Buzzard: Georgia Tech's Foray into the Open Science Grid

Mehmet Belgin*
Semir Sarajlic*
Ruben Lara
Laura Cadonati
Nepomuk Otte
Ignacio J. Taboada
Gregory L. Beyer
Norman B Bonner
mehmet.belgin@oit.gatech.edu
semir.sarajlic@oit.gatech.edu
ruben.lara@oit.gatech.edu
laura.cadonati@cos.gatech.edu
otte@gatech.edu
itaboada@gatech.edu
gbeyer3@gatech.edu
bray.bonner@oit.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Michael Brandon
Pam Buffington
J. Eric Coulter
Aaron Jezghani
David Leonard
Fang (Cherry) Liu
Paul D. Manno
Craig A Moseley
mbrandon@gatech.edu
pam@gatech.edu
j.eric@gatech.edu
ajezghani3@gatech.edu
david.leonard@oit.gatech.edu
fang.liu@gatech.edu
pmanno@gatech.edu
craig.moseley@oit.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Trever C. Nightingale
Ronald Rahaman
Kenneth J. Suda
Peter Wan
Michael D. Weiner
Deirdre Womack
Dan Zhou
Marian Zvada
tnightingale6@gatech.edu
rrahaman6@gatech.edu
ksuda3@gatech.edu
peter.wan@oit.gatech.edu
mweiner3@gatech.edu
dwomack30@gatech.edu
dzhou62@gatech.edu
zvada@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Andre C McNeill
andre.mcneill@honeywell.com
Honeywell International INC
Atlanta, GA, USA

Neil C. Bright
neil.bright@nesi.org.nz
New Zealand eScience Infrastructure (NeSI)
Auckland, NZ

Robert W Gardner
Paschalis Paschos
rwg@uchicago.edu
paschos@uchicago.edu
University of Chicago Enrico Fermi Institute
Chicago, IL, USA

Lincoln Andrew Bryant
Judith Lorraine Stephen
lincolnb@uchicago.edu
jlstephen@uchicago.edu
University of Chicago Enrico Fermi Institute
Chicago, IL, USA

James Alexander Clark
Peter F. Couvares
james.clark@ligo.org
pcouvare@caltech.edu
LIGO Laboratory, California Institute of Technology
Pasadena, CA, USA

Brian Hua Lin
Todd Tannenbaum
Gregory Thain
blin@cs.wisc.edu
tannenba@cs.wisc.edu
gthain@cs.wisc.edu
University of Wisconsin-Madison
Madison, Wisconsin, USA

ABSTRACT

Open Science Grid (OSG) is a consortium that enables many scientific breakthroughs by providing researchers with access to shared High Throughput Computing (HTC) compute clusters in support of large-scale collaborative research. To meet the demand on campus, Georgia Institute of Technology (GT)'s Partnership for an Advanced Computing Environment (PACE) team launched a centralized OSG support project, powered by Buzzard, an NSF-funded OSG cluster. We describe Buzzard's unique multi-tenant architecture, which supports multiple projects on a single CPU/GPU pool, for the benefit of other institutions considering a similar approach to support OSG on their campuses.

*Corresponding author

PEARC '22, July 10–14, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9161-0/22/07.
https://doi.org/10.1145/3491418.3535135

CCS CONCEPTS

• **Computer systems organization** → **Distributed architectures**; **Grid computing**; • **General and reference** → *Design*.

KEYWORDS

Open Science Grid, HTCondor, High Throughput Computing, StashCache

ACM Reference Format:

Mehmet Belgin, Semir Sarajlic, Ruben Lara, Laura Cadonati, Nepomuk Otte, Ignacio J. Taboada, Gregory L. Beyer, Norman B Bonner, Michael Brandon, Pam Buffington, J. Eric Coulter, Aaron Jezghani, David Leonard, Fang (Cherry) Liu, Paul D. Manno, Craig A Moseley, Trever C. Nightingale, Ronald Rahaman, Kenneth J. Suda, Peter Wan, Michael D. Weiner, Deirdre Womack, Dan Zhou, Marian Zvada, Andre C McNeill, Neil C. Bright, Robert W Gardner, Paschalis Paschos, Lincoln Andrew Bryant, Judith Lorraine Stephen, James Alexander Clark, Peter F. Couvares, Brian Hua Lin, Todd Tannenbaum, and Gregory Thain. 2022. Buzzard: Georgia Tech's Foray into the Open Science Grid. In *Practice and Experience in Advanced Research Computing (PEARC '22), July 10–14, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3491418.3535135

