OSG and the Campus

Open Science Grid

Rob Gardner • University of Chicago
Research Professor of Physics, Enrico Fermi Institute
Senior Fellow, Computation Institute

Towards Security Assured Cyberinfrastructure in Pennsylvania (SAC-PA) CI Cybersecurity Workshop, June 22, 2017

What is the Open Science Grid?

- **Helps researchers speed up their research using high throughput computing methods**
- Helps campus HPC administrators **share resources** for multi-campus and national collaborative research
- Last 30 days: 100M core-hours
- Last 12 months: 200 Million jobs consumed **1 Billion hours** of computing involving 1.5 Billion data transfers to move **>200 Petabytes**
- Accomplished by **federating 114 clusters** providing 1h-100M hours each

OSG is Open to All

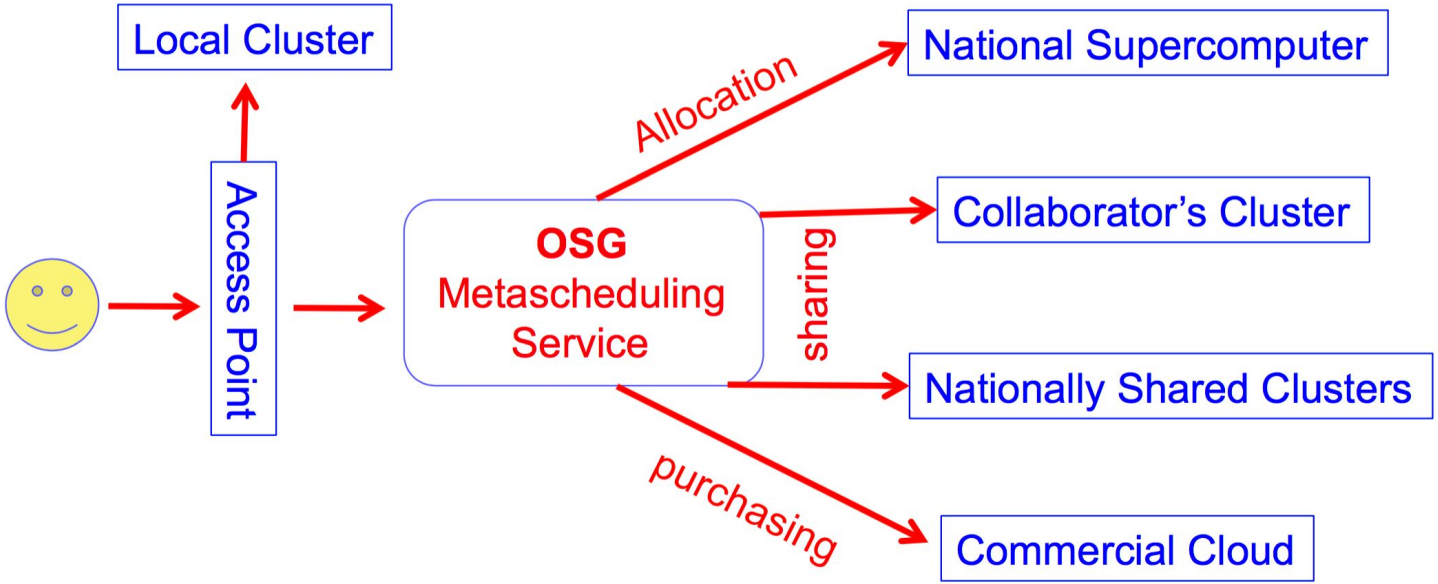
- Open to providers at all scales
  - from small colleges to large national labs
- Open to user communities at all scales
  - from individual students to large research communities
    - domain science specific and across many campuses
    - campus specific and across many domain sciences
- Open to any business model
  - sharing, allocations, purchasing
  - preemption is an essential part of operations

OSG Magic

We create a uniform environment across a heterogeneous set of resources that is distributed globally

Submit locally – Run Globally

OSG supports computing across different types of resources



Seamless integration is they key to our success!

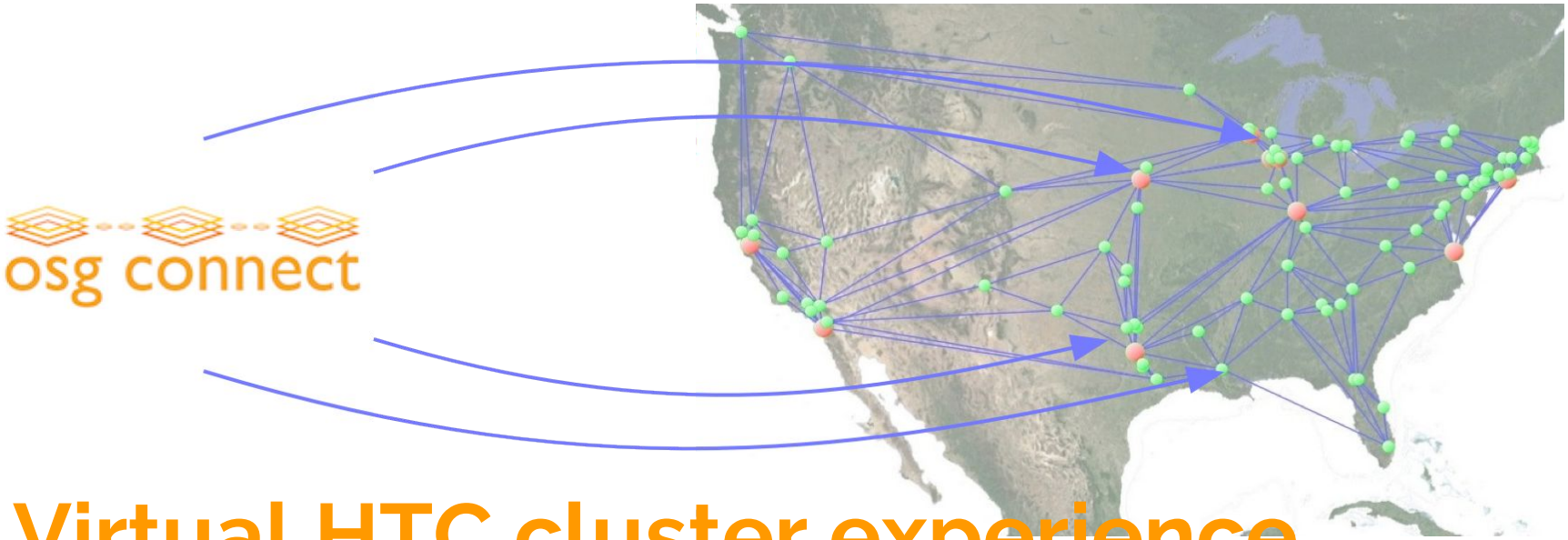
OSG Tools to Match Diversity of Scale

- OSG Connect
  - OSG hosts the service on OSG hardware
- OSG Cluster in a Box
  - OSG manages services on hardware placed inside campus SciDMZs
- OSG Compute Element
  - Gateway software that campuses deploy or OSG hosts

In all cases seamless integration is key

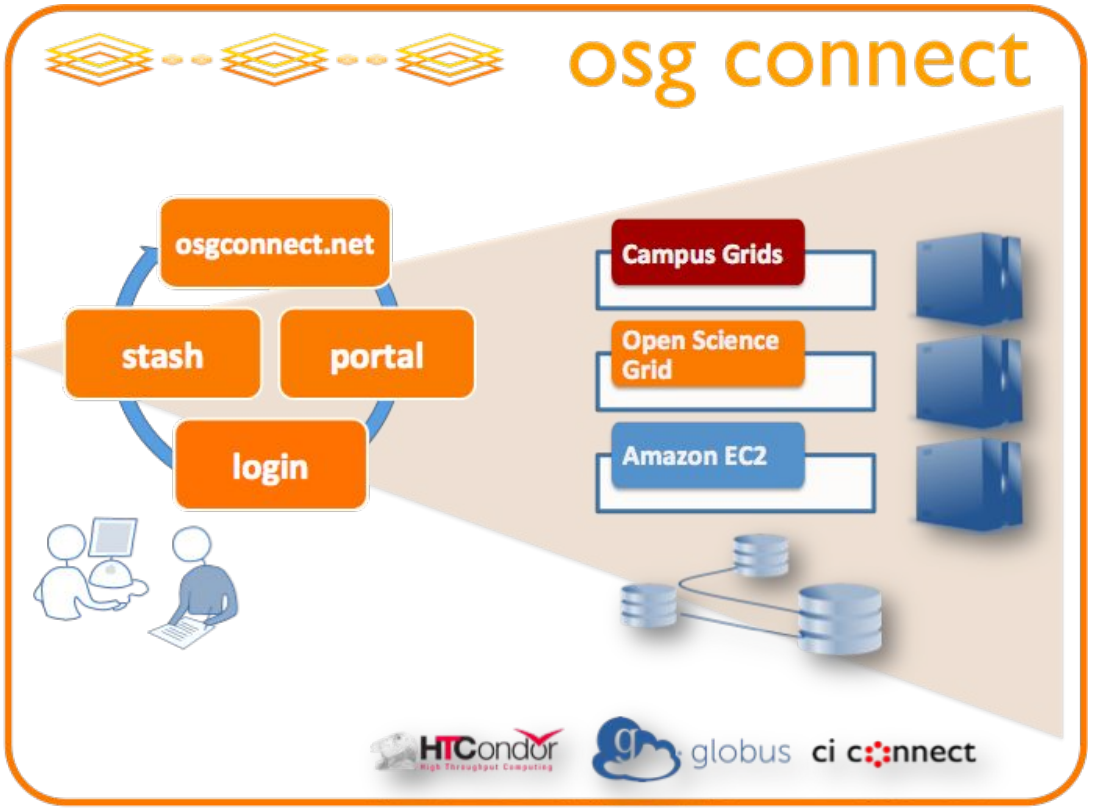
OSG Connect Service (login.osgconnect.net)

Campus identity (CILogon) ‣ OSG Connect identity (Globus) ‣ virtual organization (OSG) ‣ HTCondor to sites



