Doing more with less: Growth, improvements, and management of NMSU's computing capabilities

Strahinja Trecakov
trecakov@nmsu.edu
New Mexico State University
USA

Nicholas Von Wolff
nvonwolf@nmsu.edu
New Mexico State University
USA

ABSTRACT

Deployed in 2015, Discovery is New Mexico State University's commonly-available High-Performance Computing (HPC) cluster. The deployment of Discovery was initiated by Information and Communication Technologies (ICT) employees from the Systems Administration group who wanted to help researchers run their computations on a more powerful system than one they had sitting in their offices. Over the years, the cluster has grown 6 times, and as of March 2021 has 52 compute nodes, 1480 CPU cores, 17 Terabytes of RAM, 30 GPUs, and 1.8 Petabytes of usable storage. Discovery's hardware is acquired using a combination of university funds, condo-model based funds, and grant funds, causing Discovery to be a heterogeneous system containing several CPU generations.

This paper discusses our growth and administration experiences on this heterogeneous system, as well as our outreach and contribution to the HPC community.

CCS CONCEPTS

• **Social and professional topics → Management of computing and information systems**.

KEYWORDS

High-performance computing, high-throughput computing, heterogeneous system, system administration

ACM Reference Format:
Strahinja Trecakov and Nicholas Von Wolff. 2021. Doing more with less: Growth, improvements, and management of NMSU's computing capabilities . In *Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465610

1 INTRODUCTION

1.1 Overview

Many HPC clusters are deployed as one large homogeneous solution and serve customers during a 5 to 7-year lifecycle. These clusters rarely add new hardware and only minor changes are made to their configuration. Heterogeneous clusters, like Discovery, usually grow in batches and consist of different CPU generations, GPU architectures, networking technologies and vendors.

Our biggest issue with growth, and existence, is the lack of a sustainable budget. Discovery's size increases have been accomplished through a combination of university grant and condo-model funding. The condo-model allows researchers to spend their available time on their research instead of managing and administering a system without the required knowledge.

Discovery is maintained by 1.2 FTE (a full-time and a part-time) system administrators. The rest of the HPC team, graduate assistants, work on outreach, on-boarding, and small technical projects. To optimize server management, we have automated most of our processes using multiple open source tools.

Discovery currently serves over 350 users from New Mexico State University (NMSU) and other educational institutions in the state of New Mexico. We expect a significant increase over the next year due to additional focus on academic and classroom use.

With this paper, we intend to motivate other institutions to keep pushing the boundaries and do their best to contribute to their institutions and to the HPC community. We discuss 1. our motivation and the future of NMSU's HPC; 2. system management and administration procedures, as well as monitoring of this heterogeneous cluster; 3. outreach and our effort to contribute computing resources to the Open Science Grid (OSG) community.

1.2 Motivation and Goals

Every HPC cluster requires a professionally managed, physically secure, and temperature-controlled facility. The facility should have many-layered uninterruptible redundant power and cooling systems, off-site data backup facilities, a redundant high-speed network to ensure high availability, and system administration staff. At NMSU some researchers purchase expensive hardware that ends up sitting in closets or offices without proper housing, management, and administration. The existence of multiple, independent, small clusters like these does not increase community computational resources, research productivity, or usage. As an attempt to remediate these types of issues, NMSU's ICT Systems Administration group deployed our first centrally managed HPC cluster.

Our mission is to facilitate computational research at New Mexico State University, and that entails:

(1) Helping researchers use computational science and high-performance computing in their fields.
(2) Educating and encouraging instructors to utilize our HPC and HTC resources for classroom activities.
(3) Hosting training sessions on different topics about how to take full advantage of HPC and HTC systems.

PEARC '21, July 18–22, 2021, Boston, MA, USA

ACM ISBN 978-1-4503-8292-2/21/07…$15.00
https://doi.org/10.1145/3437359.3465610

(4) Acting as the first point of contact for researchers looking to find external HPC and HTC computing resources.
(5) Contributing unutilized resources to the computational community such as OSG.
(6) Establishing a sustainable budget for HPC staff and hardware growth.

