Short fragment sequence alignment on the HP-SEE infrastructure

Miklos Kozlovszky*, Gergely Windisch**, Ákos Balaskó*

* MTA SZTAKI/Laboratory of Parallel and Distributed Computing, Budapest, Hungary

* Óbuda University/John von Neumann Faculty of Informatics, Budapest, Hungary

m.kozlovszky@sztaki.hu

Abstract - The recently used deep sequencing techniques represent a new data processing challenge: mapping short fragment reads to open-access eukaryotic genomes at the scale of several hundred thousand. This problem is solvable by BLAST, BWA and similar sequence alignment tools. BLAST is one of the most frequently used tool in bioinformatics and BWA is a relative new fast light-weighted tool that aligns effectively short sequences. Local installations of these algorithms are typically not able to handle large problem size therefore the sequence alignment process runs slowly, while web based implementations cannot accept high number of queries. HP-SEE infrastructure allows accessing massively parallel supercomputing infrastructure. With gUSE/WS-PGRADE we have created successfully an online Bioinformatics eScience Gateway, which is capable to serve the short fragment sequence alignment demand of the regional bioinformatics communities within the SEE region. Using workflows we have ported algorithms (BLAST and BWA) to the massively parallel HP-SEE infrastructure. In this paper we describe the created Bioinformatics eScience Gateway, and show as case study how we have implemented the ported BLAST workflow using parameter study. With our online service, researchers can do high throughput sequence alignments against the eukaryotic genomes to search for regulatory mechanisms controlled by short fragments on HP-SEE's supercomputing infrastructure.

Keywords

Application porting, sequence alignment workflow, HP-SEE, gUSE

I. Introduction

Nowadays, bioinformatics provides key in-silico research procedures to match biological sequences against large sequence databases. DNA and protein sequences are built up from IUPAC codes (normal alphabetic characters), which can compared (aligned) with other sequences. The recently used deep sequencing techniques present a new data processing challenge: mapping short fragment reads to open-access eukaryotic genomes at the scale of several hundred thousand. Since in most of the cases searching on large data set means data-, and compute-intensive challenge, solutions to decrease the computational time are highly needed. eScience Gateways have been used worldwide to solve computational problems in bioinformatics like the MOSGrid portal [1,2] or ProGenGrid [3].

This paper shows how we have built up with the grid User Support Environment (gUSE)[4] the portal like HP-SEE Bioinformatics eScience Gateway and as a case study how one of the sequence alignment tools has been ported to HP-SEE's supercomputing infrastructure.

A. Motivation

Our work was carried out mainly from the collaboration between the Grid Application Support Centre of MTA SZTAKI and the Obuda University. The Grid Application Support Centre (GASuC) [5] at MTA SZTAKI has already a quite long track record with successfully ported applications to Distributed Computing Infrastructures (DCIs). GASuC is supporting large EU projects, where the Centre is providing knowledge transfer and development tools for DCI porting tasks. Our solution is heavily needed for some national research labs in Hungary to do short sequence alignments for tiRNA searches against large human datasets and with our service also researchers from the SEE (South East European) region can do sequence analysis with high throughput short fragment sequence alignments against the eukaryotic genomes to search for regulatory mechanisms controlled by short fragments.

II. The HP-SEE project and its infrastructure

The High-Performance Computing Infrastructure for South East Europe (HP-SEE) [6] project links existing and upcoming HPC facilities in South East Europe in a common infrastructure, and it provides support for the active regional research communities (shown in Fig. 1.). The HP-SEE project started in 2010 with duration of 24 months and involves 14 countries mainly from the SEE region. It receives EC support through FP7 under the "Research Infrastructures" action.

