Open Science Grid in Adolescence: 2012-2016

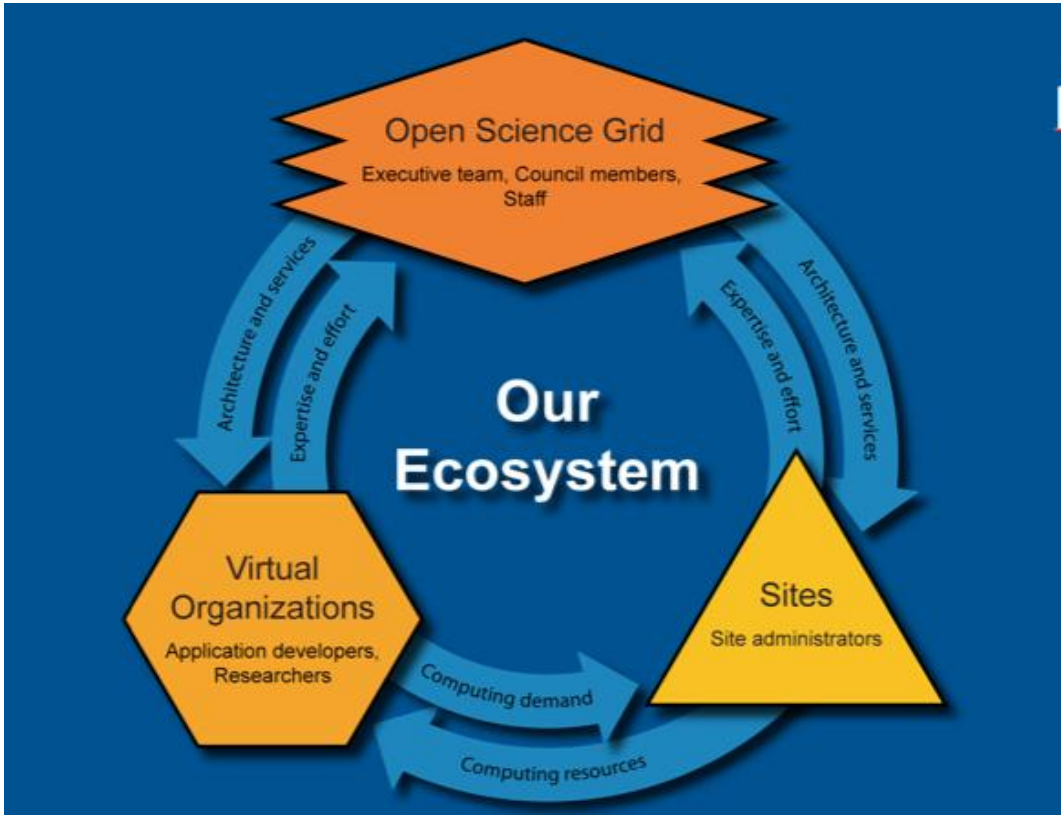
Ruth Pordes
Fermilab

The OSG Eco-System

The Open Science Grid (OSG) advances science through open distributed computing.

The OSG is a multi-disciplinary partnership to federate local, regional, community and national cyberinfrastructures to meet the needs of research and academic communities at all scales.

- Consortium
  - Sites/Resource Providers, Science Communities, Software Providers/Projects, other eco-systems.
- Project
  - staff, deliverables, operations
- Satellites
  - research, development, extensions, loosely coupled



