XSEDE Cloud Survey Report

David Lifka, Cornell Center for Advanced Computing
Ian Foster, ANL and The University of Chicago
Susan Mehringer, Cornell Center for Advanced Computing
Manish Parashar, Rutgers University
Paul Redfern, Cornell Center for Advanced Computing
Craig Stewart, Indiana University
Steve Tuecke, ANL and The University of Chicago

A National Science Foundation-sponsored cloud user survey was conducted from September 2012 to April 2013 by the XSEDE Cloud Integration Investigation Team to better understand how cloud is used across a wide variety of scientific fields and the humanities, arts, and social sciences. Data was collected from 80 cloud users from around the globe. The project descriptions in this report illustrate the potential of cloud in accelerating research, enhancing collaboration, and enriching education. Cloud users provided extensive data on core usage, preferred storage, bandwidth, etc. and described cloud benefits and limitations for their specific use cases. Educators, research administrators, CIOs, and research computing practitioners may find value in this data when considering the use and/or deployment of public, private, or hybrid clouds to complement current cyberinfrastructure.

XSEDE
Extreme Science and Engineering Discovery Environment

September 2013

Contents

Executive Summary

The XSEDE Cloud Integration Investigation Team was asked by the National Science Foundation to conduct a cloud use survey in order to get a better understanding of how cloud is being used today in research and education. Eighty projects from around the globe participated in the survey. The participants represent a wide range of science and engineering disciplines as well as the humanities, arts, and social sciences.

Several characteristics of the *XSEDE Cloud Survey Report* make it unique:

- Unlike most cloud surveys conducted to date, this report is focused solely on the use of clouds for research and education rather than administrative or business IT
- Twenty-two sets of quantitative data were collected on each education and research project, e.g., preferred cloud development environment, cloud use regularity, data movement, bandwidth into/out of the cloud, etc.
- Qualitative data was collected from follow-up interviews and the analysis of associated documentation/publications in order to provide a more in-depth understanding of the user experience.

This report is intended to help educators, research administrators, CIOs, and research computing practitioners envision what role cloud might play in research, teaching, and learning at their respective institutions. While cloud technology is still maturing, it is our belief that it is here to stay. Academic institutions need to ascertain how cloud fits in their cyberinfrastructure (CI) strategy and plan and adapt accordingly.

Survey Finding #1: Top 3 Reasons Researchers and Educators use the Cloud

According to the survey data, the top three reasons researchers and educators use the cloud is:

1. On-demand access to burst resources
2. Compute and data analysis support for high throughput scientific workflows
3. Enhanced collaboration through the rapid deployment of research team web sites and the sharing of data.

Survey Finding #2: Applications Identified as Good Candidates for the Cloud

Survey participants identified several applications and programming models as good candidates for the cloud:

- *MapReduce* – for processing and analyzing large data sets. MapReduce was cited by the survey participants as the most frequently used special feature available from their cloud service providers that enabled their research.
- *High throughput, embarrassingly parallel workloads* – for analyzing thousands of molecules, particle collisions, etc. Examples include large scale data mining, BLAST searches, Monte Carlo simulations, (Value-at-Risk, supply chain networks, etc.), image analysis (digital pathology, tomography, etc.), and other loosely coupled workloads.
- *Academic labs and teaching tools* – for scaling educational experiences to dozens, hundreds, or even, thousands of students. Cloud-based labs are either always on or provisioned on-demand. Examples are freshman biology students accessing highly visual, interactive cloud-hosted teaching tools to learn population genetics and the mathematics behind it or data management students learning how to write applications or use Hadoop [1], [2]. Benefits noted by faculty included overcoming resource limitations in existing lab environments and preparing students for

their future in a "cloud computing world." The convergence of mobile and cloud services will likely accelerate the design and deployment of cyberlearning experiences, e.g., faculty-developed digital textbooks, interactive classroom simulations, MOOCs, etc.

- *Domain-specific computing 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. For example, Cloud BioLinux provides instant access to a range of pre-configured command line and graphical software applications including a full-featured desktop interface, documentation, and over 135 bioinformatics packages [3].
- *Commonly requested software* – 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. Globus Online, a software service on XSEDE, uses a set of SaaS components to make it easy to move massive amounts of data without requiring custom end-to-end systems.
- *Science Gateways* – the rapid elasticity of cloud-based gateways can reach large communities of researchers and citizen scientists with on-demand services. Zooniverse, the largest citizen science gateway in the world, uses 700,000 cloud core hours per year and 100TB of data to support nearly a dozen websites on space, climate, and the humanities [4].
- *Event-driven science* – applications that must scale quickly to respond to real-time events are another good candidate for the cloud. California volunteers are helping scientists gather seismic data by hosting hundreds of small seismometers in their homes and offices. During quiescent periods the only data sent over the Community Seismic Network is control traffic; during an event, the ground motion intensity data is substantial [5].

These types of applications are increasing rapidly. Unlike traditional HPC workloads, most require many cores rather than fastest performance per core. The *NSF Cyberinfrastructure for 21st Century Science and Engineering Advanced Computing Infrastructure Vision and Strategic Plan* recognizes the growth of these applications and calls for a more comprehensive and balanced cyberinfrastructure to support the entire spectrum of NSF-funded communities [6].

Survey Finding #3: Cloud Benefits Reported by the Survey Participants

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 [7]."

If an application is cloud-friendly and if system utilization projections do not justify purchasing on-premise servers, i.e., usage is intermittent or "spikey," clouds can reduce capital expenditures and associated operation and maintenance costs.

Clouds provide small labs, departments, and budget-constrained colleges and universities access to computing capabilities that they might otherwise not have. They democratize access and, in the case of Platform as a Service and Software as a Service, mask computing complexities. As such, clouds help to address the "long-tail research" problem by providing resource-limited organizations with on-demand access to tools for data discovery, collection, and analysis.

It is important to increase the number and diversity of researchers, educators and students participating as creators and users of cyberinfrastructure. The addition of clouds or cloud access to campus, regional, and/or national cyberinfrastructure can complement essential investments in high-end computing and enable a wider class of researchers to take risks and innovate. The on-demand, feature-rich environments offered by the cloud may help to increase CI participation by underrepresented groups as well.

Survey Finding #4: Cloud Challenges Reported by the Survey Participants

Survey participants reported several challenges in using the cloud, e.g., learning curve, virtual machine performance, data movement costs, etc.

Like any new technology, there is a learning curve with the cloud. Creating, deploying, and managing a cloud instance, for example, is a new experience for many researchers and faculty. Investment in cloud training, therefore, is important so that researchers can focus on the science rather than the technology enabling it. Systems administrators need to be cloud savvy as well.

Many applications, such as those listed in Survey Finding #2, run efficiently and cost-effectively in a virtual machine environment. Performance for these applications, however, may be somewhat less than optimal. This is often compensated for by running slightly longer or by adding cores. Tightly coupled HPC workloads tend to not scale well in a virtual machine environment. Competing for CPUs, memory, disk, and network I/O in a shared cloud environment is not the same computing experience as running on a dedicated cluster. Databases also may have scalability and performance issues since they are highly dependent on I/O speeds. Some cloud providers offer dedicated bare metal clusters and database servers to address these performance limitations albeit at a higher price point.

When analyzing the appropriateness of a particular cloud service for a given application, it is important to make the distinction between virtual cloud resources (a shared virtual machine environment) and physical cloud resources (a dedicated bare metal cluster on the network). 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. Time to access and overall cost-performance are other factors worth considering.

Several survey respondents reported that they were surprised by the cost to move data when they received their monthly bill. Most cloud service providers charge by the GB to move data out of the cloud. To avoid or minimize these costs, some researchers generate their data in the cloud and leave it there; others take advantage of community data sets that are already available in the cloud. If a lot of data must be regularly moved out of the cloud, an on-premise resource may be a best solution.

Surprisingly, the educators and researchers surveyed were not overly concerned about cloud security. This may be because unlike businesses that have very real concerns about protecting IP and customer data, much of academic research is publicly-funded and is, therefore, required to be made publicly-available. An exception noted was HIPAA data which due to its stringent security requirements may be best served by a private cloud environment, although public clouds are actively working on hosting solutions to secure this data type. A right-sized, on-premise private HIPPA resource could potentially cascade to a regional HIPAA cloud, or even a public cloud, providing the hybrid architecture was HIPAA compliant.

Survey Finding #5: Continued Investment Needed

While clouds can clearly provide value to researchers and educators today, survey findings suggest that continued investments in basic, applied, and experimental cloud computing research are needed to address cloud challenges. 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.

Research in cloud computing is an important technology frontier. Survey participants identified many areas of research interest such as domain-specific applications, dynamic provisioning of images, network support for clouds, data portability, and aggregating heterogeneous resources as services. Other CS research possibilities noted included cloud-hosted real-time intelligence systems, multiparty security dataflow solutions for OpenFlow networks, and big-data machine learning algorithms for rapidly evolving data sets [8].

A strong interest in multi-clouds was also expressed. 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 [9]. The challenge will be implementing a management framework that can span all cloud environments.

Introduction

The goal of the Extreme Science and Engineering Discovery Environment (XSEDE) is to enhance research productivity. NSF through the XSEDE integrating fabric is committed to promoting a diversity of computing resources, inclusive of clouds, and, in addition, recognizes the opportunity for cloud to play a significant role in many other parts of a scientific workflow. XSEDE must embrace cloud, identify complementary areas that cloud can support, and have a clear strategy for integrating cloud into national cyberinfrastructure.

To achieve this objective, a clear understanding of cloud use cases in research and education was needed. Since this use case data was not readily available except for a few public cases and, even then, not to the level of detail desired, the NSF Directorate for Computing and Information Science and Engineering (CISE) Division of Advanced Cyberinfrastructure (ACI) asked the XSEDE Cloud Integration Investigation Team to conduct a survey focused on the use of cloud for research and education in science and engineering and the humanities, arts, and social sciences.

The goal of the survey was to help XSEDE management understand the cloud computing experiences of this user population so that they can better plan for integrating cloud into the XSEDE architecture.

Collecting Cloud Use Data

The XSEDE Cloud Survey [10] was conducted from September 2012 to April 2013. Cloud use data was collected from eighty research and education projects from around the globe through an extensive online survey, follow-up interviews, and a literature search focused on research and education projects that use the cloud. The projects surveyed represent twenty-one science and engineering disciplines as well as disciplines from the humanities, arts, and social sciences.

The survey data provides a detailed view of how cloud computing was used to enable each research and education project. The data collected included:

- cloud use cases
- service providers
- special features available from the cloud provider that enabled the research
- preferred development environments
- cloud use regularity
- number of cores used peak and steady state
- number of core hours used per year
- reasons for storage access
- preferred storage models
- amount of storage used during program execution
- short-term//long-term storage needs
- amount of data moved into/out of cloud
- bandwidth into/out of cloud
- bandwidth to storage within the cloud
- types of data moving
- data accessibility
- software used in the cloud
- cloud funding sources
- research funding sources
- comments on cloud capabilities/features
- comments on cloud problems/limitations

The summary data provided in this report is followed by individual project data organized by discipline.

Additional Notes and Analysis

Individual project data is supplemented with additional notes and references drawn from academic publications, case studies, reports, and interviews.

An analysis of cloud benefits and cloud limitations as reported by the survey participants is also featured in this report.

Potential Cloud Impact

While cloud is still in the early adopter phase of the technology adoption lifecycle, particularly in regards to its use in research computing, cloud has a strong potential to increase the number and broaden the diversity of advanced computing users.

It is our hope that this survey data will provide university administrators, research computing directors, scientists, and educators with insights into how, given the right application, cloud computing can enable more efficient research and education.

We wish to thank the project participants who graciously gave their time to complete the cloud survey and participate in follow-on discussions. This was truly a community effort and the breadth and depth of first-hand data provided will help all of us to better understand what role clouds might play in multi-level cyberinfrastructure.

XSEDE Cloud Integration Investigation Team

David Lifka, Cornell University Center of Advanced Computing (PI)
Ian Foster, Argonne National Laboratory and The University of Chicago
Susan Mehringer, Cornell University Center for Advanced Computing
Manish Parashar, Rutgers University
Paul Redfern, Cornell University Center for Advanced Computing
Craig Stewart, Indiana University
Steve Tuecke, Argonne National Laboratory and The University of Chicago

We wish to acknowledge John Towns, XSEDE Principal Investigator and Project Director, Barry Schneider, NSF Program Director, and Irene Qualters, NSF Program Director, for calling for a more in-depth understanding of the use of cloud computing in research and education and for contributing to the insightful analysis of the cloud survey data.

Thanks also to the National Science Foundation Division of Advanced Cyberinfrastructure for sponsoring this project.