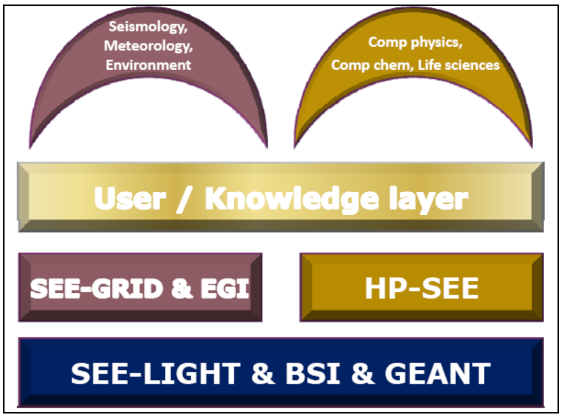
Seismology, Meteorology, Environment
Comp physics, Comp chem, Life sciences
User / Knowledge layer
SEE-GRID & EGI
HP-SEE
SEE-LIGHT & BSI & GEANT

Figure 1. Converged Communication & Service Infrastructure Model for South-East Europe [7]

The initiative opens up the South East European HPC infrastructure to a wide range of new user communities, including those of less-resourced countries, fostering collaboration and providing advanced capabilities to researchers. Three main targeted research communities have been identified: computational physics, computational chemistry and life sciences. MTA SZTAKI and Obuda University are working together in the Life Science VO of the HP-SEE project.

III. HP-SEE'S BIOINFORMATICS ESCIENCE GATEWAY

The Bioinformatics eScience Gateway based on gUSE and operates within the Life Science VO of the HP-SEE infrastructure. It provides unified GUI of different bioinformatics applications (such as BLAST, BWA, or gene mapper applications) and enables end-user access indirectly to some open European bioinformatics databases. gUSE is basically a virtualization environment providing large set of high-level DCI services by which interoperation among classical service and desktop grids, clouds and clusters, unique web services and user communities can be achieved in a scalable way. gUSE has a graphical user interface, which is called WS-PGRADE. All part of gUSE is implemented as a set of Web services. WS-PGRADE uses the client APIs of gUSE services to turn user requests into sequences of gUSE specific Web service calls. Our bioinformaticians need application specific portlets to make the usage of the portal more customized for their work. In order to support the development of such application specific UI we have used the Application Specific Module (ASM) API [8] of the gUSE by which such customization can easily and quickly be done. Some other remaining features were included from WS-PGRADE. Our GUI is built up from JSR168 compliant portlets and can be accessed via normal Web browsers (shown in Fig. 2.).

Figure 2. Login screen of the HP-SEE Bioinformatics eScience Gateway

A. *Bioinformatics eScience gateway resources*

In our Bioinformatics eScience Gateway the gUSE back-end is connected to various DCIs (shown in Fig. 3.). Because the developed workflows are independent from the DCI middleware, the resource allocation depends mainly on the scale of the problem size and the available infrastructure.

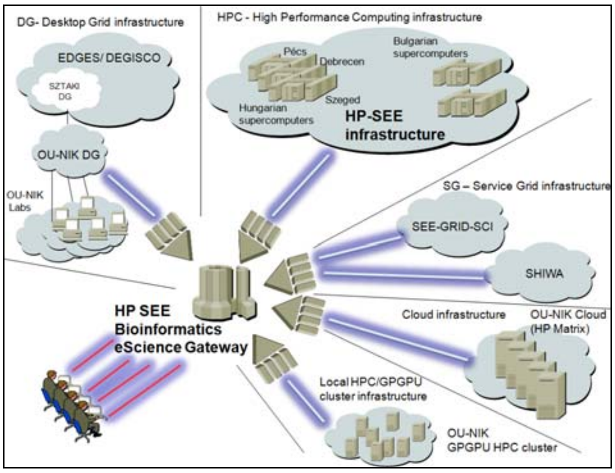

Figure 3. Computational and storage resources of the Bioinformatics eScience Gateway

IV. IMPLEMENTATION OF THE GENERIC BLAST WORKFLOW

Normal applications need to be firstly ported for use with gUSE/WS-PGRADE. Our used porting methodology includes two main steps: workflow development and user specific web interface development based on gUSE's ASM (shown in Fig. 4.). gUSE is using a DAG (directed acyclic graph) based workflow concept. In a generic workflow, nodes represent jobs, which are basically batch programs to be executed on one of the DCI's computing element. Ports represent input/output files the jobs receiving or producing. Arcs between ports represent file transfer operations. gUSE supports Parameter Study [9] type high level parallelization. In the workflow special Generator ports can be used to generate the input files for all parallel jobs automatically while Collector jobs can run after all parallel execution to collect all parallel outputs. During the BLAST porting, we have exploited all the PS capabilities of gUSE.