1.3 Discovery Growth

In 2015, New Mexico State University's Student Technology Advisory Committee (STAC) approved and funded a proposal by the Information and Communication Technology department to deploy a student-accessible compute cluster. The initial cluster was made of 7 compute nodes where one node was equipped with 2 Nvidia GPUs. The cluster was using an InfiniBand QDR 32Gb/s network and ICT's primary Storage Area Network (SAN) system. The I/O performance was not what was needed, so the decision was made to redesign our storage solution. With new funding from STAC the following year, Discovery was grown by 7 more compute nodes and a new storage solution (Lenovo E1012 and E1024). The 3rd upgrade included 2 more compute nodes and a InfiniBand interconnection upgrade from QDR to FDR 54Gb/s. In 2018, through the condo-model, we had our 4th upgrade where we added 10 compute nodes and 6 GPU enabled nodes. After this upgrade we noticed I/O issues due to a bottle neck in the disk performance of our homegrown storage solution. During 2019, we were awarded a NSF EPSCoR grant to help facilitate research on realizing New Mexico's potential for sustainable energy development. With this grant we replaced our storage system with a Lenovo DSS-G (GPFS) storage solution, expanded our FDR network, added 4 GPU enabled nodes and 2 high memory nodes. In 2020 with NSF Campus Cyberinfrastructure grant we added 10 compute nodes, 4 GPU enabled nodes, replaced our interconnect network to HDR 200Gb/s and added 600TB of storage.

2 ARCHITECTURE AND ADMINISTRATION

2.1 Management and monitoring

Managing, maintaining, and monitoring a large and complex systems requires good procedures in order to keep the system consistent, persistent, and reliable. To do so, we use multiple open source tools such as Ansible and AWX [1], GitLab [3], iPXE [4], xCAT [11], Lenovo Confluent [6], Zabbix [17], Telegraf [9], Grafana [13], and Open XDMoD [15].

The only physical, and manual, configuration that we to do is configure every node's management controller with an IPv4 address and user. We then leverage xCAT to image compute nodes with CentOS and preform basic provisioning tasks like assigning static IPv4 addresses, installing drivers, and updating firmware. For management and configuration, we use Ansible paired with AWX. This combination of xCAT and Ansible/AWX gives us complete automation and reproducibility of the configuration of the Discovery cluster.

Monitoring node health and compute-job performance is important, and for this we use a combination of open source tools: Telegraf and Grafana for node statistics, Zabbix for alerts and problem detection, and Open XDMoD for utilization, underperforming software, hardware, and Slurm job statistics.

2.2 Network

The Discovery cluster uses 2 separate networks. The first is a standard 1Gb/s Ethernet network used for imaging, management, internet access, and connecting to the campus networks. The second is our high-speed InfiniBand fabric that uses Mellanox QM8700 HDR 200Gb/s InfiniBand switches. Each cluster node is connected to the InfiniBand network using 1x HDR 200 to 2x HDR 100 split cables providing every node with 100 Gb/s networking on the InfiniBand fabric. This high-speed fabric is used for storage and inter-node communications. For security reason, access to the cluster is restricted to computers on our campus network. To facilitate remote access, we use a Cisco AnyConnect VPN service tied into our user authentication services.

2.3 Compute nodes

When initially deployed Discovery was a homogeneous cluster consisting of only Intel E5-2640 v3 CPUs. Today, after 6 hardware upgrades it has 1480 CPU cores from 5 different CPU generations (Intel Haswell, Broadwell, Skylake, Cascade Lake and AMD EPYC), 17 Terabytes of RAM, 30 Nvidia Tesla GPUs (K40/P100/V100/A100). More detailed hardware specification can be found in Table 1.

Table 1: Compute nodes hardware specifications

Count	Hardware Specifications
6	2x Intel E5-2640 v3; 64GB RAM
1	2x Intel E5-2640 v3; 64GB RAM; 2x Nvidia K40
7	2x Intel E5-2650 v4; 128GB RAM
2	2x Intel E5-2650 v4; 256GB RAM
10	2x Intel Xeon Gold 5117; 192GB RAM
5	2x Intel Xeon Gold 5117; 192GB RAM; 2x Nvidia P100
1	2x Intel Xeon Gold 5120; 192GB RAM; 2x Nvidia V100
2	4x Intel Xeon Gold 5218; 3TB RAM
4	2x Intel Xeon Gold 5218; 192GB RAM; 2x Nvidia V100
10	2x Intel Xeon Gold 6226R; 384GB RAM
4	2x AMD EPYC 7282; 512GB RAM; 2x Nvidia A100