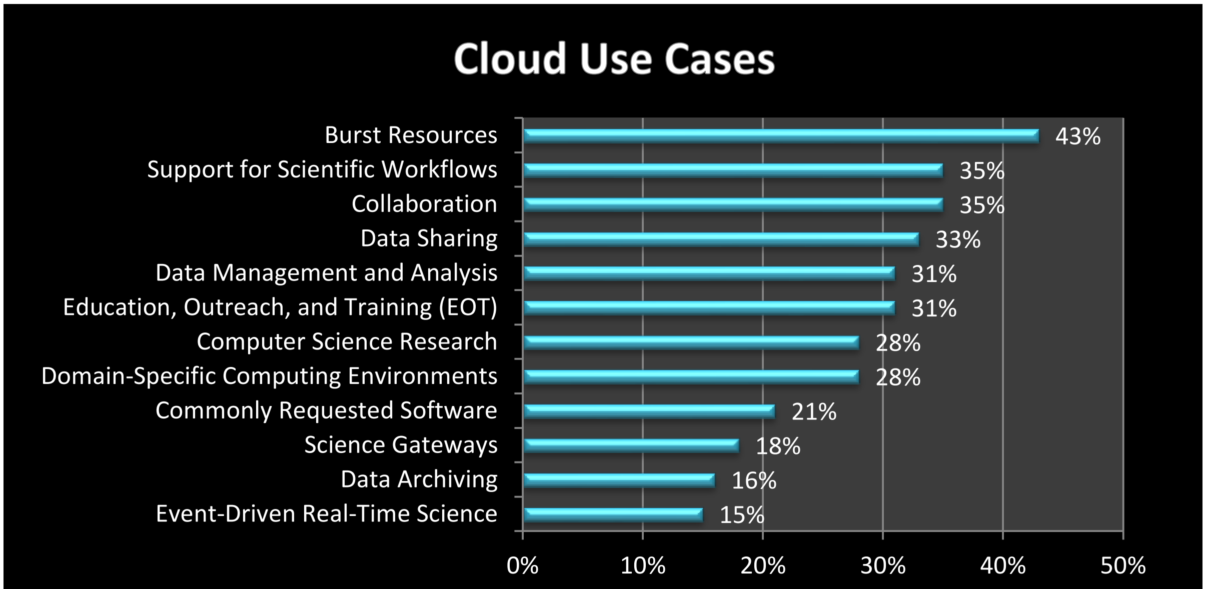
Cloud Projects Surveyed: Summary Data

Cloud Use Cases

With inputs from the HPC and cloud services community, the XSEDE Cloud Integration Investigation Team defined twelve cloud use case categories:

- *Burst Resources* – "bursting" is the addition of compute/analysis resources on demand to augment campus clusters, Open Science Grid (OSG), departmental clusters, and high-profile applications in time of need where computation or analysis is effective with distributed resources.
- *Collaboration* – collaboration can be enhanced by the rapid deployment of research team wikis and web sites for communications, project planning/coordination, documentation, and document/data sharing.
- *Commonly Requested Software* – provide economies of scale for software licenses and optimal execution environments, e.g., MATLAB, R, etc.
- *Computer Science Research* – includes topics such as cloud infrastructure, systems/middleware for cloud applications and enterprise, and web and mobile applications.
- *Computing and Data Analysis Support for Scientific Workflows* – workflows tend to be loosely-coupled parallel applications that involve a series of connected tasks. Examples are the computing and/or analysis of data generated by high-throughput gene sequencing machines, telescopes, simulations, etc.
- *Data Archiving* – data archiving requires a location where data sets and collections can be archived for their perceived useful lifetime. This has different cost and access requirements than active data that is actively being shared or analyzed.
- *Data Management and Analysis* – cloud resources provide a low-risk exposure to and testing of operating systems and application software technologies in terms of time spent, disruption of production resources, and cost that may provide a potential benefit to researchers, e.g., the use of databases for storing and analyzing research data more effectively.
- *Data Sharing* – data sharing resources provide a location where data can be efficiently and cost-effectively stored and shared with a potentially high volume of users and accessed by anyone.
- *Domain-Specific Computing Environments* – custom software environments for data analysis/pre- and post-processing stages of scientific workflows or event-driven science. Instead of a web-based interface such as a Science Gateway, these are virtual operating systems and application software that researchers log into and use remotely via SSH and/or xterms. One or more virtual servers can be booted as required to support a researcher and their collaborators. One feature that typically distinguishes these kinds of resources is interactive access as opposed to batch or web-based access. Sometimes collections of these nodes are used simultaneously as a "personal parallel computer" that does not require a scheduler. This is well-suited for supporting on-demand parallel analysis, visualization, and deployment of specialized parallel environments and tools such as Hadoop and MapReduce.
- *Education, Outreach, and Training (EOT)* – customized software/development/programming environments for EOT, e.g., all software and tools installed so that students can remote-desktop into a common environment to meet training workshop, virtual workshop, or traditional classroom course learning objectives.
- *Event-Driven Real-Time Science* – scientific events (often natural, e.g., weather, geophysical or oceanographic) that have corresponding data from sensors that scientists wish to analyze immediately as it becomes available. This results in a spike in demand for computing, storage, and data analysis by domain scientists. Once the event has passed, usage drops off.
- *Science Gateways* – domain-specific web portals that provide the community of researchers in a particular research domain access to the common features that they care about, which may include calendars of events, news, publications, data, software tools, and seamless access to simulations/data analysis, normally directly from the web portal without the researchers having to know anything about data or resource locality and the technical details of using/accessing them. They also can provide entrées into more traditional HPC environments.

From the twelve cloud use case listed above, survey participants were asked to select which cloud use cases their research or education project represented. Burst resources was cited as the most common cloud use case, followed by computing and data analysis support for scientific workflows, collaboration, data sharing, and data management and analysis. Education, outreach, and training (EOT) and the use of the cloud for computer science research were also commonly cited use cases.

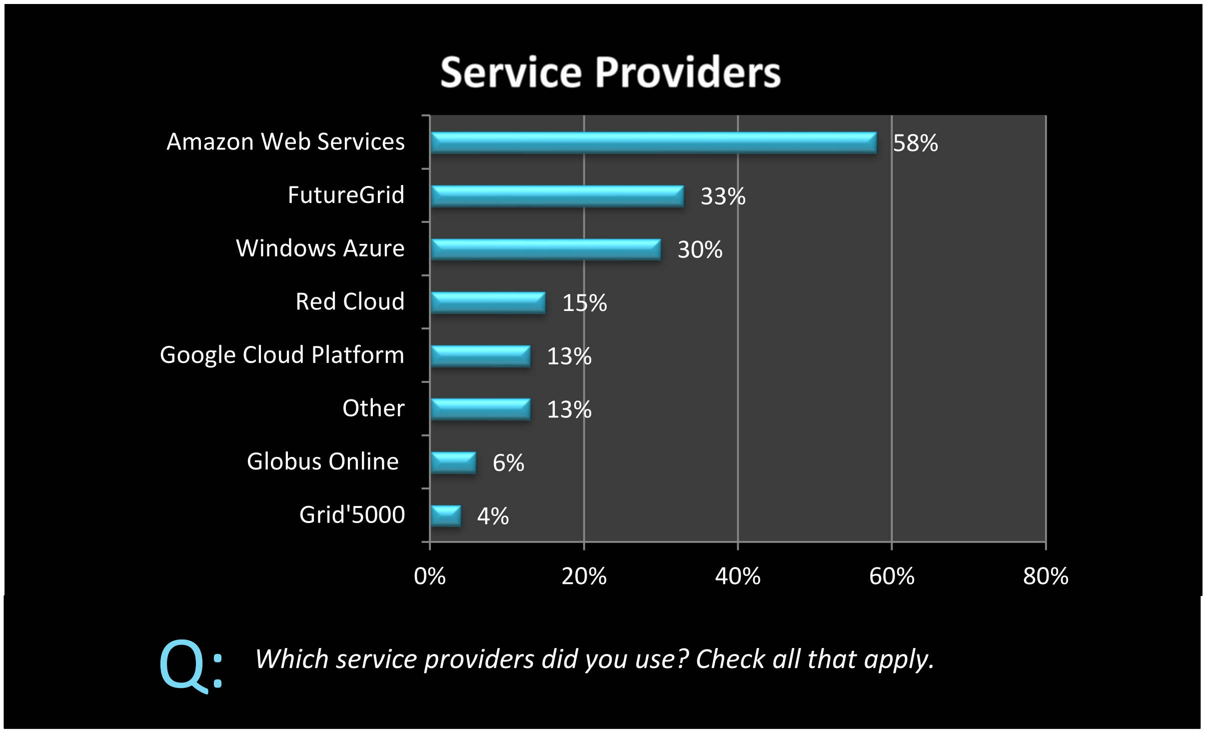


Q: *What cloud use cases are represented by your research or education project? Check all that apply.*

Cloud Service Providers

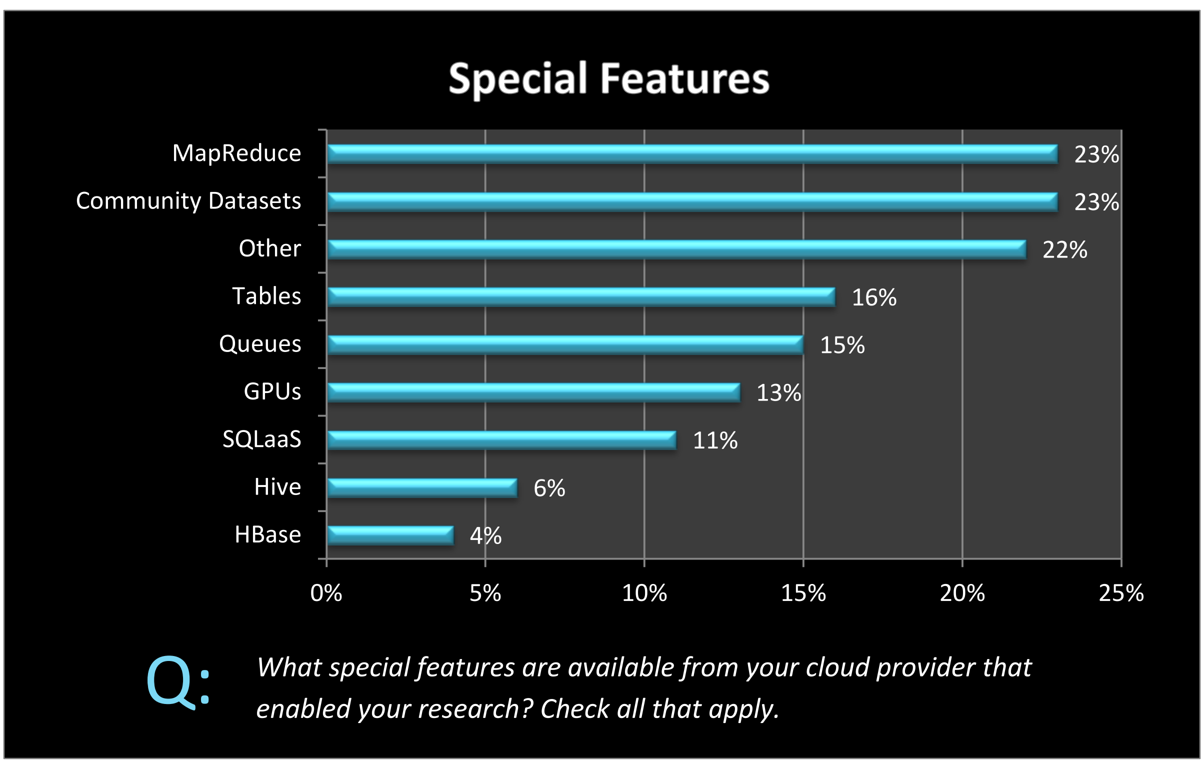
The researchers and educators surveyed used a variety of public and private cloud service providers. Fifty-eight percent used Amazon Web Services (AWS) followed by FutureGrid, Windows Azure, Red Cloud, Google Cloud Platform, and Globus Online. “Other” service providers identified by the survey participants included CloudSigma, Nimbix Accelerated Compute Cloud, Open Science Data Cloud, Open Science Grid, and Penguin On-Demand HPC Cloud Service (POD). Some service providers, such as CSC, POD, and Rackspace offer tightly coupled, non-virtualized computer clusters over the network in addition to or rather than shared virtual machine environments. It is important to make the distinction between shared virtual machines (public clouds) and dedicated, single tenancy, non-virtualized clusters on a network (hosted private clouds) when comparing cloud service offerings.

While other cloud surveys, e.g., Forrester [11], rank “big 3” usage (AWS, Azure, Google) in the same order as this survey, it should be noted that the “Service Provider” used statistics in the table below reflect the eighty research and education projects surveyed. They should not be interpreted as an indicator of overall market share or the superiority of one service over another. The goal of this survey was to collect cloud use data from as many disciplines as possible and to represent a diversity of providers. Each cloud service provider should be considered based on its own merits and the applicability of that particular service and features to the application at hand. Application requirements analysis and cost-performance comparisons are essential prior to selecting a cloud service provider and/or deploying a private cloud. OEMs such as Dell, HP, IBM, SGI, etc. and other service providers offer many cloud environments to choose from, e.g., Eucalyptus, OpenStack, VMware, etc. The Intel Cloud Finder is a useful search tool for identifying potential cloud service providers [12]. Providers are also listed in the Appendix on page 130.



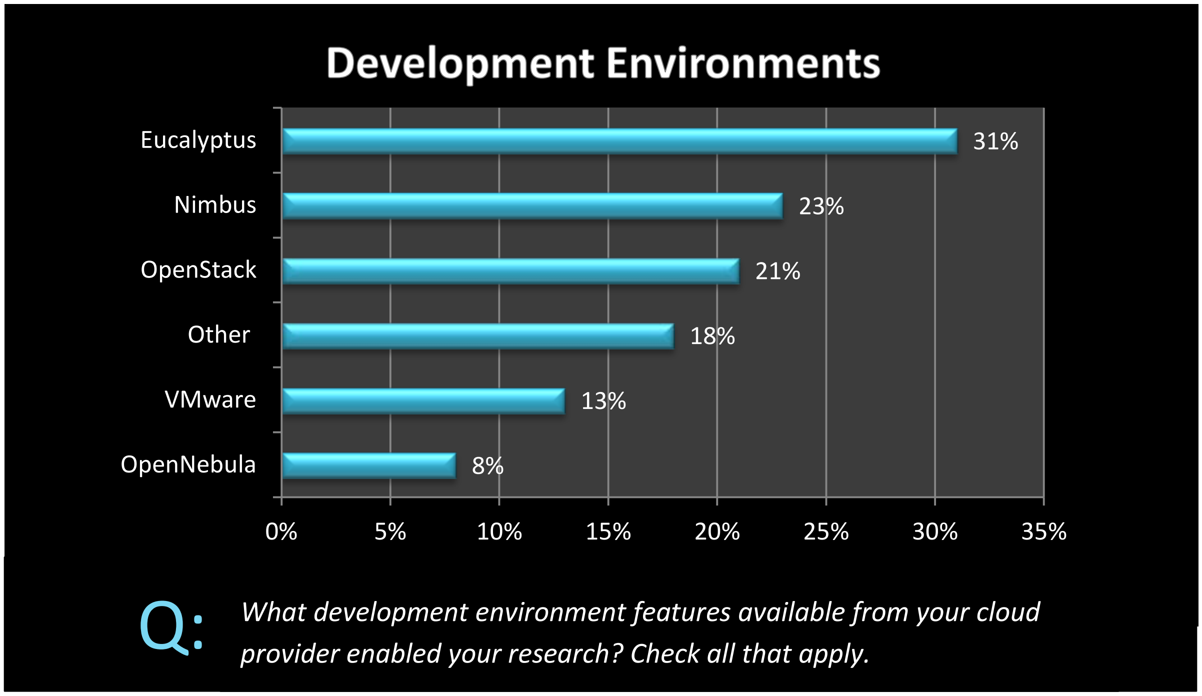
Special Features

Survey participants were asked to identify any special features provided by their cloud service provider that enabled their research. MapReduce and access to community datasets were the most highly used special feature. The “other” category included special features such as account management, root access, secure data store and computation, and web application platforms.