1 INTRODUCTION

We are witnessing drastic changes in Scientific Computing practices in response to increasingly collaborative and multi-disciplinary projects happening at national and international scale. While conventional High Performance Computing (HPC) capabilities remain critical for many domains, current trends indicate a large-scale adoption of very large numbers of relatively small and short computations, which is commonly referred to as High Throughput Computing (HTC). Open Science Grid (OSG) is a unique consortium that enables access to a wide array of member HTC compute clusters in the nation, with unified authentication, software and data capabilities [8]. OSG has been instrumental in ground-breaking scientific advancements, including the Nobel Prize-winning Gravitational Waves research (LIGO) [5] and countless other projects.

Georgia Institute of Technology (GT)'s Partnership for an Advanced Computing Environment (PACE) team [7] manages sustainable leading-edge advanced research computing resources and provides GT researchers, their external collaborators and students with research facilitation, consultation and technical support services. At PACE, we have been observing a consistent demand for HTC workloads. An analysis of the job characteristics between 2015 and 2019 yielded that 78% of the jobs submissions requested one or two cores (Figure 1(a)). A more recent analysis in 2022 confirmed continuation of this trend (Figure 1(b)). Additionally, several GT professors identified an OSG ecosystem on campus as a critical requirement for the continuation of their work. Considering all of these factors, the PACE team started a new project to establish campus-wide support for OSG, partially funded by a National Science Foundation (NSF) award "CC* Compute: Integrating Georgia Tech into the Open Science Grid for Multi-Messenger Astrophysics (1925541)" in 2019[1]. The funds from this award are used to build the "Buzzard" OSG cluster, which is fully dedicated to OSG. Buzzard is preceded by another cluster dedicated to the LIGO project, so it is not the first PACE OSG resource, yet it's the first that defines our new campus-level service that's open for all OSG projects, using a single pool of CPU and GPU nodes. Buzzard sets a milestone for

