Accelerating 3D Protein Modeling Using Cloud Computing

Using Rosetta as a Service on the IBM SmartCloud

Peter Kunszt, Lars Malmström
Institute of Molecular Systems Biology and SystemsX.ch
ETH Zürich, Switzerland

Nicola Fantini, Wibke Sudholt
CloudBroker GmbH
Zürich, Switzerland

Marcel Lautenschlager, Roland Reifler, Stefan Ruckstuhl
IBM Switzerland
Zürich, Switzerland

***Abstract*—Biology as a scientific domain needs a growing amount of computational power. However, not every researcher has access to high performance computing resources locally. Today, it is easy to buy computing resources on demand from public cloud providers like Amazon and IBM, paying only for the amount of computing that is really being used. However, the difficulty of setting up the simulation and operating the virtual infrastructure is also often a showstopper for scientists to use cloud resources. This gap is filled by innovative software as a service providers like the ETH Spin-off company CloudBroker GmbH, enabling a more direct access to commercial clouds for researchers in life science. Here we report on a joint project between the ETH Zurich, IBM and CloudBroker to perform a large-scale 3D protein model simulation using the application Rosetta on the new IBM SmartCloud Enterprise.**

Keywords-Cloud computing; HPC; SaaS; Rosetta; Modeling; Proteomics

I. INTRODUCTION

Cloud computing [1] has a lot of potential for research in general and specifically for life sciences. It is potentially a major enabling technology for scientists to perform complex in-silico experiments even if they do not have the necessary local resources. However, researchers in life science are often less apt in setting up their computing infrastructure than researchers in other domains like physics and chemistry. Also, frequently scientific applications in life science are very difficult to install and configure correctly. Moreover, researchers are often not only interested in a single application, but also in the execution of complex workflows assembled of many individual applications. It takes highly skilled experts a very long time to properly optimize the cloud resources for a certain type of workflow.

However, once an application has been set up in the cloud, it is very easy to reuse the cloud for the same purpose later. Clouds provide therefore an opportunity to research labs to render their research workflows into reproducible pay-per-use services. The business model of the ETH spin-off CloudBroker [2] is to provide exactly such a service. This lowers the entry barrier to cloud usage in scientific domains where the end-users are more interested in Software as a Service (SaaS) as opposed to the bare Infrastructure as a Service (IaaS) that most commercial cloud providers offer. CloudBroker offers a platform and APIs to set up and use applications on the cloud. Scientists only interact with the application and its parameters, most details of the infrastructure are automatically predefined as much as possible.

Not every application is suitable to be run on clouds especially applications with very heavy I/O requirements where very large datasets need to be analyzed and transferred through the internet are still difficult. But there are many applications that match the current capabilities of commercial clouds very well. The Rosetta software suite [3] focuses on the prediction and design of protein structures, protein folding mechanisms, and protein-protein interactions. Rosetta is computationally very intensive and does not need a lot of I/O. It is one of the tools that can be provided as a service by CloudBroker on the IBM SmartCloud Enterprise [4], ready to be used by the end-users. Rosetta is free for academic users and needs to be licensed for commercial entities. In our project we were interested in what impact the availability of cloud resources would have on a real research project. We have defined a maximal envelope of two weeks with 1000 CPU cores.

II. MODELING STREPTOCOCCUS PROTEINS

Detailed knowledge about the protein structures in highly virulent pathogens is essential in the fight against antibiotics-resistant bacteria [5]. In an ongoing experiment, the genome of a mutated and highly virulent strain of streptococcus (AP1, virulent for mice) was sequenced and compared to a low virulent strain (sf370, low virulence for mice), identifying proteins that have higher mutation rates than expected. The question is how the mutated proteins are structurally different to be able to investigate differences in protein-protein affinities to the host.

Predicting the protein structure from the amino acid sequence involves a lot of simulation and is very compute intensive. Each structural domain is modeled independently, and proteins can have just one or several structural domains. Some models are therefore very compute intensive and others are relatively quickly done. Finally, during scientific post-analysis, where many additional factors are taken into account, the best or most accurate models are selected and evaluated.