2.4 Storage

To keep up with the growth of our computational resources and eliminate historic I/O performance issues with our existing storage solutions, we deployed an IBM Spectrum Scale (GPFS) distributed storage solution. In 2019 we started with a Lenovo DSS-G220 solution that has 2 enclosures with 84 drives each that gave us 1.66PB of raw storage. A year later we upgraded to a DSS-G230 by adding another enclosure with 84 x 10TB drives bringing it to total of 2.5PB of raw storage. This system is fully deployed and managed with xCAT and Ansible. With direct connections to our InfiniBand fabric, via HDR 200Gb/s active optic cables, it provides the best possible I/O performance that our GPFS filesystem is capable of.

Each user by default gets 100GB of persistent home and 1024GB of ephemeral scratch storage. Scratch space only retains data for 120 days and is meant for temporary high demand usage. For collaboration between users and groups, we offer project space with a quota of 500GB. Both home and project directories are backed up

nightly and we leverage GPFS snapshots to provide quick file-level restores without having to pull data from our much slower backup solution.

2.5 Software

To achieve clean and consistent package builds, we use the Spack package manager [12]. It is designed to support different software versions, configurations, architectures, and environments. Spack is widely used in the HPC community and it has over 5200 package recipes preconfigured. It enables us to provide multiple versions of software, and even the same version, with different built flags or optimizations. Spack also allows for the use of both newer and older libraries then what is shipped with CentOS. For user access to software, we use the Lmod module system [14]. Lmod manipulates shell environmental variables, to add software binaries, and libraries, to the user's environment. This is our primary mechanism for users to discover, load, and unload, software modules. The combination of Spack for building and Lmod for access, provides a high degree of flexibility when building and deploying software.

2.6 Resource scheduling

In order to facilitate an optimal and fair use of resources, we use the Slurm Workload Manager [18]. Slurm is a clustering software that orchestrates running jobs across multiple compute nodes. Jobs are usually written as bash scripts which consist of a resource request and any number of tasks. Resource requests can be for an entire compute node, multiple nodes, or just a small portion of a node. In cases where multiple jobs are running on a single node Slurm uses Linux Control Groups (cgroups) to ensure that CPU threads, memory, and GPUs can only be used by the job they are assigned to. Jobs are confined by these cgroups and are prevented from using more resources than they requested. Slurm offers other benefits such as: cluster partitioning, job queuing, quality of service fair share policies, job histories, and much more. The Discovery cluster is a bare-metal installation consisting of one combined head/database node, two login nodes, and over 50 compute nodes. A head node oversees the scheduling of jobs, managing the Slurm configuration, and keeping track of the state of the cluster. The database node stores historical information about jobs, quality of service policies, user information, and user account assignments. Login nodes are the user-facing part of a Slurm cluster where users can login to submit jobs and do light development tasks. Login nodes typically have strict resource restrictions to prevent users from doing research directly on the login node itself. Lastly, we have the compute nodes where users are not able to login directly, but instead use Slurm to run jobs and reserve resources.

2.7 User authentication

To facilitate access to the Discovery cluster, and its supporting services, we use Microsoft Active Directory (AD) for user authentication via the Lightweight Directory Access Protocol (LDAP). For reduced overhead we use NMSUs main AD domain and join the entire Discovery cluster to it. This provides user authentication, and identity mapping for Linux without having to manually create user accounts on each individual node. To extend this authentication to our web services, we use a single sign-on (SSO) service called Keycloak [5]. Keycloak acts as a login portal for all our HPC web applications and is backed by our common AD via LDAP. The result is a single identity used for our University email, Slurm, Linux, and Web services which provides a consistent environment for our users.

3 DOCUMENTATION AND COMMUNICATION

Documentation is important in every workplace, and having accurate up-to-date documentation is a lifesaver. Every scheduled maintenance we try to improve our procedures, so we must keep notes about them current. To do this, we use a few tools to keep track of changes, procedures, policies, hardware, and cable management. We leverage the capabilities of Netbox [7] to keep track of rack layout, utilization, hardware specifications and networking inventory. For administrative procedures and user documentation, we use separate Antora [2] instances, a static HTML site generator for technical documentation. Our Antora repositories live in GitLab and are managed by our HPC team.

