High Throughput Computing on the Open Science Grid

Mats Rynge - USC Information Sciences Institute

Open Science Grid

Opening Questions

- What is *high throughput computing (HTC)*?
- What is the Open Science Grid (OSG) and how is it organized?
- What capabilities does it provide to researchers and member organizations?
- Interesting use cases?

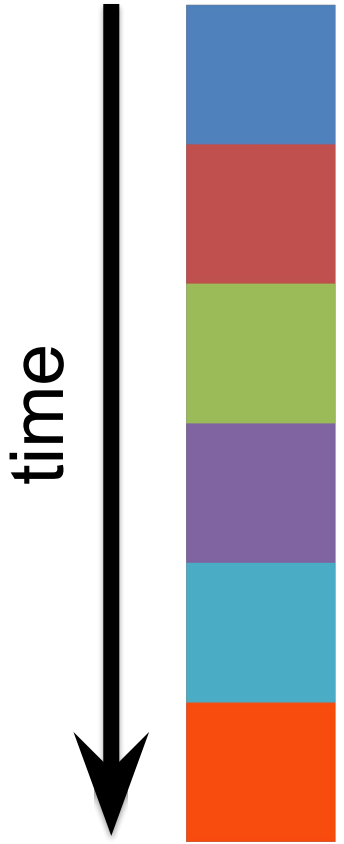
Open Science Grid

Serial Computing

What many programs look like:

- Serial execution, use one processor (CPU core)
- More complex tasks or more individual tasks → significantly longer overall compute

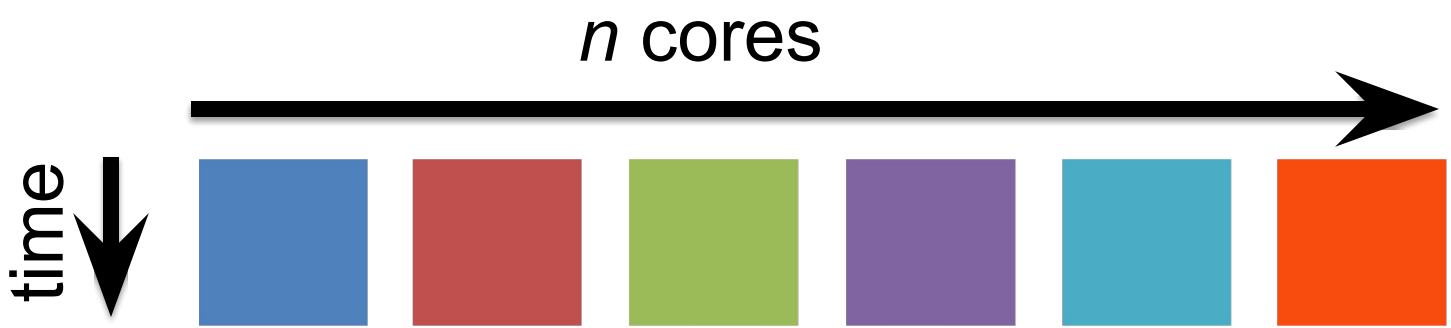
- ***How can you speed things up?***



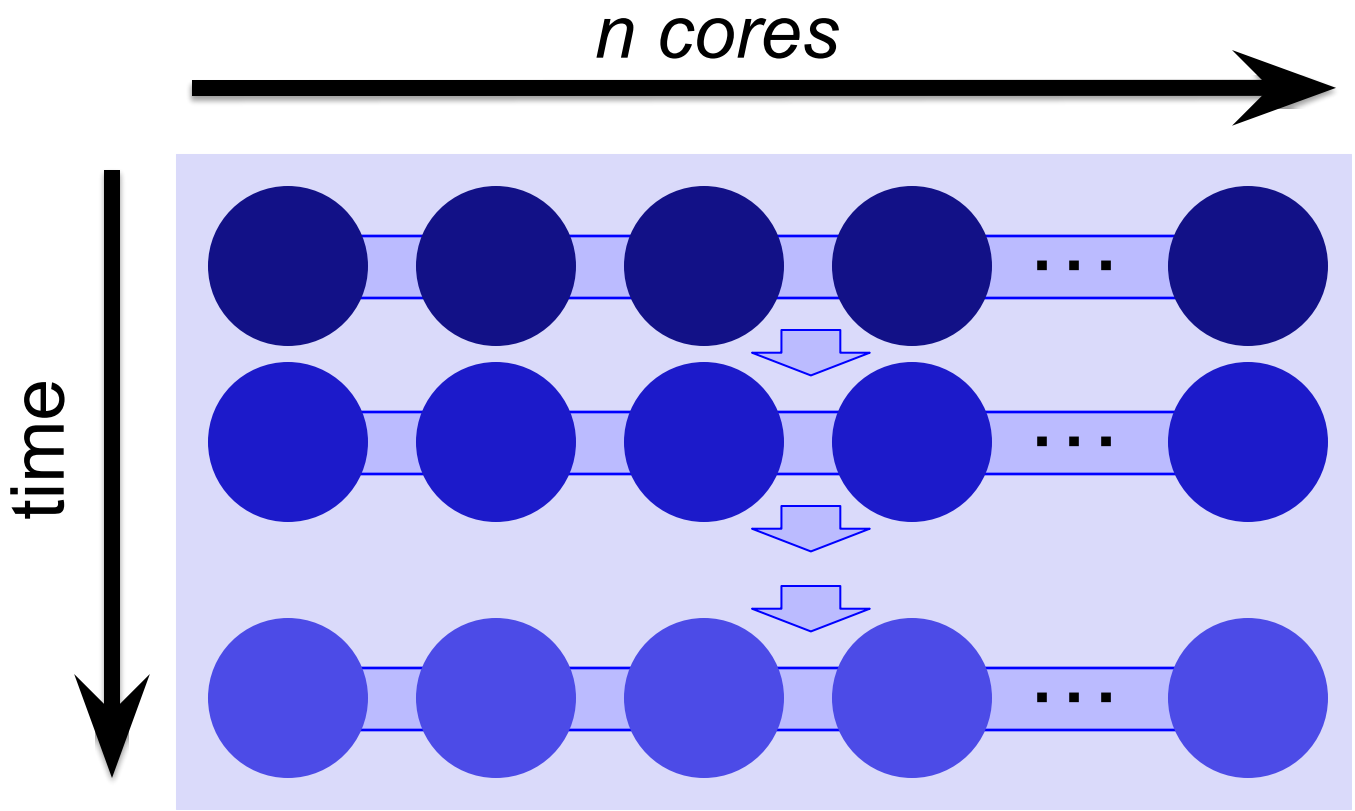
High Throughput Computing (HTC)

Parallelize!