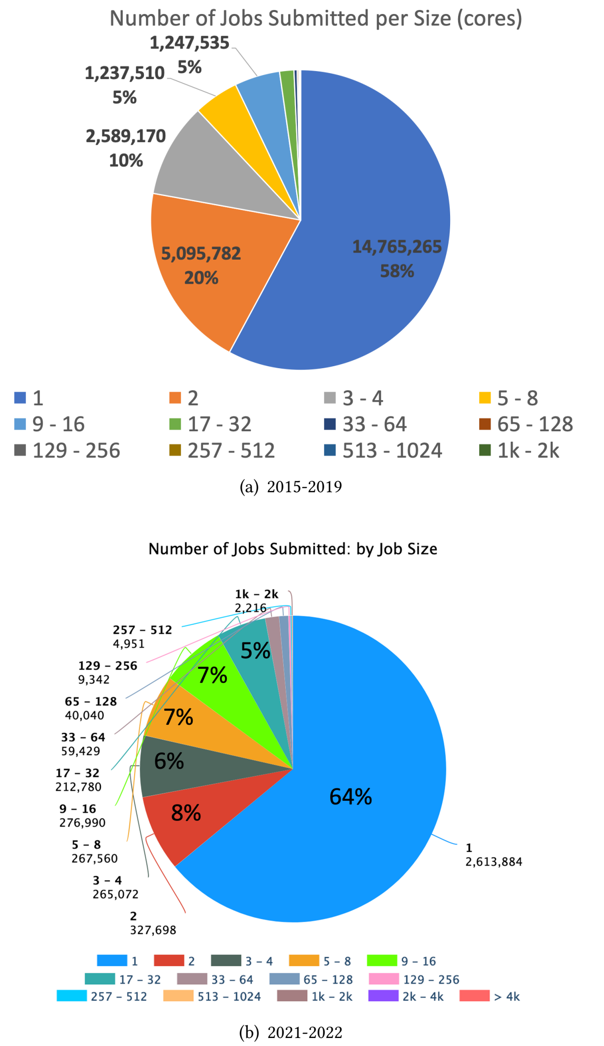


Figure 1: The number of requested cores for job submissions between (a) 2015-2019 and (b) 2021-2022 on PACE, based on XDMoD data. The majority of submissions are HTC-eligible small jobs as a continued trend.

PACE's strong commitment to support the entire OSG community, both locally and externally.

We share with the broader community our approach for supporting OSG at the campus level and beyond, including the unique 'multi-tenant' configuration of Buzzard to support multiple OSG projects with different priorities, using a single mixed pool of CPU and GPU nodes.

2 PACE'S CENTRALIZED OSG SERVICE FOR GEORGIA TECH

PACE team's involvement with OSG goes back to the deployment of a cluster dedicated to LIGO research in 2016. This cluster was a highly customized resource, deployed with an ad-hoc approach.

[1]This material is based upon work supported by the National Science Foundation under grant number 1925541. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

PACE's demonstrated success with this cluster received attention from campus and increased demand for OSG. The new project, powered by the Buzzard cluster, represents PACE's foray into a centralized OSG service.

2.1 Buzzard Cluster

The Buzzard cluster is a significant enabler of PACE's centralized OSG support, regardless of its compute power today. Buzzard readily supports multiple job pools for LIGO, IceCube, CTA/VERITAS and Open Science Pool (OSPool) Virtual Organizations, and more importantly makes addition of new resources for these pools, as well as addition of new VOs, as easy as expanding conventional HPC resources. By adding this new service in its standard portfolio, PACE is committed to refresh and grow Buzzard through new grants and/or its own budget well beyond the scope and lifetime of the original NSF project. Buzzard has already been expanded by several PIs on the proposal with new LIGO compute nodes and additional storage for the CTA/VERITAS, and we expect see further adoption from the rest of the campus.

It's important to note that 100% of the resources acquired by this award are dedicated to OSG, making no distinction between GT and external researchers, as long as they are members of the same VO. The only additional advantage GT users have is the ability to use a local submit node that mounts their private storage, so they can compute directly on their private data.

2.2 Buzzard's StashCache Service

StashCache is a distributed caching federation comprised of data origins, redirectors, and caches, deployed in geographically-distributed locations across the U.S. and Europe [9]. We deployed the first OSG StashCache in the Southeast US to provide Buzzard and other regional institutions with fast access to OSG data. Due to its geographic location, our StashCache also serves requests from several international OSG sites as well, including Puerto Rico, Brazil, Mexico, Spain and France. Buzzard users, both local and external, can access public StashCache datasets and software via the CernVM File System (CVMFS) [3], a scalable and reliable software distribution service. CVMFS can support multiple repositories for different projects, which can be enabled/disabled with trivial configuration changes. This mechanism is also used for the distribution of Singularity containers, which have been increasingly replacing conventional distributions of the OSG software.

2.3 Scientific Impact

Bringing Buzzard into OSG opens up great opportunities with significant scientific impact. IceCube simulations of neutrinos and cosmic rays make substantial use of its GPUs. IceCube PhD Students at GT are able to use standard data analysis pipelines via OSG. Buzzard is also used to execute simulations of particle cascades in the atmosphere initiated by very-high-energy gamma rays. These simulations are crucial for the analysis of data recorded with VERITAS and, in the future, with CTA. Buzzard already serves as a storage facility for billions of air showers simulated for VERITAS on the OSG, allowing members of these projects worldwide to access and analyze the same data set without having to transfer it to their local institutions. Buzzard is now contributing significant resources to the International Gravitational-Wave Network (IGWN) pool, accessible to all LIGO-affiliated researchers through shared submit nodes. Both local and remote LIGO users are receiving a large fair share allocation on the cluster, especially after the expansion of the original Buzzard with additional LIGO nodes. This has proven beneficial for debugging with direct access to local administrators and for prototyping new analyses that are expected to contribute to the analysis of LIGO data from the next observational data. Buzzard's OSPool integration serves a large number of domains, including but not limited to Physics, Biological Sciences, Astronomy, Chemistry, Engineering, Mathematics, Computer Sciences, Education, Health, Economics, Integrative Activities, Earth and Ocean Sciences and other social sciences for users from up to 70 different campuses.

