DOE Systems Biology Knowledgebase

KBASE

Data and modeling for predictive biology

Overview of Infrastructure

Tom Brettin and Rick Stevens
Oak Ridge and Argonne
National Laboratories

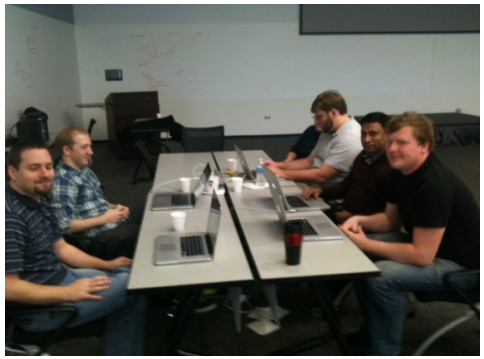
KBASE predictive biology
DOE Systems Biology Knowledgebase

Working As One Team

Communities Hackathon
Jan 2012, LBL

Plant CDM Design
and Build
Jan 2012, ORNL

First Internal Kbase Build – Feb 2012, ANL



Guiding Principles

- Enabling science
  - Scientific objectives lead to software requirements
  - The project scientific drivers provide scope
- Open access to public data, services, and software
- Community contribution of data and services

Infrastructure Capabilities

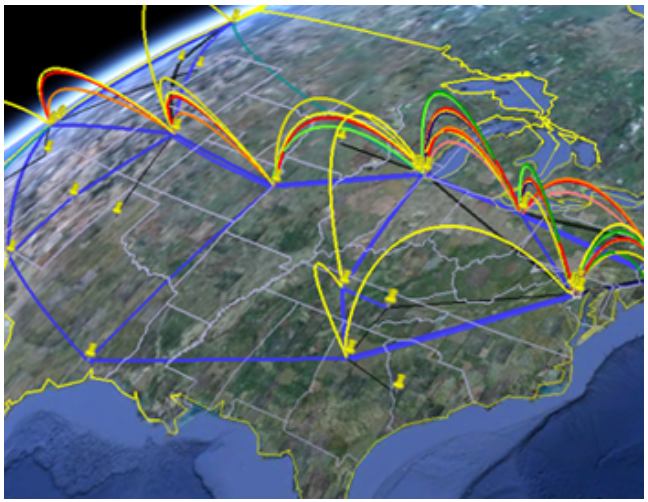
KBase is providing robust and scalable infrastructure to support new science by creating and integrating…

- High speed data transfers that will enable better response times
- Access to back end storage systems for large data sets
- Remote compute services for HPC, cluster and cloud based tasks
- Robust workflow support
- Web UX access to data and computing
- Persistent and transient data management capabilities
- Support for users, teams, projects and cross talk
- Integration with emerging commercial cloud infrastructures and DOE supercomputing