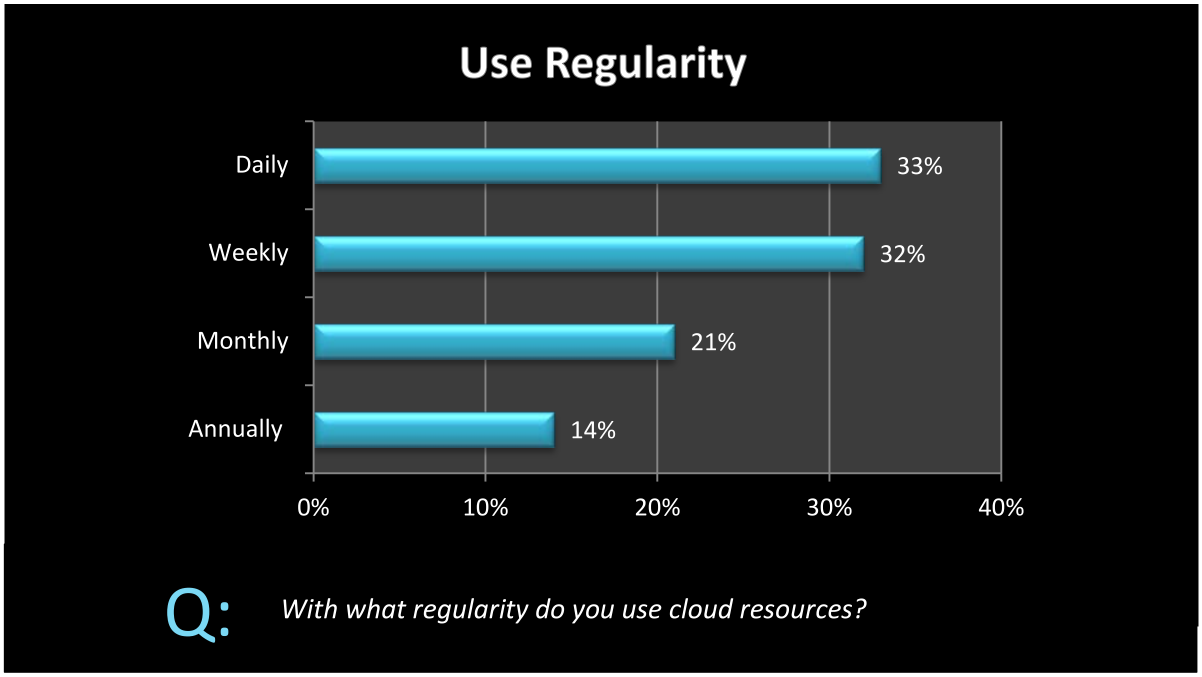
Development Environments

31% of survey respondents used Eucalyptus, the open source, AWS-compatible cloud development environment, followed by Nimbus (23%) and OpenStack (21%). “Other” development environments included CometCloud, Cooperative Computing Tools, Linux, StarCluster (MIT), VirtualBox, Windows Azure, and Xen. VMware and OpenNebula were also cited.



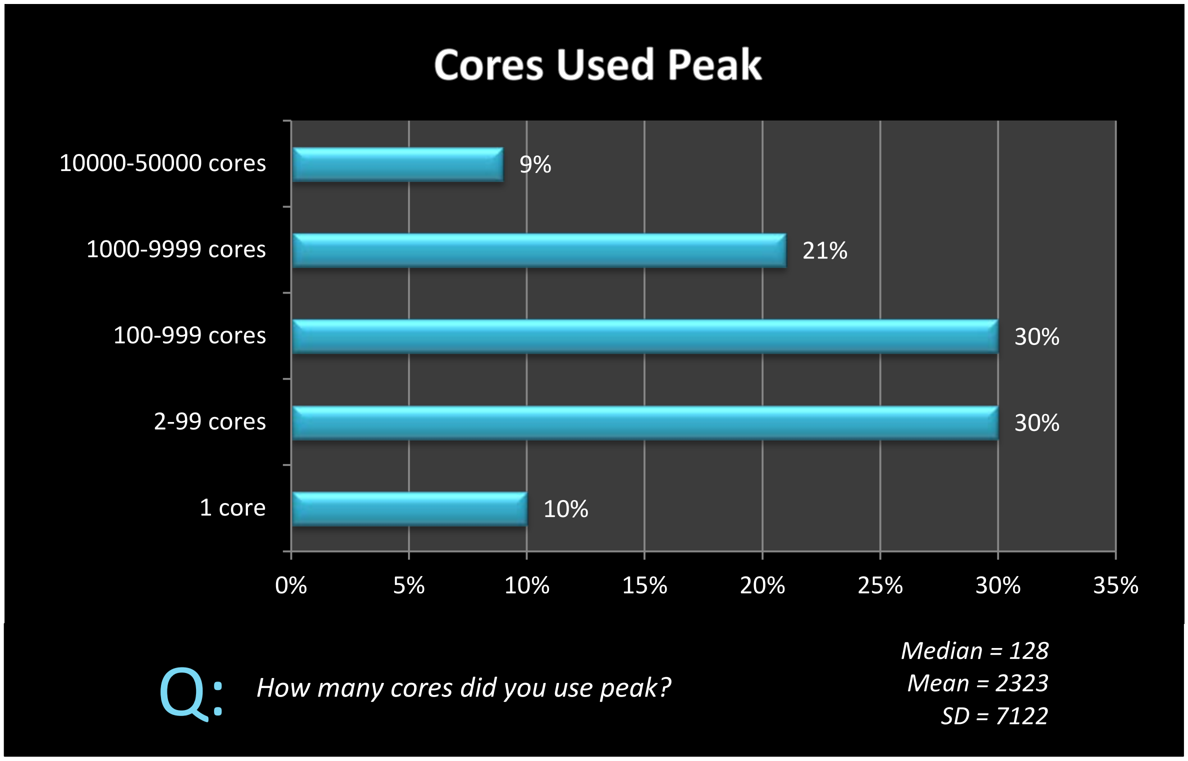
Use Regularity

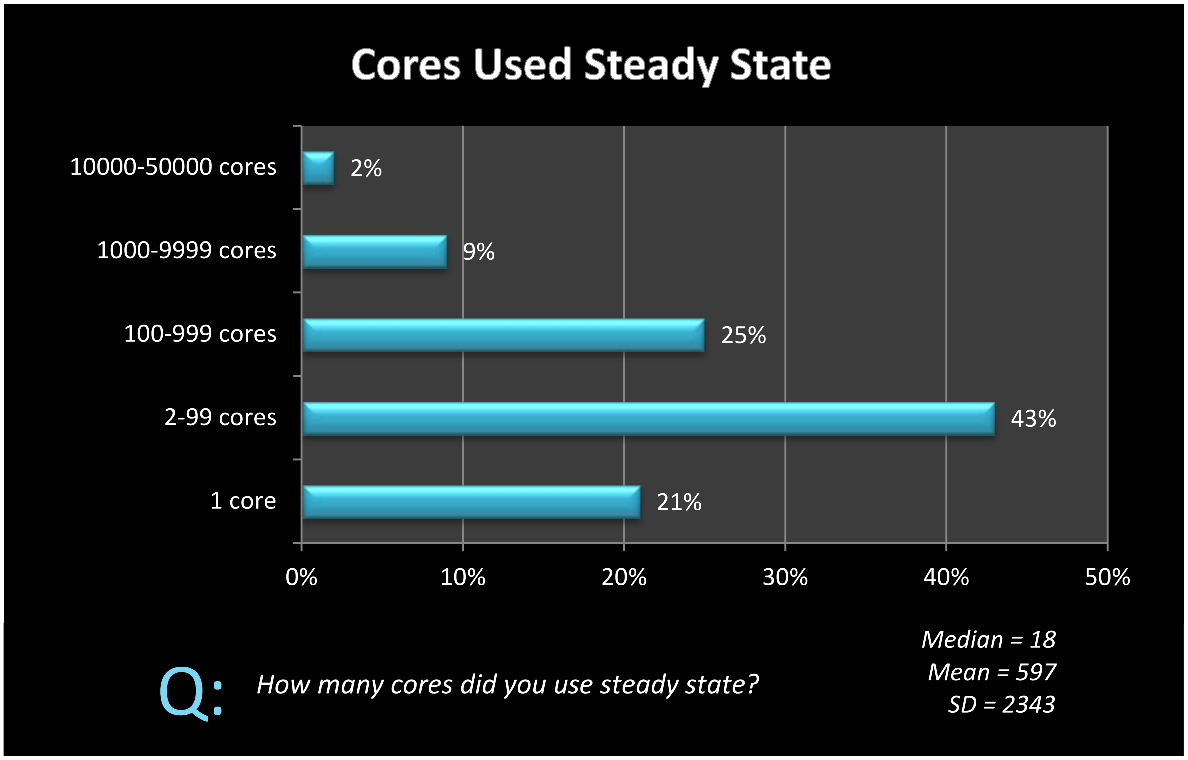
65% of the survey participants used the cloud daily or weekly.

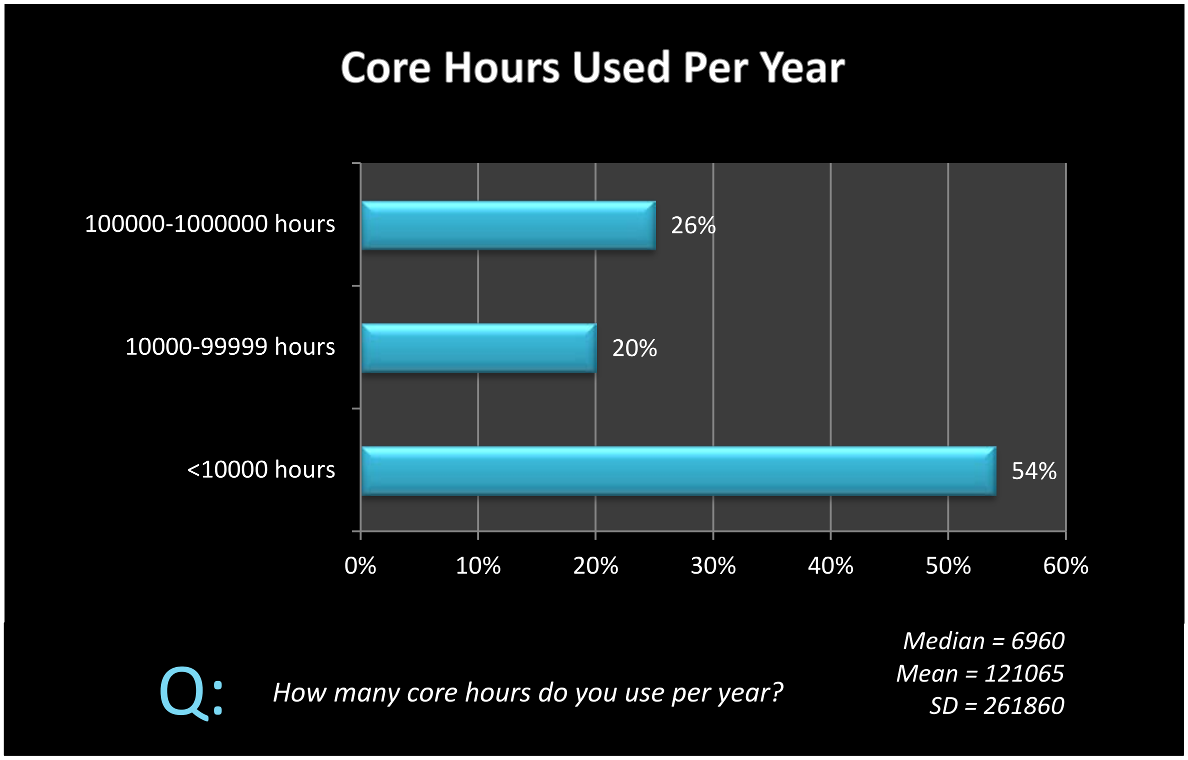


Core Usage Data

The median number of cores used peak was 128; the median number of cores used steady state was 18; and, the median number of cores used per year was 6960. The majority of researchers and educators surveyed used less than 1000 cores peak, 100 cores steady state, and 10000 core hours per year.

