Análise Computacional de uma Ferramenta de Bioinformática em Arquiteturas de Memória Compartilhada do Santos Dumont Usando Perfilador Intel VTune.

Albert Emidio[1] [2], Reiglan Soares[1] [2], Lucas Cruz[1], Kary Ocaña[1], DiegoCarvalho[3], Carla Osthoff[1]

[1]Laboratório Nacional de Computação Científica (LNCC), RJ – Brasil

[2]Faculdade de Educação Tecnológica do Estado do Rio de Janeiro (FAETERJ-RJ), RJ – Brasil

[3]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET), RJ – Brasil

{albert,reiglan,lucruz,karyann,osthoff}@lncc.br,d.carvalho@ieee.org

Abstract. *The article presents an analysis of the computational performance of the most resource-intensive alignment stage in the Parsl-RNA Seq workflow, using the Intel VTune profiler on the Santos Dumont supercomputer. The objective of the work is to analyze the computational performance of the Bowtie2 software, which is responsible for sequence alignment and is the most computationally demanding task. Bowtie2 was executed with support from the Parsl library and monitored by VTune, aiming to evaluate its efficiency and scalability in high-performance and distributed environments.*

Resumo. *O artigo apresenta uma análise do desempenho computacional da etapa de alinhamento mais custosa do workflow Parsl-RNA Seq, utilizando o perfilador Intel VTune no supercomputador Santos Dumont. O objetivo do trabalho é analisar o desempenho computacional do software Bowtie2, responsável pelo alinhamento de sequências que é a que mais demanda recursos computacionais. O Bowtie2 foi executado com suporte da biblioteca Parsl e monitorado pelo VTune, com o objetivo de avaliar sua eficácia e escalabilidade em ambientes de computação de alto desempenho.*

1. Introdução

Os experimentos em bioinformática abordam uma grande quantidade de dados complexos, desde sequências genômicas até interações moleculares. Devido a essa complexidade, são necessárias soluções eficientes para o tratamento desses dados, o que envolve não apenas recursos computacionais, como a computação de alto desempenho (CAD), mas também a modelagem de *workflows*. [Cruz et al. 2020].

O *workflow* de transcriptômica ParslRNA-Seq foi utilizado para a execução, gerência e análise computacional. ParslRNA-Seq foi modelado na análise de experimentos transcriptômica, na expressão diferencial de genes (EDG), com a linguagem de programação Python e a biblioteca Parsl em um ambiente de computação de alto desempenho (CAD), o que facilitou a integração e automação do *workflow*. A principal atividade executada nesse experimento foi o Bowtie2, escolhido para fazer análise de desempenho devido ao seu elevado custo de CPU e suporte a *multithreading*. As demais atividades do *workflow* incluem: Sort que ordena as leitura dos genes; Split que divide os arquivos de entrada; HTSeq que conta as leituras; Merge, para indexar as contagens; e DESeq, para a análise estatística das EDGs. [Cruz et al. 2021]

Bowtie2 é a primeira atividade do *workflow* e um *software* de alinhamento eficiente utilizado para mapear sequências de DNA curtas em grandes genomas de referência. Ele usa métodos baseados em indexação e alinhamento para garantir eficiência em termos de tempo e memória. O perfilador Intel VTune é uma ferramenta de análise de desempenho que detalha o uso da CPU, utilizado para otimizar *software* em sistemas Intel. O objetivo da pesquisa é analisar o uso dos recursos computacionais na execução da atividade Bowtie2 já que ele é a atividade mais custosa do *workflow*, assim, observando a utilização da CPU pelo perfilador.

2. Metodologia

Os dados coletados para a análise são de um experimento real de RNA-Seq, extraídos do repositório público Gene Expression Omnibus, divididos em grupo de controle (SRR5445794, SRR5445795, SRR5445796) e grupo de condições das vias metabólica Wnt (SRR5445797, SRR5445798, SRR5445799), sendo o organismo em evidência é o *Mus musculus* e o GEO.ID GSE97763. Os dados de entrada apresentam tamanho entre 1.8GB e 3.0GB, ao todo sendo 13GB, a saída do Bowtie2 também ficou em 13GB . O *workflow* foi executado com os dados de entrada e a atividade Bowtie2, foi submetida ao perfilador para analisar os recursos computacionais utilizados já que ela é a atividade que mais demanda uso da CPU

Para realizar o experimento foram utilizados os nós computacionais 2x CPUs Intel Xeon Cascade Lake Gold 6252 (48 núcleos) e com 384 GB de RAM, denominado Cascade Lake e 2 CPUs Intel Xeon E5-2695v2 Ivy Bridge, 24 núcleos (12 por CPU) de 64 GB de memória RAM, denominado Ivy Bridge.

3. Resultados e Discussão

3.1 Análise do *Workflow* no Perfilador Vtune

A Figura 1 apresenta o histograma das atividades do *workflow* Parsl RNA-Seq, de seis atividades, que foi executado no nó computacional Ivy Bridge. Este experimento envolve a execução de duas atividades em *multithreading* executadas em paralelo (Bowtie e Sort) com 24 *threads*, enquanto as demais atividades são realizadas de forma sequencial, processando um arquivo por núcleo. O gráfico mostra o uso da CPU utilizado ao executar o workflow completo com 6 atividades. Já em contrapartida, as Figuras 2 e 3 apresentam o uso da atividade Bowtie2 em diferentes nós computacionais.

