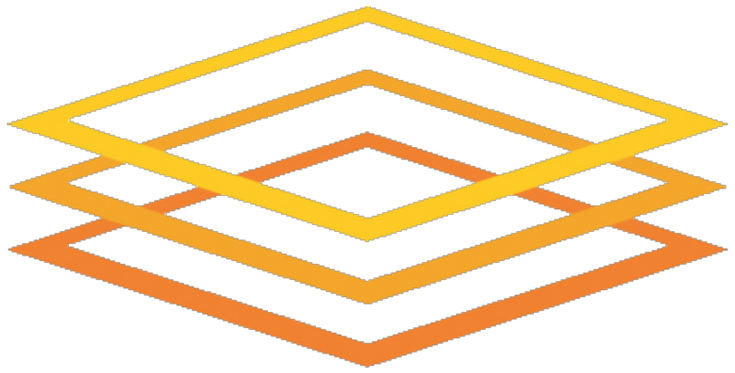
Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science



The Open Science Grid

Lothar A. T. Bauerdick, Fermilab

SuperComputing 2014

Nov 19, 2014

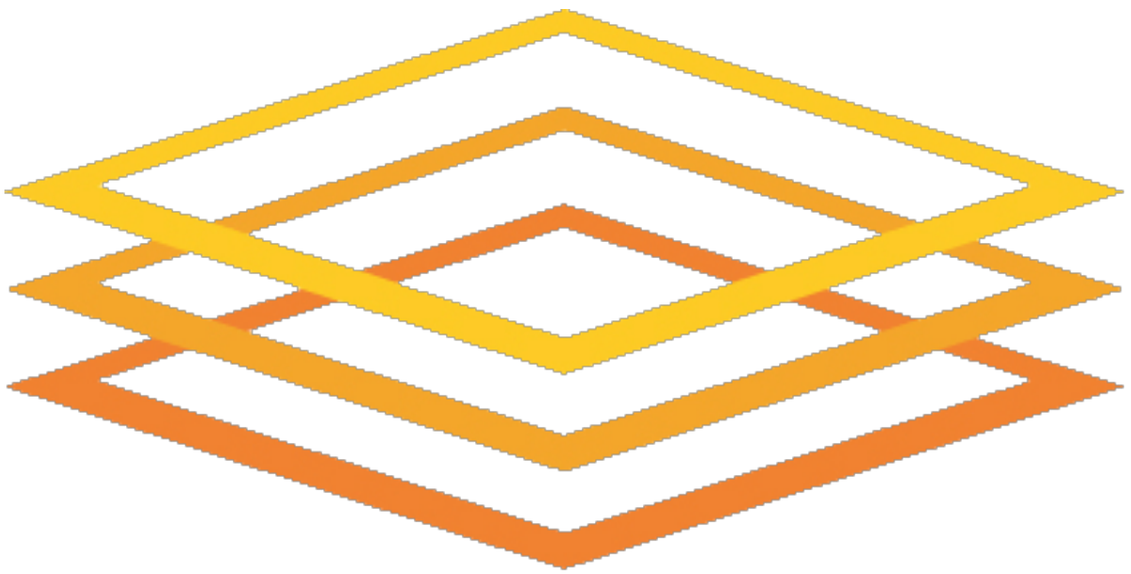
What is the Open Science Grid?

Open Science Grid

Mission: *The Open Science Grid aims to promote discovery and collaboration in data-intensive research by providing a computing facility and services that integrate distributed, reliable and shared resources to support computation at all scales.*

- **OSG is a Consortium**
  - Resource owners and campuses, scientist and research users, computer scientists and software providers, national and International partners
- **OSG is a Project**
  - Provides a fabric of services across contributed resources
  - Half into a 5-year renewed OSG project, after 6-years that accomplished much and taught us even more about what is lacking and needed
- **OSG is an Eco System**
  - Provides a framework for exploring ways of scientific discovery through the use of distributed high throughput computing
  - Domain and computer scientists collaborating for more than a decade
  - Contributing to state of the art through innovation and collaboration

Open Vision of the Open Science Grid



- Open Science
- Open Facility
- Open Software Stack
- Open Ecosystem

Open Science

Open Science Grid

- **Benefit Science in the broadest possible way**
  - Diversity in Scientific Disciplines
  - Diversity in Scale
    - Single PI to global community
    - Accelerate transition to x10 larger scale irrespective of scale the Science is at today.
  - Commonality in Method:

Distributed High Throughput Computing

the shared utilization of large ensembles of autonomous resources toward a common goal, where all elements are optimized for maximizing throughput

A diverse Fabric of Services

Open Science Grid

- **OSG platform and ecosystem of DHTC capabilities**
  - enables Virtual Organizations to run workflows and data storage systems across all OSG sites (essential for LHC and other large VOs)

- **OSG Direct** HTC facility of harvested CPU resources
  - use *job manager overlays* to effectively harvest CPU resources opportunistically from OSG sites, that stakeholder VOs would otherwise leave idle

- **OSG Connect** platform for Campus Grids “as a service”
  - provides a login and group management service for VOs and individuals with HTC workloads
  - connect campus users, effectively providing **Campus Grids** as a service

OSG in Numbers

Open Science Grid

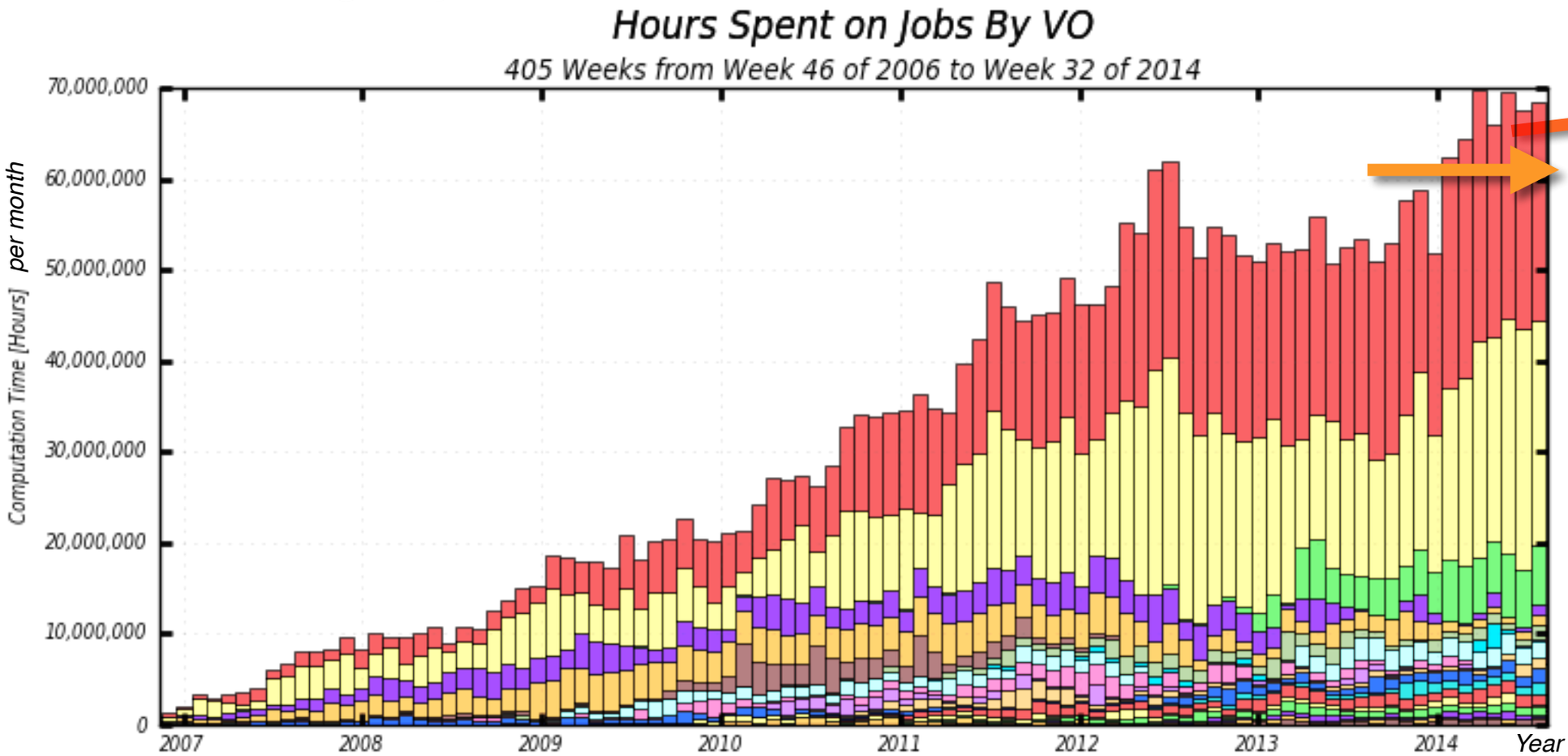
- OSG Delivers up to 2.5 Million CPU hours every day
  - ✦ almost 750M hours of Distributed High Throughput Computing per year
    - about 60% go to LHC, 20% to other HEP, 20% to many other sciences
- OSG has a footprint on ~120 campuses and labs in the U.S.
  - ✦ Supports active community of 20+ multi-disciplinary research groups

