Services in Support of Multi-Institutional Science

2019 Joint HSF/OSG/WLCG Workshop
March 18-22 2019

Pascal Paschos, Ph.D.
University of Chicago
Outline

- Services at Scale
- OSG Connect, *-Connect Services
- Hosted CE Services
- Data Management Service: Rucio
- StashCache Data Federation
- VC3: Virtual Clusters for Community Computation
- SLATE
- Stratum-R
- GRACC
- Expertise as a service
- Summary
Services at scale

- Multi-institutional and multi-disciplinary academic research is complex and stratified creating barriers for integrated service deployment and support
  - Policies and infrastructure vary from site to site
  - Lack of strategic vision and fear of complex solutions
  - Research is collaborative but often siloed in

- Services that enable and support research must be engaging, innovative, adaptable, elegant and have strategic depth in vision and capacity, while respecting the individuality and priorities of local institutions

- Open Science Grid and its partners can bring infrastructure supported science to academic research and enable coordination between hubs of scientific research at scale: from a single user to single user engagement to large collaborative projects
OSG Connect, *-Connect

- Help Desk and Collaboration Support
- Software Support
  - OASIS (OSG Application Software Installation Service via CVMFS)
  - Singularity containers
- Data Management: Storage and Transfer
  - HTCondor, StashCache, GridFTP, Globus
- Collaborative or Individual Projects
  - Consultation services
  - XSEDE user access of OSG infrastructure through the OSG-XD login
- Submission instances deployed for specific collaborations
  - XENON, SPT-3G
  - CMS Connect, ATLAS Connect
Hosted CE services

- An OSG Compute Element (CE) allows factory submitted pilot jobs from the grid to land and execute on a local cluster. A gateway installed software authenticates and authorizes incoming pilot job containers which then run on the local resource.
- Setting up a CE at the front end of a local cluster is *not* an easy process.
- Academic Clusters do not necessarily use HTCondor as a scheduler and local IT teams are unlikely to adopt a scheduler they are not familiar with.
Hosted CE services

- To facilitate onboarding communities of researchers to OSG without heavy local admin involvement a facility can request an OSG Hosted Compute Element
- Local cluster installs HTCondorCE which runs condor_job_router to transform jobs for the local batch system
- The Hosted CE then submits pilot jobs to the local cluster over ssh
- There are at present about 20 Hosted CE services at UC connecting OSG to institutions and facilities

Requirements:
- Single OSG service account has access to the login server
- Cluster scheduler server and nodes have outbound IP connectivity
- Supported operating system: CentOS/RHEL 6.x, 7.x
- Cluster login should have common batch scheduler deployed (e.g. SLURM, PBS, HTCondor)
GRACC reporting of select Hosted-CEs

Total Wall Hours: 246 K
Total CPU Hours: 169 K
Total Jobs: 359 K
For XENON1T, needed to work with storage providers from EU, US, and Israel

Needed data management capabilities for the long term

Important to adopt/join an open source community
Data Management Service: Rucio

- A scalable data management solution for multi-disciplinary research (e.g., HEP, astronomy, biology).
- Open source, developed for ATLAS experiment at CERN but driven by a community of scientific collaborations
- Works for OSG, EGI, and, cloud storage providers
- Provides namespace (file catalog), reliable transfer service (FTS), storage endpoints, consistency service
- Subscription model for data placement with declarative "rules"
- Adopted by data driven communities for their data management and replication services
- Deployed or evaluation instances for Xenon, CMS, IceCube, LIGO, LSST, FIFE etc
Production Rucio for the Xenon Experiment

- EGI storage distributes data during large data collection periods at the INFN Gran Sasso National Laboratory
- Jobs launched from the xenon node at UChicago. Jobs move through glideinWMS to OSG.
- If lands at EGI, pulls data from EGI; Same with OSG
- Storage and compute pool is expandable
- Rucio/FTS automates the data movement
XENON OSG/EGI/XSEDE Usage

- Combined resources from shared OSG cyberinfrastructure, XSEDE allocations, and allocations on European grid sites
- Up to 2M hours per month!
StashCache data federation enables the scalable distribution of research data to large volume of jobs that need them negating the need to prestage data or deploy expensive local storage infrastructure.

