DEVELOPING EFFICIENT SCIENTIFIC GATEWAYS FOR BIOINFORMATICS IN SUPERCOMPUTER ENVIRONMENTS SUPPORTED BY ARTIFICIAL INTELLIGENCE

Ocaña, K.[1], Coelho, M.[1], Terra, R.[1], Freire, G.[1,2], Santos, M.[1,2], Cruz, L.[1,2], Galheigo, M.[1], Carneiro, A.[1], Fagundes, B.[1], Carvalho, D.[2], Cardoso, D.[2], Meneses, E.[3], Gadelha, L.[1], Osthoff, C.[1]

1 - National Laboratory of Scientific Computing

2 - Federal Center for Technological Education Celso Suckow da Fonseca

3 - National High Technology Center in Costa Rica

The BioinfoPortal Project: Transitioning Bioinformatics Scientific Gateway to HPC Resources

We introduce a discussion about the challenges and research opportunities of integrating the BioinfoPortal gateway (https://bioinfo.lncc.br/) **[Ocaña, K. et al., 2020]** and CSGrid middleware in the Brazilian National High-Performance Computing (HPC) System, which manages the Santos Dumont supercomputer (SDumont, https: //sdumont.lncc.br/), the largest supercomputer in Latin America with 5.1 Petaflops and 36,472 computational cores distributed in 1,134 computational nodes. The three-step roadmap includes: (1) Bioinfo-Portal four-layer architecture, functionalities, and integration to SDumont; (2) findings of performance of three HPC bioinformatics applications; (3) proposing the BioinfoPortal API component (ML 5th-layer) based on machine learning (ML).

The Scientific Gateway BioinfoPortal

BioinfoPortal presents a four-layer architecture (Figure 1). At the bottom, Resource Layer with several types of SDumont clusters provided with bioinformatics applications. Next, Data Layer with data resources. Both layers above are connected/utilized through CSGrid middleware services (white) in Management Layer. At the top, User Interface Layer contains Web interfaces for user interaction. To further drive our science gateway, BioinfoPortal construction and performance need enhancement in terms of resource provision and task scheduling. ML could be used in this regard, based on computational efficiency data gathered from previous experiments which could be generalized by a learning model, enabling a smarter allocation of resources for the next experiments, aiming at the maximization of system usage and throughput. The proposed API for ML (5th-level) is expected to be capable of adaptively distributing the workload, assuring fair response times, managing load balance, and ultimately providing an overall system improvement.