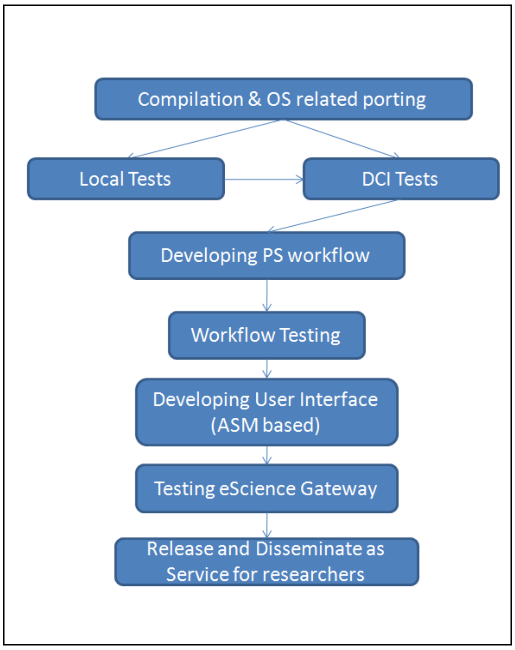
Compilation & OS related porting
Local Tests
DCI Tests
Developing PS workflow
Workflow Testing
Developing User Interface (ASM based)
Testing eScience Gateway
Release and Disseminate as Service for researchers

Figure 4. Porting steps of the application

Parallel job submission into the DCI environment needs to have parameter assignment of the generated parameters. gUSE's PS workflow components were used to create a DCI-aware parallel BLAST application and realize a complex DCI workflow as a proof of concept. Later on the web-based DCI user interface was created using the Application Specific Module (ASM) of gUSE. On this web GUI, end-users can configure the input parameter like the "e" value or the number of MPI tasks and they can submit the alignment into the DCI environment with arbitrary large parameter fields.

During the development of the workflow structure, we have aimed to construct a workflow that will be able to handle the main properties of the parallel BLAST application. To exploit the mechanism of Parameter Study used by gUSE the workflow has developed as a Parameter Study workflow with usage of autogenerator port (second small box around left top box in Fig 5.) and collector job (right bottom box in Fig. 5). The preprocessor job generates a set of input files from some pre-adjusted parameter. Then the second job (middle box in Fig. 5) will be executed as many times as the input files specify.

The last job of the workflow is a Collector which is used to collect several files and then process them as a single input. Collectors force delayed job execution until the last file of the input file set to be collected has arrived to the Collector job. The workflow engine computes the expected number of input files at run time. When all the expected inputs arrived to the Collector it starts to process all the incoming inputs files as a single input set. Finally output files will be generated, and will be stored on a Storage Element of the DCI shown as little box around the Collector in Fig 5.

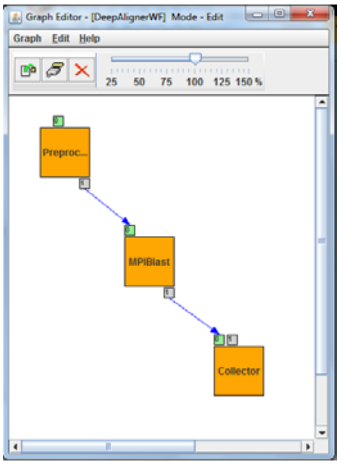

Figure 5. Internal architecture of the generic blast workflow

Due to the strict HPC security constraints, end users should posses valid certificate to utilize the HP-SEE Bioinformatics eScience Gateway. Users can utilize seamlessly the developed workflows on ARC based infrastructure (like the NIIF's Hungarian supercomputing infrastructure) or on gLite/EMI based infrastructure (Service Grids like SEE-GRID-SCI [10], or SHIWA[11]). After login, the users should create their own workflow based application instances, which are derived from pre-developed and well-tested workflows.