Independent tasks run on different cores

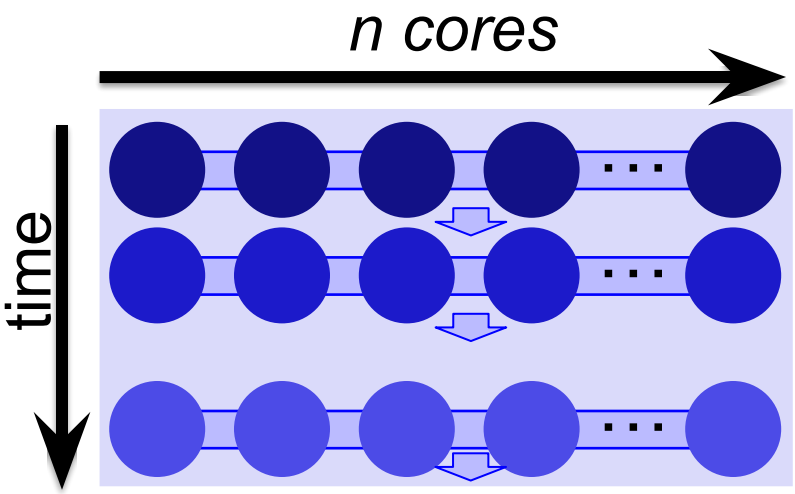


High Performance Computing (HPC)



High Performance Computing (HPC)

- Benefits greatly from:
  - CPU speed + homogeneity
  - Shared filesystems
  - Fast, expensive networking (e.g. Infiniband) and servers co-located
- Scheduling: **Must wait until all processors are available**, *at the same time* and *for the full duration*
- Requires special programming (MP/MPI)



High Throughput Computing (HTC)

- Scheduling: only need **1 CPU core for each** (shorter wait)
- Easier recovery from failure
- No special programming required
- Number of concurrently running jobs is *more* important
- CPU speed and homogeneity are *less* important

HTC: An Analogy

Question: How do you bake the world's largest cake?

HTC: An Analogy

Answer: HTC-Style!

Many small cakes baked separately, joined together

High *Throughput* vs High *Performance*

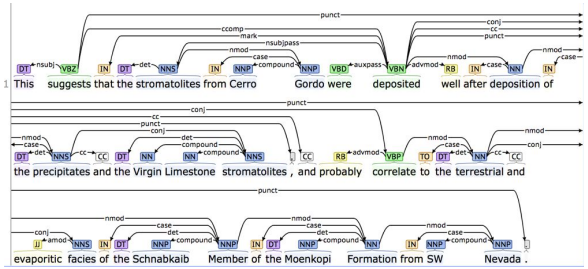
HTC

- Focus:
  - Large workflows of ***numerous***, ***relatively small***, and ***independent*** compute tasks
- More important:
  - maximized number of running tasks
- Less important:
  - CPU speed, homogeneity

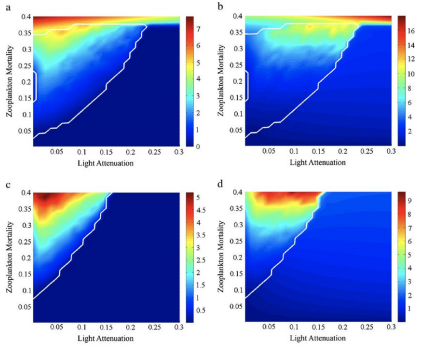
HPC

- Focus:
  - Large workflows of ***highly-interdependent*** sub-tasks
- More important:
  - persistent access to the *fastest* cores, CPU homogeneity, special coding, shared filesystems, fast networks

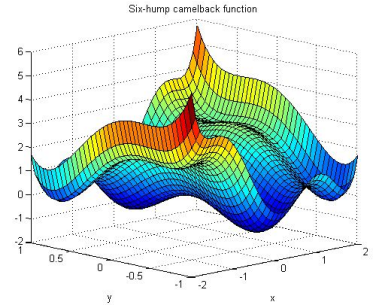
HTC Examples



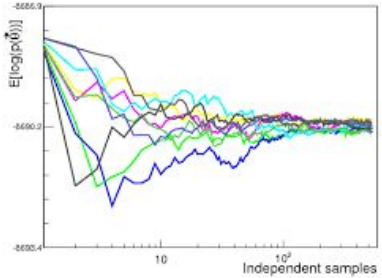
text analysis (most genomics …)



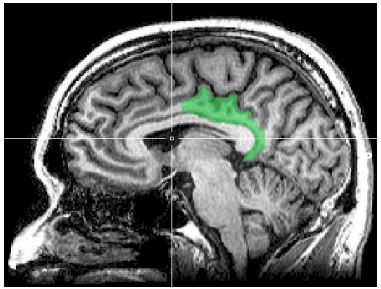
parameter sweeps



multi-start simulations



statistical model optimization (MCMC, numerical methods, etc.)



multi-image and multi-sample analysis

Open Science Grid

- **HTC for Everyone**
  ~120 member orgs
  distributed HTC (dHTC) system

- Submit jobs locally, they backfill across the country
- Jobs can be interrupted at any time (but not too frequent)

