Final Report

The Future of Cloud for Academic Research Computing

Results of an NSF-Supported Workshop, Entitled "*Cloud Forward*"
Supported by NSF ACI/CSE Award 1632037

Authors

Jim Bottum, Dustin Atkins, Alan Blatecky, Rick McMullen, Todd Tannenbaum, Jan Cheetham, Jim Wilgenbusch, Karan Bhatia, Erik Deumens, Barr von Oehsen, Geoffrey Fox, Marcin Ziolkowski, Asbed Bedrossian, Dan Fay

Workshop Contributors

Dustin Atkins (Clemson University), Asbed Bedrossian (University of Southern California), Karan Bhatia (Google), Alan Blatecky (RTI), Jim Bottum (Clemson University & Internet2), Thomas Cheatham (University of Utah), Jan Cheetham (University of Wisconsin-Madison), Erik Deumens (University of Florida), Dan Fay (Microsoft), Geoffrey Fox (Indiana University), Steve Gallo (University of Buffalo), Jill Gemmill (Clemson University), John Hicks (Internet2), Kate Keahey (University of Chicago and Argonne National Laboratory), Gail Krovitz (Internet2), Ruth Marinshaw (Stanford University), Fritz McCall (University of Maryland College Park), Rick McMullen (Texas A&M University), John Moore (Internet2), Sean O'Brien (Internet2), Sanjay Padhi (Amazon Web Services), Joseph Ryan (University of Denver), Dan Stanzione (University of Texas at Austin and Texas Advanced Computing Center), Todd Tannenbaum (University of Wisconsin-Madison), Esen Tuna (Indiana University), Barr von Oehsen (Rutgers University), Nick Weber (National Institutes of Health), Karen Wetzel (Educause), Jim Wilgenbusch (University of Minnesota and Minnesota Supercomputing Institute), Boyd Wilson (Omnibond), Marcin Ziolkowski (Clemson University), and Joel Zysman (University of Miami)

1. Introduction and Purpose of the Workshop

Cloud computing is in the early stages of becoming a disruptive force in the area of research computing, and there are deep questions about how this new method of computing will impact the business and operating model for academic research computing. This workshop brought together thought leaders from academia – including researchers, CROs, CIOs – government, and the private sector to discuss key issues surrounding the future of cloud computing's impact on computing in the research community. Though cloud computing can take many forms, for purposes of this workshop, cloud computing is meant mostly as commercial cloud computing – including Amazon Web Services, Google, and Microsoft Azure. The workshop organizers sought a broad array of cloud computing use cases, providers, and research groups to best chart the future of this important evolution. Cloud-based computing in this context on campuses is evolving at an increasingly rapid pace, and the goal of this workshop was to explore how the advent of cloud technologies will impact researchers on campuses.

Academic researchers and their host institutions are facing a growing complexity of options for accessing and provisioning[1] computational and data handling resources. This rapidly changing environment presented several key motivations for this workshop including the increasing availability of commercial and non-commercial cloud services that are appropriate for cloud computing in the context of academic research use for individual researchers and teams. Several recent major reports on clouds and the future of computing identified some of the issues created by the disruptive effects of clouds. As a result, it is becoming more important than ever for academic research computing to enumerate the current enablers and barriers to adoption of cloud services. The workshop brought together the necessary community leaders needed to identify a path to acquire and use cloud computing services more broadly and in novel ways as well as serving the needs of academic researchers in a wide variety of disciplines. One of the most critical issues identified by the workshop, is the need for a prepared workforce to support the research community.

The workshop involved campus researchers, directors of research computing, Chief Information Officers, industry leaders, and other practitioners from around the country to engage this discussion. Workshop attendees included the following: Dustin Atkins (Clemson University), Asbed Bedrossian (University of Southern California), Karan Bhatia (Google), Alan Blatecky (RTI), Jim Bottum (Clemson University & Internet2), Thomas Cheatham (University of Utah), Jan Cheetham (University of Wisconsin-Madison), Erik Deumens (University of Florida), Dan Fay (Microsoft), Geoffrey Fox (Indiana University), Steve Gallo (University of Buffalo), Jill Gemmill (Clemson University), John Hicks (Internet2), Kate Keahey (University of Chicago and Argonne National Laboratory), Gail Krovitz (Internet2), Ruth Marinshaw (Stanford University), Fritz McCall (University of Maryland College Park), Rick McMullen (Texas A&M University), John Moore (Internet2), Sean O'Brien (Internet2), Sanjay Padhi (Amazon Web Services), Joseph Ryan (University of Denver), Dan Stanzione (University of Texas at Austin and Texas Advanced Computing Center), Todd Tannenbaum

(University of Wisconsin-Madison), Esen Tuna (Indiana University), Barr von Oehsen (Rutgers University), Nick Weber (National Institutes of Health), Karen Wetzel (Educause), Jim Wilgenbusch (University of Minnesota and Minnesota Supercomputing Institute), Boyd Wilson (Omnibond), Marcin Ziolkowski (Clemson University), and Joel Zysman (University of Miami).

Through exploration of the evolving nature of academic research computing over time, the workshop explored topics and challenges for academic research cloud computing in a variety of contexts – including applications, support, data movement, administrative, legal, and financial. The workshop was invitation-only to key stakeholders and those who were able to contribute to this final report detailing key issues and recommendations moving forward.

The term ***cloud*** in this report is used in two ways. The first way is to designate the infrastructure operated by commercial companies like AWS, Google, and Microsoft. The second way is the style and workflow of computing that has the following features that commercial cloud providers offer on their infrastructure: self-provisioning (or easy, quick provisioning) of resources, interactive timescale to get access to resources (including tying local compute with cloud compute rather than access through queueing and a scheduler), the ability to customize the computing environment (through VM or containers, as opposed to running in a pre-determined environment), and the ability to run continuous services like web service and database service through defined APIs.