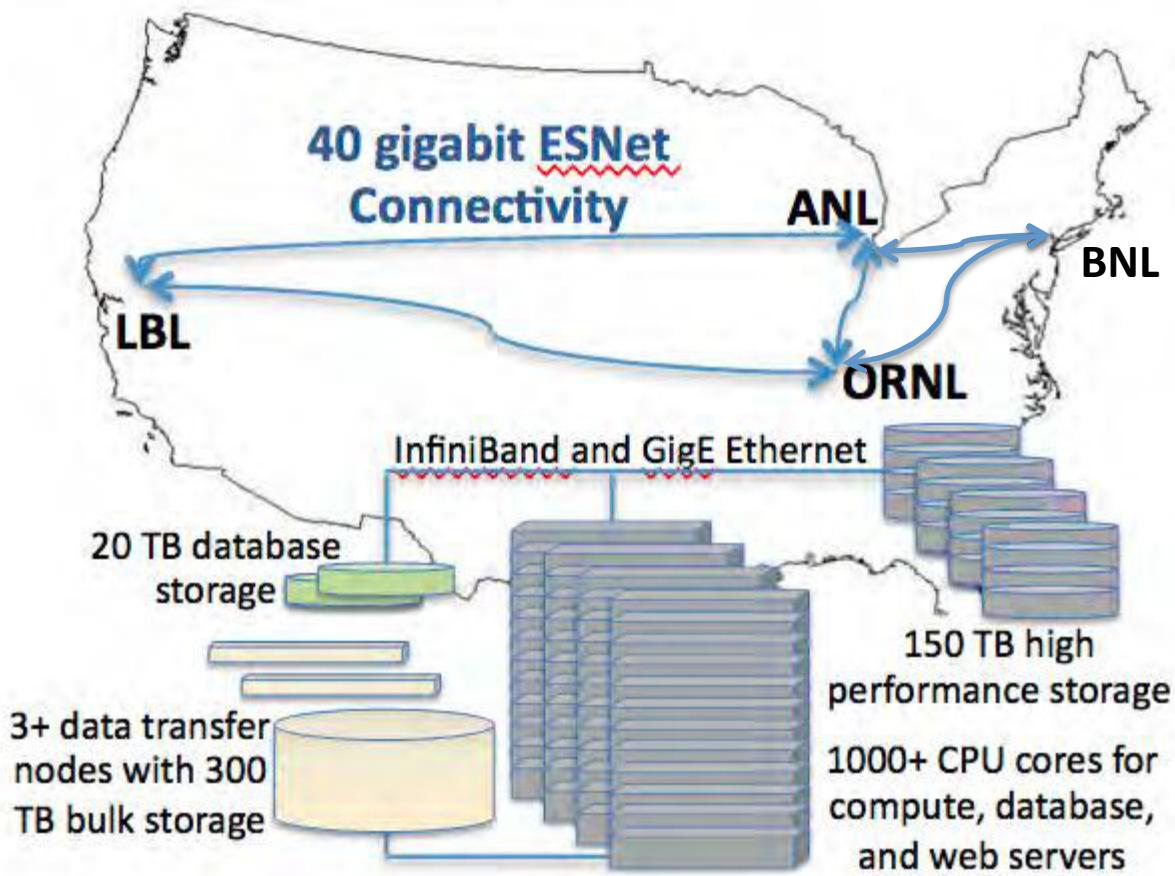
Department of Energy is without doubt leading the world in both it's production quality scientific computing infrastructure as well as new R&D to keep the United States at the forefront of the computational sciences.

A NATIONAL PHYSICAL INFRASTRUCTURE

Energy Sciences Network (ESnet)



KBase leverages ESNet for 10+ Gb/s data transfer between all nodes



- **ESnet backbone (ESnet4) is a national 10 Gbps optical circuit infrastructure**
- **GridFTP and Globus Online enabled data transfers**
-
- **National failover and redundancy**