Maturing

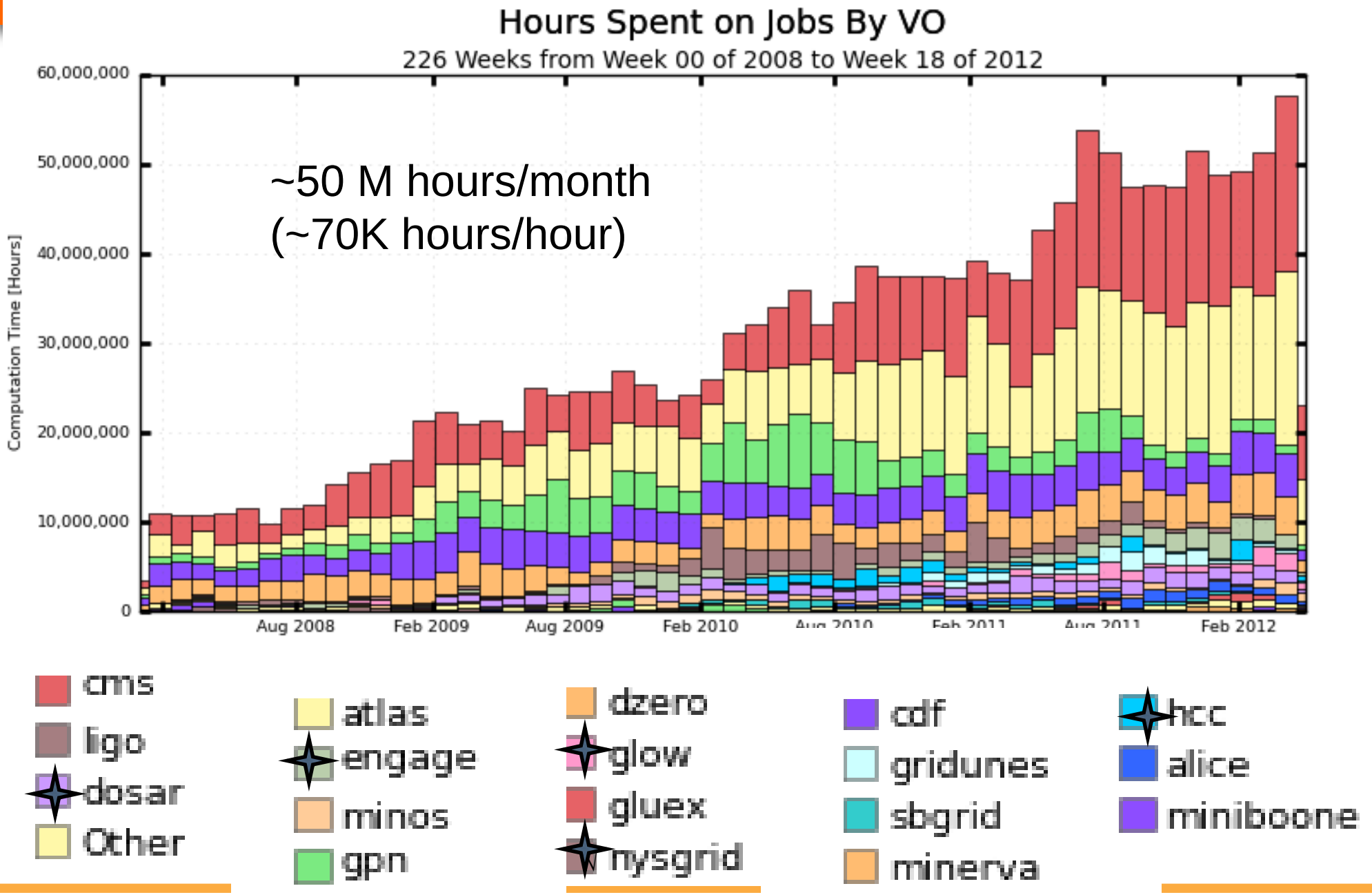
- OSG is being supported for another 5 years.
  - Strong support from DOE and NSF.
- Endorsement to not only continue focus on physics but also continue broad engagement of other sciences making use of OSG.
  - Sustain services for LHC and make significant contributions to LHC upgrade.
  - Extend, simplify, and adapt services and technologies for other sciences.
  - Continue community partnerships and facilitation of peer infrastructures in Europe, South America, Asia, Africa.

Accepted in the National Landscape

- Regarded as Effective Partner by Funding Agencies
  - A Service Provider in NSF XD- XSEDE.
  - Strong NSF OCI funding for growth into the Campuses.
  - Extend support for collaborative extreme-scale science in DOE.
  - Accept responsibility for “orphaned” software and services.