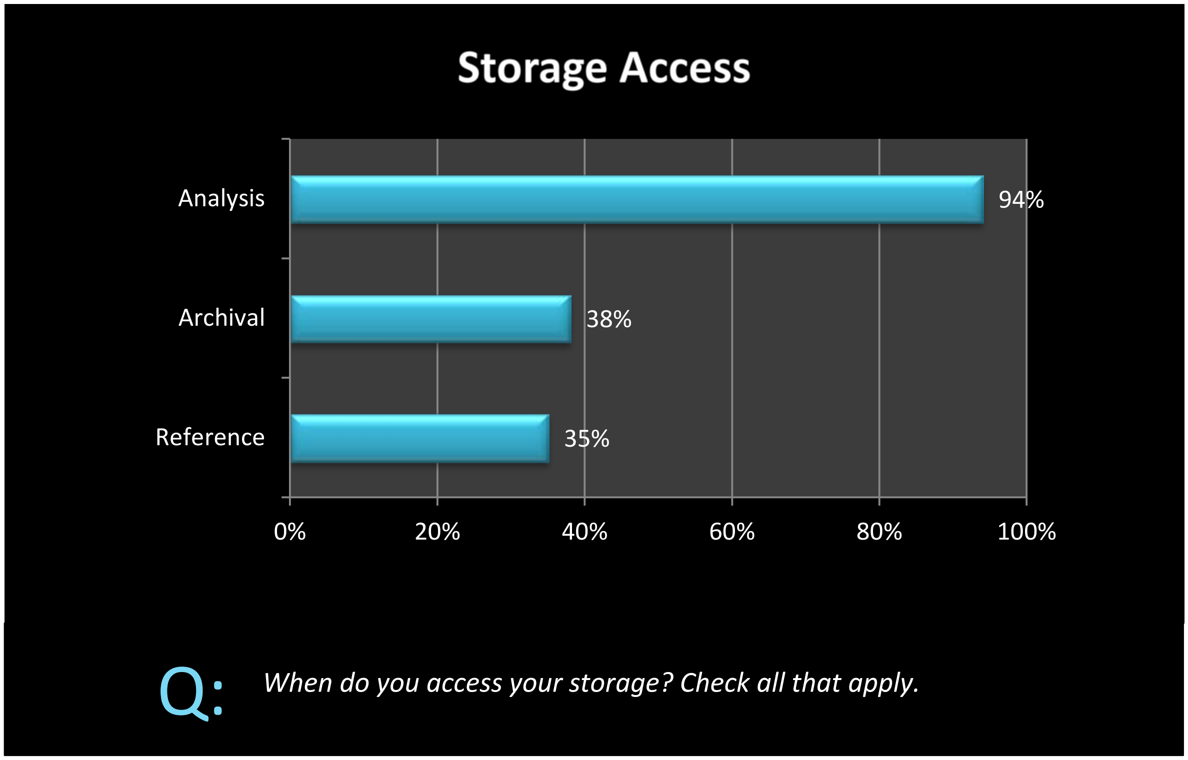






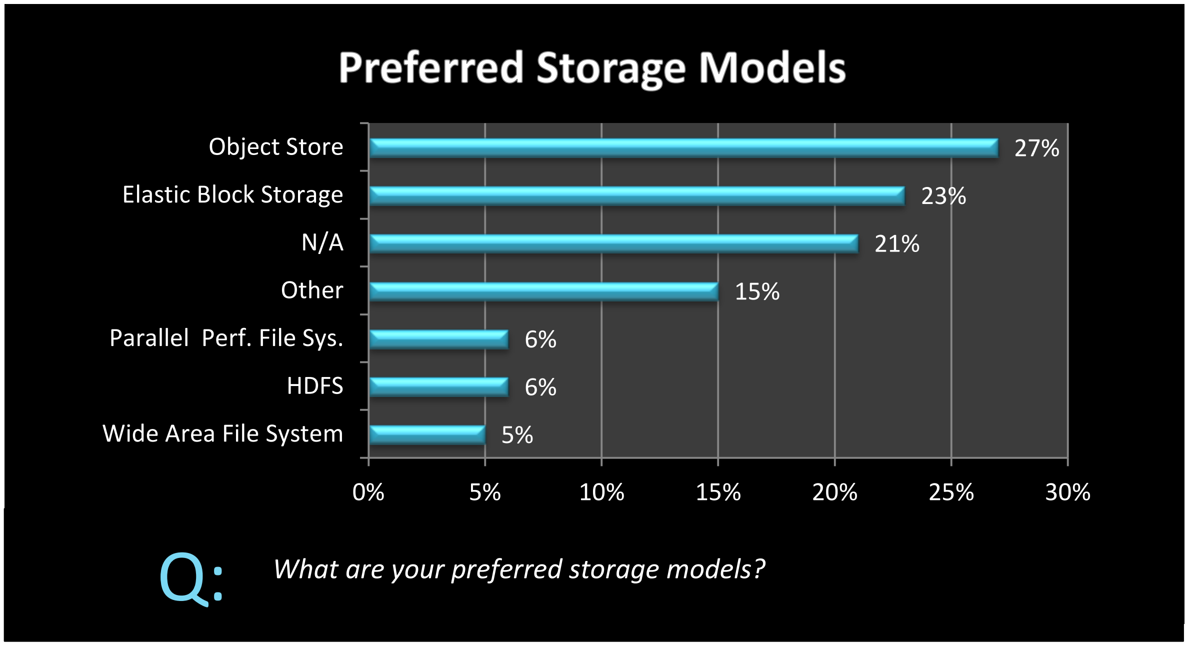
Storage Access

The vast majority of users surveyed said that they accessed cloud storage for the purpose of data analysis. 38% used the cloud for archival data storage.



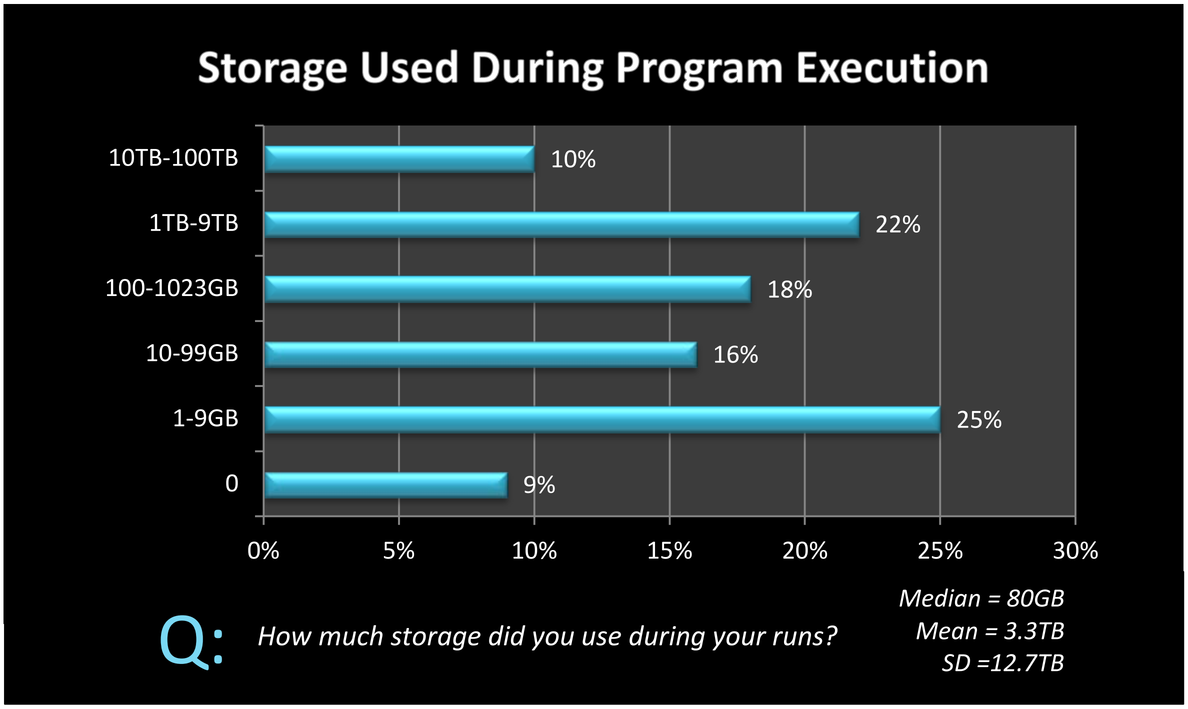
Preferred Storage Models

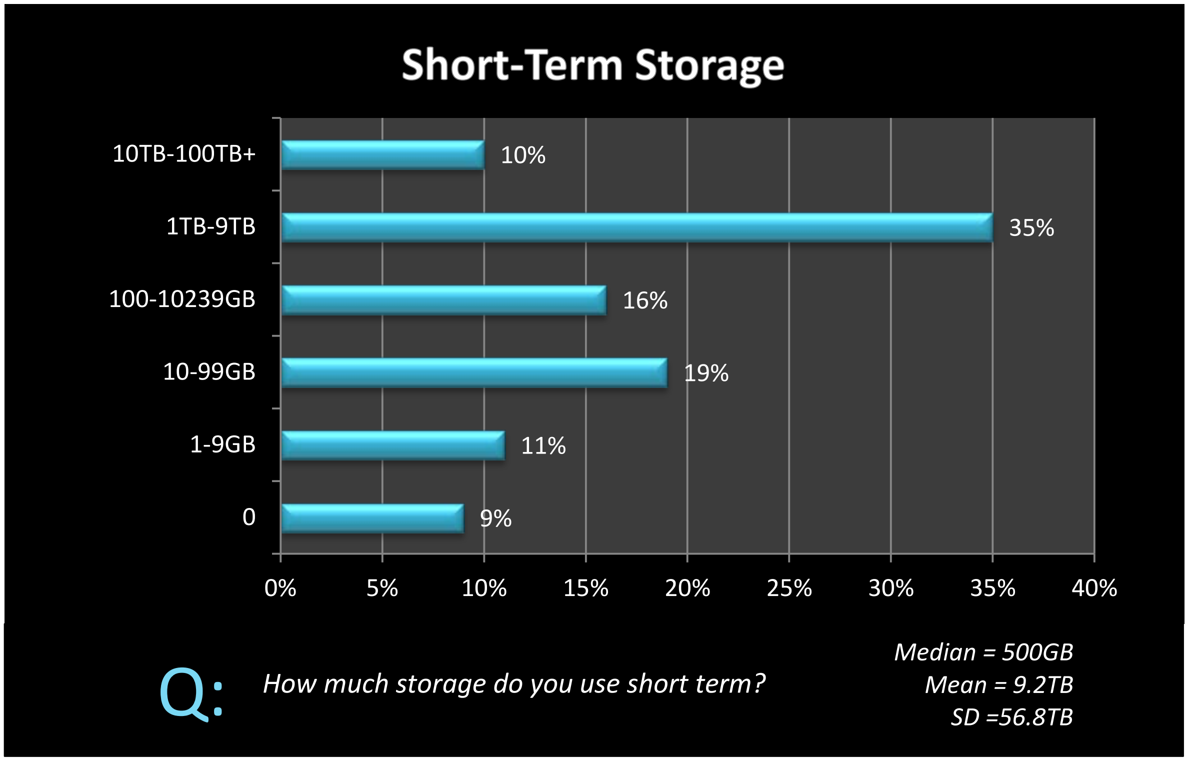
Object Store, e.g., Amazon S3 and OpenStack Swift, and Elastic Block Storage were the preferred storage models. “Other” models included conventional file systems, GlusterFS, NAS, RDMS, TomusBlobs, self-written unified image registry for clouds, and Windows Azure storage. Parallel performance file systems, HDFS, and Wide Area File Systems were used by a smaller percent of users.



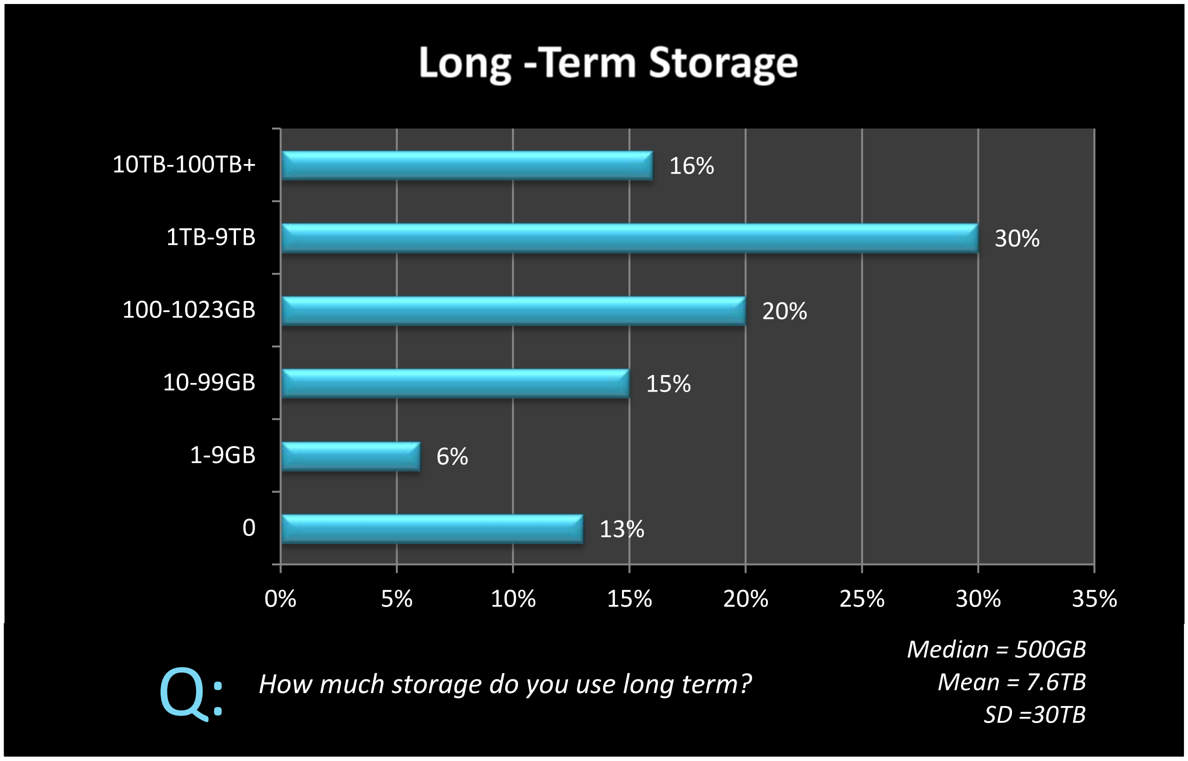
Storage Used: During Program Execution and Short-Term/Long-Term

The median amount of storage used during program execution was 80GB. Due to some very large storage users, e.g., macromolecular modelers, the mean amount of storage used during runs was 3.3TB.