Past year:
>420 million jobs
>1.6 billion CPU hours
>200 petabytes transferred

https://www.opensciencegrid.org/

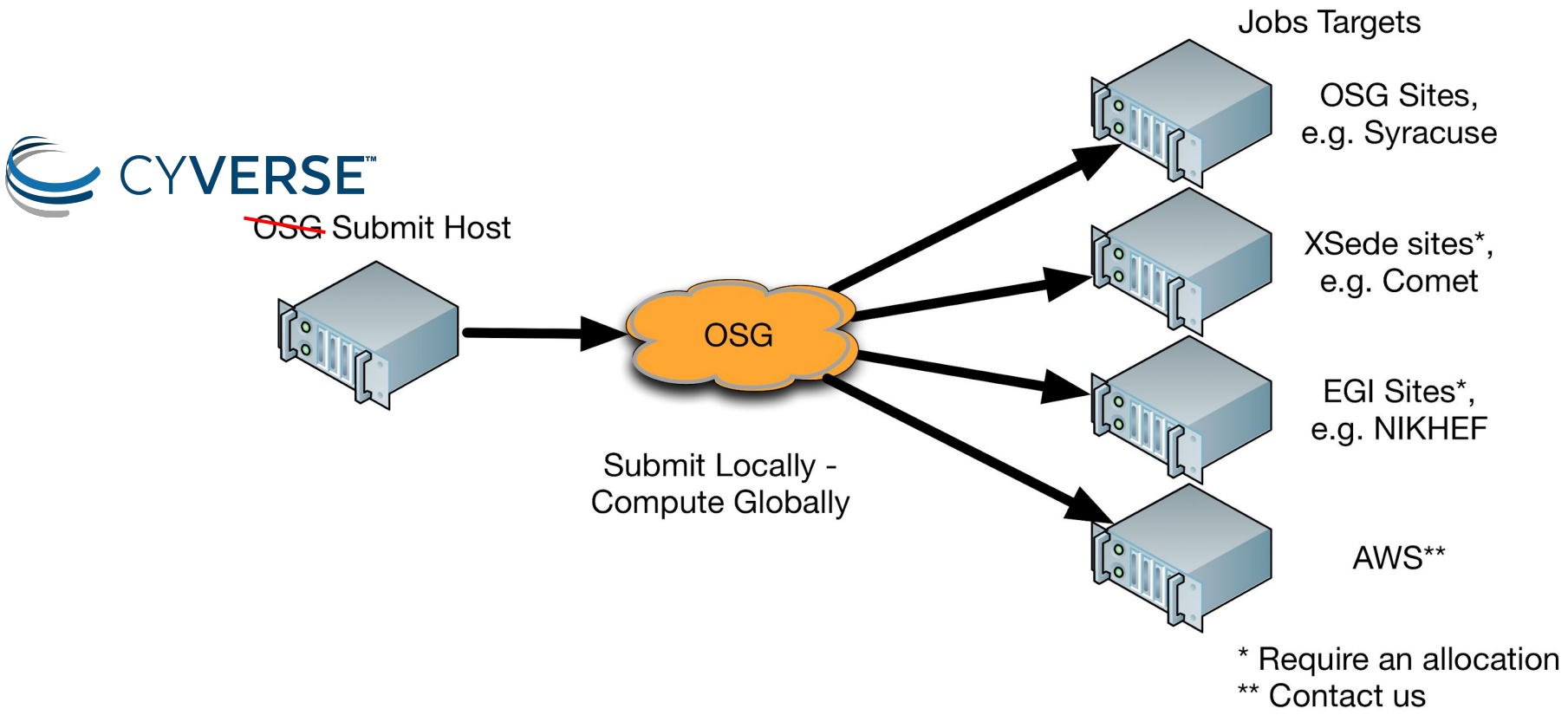
Open Science Grid

A **framework** for large scale distributed resource sharing addressing the technology, policy, and social requirements of sharing computing resources.

OSG is a **consortium** of software, service and resource providers and researchers, from universities, national laboratories and computing centers across the U.S., who together build and operate the OSG project. The project is funded by the NSF and DOE, and provides staff for managing various aspects of the OSG.

Integrates computing and storage resources from over 120 sites in the U.S.

"Submit Locally, Run Globally"



Some OSG Use Examples

- Single researcher using OSG resources from a “public” access point (OSG Connect)
- Member of a project (e.g. Cyverse) or campus (e.g. Wisconsin) submitting jobs from an organizational access points
- Campus or lab providing their resources to be backfilled by OSG jobs
- Large project (e.g. CMS, LIGO, …), provides and use resources