2. Previous Reports on Cloud in Research Computing

Several reports and working groups preceded this workshop's proceedings, and informed the discussions on the future of cloud for academic research computing. The Magellan Report, 2011[i], reported the findings of a two-year study funded by the Department of Energy's Office of Advanced Scientific Computing Research through American Recovery and Reinvestment Act. It investigated the potential of a private cloud for mid-range and data intensive computing workloads by DoE investigators and collaborators at Argonne and NERSC, with an emphasis on virtualization, the applicability of the MapReduce programming model using Hadoop, and portability of software stacks in a number of application used by physicists, climate scientists, and genome scientists. Gains have been made in some of the gaps identified in the report, including addressing performance and reliability limitations in open-source virtualized cloud software stacks for production science use. Although the cost benchmarking in the report is nearly six years old, the 2016 National Academies Press publication, Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020[2] [ii]indicated that the cost/pricing comparison between the Magellan testbed and AWS service offerings in the 2011 Magellan report is still applicable in 2016. Several of the *Cloud Forward* workshop institutions indicate they are building, planning, or envisioning services for their researchers that align with some of the Magellan report recommendations, including providing standardized images and programming assistance for researchers to help them move applications into the cloud

and/or building services around customized pipelines for collaborative research that involves version, data, and library dependencies.

In 2015, EDUCAUSE published an ECAR working group paper, "*Research Computing in the Cloud: Functional Considerations for Research.*"[3] [iii]This report examined the technical capabilities of commercial cloud services for different types of research computing as well as policy and cost considerations. The flexibility of cloud services, including scalability and elasticity, for enabling multiple approaches for solving problems in research was noted as a key benefit. The authors identified HTC (High Throughput Computing) workflows involving independent sequential computational steps as a particularly good fit with the resources available through cloud services, but found that HPC workflows requiring tight coupling and high memory bandwidth were not well supported by the architectures available in commercial clouds. Other challenges with use of the cloud for research were noted, including cost issues of maintaining data in the cloud long term, past the grant-funded period and into later phases of the data's lifecycle, when it is frequently useful for analysis, data-mining resulting in secondary findings, and compliance with public access requirements. In addition, vetting the security of cloud providers for sensitive and restricted data and extending software licenses for use in the cloud has presented some institutions with challenges.

Lastly, a report release by XSEDE in 2013[4][iv] summarizes survey data collected between September 2012 and April 2013, on cloud usage by 80 research groups. The survey questions focused exclusively on research related activities in the cloud and collected 22 quantitative attributes for each research project. The report was commissioned by the National Science Foundation and the survey and data summarization was conducted by the XSEDE Cloud Integration Investigation Team. Survey findings were broken into the five high level categories:

- Top 3 Reasons Researchers and Educators use the Cloud
- Applications Identified as Good Candidates for the Cloud
- Cloud Benefits Reported by the Survey Participants
- Cloud Challenges Reported by the Survey Participants
- Continued Investment Needed

Some of the findings in this report overlapped with the findings and the general discussions that transpired over the two-day Cloud Forward Workshop. In particular, the following findings, taken directly from the XSEDE report closely overlap with some component of the Cloud Forward Workshop findings, including:

- Investments that facilitate access to production cloud resources, cloud training, and cloud user consulting are needed as well, whether the clouds are public, private, or national CI or, more likely, some combination thereof.

- Although in their infancy, hybrid clouds hold the promise of enabling modest size private clouds used for steady-state workloads to burst to public, community, or national CI during peak workloads. Most private clouds are expected to become hybrid clouds in the future. The challenge will be implementing a management framework that can span all cloud environments.
- Executing a tightly coupled HPC application in a virtual machine environment may not be the best use of production resources. It is important to pick the environment best suited to your application.
- Software as a Service (SaaS) environments such as MATLAB and R provide researchers and educators with economies of scale in software licenses and more optimal execution environments.
- Science as a Service provides researchers with rich web applications and platform components that reduce time to science by hiding platform complexities and by offering special performance features desired by specific research communities, i.e., GPGPUs, shared datasets, etc.

However, other points reflected in the XSEDE report are noticeably different from findings from the Cloud Forward workshop. The XSEDE report tended to deemphasize the barrier that can be created by having to pay for cloud services. In part this is likely because the survey participants are already using the resources and in many cases were early adopters of these services. The report states, “Pay as you go, compute elasticity, and data elasticity are among the cloud benefits reported by the survey participants. As one scientist said, “clouds promise to scale by credit card, that is, scale up immediately and temporarily with the only limits imposed by financial reasons, as opposed to the physical limits of adding nodes to clusters ... or the financial burden of over-provisioning resources.” The Cloud Forward workshop, on the other hand, started with the assumption that cloud services will become ever more important to the research enterprise, with a focus on what that change will mean to the campus research communities.

3. Current Supported Cloud Resources and Initiatives

NSF continues to support a number of cloud and national resources that are available to researchers through XSEDE. Proposals justifying the use of XSEDE resources can be submitted for review, quarterly, by the XSEDE allocation committee. (XRAC). Researchers with active funding are given compute resources in the form of service units (SU). Requests for storage resources are also available at some of the XSEDE sites. The resources are open to anyone, even researchers without NSF funding, but having a funded project increases the chances of allocation awards.

- **Blue Waters**: Operated by NCSA and serves special HPC batch needs for parallel jobs needing 10,000 cores and up, using MPI and hybrid parallel programming styles with multi-core and GPUs.

- **Bridges:** Operated by PSC and is designed to support familiar, convenient software and environments for both traditional and non-traditional HPC users

and contains VMs, CPU and GPU nodes, very large memory nodes, and supports gateways, databases, and data movement.

- **Comet:** Operated by SDSC and provides HPC and HTC batch processing with ability to create and manage virtual clusters with Linux VMs, so that researchers have more control over the computing environment, operating system, middleware, and application stack.