rmilab

OSG "Size" is Increasing in Terms of CPU Resources

Open Science Grid

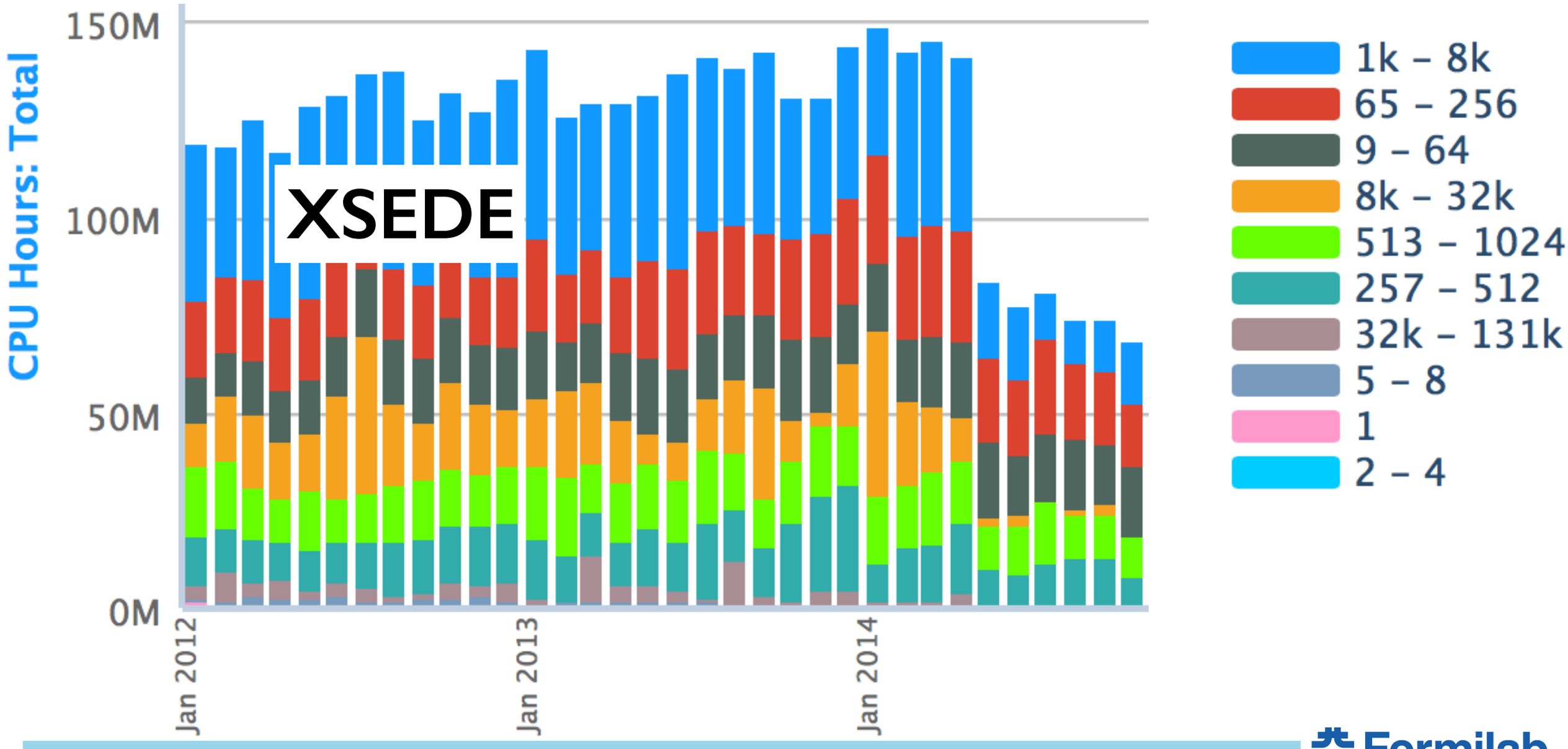
- Since 2007, 3.1 Billion CPU hours delivered, more than a billion jobs run!
- LHC plans increase in CPU resources at U.S. Tier-1 and Tier-2 over coming 2 years



