Otimizações em um *Workflow* Científico de Alto Desempenho

Lucas Cruz[1,2], **Micaella Coelho**[2], **Luiz Gadelha**[2], **Carla Osthoff**[2], **Kary Ocaña**[2]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET/RJ)
Petrópolis – RJ – Brasil

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

{lucruz, micaella, lgadelha, osthoff, karyann}@lncc.br

Abstract. *The article brings discussions about the choice of modifications in the execution format of the workflow ParslRNA-Seq, which lead to improved performance and computational scalability, based on the reduction of expenses with I/O operations with the use of SSD in regarding the Lustre parallel file system on the Santos Dumont supercomputer.*

Resumo. *O artigo traz discussões sobre a eleição de modificações no formato de execução do workflow ParslRNA-Seq, que levam a melhora do desempenho e escalabilidade computacional, baseado em redução de gastos com operações de E/S com o uso de SSD em relação ao sistema de arquivos paralelos Lustre no supercomputador Santos Dumont.*

1. Introdução

Na bioinformática, a modelagem de experimentos de sequenciamento RNA é um desafio devido à complexidade e manipulação de grandes volumes de dados. No supercomputador Santos Dumont (https://sdumont.lncc.br), o sistema de arquivos paralelos Lustre, é implantado por meio de o CRAY ClusterStor 9000 v3.3, com um servidor de metadados (MDS) – responsável pelo gerenciamento dos metadados do arquivo – e 10 *Object Storage Servers* (OSS) – responsáveis por gerenciar os dados do arquivo. Cada nó computacional possui um disco local de tecnologia *Solid State* que pode ser utilizado para armazenamento temporário de dados.

O presente trabalho é uma continuação de [Cruz et al. 2021] para a otimização de desempenho do *workflow* científico desenvolvido, chamado ParslRNA-Seq, e apresenta uma nova proposta quanto ao formato de execução do mesmo para o seu processamento em ambientes distribuídos de alto desempenho. Além da nova proposta de execução, o presente trabalho apresenta também uma análise comparativa acerca do ganho computacional da execução desse *workflow* fazendo utilização do sistema de Entrada e Saída Paralelo (Lustre) e do disco local do nó computacional, chamado de *Solid State Drive* (SDD). As estimativas sugerem um ganho de tempo de processamento para cada pipeline de tarefas do *workflow* em cerca de 2 minutos, o que para um ambiente paralelo e distribuído com múltiplos usuários pode ser bastante significativo.

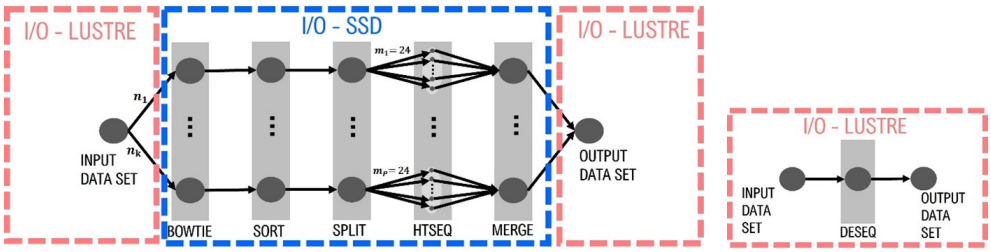
2. Trabalhos Relacionados

[Cruz et al. 2021] levou as execuções do *workflow* científico ParslRNA-Seq a alcançarem um ganho em tempo computacional maior do que 65% em relação a versão apresentada

em [Cruz et al. 2020], com três atividades (bowtie, htseq e deseq). No novo trabalho foram adicionadas 3 atividades (sort, split e merge) de forma estratégica sugerindo o uso da técnica de divisão e conquista: com o particionamento do dado, processamento paralelo de dados e combinação de resultados gerando a diminuição de um Tempo Total de Execução (TTE) de cerca de 3 dias para cerca de 24 minutos.

3. Metodologia

A nova proposta no formato de execução do *workflow* se refere à: forma de alocação de nós computacionais; a desacoplagem da atividade que detém a barreira de sincronização de dados (deseq); e, a utilização da memória SSD. Quanto à forma de alocação dos nós o presente trabalho sugere que para cada dado a ser processado em uma *pipeline*, seja alocado um nó computacional para seu processamento. Assim, temos um *throughput* de rede maior. Quanto a desacoplagem sugerida, é referida acerca da atividade deseq, que necessita de uma sincronização de dados, pois irá utilizar todos os dados de entrada de uma só vez e, sendo de baixo custo computacional, só irá utilizar um único nó computacional (Figura 1). Isso irá permitir que os recursos computacionais alocados para executar o *workflow* sejam liberados mais cedo. Além disso, é sugerida também a utilização do SSD munido da seguinte estratégia: os dados de entrada são lidos diretamente do Lustre e escritos no SSD na primeira atividade componente do *workfow* (bowtie) e o processamento se segue, escrevendo e lendo, no SSD até a atividade merge, que escreve sua saída diretamente no Lustre. Essa saída é então, utilizada como dado de entrada pela atividade deseq, que é a última atividade do *workflow*. O *workflow* passa então, a ser dividido em duas partes: a primeira, descrita na Figura 1(a); e, a segunda, descrita na Figura 1(b).