The DOE KBase Cloud

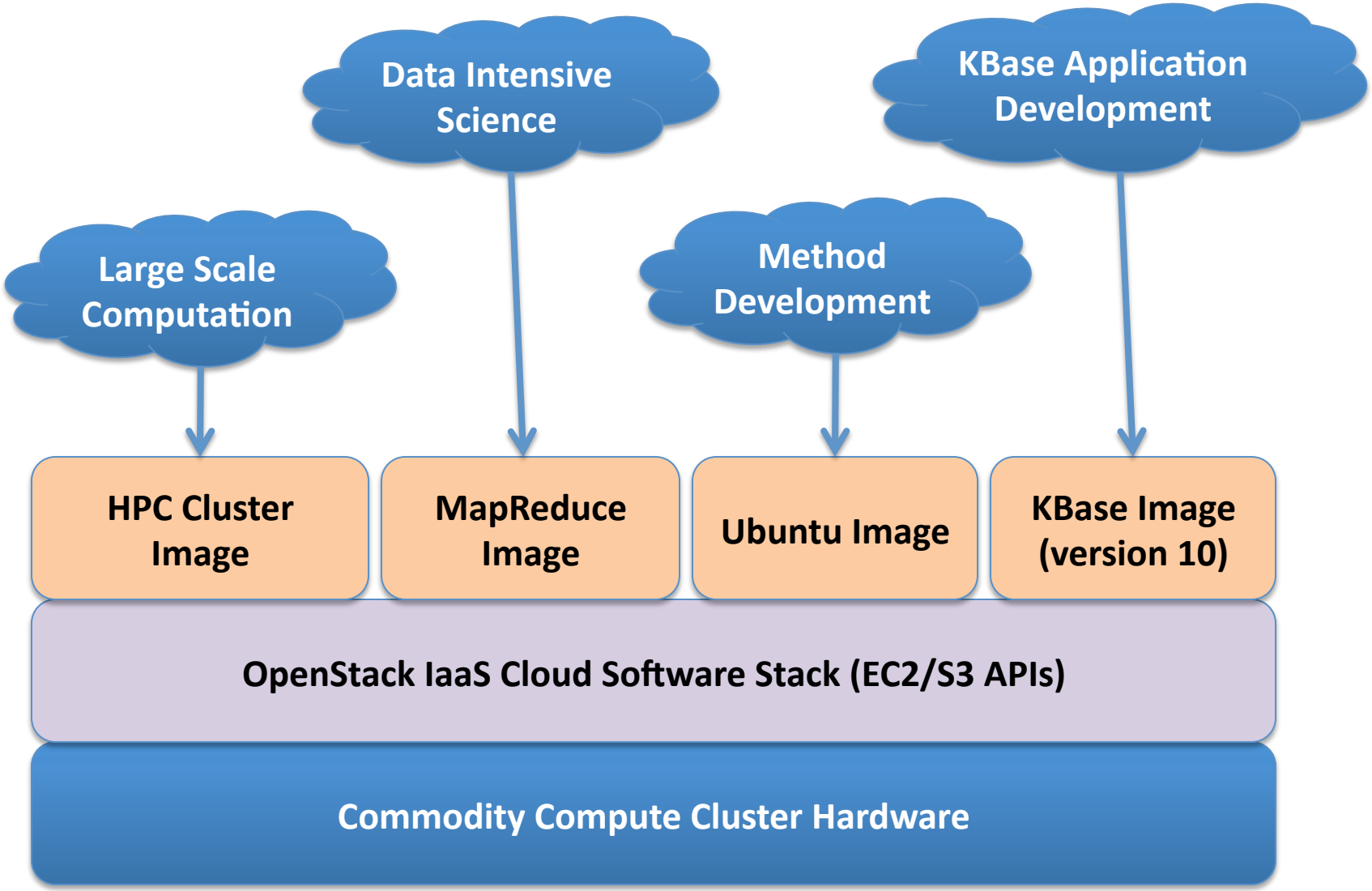
Built on the DOE ASCR investment in the Magellan cloud infrastructure

Current configuration of 700 nodes homed at ANL optimized for heterogeneous applications

- Open Stack Cloud @ Argonne
- Open Stack Cloud @ Oak Ridge
- Cluster system @ Berkeley
- Cluster system @ Brookhaven

KBASE predictive biology
DOE Systems Biology Knowledgebase

The Kbase Cloud Architecture



The software system architecture can be best summarized by two modern architectural idioms; those are Services Oriented Architecture and System of Systems. These two paradigms are captured in modern cloud technology such as SaaS.

SOFTWARE ARCHITECTURE

The KBase Services

Services Oriented Architecture: The KBase Unified API access to a highly diverse set of services ranging from quick retrieval of simple data to massive computations on the KBase Cloud.