- **Corral**: Operated by TACC and offers the ability to create and run various services like databases and web portals for research groups or research communities. This allows public or focused user communities to access data as well as software and algorithms ready to run on existing datasets or on datasets provided by the user.

- **Jetstream:** Operated by Indiana University, TACC, and University of Arizona and provides cloud-style access with the ability to run and control interactive workflows with Windows VMs.

- **Lonestar:** Operated by TACC and offers processing capability for visualization of data produced on other XSEDE resources such as Stampede and Blue Waters.

- **Stampede:** Operated by TACC and serves traditional HPC and HTC batch needs.

In addition to production resources, the NSF has also funded other large cloud testbeds.

- **Aristotle:** A collaboration between Cornell University, University of Buffalo and UCSB, this is a federated cloud that can burst to the public cloud and supports big data. Each site operates standard cloud infrastructure components augmented with DIBBs storage assets including data analysis servers, scalable storage, and a Globus file transfer and sharing endpoint, all connected to the Internet at 10Gb.

- **Chameleon:** A collaboration between the University of Chicago and TACC - Chameleon is a configurable experimental environment for large-scale cloud research on bare metal resources. Chameleon is deployed at the University of Chicago and the Texas Advanced Computing Center and consists of 650 multi-core cloud nodes, 5PB of total disk space, and leverage 100 Gbps connection between the sites. While a large part of the testbed will consist of homogenous hardware to support large-scale experiments, a portion of it will support heterogeneous units allowing experimentation with high-memory, large-disk, low-power, GPU, and co-processor units. The project will also leverage existing FutureGrid hardware at the University of Chicago and the Texas Advanced Computing Center in its first year to provide a transition period for the existing FutureGrid community of experimental users

- **Cloudlab:** Developed and operated from a collaboration of the University of Utah, Clemson University, and the University of Wisconsin, CloudLab clusters have almost 15,000 cores distributed across three sites around the United States: Utah, Wisconsin, and South Carolina. Each cluster has a different focus: storage and networking (using hardware from Cisco, Seagate, and HP), high-memory computing (Dell), and energy-efficient computing (HP). CloudLab is interconnected with nationwide and international infrastructure from Internet2, and is built from the software technologies that make up Emulab and parts of GENI, so it provides a familiar, consistent interface for researchers.

In addition to these, there are cloud and computational resources from other funding agencies

- **Department of Energy (DOE):** The DOE Office of Science provides a portfolio of national high-performance computing facilities housing some of the world's most advanced supercomputers. These leadership computing facilities enable world-class research for significant advances in science.

  Open to researchers from academia, government labs, and industry, the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program is the major means by which the scientific community gains access to some of the fastest supercomputers. The program aims to accelerate scientific discoveries and technological innovations by awarding, on a competitive basis, time on supercomputers to researchers with large-scale, computationally intensive projects that address "grand challenges" in science and engineering.

- **National Institutes of Health (NIH):** Due to the expansion in the volume and complexity of given the expansion in the volume and complexity of genomic data generated by the research community, the National Institutes of Health (NIH) is now allowing investigators to request permission to transfer controlled-access genomic and associated phenotypic data obtained from NIH-designated data repositories under the auspices of the NIH Genomic Data Sharing (GDS) Policy to public or private cloud systems for data storage and analysis. NIH expects cloud computing systems to meet the same data use and security standards outlined in *NIH Security Best* institution's own IT security requirements and policies. NIH has also launched an electronic "Commons," a community-controlled cloud infrastructure that would support collective uses of computing, storage and data for biomedical research by NIH and its academic and industry collaborators.

- **National Center for Atmospheric Research (NCAR):** The NCAR-Wyoming Supercomputing Center (NWSC) provides advanced computing services to scientists studying a broad range of disciplines, including weather, climate, oceanography, air pollution, space weather, computational science, energy production, and carbon sequestration. It also houses a landmark data storage

and archival facility that will hold, among other scientific data, unique historical climate records.

- **Open Science Grid (OSG):** The OSG facilitates access to distributed high throughput computing for research in the US. The resources accessible through the OSG are contributed by the community, organized by the OSG, and governed by the OSG consortium. The Open Science grid consists of computing and storage elements at over 100 individual sites spanning the United States. These sites are primarily at universities and national labs and range in size from a few hundred to tens of thousands of CPU cores.

4. Evolution and Current State of Academic Research Computing

Academic computing grew with the creation of computers in the 1960s. Initially researchers used mainframes that were operated from 8 to 5 for business purposes of universities and research organizations, during off hours. During the 1980s and 1990s, research computing moved to mini computers, like DEC VAX 11/780, and then to workstations and clusters of workstations. The high-end computing moved to supercomputers operated as national resources. During the 2000s, the basic building block for computing became the server as a node in a cluster. Research groups operated small clusters, some universities operated shared clusters, national labs and supercomputer centers operated high-end clusters, commercial companies like AWS, Google, and Microsoft operated large clusters to provide a wide range of consumer services. These companies then developed infrastructure that allowed anyone to configure and allocate virtual machines and virtual clusters and operate, manage, and use them for computational tasks without the need to pay attention to any aspect of the hardware underlying these systems: Cloud computing was born.

Since 2010, the number of service providers offering cloud services has grown and the systems available have grown to meet almost any specification for research computing systems: Infiniband interconnect, parallel file systems, high-performance input/output, solid-state disks, accelerators like GPUs and FPGAs. Research groups that developed the competence and expertise to operate clusters for their projects during the 1990s and 2000s can now use the commercial cloud and recently NSF-funded resources like Jetstream to operate clusters without touching any hardware. Also since 2010, an increasing number of universities have built a centrally funded and operated facility to support research computing. The goal is to consolidate the distributed clusters operated by individual research groups, institutes, and departments, both to increase the efficiency of the researchers, and to reduce the risk to the university by having professionals manage the systems instead of relying on researchers and graduate students.