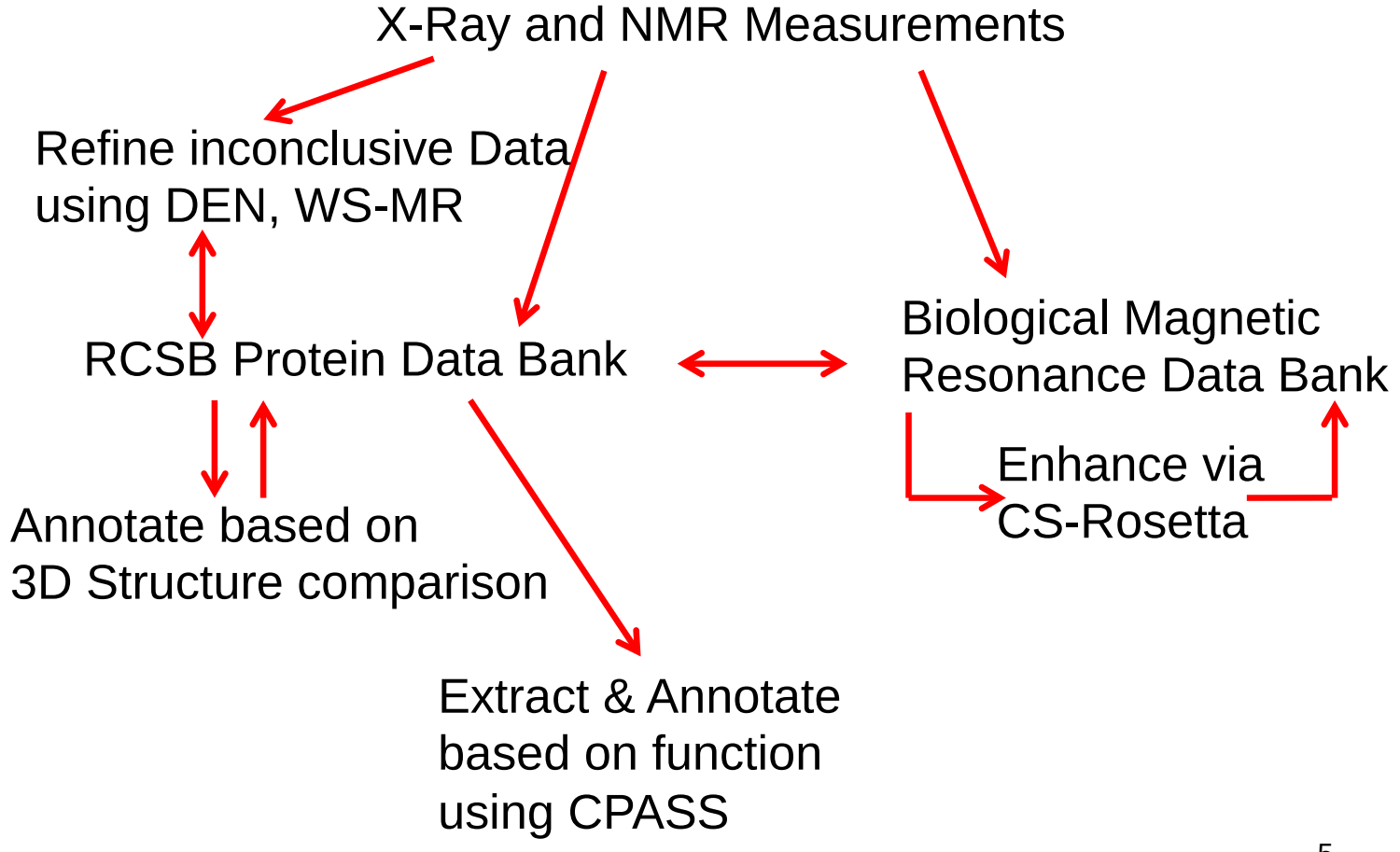
Growing In Size

Growing in Use



25% usage by non-HEP*
e.g. Protein Structure - a Key Application

Protein Structure From Experiment



Diverse Communities

Protein structure determination using CS-Rosetta at the BioMagResBank

U of Wisconsin Madison

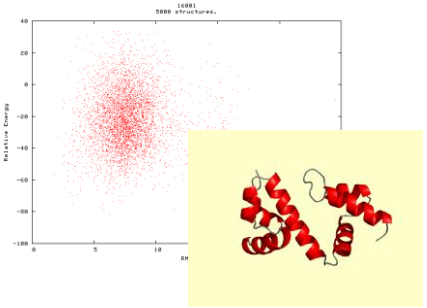
CS-Rosetta uses Monte Carlo simulations based on Nuclear Magnetic Resonance (NMR) data and a database of known structures to predict the structure of proteins.

Determining these structures through lab work alone is very time consuming and not always possible.

The question: how accurate is CS-Rosetta and which types of proteins does it work best on?

OSG allows us to run CS-Rosetta simulations on the entire BMRB database (a public database which holds NMR data). This allows us to study the predictive validity of CS-Rosetta as well as provide additional information for users studying proteins in our database.

Without OSG, there would be more new data submitted to the BMRB than CPU hours available to run the simulations and we would never be able to complete this project!



Jon Wedell - BioMagResBank

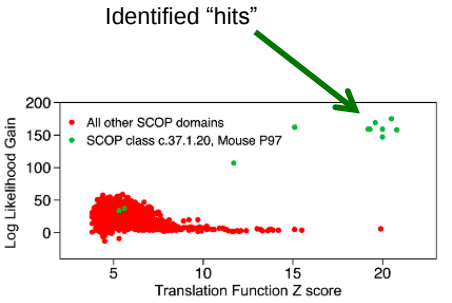
X-ray Crystallography: Wide Search Molecular Replacement

Harvard Medical School

WS-MR success- fully identifies the closest structural homologues from a large family of candidates.

WS-MR is used to determined new X-ray structures using phasing information from identified structural homologues.