Communication with our users is key to our success and growth. We always welcome innovative ideas, issues, and requests that help our researchers. To offer the best quality of service, we use a variety of platforms to communicate with our end users. Support requests can be submitted through the contact form on our website, or by sending an email to our group. This way every member of our team is notified and can assist the customer as soon as possible. Moreover, all our users are part of a Microsoft Teams Group where they can easily send us a message or share their experience/issues with other users.

For announcements, we use 3 different platforms (Microsoft Teams, an email list, and the website) to make sure that most, if not all, of our users receive the message. This helps users be informed about planned and unplanned/emergency downtimes, and updates.

4 OUTREACH, TRAINING, AND USER GROWTH

As the need for HPC resources and skills continues to grow within STEM (Science Technology Engineering and Math), NMSU is actively looking to increase the presence of other fields as well as bring HPC skills inside the classroom. To encourage members of the academic community to take advantage of Discovery, and to inform them about its capabilities, we provide presentations on a wide variety of topics. These training sessions provide access to HPC usage experiences such as parallel computing (OpenMPI/MPICH), Open OnDemand capabilities, and other popular software such as Matlab, R, Jupiter Notebook, and Anaconda among others.

Each new user must go through an on-boarding process, either one-on-one or a Canvas learning management system course, with quizzes, to get an account on Discovery. The on-boarding ensures that new users get basic training on the Linux terminal, Slurm scheduler, and software module system, setting the users up for success on Discovery.

In 2017 NMSU's High-Performance Computing group prioritized the task of outreach which included attending departmental meetings, handing out flyers, and putting posters about Discovery in prominent locations, as well as the presentations listed above as

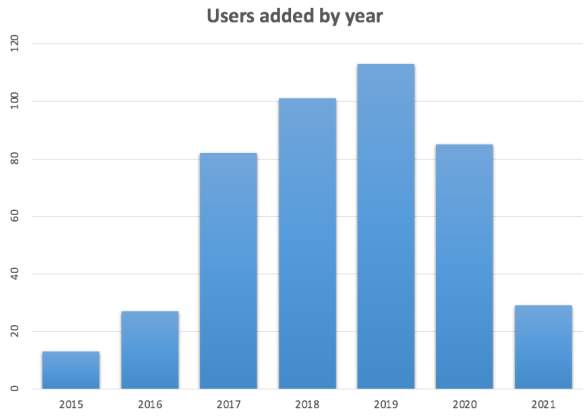

Figure 1: The unique number of user account requests received each year.

another way to improve visibility of the HPC team and the HPC resources. In Figure 1, you can see the increase of users each year. An increasing number of users each year was interrupted by the pandemic as we could not hold in person workshops and do our normal departmental outreach.

5 MORE THAN DISCOVERY

5.1 OSG

In 2018, we stood up a decentralized High-Throughput computing cluster called Aggie-Grid that consists of student lab machines that join the cluster when they are idle. That year NMSU started contributing its computational hours to the Open Science Grid [16] from both the Aggie-Grid and Discovery clusters. We have contributed over 10.6 million core hours and more than 2.52 million jobs have run on our resources since 2018 (Figure 2), including 807K core hours of COVID 19 related research.

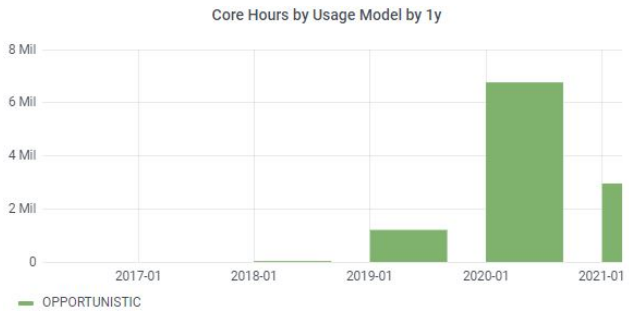

Figure 2: Total Core Hours per year [8].