Today, many campuses are looking at cloud computing as an important addition to current research computing offerings. Unfortunately, there is overhead associated with not only incorporating these services into the current environment, but in educating and supporting the research community on how to best utilize these services within their

workflows. The consensus of this workshop is that having people in place to bridge the gap between the research community and the resources is absolutely essential to the successful adoption of cloud services. Since this is a relatively new area for university research computing, most do not have the people or the bandwidth to fill this gap.

For this reason, it is necessary that university advanced computing efforts hire new people with the expertise, including ACI-REFs, system administrators (for creating containerized workflows), or reeducate current staff. There has been some movement with regard, for instance, to how to address experimentation - changing approaches or testing theories – in some existing industry programs, as well as in CloudLab, Chameleon, and Jetstream. Additionally, some training on integration of cloud for IT staff and researchers has occurred on several campuses, but is not yet widely formalized or adopted across most campuses and organizations. Some campuses have begun expanding local computing environment to the cloud, but issues such as users not knowing where their jobs will run, interoperability between local resources and commercial clouds, and billing challenges have proven to be significant barriers to further adoption to this point.

5. Challenges and Opportunities for Cloud in Academic Research Computing

During this workshop, several key applications used by research disciplines were identified as *cloud ready* or are already actively running in the cloud. The first case are those applications that are called **Pleasingly Parallel**, where little effort is needed to separate the problem into a number of parallel tasks, or where there is little or no dependency or the need for communication between those tasks. Users who have Pleasingly Parallel problems can easily take advantage of High Throughput Computing resources (such as commercial clouds or HTCondor based environments) to run their jobs. A second case of applications are called **Gateways**, which use some sort of front end or portal such as a web interface that allows users to easily use complex compute resources. A gateway is usually operated and maintained by a group that is familiar with the domain science being conducted. A third case are those applications know as **Software as a Service (SaaS)**, which has become a common delivery model for many business applications. Software is centrally hosted and is accessed by a thin client such as a web browser; in some ways, it is quite similar to how Gateways operate.

In this regard, there is a significant need to distinguish between an expert or systems administration view, and a user view. For example, iPlant software is aimed at simplicity for the more novice user, whereas some cloud interfaces are aimed at experts but marketed to everyone, including beginners. Understanding these differences is essential if users are expected to migrate from gateways that have been designed for their use to cloud interfaces. Furthermore, while a cloud bursting model where local resources are linked to commercial clouds for overflow is attractive, it must be supported continually, and more work is needed to support this model on campuses for end-users that may be taking advantage of such a model.

There are a host of other issues and concerns that currently affect cloud application use. The first concern arises in that cloud services can sometimes lead to vendor lock-in and perhaps publication difficulties as there is sometimes confusion as to which algorithm is being used. However, there are available open-source versions or alternatives for the most important software applications, though open source versions may still not be fully satisfactory to researchers, or can be less functional than matching commercial systems. However, sometimes the cloud-provided software does not provide the capabilities needed; for example, Google Tensorflow did not support distributed memory parallelism in its first release.

Another example is that of Apache Beam, which is the open source version of Google's production Cloud Dataflow. There are two caveats regarding Apache Beam; first, it has little user community uptake at the moment and it is competing with other well-known and well used research workflow systems like Galaxy, Kepler and Pegasus. Second, Apache Beam does not have an execution engine built in, and a researcher has to use either Spark or Flink currently.

Lastly, there are issues with non-local clouds, including the necessity of having the expertise to setup non-trivial specialized images and getting allocations on such systems. And, while all clouds have data access issues for datasets that are only available on particular resources, there are no general rules governing this. It is estimated by the workshop participants that only 5-10% of users (not necessarily reflective of disciplines) are cloud-ready, while the rest need either support and/or extensive training to be successful. Finally, a stronger way to limit unintended compute use is needed. At present, a graduate student mistake can wipe out a research group's allocation over the weekend.

6. Challenges and Opportunities for Cloud in Research Support and Engagement

Additionally, as part of this workshop, the participants were asked to address challenges and opportunities associated with wide scale adoption of cloud resources. That is, what sort of issues do universities need to address if they depend almost solely on cloud resources to support their research instead of depending on just university-owned resources. Four questions were raised and generated lively discussions.

1. *What characteristics make certain classes of problem cloud-ready (or not)? What are the problems in a discipline that make it difficult or not an ideal candidate to be cloud-ready?*

. Nearly all disciplines have some applications and problems that are "ready"; that is, parts of many disciplines (especially Genomics) are already using some cloud services or are looking at doing so. Many disciplines already have key software available; some as SaaS, others as open source. Even disciplines often listed in the "long tail" of science, such as the humanities, are considering cloud resources because of ease of use. However, it is important to recognize the difference between technology being ready and the disciplines being ready.

Research that depends upon large data sets that are typically stored in a central location outside the cloud resource may be cloud-ready, but getting the data may be problematic, in terms of both speed cost. And, while in the past research computing had to be loosely-coupled and/or parallel in order to be cloud-ready, that is becoming much less of an issue as clouds begin to provide new capabilities and software. However, when the research requires the use of large geospatial data, high fidelity modeling for efforts such as climate research or multi-source data analytics to improve food production, use of cloud resources becomes more difficult. But, it should be noted that significant improvements in cloud capabilities in just the last 2 years have significantly reduced some of these difficulties. In another 2-3 years, these challenges may not be much of an issue at all.

Another issue to be considered in being cloud-ready include the willingness of students and faculty to experiment and explore new approaches. This willingness can depend on whether the campus already has some cloud experience or expertise/support that can be leveraged and how easy is it for a student or faculty member to open a cloud-account; does the campus have procedures in place (along with some support staff) to enable cloud computing.