Individual WS-MR search is performed on ~100,000 domains derived from the Protein Data Bank.

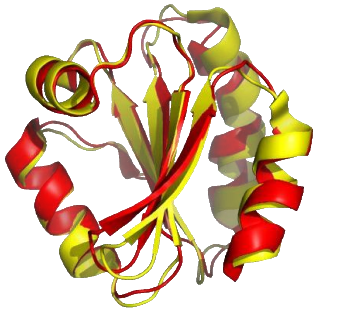


Stokes-Rees, I., and Sliz, P. (2010). Protein structure determination by exhaustive search of Protein Data Bank derived databases. PNAS. *107*, 21476–21481.

6

Calculating Pairwise Similarities of Proteins in RCSB Protein Data Bank (PDB)

- Detection of similarities in protein structure is important to infer functional and evolutionary relationships between protein families.
- A 3D structure comparison of 140 Million pairs is under way, resulting in annotation of PDB information that is then available to the biomedical science community worldwide.
- PDB is used by 200,000 unique



Protein Structure from Computation

U of Washington
U of Chicago

- Given an amino acid sequence, e.g.,

MDPNCSCAAAGDSCTCANSCTCLACKCTSCK

we can use computation to predict protein folding into a 3D structure.

- Long history: more than 30 years as a "grand challenge" problem in computing
- Useful for Drug design, Enzyme design, Function annotation, Target selection
- On OSG, we presently have two distinct efforts that do this:
  - Baker Lab using Rosetta (see separate talk)
  - F.Zhao using RaptorX (see next slide)

SDSC
RENCI

lic (PDB/SDSC), C.Bizon (RENCI)

OSG Leadership Strengthened

- The Leadership Team remains in place and extends:
  - Miron, Michael Ernst, Frank.
  - Lothar as Executive Director.
  - Ruth as Council Chair.
  - Alain, Brian, Chander, Dan, Mine, Rob with significant responsibilities and authorities as Area/Team Leads.