Fermilab

CPU-hour/month comparison with XSEDE

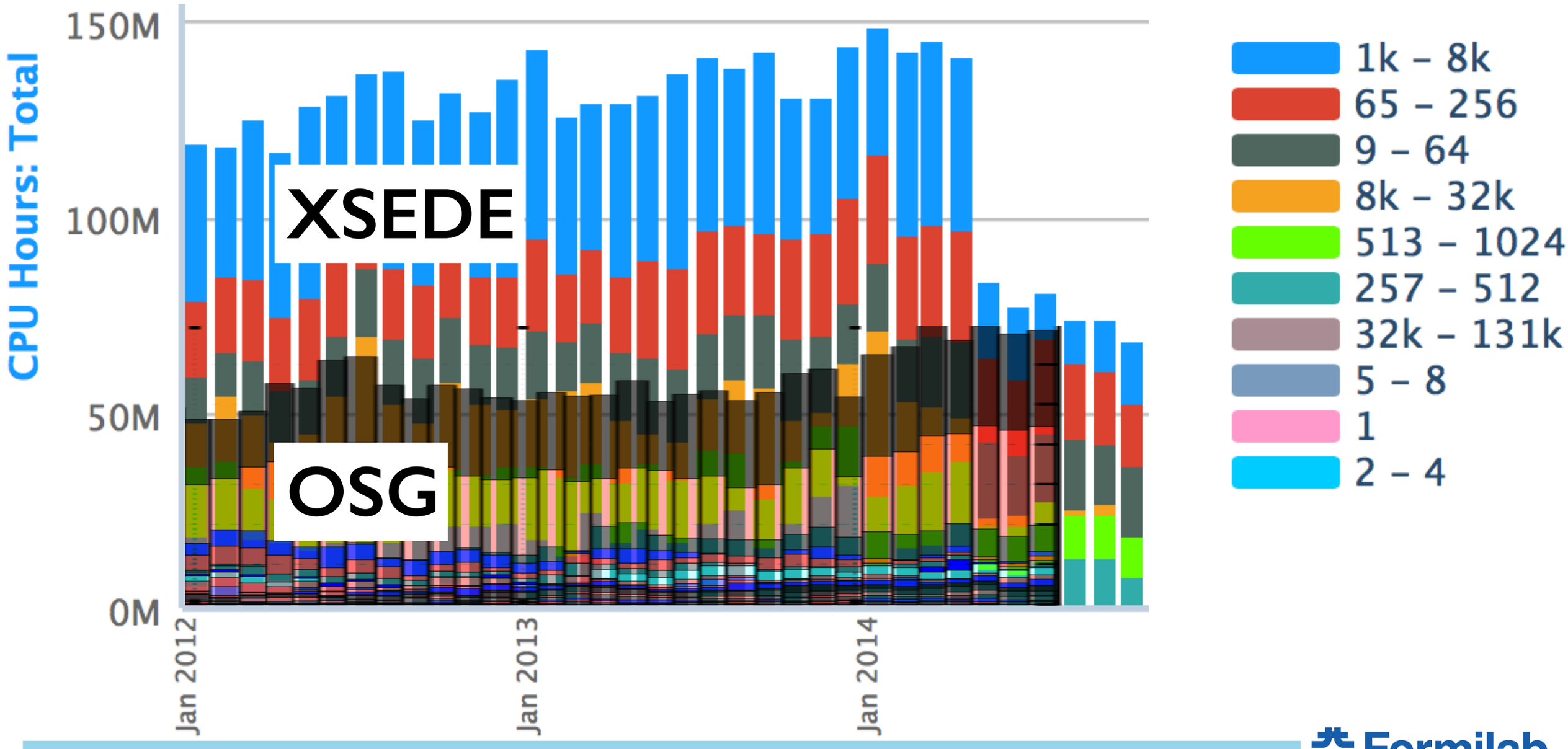
Open Science Grid



CPU-hour/month comparison with XSEDE

Open Science Grid

- OSG High-Throughput Computing CPU hours provided ~ XSEDE CPU hours for the “smaller” job sizes

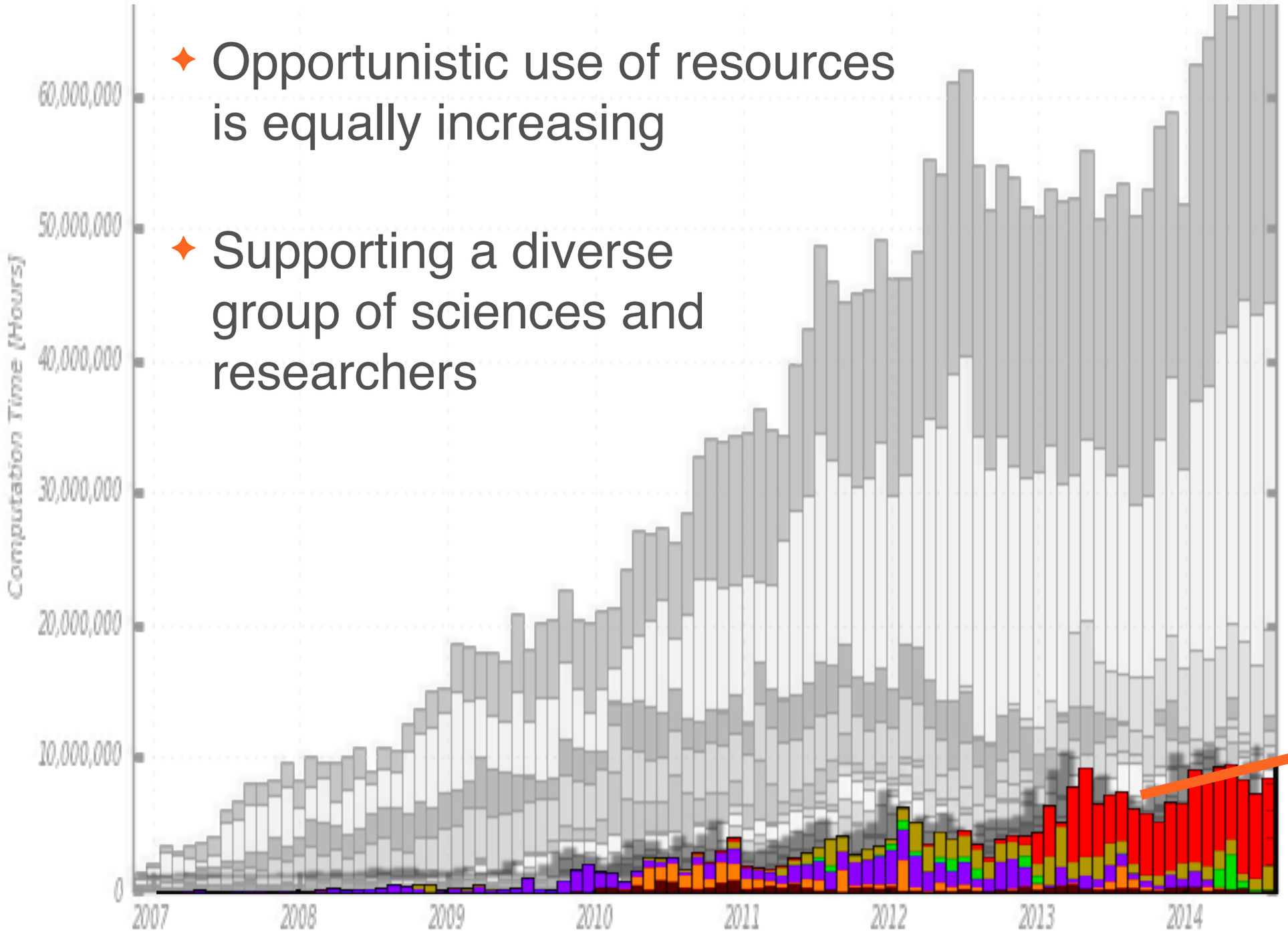


Fermilab

Growing Use of "Owned" and of "Opportunistic" Resources

Open Science Grid

- Opportunistic use of resources is equally increasing
- Supporting a diverse group of sciences and researchers