The primary objective of this project is to raise awareness and knowledge of OSG services in the entire campus and onboard as many users as possible to benefit from this significant resource as a whole, regardless of their usage of Buzzard. A crucial part of this service is, therefore, training and documentation for effective onboarding of new users. PACE offers an OSG Orientation class coupled with weekly consulting sessions, along with communications that feature OSG in articles, blog posts and symposiums. We have also identified research groups with usage patterns that reflect workflows conducive to OSG, based on their use of existing campus resources, and conducted targeted outreach to those who would benefit most.

3 BUZZARD AS A MULTI-TENANT SINGLE CPU/GPU POOL OSG CLUSTER

3.1 Buzzard Architecture

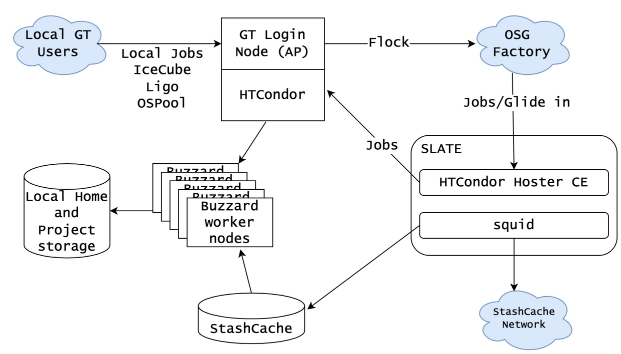


Figure 2: Buzzard Architecture to support local and external jobs for multiple VOs.

Funds from the NSF award allowed PACE to deploy Buzzard to kickstart a campus-wide OSG service, in strong collaboration with the University of Chicago's Maniac Lab [6]. The initial configuration included 12× 24-core 384GB CPU nodes, 11× 24-core 384GB GPU nodes equipped with 4× RTX6000 GPUs each, and a storage server with 200TB of usable space.

Figure 2 depicts the current design of Buzzard, which is able to support both locally-submitted jobs that either run locally or 'flock out' to OSG's central 'Glidein WMS' factory, as well as externally-submitted jobs that 'glide in' to Buzzard. The worker nodes constitute a single, mixed pool of GPUs and CPUs. This configuration allows for a local GT user to (1) login locally to the GT login node (Access Point) and submit a local job to run on Buzzard, using local storage; (2) submit an external job that flocks out to the OSG factory, which may run on other sites or come back to Buzzard depending on utilization levels; and (3) login to an external OSG login node and submit a job to target Buzzard specifically. OSG factory-routed jobs that glide into Buzzard use VO-specific service accounts, namely *osg*, *ligo* and *icecube*. We utilize the SLATE CI framework [2] by the Maniac Lab of University of Chicago to deploy and manage our OSG Compute Entrypoint (CE), as well as the Frontier Squid Service as a caching proxy.

3.2 Buzzard's Unique Multi-Tenant Mixed CPU/GPU Configuration

While there are many OSG clusters supporting OSPool, LIGO, and IceCube VOs, on CPU and GPU resources, Buzzard is unique in its ability to allow for multiple VOs sharing a *single* collection of CPU and GPU nodes. Buzzard uses HTCondor [1] for job scheduling, but unlike most other institutions, it is not integrated with a local HPC scheduler because it's an OSG-only resource.