5.2 Slate

To support collaboration across different institutions and facilities we have partnered with SLATE (Service Layer at the Edge) [10] institutions and took a part in this project by deploying a SLATE cluster at our main campus. With limited staff, SLATE provides us with community-tested applications that our team would otherwise not have time to develop, deploy and maintain. The containerized services that we run are OSG HTCondor-CE and OSG Squid Proxy instances.

6 SUMMARY

Although there are areas that can be improved, we believe that the Discovery cluster is managed well with no sustaining budget and few staff. Our outreach, and training, helps researchers successfully navigate their research goals on our HPC. Outreach also helps to growing the user base, as well as the Discovery cluster itself via the condo-model. Institutions in the same position may find this paper helpful for growing their clusters and managing heterogeneous cluster.

ACKNOWLEDGMENTS

New Mexico State University Discovery is directly supported by the National Science Foundation (OAC-2019000), the Student Technology Advisory Committee, and New Mexico State University and benefits from inclusion in various grants (DoD ARO-W911NF1810454; NSF EPSCoR OIA-1757207; Partnership for the Advancement of Cancer Research, supported in part by NCI grants U54 CA132383 (NMSU)). We thank all past and current members of the NMSU's High Performance Computing Group for their hard work over the years. We are grateful to Curtis Ewing and Diana Toups Dugas for reviewing this paper.

REFERENCES

[1] [n.d.]. *Ansible.* https://ansible.com.
[2] [n.d.]. *Antora.* https://antora.org.
[3] [n.d.]. *GitLab.* https://about.gitlab.com/.
[4] [n.d.]. *iPXE.* https://ipxe.org.
[5] [n.d.]. *Keycloak.* https://www.keycloak.org/documentation/.
[6] [n.d.]. *Lenovo Confluent.* https://hpc.lenovo.com/users/documentation/.
[7] [n.d.]. *Netbox.* https://github.com/netbox-community/netbox/.
[8] [n.d.]. *OSG Site usage.* https://bit.ly/3uXircg.
[9] [n.d.]. *Telegraf.* https://www.influxdata.com/time-series-platform/telegraf/.
[10] Joe Breen, Lincoln Bryant, Gabriele Carcassi, Jiahui Chen, Robert W Gardner, Ryan Harden, Martin Izdimirski, Robert Killen, Ben Kulbertis, Shawn McKee, et al. 2018. Building the SLATE Platform. In *Proceedings of the Practice and Experience on Advanced Research Computing*. 1–7.
[11] IBM Corporation. 2015. *xCAT: Extreme Cloud Administration Toolkit.* https://xcat.org.
[12] Todd Gamblin, Matthew LeGendre, Michael R Collette, Gregory L Lee, Adam Moody, Bronis R de Supinski, and Scott Futral. 2015. The Spack package manager: bringing order to HPC software chaos. In *SC'15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–12.
[13] Grafana Labs. 2018. *Grafana Documentation.* https://grafana.com/docs/.
[14] Robert McLay, Karl W Schulz, William L Barth, and Tommy Minyard. 2011. Best practices for the deployment and management of production HPC clusters. In *SC'11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–11.
[15] Jeffrey T Palmer, Steven M Gallo, Thomas R Furlani, Matthew D Jones, Robert L DeLeon, Joseph P White, Nikolay Simakov, Abani K Patra, Jeanette Sperhac, Thomas Yearke, et al. 2015. Open XDMoD: A tool for the comprehensive management of high-performance computing resources. *Computing in Science & Engineering* 17, 4 (2015), 52–62.
[16] Ruth Pordes, Don Petravick, Bill Kramer, Doug Olson, Miron Livny, Alain Roy, Paul Avery, Kent Blackburn, Torre Wenaus, Frank Würthwein, et al. 2007. The open science grid. In *Journal of Physics: Conference Series*, Vol. 78. IOP Publishing, 012057.
[17] Lambert M. Surhone, Miriam T. Timpledon, and Susan F. Marseken. 2010. *Zabbix.* Betascript Publishing, Beau Bassin, MUS.
[18] Andy B Yoo, Morris A Jette, and Mark Grondona. 2003. Slurm: Simple linux utility for resource management. In *Workshop on job scheduling strategies for parallel processing*. Springer, 44–60.