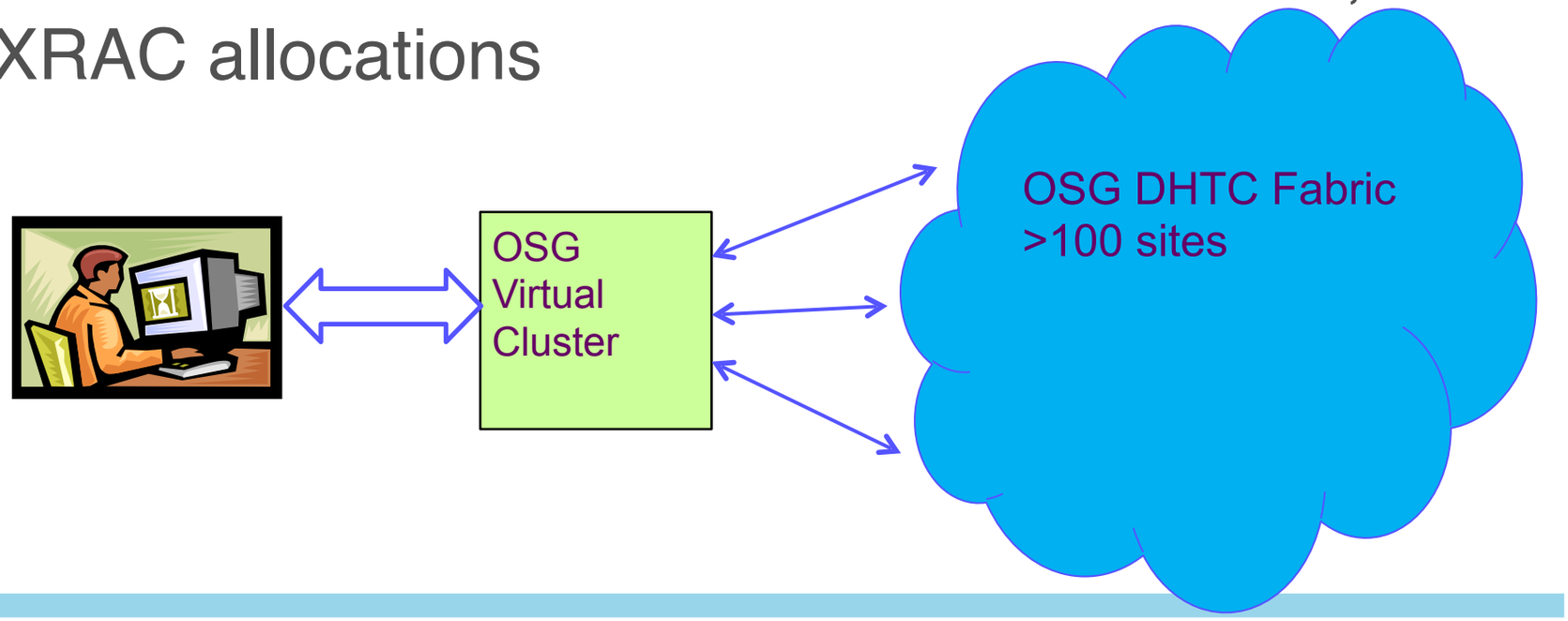
Opportunistic Use?

>100M CPU hours opportunistic use past 12 Months

Open Facility

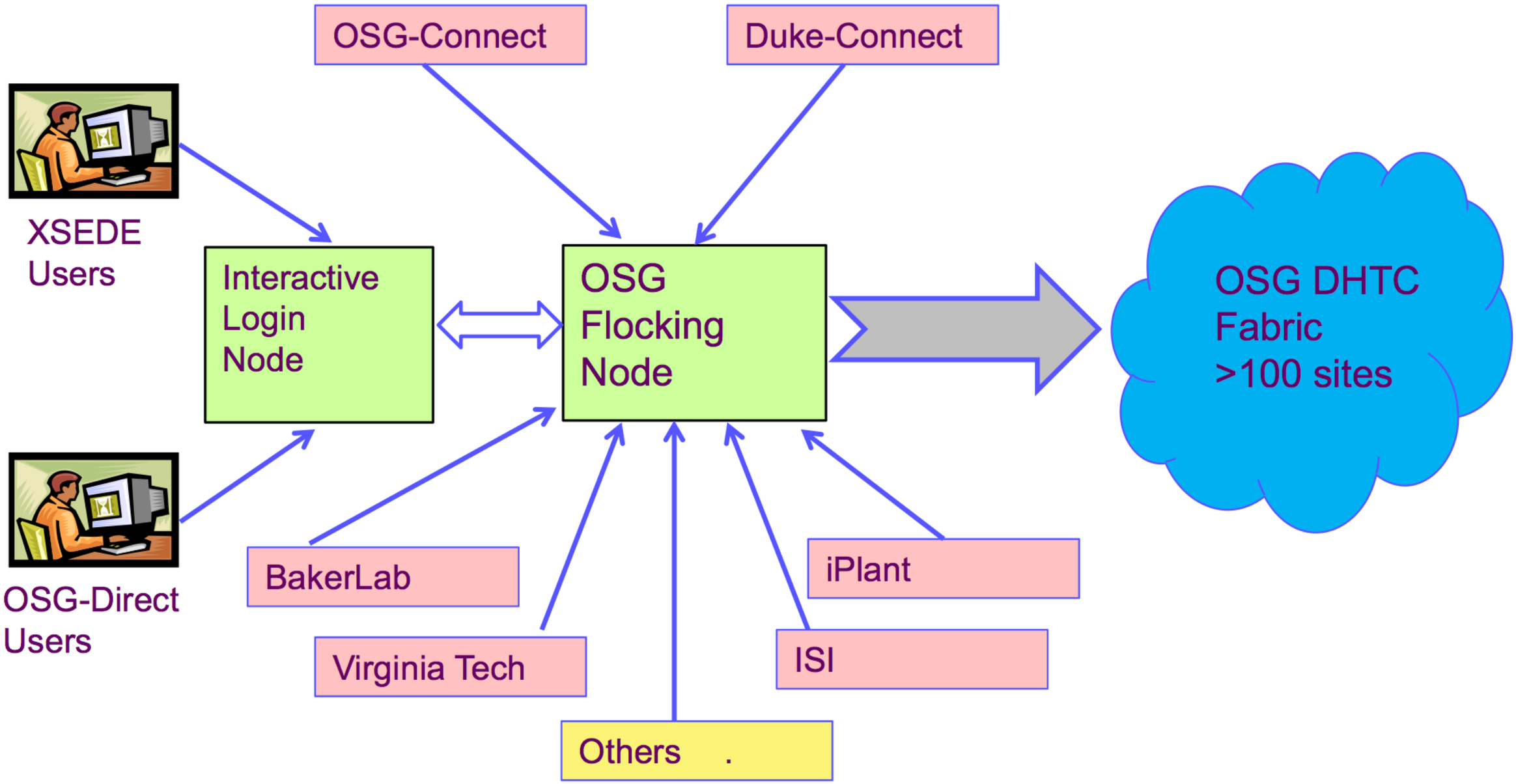
Open Science Grid

- A Distributed High-Throughput Computing Facility based on harvesting otherwise idle resources
  - **job manager overlays** using capabilities of HTCondor and GlideinWMS
  - effectively harvest CPU resources opportunistically from all OSG sites
- Available to a large and diverse community of researchers
  - backend for specialized science gateways e.g. for biology and medical applications using web portals such as Galaxy
  - is basis for OSG as an Level-2 XD Service Provider, access for PIs through XRAC allocations



Access to OSG Fabric via OSG-VO

Open Science Grid



All access operates under the OSG VO using glideinWMS

The OSG Harvesting Free CPU Cycles :-)