- **Origin**: Authoritative copy of the data.
- **Cache**: Transfers data for jobs. A number of caches are operationalized across the OSG aiming for proximity to a compute site which can operate it’s own cache.
- **Redirector**: Centrally run service by OSG which instructs the cache to access the proper origin for data.
StashCache - over 800 TB read last month
Virtual Clusters for Collaborations

- VC3 ([virtualclusters.org](http://virtualclusters.org)) is a platform for provisioning cluster frameworks for collaborative science teams.
- Collaborations who have existing HTCondor pools can extend them by adding more worker nodes via VC3 (manually at present).
- Add XSEDE resources, Open Science Grid, and campus HPC clusters.
- End-users can transparently use additional resources.
- South Pole Telescope (SPT-3G), XENON1T Analysis Framework, IceCube Simulation Framework early users.
VC3 Deployment example

- Tier-3 campus deployment via VC3 which allows the provisioning at user-level - no root access - of:
  - The CERN File System (CVMFS) (via parrot virtual filesystem)
  - The OSG grid environment on the worker nodes (via CVMFS)
  - Customized Operating Systems (via singularity)
  - Integrated OSG Compute Element (CE) with the VC3 submit host
SLATE: Services Layer At The Edge

- A DevOps friendly Kubernetes based distributed service orchestration platform
- **Federation software & model** for independent Kubernetes clusters
- Site autonomous, allows for user controlled Kubernetes configuration
- Single entrypoint using institutional identity with UNIX-like permissions model
- Scale resources on demand
- Applications are curated in a catalog
- An infrastructure, and software
Security, Policy on Federated Edge Systems

- Application developers create Helm charts and Docker images
- The SLATE team is developing a trust model for acceptance of these containerized "apps"
- Curated application catalog provides point of review for widespread deployment
SLATE Deployment

- **Trusted group model**: experts ("app admins") from a stakeholder or a service organization can deploy and operate services across multiple sites
- Early deployments for ATLAS XCache - considered for WLCG "Lightweight Sites" (c.f. Thurs session)
- Working with OSG & PRP to **containerize & curate** services to more easily integrate campus resources
  - frontier-squid
  - HTCondor services: CE, submit
  - StashCache
- Other "apps" e.g. JupyterHub useful for training events
Stratum-R overview

- Allow worker nodes to access CVMFS repos on sites without CVMFS client
  - Repos unpacked by server and stored as Linux FS at target site
  - Particularly useful for HPC sites where CVMFS client impossible
- Problem: Direct replication of CVMFS repos to Linux FS not feasible
  - Large repo size: Limited by network latency, load on server
- Solution: Build Stratum-1 server, add local client to use server as source
  - Parallel rsync to unpack repos into FS (or container images) Second rsync from Stratum R to HPC edge node
  - FS mounted on all worker nodes, symlink between repo area and cvmfs
  - Voila! User can log into remote site and use repos as normal
Stratum-R deployment Example (US-ATLAS on Stampede2)

- Uchicago Stratum-R EL6 server
  - 8 CPU, 32 GB
  - 10GB NIC Jumbo Frames MTU 9000
  - 64 TB on a MD1000
- Unpack common CVMFS software repos into FS

X509 parallel sync

- Repo mounted to an edge node
- Users access to cvfms, e.g. /cvfms/atlas.cern.ch/repo
- Repos (ATLAS, CMS, OSG, IceCube…) updated via cron jobs
GRACC Grid Accounting Service

Total Core Hours per Month

Active Virtual Organizations: 56
Active Sites: 2018
Represented Fields of Science: 24
GRACC Grid Accounting Service

- Provides a view of grid utilization metrics which are collected, indexed and displayed over a Grafana interface
- Allows for inspecting summary statistics at scale, from individual users to projects, to facilities and collaborations
- Enhances multi-institutional engagements by allowing for a data driven evaluation of Grid resource utilization, performance, comparison and added-value in leveraging a shared resource
- Enables research domain specific metrics to be captured and repurposed in support of proposals and strategic planning by facilities, administrators and researchers
Central Collector Service

The central collector services allows for a real time monitoring of the status of pilot jobs on the HTCondor-CEs in the OSG.

Contains information about site queues - number cores, max memory, walltime, etc.
Expertise as a service

- Availability of technical expertise safeguards the successful deployment and operationalization of other services
- Training of researchers in expertise development improves efficiency and coordination between end-points and centrally managed services
- Research and the path to discovery is not enabled by just a single technological solution but by architecting a synergy of service elements
- Example: South Pole Telescope (SPT) experiment
Example: SPT (South Pole Telescope)

- OSG staff setup and maintain infrastructure for the collaboration at the South Pole
- Provision and maintain Stash node, software repo, manage data transfers, online analysis and storage capacity
- These services come together in enabling the success of this multi-institutional project and streamline the operational robustness of the computational and data management workflow
Summary

- Services presented today
  - Aim in advancing shared infrastructure deployment and enable collaborations
  - Remove the visibility of layers of complexity, promote and empower academic engagement and foster the potential of growth in a restrictive funding landscape
- Seek improvement and increased adoption rates through success stories