer. For analysing large Dengue virus data sets, the use of one or more GPUs proved more efficient than using multi-core.*

Resumo. *Algoritmos filogenéticos Bayesianos são computacionalmente intensivos. O software de inferência Bayesiana BEAST acoplado à biblioteca de alto desempenho BEAGLE 3 foi testado nos recursos híbridos do supercomputador Santos Dumont. Para analisar grandes conjuntos de dados do vírus da dengue, o uso de uma ou mais GPU provou ser mais eficiente do que o uso de multi-core.*

1. Introdução

Plataformas de Computação de Alto Desempenho (CAD) como *clusters*, grades, nuvens ou supercomputadores [Yin et al. 2017] permitem executar de uma maneira eficiente tarefas com grande demanda de processamento. O supercomputador Santos Dumont (SDumont) possui uma arquitetura híbrida com CPU *multi-core* e dispositivos com a chamada arquitetura *many-core*, GPU e MIC. SDumont é usado de forma intensiva por vários grupos de pesquisa brasileiros das mais diversificadas áreas científicas. Fruto da colaboração entre os centros de pesquisa CENAPAD e LABINFO pertencentes ao LNCC são realizadas diversas pesquisas no apoio às análises computacionais envolvendo bioinformática, biologia computacional e análises de CAD.

Mais especificamente, o presente trabalho visa levantar um estudo de desempenho de aplicações de filogenômica e evolução molecular computacional em multi-GPU. Dessa maneira, pretende-se reforçar e estender as pesquisas do artigo prévio publicado no *workshop* BreSci 2020 [Ocaña et al. 2020] que apresenta uma análise comparativa do *software* BEAST/BEAGLE [Jin and Bakos 2013] executado em 1 nó com CPU *multi-core* e 1 GPU do SDumont e usando dados biológicos de diferente natureza. No presente trabalho apresentamos resultados de desempenho escalando BEAGLE/BEAST até 8 GPU, os quais se mostram mais eficientes a àqueles apresentados em [Ocaña et al. 2020].

O uso de um ou mais GPU pode se mostrar mais eficiente do que o uso de CPU *multi-core* para analisar grandes conjuntos de dados de nucleotídeos. Para dados de entradas com alinhamento de várias partições como no vírus Dengue é altamente recomendável dividir os cálculos de probabilidade em vários dispositivos GPU.

2. Aplicação BEAST 1.8 com BEAGLE no SDumont

BEAST é uma aplicação de análise filogenômica baseada em inferência Bayesiana. A biblioteca BEAGLE 3 [Jin and Bakos 2013] torna mais eficiente a paralelização em escala fina de cálculos métodos de Markov chain Monte Carlo (MCMC) realizados pelo BEAST.

BEAST 1.8, BEAGLE 3 e as respectivas bibliotecas foram instaladas no ambiente do SDumont. Foram levantados cenários, para as diferentes chamadas CPU, GPU e multi-CPU/GPU acoplando a devida parametrização do BEAST e BEAGLE e levando em consideracão as características e natureza dos dados. Os dados de entrada foram arquivos de sequências biológicas do vírus da Dengue no formato XML.

3. Resultados

Esta seção apresenta detalhes sobre configuração do experimento, ambiente computacional e análises de desempenho.

3.1. Ambiente computacional

O SDumont possui 36.472 núcleos de CPU, distribuídos em 1.134 nós computacionais, na sua maioria CPU com arquitetura *multi-core*. O SDumont possui um nó diferenciado, o MESCA2, com um total de 240 núcleos. No presente experimento foi utilizado um nó computacional do SDumont composto por 8 GPU *Nvidia* V100 com 16 GB e *NVLink* e 2 CPU *Skylake GOLD* 6148, que somam no total 40 núcleos com 384 GB DDR4 RAM.

3.2. Configuração do experimento e análise

Dados moleculares do vírus da Dengue em formato XML foram extraídos do diretório *benchmark* do BEAST 1.8 e usados nas análises [Yin et al. 2017]. O parâmetro *chainLength* do BEAST 1.8 está relacionado ao cálculo das cadeias MCMC. O incremento no valor do *chainLength* fornece consistência às análises bayesianas, mostrando-se proporcional ao incremento no tempo computacional. As execuções foram realizadas usando CPU 24 *threads*, CPU 40 *threads*, 1 GPU, CPU 40 *threads*/1 GPU, CPU 40 *threads*/8 GPU, e 8 GPU.