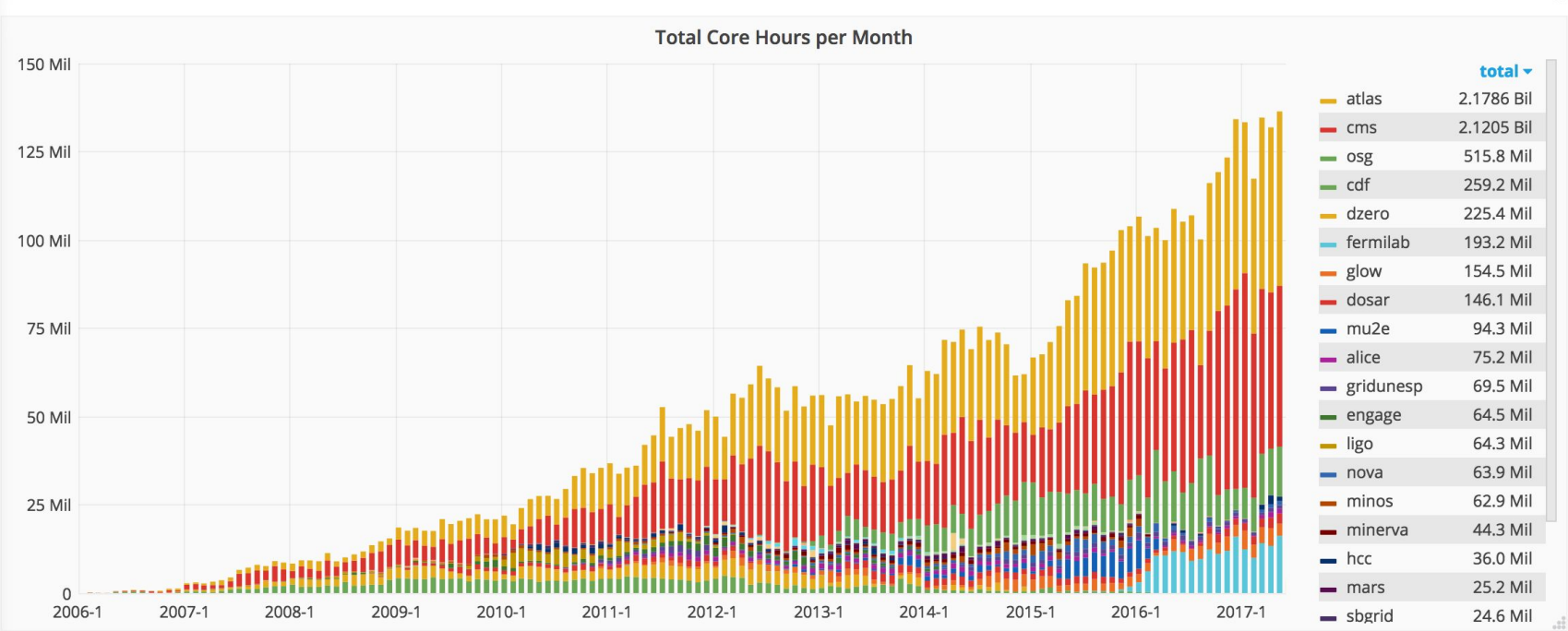
Open Science Grid

In the last 24 Hours	
278,000	Jobs
3,614,000	CPU Hours
13,859,000	Transfers
499	TB Transfers
In the last 30 Days	
8,586,000	Jobs
127,120,000	CPU Hours
0	Transfers
0	TB Transfers
In the last 12 Months	
121,418,000	Jobs
1,623,768,000	CPU Hours
681,704,000	Transfers
68,777	TB Transfers

OSG delivered across 118 sites

~ 3.5 million CPU hours delivered per day

gracc.opensciencegrid.org



Active Virtual Organizations	Active Sites	Represented Fields of Science
238	113	41

Best* dHTC Jobs for OSG

- **Runtime: 1-12 hours per-job**
- **Single-threaded**
- **Memory: <2 GB RAM**
- **Total disk I/O: <10 GB**

Rule of thumb:
What can you run on a laptop overnight?

***These are not hard limits!**

- Job checkpointing can allow for longer runtimes.
- More per-job memory and disk is certainly available, but in less numerous compute ‘slots’.
- Similarly, few-core slots are available, just in fewer numbers than single-core slots (such that single-core jobs are always advantageous).
- Extra infrastructure available for larger data (up to 10s of GBs per job).
- Software: OSG’s “OASIS” modules, existing support for Docker/Singularity.

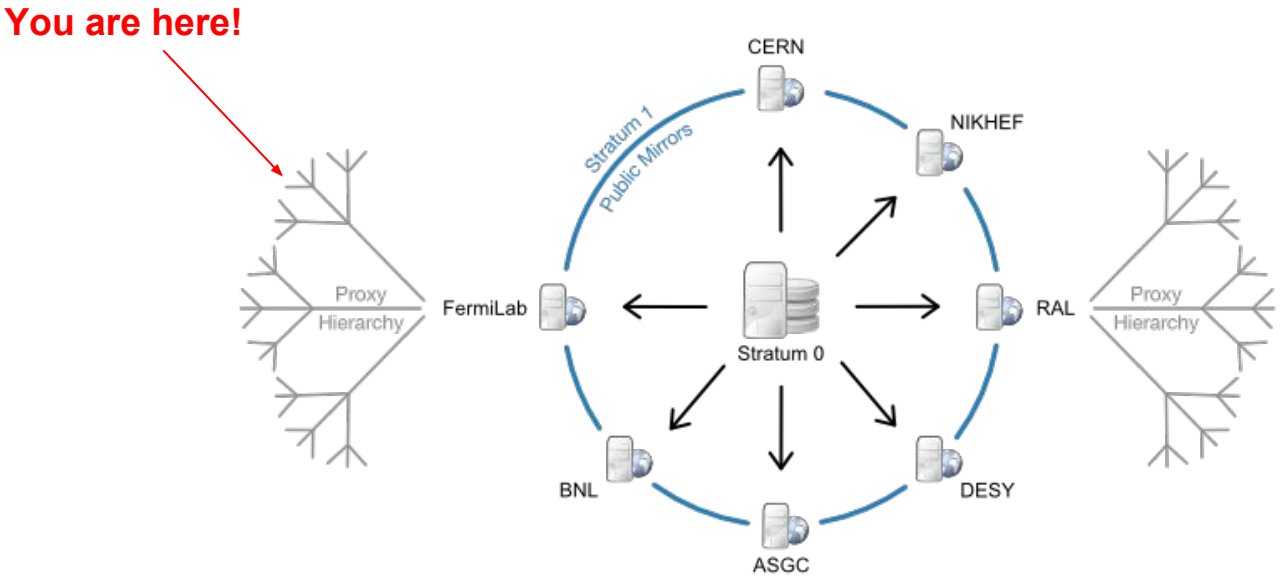
CVMFS - CERN Virtual Machine File System

"The CernVM File System provides a scalable, reliable and low-maintenance software distribution service. It was developed to assist High Energy Physics (HEP) collaborations to deploy software on the worldwide-distributed computing infrastructure used to run data processing applications. CernVM-FS is implemented as a POSIX read-only file system in user space (a FUSE module). Files and directories are hosted on standard web servers and mounted in the universal namespace /cvmfs."

Used for software and data!