⇒ Virtual HTC cluster experience

OSG Connect - easy way to get started



OSG as a **campus cluster**

- ★ Login host
- ★ Job scheduler
- ★ Software
- ★ Storage

OSG Connect Service

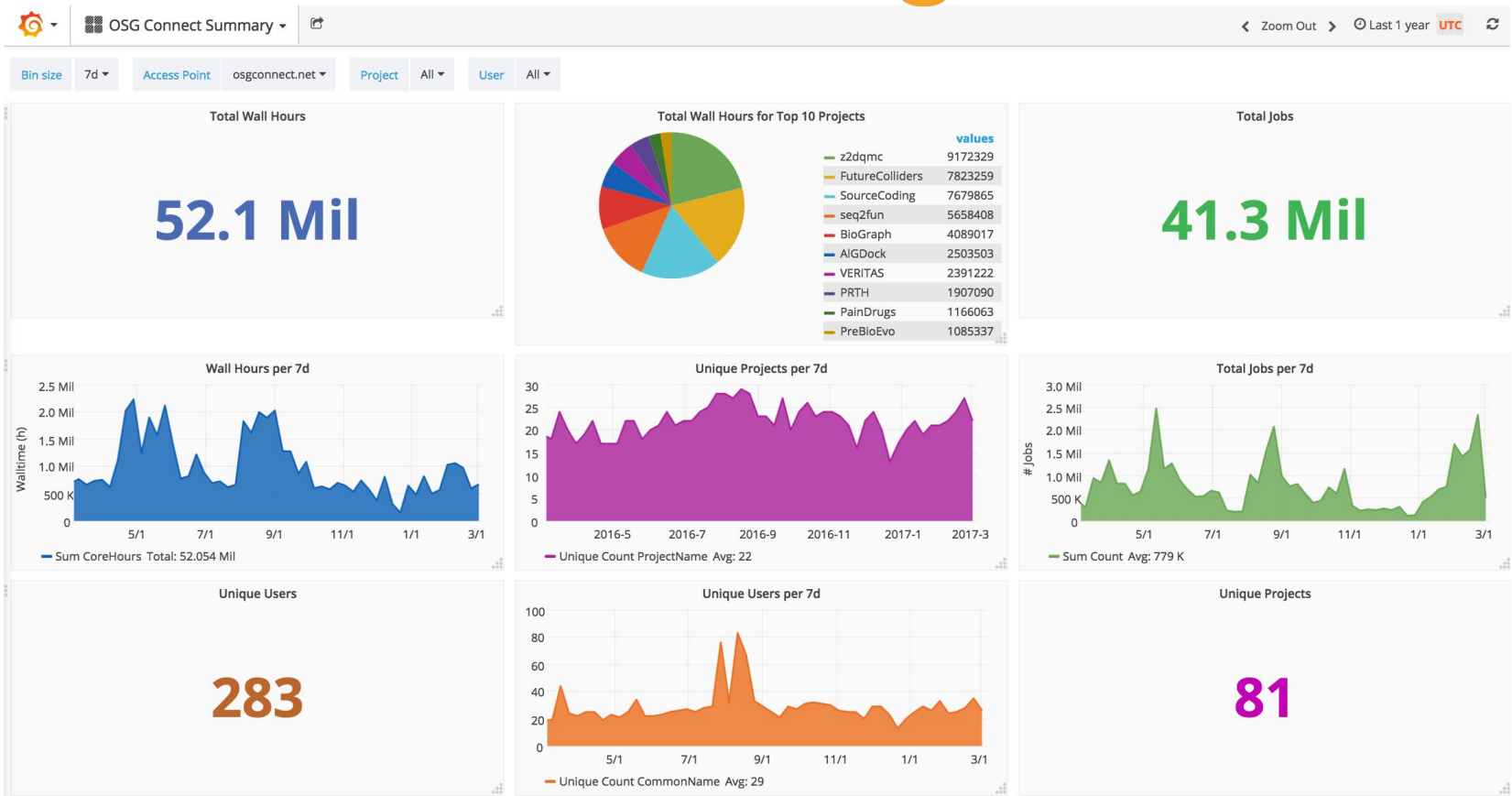
- For users without an institutional submission point
- login node for job management, `login.osgconnect.net`
- Stash is a temporary storage service
  - Globus Online, HTTP, Xrootd
  - Posix accessible from login nodes
  - Origin server for StashCache
- Uses OASIS software repository for user-installed software

Applications Repository: OASIS

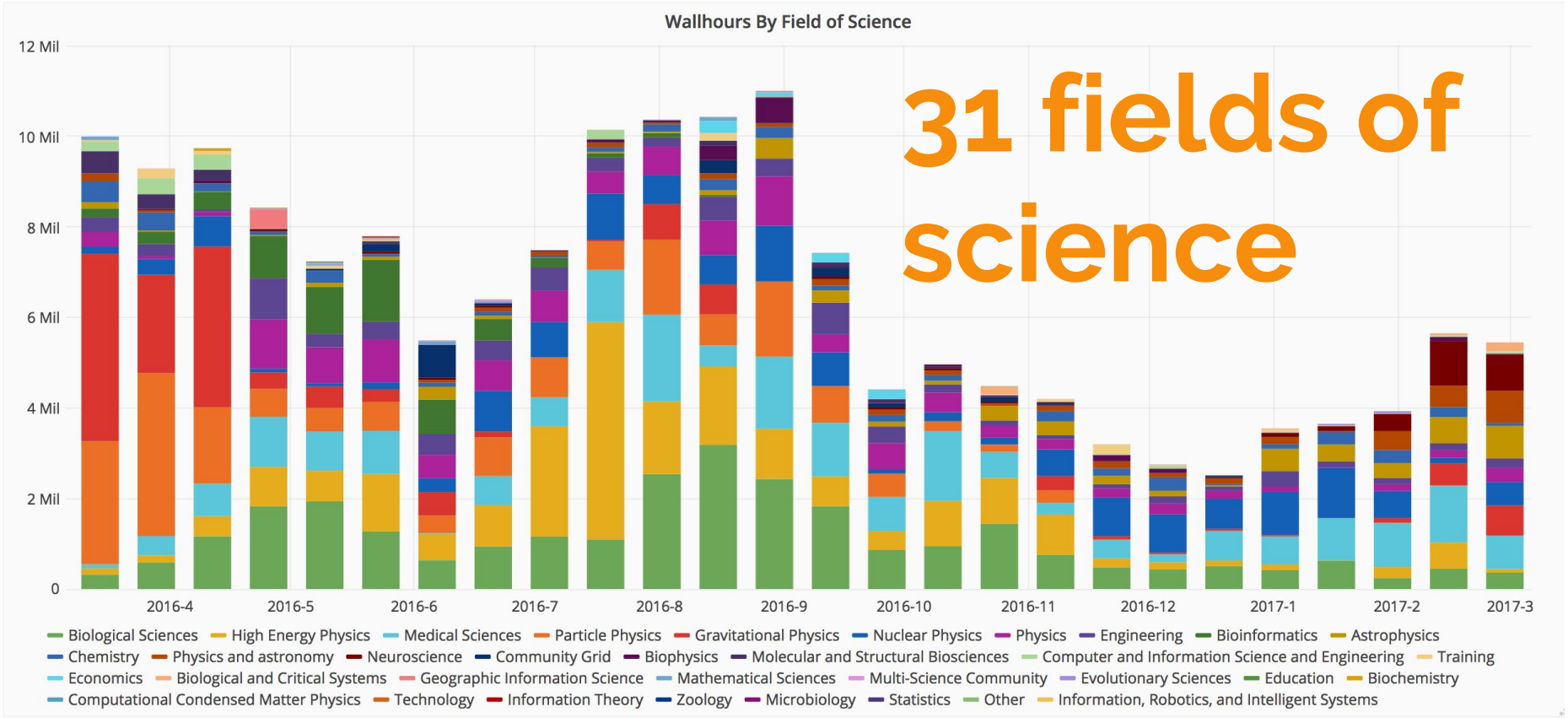
- Repository for common user software
- Accessed with a **`module`** command
  - identical software on all clusters
  - apps/libraries installed


```
#!/bin/bash
switchmodules oasis
module load R
module load matlab
...
```

osg connect

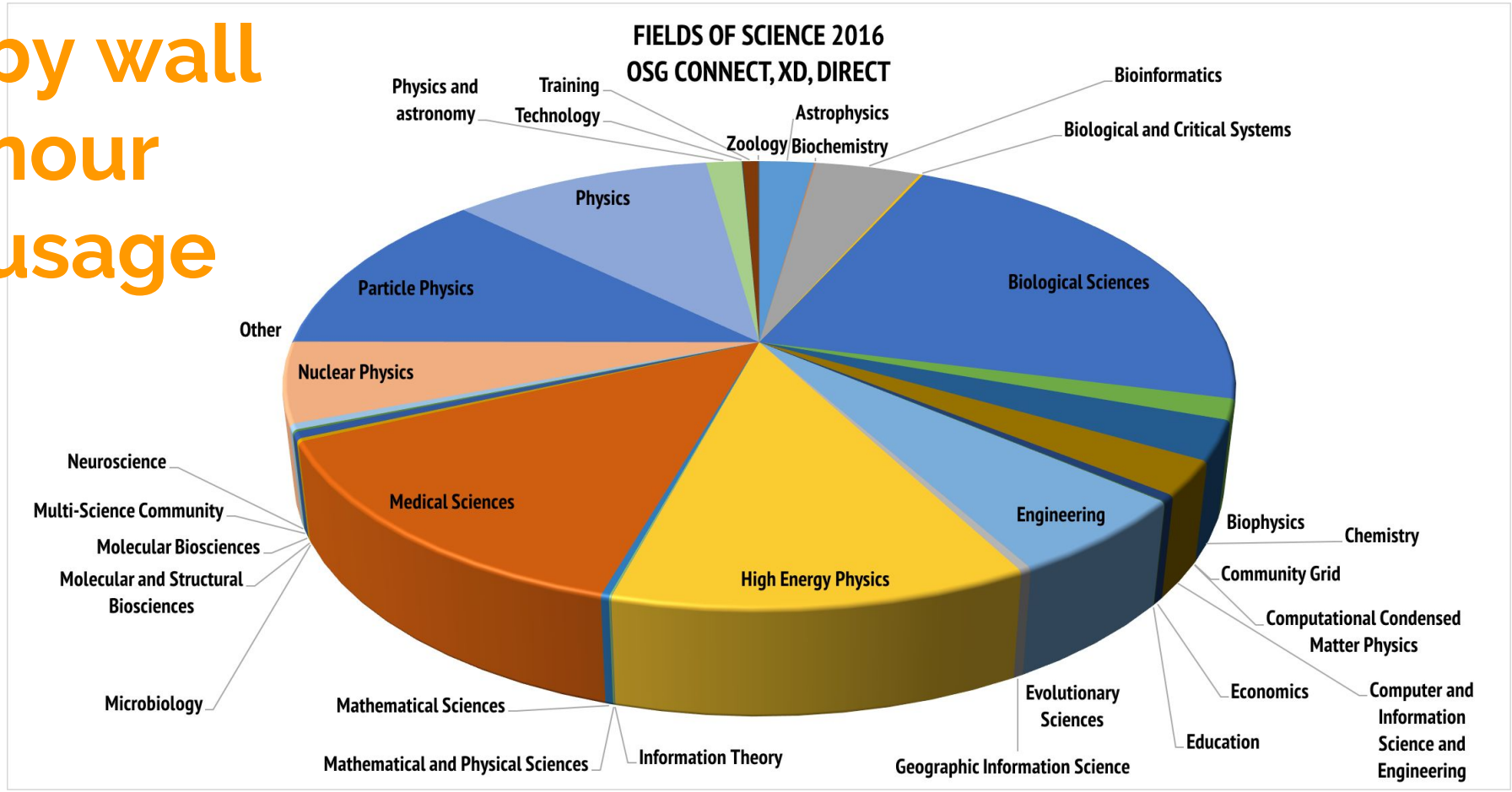

OSG Connect Summary
Zoom Out
Last 1 year
UTC
Bin size
7d
Access Point
osgconnect.net
Project
All
User
All
Total Wall Hours
52.1 Mil
Total Wall Hours for Top 10 Projects
values
z2dqmc 9172329
FutureColliders 7823259
SourceCoding 7679865
seq2fun 5658408
BioGraph 4089017
AlGDock 2503503
VERITAS 2391222
PRTH 1907090
PainDrugs 1166063
PreBioEvo 1085337
Total Jobs
41.3 Mil
Wall Hours per 7d
Walltime (h)
2.5 Mil
2.0 Mil
1.5 Mil
1.0 Mil
500 K
0
5/1
7/1
9/1
11/1
1/1
3/1
Sum CoreHours Total: 52.054 Mil
Unique Projects per 7d
30
25
20
15
10
5
0
2016-5
2016-7
2016-9
2016-11
2017-1
2017-3
Unique Count ProjectName Avg: 22
Total Jobs per 7d
Jobs
3.0 Mil
2.5 Mil
2.0 Mil
1.5 Mil
1.0 Mil
500 K
0
5/1
7/1
9/1
11/1
1/1
3/1
Sum Count Avg: 779 K
Unique Users
283
Unique Users per 7d
100
80
60
40
20
0
5/1
7/1
9/1
11/1
1/1
3/1
Unique Count CommonName Avg: 29
Unique Projects
81