(a) Primeira parte. (b) Segunda parte.

Figura 1. Nova proposta de processamento do ParslRNA-Seq.

O conjunto de dados de entrada utilizados são os mesmo de [Cruz et al. 2021] e contém ao todo seis arquivos, pertencentes a um experimento real de sequenciamento RNA, com tamanhos variando entre 1.8 GB e 3.0 GB. O ambiente computacional utilizado foi o SDumont e foram alocados seis nós computacionais para execução do *workflow*, dos quais se compõem por duas CPUs Ivy Bridge Intel Xeon E5-2695v2 (12c @2.4GHz) e 64 GB de memória RAM e um SSD de 128 GB.

4. Resultados e Análise

Análise comparativa das execuções entre uso do Lustre e do SSD. Os TTEs apresentados na Tabela 1 dizem respeito a cinco execuções completas da primeira parte do *workflow* (Figura 1(a)) feitas com o uso do SSD e uso do Lustre de modo a comparar a diferença no TTE do *workflow*. Na tabela, estão os dados do tamanho do dado de entrada processado

e seus respectivos tempos de execução, bem como o desvio padrão e a média entre esses tempos. A estratégia de desacoplagem citada na seção 3, serve para impedir que os nós computacionais fiquem ociosos durante tempo demasiado, já que o TTE do *workflow* será sempre definido como o tempo que o maior arquivo de entrada leva para ser processado, o que inclui os tempos de execução das tarefas e das transferências de dados no decorrer da execução. Dessa forma, o TTE do *workflow* passa a ser o tempo de processamento das atividades da primeira parte, acrescidos do tempo da segunda parte.

Tabela 1. TTE da primeira parte do *workflow* usando o Lustre e o SSD.

EXECUÇÕES NO LUSTRE							
TAMANHO	TEMPO TOTAL DE EXECUÇÃO DA PRIMEIRA PARTE DO *WORKFLOW* (minutos)					DESVIO PADRÃO	TEMPO MÉDIO (minutos)
1.8 G	11,6167	12,0833	11,95	12,0667	11,5667	0,2479	11,85668
3.0 G	17,6	17,4	17,7667	17,7667	17,6333	0,1509	17,63334
EXECUÇÕES NO SSD							
TAMANHO	TEMPO TOTAL DE EXECUÇÃO DA PRIMEIRA PARTE DO *WORKFLOW* (minutos)					DESVIO PADRÃO	TEMPO MÉDIO (minutos)
1.8 G	10,0661	10,0328	10,0817	10,1988	10,0816	0,0628	10,0922
3.0 G	15,8649	15,8978	15,5828	15,6994	16,1159	0,2035	15,83216

Nesse cenário é possível ter como base o tempo do arquivo de maior tamanho, de 3 GB, que levará maior tempo para ser processado, e, portanto, o tempo dele determina o TTE do *workflow*. Pela Tabela 1, o tempo médio de execução da primeira parte do *workflow* usando o Lustre é de cerca de 17 minutos. Já usando o SSD, a execução dura em média cerca de 15 minutos. O tempo médio da atividade deseq é de cerca 1,4 minutos [Cruz et al. 2021]. Ou seja, usando o Lustre o *workflow* leva cerca de 19 minutos para finalizar a execução e usando o SSD ele leva cerca de 17 minutos. É possível ainda notar, pelo tempo de execução do arquivo de menor tamanho, de 1.8 GB, que através da estratégia de alocações de nós citada na seção 3, há a liberação de um nó computacional cerca de 6 minutos mais cedo em relação ao arquivo de maior tamanho.

5. Conclusão

Nesse trabalho, o ParslRNA-Seq se beneficia: do uso do SSD; da estratégia de alocação de nós computacionais; e, do desacoplamento da atividade que impede a liberação de recursos ociosos. A exploração desses fatores, levaram a uma redução no TTE do *workflow* de cerca de 24 para 19 minutos. Comparativamente a [Cruz et al. 2021] e o presente trabalho, podemos notar que alocações de nós em blocos únicos, não é compensatória devido à ociosidade dos nós durante tempo excessivo quando terminam de processar os dados de uma *pipeline*. Com a nova proposta, há a liberação de recursos cerca de 6 minutos mais cedo antes do fim do TTE do *workflow*. Como trabalhos futuros, iremos desenvolver estudos para avaliar o cenário em que o sistema de arquivos é compartilhado por outras aplicações, utilizando toda a banda de entrada e saída de dados.

Referências

Cruz, L., Coelho, M., Gadelha, L., Ocaña, K., and Osthoff, C. (2020). Avaliação de desempenho de um workflow científico para experimentos de rna-seq no supercomputador santos dumont. In *Anais Estendidos do XXI Simpósio em Sistemas Computacionais de Alto Desempenho*, pages 86–93, Porto Alegre, RS, Brasil. SBC.

Cruz, L., Coelho, M., Terra, R., Carvalho, D., Gadelha, L., Osthoff, C., and Ocaña, K. (2021). *Workflows* científicos de rna-seq em ambientes distribuídos de alto desempenho: Otimização de desempenho e análises de dados de edg. In *Anais do XV Brazilian e-Science Workshop*, pages 57–64, Porto Alegre, RS, Brasil. SBC.