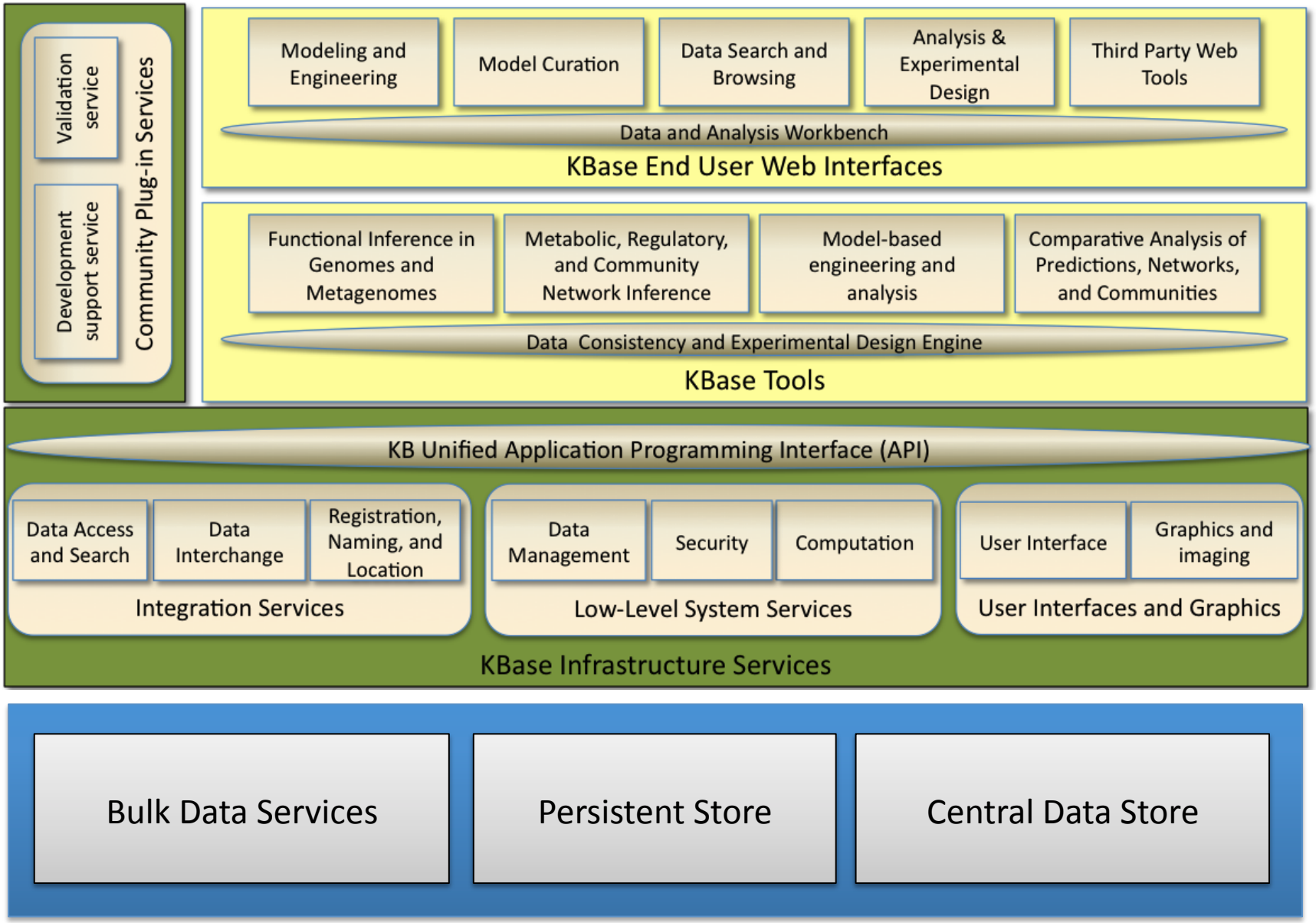
- In a SOA the system is functionally decomposed into many services each of which is implemented as one or more servers.
- Our long-term goal includes community developed and contributed services. Our initial set of services will be backed by the following example servers:

Genomic Servers	***Protein Family Servers***	***Phenotype Servers***	***Polymorphism Servers***
Compound and Reaction Data Servers	***Metabolic Modeling Servers***	***Expression Data Servers***	***Regulatory Models Servers***

