Developing A Scalable Platform For Next-Generation Sequencing Data Analytics Over Heterogeneous Clouds and HPCs : A Case for Transcriptomes and Metagenomes

Shayan Shams
Center for Computation and Technology
Louisiana State University
Baton Rouge, LA 70803
Email: sshams2@cct.lsu.edu

Nayong Kim
Center for Computation and Technology
Louisiana State University
Baton Rouge, LA 70803
Email: nykim@cct.lsu.edu

Jian Tao
Center for Computation and Technology
Louisiana State University
Baton Rouge, LA 70803
Email: jtao@cct.lsu.edu

Ming Tai Ha
Electrical and Computer Engineering Busch Campus,
Rutgers University
Piscataway, NJ 08854
Email: ming.tai.ha@gmail.com

Shantenu Jha
Electrical and Computer Engineering Busch Campus
Rutgers University
Piscataway, NJ 08854-8058
Email: shantenu.jha@rutgers.edu

Ramesh Subramanian
School of Vet. Medicine
Louisiana State University
Baton Rouge, LA70803
Email: ramji@lsu.edu

Vladmir Chouljenko
School of Vet. Medicine
Louisiana State University
Baton Rouge, LA70803
Email: vchoul1@lsu.edu

Konstantin Gus Kousoulas
School of Vet. Medicine
Louisiana State University
Baton Rouge, LA 70803
Email: vtgusk@lsu.edu

Ram Ramanujam
Center for Computation and Technology
Louisiana State University
Baton Rouge, LA 70803
Email: ram@cct.lsu.edu

Seung-Jong Park
Center for Computation and Technology
Louisiana State University
Baton Rouge, LA 70803
Email: sjpark@cct.lsu.edu

Joohyun Kim*
Center for Computation and Technology
Louisiana State University
Baton Rouge, LA 70803
Email: jhkim@cct.lsu.edu

***Abstract*—A novel scalable pipeline for metagenome/ transcriptome is presented. Thanks to the underlying distributed computing platform, a significant roadblock in Next-Generation Sequencing (NGS) data analytics, associated with ever-growing and noisy data sets, can be effectively resolved. On top of the core feature for accessing and utilizing heterogeneous distributed computing resources including HPCs and Clouds (EC2, OpenStack-based, and IBM Bluemix), the distributed application runtime environment is built for efficient management of massive distributed workloads and data processing tasks by leveraging high-end HPC technologies as well as emerging Hadoop-based software models and DOCKER. Thus, the available repertoire of options providing flexible and scalable runtime scenarios constitutes capacities of the pipeline. As a result, the developed pipeline can handle any size of data sets and run the target analysis over distributed computing resources effectively. Last but not least, the scalable platform is capitalized with the introduction of the novel method for de novo genome sequence reconstruction dubbed as Multiple Assembly Multiple Parameter (MAMP). Preliminary results indicate potentials of MAMP.**

I. Backgrounds and motivations

NGS technologies are now ubiquitously applied for not only basic biological discoveries but also a broad range of biomedical and translational research topics. In fact, the emergence of genomics medicine or BigData-scale public health challenges can not be comprehensible without these technologies. NGS is high-throughput sequencing methods generating intrinsically large data sets and are notoriously challenging in data analytics due to complicated artifacts with multiple sources of errors. This is becoming even worse as fast growing interests on population-size genomics as well as metagenomic studies for environmental samples. Therefore, addressing such fundamental aspects of roadblocks with NGS data analytics is increasingly becoming significant, but existing solutions are

* Corresponding author

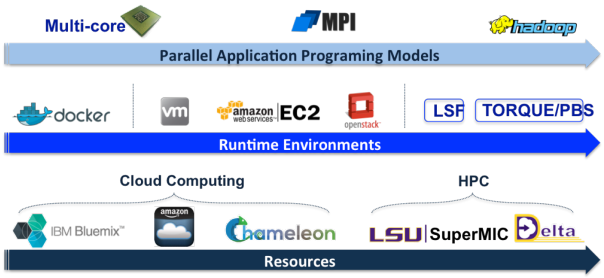

Fig. 1. Distributed heterogeneous computing environments and their hierarchical nature for a distributed application.