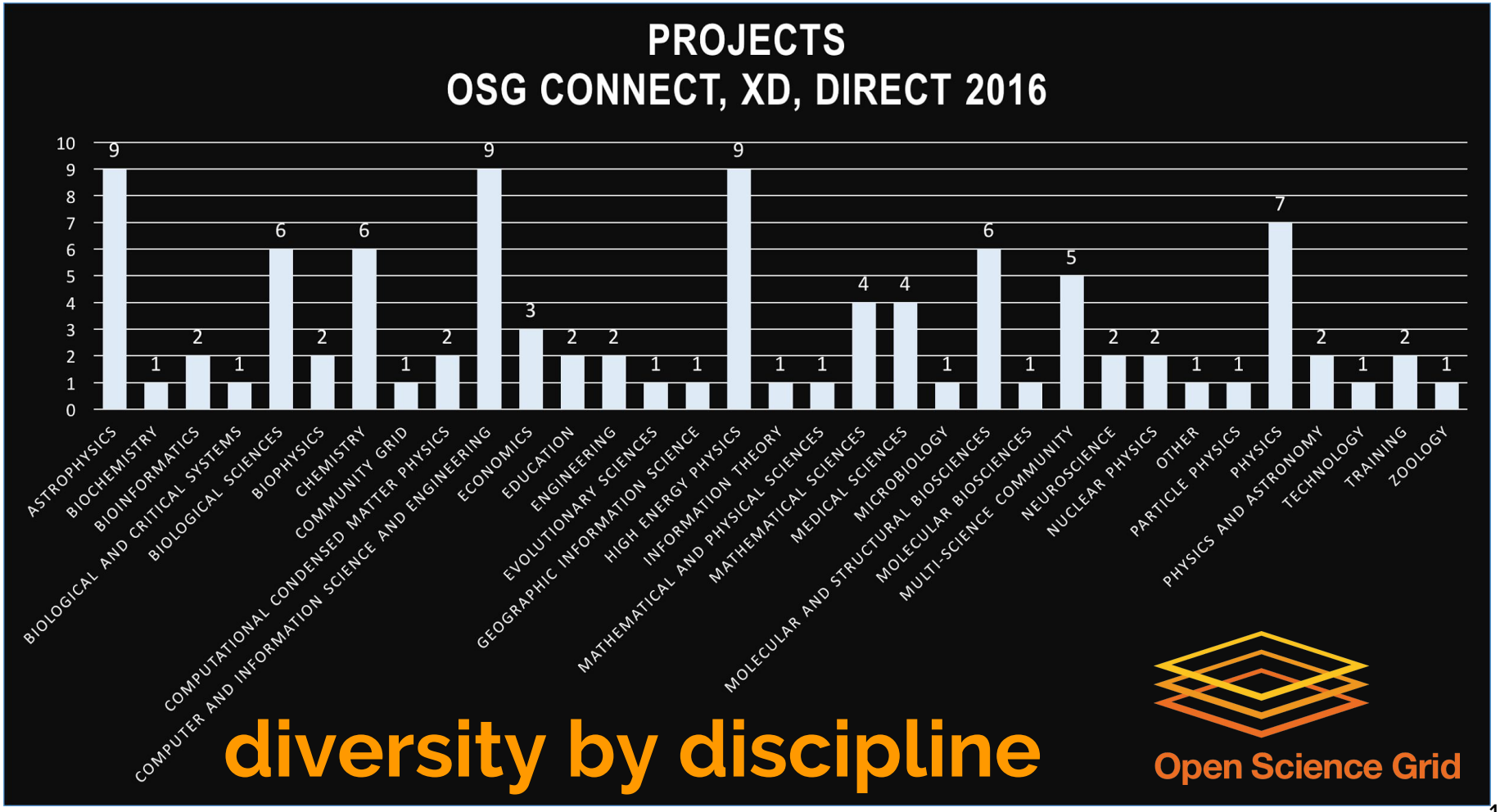
By Field Of Science

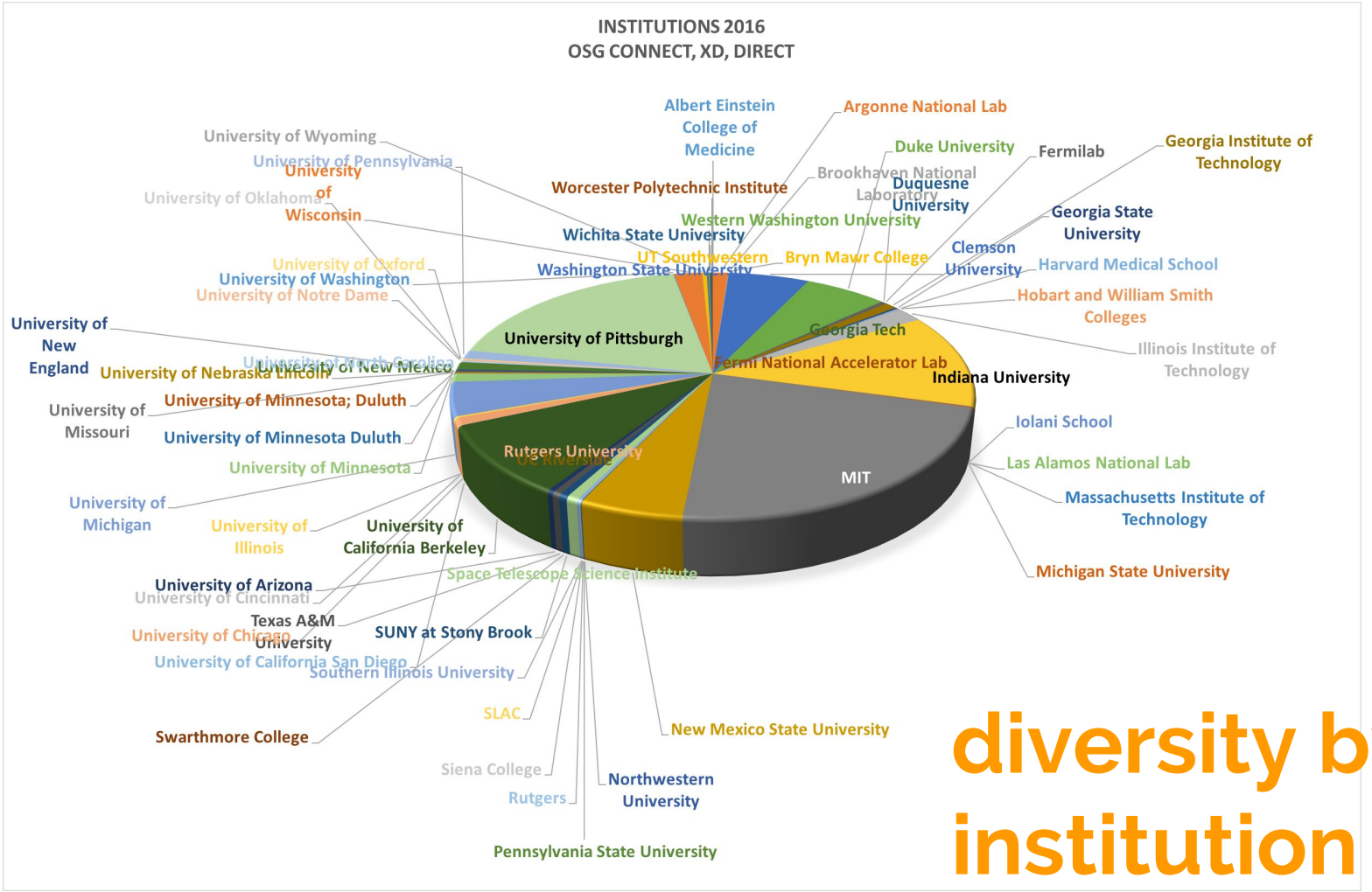


31 fields of science

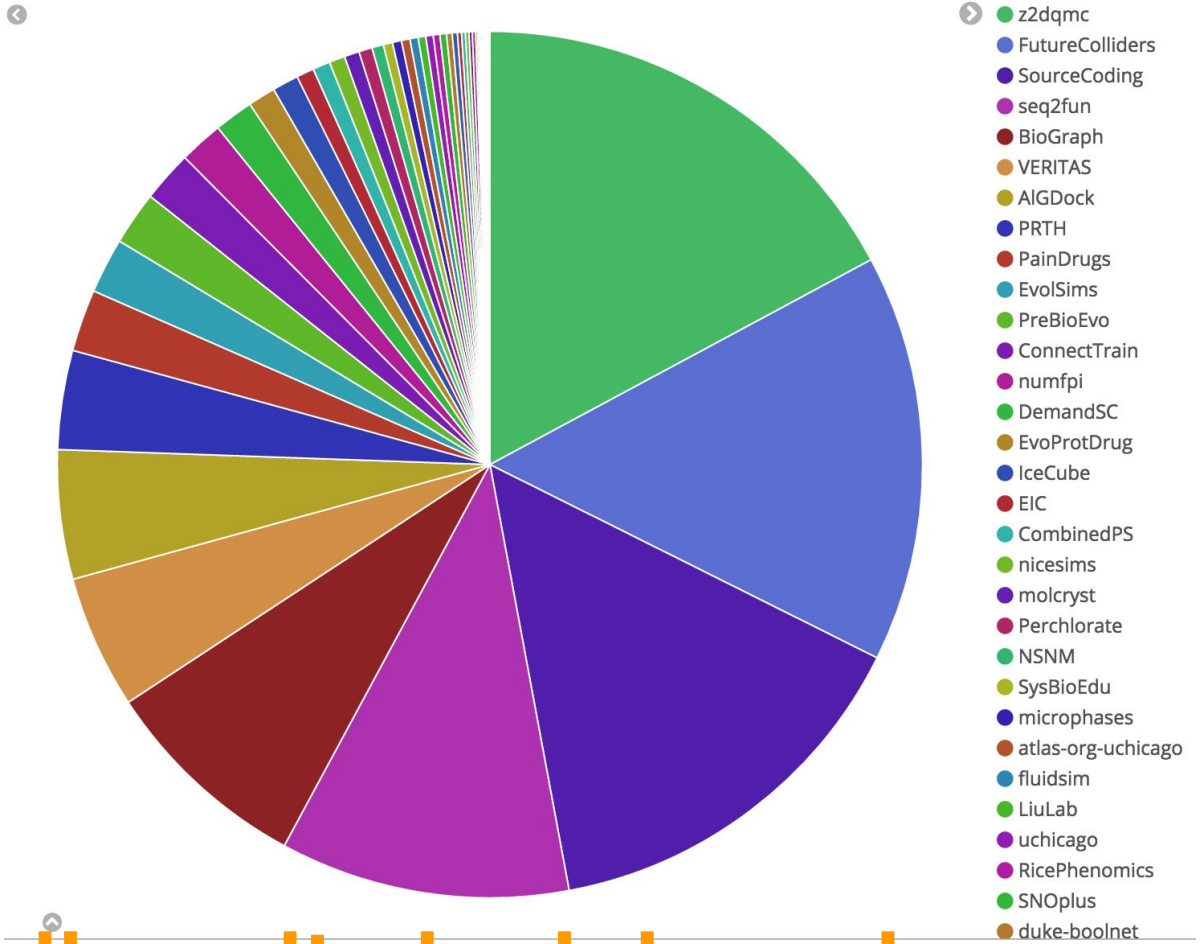
by wall hour usage



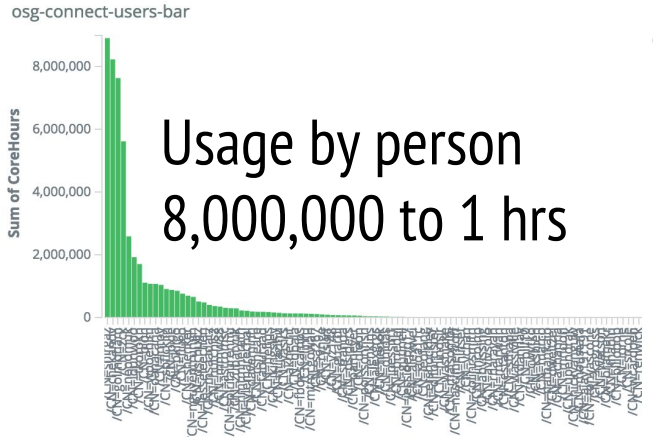

PROJECTS
OSG CONNECT, XD, DIRECT 2016
10
9
8
7
6
5
4
3
2
1
0
9
1
2
1
6
2
6
1
2
9
3
2
2
1
1
9
1
1
4
4
1
6
1
5
2
2
1
1
7
2
1
2
1
ASTROPHYSICS
BIOCHEMISTRY
BIOINFORMATICS
BIOLOGICAL AND CRITICAL SYSTEMS
BIOLOGICAL SCIENCES
BIOPHYSICS
CHEMISTRY
COMMUNITY GRID
COMPUTATIONAL CONDENSED MATTER PHYSICS
COMPUTER AND INFORMATION SCIENCE AND ENGINEERING
ECONOMICS
EDUCATION
ENGINEERING
EVOLUTIONARY SCIENCES
GEOGRAPHIC INFORMATION SCIENCE
HIGH ENERGY PHYSICS
INFORMATION THEORY
MATHEMATICAL AND PHYSICAL SCIENCES
MATHEMATICAL SCIENCES
MEDICAL SCIENCES
MICROBIOLOGY
MOLECULAR AND STRUCTURAL BIOSCIENCES
MOLECULAR BIOSCIENCES
MULTI-SCIENCE COMMUNITY
NEUROSCIENCE
NUCLEAR PHYSICS
OTHER
PARTICLE PHYSICS
PHYSICS
PHYSICS AND ASTRONOMY
TECHNOLOGY
TRAINING
ZOOLOGY
diversity by discipline
Open Science Grid