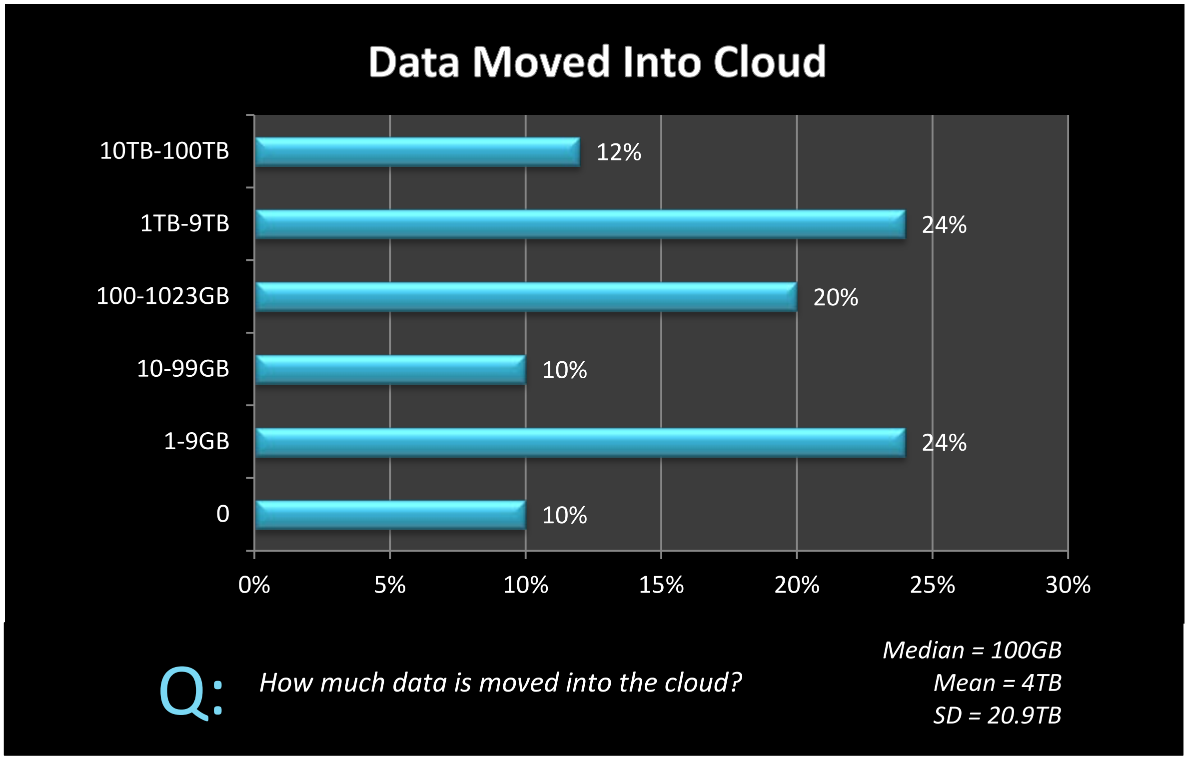


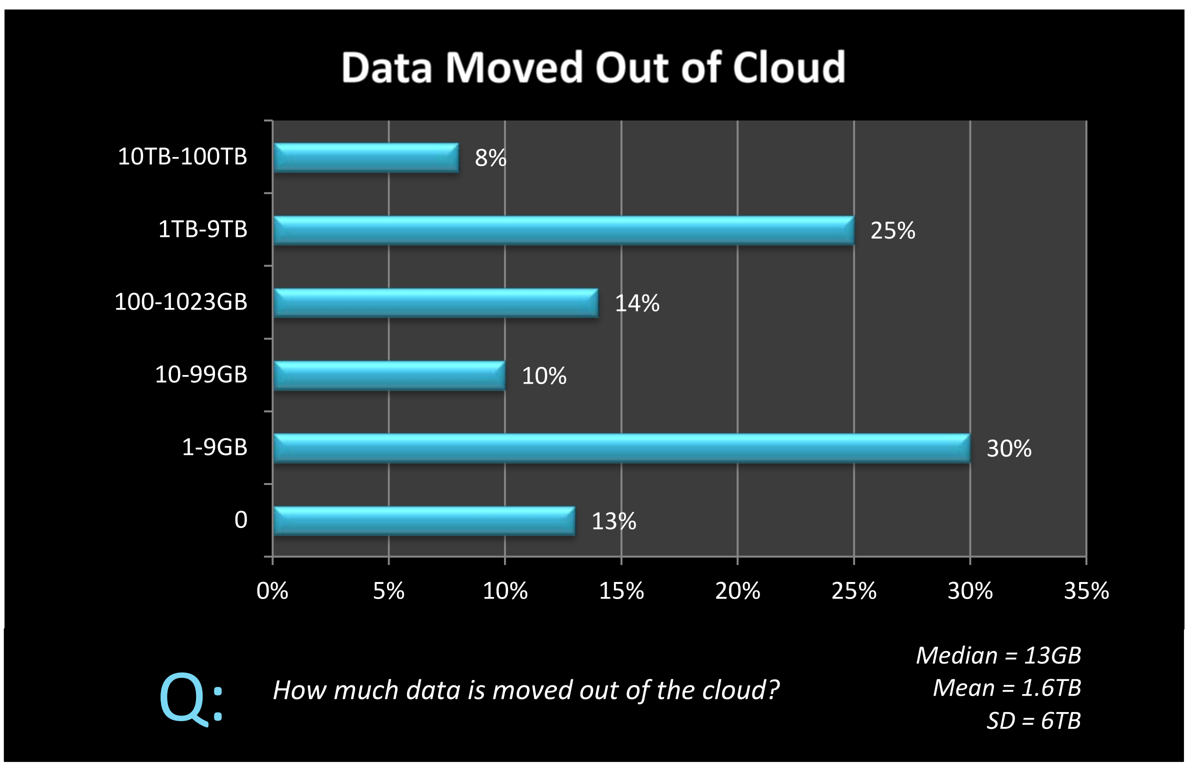
The median amount of data stored short-term was 500GB. The median amount of data stored long-term was also 500GB. The mean amount of data stored, both short-term and long-term, is considerably higher because of a subset (10%-16%) of scientists storing 10TB to 100TB+.



Data Movement

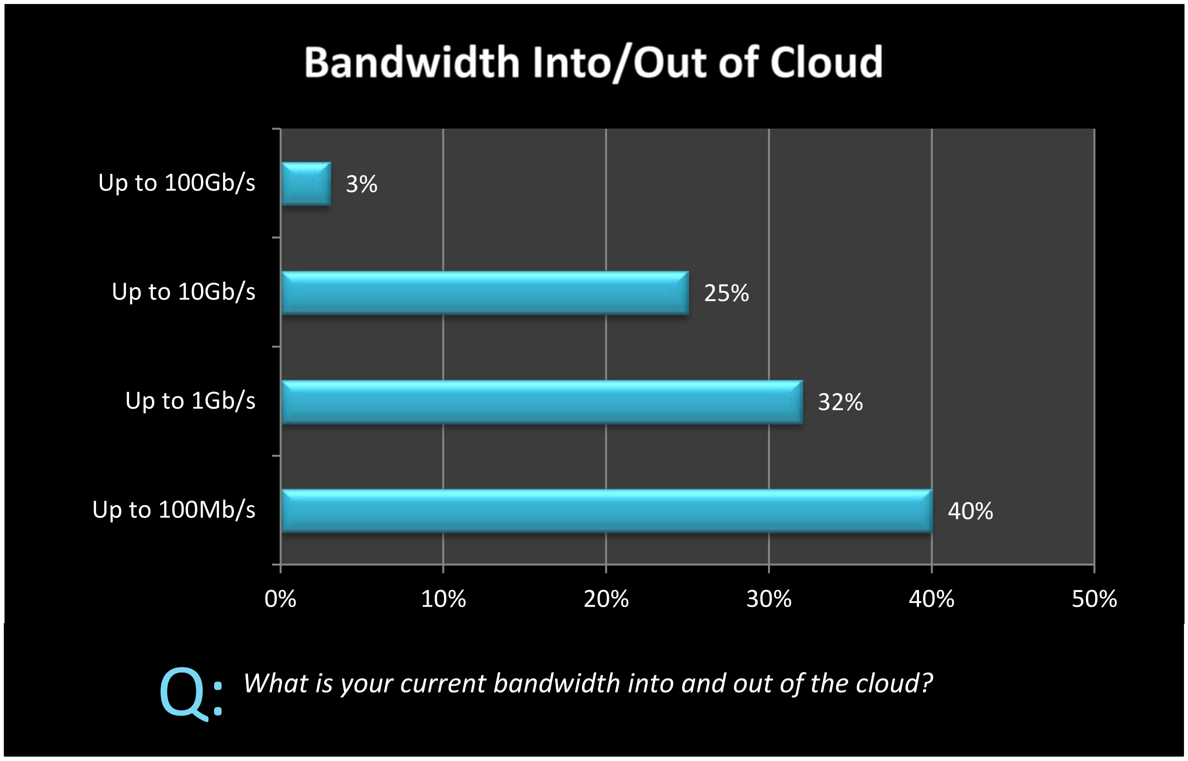
The median amount of data moved into the cloud was 100GB. The median moved out of the cloud was 13GB.

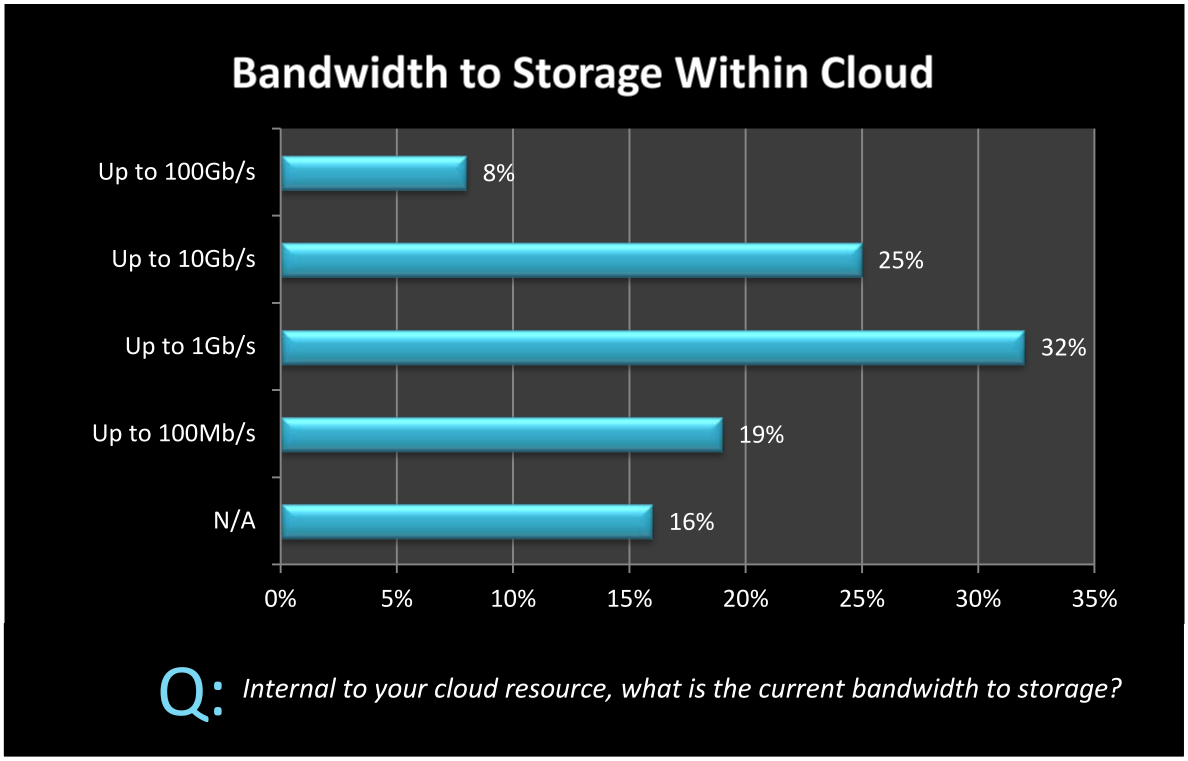




Bandwidth

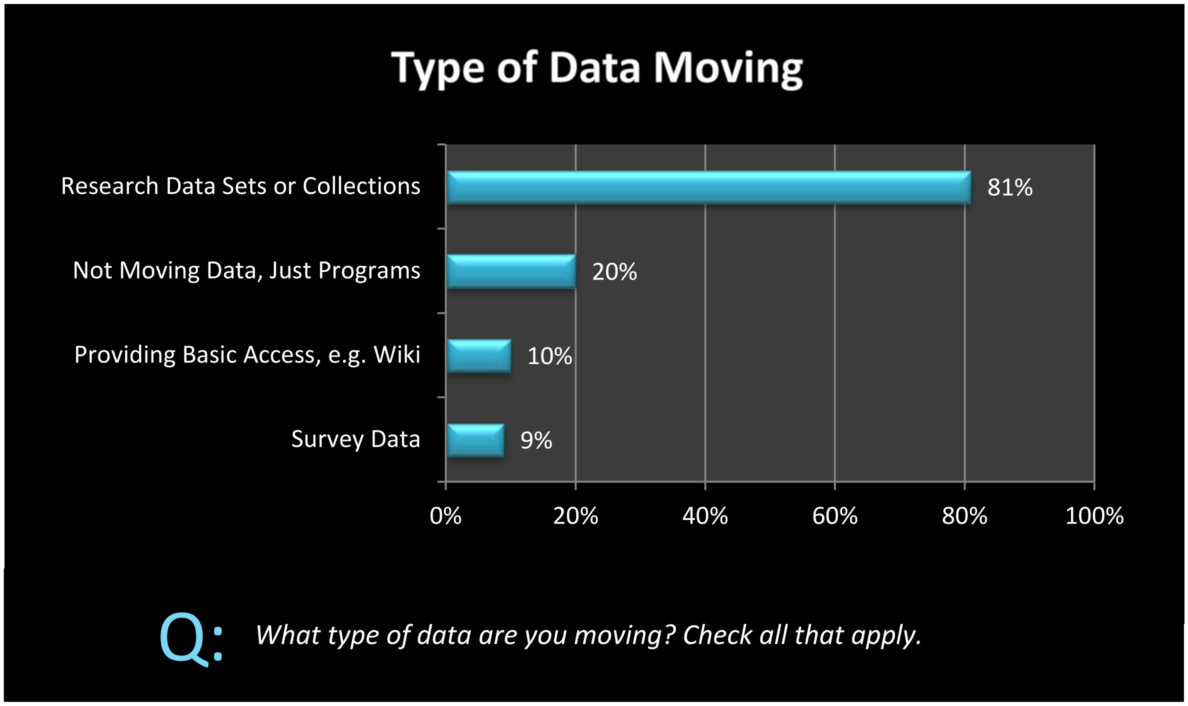
Bandwidth speed into/out of the cloud was slower than bandwidth to storage within the cloud.