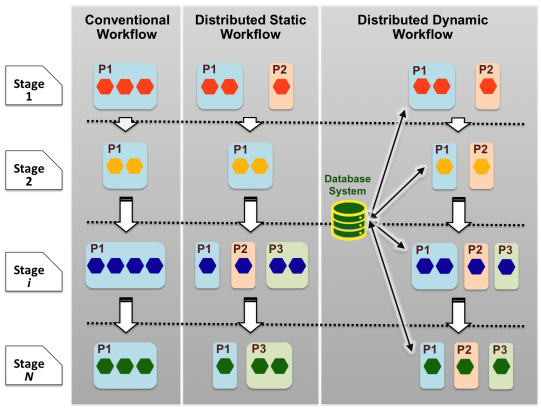

Fig. 2. Schematic illustrations of execution patterns for workflows of distributed applications

often ineffective because of the lack of an integrative approach that combines advances in algorithms, infrastructure, and computational models altogether. To overcome such challenges, we have been focusing on a distributed computing platform and in this work, with a use case for transcriptome/metagenome, we present a novel pipeline. The pipeline is built upon core ideas of the platform and thus other omics applications with NGS are expected to be beneficial with relatively low developmental efforts with reusable components and abstraction.

II. Summary of our contributions

The platform is composed of methodologies to access multiple heterogeneous distributed resources and, as a key component, the distributed application runtime environment (DARE)[1]. New features has been steadily added to DARE over the past years after identifying them as essential for various life science data analyses. [2–4]. The core feature for accessing and utilizing heterogeneous distributed computing resources including HPCs and Clouds is implemented by employing the pilot system, Radical Pilot, developed by Jha and coworkers. Empowered with the pilot system, DARE is in charge of providing efficient management of massive distributed workloads and data processing tasks. In this work, we report recent developmental efforts utilizing cloud systems such as Chameleon (OpenStack) and IBM Bluemix cloud, and HPC systems such as DELTA (IBM Power8-based cluster) and SuperMIC (Intel cluster), adding to Amazon EC2 previously reported[2]. Consequently, multiple local runtime environments for HPCs with schedulers such as PBS/SGE/TORQUE and LSF are seamlessly integrated into a pool of supported resource environments. Equally importantly, virtualization cloud environments such as Amazon EC2 and OpenStack are now fully supported. Finally, we added the alternative options based on DOCKER, altogether resulting in the comprehensive platform to manage distributed applications implemented with MPI and Hadoop. Overall, these new outcomes underscores the scalability of the platform (see Fig. 1).

While our platform is designed to be generic to support any NGS data analytics, we focus on the pipeline for transcriptome/metagenome. In brief, the pipeline allows an individual researcher to conduct essential tasks such as preprocessing of sequencing read data sets, de novo assembly, and post-processing. Unlike most of tools for the same purpose, our pipeline is capable of any size of data volumes and carries out efficient executions using various distributed computing resources. To this end, depending upon input size and estimated intermediate data sets, appropriate resources as well as programming models for distributed applications can be chosen. This scalability for large data set is further strengthened by the capacity of DARE for massive distributed tasks. For example, in Fig. 2, three common execution patterns supported by DARE for executing a pipeline application are shown. This aspect is greatly important since running multiple jobs effectively enables to design a novel method based on an ensemble approach for dealing with noisy data. Ensemble approaches are known to achieve higher accuracy in data analyses particularly with machine learning problems. In our case, for the genome sequence reconstruction, which is the integral part of transcriptome and metagenome studies but very challenging, the task of clustering and merging of contigs obtained with multiple assemblers with multiple parameters such as the number of k-mers is introduced. Consequently, we developed a MAMP-based method. Note that an increased number of tasks with MAMP can be easily managed by the scalable nature of the platform. Our preliminary results with MAMP, when compared to other existing strategies, showed comparable performance (presented in the table of the poster), which highlights its potential with the fact that its main advantage is in that a systematic search of further optimization is possible.

References

[1] Sharath Maddineni, Joohyun Kim, Yaakoub El-Khamra, and Shantenu Jha. Distributed application runtime environment (dare): A standards-based middleware framework for science-gateways. *Journal of Grid Computing*, 10(4):647–664, 2012.

[2] Shayan Shams, Nayong Kim, Xiandong Meng, Ming Tai Ha, Shantenu Jha, Zhong Wang, and Joohyun Kim. A scalable pipeline for transcriptome profiling tasks with on-demand computing clouds. In *HiCOMB, IEEE IPDPS*, 2016.

[3] Anjani Ragothaman, Sairam Chowdary Boddu, Nayong Kim, Wei Feinstein, Michal Brylinski, Shantenu Jha, and Joohyun Kim. Developing eThread Pipeline Using SAGA- Pilot Abstraction for Large-Scale Structural Bioinformatics. *BioMed Research International*, 2014:348725, 2014.

[4] Joohyun Kim, Sharath Maddineni, and Shantenu Jha. Building Gateways for Life-Science Applications using the Dynamic Application Runtime

Environment (DARE) Framework. In *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*, TG '11, pages 38:1–38:8, New York, NY, USA, 2011. ACM.