diversity by institution



- z2dqmc
- FutureColliders
- SourceCoding
- seq2fun
- BioGraph
- VERITAS
- AlGDock
- PRTH
- PainDrugs
- EvolSims
- PreBioEvo
- ConnectTrain
- numfpi
- DemandSC
- EvoProtDrug
- IceCube
- EIC
- CombinedPS
- nicesims
- molcryst
- Perchlorate
- NSNM
- SysBioEdu
- microphases
- atlas-org-uchicago
- fluidsim
- LiuLab
- uchicago
- RicePhenomics
- SNOplus
- duke-boolnet
- OSG-Staff
- freesurfer
- GeoTunnel
- holosim
- UserSchool2015
- SBGrid
- NSLS2ID
- ContinuousIntegration
- EvolvingAI
- TDAstats
- N/A
- MS-EinDRC
- SouthPoleTelescope
- mab
- duke-CMT
- xenon1t
- Unknown
- Paniceae-trans
- PathSpaceHMC



Usage by person
8,000,000 to 1 hrs

diversity by job scale

How can OSG can help?

- We can provide software and services that allow you to share your resources with a specific set of other institutions, or the nation at-large. Who you share with is entirely under your control. In some cases OSG can host these services on your behalf
- We can provide software and services that allow your scientists access to shared resources at a specific set of other institutions, or the nation at large. Whose resources your scientists access is under the control of the scientists, once enabled by you and us.
- We can help you with your perfSONAR configuration - to include in mesh testing with other universities and archival of measurements for troubleshooting

OSG User Support

Team

http://support.opensciencegrid.org/

opensciencegrid help desk

Welcome
Login Sign up

Home | Solutions

How can we help you today?

Enter your search term here... SEARCH

New support ticket

Check ticket status

User Support

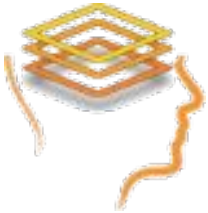
Training

Intro to HTC on OSG Connect

software carpentry

osg connect

Thank you!



opensciencegrid

user-support@opensciencegrid.org

support.opensciencegrid.org

www.opensciencegrid.org/links

Science sampler

With apologies for the many projects not included....

Large Scale Genomics

Genomics

- FASTQ files are mapped to a reference genome and converted to a BAM alignment file.
- BAM files can be mined for gene expression vectors that can be bundled into a gene expression matrix (GEM).
- GEMs are a stable data structure that can be mined for differentially expressed genes (DEGs) or used to construct Gene Co-expression Networks (GCNs)

Raw DNA Sequence →
Gene Interaction Graph

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

OSG-GEM

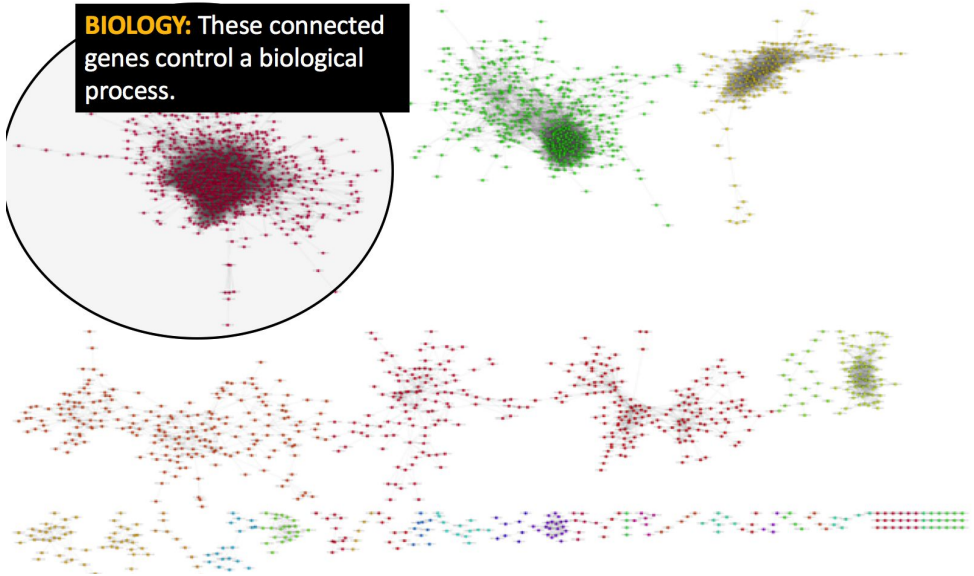
GeneID	ERR712353	ERR712355	ERR712362	ERR712366
GRMZM2G059856_T01.v6a	9.40	20.25	4.65	0.00
GRMZM5G811273_T02.v6a	0.00	0.38	0.00	0.45
GRMZM5G811273_T01.v6a	5.84	0.00	0.00	9.06
GRMZM2G059778_T01.v6a	0.45	1.65	0.00	0.88
GRMZM2G394757_T01.v6a	1.31	1.38	1.57	0.00
GRMZM2G093399_T01.v6a	9.00	1.08	1.00	0.65
GRMZM5G809743_T01.v6a	5.21	1.66	21.98	1.19
GRMZM5G833153_T01.v6a	2.98	30.50	19.33	15.75
AC177838.2_FGT012.v6a	0.29	0.24	0.00	0.00

OSG-KINC

Large Scale Genomics..

Genomics

COMPLEX GENETIC SYSTEMS WOVEN INTO GRAPHS



Gene Expression Matrix HPC Challenges

- Memory Intensive
  - Some Steps Require >16 GB RAM/dataset
- Storage Intensive
  - ~37 GB of intermediate files/dataset
- Dataset Size Varies Widely
  - 10M to 20M sequences/dataset
- Data Transfer Issues (Tera-/Petabytes)