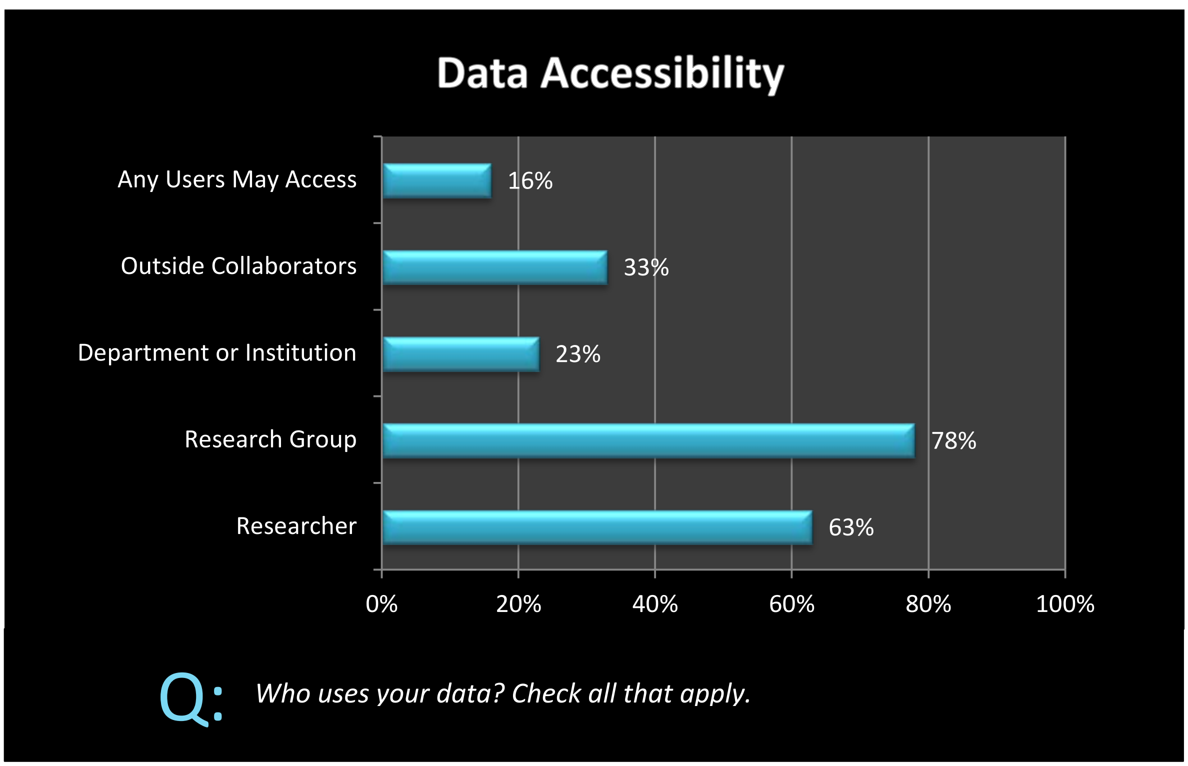




Type of Data Moving and Data Accessibility

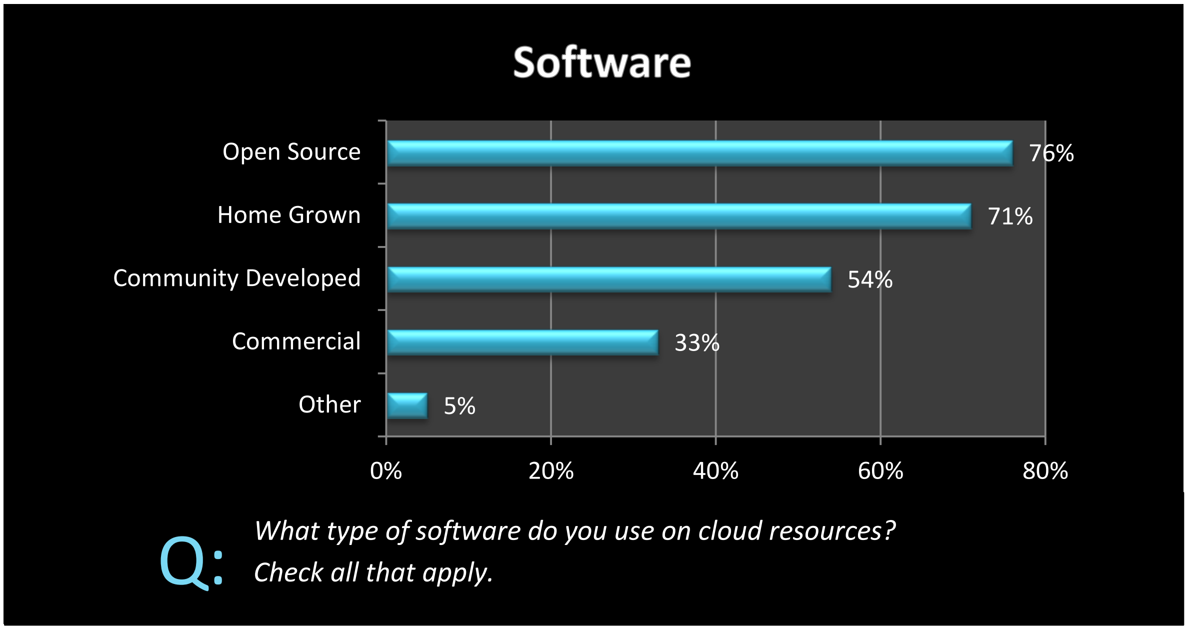
The vast majority of data being moved was research data sets or collections. 78% used the cloud to share data within their research group and 33% used the cloud to share data with outside collaborators.





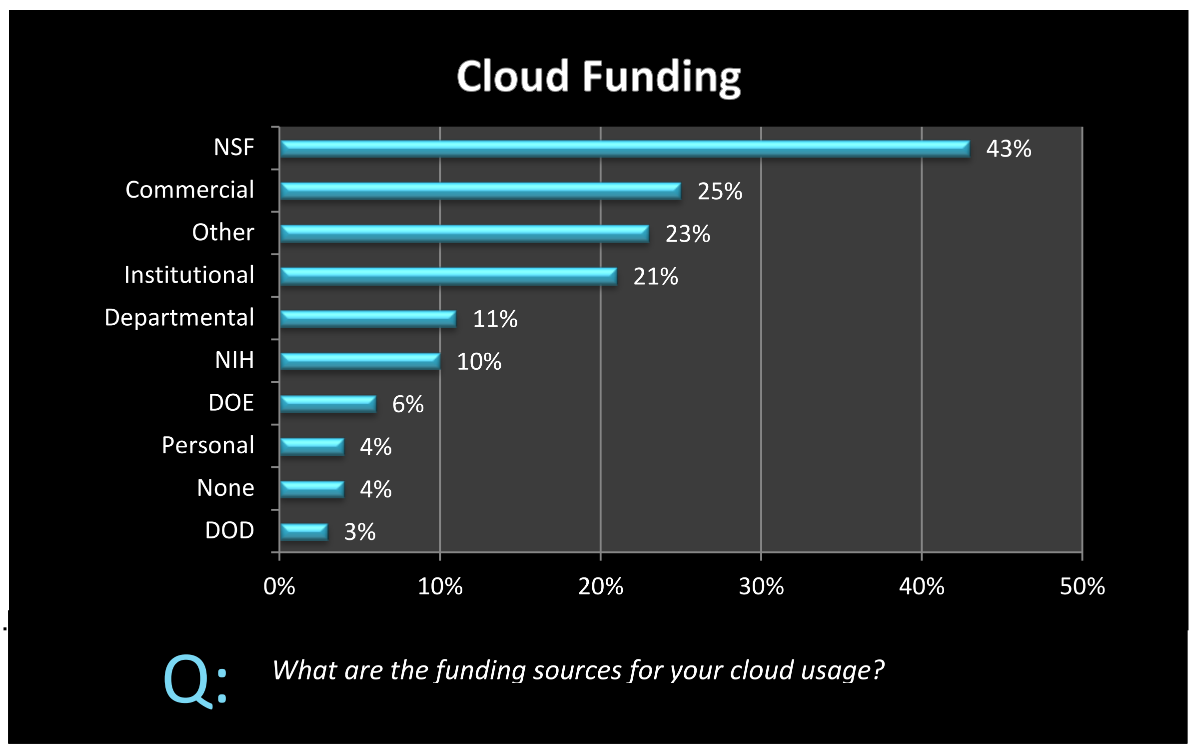
Software

While use of open-source and home-grown software dominated, 33% used commercial software in the cloud. Specific packages and tools identified included AMBER, CometCloud, CycleCloud, CycleServer, e-Science Central, GNU Wget, Illumina, LibSVM, MATLAB, MapReduce, MediaWiki, PostgreSQL, Rosetta, Redmine, Venus-C, and Window Azure SQL.



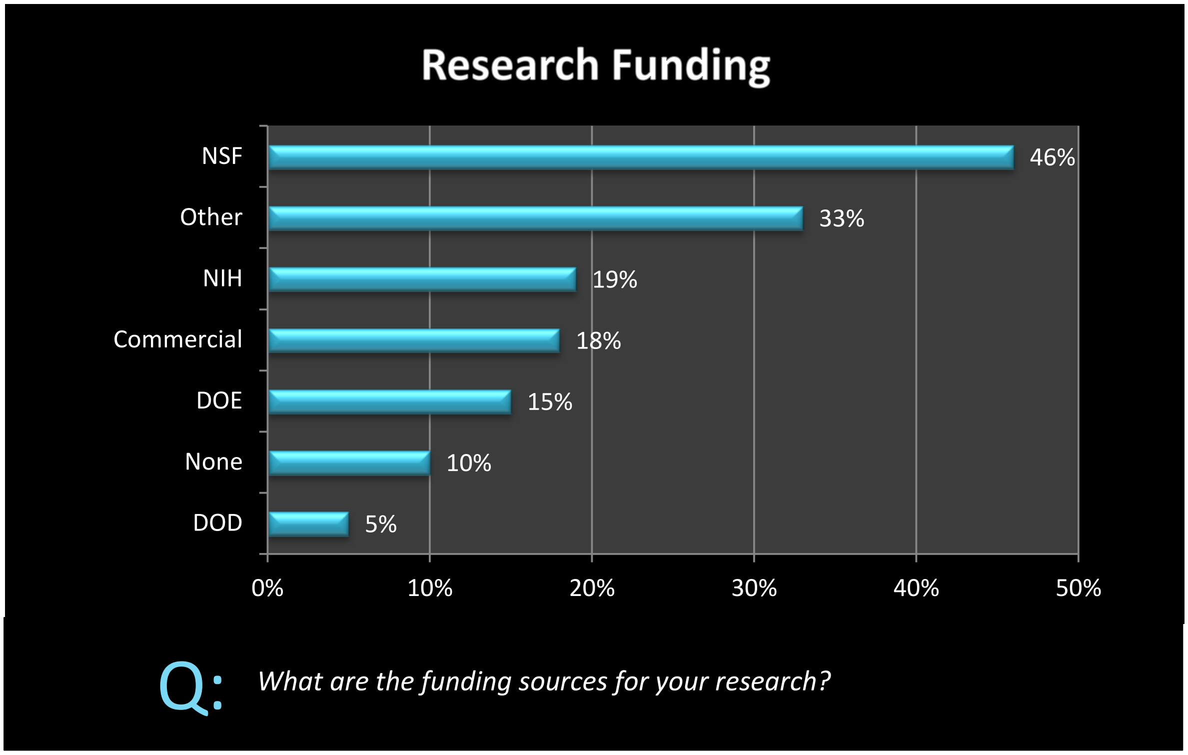
Cloud Funding

43% of the survey participants received cloud funding from the NSF. Commercial companies, mainly cloud service providers, provided free access to select researchers to try out their products, provide feedback, and/or collaborate on projects of interest to them. The "other" cloud funding category included the Alfred P. Sloan Foundation, ESA Science Mission, European Union, and cost recovery.



Research Funding

46% received research funding from the NSF, followed by “other,” commercial, and NIH funding. “Other” funding sources included DARPA, the ESA Science Mission, the European Union, the Gordon & Betty Moore Foundation, Microsoft Research, institutional, and personal funding.



Cloud Benefits Reported by Survey Participants

Benefit #1: Pay as You Go

"Pay as you go and elasticity are critical." – Architecture Services CTO

Pay as you go is a key feature cited by researchers and educators who are using the cloud to chart the galaxies, analyze tropical cyclone images, and educate undergraduate students in computational methods such as data management. Researchers place a high value on rapid access to computing and data analytics platforms. The ability to ramp resources up and down quickly also creates cost efficiencies for the lab, department, or institution.

"…cloud enabled the scientific community to access this genome resource quickly without researchers having to procure, deploy, and maintain their own data server." – Science Gateway Developer

"You only pay for what you use – when you're not using your 10,000 node Hadoop cluster, you don't pay for it." – Citizen Science Portal Developer

Benefit #2: Lower Costs

Cloud computing is a disruptive technology that has the potential to provide cost-effective alternatives to traditional research computing expenditures. Assuming an application is cloud-friendly, running in the cloud rather than deploying on-premise infrastructure can reduce capital expenditures. Use of the cloud may generate additional savings in recurring operation and maintenance costs, i.e., space, power, and cooling.

"Maintenance and administration cost savings are a plus for the cloud." – Systems Biologist

Building internal compute infrastructure for maximum load is costly, particularly for applications that tend to be cyclical.

"…our load CPU demand over a year isn't constant. There are peaks and there are troughs. If we priced our purchase to satisfy our peak needs, we'd find that our system would lay idle for some fraction of the year." – Particle Physicist

"There is no need to purchase an upfront data center for the 5-year mission, as it would be under-utilized most of the time." – Space Agency Operations Manager

Estimating how much a piece of hardware will actually be used, i.e., percent of utilization, and its associated costs (Total Cost of Ownership) vs. pay as you go cloud fees is an important consideration when deciding whether to buy on-premise hardware or to compute in the cloud.

If a decision to run in the cloud is made, standardizing compute resources used, tracking usage trends, planning batch workloads, and other capacity management strategies can optimize cost [13].

Use of the cloud may free up CI staff to focus on higher-order researcher needs such as data analysis, algorithm development, optimization, etc.

Benefit #3: Compute Elasticity

Compute elasticity, i.e., seamlessly adding compute on-demand, enables scientists and engineers to reduce run-times. This "bursting" capability can accelerate research productivity particularly for share-nothing, parallelizable applications and increase the potential for new insights and discoveries. When internal resources are maxed out, the cloud is an option for handling the overflow, e.g., ATLAS Google project [14]. Compute elasticity also better enables university and industry entrepreneurs to launch new companies by reducing initial capital expenditure requirements and subsequent R&D cycle times. The barrier to entry is much lower.

"Our 50,000-core compute ran across all 7 Amazon regions using on-demand and spot instances for a computational docking application….The experiment—the equivalent of 12.5 processor-years—was conducted in a mere 3 hours. Previously, it would take…about 11 days to run a similar analysis on its in-house 400-core cluster—stopping all other work in the process" – Software Developer

Cloud elasticity helps level the computational playing field for small labs, departments, and other resource-constrained organizations and individuals, enabling more risk taking and innovation.

"We calculated similarity scores for 8.6 trillion data pairs….and reduced our run-time processing for a job analyzing 3.8 million ScienceDirect articles from 100 days on our infrastructure down to just 5 days of processing time on AWS." – Data Mining Specialist

Innovative service models such as spot instances are an option for researchers who have time-flexible, interruption-tolerant tasks to compute at spot prices that are often significantly lower than on-demand prices [15].