Identification of mutations in the protein structure between low virulent strains (sf370) and high virulent strains

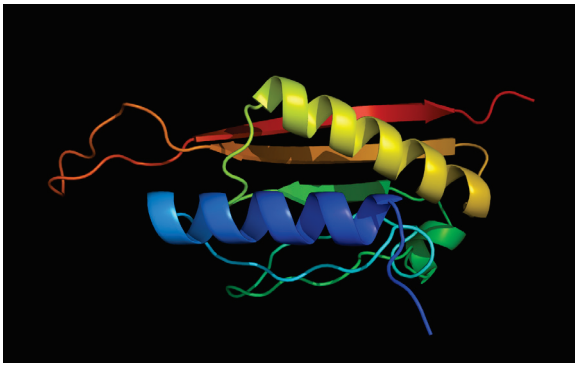

Figure 1: 3D model of a protein. The aim is to calculate 23 million such models.

(AP1) is done by sequencing and comparing both genomes. The 3D protein structures of all proteins that the genome is coding for are modeled based on their amino acid sequence-based predictions. They are compared as well, identifying relevant docking site changes. For Streptococcus we can identify 1697 proteins. There are 832 structural domains identified for de-novo modeling and 1440 structural domains identified for homology modeling. We focus on the de-novo domains first. On average there are 10'000 models for each structural domain, which gives almost 23 million models to calculate. On today's processors, this can be estimated to correspond to roughly 800'000 CPU-core hours for all models to be calculated.

III. Running Rosetta on the Cloud

At ETH Zurich, there is access to a central cluster for such purposes named Brutus, which is a very large heterogeneous multi-purpose resource. However, this resource has to be shared with many other groups. Thus at the average rate of jobs going through the scheduling system, it would take several years to calculate all the models. One could of course purchase a larger share of the cluster, but this resource would not be needed beyond this experiment. On a commercial cloud, one can simply allocate a large virtual cluster for a few days and get the necessary computations done at once.

For our experiment, we could make use of the new IBM SmartCloud Enterprise, running up to 1008 virtual CPU cores simultaneously during a period of two weeks. After some preparation and testing, almost 250'000 net CPU hours were utilized during that time, which amounts to one third of the whole work. By pre-choosing the most interesting targets, enough results are ready now to proceed with the post-analysis steps. During the two-week run, we could model 249 out of the 832 de-novo structural domains, producing 2.3 million models. Without this project, it would have taken many months to get to the same point, so the research project was indeed accelerated considerably.

The deployment of Rosetta into the cloud infrastructure is relatively straightforward. A virtual machine instance needs to be built where the Rosetta software is compiled. The image of this VM is then used inside the IBM cloud.

IV. Experiment Architecture

We had three resource layers in our setup:

- IaaS layer: This is the IBM SmartCloud Enterprise, which offers virtual machines and associated storage on demand.
- SaaS layer: This is the CloudBroker Platform, providing a web user interface and a web service API to interact with.
- End-user layer: The user in the lab can use either a standard web browser or a set of client tools provided by CloudBroker to interact with the system.

Figure 2 shows the setup in detail. The IBM SmartCloud Enterprise infrastructure in the bottom provides an API and a GUI to its users. This is being used by the CloudBroker Platform to deploy the Rosetta software, to control the automatic provisioning of the virtual machine instances, and to launch the Rosetta jobs.

The CloudBroker Platform manages the Rosetta jobs automatically and monitors the execution. Fail-safes are built into the system with automatic restarts and shutdowns should errors occur. The end-user is given a CloudBroker Rosetta client that can be used on the command line very similarly to the Rosetta software itself. It is a smart client that automatically creates the jobs from the input files as necessary. The smart client was used at IMSB to provide the parameters for the protein structure modeling. The jobs can be monitored using the web GUI, which also allows to manually steer the execution, to submit or to cancel jobs and to upload and download data. Administrators can also use the interfaces directly to monitor the execution or to intervene in case of problems. In our experiment, the smart client was used to submit all the data to the cloud. It created over 36'000 jobs that were then managed by the platform and executed on the IBM cloud. Its invocation is just a single command-line command with some parameters to steer the job distribution. The same client was used to retrieve the results. The monitoring could be done through the CloudBroker platform web interface.