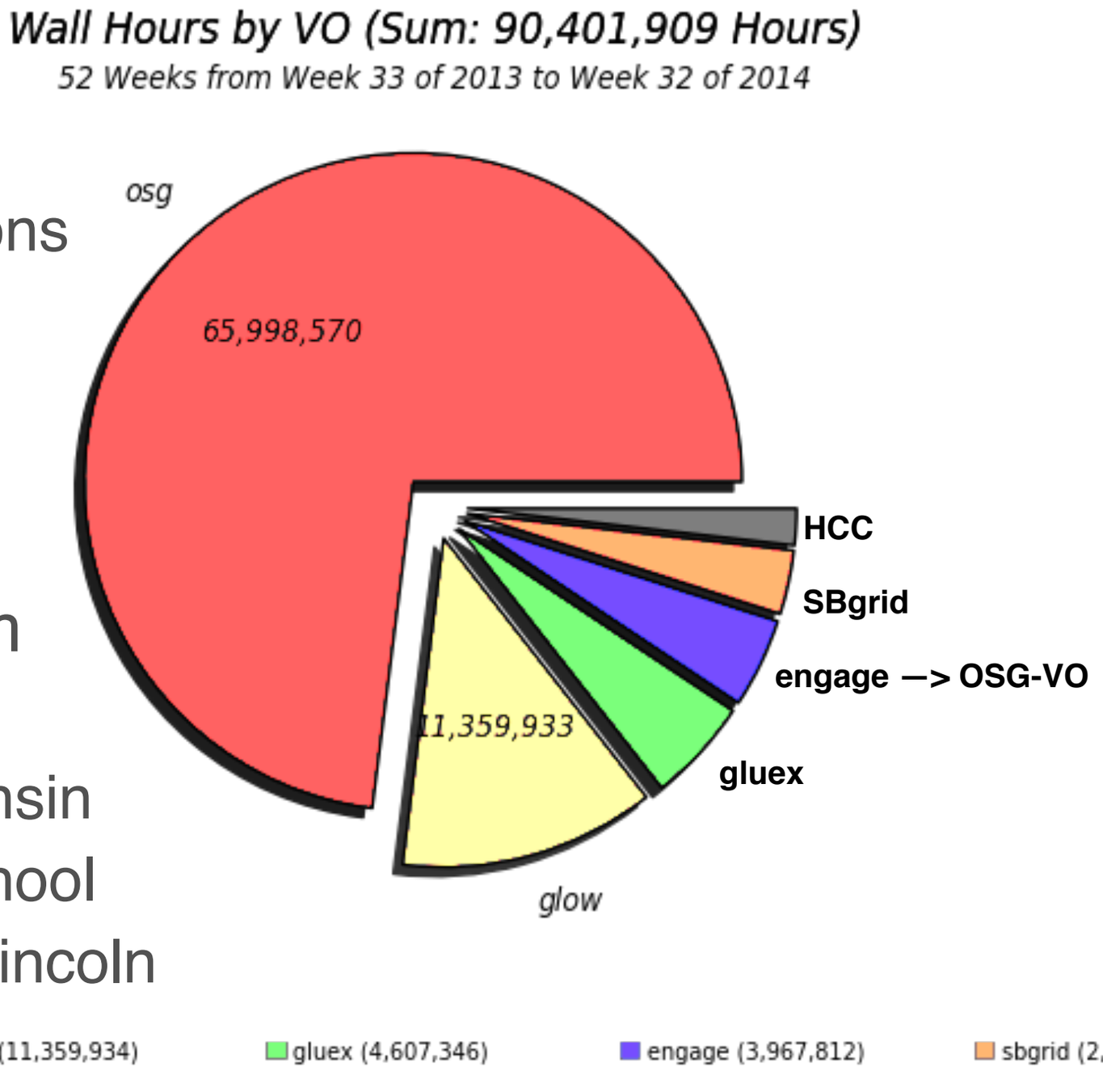
Open Science Grid

Fermilab

Use of OSG Direct "Open Facility"

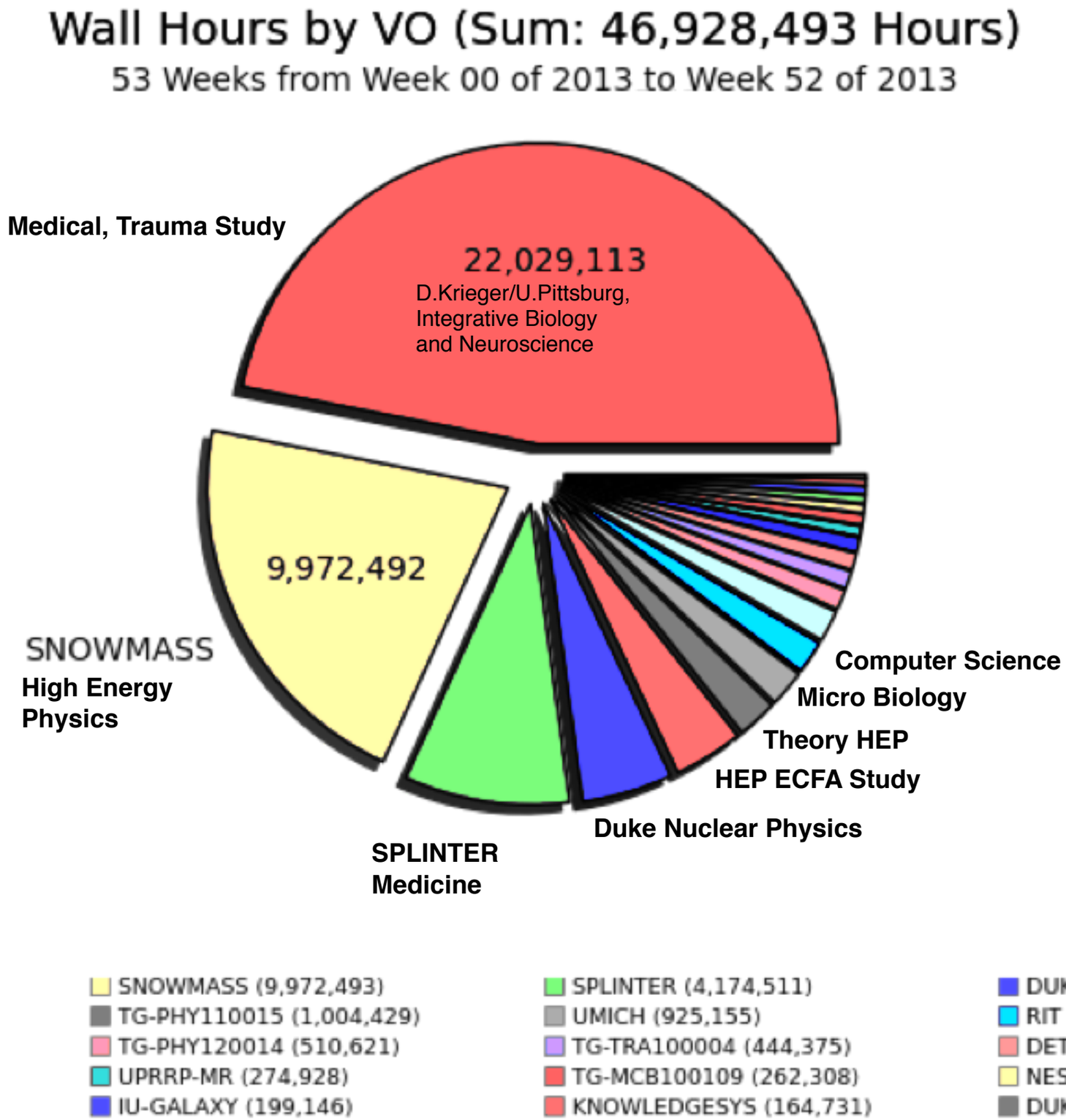
Open Science Grid

- OSG VO:
  - NSF XD Service Provider
    - about 30M hours
  - a number of science applications
  - OSG Connect users ("retail")
  - individual science groups, who also come in through the engage VO
- Users flowing into OSG from campus grids:
  - Glow campus grid at U.Wisconsin
  - SBgrid at Harvard Medical School
  - HCC campus grid at Omaha/Lincoln



OSG Direct "Open Facility" Science Users in 2013

Open Science Grid



Operating the OSG

Open Science Grid

- **Distributed Operations Team** w/ Grid Operations Center at IndianaU
  - running a world-class unique and diverse set of services, enabling more than 100 sites!
    - to provide the OSG platform/eco system of DHTC services, sites, software to enable VOs to run workflows and data systems across OSG sites
    - includes infrastructure services, operations support, cyber security and incident response etc

- **OSG Direct Open Facility**
  - harvesting resources opportunistically from OSG sites, delivering to XD, to science gateways e.g. for biology or medical applications etc

- **Other Services**
  - user and host certificates (OSG CA, the follow-up of DOEgrids CA)
  - software distribution services (OASIS based on CVMFS)
  - network monitoring and dashboard, etc

OSG Support for Users, VOs and Campuses

Open Science Grid

- **OSG User Support**
  - consulting on technologies, architectures and user support
  - spreading knowledge on HTC as a science problem solver
- **OSG Technologies and Software**
  - developing concepts and blueprints, deliver an evolving software stack
  - software “factory”: packaging, system testing, patching
- **Campus Grids**
  - OSG Connect service
  - Campus Infrastructure Community

How About Data?