Heavily cached, read-only

Available across OSG, EGI, some XSEDE resources



CVMS Repositories

```
/cvmfs/
      ams.cern.ch
      atlas.cern.ch
      cms.cern.ch
      connect.opensciencegrid.org
      gwosc.osgstorage.org
      icecube.opensciencegrid.org
      ligo-containers.opensciencegrid.org
      nexo.opensciencegrid.org
      oasis.opensciencegrid.org                <- "modules" software
      singularity.opensciencegrid.org          <- containers (next slide)
      snoplus.egi.eu
      spt.opensciencegrid.org
      stash.osgstorage.org                     <- ~1PB of user published data
      veritas.opensciencegrid.org
      xenon.opensciencegrid.org
```

500k container instances / day

OSG stores container images on CVMFS in extracted form. That is, we take the Docker image layers or the Singularity img/simg files and export them onto CVMFS. For example, ls on one of the containers looks similar to ls / on any Linux machine:

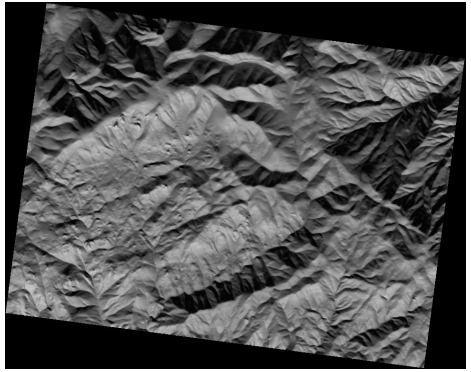
```
$ ls /cvmfs/singularity.opensciencegrid.org/opensciencegrid/osgvo-el7:latest/
cvmfs  host-libs  proc  sys  anaconda-post.log     lib64
dev    media      root  tmp  bin                  sbin
etc    mnt        run   usr  image-build-info.txt  singularity
home   opt        srv   var  lib
```

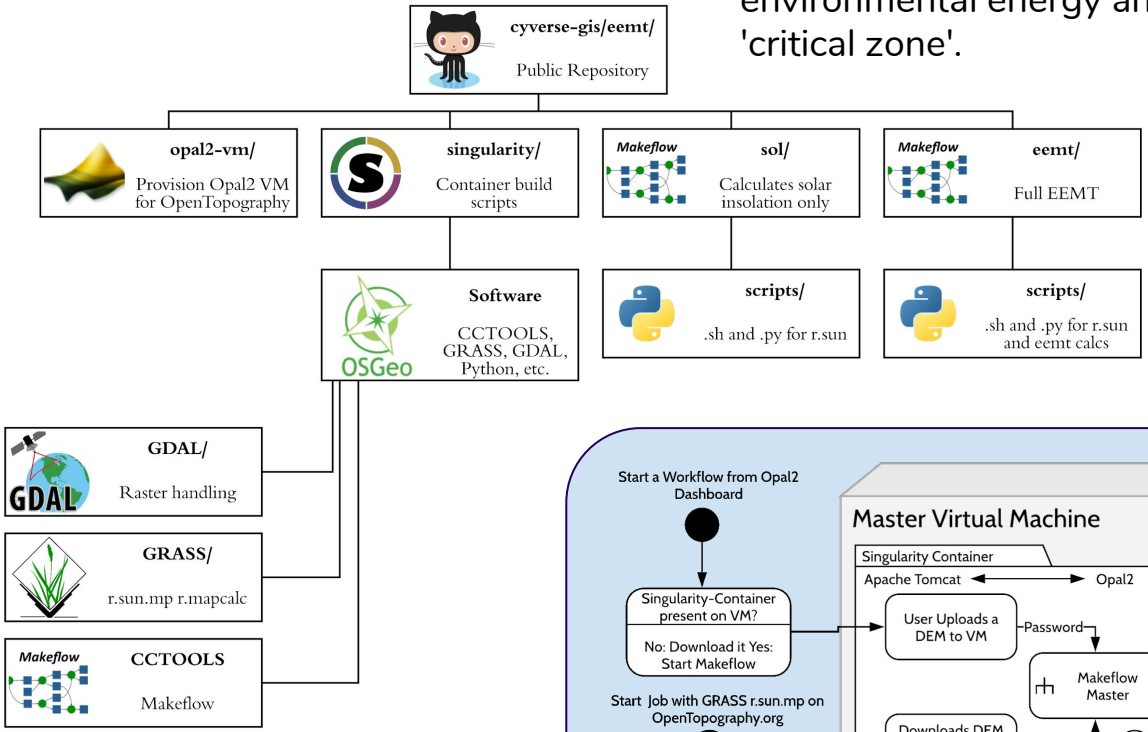
Result: Most container instances only use **a small part** of the container image **(50-150 MB)** and that part is **cached** in CVMFS!