Since the IBM SmartCloud Enterprise does not yet have a global cloud storage space, a dedicated storage instance was set up in the cloud to keep track of the persistent data. The smart client uploaded the input datasets to this instance and the CloudBroker Platform managed the distribution of the input data to the various instances, taking care of data and job placement. Result data were also copied back to the storage instance, from which they could be retrieved again. This solution would not scale to large datasets, but for Rosetta it was sufficient as it requires only modest input and output data. IBM plans to set up a cloud storage service much like the Amazon Simple Storage Service S3, which would then assure that the storage is persistent and scalable

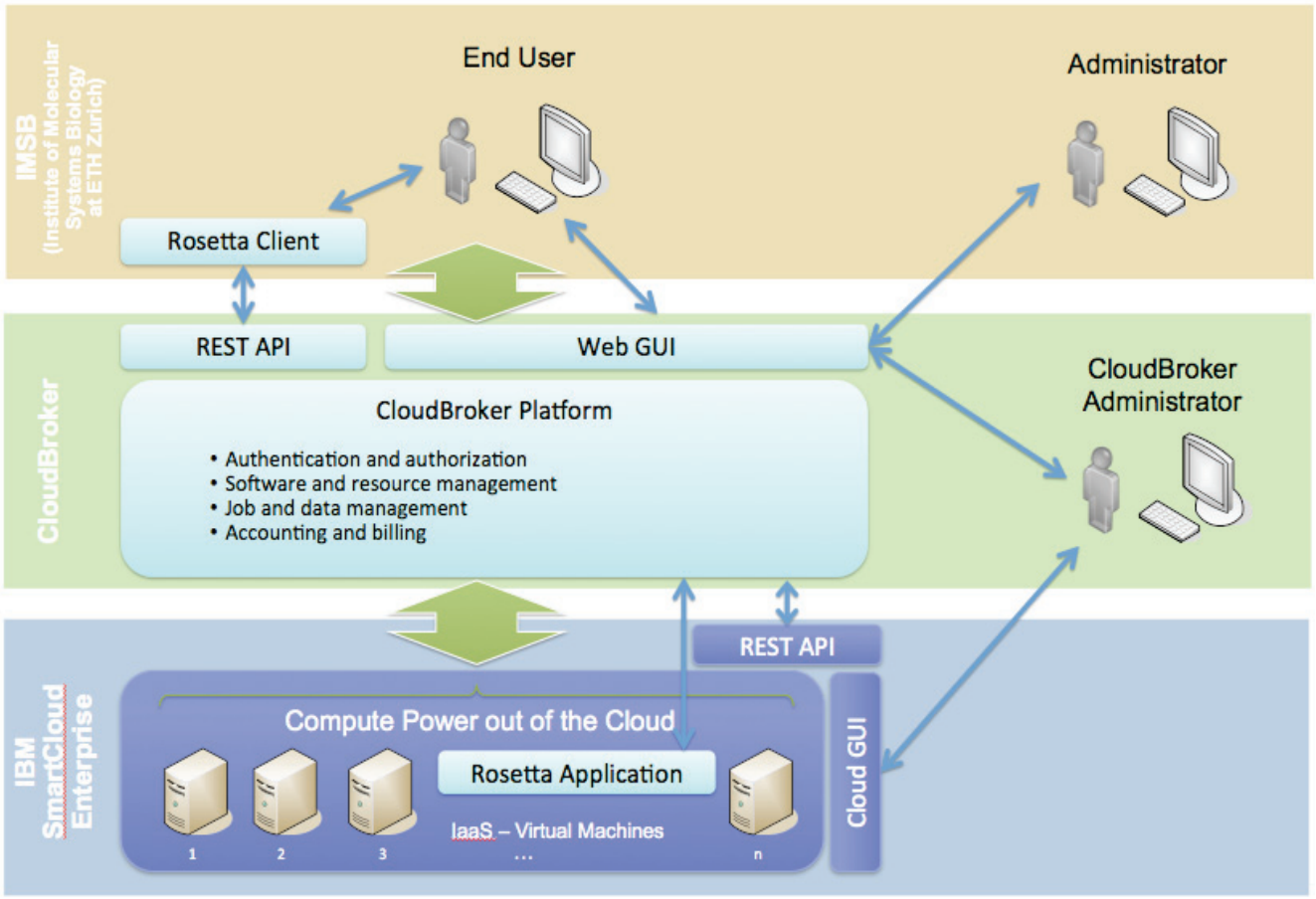

Figure 2: Setup of the cloud infrastructure

inside the IBM cloud as well. The CloudBroker platform automatically managed all the cloud resources, including instance and job failures.