KBASE predictive biology

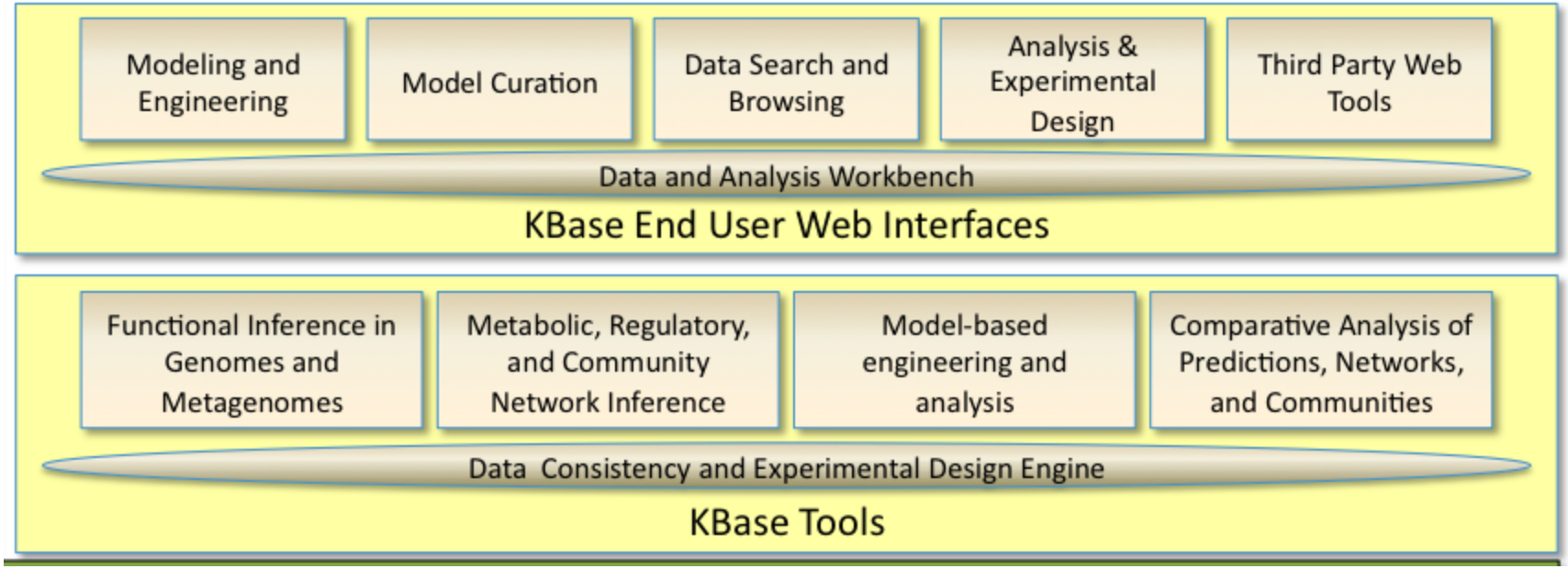
DOE Systems Biology Knowledgebase

The KBase Infrastructure and Services



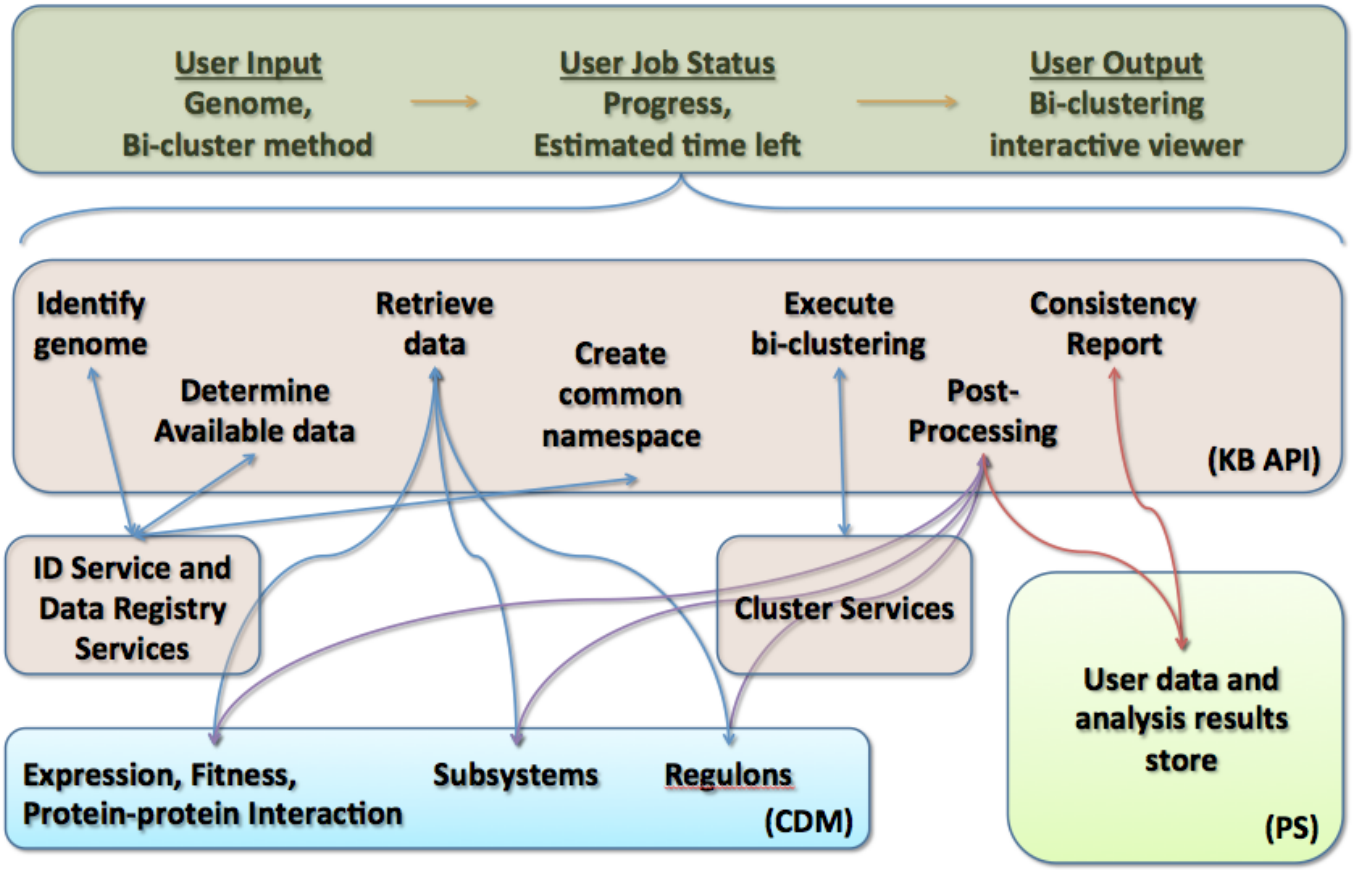
Building a KBase User Experience

- Most of the code development will be layered on top of the KBase Unified API
- Supports a layered architecture consisting of collections of KBase Tools and End User Web Interfaces
- Each layer extends the API by providing additional support for high-level functions



Scientific Workflows

These are the concrete use-cases that drive development of the infrastructure. Workflows from each of the three science domains (microbes, communities, and plants) will be used to demonstrate the capabilities provided by the first production release.



Building a Developer Community

Two ways to develop with KBase:

- Extend from below with additional services that ***extend the KBase API***
- Extend from above by plugging in applications that ***use the KBase API***

A series of system builds occurring every quarter will enable a graded process. Successive builds will expand community involvement.

DEVELOPMENT SCHEDULE

KBASE predictive biology

DOE Systems Biology Knowledgebase

Development Schedule

Feb 2012 – Development release (internal target)

- debug release engineering, prototype deployments, initial data models and data loads, non-unified API, performance testing, architecture refinement

May 2012 – Alpha release (internal target, invited testing)

- draft tutorials, v0.0 database loads, draft API (performance and ubiquity unified prototypes), draft UI library, domain workflow drafts, cloud and cluster services

Aug 2012 – Beta Release (early adopter beta testing)

- workflow function complete, API refinement, v1.0 database loads, prototype plug-in interfaces, prototype galaxy support, performance debugging

Nov 2012 – Production Release Candidate (public beta testing)

- draft website, draft documentation, full functionality API, draft UI v1.0, database loads v2.0, significant number of external beta test users

Feb 2013 – KBase Production Release

- public website, unified API, initial production UI, database loads v3.0 (microbes, community, plant databases), production demonstration workflows, and replication