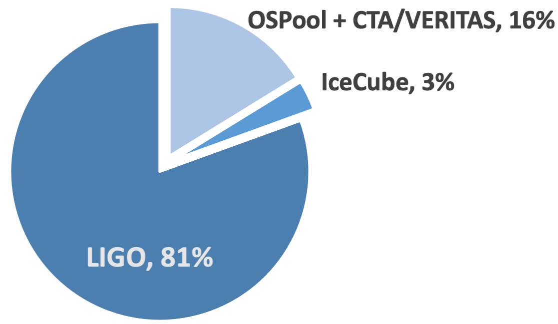
Since its initial deployment, Buzzard had been expanded with an addition of 25× 24-core 384GB CPU nodes using external funds for the LIGO project. The CTA/VERITAS workflow currently uses the OSPool, so the allocations for this project and OSPool ended up being merged. The resulting VO-specific resource allocations after these changes are shown in Figures 3(a) and 3(b).

HTCondor makes implementation of (and changes to) fairshare allocations for each VO very easy, without needing to compartmentalize any of the resources. We do not allow CPU-only jobs on GPU nodes, and instead allocate 6 cores for each GPU request by default, since each GPU node comes with 24 cores and 4 GPUs. The fairshare configuration for these allocations are kept in a file named '`condor_group_quotas.conf`' under the '`condor.d`' configuration folder, utilizing HTCondor's *negotiators* feature:

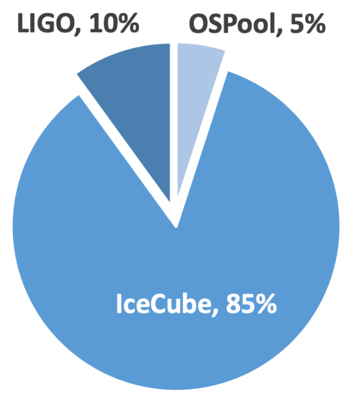
```
USE Feature: AssignAccountingGroup(/etc/condor/accounting/gt-user.map)
...
NEGOTIATOR_SLOT_CONSTRAINT=(isUndefined(TotalGPUs))
NEGOTIATOR_GPU.NEGOTIATOR_SLOT_CONSTRAINT=(TotalGPUs > 0)

# Set up group quotas for the CPU negotiator
GROUP_NAMES = icecube, ligo, osg
GROUP_QUOTA_DYNAMIC_icecube = 0.03
GROUP_QUOTA_DYNAMIC_ligo = 0.81
GROUP_QUOTA_DYNAMIC_osg = 0.16
# Set up group quotas for the GPU negotiator
NEGOTIATOR_GPU.GROUP_NAMES = icecube, ligo, osg
NEGOTIATOR_GPU.GROUP_QUOTA_DYNAMIC_icecube = 0.85
NEGOTIATOR_GPU.GROUP_QUOTA_DYNAMIC_ligo = 0.10
NEGOTIATOR_GPU.GROUP_QUOTA_DYNAMIC_osg = 0.05
```

Here, we associate each VO-specific resource pool with a group, namely *icecube*, *ligo* and *osg*. The 'osg' group represents the merged CTA/VERITAS and OSPool allocations. HTCondor doesn't support using POSIX groups for user access control, so we manage a file '`gt-user.map`' under condor's `accounting` folder, as specified in the 'Use Feature:' line above. We automated the management of this list by using a simple maker script that creates this file on a regular basis, based on the memberships we keep using POSIX groups, in the following syntax (usernames anonymized) [4].



(a) CPU Resources



(b) GPU Resources

Figure 3: Fairshare CPU and GPU allocations for each Virtual Organization (VO) supported by Buzzard. The large LIGO portion is due to the additional LIGO nodes added after Buzzard's initial deployment.

```
* user1  osg
* user2  icecube,ligo,osg
* user3  icecube
```

3.3 Drawbacks and Lessons Learned

We experienced several drawbacks during the course of this large-scale collaborative project, including COVID-19 related supply chain issues, incorrect initial design assumptions and staff changes in the team. Several lessons we've learned from this experience are (1) use virtual machines for exploring configurations early on (2) start independent tasks as early as possible (3) document as you implement (4) don't wait too long to engage experts (5) communicate expectations and deliverables in written form and (6) provide external collaborators with access to systems as possible.