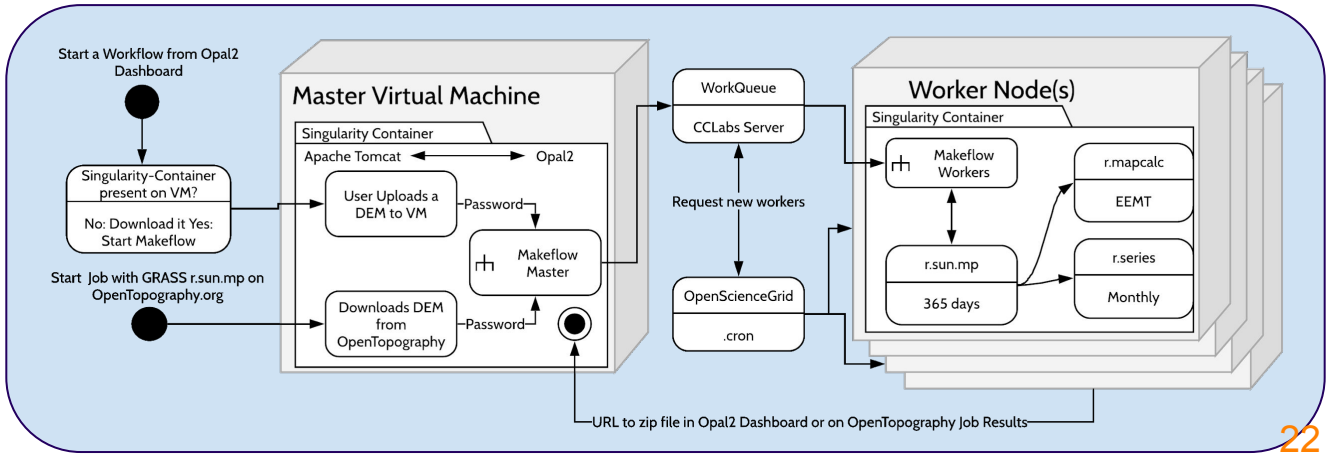
EEMT / SOL

Tyson Swetnam / Jon Pelletier

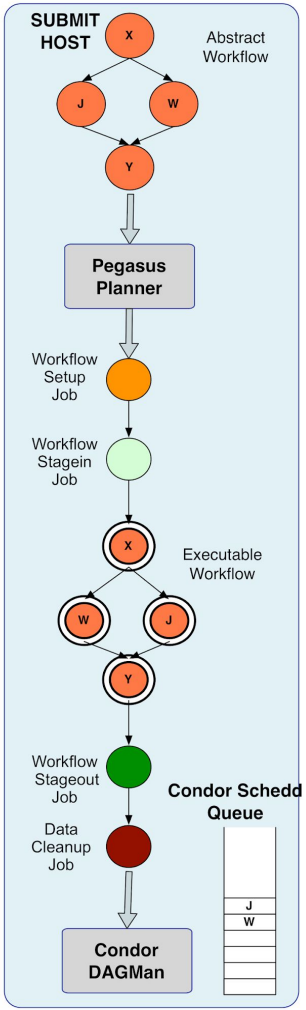
Effective Energy and Mass Transfer - a representation of environmental energy and mass transfer doing work on the Earth's 'critical zone'.

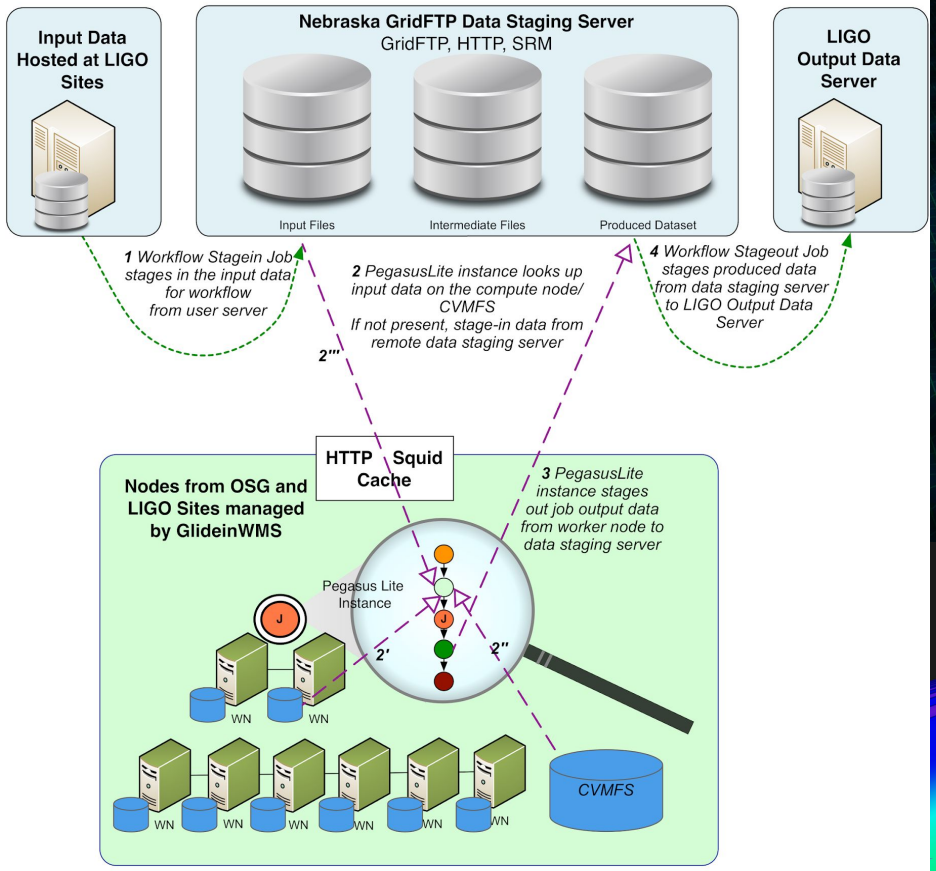






Data Flow for LIGO Pegasus Workflows in OSG





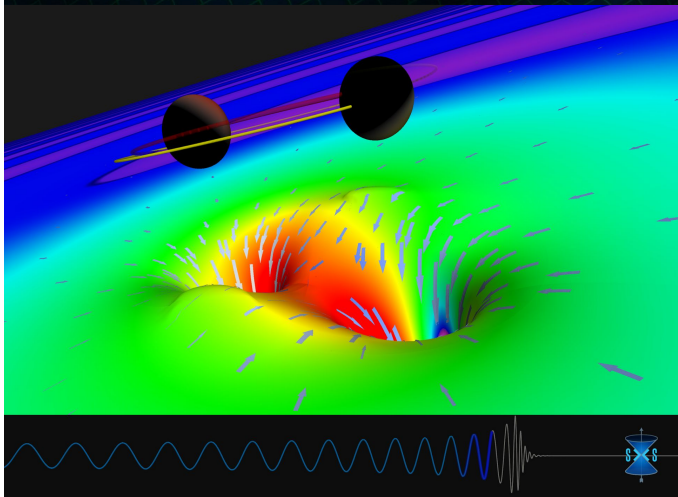
LEGEND

- Directory Setup Job
- Data Stageout Job
- Pegasus Lite Compute Job
- Data Stagein Job
- Directory Cleanup Job
- Worker Node

Advanced LIGO – Laser Interferometer Gravitational Wave Observatory

60,000 compute tasks
Input Data: 5000 files (10GB total)
Output Data: 60,000 files (60GB total)