Benefit #4: Data Elasticity

"The stochastic nature of our simulator requires simulating the same input multiple times, so with 'unlimited' cloud resources, researchers can gather and analyze larger amounts of data and investigate new sets of problems…" – President, Bioinformatic Research Consortium

IT directors and researchers alike are grappling with how to store, share, and protect large-scale data produced by simulations and experimental resources such as colliders, earthquake sensors, and gene data banks. Cloud-based science gateways, supported by providers such as Amazon Web Services, Globus Online, and SDSC Cloud Storage, are a viable alternative for providing communities of scientists access to vast amounts of data with readily-available analysis tools.

"We have an international audience, and we need our system to be reliable and available to all our users on a 24/7 basis. As our platform grows, we anticipate very large datasets to be contributed, so being able to scale quickly is important." – Supervisor, Energy Science Gateway

Researchers can scale large datasets with services such as Amazon Elastic MapReduce, SQL Database hosted on Windows Azure, MongoDB, or Google BigQuery without deploying on-premise Hadoop clusters or SQL servers. Large memory instances for database applications are available today and these options will likely grow based upon user demand. Domain-specific software/tool environments and workflows may be required to ensure the timely availability and analysis of Big Data projects that exceed the capacity of database management systems. Cloudera is among the companies developing Apache Hadoop distributions with analysis and management tools.

"The ability to instantiate clusters on demand with the software/environment specific to the analysis at hand enhances research productivity." – Shared Regional Data Center Researcher

Infrequently accessed data may be archived in the cloud. Advantages include geographic distribution in locations distinct from on-premise systems and lower cost due to massive economies of scale that cloud service providers offer. Hurricane Sandy motivated several academic institutions to consider adding cloud-based backup systems.

"Pay as you go and elasticity are critical. Services such as Amazon Glacier may mean we can leave data in the cloud rather than uploading it every 6 months." – Astrophysicist

Data download size and frequency need to be carefully considered in any data storage cost/benefit analysis.

Benefit #5: Software as a Service

Two benefits of Software as a Service (SaaS), e.g., MATLAB [16], R [17], cited by survey respondents were convenience and scalability. The ability to access software on-demand seamlessly from the desktop empowers researchers to experiment at a faster, more extensive scale while negating the need for server installation and software upgrades. Researchers can focus on the science rather than software availability and support. Analysts predict 50% of organizations will have a strategy for implementing Software as a Service by 2015 [18].

"…we run Parallel Computing Toolbox codes on an optimal number of cores in the Cloud rather than procure dedicated hardware/software for only periodic use… the Cloud provides the software we need when we need it, enabling us to develop simulation optimization and feasibility determination algorithms faster and more efficiently. " – Operations Research Engineer

"Science as a Service" providers are developing turnkey tools and software suites to make researchers more efficient. Integration of these capabilities across all levels of cyberinfrastructure will help build a more complete and collaborative ecosystem for research and education.

"…simulations often are too large to execute effectively on desktop workstations (requiring hours to days to weeks to complete), but can be completed in an interactive timeframe (minutes to hours) on Red Cloud with MATLAB. The results of these moderately complex simulations then often guide the construction of larger-scale simulations for which efficient parallelization and high-end computational resources are absolute necessities." – Neuropsychologist

Capabilities such as research data management may also be delivered to users as hosted Software as a Service. e.g., Globus Online uses the SaaS model via Amazon Web Services infrastructure to deliver a high-performance file transfer service [19].

Benefit #6: Education as a Service

The convergence of cloud and mobile services and devices will have a dramatic impact on what learning resources are accessed when, where, and by how many. Education as a Service can scale to dozens, hundreds, or even thousands of users, delivering interactive simulations and other learning experiences that encourage experimentation and discovery.

"We use cloud cyberinfrastructure to address successfully the dual issue of scalability (serving thousands of users at a fairly reasonable quality of service) and sustainability (providing accessibility and availability beyond the classroom)." – Teaching Tool Developer

Physical textbooks are beginning to be replaced by digital alternatives. The majority of university presidents predict within the next decade 50% of undergraduate textbooks will be digital [20]. The cloud may emerge as a platform of choice for professors who wish to collaboratively write online textbooks that feature cyberlearning tools and experiences that actively engage learners.

"I am assembling a collection of open-source tools to support further educational development: Calliope for optimization formulations, Octave for MATLAB-type programming and more." – Operations Research Professor developing online textbook

Cyberlearning use cases range from supporting classroom education to delivering asynchronous labs accessed anywhere there is an Internet connection.

"Hosting security lab exercises in the cloud brings us two main benefits…we can better prepare our students for their future careers in a cloud computing world…we can effectively address the resource

limitation of our existing lab environments and meanwhile ease the burden on our IT professional….28 students, one instructor, and one teaching assistant have amazingly used only $289 for four lab exercises in a semester, much less than the originally expected cost ($3600 budgeted). Through a survey answered by our students, we found that the majority of students are in favor of learning and using a leading cloud computing platform." – Computer Science Professor

Computer Science professors have been early adopters of the cloud for education. STEM fields will likely follow. Technically-oriented students like to embrace new technologies, particularly if those technologies are evolving rapidly and consistently deliver the latest applications, tools, and experiences. One university had over 120,000 students access a single class using the cloud [21].

Benefit #7: Broader Use

"Our cloud solution is primarily aimed at domain scientists who do not have advanced IT skills." – Chemistry Research Associate

Cloud appeals to a broad class of researchers, many of whom are not traditional HPC users. As the *NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging Report* noted, "computational performance alone is not an accurate indicator of computational utility [22]." By including cloud as part of a comprehensive cyberinfrastructure portfolio, more researchers and educators will be able to discover the value of advanced computation in stimulating discovery and innovation.

"The availability of platform services such as storage and programming abstractions such as .NET or MapReduce reduces the overhead of installing, monitoring and managing such services locally." – Energy Informatics Director

Platform as a Service and Software as a Service clouds offer features that mask computing complexities for less sophisticated users. IDC predicts that domain-specific, i.e., industry focused PaaS, will increase tenfold by 2016 [23].

Researchers who do not have a team of IT experts or capital budget available to rapidly architect, install, and run on-premise infrastructure at scale find the cloud particularly appealing and, at times, the only alternative. The sweet spot for clouds may be mid-scale CI, e.g., between NSF Major Research Instrumentation (MRI) and Major Research Equipment and Facility Construction (MREFC) grants. Clouds may decrease barriers to entry for small to midsize educational institutions that are not in the top tier of the research hierarchy [24].

Barriers to entry may be decreased for small to medium-sized businesses as well. The UberCloud Experiment is exploring the benefits and challenges of accessing the cloud for CAE and other simulation applications that given additional compute resources could speed up product design or improve product quality [25]. The Council on Competitiveness *Make: An American Manufacturing Movemen*t report notes that cloud computing has the potential to be a game-changing technology for manufacturing firms by providing agile services that are accessible regardless of company size or location [26].

Benefit #8: Scientific Workflows

High-throughput workflow applications such as the analysis of thousands of molecules or particle collisions are good candidates for the cloud. These applications can be divided into many independent tasks. The ability to ramp usage up and down for these types of applications is also appealing from a cost perspective.

Clouds promise to 'scale by credit card' ….Our projects utilized this new resource to execute scientific workflow applications in a fast and cost efficient way." – Computer Science Researcher

MapReduce is available from many cloud providers for high throughput computing and data analysis. High throughput applications such as BLAST as compared to MPI-based applications using, for example, partial differential equation solvers on an HPC machine, run efficiently in virtual machine environments.

"For highly performance driven applications that operate on a tightly coupled model, purchasing and managing a rack with ~50 cores is a better model than Cloud resources…..However, much of the research in our group deals with large scale problems rather than high performance problems. In such a scenario, on-demand access to a large number of virtual machines is more useful than round the clock availability of a captive cluster." – Associate Director, Energy Informatics

Several cloud service providers are developing or enhancing HPC cloud offerings to improve I/O, latency, and scalability issues that can be experienced with cloud-based HPC platforms using virtual machines. Other services, such as Penguin's On-Demand HPC Cloud Service (POD), offer access to tightly-coupled, non-virtualized compute cluster utilities over the network or "HPC on-demand" that feature typical HPC components such as low-latency interconnects.

Some researchers surveyed customized public cloud services with special features that their particular user community desired. The development of problem-specific workflows may be necessary in order to facilitate and stimulate cloud adoption within certain scientific domains.

"We have developed a python command line and web front end to Amazon EC2. This makes it very easy to run jobs on EC2 instead of local or remote clusters. The script handles all uploads and downloads and functions similar to how a queuing system works." – Chemistry Researcher

Computer and computational scientists are enhancing the cloud with capabilities derived from basic and applied research. For example, FutureGrid [27], which is part of XSEDE, is a robust, reproducible research test-bed ("Computer Testbed as a Service") with a cloud focus. Middleware and application users can customize bare-metal or VM/hypervisor cloud, grid, and/or parallel computing environments to investigate interoperability, functionality, performance or evaluation issues. The FutureGrid team has developed tools for dynamic provisioning and image management, virtual networks, monitoring, etc. and conduct and support educational workshops and other learning venues.

As cloud usage widens, discipline-specific R&D, e.g., custom interfaces, workflows, etc., will be essential in order to address the needs of a growing body of users who are not computational scientists.

Benefit #9: Rapid Prototyping

Cloud access enables small labs and departments without compute resources to try out new ideas and classes of problems without deploying hardware or competing for access to on-premise or national resources that may be saturated with priority projects. Clouds can provide research agility, i.e., the quick-testing ("fast-failing") of ideas and the ability to do the unexpected [28].

"From a cost and scalability point of view, we would definitely consider requesting funding for cloud resources. The cloud enables us to explore different classes of problems rapidly opening new doors to research." – Biological Systems Researcher

Relatively instant compute access means researchers with PC-only capabilities can take more risks, experimenting with new concepts without undue concern for compute availability or cost.

"We use the cloud for rapid prototyping. It is also affordable for constant use of small instances for things like MediaWiki and Redmine. Our use is generally data intensive and access to Red Cloud and GlusterFS avoids the data transfer dilemma." – IT Director, Biotechnology Core Facility

Benefit #10: Data Analysis

The researchers surveyed are leveraging the cloud not only for computation, but for data analysis. Motivations include low cost vs. the cost to procure and maintain on-premise database servers and associated data storage hardware.

"A steep drop in the cost of next-generation sequencing during recent years has made the technology affordable to the majority of researchers, but downstream bioinformatic analysis still poses a resource bottleneck for small laboratories.…We can enable researchers without access to local computing clusters to perform large-scale data analysis, by tapping into a pool of on-demand Cloud BioLinux VMs that can be rented at low cost.…Renting servers in the cloud can work as a better model for smaller research laboratories, where the cost for hardware and data center maintenance, cannot be justified to support only a few experiments." – Bioinformatics Engineer

Hadoop in the cloud is used by many researchers for data-centric applications such as digital pathology imaging analysis and the analysis of weather data, e.g., analyzing 300,000 satellite images of tropical cyclones [29]. Public datasets in the cloud, such as the NIH/AWS 1000 Genomes Project, make data more widely available and provide a framework for researchers to add tools to improve data usage [30].

Surprisingly, while data security is a chief concern of commercial enterprises, users in this survey did not express a similar concern. More often, the cost of data movement and scaling performance were objects of concern, particularly when using public clouds shared by a multitude of users with potentially conflicting usage patterns.

Cloud Challenges Reported by Survey Participants

Challenge #1: Learning Curve

Like any new technology, there is a learning curve with cloud although most survey respondents describe it as minor.

“The start-up, programming, and configuration are more challenging than an in-house local cluster; however…it isn’t difficult to learn.” – Biomechanics Researcher

Creating, deploying, and managing a cloud instance in an Infrastructure as a Service (IaaS) environment is a new experience for many researchers and, depending upon the application, can be time consulting. Even Platform as a Service environments designed to mask cloud complexities have a learning curve.

“The platform may provide the best platform for conducting our research but results are significantly delayed by initial development time.” – Science Gateway Developer

Cloud Wikis, how to documents, and online training can shorten the learning curve. With adequate investments in end-user training and consulting by federal agencies and academic CI facilities, researchers can focus on the science rather than the technology enabling it. Consulting support for research computing is not readily available from many public cloud service providers. Investments in user training may be necessary to facilitate the transition of academic communities to the cloud. The availability of pre-configured instances would be helpful as well.

Systems administrators and research computing consulting staff also need to be cloud savvy. HPC facility staff, for example, may not have the expertise to deploy a private cloud that bursts to public or national CI resources or to help a researcher build a virtual machine image. Federal agency investments in cloud training focused on the deployment of research and education applications may be required to accelerate adoption and overcome cultural barriers, i.e., resistance to service-based vs. deploying and operating on-premise systems.

A variety of cloud training classes/certifications are available for systems personnel, e.g., AWS Certified Solution Architect-Associate Level; Eucalyptus Design, Build and Manage (DBM) training classes; Google Apps Certified Deployment Specialist; Hanu Software’s Windows Azure IaaS Accelerator Workshops (supporting mixed platforms such as SQL Server 2012 or Linux); IBM Certified Solution Architect-Cloud Computing Infrastructure; Rackspace Training for OpenStack; and, VMware Certified Professional-Cloud (VCP-Cloud).

The National Institute of Standards and Technology (NIST) has defined cloud terminology and is facilitating and leading the development of cloud computing systems standards in areas where gaps exist, e.g., interoperability, portability, etc [31].

Challenge #2: Virtual Machine

A few cloud environments, e.g. Red Cloud, guarantee each virtual machine instance exclusive access to the CPU cores with which it is configured.

In most public cloud environments, however, CPU cores are shared by multiple instances which can hurt CPU performance. Cloud users can compensate by adding more virtual machines or by running longer. Competing for memory, disk, and network I/O in a shared cloud environment is not the same computing experience as running on a dedicated cluster. Some HPC workloads simply don’t scale well in virtual machine environments even with HPC instances; they need specialized hardware.

“The virtual machine nature of cloud tends to be detrimental to performance.” – Computational Chemist

While survey respondents found cloud management and identity tools such as Amazon's AWS Management Console convenient and easy to use, several said that they would like more control over compute instances and hardware layers to manage shared resources.

"It would be great if the compute instances could be managed in a more flexible and fine-grained manner." – Computer/Network Security Professor

Challenge #3: Bandwidth

Variability in network bandwidth can be an issue when transferring data from Local Area Networks to the cloud.