2. *What new dynamics between CIOs, CROs, IT support and research support staff are needed to make effective use of cloud computing services in university research programs?*

Perhaps the most important dynamic is the willingness of the campus to invest in workforce training including augmenting the existing IT and compute staff across the campus. Recognition that cloud services are important for both research as well as the enterprise is important to develop research and education computing strategies. While the CRO primarily focuses on research capabilities, the CIO typically focuses on enterprise needs. But because cost, budget, operations, security, privacy and compliance issues for clouds cross the entire campus enterprise, the CRO and CIO will need to collaborate more extensively than they have in the past. Institutional cloud agreements in support of research are often quite different than institutional agreements for IT support and will need to involve both the CRO and CIO. In some cases, IT groups tend to be slower in adoption of cloud services than the research community, which is another reason for closer cooperation.

3. *What are the workforce development implications for those charged with supporting research with the advent of cloud? What can campuses do to address these? What about organizations above the campus level? What are the roles of currently available cloud computing services in basic and applied research, research training and STEM education conducted outside of government labs and private industry? Who trains the cyberpractitioners on how to do this stuff? What are the roles of cloud providers (public/private/other) in this training?*

One of the most critical issues facing the research community over the next decade is the need for a prepared workforce. The rapid growth and importance of data, environment-changing technologies such as IoT and G5, and the capabilities of cloud computing will transform how we do research, conduct business, teach, and live our lives. The workforce will need to include people who know how to use these capabilities, people who know how to develop these capabilities, and people who know how to integrate them to conduct new research. It is also critical to develop a cyber-savvy corps that are able to not only provide support, but help users address challenges and opportunities involving data science, analytic tools and workflows. The ACI-REF project is a good model for how to establish cyberpractitioners who do more than just provide support; they also serve as evangelists and help users adopt and adapt new approaches and techniques. Other organizations providing support services include XSEDE and OSG.

Although there a number of tutorials and education opportunities at conferences such as Super Computing, the options and availability are limited. There is clearly a need for a qualified and certified workforce, as well as the development of well understood career paths, but there are few options available today. However, organizations such as the Campus Research Computing Consortium (CaRC), which received RCN funding in late 2016, are directly addressing these same issues.

4. *How do individual researchers and research teams collaborate effectively across different cloud vendors?*

The short answer is not very well at this point. Since cloud providers have their own proprietary approach about how to manage and provide cloud services, it is difficult to collaborate across different compute platforms and approaches. However, use of higher-level abstractions such as containers, storage models, gateways and community portals, will enable more collaboration. For example, use of a central location for data would allow researchers to use different cloud providers, and the results from the computations will be available to be shared at the central location. If a set of common tools can be identified at the outset, it will significantly increase the ability to collaborate. Likewise, if the research plan is transparent when it is set-up, with a clear description of the architecture, the workflow, and the tools to be employed, collaboration is significantly enhanced.

7. Challenges and Opportunities for Cloud in Data Movement and Handling

There are further challenges and opportunities for cloud for research computing with respect to data movement, data handling, and transport. Almost by definition “the cloud” is not on the campus network. As a result, access to cloud services gains renewed and heightened importance, as issues such as internet speed, latency, security, and identity and access management -- generally issues that are not on the forefront of the common service user’s attention -- become of paramount consideration. There are technical, policy and compliance considerations when moving and storing data in “the cloud”. To

ensure compliance, upfront thought needs to be given to the level of data security required and implemented on the wire and while the data is at rest. In addition, thought needs to be given to how long the data will be retained and how the long term data costs will be covered.

Campus and individuals must also consider the network implications of deploying cloud services and resources. Research institutions have over time built large network pipelines, generally between research institutions and organizations. They take advantage of NRENs, Internet2, GEANT, and other state and regional high performance networks to enable inter-institutional collaboration between their researchers. Commercial cloud service providers have also invested heavily in large network pipes to provide for the capacity needs of their customers. The two groups have come to the Internet from different angles: higher education and government sponsored high speed networks for the former, and commercial internet for the latter group.

As a result, for the most optimal network connectivity between research institutions and the commercial cloud providers, special attention is needed by the network groups of both sides, in order to provide a properly plumbed, configured route between the institutional user and the cloud infrastructure. Ad-hoc, commercial routes can easily kill bandwidth along the route from source to destination, through incompatible configurations in TCP window sizes, buffer sizes, MTUs, etc. In the past 2 years, more commercial cloud providers have started directly peering with research institutions, or with organizations such as Internet2, to provide optimized network throughput to cloud services. It should be noted that large campus network border throughput statistics do not readily translate to friction-free network traffic between the researcher and cloud services.

Further, security of data is a key aspect of many research projects and must be considered in the context of the provisioning and use of cloud resources. There is a certain, albeit perhaps misplaced, sense of security that researchers may feel when they use a campus network: “the Information Security group should be keeping the network safe!” Whether this is true or false, it definitely cannot be said of common Internet traffic. Whatever providers, routers, switches, or even countries that a researcher’s data may travel through on its way from the institution to cloud providers, is very dynamic and hard to manage, if not completely unknown. The legal, compliance and security consequences and concerns are severe. Many laws, regulations and policies require due IT diligence to secure the data with industry best practices. Export control laws require all data never to leave US soil, patient health protection requires appropriate security (typically encryption in flight and at rest, and strong key management practices), family education privacy laws require robust understanding of institutional education record management, and more.

As a result, researchers and campus IT groups must be aware of the security needs, and be ready to facilitate research by reducing the barriers that security technologies present to researchers. Such technologies as firewalls, virtual private networks (VPNs), virtual private circuits (VPCs), encryption, key management, and IPAM (IP address

management, public vs. RFC 1918 address spaces) are jargon to most researchers, and need streamlined management. In addition, sometimes central IT managed configurations and coordination do not always aware if network shaping and traffic management concerns of the researcher.