V. CONLUSION

In this paper we have described the HP-SEE Bioinformatics eScience Gateway, which operated by Obuda University. The web based portal gives HPC support and increased parallel processing power trough the HP-SEE supercomputing infrastructure to align short biological sequences. For the researchers the eScience Gateway provides high throughput sequence alignment service with parallel BLAST and BWA algorithms, however because the solution is easily extendable we are planning to include other ported alignment algorithms as well. Our project work addressed all issues how to port existing environments (such as bioperl, blast, bwa) to the HP-SEE infrastructure, howto build up effective workflows with gUSE -a stand-alone DCI focused application development environment- and howto create the GUI interface for the researchers using ASM in a seamless way. On the client side the web browser provides for eScience gateway users and workflow developers a native application programming interface to launch time-consuming tasks transparently on the available HPC resources. To demonstrate and test our solution we have used successfully the created workflow based HPC application to do short sequence alignments for tiRNA searches against large human datasets on HP-SEE infrastructure. According to our performance tests using 96 CPUs in HP-SEE's supercomputing infrastructure, thus even with moderate large datasets end users were able to achieve significant speedup in parallel alignment with DCI-enabled version comparing to single multi-core computers. The developed workflow based Bioinformatics eScience gateway is generic enough to host seamlessly later on more processing algorithms helping to solve large scale alignment problems for HP-SEE's Life Science community.

VI. ACKNOWLEDGMENTS

The authors would like to thank three projects supported by the European Commission's 7th Framework Programme for their financial support. This work was supported by the EGI –InSPIRE (Integrated Sustainable Pan-European Infrastructure for Researchers in Europe, under contract number: RI-261323), by the SHIWA (SHaring Interoperable Workflows for large-scale scientific simulations on Available DCIs, under contract number 261585) and by the HP-SEE (High-Performance Computing Infrastructure for South East Europe's Research Communities, under contract no. RI-261499) projects.

REFERENCES

[1] S. Gesing, R. Grunzke, A. Balasko, G. Birkenheuer, D. Blunk, S. Breuers, A. Brinkmann,G. Fels, S. Herres-Pawlis, P. Kacsuk, M. Kozlovszky, J. Krüger, L. Packschies, P. Schaefer, B. Schuller, J. Schuster, T. Steinke, A. Szikszay Fabri, M. Wewior, R. Müller-Pfefferkorn, and O. Kohlbacher. Granular security for a science gateway in structural bioinformatics. In Proc. IWSG-Life 2011, 2011.

[2] [MOS2] S. Gesing, J. van Hemert, P. Kacsuk, and O. Kohlbacher. Special issue: Portals for life sciences|providing intuitive access to bioinformatic tools. Concurrency and Computation:Practice and Experience, 23(3):223{234, 2011.

[3] Aloisio, G.; Cafaro, M.; Fiore, S.; Mirto, M.; , "ProGenGrid: a workflow service infrastructure for composing and executing bioinformatics grid services," Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on , vol., no., pp. 555-560, 23-24 June 2005;doi: 10.1109/CBMS.2005.90

[4] Kacsuk, P. Karoczkai, K. Hermann, G. Sipos, G. Kovacs, J.; WS-PGRADE: Supporting parameter sweep applications in workflows. In: Proc. of 3rd Workshop on Workflows in Support of Large-Scale Science, In conjunction with SC 2008, Austin, TX, USA, 17 Nov. 2008, pp.1 – 10, ISBN: 978-1-4244-2827-4

[5] Accessible online at: http://www.lpds.sztaki.hu/gasuc/

[6] Accessible online : http://www.hp-see.eu/

[7] HP-SEE core presentation, http://www.hp-see.eu/files/HPSEE-WP1-GR-027-CorePresentation-a-2010-11-22.pdf

[8] Á. Balaskó, M. Kozlovszky,A. Schnautigel,K. Karóczkai, ,I. Márton,T. Strodl,P. Kacsuk; Converting P-GRADE grid portal into e-science gateways, Proceedings of International workshop on science gateways. IWSG 2010. Catania, 2010. pp.:1-6

[9] P. Kacsuk, Z. Farkas, G. Sipos, G. Hermann, T. Kiss: Supporting Workflow-level PS Applications by the P-GRADE Grid portal, Towards Next Generation Grids Proceedings of the CoreGRID Symposium 2007

[10] Accessible online at: http: //www.see-grid-sci.eu

[11] Accessible online at: http://www.shiwa-workflow.eu