- Added new/old faces:
  - Von Welch as lead of the OSG PKI / Identity Management project (more later).
  - Computer Scientists at ISI - Ewa, Mats - and Notre Dame – Doug Thain – increasingly in the conversations.

An Anchor for Distributed High Throughput Computing (DHTC)

Blueprints, Outreach

Satellite Projects

Technology Investigation, Software

Integration, Production, Operations

OSG

Science Needs

Computer Science Research

Technology and Software

Deploy, Commission and Us

Job Overlay Infrastructure in place - Key Usability Improvement

- Approaching 100% job overlays for all Vos
  - >50% of OSG cycles through glideinWMS,
  - ~25% through Atlas PanDA WMS
- tremendous boost to usability, usage and effectiveness
- Easier for VOs to access cycles, global queue to implement priorities
- Site failures are not seen by the end user

- OSG provides distributed operations
  - "Glidein Factories" at IU, CERN and UCSD
  - perform routine maintenance on the Factory as well as monitor glideins to ensure they are running on sites without error
- Provide future path to flexibly provision compute resources
- Supports access to clouds including Amazon EC2

Sustain Services & Expand Science Communities

- LHC
  - Continued focus on LHC – support for ATLAS, CMS, ALICE USA distributed computing in the US.
  - Active /proactive contributions on behalf of US LHC to WLCG – to TEG reports and implementation follow ons.
  - Prepare for LHC shutdown and upgrade.

- Embrace future physics, nuclear physics, astrophysics experiments: Belle II, DES, EIC, LSST, SuperB…

(will explain these..)

Transition of some support from External Projects

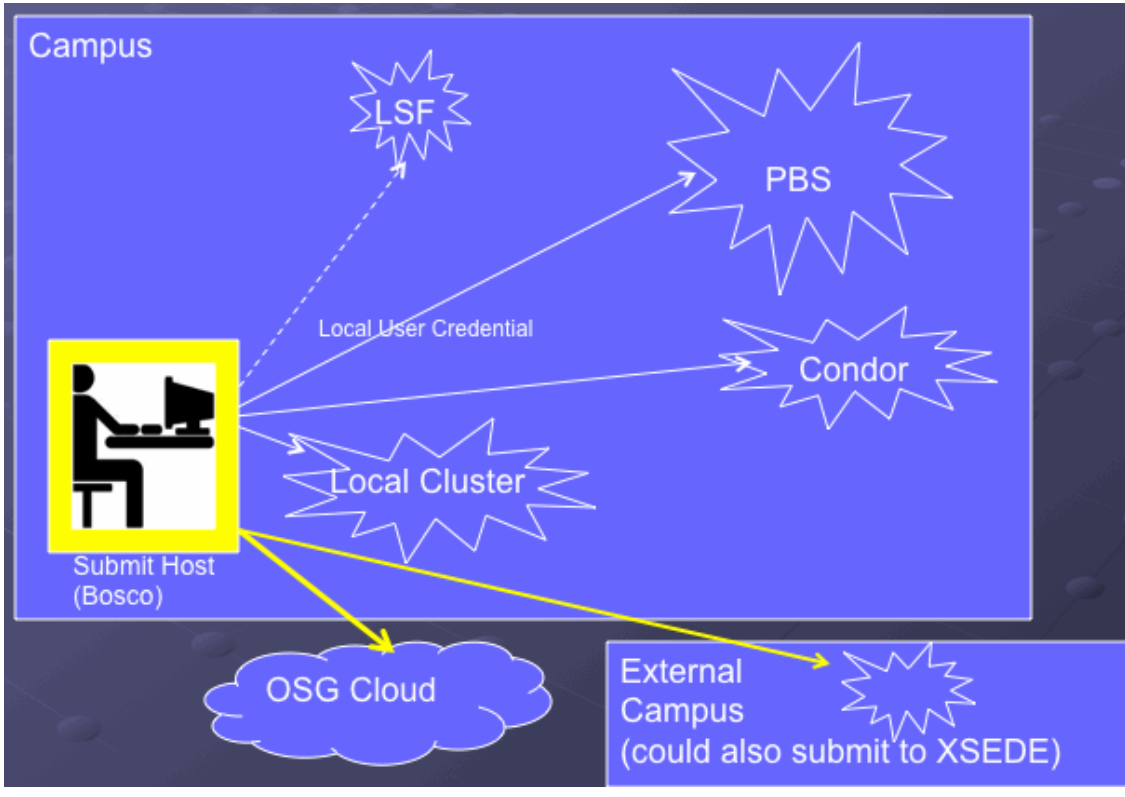
- Bestman support now under the OSG Software Area.