Identity and Access Management (IAM) is important to institutions and IT groups, because they need to know who gets access to online services, and this is especially a consideration when looking at the adoption of cloud services and resources. Information Security groups are interested in IAM in order to monitor use, detect and mitigate misuse, minimally shut off misbehaving accounts, and report abuse. Government grant-making organizations (NSF, NIH, DoE, DoD, etc.) are interested in IAM for the same reasons plus the added assurance that the account logging into research accounts is in fact the researchers who were approved for the grant. Researchers, meanwhile, are interested in IAM because they want to collaborate across institutions, in order to build on each other's work and achieve their results. On the surface, these are not diametrically opposed interests, but the goals and rewards for the various stakeholders are not aligned to make for simple IAM environments. As a result, most institutions do not yet have mature IAM deployed across their institutions, but this is a key to success for any use of secure cloud services, and especially so for academic research groups.

Many R1 institutions are members of the InCommon Federation and use Shibboleth as their Single Sign-On (SSO) solution. Shibboleth is a reference implementation of the SAML standard, which enables Federated Identity services. Other institutions use CAS, and still others use Microsoft's Active Directory (AD) with Active Directory Federation Services (ADFS) for SSO. Both CAS and ADFS support the SAML standard. Many commercial cloud providers support oAuth as their SSO standard, and SAML is slowly evolving to integrate with oAuth, which will provide more seamless integration between research institution and commercial cloud provider IAM platforms. Many institutional IAM platforms rely on their enterprise (LDAP) directories to act as attribute stores for their SSO. Cloud services often necessitate rich group and entitlement information to be available about an account at the time when they login to the services, so a mature campus Federated IAM service must be deployed in order to allow researchers from multiple campuses to login to their cloud services in order to collaborate on a project.

Cloud resources (e.g.: VMs, storage, etc.) are typically managed through console access. In order for an institution to properly authenticate, authorize, and monitor cloud services, console access should be integrated with the campus Federated IAM platform. All logins to the console need to be logged. The cloud service account's network topology should provide for a public subnet, a private subnet, and a protected subnet, which allows for secure, tunneled (VPN) access back to campus online resources. The cloud account should also provide for the campus Information Security group to log and monitor cloud service use in order to meet the institution's compliance and security needs. Often, these needs necessitate escrowed root key management and special cloud IAM permissions and roles, in order to allow the researchers to do their work in the cloud, yet meet their IAM and security needs through their services.

8. Administrative, Financial, and Legal Issues in Cloud

With much of the basic hardware of research computing converging to clusters of nodes, resources operated by researchers, universities, national centers, and commercial cloud providers are increasingly comparable and require the same skillsets and expertise to configure and operate. While they offer the same general characteristics, commercial cloud operators can provide orders of magnitude greater capability and performance beyond what the campus can offer. In the support of research computing, resources need to be evaluated against more criteria than hardware and capability. For researchers, there is a growing need is to develop and support complex workflows in cross-disciplinary and multi-institutional collaborations with constant or shrinking budgets. Because of the costs, it is vitally important for campuses to develop a business model to help determine the best mix of resources to be used from university centers, national centers, and commercial cloud providers.

The appropriate business model may depend on both the research needs of the campus as well as the administrative approach used by the campus. In some cases, a campus may issue a BAA or an RFI to determine what vendor best meets the research and educational needs of the campus. In other cases, in particular large research universities, the business model will need to integrate with other cyberinfrastructure services that are centrally managed and operated. However, no matter what business model will be deployed, the model must also address a wide range of other issues beyond acquisition and use of cloud resources.

These issues include determining what level of administrative and technical support the campus will make available to students and faculty. For example, will the campus provide HelpDesk support or consulting services to help researchers customize their workflows? Will the campus provide a campus portal or gateway to enable access and also to insure that compliance and access requirements such as identity management, accurate billing, and data security issues are met? And finally, will the campus have an active outreach program to support novice users and provide information on new capabilities or best practices? These issues may have been formally or informally addressed when researchers used on-premise campus compute resources and governance, compliance and administrative procedures and processes developed as usage of the resources grew. However, when campus researchers begin to use resources external to the campus, the existing processes and procedures may not be adequate.

9. Current Federal Funding Mechanisms for Academic Research Computing

NSF currently provides several mechanisms for awardees to obtain research computing resources for their funded projects: 1) Individual proposals for small research groups, to complex collaborations, can put cost for buying hardware in the equipment portion of the budget and for buying computing services in the service part of the budget. The hardware purchases in recent years have gone to university centers using the condo-model to provide high quality computing infrastructure or acquiring department clusters.

Computing services have included buying resources from commercial cloud providers. 2) NSF has funded national resources that address research computing needs and in recent years several resources have been added that fit the cloud style of computing. Researchers with NSF funding, and others without such funding, can apply for resource allocations on these national resources. No funds are transferred to the institutions of the researchers, they and their collaborators are given access to the resources directly. 3) Special NSF programs fund instruments for specific research activities. These include, with proper justification, buying general purpose hardware to build specialized infrastructure. The infrastructure may be operated by special staff in research groups, institutes, or departments, or by information technology staff in a central research computing departments. This includes other efforts such as the Major Research Instrumentation (MRI) program and initiatives such as the NSF BIGDATA program.

It should be noted that although NSF funds a significant amount of academic open science research computing, other federal agencies such as NIH and DOE also fund some academic research computing. However, that support is primarily focused on the mission of the agency rather than support of broad open science.

10. Campus-Level Administrative Requirements

In the context of cloud service and resource offerings to campuses, there arises the question of and need for certain administrative requirements, policies, and procedures governing the use of such services and resources. When researchers purchase computing resources, the host institutions assume responsibilities and risks. As scrutiny of auditors has increased in recent years, in all areas including financial records, risk management and risk assessment, and compliance with laws and regulations pertaining to management of data, universities have put in place more detailed policies and business processes to mitigate the risk of failing any audit. This includes standing up research computing centers to consolidate operation, maintenance and support of resources.