Observa-se um uso intensivo da CPU no nó Ivy Bridge, em que seis tarefas ativam simultaneamente 24 *threads*, maximizando a capacidade de processamento do nó e resultando na menor duração do experimento. As atividades são iniciadas a partir de *tasks* e, ao entrarem em uma etapa que suporta *multithreading*, alocam o número de *threads* permitido pelos núcleos do nó computacional. A seguir iremos apresentar uma análise da atividade bowtie de forma a comprovar que a utilização dos 24 cores da Figura 1 é relativa a aplicação Bowtie2.

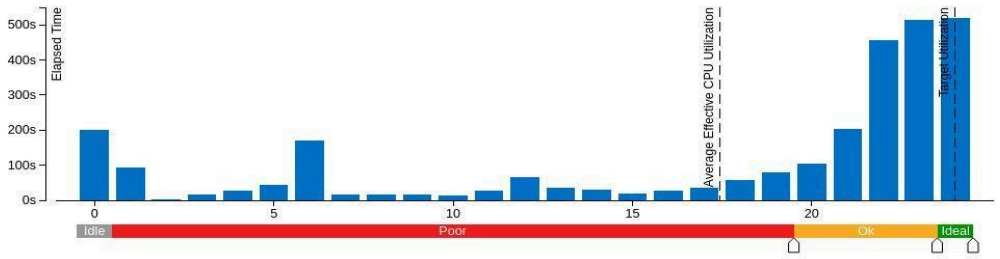

Figura 1. Uso da CPU no Ivy Bridge 24 *cores*

3.2 Análise da Atividade Bowtie2 no Perfilador Vtune

A distribuição do consumo de CPUs ao longo de processo mostra que ambos o nó computacional Ivy Bridge(Figura 2) e nó computacional Cascate Lake (Figura 3) apresentam uma alta utilização do número máximo de cores do nó computacional durante toda a execução da aplicação. A Figura 2 tem como média o uso da CPU a partir das 20 *threads* e isso quer dizer que todos os trabalhos usaram mais que essa quantidade para executar a atividade. Já na Figura 3 essa média se dá a partir das 35 *threads.*

A Figura 2 utilizou de forma eficiente e distribuída seus 24 *cores,* sem ociosidade, mostrando desempenho superior a partir do uso de 20 cores; porém, devido à sua tecnologia de memória mais antiga, apresentou tarefas mais lentas, resultando em tempos de execução maiores. A Figura 3 se destacou pelo uso intensivo de processamento, mostrando desempenho superior a partir do uso de 40 cores, com tempos de execução mais rápidos. A duas arquiteturas se destacam pelo motivo da sua média de uso da CPU estar localizadas em um alto nível de uso, já que todas as threads estão sendo usadas para executar a atividade.

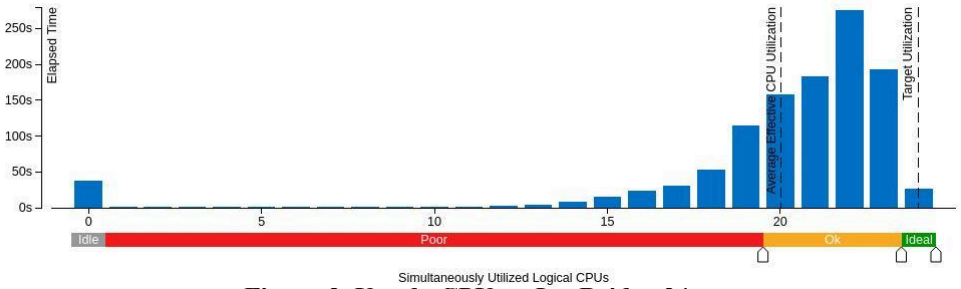

Figura 2. Uso da CPU no Ivy Bridge 24 *cores*

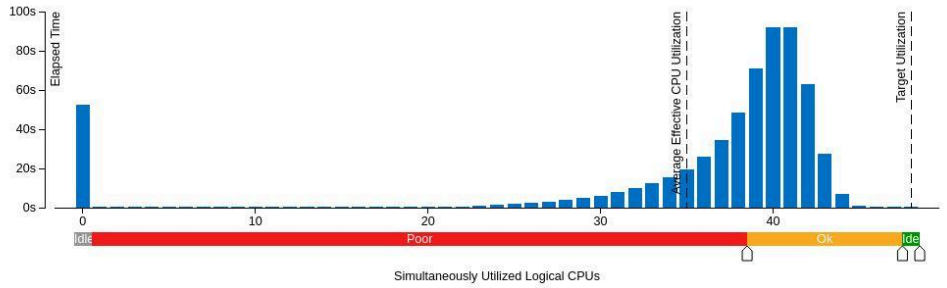

Figura 3. Uso da CPU no Cascade Lake com 48 *cores*

4. Conclusão

O Intel VTune permitiu comprovar a eficiência da paralelização do código Bowtie2 em ambas arquiteturas Ivy Bridge e Cascate Lake, mostrando o uso do número máximo de *cores* em grande parte do tempo em ambas as arquiteturas. Ao analisar o histograma das duas arquiteturas, é possível observar uma semelhança em termos de comportamento, com uma tendência de crescimento, confirmando a alta utilização dos recursos de *hardware* e da eficiência do sistema durante a execução de tarefas. Esse padrão de utilização demonstrou o bom desempenho do *software* Bowtie2, já que a maximização do uso dos núcleos contribui diretamente para a redução do tempo de processamento.

5. Referências

L. Cruz et al." Workflows Científicos de RNA-Seq em Ambientes Distribuídos de Alto Desempenho: Otimização de Desempenho e Análises de Dados de Expressão Diferencial de Genes", in Anais do XV Brazilian e-Science Workshop, 2021, pp. 57-64, doi: https://doi.org/10.5753/bresci.2021.15789