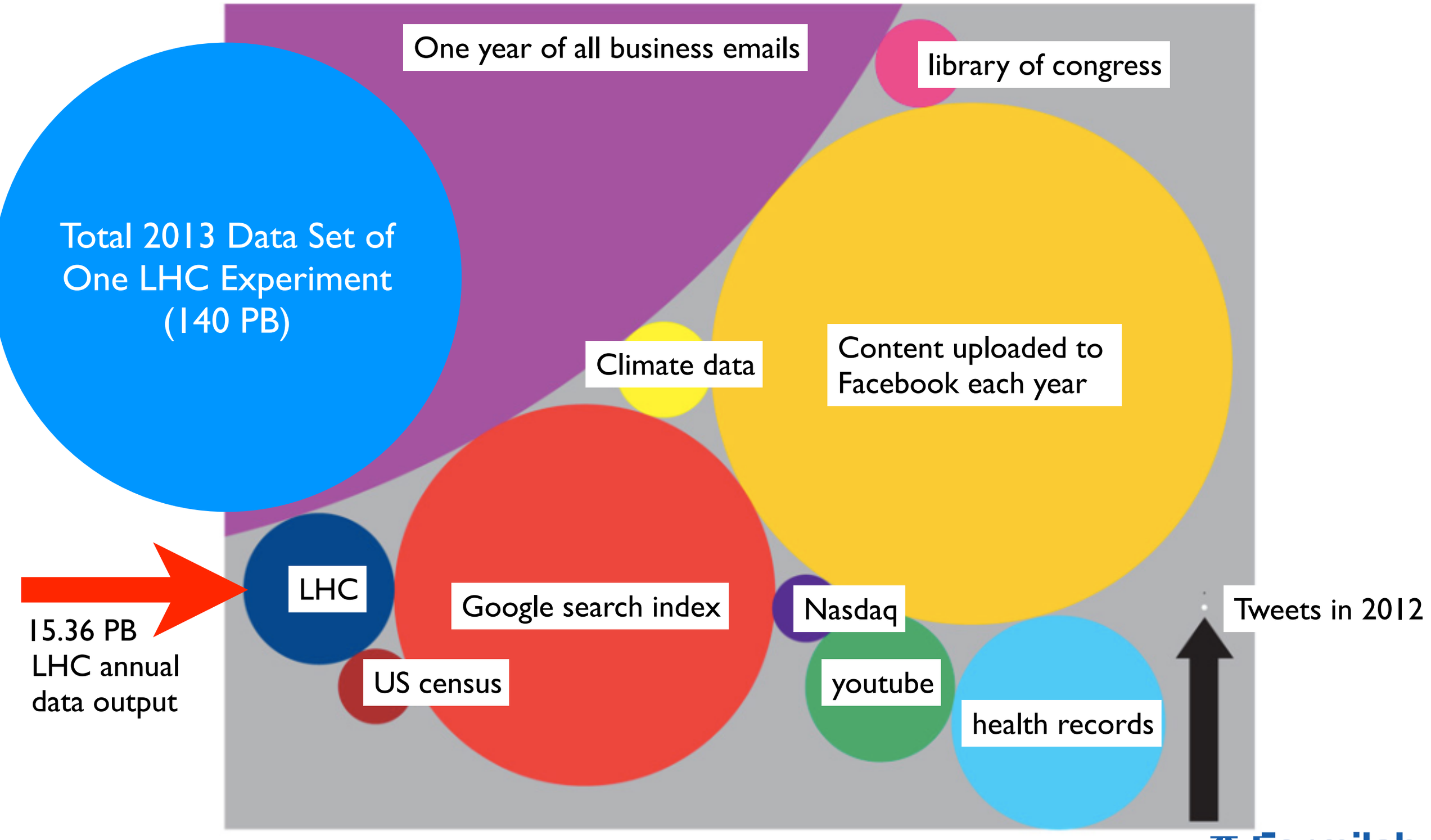
Open Science Grid

- Large VOs like the LHC experiments offer users sophisticated data infrastructures
  - dynamic data placements and high throughput robust data transfers
  - OSG transfers ~1 PetaByte of data every day
  - data discovery and metadata tools

- The real killer feature:
  Connectivity and Network Throughput
  - including high-speed networks to resources, distributed IDM, etc
  - enabling remote data access,
  - federating storage systems
  - allowing global high-throughput access to locally managed data

The Big Data Frontier from Wired Magazine

Open Science Grid



http://www.wired.com/magazine/2013/04/bigdata/

Fermilab

Opportunities for the Long Tail of Science

Open Science Grid

- OSG to prevent growth of a capability gap b/w researchers w/ compute problems at different scales: LHC vs. the **Long Tail**
  - researchers that don't own large resources but profit from DHTC
  - while "exascale" problems need to be solved for the large collaborations, it is equally necessary to ensure solutions are available for the many scientists challenged at the terascale and petascale on their passage to the exascale

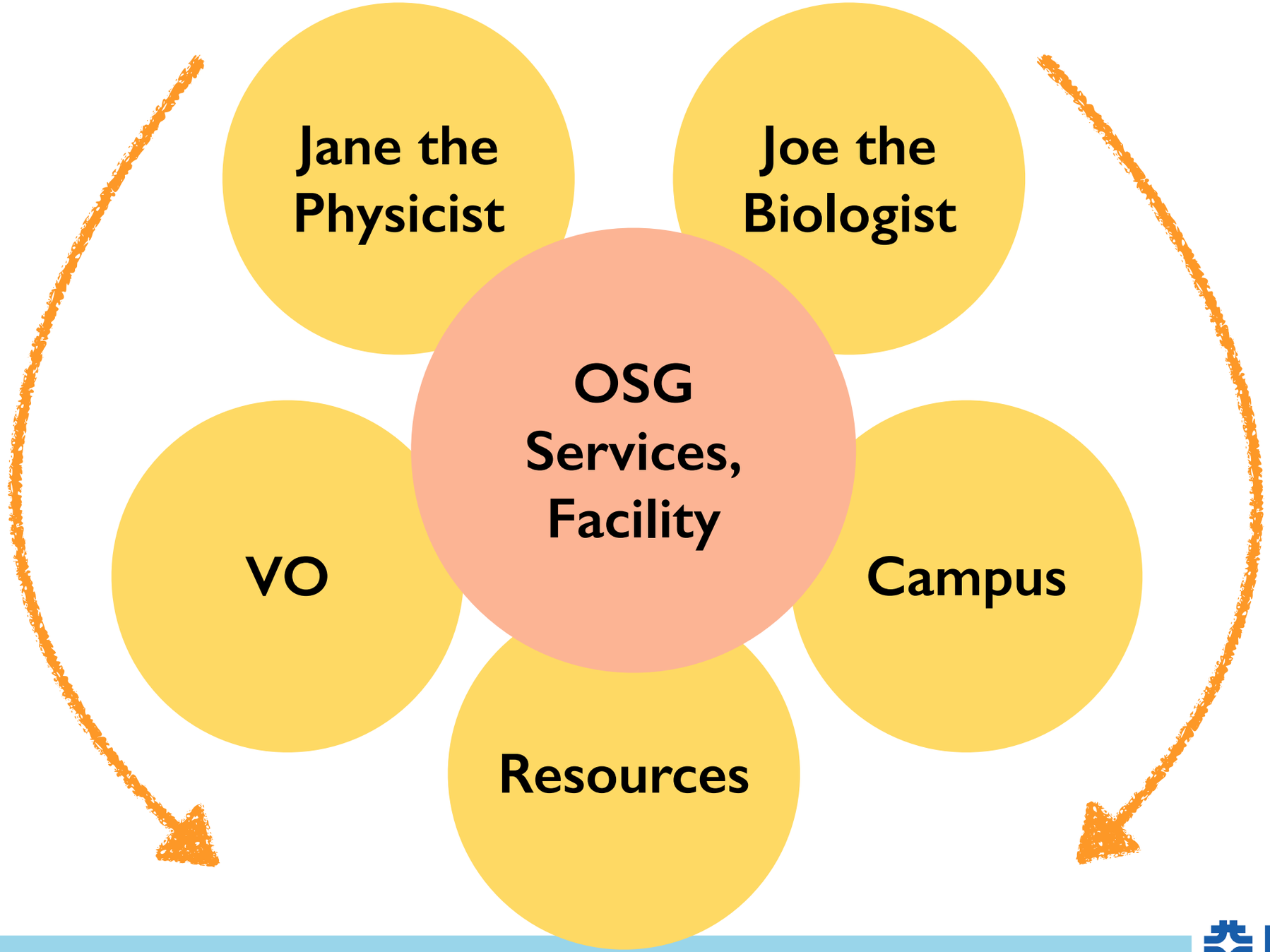
- Keep up w/ increasingly dynamic, heterogeneous environments
  - ensure that domain scientists w/ limited expertise can use them