Individual researchers purchasing general purpose computing equipment is often reviewed on many campuses and needs approval by central IT/research computing or the research arm of the university. Alternatively, the purchase of services from commercial cloud providers on purchase-cards is notoriously hard to track and poses problems for the university. Documentation of the services may be lacking and review of compliance of the researcher workflow, if restricted data is involved may be missed. An incentive for central resources for research computing that is implemented at numerous universities in some form, is the ability for researcher to submit work for the non-cost harvest-idle-cycles queue available through OSG.

Because commercial cloud providers charge for all resources used, flexibility is limited for researchers to do exploratory research computations that use large numbers of core hours or transfer large amounts of data from outside the cloud provider's network. Finding a way for university centers to purchase a fixed allocation that can then be allocated to the burst capacity within a fixed-cost bound would be important to

incentivize this mode of computing. The workload from production research computing, where the total computation and data movement is known in advance, and therefore the cost can be accurately estimated advance, can more easily be moved to the commercial cloud.

The current state of readiness for accessing, purchasing and provisioning of cloud services by researchers at the campus level varies greatly between research communities. Many production research computing workflows can be clearly assessed for the need of resources and can be ported and migrated to the commercial cloud. However, many "exploratory" research computing workloads may not be predictable and will need unspecified levels of resources. They may run best on hardware-constrained resources accessed under the "harvesting idle cycles" mode. The code in this work load also changes more often, but with modern tools that does not pose a serious obstacle to migrate to the commercial cloud; the business model, specifically the lack of idle cycle harvesting, is the main obstacle.

11. Enablers in Adoption of Cloud Services in Academic Research Computing

There are several enabling factors for the adoption of cloud services in research noted during the course of the workshop, including the following main areas.

Flexibility and Scaling: Cloud services offer several advantages to research communities: services for diverse communities can be implemented and optimized, leading to higher satisfaction at lower cost; scaling services up or down (in any metric) is possible as needs change; providers offer costs based on economies of scale not achievable by a researcher, campus, or research collaboration; and costs are tied to actual usage.

Collaboration: Collaborations within a campus, or across institutions are significantly enabled by having a common platform for implementation and delivery of computing, storage and other SaaS. This further enables collaborations to share data, computational tools and workflows. Implementation specifics, however, are important in determining the impact on collaboration. Multi-institutional collaborations based on shared cloud resources can provide some degree of self-support, with less dependence on the local research support capabilities on each campus.

Robustness: Cloud services offer an intrinsic level of robustness in several ways: vendors focus on reliability and recoverability of services and data, and provide periodic "free" upgrades to hardware and services in the course of their own technology refresh processes. Encryption for data in motion and at rest are available, and, in general, tooling and system management capabilities are as good as or better than what is available on campus.

Researcher and Community Readiness: Perhaps the most significant enabler for the adoption of cloud services in research is that uptake by individual researchers is growing due to support from private cloud facilities funded by the NSF, the diversity and

capability of machine instances now offered, the growing list of research software suites and tools available as cloud services, and improved training for research users from vendors.

Further, the growing opportunity in public (commercial or academic) clouds is attractive for Minority Serving Institutions (MSIs) and smaller universities. These institutions typically find it hard to support their faculty and students with modern cyberinfrastructure using local funding and don't have local expertise to support the research. The national community has made extensive outreach efforts in the past with a series of "Cyberinfrastructure days" organized by XSEDE/TeraGrid and the Minority Serving Institutions (MSI)- Cyberinfrastructure Empowerment Coalition. These were successful in bringing the power of cyberinfrastructure to the attention of MSI's but follow through was hard as large NSF facilities were not well matched to most MSI needs. However, clouds seem well suited to support many MSI requirements, but this will only succeed if there is a support organization. This needs to be set up in collaboration with minority communities, so it satisfies their needs. It would need to be a resource that helped MSI faculty set up their needed cyberinfrastructure resources on clouds. A similar organization could be very helpful for many smaller universities.

12. Barriers for Cloud Services in Research

Workshop participants identified a number of barriers that need to be addressed to encourage the use and adoption of cloud computing. While some of the barriers are related to technology, software and training, other barriers are more complex and involve funding, policy, or licensing, making them more difficult to address. The workshop identified 4 main categories of barriers, funding, legal/compliance, technology and training/support.

Business Model: While pay-for-use is the standard model for acquiring cloud computing resources, this approach raises a number of issues for campuses. Because of the way campuses assign cost and operate, it is often much easier for a campus to get additional funds for hardware than it is to get additional funds for operations. And, because of the way cost accounting is done (direct, indirect, overhead), a campus cannot apply capital funds they were planning to spend on hardware to operations so that they can purchase cloud services; sometimes, indirect is charged for cloud services even though the systems do not sit on the campus. As a result, it is difficult for campuses to scale up cloud computing to meet their research needs until these issues are resolved. Other funding issues include billing complexities as campuses will need to know who is using the resources for chargeback and measuring use. Other issues include not-to-exceed capabilities so that misuse (inadvertent or otherwise) does not result in cost overruns and being able to monitor and manage use over the entire year.

Research funding uses a pay-in-advance fixed-cost model, unlike the majority of commercial and industrial activities that charge for products and services and can raise the price when creating the products and services becomes more expensive. For the commercial cloud to play a significant role in research computing infrastructure a fixed

cost business model must be developed. The commercial cloud is the ideal infrastructure to tie multiple data centric research projects across the world together; the infrastructure exists. But the data is spread across multiple vendors and ingress and egress charges to perform the most valuable cross-data-set analyses will quickly outrun any budget. This is exacerbated by the fact that the research is exploratory and the algorithms are at best optimized, and most are experimental and untested. A consortium approach to cloud infrastructure for research might work, just like organizations like Internet2 provides stable, large, and affordable network bandwidth to its members. The consortium can negotiate the required fix-cost contracts with cloud providers and other resource providers who are interested.