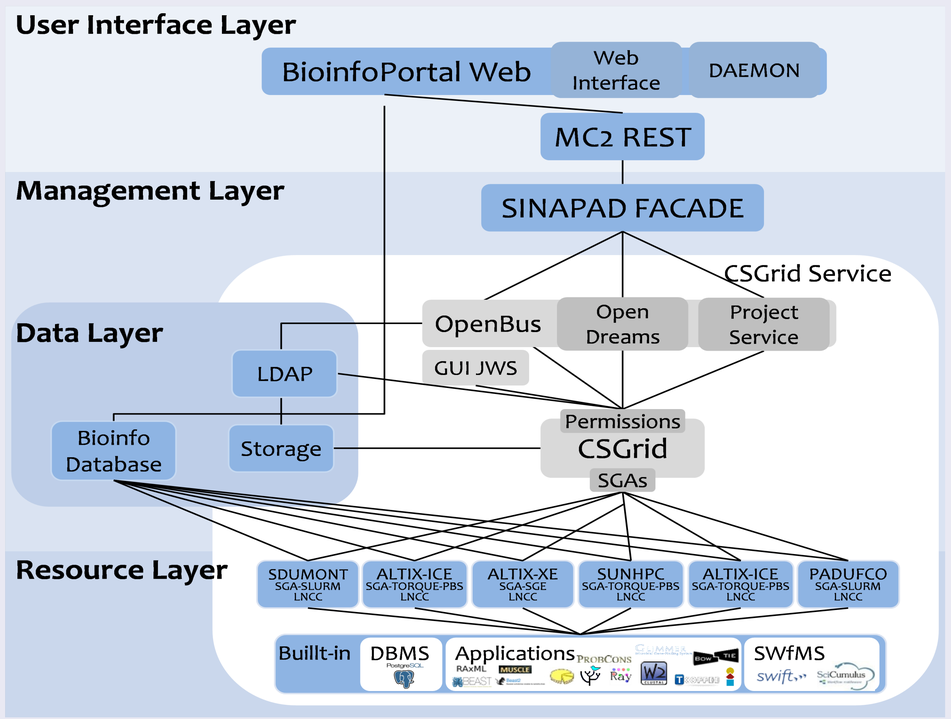


Figure 1: The Layered Architecture of the Scientific Gateway Bioinfo-Portal

First Experiment: RAxML on SDumont CPU Resources

Phylogeny aims to infer the tree of life evolutionary history. Executions of RAxML versions (serial, PThread, MPI, Hybrid) were performed on SDumont. RAxML Hybrid overperformed and was used with genomic data files (D1-D4) of several sizes and RAxML bootstrap (100 to 2000) of several values. RAxML Hybrid was executed in SDumont nodes, where each node consists of two Intel Xeon E5-2695v2 Ivy Bridge CPUs, with 24 cores (12 per CPU). Figure 2 shows speedup of D1 (smallest) and D4 (largest) genomic data. Execution of D4 with RAxML bootstrap of 2000 presents an almost linear speedup, which means the largest files benefit better from SDumont HPC resources.

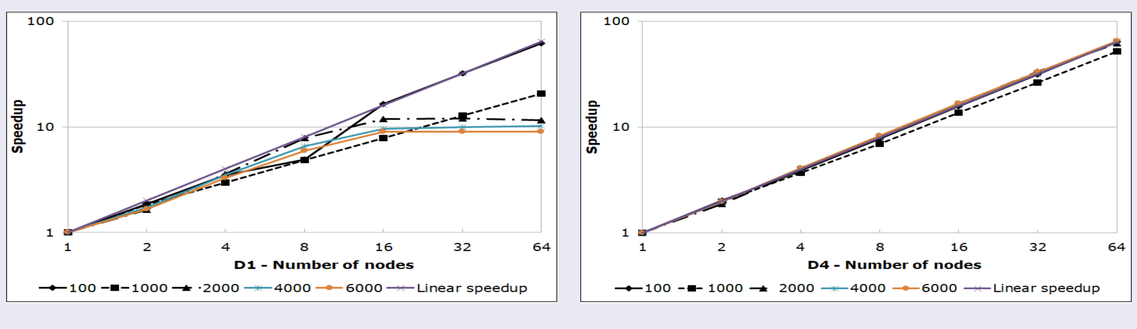
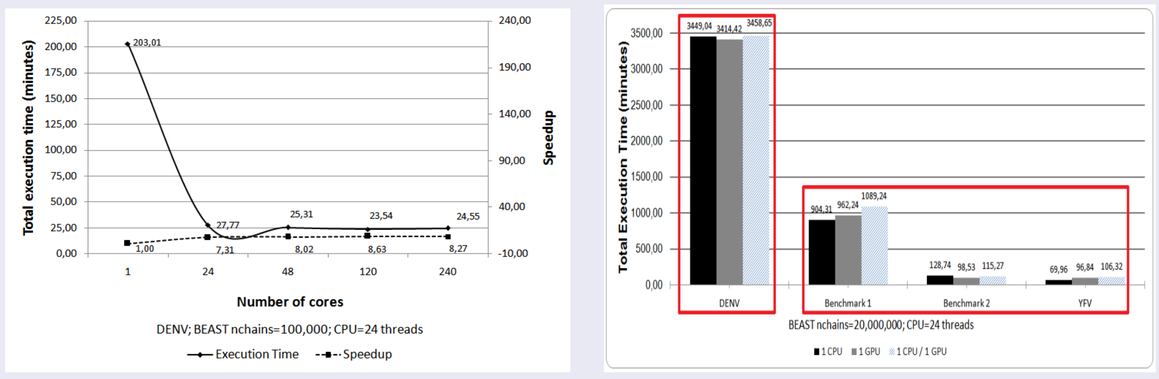


Figure 2: Speedup of RAxML Hybrid on SDumont CPU Nodes

Second Experiment: BEAST on SDumont GPU Resources

Viral evolutionary studies performed with BEAST 1.10 and BEAGLE3 HPC library in SDumont used GPU nodes (24 cores of Intel Xeon E5-2695v2 Ivy Bridge CPUs, 2 NVidia Tesla K40) and MESCA nodes (240 cores of 16 Intel CPUs Ivy). Bench.1, Bench.2, Dengue and yellow fever virus were executed with BEAST chainLength values from 100,000 to 20,000,000. Figure 3(a) shows TET for MESCA using DENV with a performance gain of 7.31 in 24 threads and Figure 3(b) shows TET on multi-thread GPUs. Best results were obtained with GPUs for partitioned data and CPUs for non-partitioned data, which overperformed the hybrid (GPU/CPU) version.