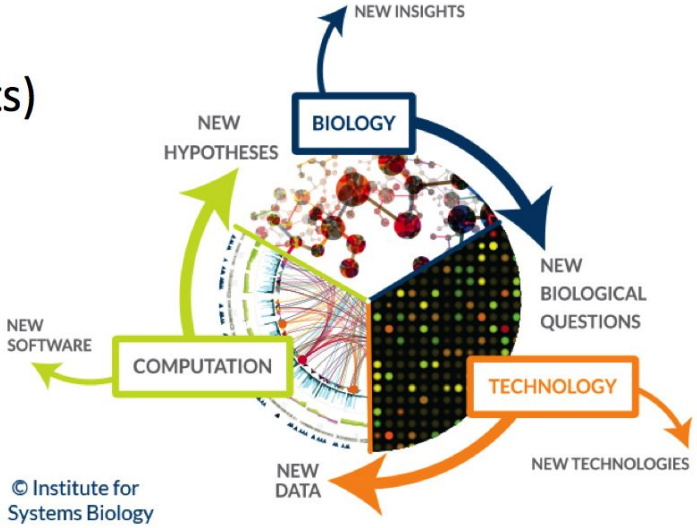
```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

William Poehlman, Alex Feltus • Clemson University •• Stephen Ficklin, Washington State University

Student Training w/ OSG

Genomics

- Introduction to Systems Biology
- WSU HORT 503 Special Topics (3 credits)
- Offered Fall 2017
- 9 Students
  - Crop & Soil Sciences
  - Molecular Plant Sciences
  - Entomology
  - Biological Systems Engineering
- Why OSG?
  - Expose students to national computing resources
  - Anticipate large-scale genomics of the future.



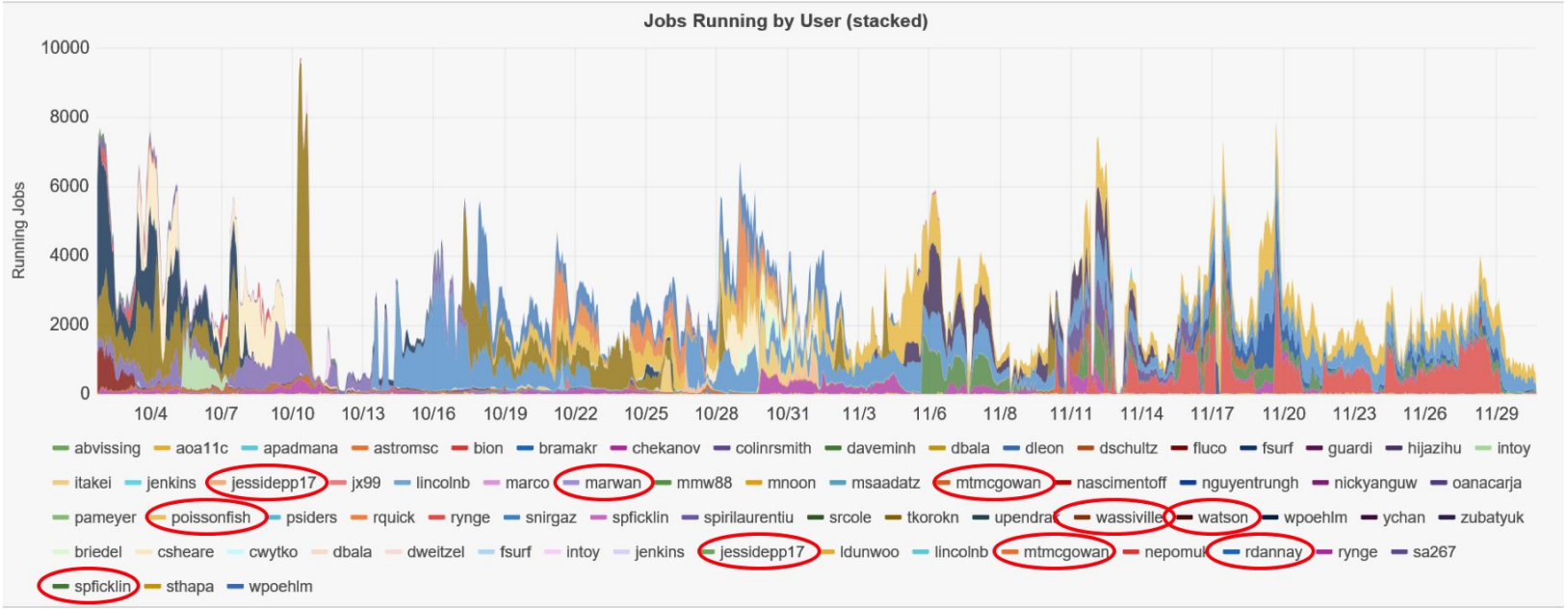
WASHINGTON STATE UNIVERSITY

William Poehlman, Alex Feltus • Clemson University •• Stephen Ficklin, Washington State University

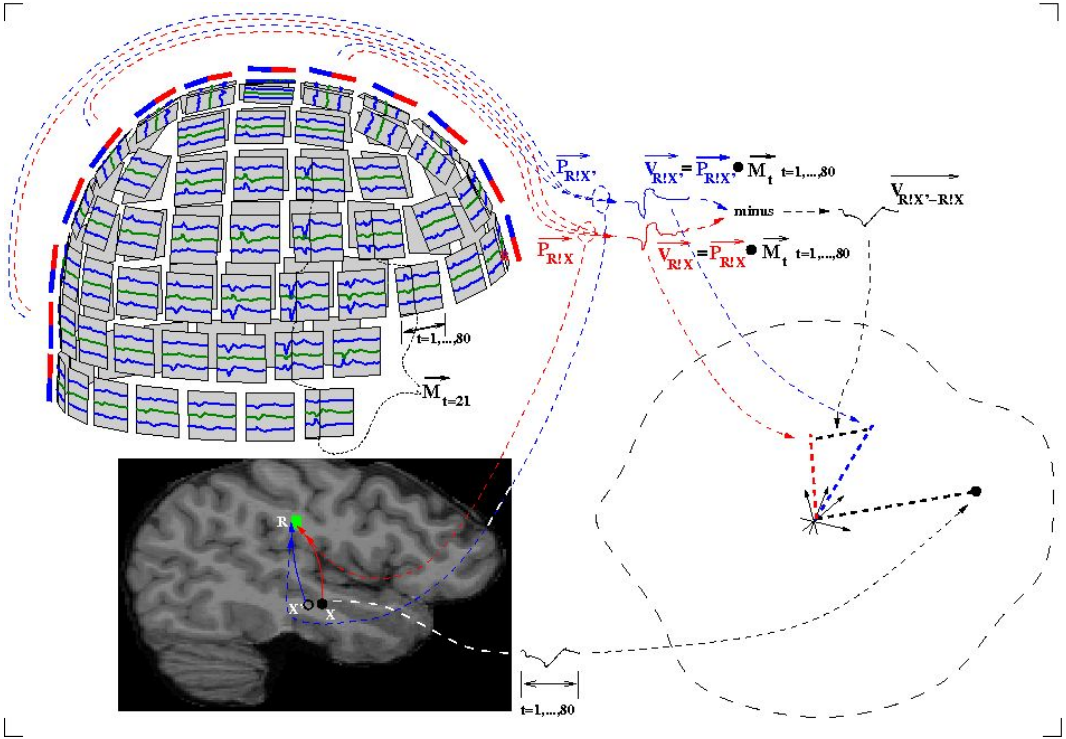
Student Training w/ OSG

Genomics

- Success: SysBioEdu Usage from Oct-Dec 2016
  - Students launched jobs and they ran!



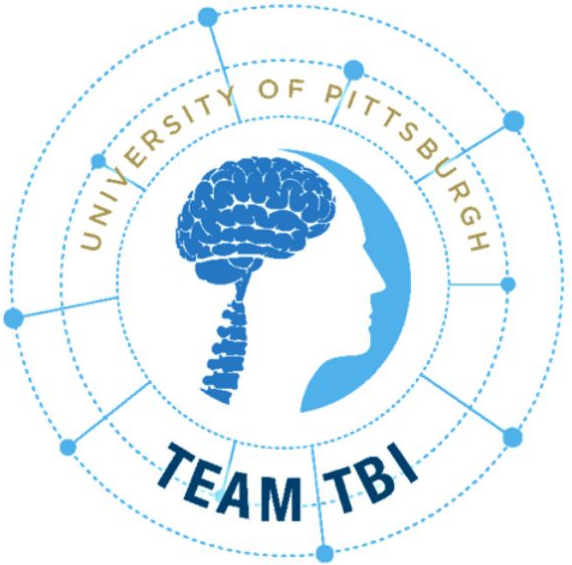
Functional Neuroimaging



Functional Neuroimaging..

- Don Krieger has been working with TEAM TBI at the University of Pittsburgh
  - **T**argeted **E**valuation, **A**ction and **M**onitoring of **T**ramatic **B**rain **I**njury

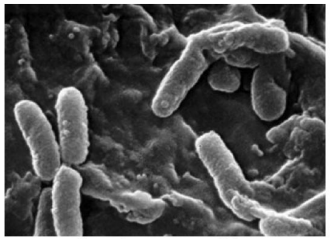
- TEAM TBI investigates the complexity of brain injury, and how targeted interventional strategies may improve outcome and function.



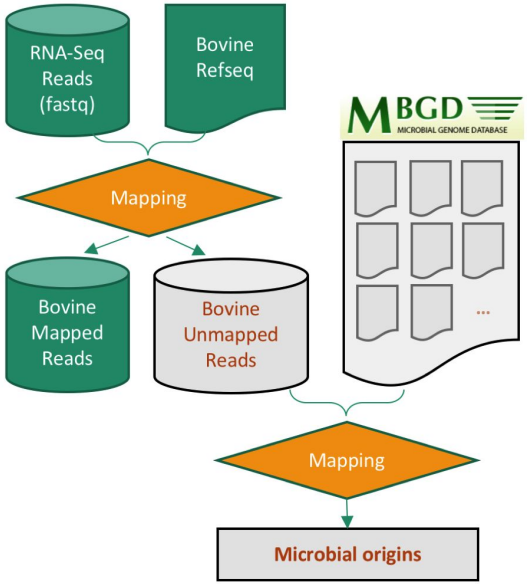
Large Scale Metagenomics.

Motivation

- In a pervious project, we have isolated exosomes from the commercial cow's milk and assessed the bovine molecules inside the exosomes.
- Moreover, we also found many unmapped reads are from microbial species.
- Thus, we designed a follow-up study to **understand the origin of microbial sequences in bovine milk exosomes.**
  - **Metagenomics analysis**



Pseudomonas cedrina



Large Scale Metagenomics..

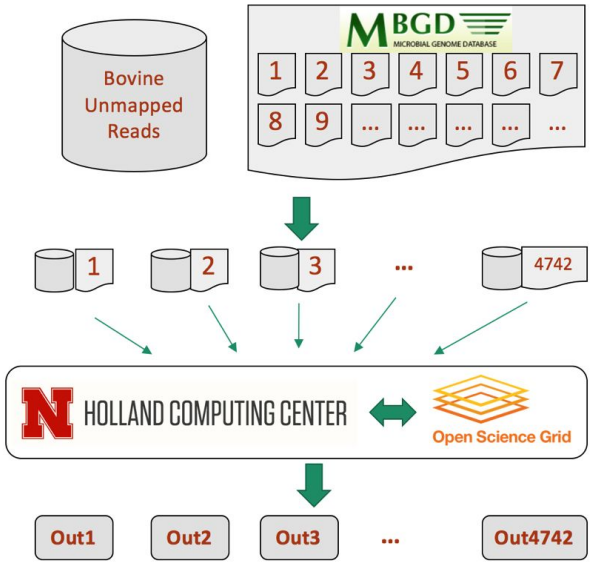
Computational challenges

- A large number of target genomes
  - 4,742 genomes (size: 100KB ~ 58MB)
- Six samples contain over 100 million of bovine unmapped reads
  - Increasing the computing time
- In total, 6 x 4,742 = ***28,452*** mapping tasks

Question: Where to execute this many of jobs?
- Impossible for the lab-server (32 cores)
- Long pending time if submitted it to HCC clusters
  - Dynamic priority scheduling of users/groups
  - More jobs completed -> longer queue time

Perfect Fit of Open Science Grid (OSG)

- The tasks are independent to each other
- Limited file transfer
  - Total size of transferred files ~1GB
- Small memory consumptions
  - Memory < 2GB
- Short running time for each task
  - Maximum: 3 hours (HCC@UNL-Crane)
- Software is available on OSG
  - Pre-installed Bowtie and Tophat
  - No further configuration needed



Large Scale Metagenomics…

Results

- Several microbial species were identified in bovine milk exosomes:
  - *Pseudomonas fluorescens, Pseudomonas chlororaphis, Pseudomonas poae, Enterobacter cloacae,* etc.
- Although some species have been reported in cow's milk before, this is the first time of identifying microbial sequences in milk exosomes
  - Potential transportability to other species through exosomes
- Based on the findings from this analysis, we have designed two experiments to further our understanding in this subject

Counterfactual Analysis.

- Economic analysis & public policy
- Considering "what if" scenarios in microeconomics
- Simulate firm/consumer behaviors

Using OSG to Evaluate Policy

We need to solve