- **Campuses** to play role filling this gap, supporting researchers
  - for our user community, including LHC, it is crucial we continue to increasingly include the campuses!
  - strategic importance to partner w/ projects supporting campuses

OSG Services and Facility Ecosystem

Open Science Grid

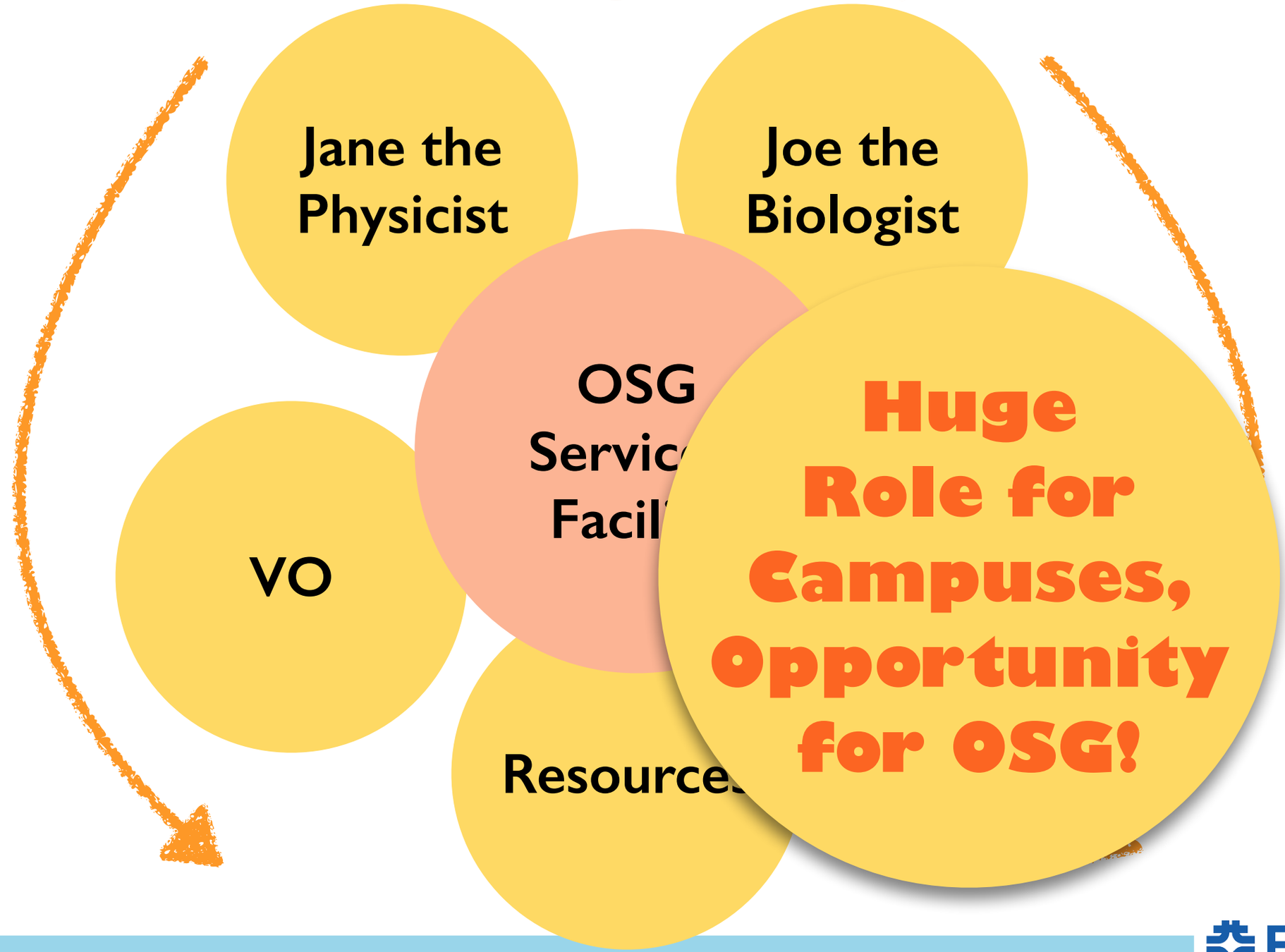
VO-centric Model ⊕ Campus User Model



OSG Services and Facility Ecosystem

Open Science Grid

VO-centric Model ⊕ Campus User Model



OSG Mechanisms to Support single PIs

Open Science Grid

- Campus VOs fully operated by a University IT organization.
- OSG VO providing the deep layer of VO middleware support.
- OSG XD providing an allocation-based entry point
  => reaching new users this way!
- OSG Direct providing an on-ramp to the Open Facility from departmental clusters and workstations
  => submit local – compute global
- OSG Connect = OSG as a service

OSG Connect – OSG as a service

Open Science Grid

- Within the last 5 years we learned that not all Campuses have IT organizations willing and capable of operating the necessary VO middleware layer.

- We thus started to offer OSG as a service that seamlessly integrates campus infrastructure with grid & cloud.