Legal and Compliance: Legal and compliance issues include management of software and database licenses, analytic tools and applications. Associated restrictions such as the ability of the campus to monitor usage becomes important. The use of "outside" cloud resources may create additional issues in identity management, authorization, and access for student and faculty researchers (both official and personal use). There are also a number of issues associated with data, especially related to privacy, security, and encryption, including network security, as data is moved from campus to the cloud provider. In some cases, the physical location of the data can be an issue if the cloud provider routinely stores the data in centers outside the US. While trust relationships between industry providers and campuses are not a legal issue, effective usage requires that some level of trust be established.

Optimization of Resource: Many of the issues and barriers associated with technology arise primarily because of rapid change and growth of cloud computing. Cloud industries are pushing capabilities to provide better or faster computing services, or are expanding and developing new applications to provide services to new users and domains. Universities able to couple and partner with cloud providers can ride these developments to support new research directions. For example, the cloud industry has significant efforts in green energy, in developing more robust infrastructure, disaster recovery capabilities, and economies of scale. These areas important to higher education but riding this wave will require universities to be more agile in providing the education and support infrastructure to ensure that these capabilities can be put into service by their scientists.

Training and Support: The challenges required to provide adequate training and support services for cloud computing infrastructure is critical. The rapid growth in cloud capabilities requires significant expertise on the campus to help users understand new capabilities and applications, and how those fit their requirements. Without significant effort and support from the campus, students and researchers will not be able to utilize the latest capabilities or resources. Other issues include providing adequate portals and gateways; supporting collaboration (teams as well as communities), flexibility and elasticity, and scaling and optimization. Lastly, it will be important to have some sort of cloud research help desk/team or some sort of body of knowledge on campus that students and faculty can go to.

13. Findings and Path Forward for Cloud in Academic Research Computing

The workshop enumerated a number of findings that relate to the future of cloud for academic research computing, the impacts on the future of research and the way researchers conduct their work, and the path forward for the inclusion of cloud technologies and services in using cloud computing for research at both the campus and national level. Further, these findings help set forth strategies on how to enable the use of cloud technologies and services more broadly across the overall research community, as well as helping campuses/campus research computing organizations respond to the growing advent of cloud. The list of specific findings from the workshop are as follows:

- The emerging conversation is not about whether academic research computing will take place in the cloud as has been the case with many previous reports and meetings, but rather how best to support it.

- Having the right people in place to support the use of cloud services is essential to successful adoption

- It is important to establish some sort of forum for the exchange of knowledge and ideas in cloud and use this as an ongoing mechanism to assess progress of the community toward cloud adoption.

- Services and tools that enable further adoption of cloud technologies for research computing should be sustained and expanded as much as possible.

- Establishing and developing shared training pilots among campuses and all cloud providers (including industry partners) to train and educate cyberpractitioners and researchers will be very useful.

- More effort need to be spent on documenting and communicating a clear, centralized summary of currently available NSF and federally-funded cloud programs in a service catalog fashion. It would be useful if this included summaries of relevant successful use cases for research.

- More effort and energy should be spent on documenting and testing research examples to explore interoperability issues, particularly across multiple data sets and cloud providers.

- Differing business models across campuses and research groups need to be more thoroughly analyzed including more discussion about indirect cost considerations and other limiting factors. These models should publicize those that are successful, especially addressing those that illustrate the value of aggregation and collaboration.

- Participants noted that the cloud provides an opportunity for enhanced collaboration and data sharing, but best practices for data movement, handling, and provenance need to be established to help guide these efforts.

- Explore and encourage the establishment of a follow-on meeting for smaller institutions or those without a large local research computing practice to consider cloud adoption issues in this context, including potential economic issues, local support needed, available resources and so forth.

- The NIH pilot initiative(s) to use cloud computing looks very promising; the pilot should be monitored to see if similar pilots can be implemented in other research disciplines

References

[1] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, Nancy Wilkins-Diehr, "XSEDE: Accelerating Scientific Discovery", Computing in Science & Engineering, vol.16, no. 5, pp. 62-74, Sept.-Oct. 2014, doi:10.1109/MCSE.2014.80

[2] Stewart, C.A., Cockerill, T.M., Foster, I., Hancock, D., Merchant, N., Skidmore, E., Stanzione, D., Taylor, J., Tuecke, S., Turner, G., Vaughn, M., and Gaffney, N.I., Jetstream: a self-provisioned, scalable science and engineering cloud environment. 2015, In Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure. St. Louis, Missouri. ACM: 2792774. p. 1-8. http://dx.doi.org/10.1145/2792745.2792774.

Endnotes

1. U.S. Department of Energy, Office of Advanced Scientific Computing Research. (December, 2011) The Magellan Report on Cloud Computing for Science. Retrieved from https://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Magellan_Final_Report.pdf.

2. National Academies of Sciences, Engineering, and Medicine. (2016) Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020. Washington, DC: The National Academies Press. doi:10.17226/21886.

3. Ames, G., Anderson, C., Hillegas, C., Lance, T., Lane, R., Lynch, C., Marinshaw, R., Monaco, G., Zaborowski, E., and Zottola, R. (July, 2015) Research Computing in the Cloud: Functional Considerations for Research. EDUCAUSE Center for Analysis and Research (ECAR. Retrieved from https://library.educause.edu/resources/2015/7/research-computing-in-the-cloud-functional-considerations-for-research

4. Lifka, D., Foster, I., Mehringer, S., Parashar, M., Redfern, P., Stewart, C., and Tuecke, S. XSEDE Cloud Survey Report. (September, 2013) Retrieved from https://www.ideals.illinois.edu/handle/2142/45766.