$$\frac{\partial \Pi_{jt}}{\partial p_{jt}} + \beta \left[\frac{\partial \boldsymbol{s}_t}{\partial p_{jt}} \right]' \mathbb{E}_t \left[\frac{\partial V_j(\boldsymbol{s}_t, \boldsymbol{X}_t)}{\partial \boldsymbol{s}_t} \right] = 0,$$

for every firm and combination of $\boldsymbol{s}_t$ and $\boldsymbol{X}_t$.

Need to compute $\mathbb{E}_t \left[\frac{\partial V_j(\boldsymbol{s}_t, \boldsymbol{X}_t)}{\partial \boldsymbol{s}_t} \right]$. How? Forward simulation.

OSG is what makes this possible for large state spaces.

Counterfactual Analysis..

Using OSG to Evaluate Policy

Outcome: making people more active decreases prices and it does not affect returns.

Case	Mean and 95% CI
Base simulation	6.195% [6.181%,6.210%]
No enrollment cost	3.666% [3.660%,3.671%]
No decision cost	3.837% [3.833%,3.842%]
No switching costs	2.607%

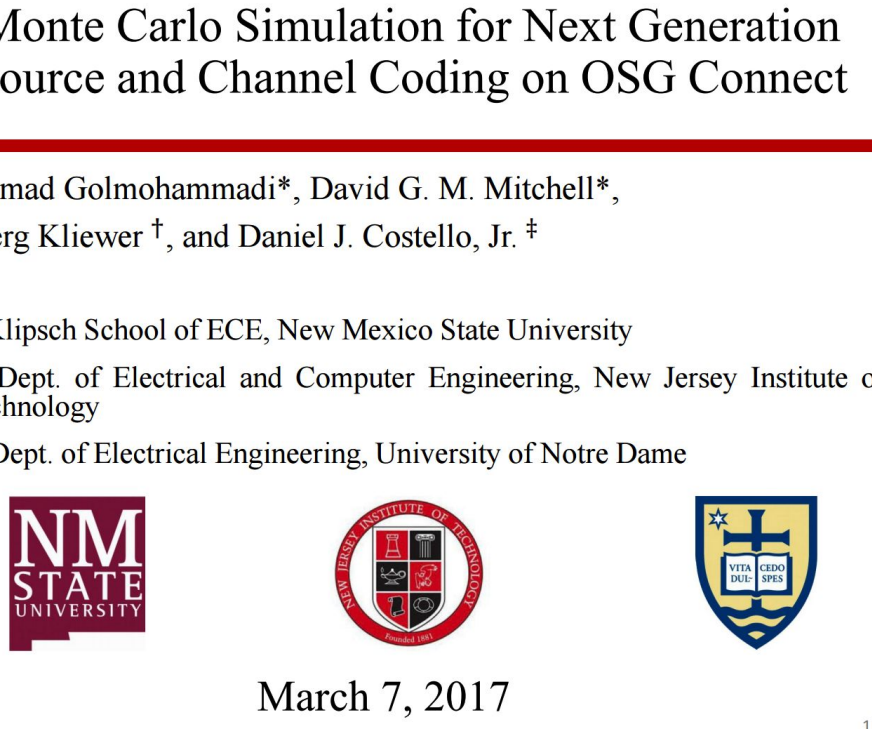
Note: The table reports the mean expected fees and 95% confidence intervals for the different scenarios under study using 10,000 random initial states.

Conclusions

1. Policy/transaction evaluation is critical in modern microeconomics.
2. Evaluations often require simulating consumer/firm behavior over a large state space.
3. In particular in the case of dynamic games, OSG could become a crucial tool, making the difference between being able to do something and not.

Simulating Source Coding.

- Data deluge - much of it mobile traffic
- Optical data compression
- Important for digital space and satellite communication & wireless data transmission



Monte Carlo Simulation for Next Generation Source and Channel Coding on OSG Connect

Ahmad Golmohammadi*, David G. M. Mitchell*,
Joerg Kliewer †, and Daniel J. Costello, Jr. ‡

* Klipsch School of ECE, New Mexico State University

† Dept. of Electrical and Computer Engineering, New Jersey Institute of Technology

‡ Dept. of Electrical Engineering, University of Notre Dame

March 7, 2017

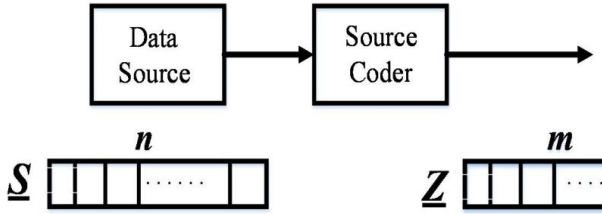
1

Simulating Source Coding..

- Whole system simulations: transmitter, decoder, receiver & stochastic noise
- Data compression & reconstruction algorithms

Data Compression

❑ In order to handle the vast amounts of data society will produce, we need efficient and low-complexity algorithms to reliably compress and reconstruct data.



Compression rate

$$R = \frac{m}{n}$$

❑ We want to reconstruct the sequence with as little distortion as possible with a practical scheme (lossy source coding)

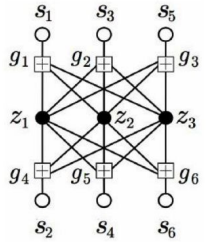
❑ The ultimate limit of compression is known (Shannon 1958)

Simulating Source Coding...

- Sparse graphs can approach fundamental limits
- To verify the results, large Monte Carlo samples needed - "not possible without the OSG"

LDGM Codes for Lossy Source Coding

- We construct codes for lossy source coding using a small structured graph (protograph)
- A large graph can be obtained from a protograph by graph lifting with lifting factor M
  - The graph is copied M times and the edges randomly permuted following the graph structure
- Low complexity algorithms based on belief propagation can be defined for the sparse graph
  - Here, messages are passed iteratively forward and backward in the graph until we converge to a codeword $\underline{\boldsymbol{z}}$



(3, 6)- regular

Evolving Strategies for Life.

- Understanding evolution at molecular scale in DNA with combination of mathematical modeling and simulation
- How quickly does a genome fix a mutation?