V. Discussion

The aim of our project was to prove that a lot of time can be saved by using cloud services in life science research.

The IBM SmartCloud Enterprise is still relatively new, but it has all the necessary features of elasticity and on-demand usage, and it has a programmable API to allow other tools to drive the system. The IBM SmartCloud is for enterprise users only, not for private persons or individual users, which results in a very professional environment. As to expect for an IaaS cloud offering, one can choose from a range of virtual machines from small single-core 'Copper' to 16-core 'Platinum' instances. As of this writing, there are six IBM datacenters around the globe that can be used. For our two-week experiment, we have chosen the datacenter in Toronto, Canada.

The interesting part of the project from the point of view of the research user was the tool provided by CloudBroker to drive the computation. Using the smart client was very straightforward, and the experiment could be set up very quickly. The provisioning of the cloud infrastructure, submission of the workload and the data management was done automatically in the background. Monitoring could be performed through the CloudBroker Platform's web interface. This is where we have seen another considerable acceleration in terms of output – models could be run immediately, without the need to apply for resources, or to set up and test the software, etc.

We have chosen the envelope of two weeks and 1'000 CPU cores because it is still relatively small to be provisioned on the IBM side (we have been using up to 63 Platinum instances), but it is considerably larger than anything that can be provided to an individual user through a regular university cluster, where the maximal queue length is 8 to 48 hours. Also, it is a relatively long time so that we could test the robustness of the cloud. The CloudBroker platform automatically manages failures: nonresponsive and failed jobs are stopped and restarted. There have been only very few such instances. We estimate the cost of such an extended calculation between $25'000 and $35'000 depending on parameterization, failure rates and the Rosetta licensing costs (free for academic institutions).

Rosetta has already been tested on large public infrastructures before. The Rosetta@home project [6] is one

of the largest distributed computing projects on the BOINC platform. Rosetta has also been run successfully on the Amazon cloud before [7].

VI. Outlook

The streptococcus modeling project now can continue with the scientific evaluation, and certainly also the remaining proteins will be modeled. Rosetta can do many more calculations other than 3D protein structure modeling, and we intend to set up different protocols on the CloudBroker Platform to be used by the research labs.

For example, one can set up a separate application per protocol, which is usually a complex combination of a chained execution of several Rosetta modules. Often it is difficult to reproduce protocols or results that are published in the literature. Providing published protocols as ready-to-use software as a service will enable the scientists who develop the protocols to have a much larger impact, by making their algorithms easily accessible to other users. Also the whole community can profit by being able to apply the published protocols to known and new targets.

Acknowledgments

We would like to thank for the support given by the IBM and CloudBroker technical teams, and especially IBM Switzerland for making this experiment possible. We also thank the SyBIT project of SystemsX.ch, the Swiss initiative in systems biology, for its support.

References

[1] Peter Mell, Tim Grance, "The NIST Definition of Cloud Computing", National Institute of Standards and Technology, Special Publication 800-145 (Draft), January 2011

[2] http://www.cloudbroker.com

[3] http://www.rosettacommons.org, "Portfolio Highlight: Rosetta++ Software Suite". UW TechTransfer – Digital Ventures. Retrieved September 7, 2008.

[4] http://ibm.com/smartcloud

[5] World Health Organization, "Policy Package to Combat Antimicrobial Resistance", World Health Day 2011: Policy Briefs, April 2011

[6] "What is Rosetta@home?". *Rosetta@home forums*. University of Washington. Retrieved September 7, 2008.

[7] http://www.bleedingedgebiotech.com/blog/bioengineering/antibody-docking-on-the-amazon-cloud/