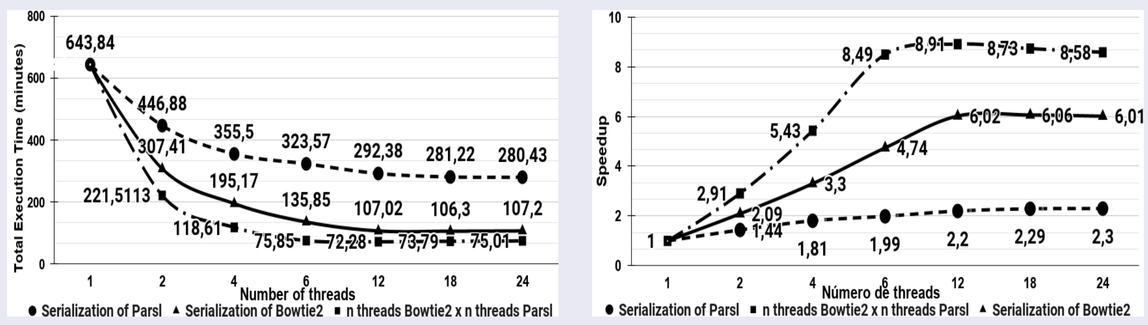


(a) (b)

Figure 3: (a) TET BEAST/BEAGLE on MESCA (CPU) Nodes and (b) TTE on CPUs and GPUs

Third Experiment: Scientific Workflows on SDumont

ParslRNA-Seq is a HPC RNA-Seq workflow managed by Parsl. Figure 4(a) shows three multithreading option scenarios executed on SDumont: or calling Bowtie or Parsl or Both Bowtie/Parsl (the latter option with the best performance). Figure 4(b) shows the ParslRNA-Seq best performance gain is achieved calling the Bowtie/Parsl multithreading with 8.91 using 12 threads; then, adding more threads does not return more gain in performance in one node of SDumont, which consists of two Intel Xeon E5-2695v2 Ivy Bridge CPUs, with 24 cores (12 per CPU).



(a) (b)

Figure 4: (a) TET and (b) Speedup of ParslRNA-Seq on SDumont

Ongoing Steps for Data Analytics

An illustration of a decision tree inducted from a data sample resulting from some bioinformatics experiments (Figure 5) shows input variables regard the environment configuration; the target variable aims to represent overall execution quality. Preliminary results evidentiate an ML-approachable dependency between experiment parameters and computational efficiency, whose predictability enables its exploitation. Besides decision trees, models capable of learning more complex relationships are considered as Neural Networks, Random Forests, and Support Vector Machines.

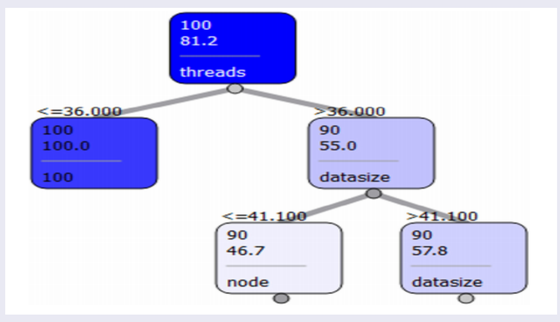


Figure 5: Decision Tree of RAxML Performance: Parameters Genomic Size, SDumont Node Number

Bibliography

1 Ocaña, K., et al. Bioinfo-Portal: A scientific gateway for integrating bioinformatics applications on the Brazilian national high-performance computing network. FGCS. 2020. 107, C, 192–214. https://doi.org/10.1016/j.future.2020.01.030

SPONSORS

CNPq CAPES European Economic and Social Committee