- ESNET DOEGrid PKI infrastructure/Certificate Authority -> OSG
  - Receiving additional support from ESNET to implement.
  - Need to have no disruption in production service.
  - Deploy OSG PKI infrastructure for User and Host/Service certificates based on DigiCert commercial backend by Oct. 2012.
  - US ATLAS and US CMS contributing to the 1st 2 year license to a defined threshold. It is not yet known if this is sufficient.
  - Feb. 2013 DOEGrids ceases to issue new certificates.
  - Feb. 2014 DOEGrids ceases to publish CRL lists.

- Understanding of support needed for Globus Toolkit
  - GRAM, GridFTP

Extending support in the Campuses

- Intra- and inter- campus sharing of resources.
- US LHC Tier-3s "beachheads" for their campus.
- Support the revolution in biology - - - sequencer on every desktop.
- Integration of popular applications frameworks e.g. "galaxy".



Interface into New Resource Types

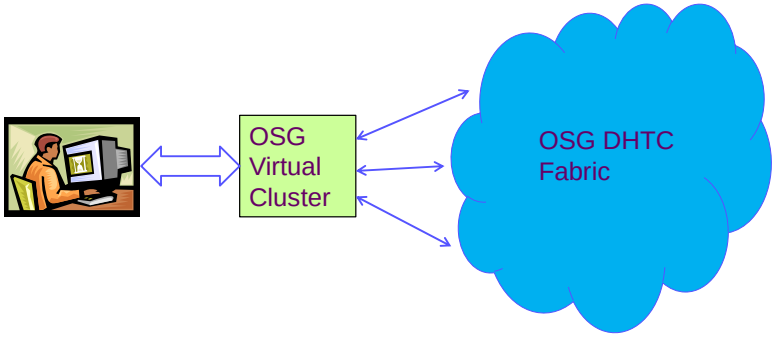
- High-core count nodes
  - Condor moving to support: Resource reservation, Access to partial nodes, Automated node draining.
  - Applications being adapted.
  - Operations/administrative services being extended.
- Support for applications using GPUs
- Science Clouds
- Commercial Clouds – Amazon EC2

Data Services for non-LHC experiments

- Large communities typically have their own implementations and deployments of Data Services.
- XROOTD deployment (Anydata Anytime Anywhere project) scope includes generalizable Data Access/Remote I/O services for use by other OSG communities/
- Globus Online being evaluated by several VOs.
- IRODS being readied for deployment to manage shared storage/data spaces.

Part of the eXtreme Digital (XD) Generation

- As an XSEDE Service Provider support users given allocations by the (NSF supported) XSEDE request and allocation process.
- First users in summer 2012 through dedicated Submit interface.
- By 2015 expect this to be in full production and to support multi-stage workflows across OSG and other XSEDE Service Providers.



XSEDE users "login" to the "OSG Virtual Cluster" that provides an abstraction layer to access the distributed OSG fabric. This interface allows XSEDE users to view the OSG as one resource where they submit their jobs, provide the inputs and retrieve the outputs.

Looking towards and beyond 2015 – Computer Science Research

- OSG's existing capabilities are effective but basic and primitive.
  - Improvements will rely on external research, development and contributions.
- Integrate static resources with dynamically allocated resources (like clouds).
- New globally capable, usable, and integrated frameworks for collaborative environments : data, security, workflows, tools for transparency, diverse resource resources.

- http://osg-docdb.opensciencegrid.org/0011/001106/001/OSG-CSresearchNeeds.pdf

Summary

Offers to participate in this unique adventure welcomed!