Science as a Service: How On-Demand Computing Can Accelerate Discovery

Ian T. Foster
University of Chicago
and
Argonne National Laboratory
foster@anl.gov

Ravi K. Madduri
University of Chicago
and
Argonne National Laboratory
madduri@anl.gov

Categories and Subject Descriptors
H.m [Information Systems]; C.1.4 [Computer Systems Organization]: Parallel Architectures - *Distributed architectures*; J.3 [Computer Applications]: Life and Medical Sciences - *Biology and genetics*

Keywords
Globus, service oriented science, scientific workflows, cloud computing

Abstract
We originally posited the notion of science as a service in 2005 as a means of publishing and accessing scientific data and applications through internet accessible services [1]. At that time, researchers were only just grasping the benefits of employing the same service oriented architectures commonly used in other domains. Since this time we have indeed seen a huge uptake in researchers leveraging services to disseminate and share data and applications in fields as diverse as genomics [2], climate science [3], and physical sciences [4]. In addition, commercial software as a service (SaaS) products like Google Docs and Gmail are now used by many researchers in everyday activities. The major benefit of a SaaS approach is that researchers are able to invoke applications or access data remotely over the internet without needing to know the inner workings of the service.

Our vision of science as a service worked well in a world when computing resources were scarce; when we needed to federate heterogeneous resources and make them accessible to researchers; when different tools and data were provided using different interfaces and representations; and when research problems involved datasets that could be hosted and analyzed on a single computer. In this talk we re-examine our vision of science as a service in a world in which computing resources are now commoditized; researchers are increasingly facing 'big data' challenges; cloud providers, such as Amazon, have become viable alternatives to purchasing dedicated infrastructure; and reliable infrastructure for scientific problems is only an API call away.

Computation and automation have become vital for discovery in many scientific domains. For example, decreased sequencing costs in biology have transformed the field from a data-limited to computationally limited discipline. Increasingly, researchers must process hundreds of sequenced genomes to determine statistical significance of variants.

Science Cloud'13, June 17, 2013, New York, NY, USA.
ACM 978-1-4503-1979-9/13/06.

Small datasets could be analyzed on personal computers in modest amounts of time: a few hours or perhaps overnight. However, this approach does not scale to large Next Generation Sequencing (NGS) datasets. Instead, researchers require high-performance computers and parallel algorithms if they are to analyze their data in a timely manner.

We use an example to illustrate the problems and opportunities. In 2010, we developed, in collaboration with researchers from the National Cancer Institute, a Lymphoma prediction workflow that linked common bioinformatics tools hosted on small clusters at different institutions [5]. This work was successful, in that the workflow was shown to classify unknown lymphoma tissues automatically. However, the work involved was substantial, requiring months of development time by a skilled team and much interaction with staff at participating institutions. In contrast, the cloud-based Globus Genomics system [6] that we developed in 2013 allows similar analytical workflows to be developed in hours—while at the same time providing access to cutting edge tools and on-demand computing power to scale analyses to large scale datasets. Thus, researchers can spend more time on interesting science through downstream analysis and less time building complex analysis infrastructures, resulting in accelerated time to discovery. As we explain below, a key difference is that in Globus Genomics, software runs entirely on commercial cloud resources. Thus, researchers do not require knowledge of the underlying infrastructure and tools—or even workflows if they are happy to use pre-packaged applications.

On-demand access to infrastructure, science services, and application software is particularly important within the small and medium labs in which most research is performed. Indeed, such access can reduce the competitive disadvantage that smaller labs may otherwise experience relative to large well-funded labs, by making tools formerly available only in large labs and specialist researchers accessible to all; providing scalable infrastructure and platform services to create on-demand scientific services; and automating previously manual data-processing and analysis tasks. These approaches will enable rapid scientific advances by automating routine information technology functions; allowing infrastructure experts to develop and manage good infrastructure through providers such as Amazon; enabling information technology (IT) experts to develop high performance yet efficient platform services on which researchers can build scientific applications; and empowering scientists to develop and manage valuable scientific services that other researchers can leverage to conduct their own research.

In a sense the growing number of science cloud services are analogous to building blocks. As we assemble bigger and better building blocks—blocks that are well-managed and maintained, and reliable and scalable—scientific entrepreneurs can more easily build applications that solve specific scientific problems without having to create and support monolithic software stacks themselves.

At the Computation Institute at the University of Chicago and Argonne National Laboratory, we have embraced these philosophies in our recent development of Globus Genomics (http://globus.org/genomics/), an end-to-end hosted service designed to efficiently and easily analyze large quantities of NGS data using state of the art algorithms, efficient data management tools, a graphical web-based workflow environment and on-demand computing infrastructure.

