The KBase Architecture and Infrastructure Design

Tom Brettin[*1] (brettints@ornl.gov), Bob Olson[2], Ross Overbeek[2], Terry Disz[2], Bruce Parello[2], Shiran Pasternak[5], Folker Meyer[2], Michael Galloway[1], Steve Moulton[1], Dan Olson[2], Shane Canon[3], Dantung Yu[4], Shiran Paternak, Pavel Novichkov[3], Daniel Quest[1], Narayan Desai[2], Jared Wilkening[2], Miriam Land[1], Scott Deviod[2] **Adam Arkin[3]**, Robert Cottingham[1], Sergei Maslov[4], and Rick Stevens[2]

1 Oak Ridge National Laboratory, Oak Ridge, TN; 2 Argonne National Laboratory, Argonne, IL; 3 Lawrence Berkeley National Laboratory, Berkeley, CA; 4 Brookhaven National Laboratory, Upton, NY; 5 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
http://kbase.us/

Project Goals: The Systems Biology Knowledgebase (KBase) has two central goals. The scientific goal is to produce predictive models, reference datasets and analytical tools and demonstrate their utility in DOE biological research relating to bioenergy, carbon cycle, and the study of subsurface microbial communities. The operational goal is to create the integrated software and hardware infrastructure needed to support the creation, maintenance and use of predictive models and methods in the study of microbes, microbial communities and plants. The driving objectives of the KBase architecture and infrastructure design focus on creating an unprecedented user experience. The integrated software and hardware infrastructure supporting the user experience comprises a continuously expanding collection of software and services. These are hosted on a physical infrastructure consisting of high speed wide area networking, cloud computing resources, and state of the art cluster computing resources.

Achieving our goal of an integrative architecture to support predictive modeling requires enabling a user experience that covers a range of users. These include senior biologists determined to understand and create biological models, computational biologists developing new algorithms and statistical models that form a basis for the biological models, and bioinformaticists chaining together complex workflows to generate, summarize and integrate data that feed into the biological models. These user activities are enabled at various levels of abstraction, including a) knowledge creation, reproduction and sharing; b) rich web applications; c) programmatic Application Programming Interface (API) libraries and scripts; and d) wire level communication access.

The development of the KBase Unified API is based on a service-oriented approach to deliver both functionality and data to the community. Behind this lies a set of services backed by servers. Initially, these services will be developed by the KBase infrastructure team and will support a long term goal of community developed and contributed services. Our initial set of services will be backed by the following servers:

1. **Genomic Servers** that provide access to a rapidly growing set of genomes, features of those genomes, and annotations of both genomes and features.
2. **Expression Data Servers** creating access to a growing body of expression data, along with the underlying encoding of metadata needed to support interpretation.
3. **Protein Family Servers** supporting access to a variety of the existing collections of protein families.
4. **Polymorphism Servers** capturing various genetic polymorphisms such as single nucleotide polymorphisms, tandem repeats, and copy number variations.

5. **Phenotype Servers** enabling the relationships between genotype and phenotype to be understood.
6. **Compound and Reaction Data Servers** supporting a unified and maintained representation of reaction networks.
7. **Metabolic Modeling Servers** that support the construction and maintenance of metabolic models.
8. **Regulatory Models Servers** that support the construction and maintenance of regulatory models.

Our KBase physical infrastructure builds on the successes of DOE investment in our national scientific cyber infrastructure and can therefore leverage enormous intellectual resources present in the DOE community.

Building on ESNet allows us to construct a wide area network between the partner labs that enables a virtual hardware infrastructure. In the first quarter of the project we have established 10Gbit data transfer connectivity between KBase data transfer nodes.

Enabling cloud computing on Magellan will create new opportunities that range from rapid deployment of developer environments to highly scalable production servers. The acceptance of virtualization technology is growing, and the use of machine images produced by others is already visible in our core services. In the near-term, we are establishing the infrastructure to run existing images, both community based and internally created, on which the infrastructure is dependent. Examples include images supporting microbial community models and plant genome wide association studies. For the mid-term, we plan to contribute machine images to the community by creating snapshots of parts of our environment so that others can use them on the hardware of choice. For the long term objective, the ability to host running machine images on KBase hardware is a means for promoting collaboration and community support.

Cluster Computing has long been a critical part of biological data analysis. In collaboration with computing centers created by the Office of Advanced Computing Research such as NERSC, our underlying cluster services can leverage these resources and scale to meet needs.

KBase aims to power the next wave of biological research in DOE and beyond. Enabling these capabilities requires a software and hardware infrastructure that is integrated, extensible, and scalable. The architecture is designed to meet these needs and support user functionality to visualize data, create models or design experiments based on KBase-generated suggestions.

This work is supported by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract DE-AC02-06CH11357.