- Role of randomness versus natural selection?

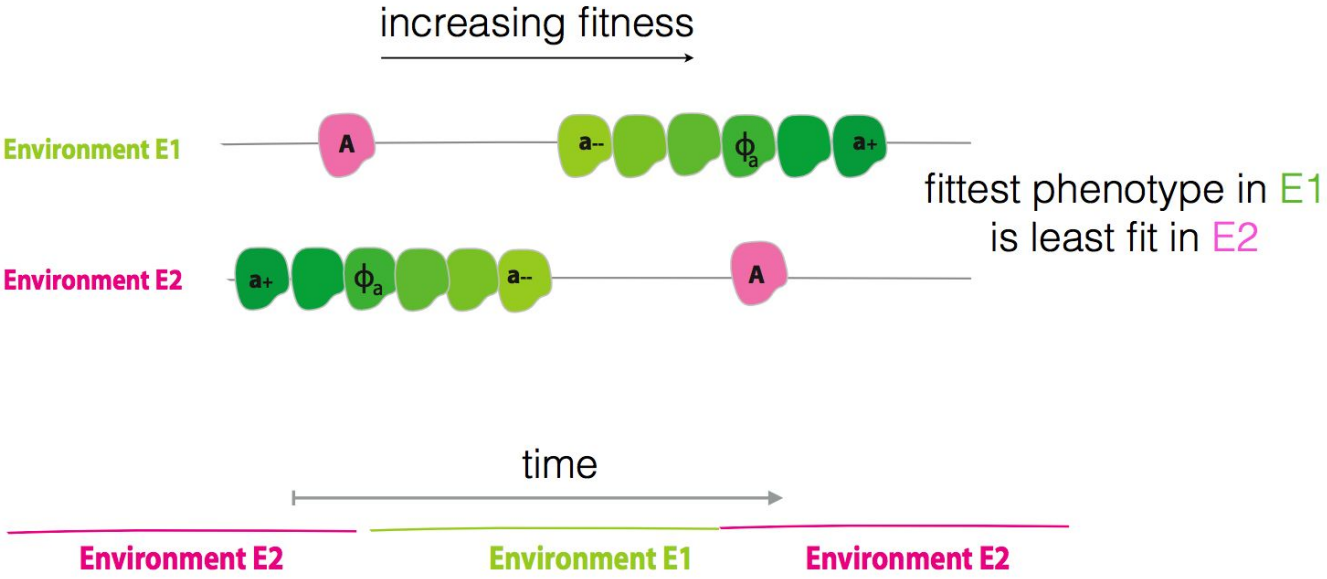
Environmental variation is commonplace yet unpredictable across biological systems from the adaptive immune system, the microenvironment in cancerous neoplasms, to populations of pathogens under drug pressure.

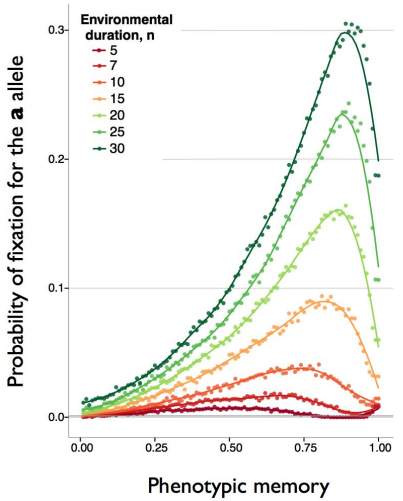
How do populations survive environmental stochasticity? How do they manage to persist and keep one's footing on an ever-changing landscape?

Can organisms prepare for this environmental stochasticity?

Can evolution prepare populations for this environmental stochasticity?

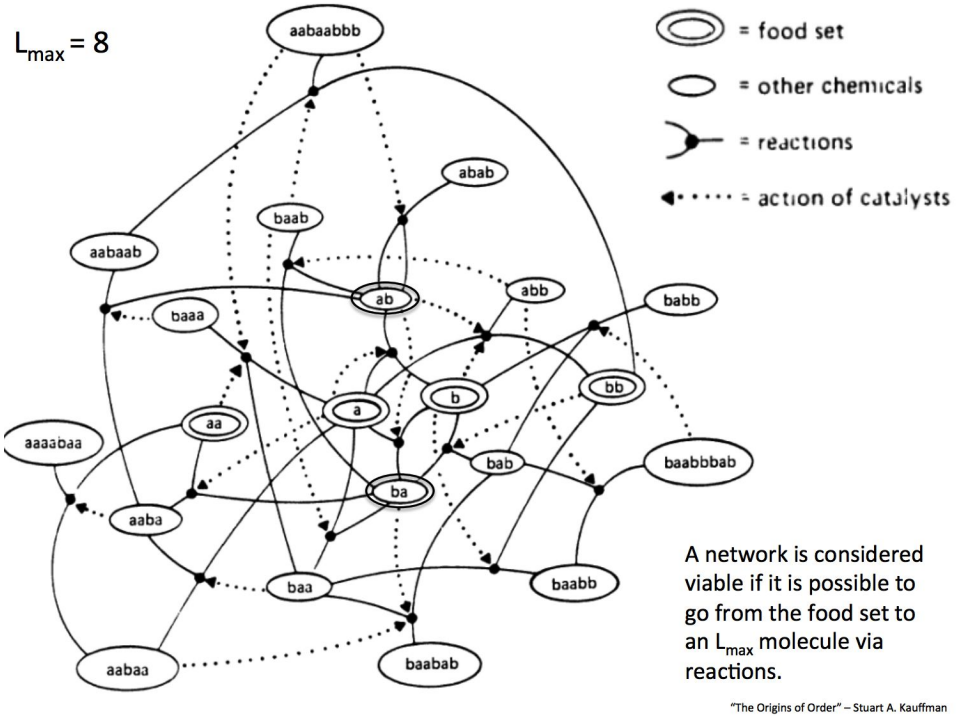
Evolving Strategies for Life..





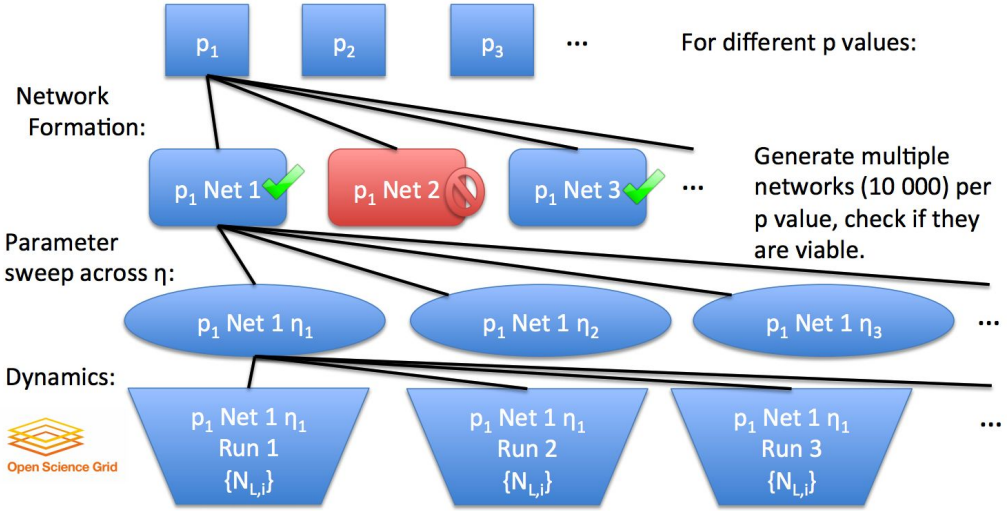
Models of Prebiotic Evolution

- Protein first origin of life model
- Network of interacting molecules assumed to be polymers
- Perhaps solve Eigen's paradox (low probability of randomly constructing "starter gene")



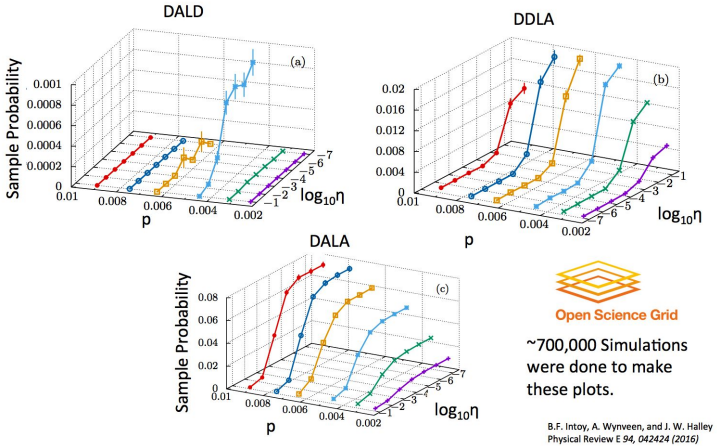
Models of Prebiotic Evolution..

Simulation General Structure



- Do multiple dynamic simulations with random initial conditions using a given viable network generated by parameter p combined with reaction rates and diffusive value η.
- A steady state is then reached with polymer length and spatial distribution $\{N_{L,i}\}$.
- Analyze the $\{N_{L,i}\}$'s to determine whether the run was lifelike or not.

Probabilities of DALD, DDLA, DALA states as a function of p and η



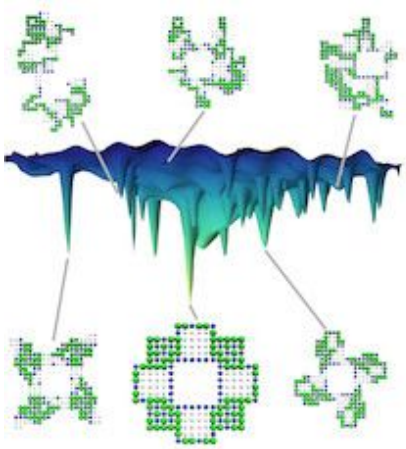
Open Science Grid

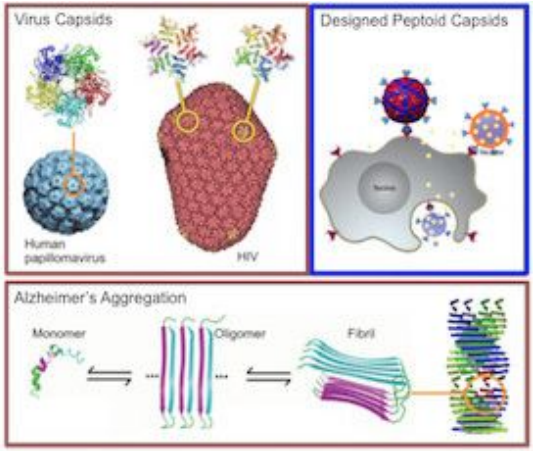
~700,000 Simulations were done to make these plots.

B.F. Intoy, A. Wynveen, and J. W. Halley Physical Review E *94, 042424 (2016)*

Protein Evolution

Understand the fundamental physical bottlenecks and dynamical behavior of protein evolution. Important questions include the extent of dominant pathways (convergent evolution) and phase transitions in evolutionary rates (punctuated equilibrium). These principals and their structural underpinnings can also be used to inform rational design of antibiotics that exploit bottlenecks in pathogen mutational response.



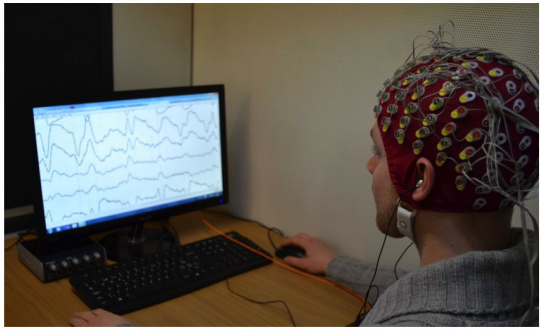


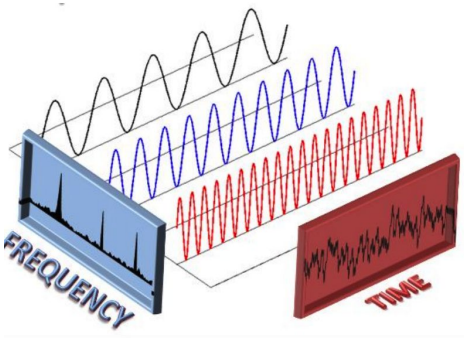
Analysis of Brain Rhythms

Neuroscience

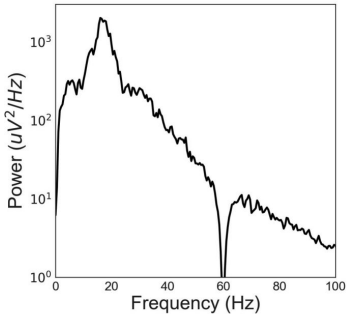
Large-scale brain recordings

- Healthy humans
- Patients
- Animals





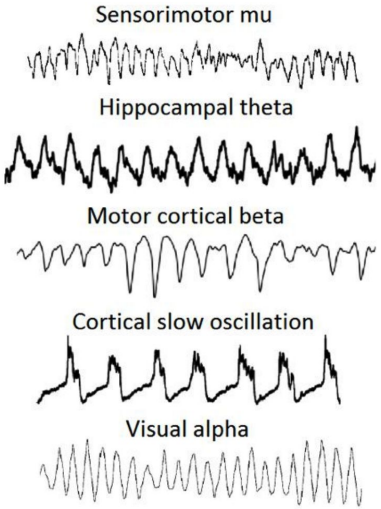
- Sampling: 500-30000 Hz
- Duration: 1 hour - 1 week
- Channels: 1 - 250+
- Several GB per subject

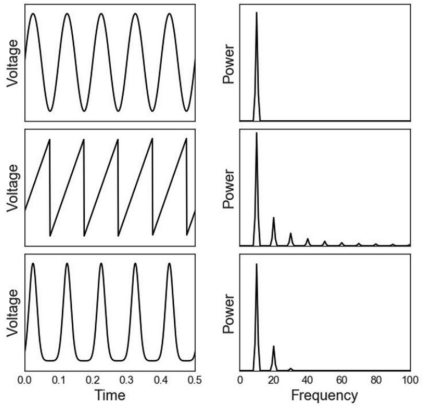


Analysis of Brain Rhythms..

Neuroscience

Fourier Transform-based analysis





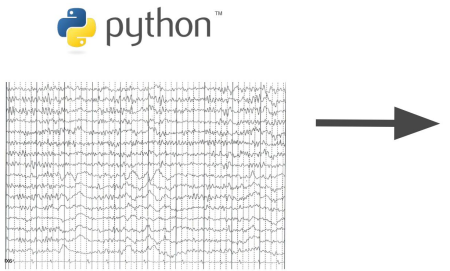
Free supercomputing for research: A tutorial on using Python on the Open Science Grid

Jan 3, 2017

Supercomputing resources typically cost money, but the Open Science Grid (OSG) provides high-throughput computing to any researcher in the US **for free**. Briefly, OSG users can run jobs on servers owned by dozens of academic institutions, whenever those servers are not actively running a job for its owners.

Besides cost, a second major barrier to entry for those who are new to supercomputing (or, specifically, Condor) is the necessary troubleshooting before we can actually run our jobs. The purpose of this tutorial is to provide a complete example for running Python jobs on the OSG. This example is nontrivial, in that it includes multiple data sets, public libraries (e.g. scipy), private libraries, and analyzing output. In complement to this tutorial, the OSG has tutorials, a structured class, and extremely helpful online support when you get stuck.