Rather than implement an entire software stack from scratch, Globus Genomics leverages a collection of existing cloud-based services. We use elastic computational infrastructure provided by Amazon Web Services, a commodity infrastructure as a service (IaaS) provider, and resell this on-demand capacity for scalable workflow execution. We use the Condor scheduler to manage a dynamically assembled pool of hosts. We outsource high performance data transfer and user, group and credential management to Globus Online [7], a platform as a service (PaaS) provider also developed and operated by our team. Finally, we host a Galaxy workflow system [8] to enable easy to use graphical workflow orchestration.

Globus Genomics users can build new analysis workflows from scratch. Equally importantly, we develop and integrate prepackaged analysis tools and best practice pipelines. Thus, users can analyze large amounts of data using computationally efficient analytical pipelines and cutting edge tools that leverage the power and flexibility of on-demand cloud computing resources—without being exposed to the complexities of managing large scale infrastructure; deploying and configuring analysis tools; transferring data between sequencers, analysis nodes and storage systems; or managing their own users and groups. Globus Genomics handles fetching data from a specified source (e.g., a commercial sequencing service); acquiring resources; scheduling computations; and shipping results to a specified destination.

While the move towards on-demand science as a service has considerable benefits for both service providers and service consumers, providers face the particular challenge of sustainability. Traditionally, owners of scientific applications have created open source software releases that can be downloaded, installed, and used by consumers on their own resources and at their own cost. However in a service-based model the software stack is supported and hosted by the service provider who in turn incurs all costs (development, infrastructure and expertise). As services become more useful and therefore more popular, providers incur increased costs associated with operating the system at larger scale. New utility models are needed to support such services—models that will involve a philosophical paradigm shift for users and the funding agencies that support them. In Globus Genomics we employ a pre-paid subscription model, in which users are charged for the compute and storage resources they use. We believe that in the near future this type of model will become commonplace as others look to recoup the costs of providing on-demand scientific services.

Acknowledgment

This work was supported in part by the U.S. Department of Energy under contract DE-AC02-06CH11357 and the NIH through the following grants: U24 GM104203 Bio-Informatics Research Network Coordinating Center (BIRN-CC) and R24HL085343 Cardiovascular Research Grid (CVRG). We also appreciate the support of the Globus Online and Galaxy teams.

References

[1] Foster, I. “Service Oriented Science,” Science, vol. 308, no. 5723, pp. 814,817, May 2005.

[2] Meyer, F., Paarmann, D., D’Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R.A.: “The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.” BMC Bioinformatics, vol. 9, no. 385, September 2008

[3] Bernholdt, D., Bharathi, S., Brown, D., Chanchio, K.., Chen, M., Chervenak, A., Cinquini, L., Drach, B., Foster, I., Fox, P., Garcia, J., Kesselman, C., Markel, R., Middleton, D., Nefedova, V., Pouchard, L., Shoshani, A., Sim, A., Strand, G., and Williams, D., “The Earth System Grid: Supporting the Next Generation of Climate Modeling Research,” Proceedings of the IEEE , vol.93, no.3, pp.485,495, March 2005

[4] Klimeck, G., McLennan, M., Brophy, S.P., Adams, G.B., and Lundstrom, M.S., “nanoHUB.org: Advancing Education and Research in Nanotechnology, Computing in Science & Engineering,” vol.10, no.5, pp.17,23, Sept.-Oct. 2008.

[5] Tan, W., Madduri, R., Nenadic, A., Soiland-Reyes, S., Sulakhe, D., Foster, I., and Goble, C. A., "CaGrid Workflow Toolkit: a Taverna based workflow tool for cancer grid." BMC Bioinformatics, vol. 11, p. 542, 2010.

[6] Madduri, R., Sulakhe, D., Liu, B., Davé, P., Lacinski, L. and Foster, I. “Experiences in building a Next-Generation Sequencing Analysis Service using Galaxy, Globus Online and Amazon Web Services,” XSEDE 2013, San Deigo, CA, USA, July 2013.

[7] Foster, I., “Globus Online: Accelerating and Democratizing Science through Cloud-Based Services,” Internet Computing, IEEE, vol.15, no.3, pp.70,73, May-June 2011.

[8] Goecks, J., A. Nekrutenko, and J. Taylor, “Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.” Genome Biology, vol. 11, no. 8, pp. R86, 2010.

Bio

Ian Foster is the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago and an Argonne Distinguished Fellow at Argonne National Laboratory. He is also the Director of the Computation Institute, a joint unit of Argonne and the University.

Ravi Madduri is a Fellow at the Computation Institute at the University of Chicago and Project Manager in the Mathematics and Computer Science Division at Argonne National Laboratory.