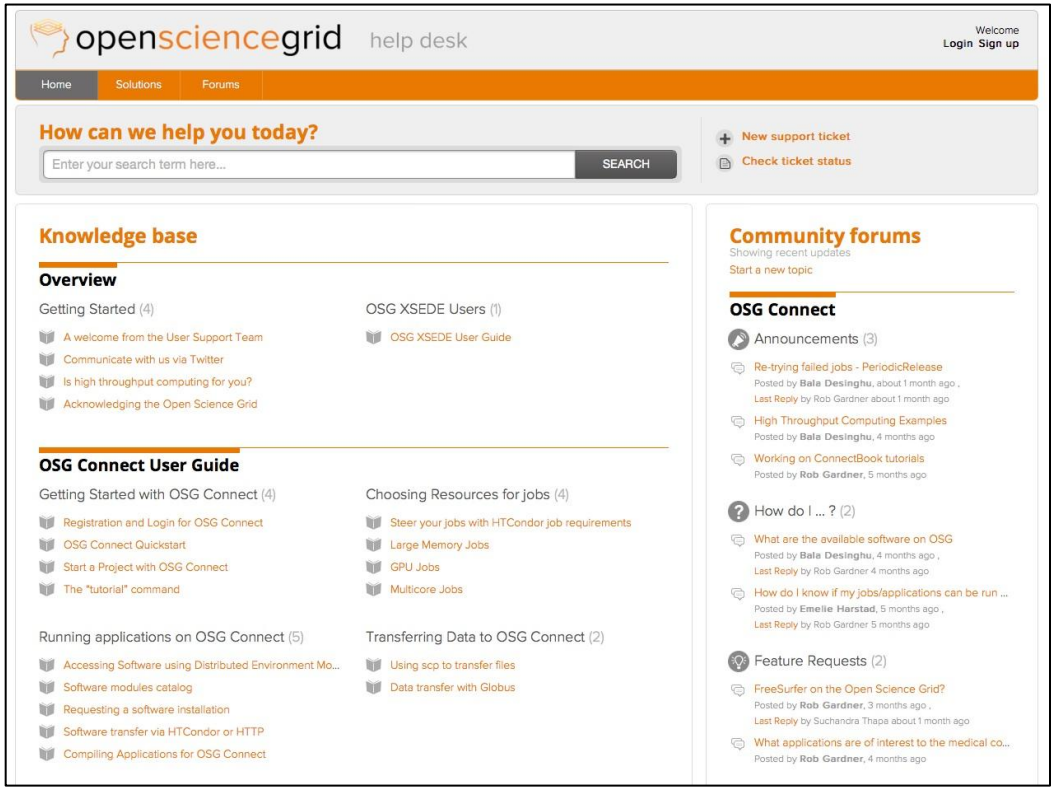
Executed on LIGO Data Grid, Open Science Grid and XSEDE



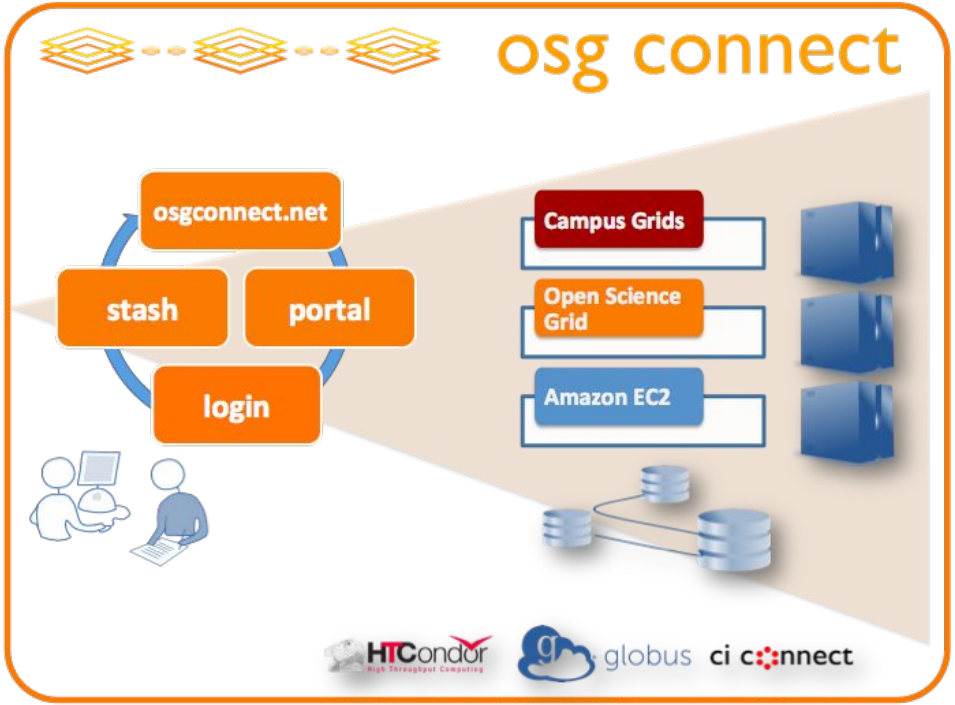
OSG User Support

https://support.opensciencegrid.org

user-support@opensciencegrid.org



OSG Connect Service



OSG Connect Provides:

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

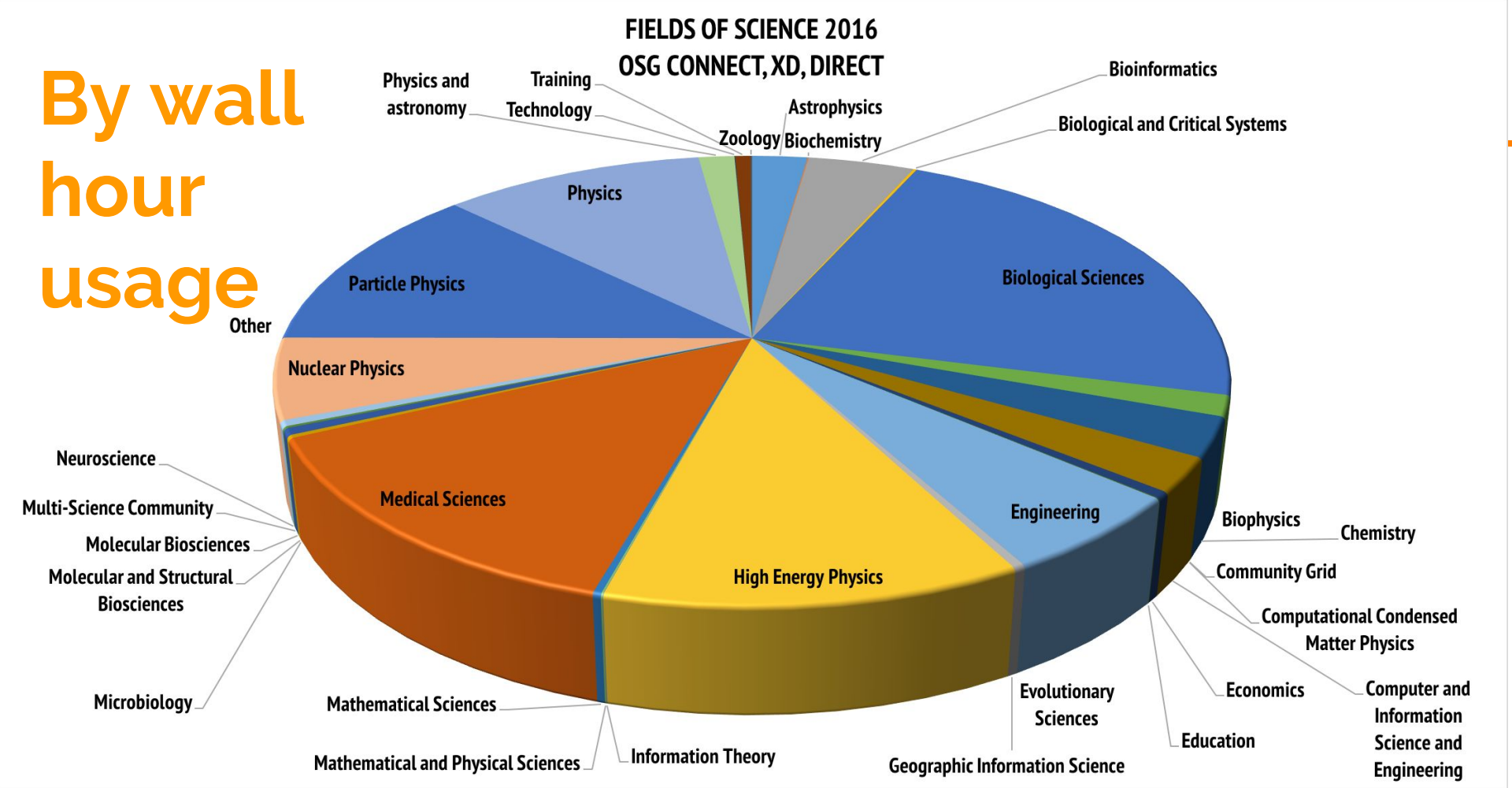
http://osgconnect.net/

Open Science Grid

OSG is Open to All

- Organization members of all scales
  - small colleges to national labs
- Research communities of all scales
  - individual students to large international research projects (e.g. CMS, LIGO, etc.)
- Open to any org's business model
  - fair-share, allocation, cost, pre-emption

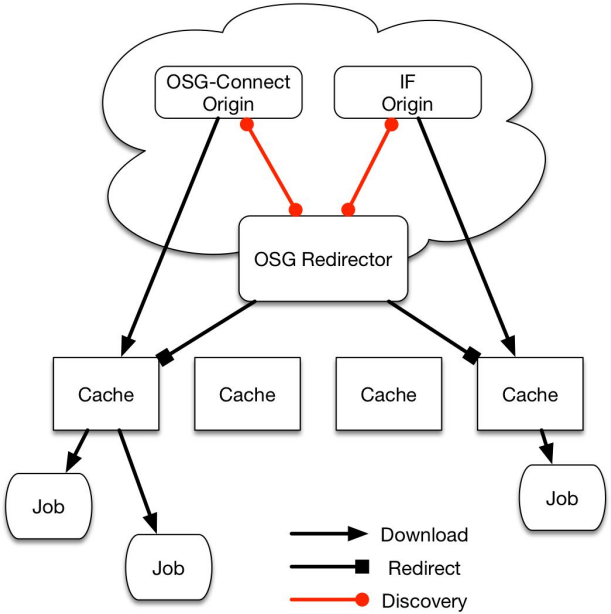
By wall hour usage

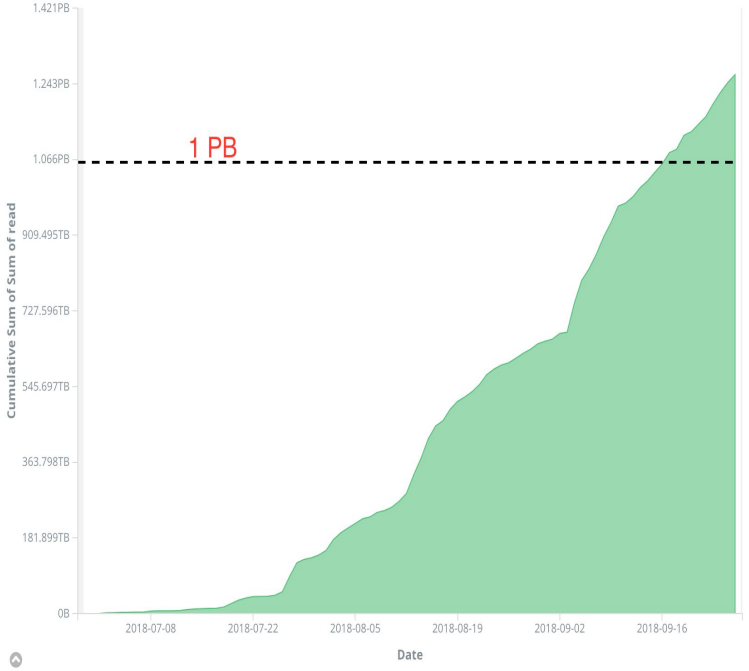


Storage service: “Stash”

- Provide a quasi-transient storage service for job input/output data
- **POSIX** access provided to the login host
- **Globus Online** Server for managed transfers from campus data services
- Personalized **http** service endpoint
- Can now handle writes!
- Connected to 100 Gbps SciDMZ (I2, ESnet)

StashCache





https://derekweitzel.com/2018/09/26/stashcache-by-the-numbers/