O melhor desempenho foi obtido usando 8 GPU. As execuções usando um valor baixo no parâmetro *chainLength* = "100000" é exploratório e requer menos gasto computacional. Quando usado um valor maior de *chainLength* = "10000000" tende a gerar uma maior consistência nos resultados, os cálculos de probabilidade requeridos tornam-se mais exaustivos levando a um gasto computacional maior. Nos nossos resultados, o incremento do *chainLength* influencia no maior tempo computacional obtido, mesmo assim ambas as execuções de *chainLength* apresentaram um comportamento muito similar no uso dos recursos CPU e GPU.

A Figura 1 apresenta os resultados de desempenho do BEAST/BEAGLE em tempo total de execução (TTE) em minutos. Para os cálculos foi usado como variabilidade o parâmetro *chainLength* fixado em "100000" e "10000000". Os experimentos sugerem que as características como tamanho dos dados e configuração de parâmetros no BEAST, como o *chainLength*, influenciam no tempo computacional. As execuções realizadas fixando o uso de recursos em 8 GPU apresentam melhor desempenho quando comparadas ao demais ambientes *multi-core* e 1 GPU, como apresentados na Figura 1.

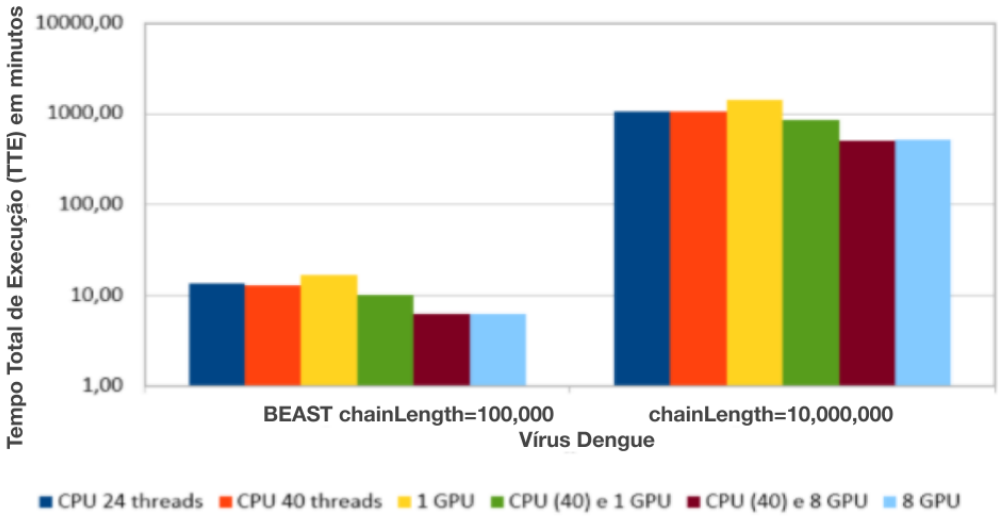

Figura 1. Análise de desempenho do BEAST/BEAGLE no CPU/GPU do SDumont

4. Conclusão

O presente estudo viabiliza a exploração e análise de desempenho do BEAST/BEAGLE em ambientes de CAD com a especificação do ambiente computacional que leve a um desempenho mais eficiente. Dessa maneira, permite que usuários possam usufruir dessas informações e realizar execuções garantindo um uso racional do ambiente do SDumont.

Análises de desempenho do BEAST/BEAGLE em múltiplas configurações de CPU/GPU no SDumont sugerem o uso da configuração híbrida CPU 40 *threads* e 8 GPU como a mais eficiente. Sobre a variabilidade no número de *chainLength* fixados em "100000" e "10000000" como esperado o tempo computacional gasto na execução é incremental, mas o comportamento é muito similar.

Como trabalhos futuros pretendemo realizar análise de desempenho usando vários nós em paralelo, focando no uso de multi-GPU. Estes passos requerem que outras versões do BEAST sejam exploradas, em especial aquela que apresente consistência com a chamada em milti-GPU em multi-nós.

BEAST/BEAGLE está integrado ao Portal-Bioinfo (https://bioinfo.lncc.br/) como uma aplicação de bioinformática. Os resultados apresentados apoiam o uso de portais científicos verdes e mais eficientes.

Referências

Jin, Z. and Bakos, J. D. (2013). Extending the beagle library to a multi fpga platform. *BMC Bioinformatics*, 14(1):25.

Ocaña, K., Coelho, M., Freire, G., and Osthoff, C. (2020). High-performance computing of beast/beagle in bayesian phylogenetics using sdumont hybrid resources. In *14º BreSci – Brazilian e-Science Workshop*. (Aceito em processo de publicação).

Yin, Z., Lan, H., Tan, G., Lu, M., Vasilakos, A. V., and Liu, W. (2017). Computing platforms for big biological data analytics: Perspectives and challenges. *Computational and Structural Biotechnology Journal*, 15:403–411.