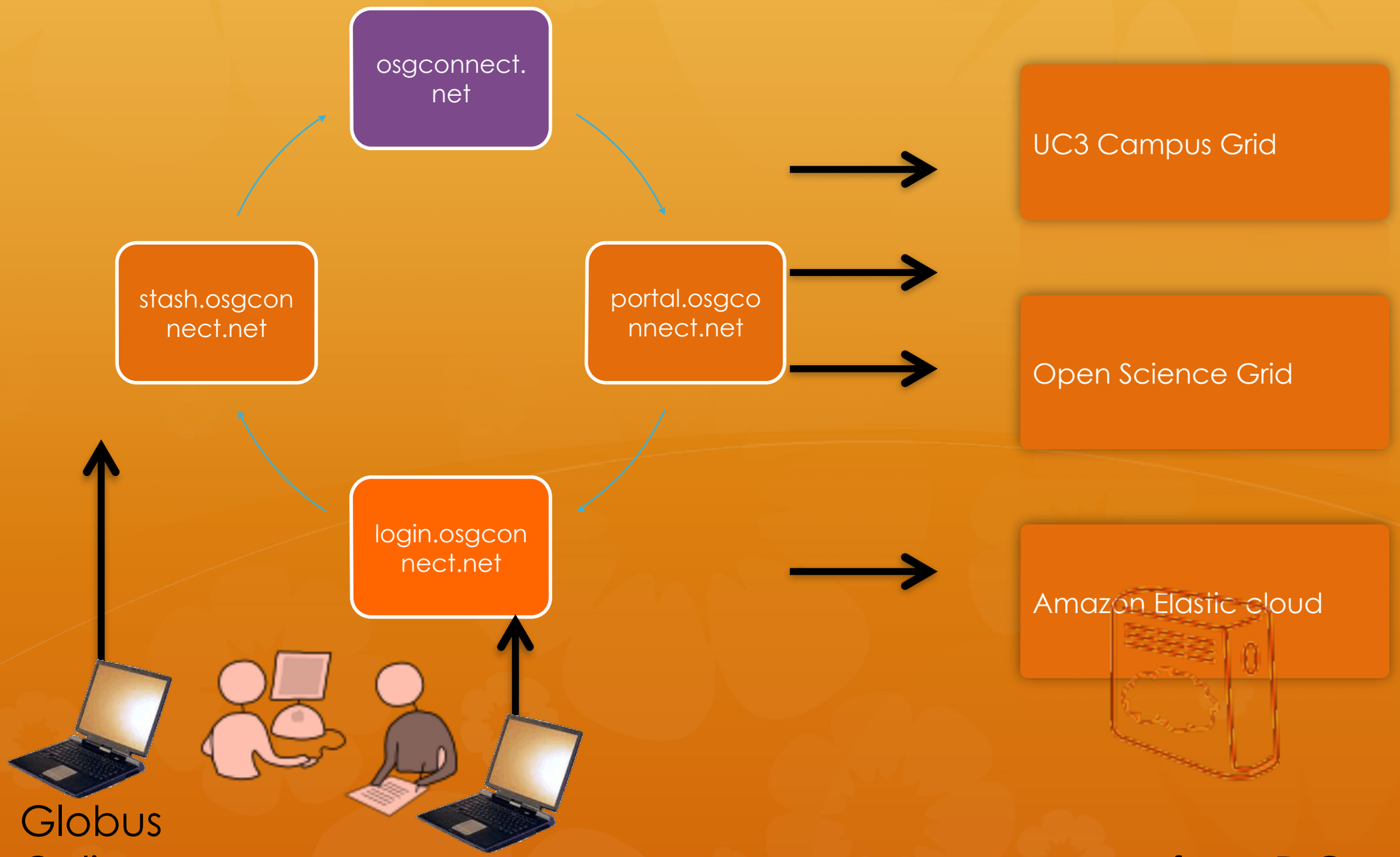
- We offer to run the service for interested Universities.

- We call this “OSG Connect”

This is the most recent addition to our toolbox

An OSG Connect Instance

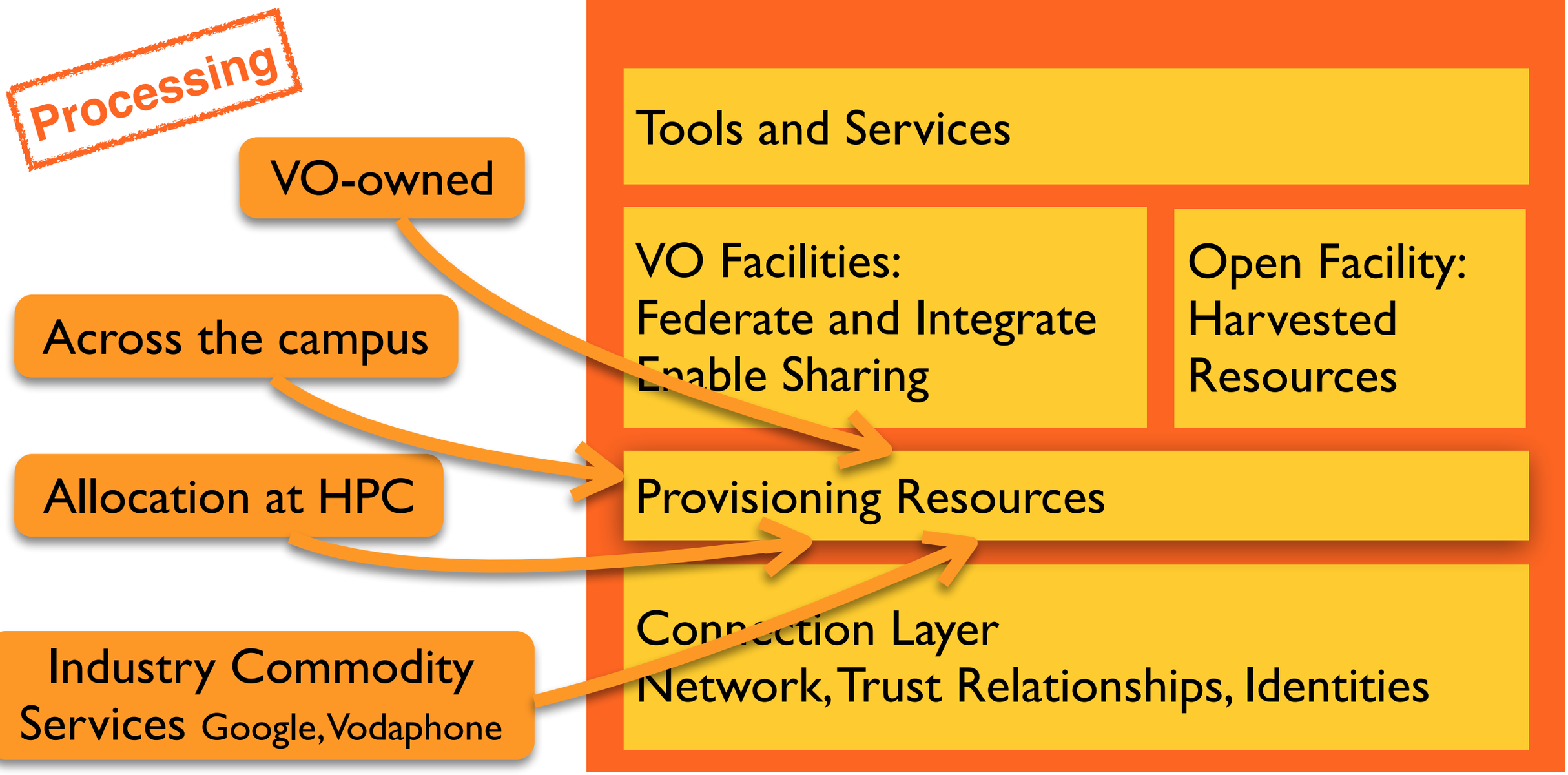
osg connect



from R.Gardner/U.Chicago

Bringing in New Types of Resources

Open Science Grid



- Focus on **Dynamic Resource Provisioning**
  - Statically federated resources need to be integrated with dynamically allocated resources causing new challenges for resource planning, acquisition, provisioning

Moving forward with the Open Vision for the Open Science Grid!

Open Science Grid

- Open Science
- Open Facility
- Open Software Stack
- Open Ecosystem