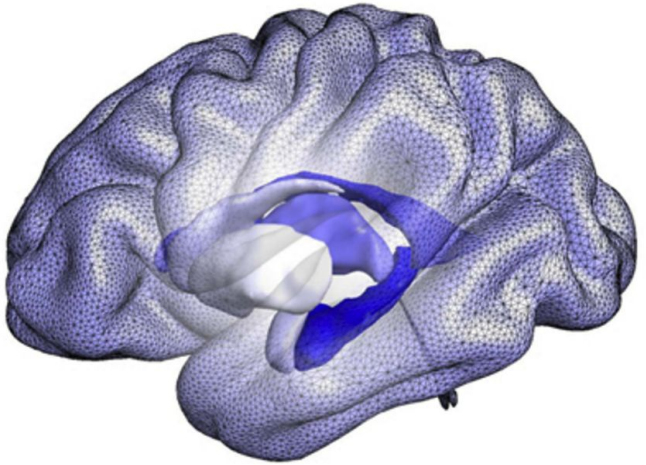
This tutorial goes through the steps of manually connecting to and running commands on the remote server, but see the **Fabfile** section at the bottom for how this can be automated on your local machine.



A FreeSurfer Workflow Service

Neuroscience

- Widely used software suite for analysis of human brain MRI scans.
- Neurophysiology of depression, examining possible anatomical differences involved in ADHD, and studying autism



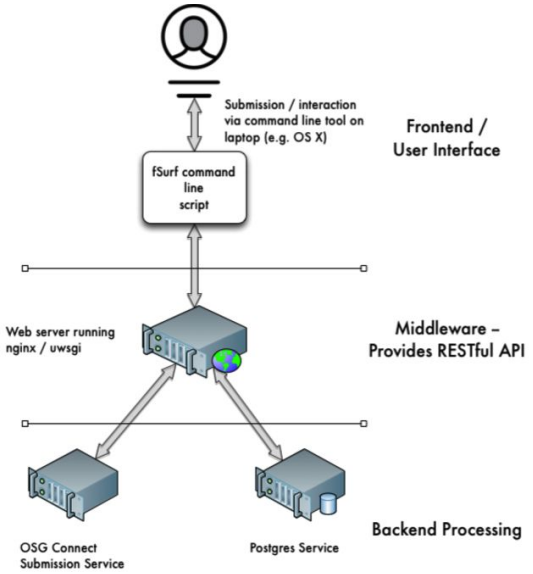
Computational model of the cortical and subcortical brain structures that form the basis of the BrainPrint, a system for representing the whole brain based on the shape, rather than the size of structures. (Martin Reuter, PhD, and Christian Wachinger, PhD, Martinos Center for Biomedical Imaging)

A FreeSurfer Workflow Service

Neuroscience

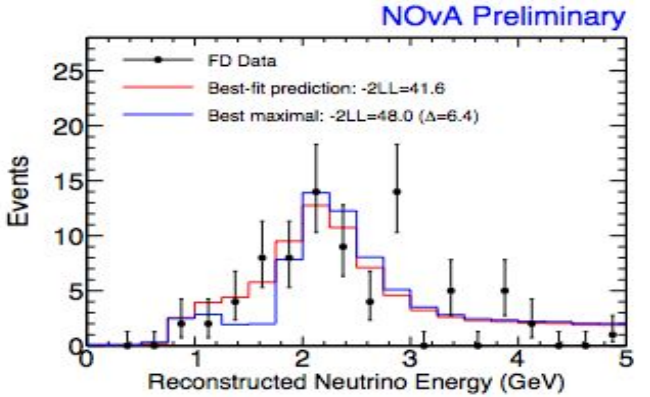
- Working with Don Krieger (Pittsburgh) to develop an OSG-based execution service
- Uses Pegasus
- Handles "standard" transforms and user options
- To be released this week!

- Front-end user interface
  - Command line script
    - Communicates to the execution service via RESTful-API commands
    - Allows user to submit, view, and remove workflows as well as download results
  - Middleware
    - NGinx / uwsgi server provides RESTful API
- Backend services (running on a single VM)
  - Postgres database
    - Tracks workflow status, locations of inputs and results
    - Stores user information
  - Pegasus submission service (using same infrastructure as OSG Connect) to



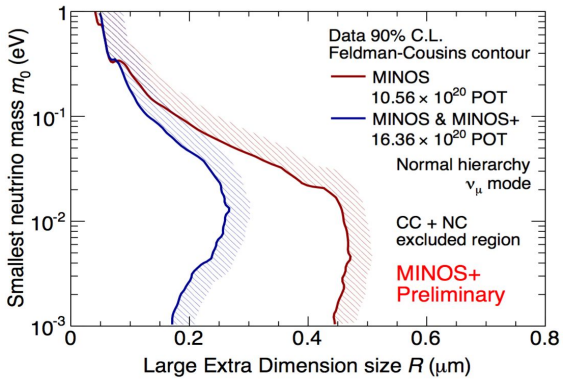

VOs

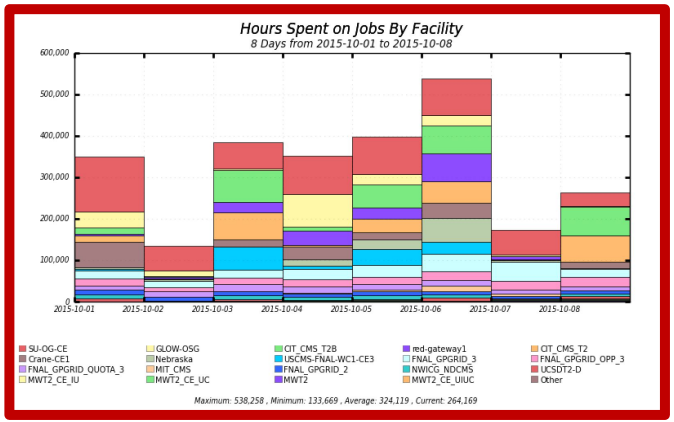
VO Highlights: From the smallest scales…



NOvA: Fermilab-based neutrino experiment

MINOS+: limits on LEDs



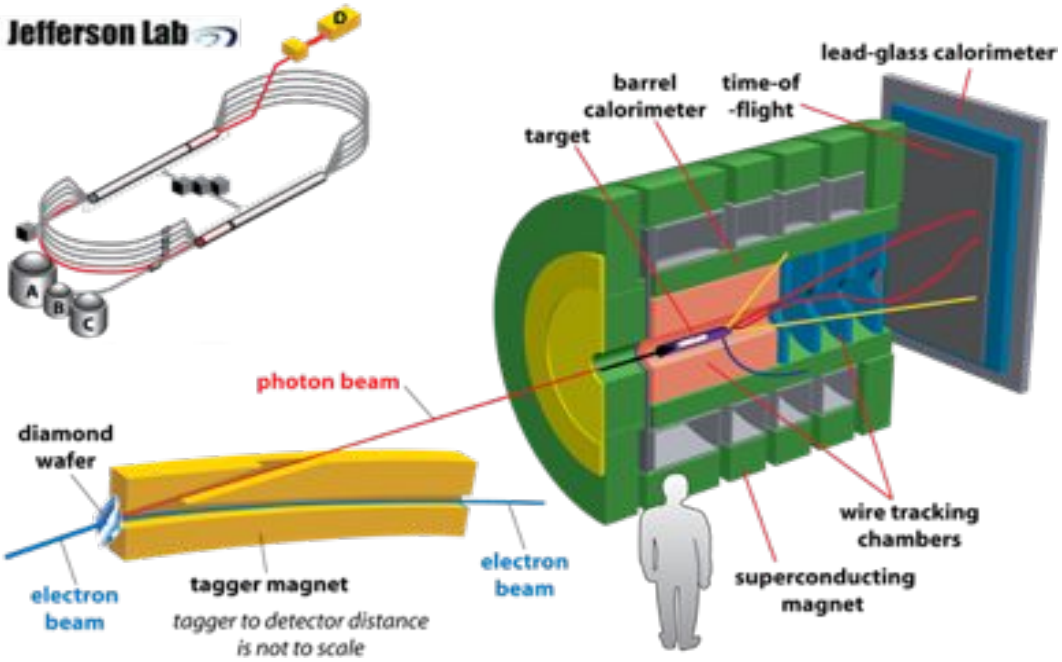


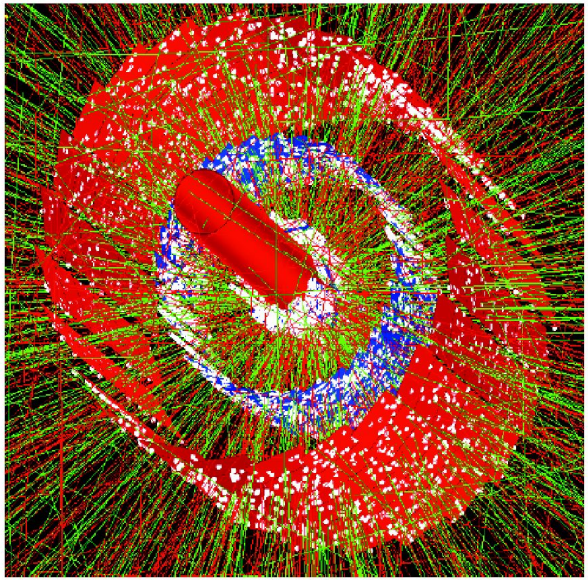
Mu2e: Lepton-flavor violation experiment

Nearly 60M opportunistic hours on OSG and counting

>500,000 in one day!

VO Highlights: From the smallest scales...



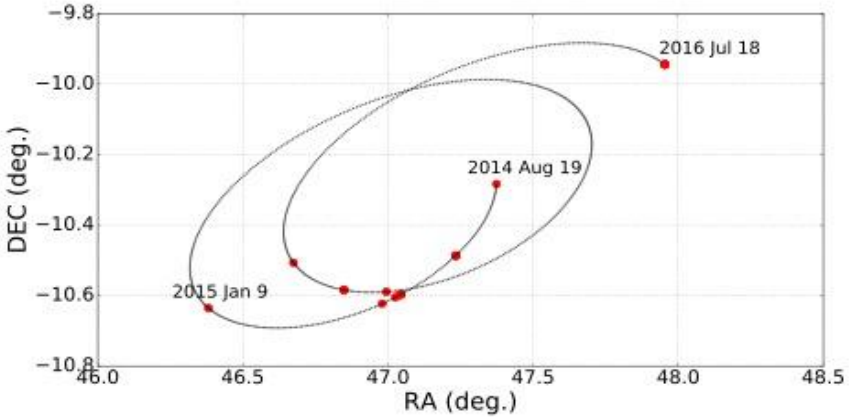


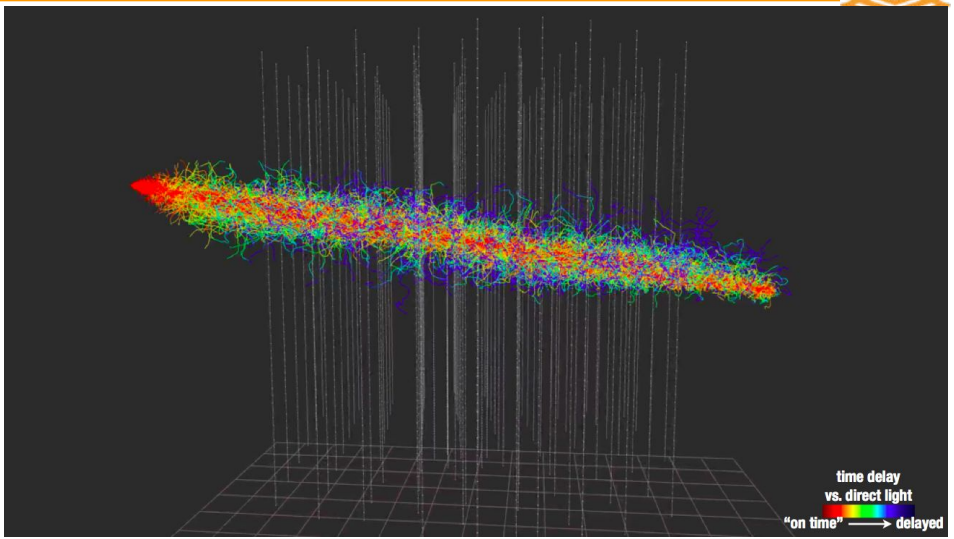
Au+Au event

GlueX: probing exotic mesons predicted by LQCD

STAR: Heavy Ion Physics

...to the largest...

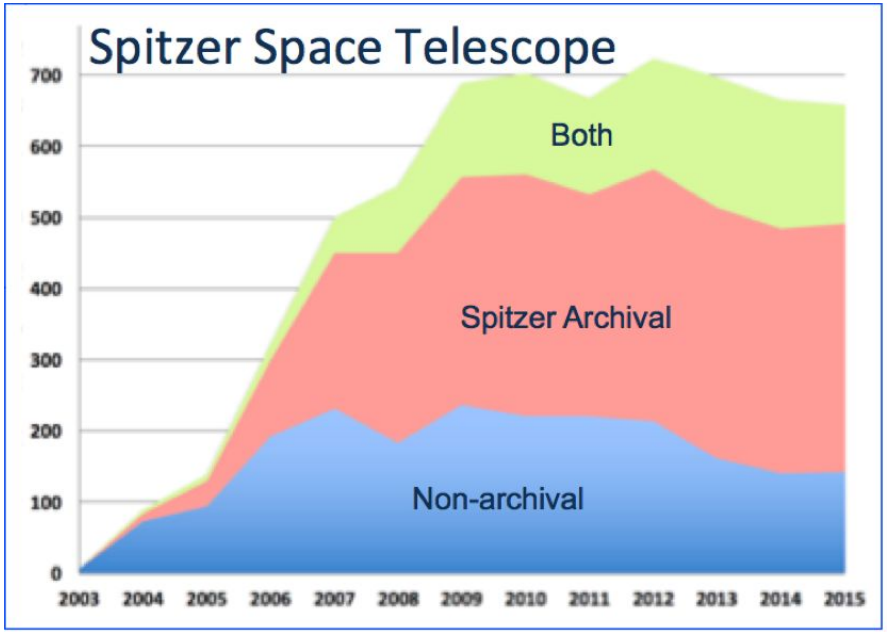




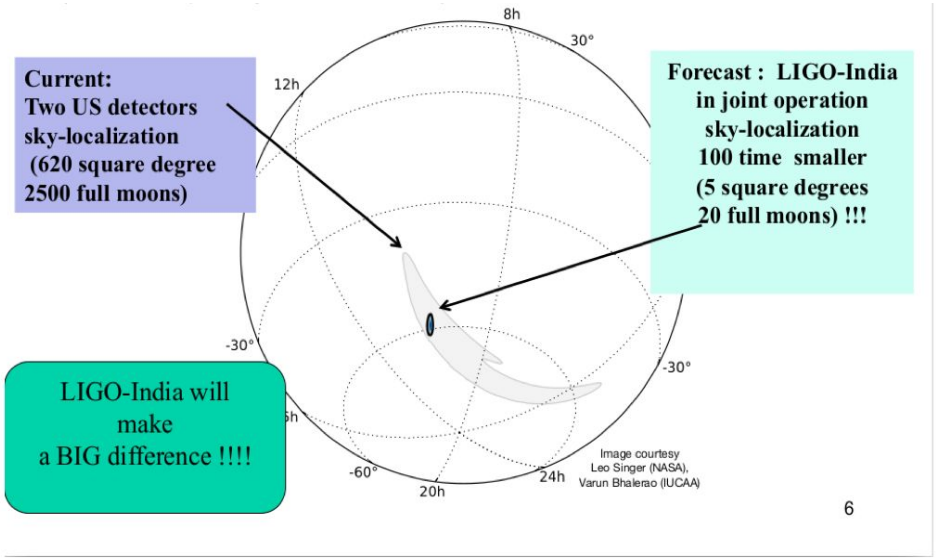
Dark Energy Survey:
Discovery of dwarf planet-
Second-most distant known object in solar system
Techniques applied to ongoing Planet 9 search

Ice Cube:
Neutrino Observatory, also sensitive to extremely high energy cosmic rays

…the completed to the still in planning…



Infrared Processing and Analysis Center:
NASA's archive for a host of IR/sub-mm astronomy missions, galaxy catalogs, Keck Observatory, and more!



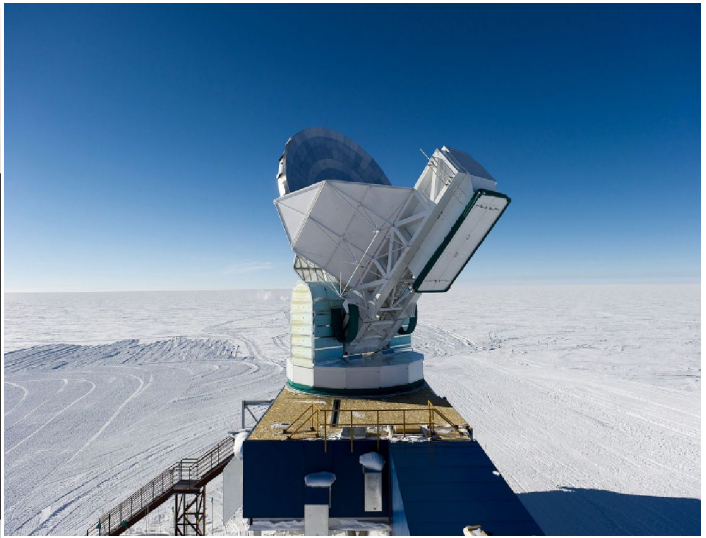
[Tarun Souradeep (2016), Beyond Gravitational wave discoveries with LIGO-India]

LIGO India:
Additional detector will greatly Improve localization of gravitational wave sources

...and working in all corners of the globe.

VERITAS:
4 12m Cerenkov telescopes for gamma ray astronomy: Arizona, USA

XENON1T:
Dark matter detector at Gran Sasso National Laboratory, Italy



South Pole Telescope:
Microwave-millimeter telescope

And in space!

Alpha Magnetic Spectrometer (AMS) mounted at the ISS
Photo credit: NASA