4 CONCLUSION

In response to increasing demand for HTC-type workloads, PACE launched a centralized OSG service for the GT campus, powered by the Buzzard OSG cluster. This new service is an high-impact addition to PACE's portfolio because it not only facilitates access to OSG by GT researchers and students but also makes it possible and convenient to grow the nation-wide OSG resources using GT-originated funds. This service comes with training, consultation, and high visibility events, and we expect an increased adoption rate from the campus in the coming years. The Buzzard cluster is unique in its ability to support multiple OSG projects on a single pool of CPUs and GPUs. It's also easily expandable as more funds become available. We present our service model and unique characteristics of Buzzard, with examples from its configuration, to the benefit of other institutions considering a similar approach to supporting OSG on their campuses.

ACKNOWLEDGMENTS

We are grateful for the extraordinary support we've received from the leadership and staff of the OSG consortium and the University of Chicago's Maniac lab, since the early stages of this project.

REFERENCES

[1] B Bockelman, T Cartwright, J Frey, E M Fajardo, B Lin, M Selmeci, T Tannenbaum, and M Zvada. 2015. Commissioning the HTCondor-CE for the Open Science Grid. *Journal of Physics: Conference Series* 664, 6 (2015), 062003. http://stacks.iop.org/1742-6596/664/i=6/a=062003

[2] Joe Breen, Lincoln Bryant, Gabriele Carcassi, Jiahui Chen, Robert W. Gardner, Ryan Harden, Martin Izdimirski, Robert Killen, Ben Kulbertis, Shawn McKee, Benedikt Riedel, Jason Stidd, Luan Truong, and Ilija Vukotic. 2018. Building the SLATE Platform. In *Proceedings of the Practice and Experience on Advanced Research Computing* (Pittsburgh, PA, USA) *(PEARC '18)*. Association for Computing Machinery, New York, NY, USA, Article 5, 7 pages. https://doi.org/10.1145/3219104.3219144

[3] Predrag Buncic, Carlos Aguado Sanchez, Jakob Bloomer, Leandro Franco, Steffen Klemer, and Pere Mato. 2009. CernVM - a virtual appliance for LHC applications. *PoS* ACAT08 (2009), 012. https://doi.org/10.22323/1.070.0012

[4] University of Wisconsin-Madison Madison WI US. Center for High Throughput Computing, Computer Sciences Department. 2022. *User Priorities and Negotiation — HTCondor Manual 9.7.0 documentation*. https://htcondor.readthedocs.io/en/latest/admin-manual/user-priorities-negotiation.html#setting-accounting-group-automatically-per-user

[5] B.P. Abbott et al. 2016. Observation of Gravitational Waves from a Binary Black Hole Merger. *Phys. Rev. Lett.* 116 (Feb 2016), 061102. Issue 6. https://doi.org/10.1103/PhysRevLett.116.061102

[6] The University of Chicago. 2020. *Maniac Lab*. https://maniaclab.uchicago.edu/

[7] PACE. 2022. *Partnership for an Advanced Computing Environment (PACE)*. http://www.pace.gatech.edu

[8] Ruth Pordes, Don Petravick, Bill Kramer, Doug Olson, Miron Livny, Alain Roy, Paul Avery, Kent Blackburn, Torre Wenaus, Frank Würthwein, Ian Foster, Rob Gardner, Mike Wilde, Alan Blatecky, John McGee, and Rob Quick. 2007. The open science grid. *Journal of Physics: Conference Series* 78 (jul 2007), 012057. https://doi.org/10.1088/1742-6596/78/1/012057

[9] Derek Weitzel, Marian Zvada, Ilija Vukotic, Rob Gardner, Brian Bockelman, Mats Rynge, Edgar Fajardo Hernandez, Brian Lin, and Mátyás Selmeci. 2019. StashCache: A Distributed Caching Federation for the Open Science Grid. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)* (Chicago, IL, USA) *(PEARC '19)*. Association for Computing Machinery, New York, NY, USA, Article 58, 7 pages. https://doi.org/10.1145/3332186.3332212