"Bandwidth in/out is an issue as is the cost model." – Citizen Science Portal Developer

As cloud use and Big Data projects increase, there is concern that bandwidth consumption will increase causing bottlenecks. According to IEEE, networks will need to support capacity requirements of 1 terabit per second in 2015 and 10 terabit per second by 2020 if current trends continue, i.e., simultaneous increases in users, access rates and services such as video on demand, social media, etc [32].

"…there is no doubt that in the next couple of years we'll see lots of nascent solutions to the fundamental problem of mobility and cloud collaboration: data movement. The data sets in our US-China project measured in the range from tens to hundreds of TBytes, but data expansion was modest at a couple of GBytes a day. For a medical cloud computing project, the data set was more modest at 35TBytes, but the data expansion of these data sets could be as high as 100GB per day, fueled by high volume instruments, such as MRI or NGS machines. In the US-China collaboration, the problem was network latency and packet loss, whereas in the medical cloud computing project, the problem was how to deal with multi-site high-volume data expansions." – Global Engineering Consultant

Cost-benefit analyses should take into account low-latency local network performance vs. higher-latency WAN connections. Future technologies may include fast, reliable Network as a Service (NaaS) or the ability to dynamically allocate network resources to computing resources, allowing both to scale or contract together, on demand.

Challenge #4: Memory Limits

Some scientists need higher memory instances for high-throughput, Big Data and memory-bound applications. For example, molecular biologists require very large memory for DNA sequencing problems such as de novo assembly of environmental microbial data.

"RAM limitations -- I need more than the maximum provided by Amazon (and most cloud providers). 300GB+ needed." – Molecular Genetics Researcher

Cloud service providers offer different types of instances but the ability to access bleeding edge resources is limited.

"The configurations are fixed so sometimes we waste memory or CPU." – Astrophysicist

On-premise hardware or dedicated hardware operated by a hosting provider may be necessary for applications that require customized configurations and/or the fastest-possible performance.

Challenge #5: Databases

Cloud service providers offer a variety of commercial and open source SQL and NoSQL databases which can run as virtual machine images or as a Database as a Service.

A few survey participants experienced unstable database performance in the cloud compared to the performance available from dedicated database servers.

"The cloud is less stable than a local server or HPC machines and may shut down unexpectedly because of upgrades or because some unmanaged exception in other processes, plus non-relational DBs require architecting and coding effort to ensure transactional operations in order to preserve consistency – your code may be shut down at any minute." – Biological Systems Researcher

Managed hosting services, e.g., Rackspace MySQL, can offer custom configurations that may include features such as redundant high performance storage and dedicated storage networks.

Challenge #6: Interoperability

Hybrid clouds, i.e., the combination of private on-premise or on-campus resources and external public cloud resources, have the potential to provide researchers and educators the flexibility to scale while protecting sensitive data and intellectual property. Few hybrid clouds are in production use; they are an emerging technology. Interest in hybrid and federated clouds, however, is very high particularly on the part of larger organizations [33].

"The time-critical nature and dynamic computational workloads of Value at Risk (VaR) applications make it essential for computing infrastructures to handle bursts in computing and storage resources needs….Integrating clouds with computing platforms and data centers, as well as developing and managing applications to utilize the platform remains a challenge." – Software Developer

The interoperability challenges of hybrid clouds include differences in platforms, tools, and APIs. Until these differences are overcome, seamless operation between private and public clouds will be a challenge. An alternative is using a cloud provider that offers physical colocation services in order to minimize interoperability issues.

Challenge #7: Security

While commercial enterprises have serious concerns about cloud security with regards to customer data and medical colleges and institutions have similar concerns about HIPAA data, in general, the lack of concern about data security in the cloud on the part of the survey participants was somewhat surprising. This may be due to the desire and, in most cases, the requirement on the part of academic researchers to share their data rather than protect it.

One cloud project hired an ethical hacker to compare the vulnerability of a set of applications running on an on-premise system vs. applications running in a public cloud environment. They concluded that most of the security issues in a public cloud are very similar to and no worse than the same security issues faced by on-premise systems.

"The issues (our ethical hacker) found were almost entirely challenges we would face and issues we would have had to protect against whether this was locally hosted, using our on-premise physical infrastructure, or remotely hosted at a public cloud provider….It is reasonable to assume further efforts may be needed if a higher level of isolation is demanded for specific confidential data. However, our results affirmed our belief that institutions such as our own can responsibly utilize cloud and public cloud providers." – Senior Fellow, Inter University Consortium for Political and Social Research

Because security is a chief concern of commercial companies, cloud service providers and managed hosting services are highly motivated to make continual improvements in security capabilities and offer security options such as data center access controls, firewall protection, data encryption, two-factor authentication, e.g., public key infrastructure (PKI), and audit tracking. Support for data encryption is also being built into software, e.g., Intel Distribution for Apache Hadoop software, with fine-grained access controls [34].

Regardless of the vigilant and critical focus on security on the part of cloud service providers, some researchers and organizations remain uncomfortable with the idea that their data is not at the same physical location that they are. They fear the possibility of unauthorized physical access to their data or machine and other security issues such as WAN vulnerability.

Most academic researchers and educators, however, seem to support an open collaboration model using community clouds that partition and share data based on user need and data owner requirements.

Challenge #8: Data Movement

Most cloud service providers charge by the GB for data movement out of the cloud; therefore, data movement costs for public clouds are an important factor to consider when deciding whether to build an on-premise cloud or use a commercial cloud service provider.

"Most of our collaborators have the following view of cloud resources: clouds are excellent at providing burst capacity and custom software environments for computation and data analytics….On the storage side, they are very concerned about the high cost of long term storage, and the risk of data loss or extreme cost retrieval. They are much more comfortable keeping their data at the local campus, where they can control and access it on demand." – Software Researcher and Designer

If a workflow can live in the cloud or data movement out of the cloud can be minimized, data movement costs can be reasonable depending on the regularity and the amount of data moved. Some HPC applications, however, can generate very large data sets. In addition, data movement is only as good as the network bandwidth enabling it.

Challenge #9: Storage

Supercomputing-class file systems are not readily available in most cloud environments.

"*I wish EBS volumes would work more like Lustre file systems, i.e., high performance, high availability, and the ability for multiple VMs to read/write to one EBS volume." – Bioinformatics Researcher*

Confidence in the security of data stored in the cloud is a greater concern when collaborating with industry.

"Due to our application requirements, we'd like the cloud to provide secure data store and allow users to customize and copy their VM instances." – Academic/Industry Research Collaboration

Challenge #10: Cost/Funding

"Opacity of cost is a problem. We were occasionally surprised by how much we were spending on certain resources." – Electrical Engineering/Computer Science Postdoc

Some survey participants expressed surprise at the cost of commercial cloud usage when they received their bill at the end of the month, in particular, the cost of moving data.

A few researchers also expressed concern about justifying the use of the cloud on their grants.

"Overall cost and charging to a grant that does not have such a cost model built in are challenges." – Shared Regional Data Center Researcher

"We find it difficult to write cloud compute resources into our grants." – Citizen Science Director

Grant reviewers may need clarity on the value and appropriateness of cloud as a potential tool to enhance research productivity and lower overall grant costs. Since clouds are evolving rapidly, it is important to consider the latest cloud technologies when assessing whether they can meet the needs of a proposed project.

The cloud is still in its infancy in many research and education communities. Amazon, Google, Microsoft and others have graciously donated cloud time and assistance to select researchers and educators in order to bridge the chasm between cloud innovators and early adopters. This support is essential in encouraging mainstream use by the research computing and education community and will hopefully continue.

Federal agency support is vital as well. NSF, NIH, DOE, etc. have made strategic investments to support the cloud innovators and early users. Further investments in cloud access, cyberinfrastructure, training, and research will be needed to widen the use of clouds.

Science & Engineering Cloud Projects Surveyed: Complete Data

Astronomy: Citizen Science

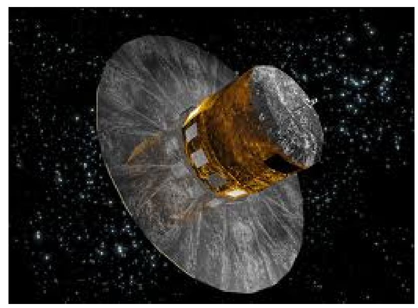
Project	Zooniverse
Cloud Use Cases	Burst resources; collaboration; computing and data analysis support for scientific workflows; data sharing; education, outreach, and training (EOT)
Primary Researchers	Arfon Smith, Adler Planetarium; Chris Lintott, University of Oxford; Lucy Forston, University of Minnesota
Abstract	The Zooniverse is home to the internet's largest, most popular and most successful citizen science projects [35]. The Zooniverse and the suite of projects it contains is produced, maintained and developed by the Citizen Science Alliance. The member institutions of the CSA work with many academic and other partners around the world to produce projects that use the efforts and ability of volunteers to help scientists and researchers deal with the flood of data that confronts them. The Zooniverse has nearly a dozen websites on space, climate, humanities, and nature, including Galaxy Zoo, The Milky Way Project, Solar Stormwatch, CycloneCenter, Ancient Lives, and Planet Hunters. The Milky Way Projects asks users to analyze data from the Spitzer Space Telescope. CycloneCenter asks users to analyze the intensities of tropical cyclones from nearly 300,000 satellite images. Planet Hunter volunteers search for planets. In 2012, they discovered PH1, the first-ever planet with four suns some 5,000 light years away [36].
Cloud Providers	Amazon Web Services
Special Features	Community datasets or collections; MapReduce; tables
Use Regularity	Daily
Cores Used Peak	1000
Cores Steady State	30
Core Hours in a Year	700000
Access Storage For	Analysis; reference; archival
Preferred Storage	Object store
Accessed During Run	100TB
Short-Term Storage	100TB
Long-Term Storage	10TB
Data Moved Into Cloud	10TB
Data Moved Out Cloud	10TB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections; survey data
Data Accessed By	Researcher; research group; outside collaborators; any users may access the data collection or survey results
Software	Home-grown; community developed; open source
Capabilities/Features	"Cloud computing platforms such as Amazon Web Services provide a level of service and reliability unlike any academic service I have encountered."
Problems/Limitations	"Bandwidth in/out is an issue as is the cost model. We find it difficult to write cloud compute resources into our grants."
Additional Notes	"Elastic MapReduce is a web service built on top of the Amazon cloud platform. Using EC2 for compute and S3 for storage, it allows you to easily provision a Hadoop cluster without having to worry about set-up and configuration. Data to be processed is pulled down from S3 and processed by an auto-configured Hadoop cluster running on EC2. Like all of the Amazon Web Services, you only pay for what you use – when

More than 850,000 citizen scientists have participated in Zooniverse

	you’re not using your 10,000 node Hadoop cluster, you don’t pay for it [37].”
Cloud Funding	NSF; Sloan Foundation
Research Funding	NSF

Astronomy: Galaxy Charting

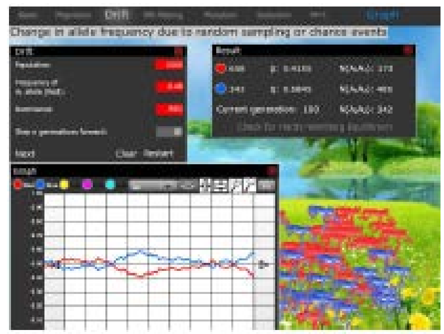
Project	Gaia astrometric global iterative solution in the cloud
Use Cases	Burst resources; commonly requested software; computer science research; computing and data analysis support for scientific workflows; domain-specific computing environments
Primary Researchers	Paul Parsons, The Server Labs; William O’Mullane, ESA
Abstract	The Astrometric Global Iterative Solution (AGIS) will process all the observations produced by the satellite (1 billion stars x 80 observations x 10 readouts). This requires a tremendous amount of data processing. As an example of the magnitude of this project: if it took one millisecond to process one image, it would take 30 years of data processing time on a single processor. Thus, the ESA Gaia Team developed their own grid/distributed computing system based on data processing trains [38]. The fact that the processing for AGIS is not continuous made it an ideal candidate for the cloud. Every 6 months we need to process all the observations in as short a time as possible (typically two weeks) so the cloud is the perfect solution [39].
Cloud Providers	Amazon Web Services; CloudSigma
Use Regularity	Annually
Cores Used Peak	800
Cores Steady State	1
Core Hours in a Year	580000
Storage Accessed For	Analysis
Preferred Storage	RDBMS, S3
Accessed During Run	500GB
Short-Term Storage	500GB
Long-Term Storage	0
Data Moved Into Cloud	500GB
Data Moved Out Cloud	50GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Department or institution
Software	Home-grown; community developed; open source; commercial
Capabilities/Features	“Pay as you go and Elasticity are critical. Services such as Amazon Glacier may mean we can leave the data in the cloud rather than uploading it every 6 months.”
Problems/Limitations	“In Amazon, the configurations are fixed so we sometimes waste memory or CPU.”
Additional Notes	“Bursty load profile EC2 based solution is cheaper, 350k vs. 750K euro in-house….there is no need to purchase an upfront data center for the 5 year mission, as it would be under-utilized most of the time. Ability to quickly launch and shutdown the application on demand. Ability to scale up or down on the size of the data set [40].”
Cloud Funding	ESA Science Mission
Research Funding	ESA Science Mission



The goal of Gaia is to chart one billion stars

Biology: Cloud-Enabled Learning Tools

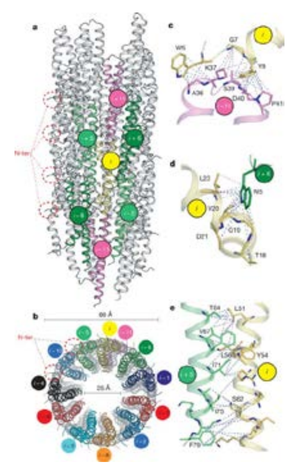
Project	CI-TEAM: A cloud-enabled evolutionary genetics learning tool for engaging the NET-savvy generation
Cloud Use Cases	Computing and data analysis support for scientific workflows; education, outreach, and training (EOT); science gateways
Primary Researcher	Bina Ramamurthy, University of Buffalo
Additional Researchers	Jessica Poulin and Katharina Dittmar, University of Buffalo
Abstract	“To help reduce the number of dropouts in freshman biology courses, professors at the University of Buffalo have turned to the power of collaboration and cloud computing to build an online teaching tool designed to explain concepts better than a textbook can. The tool [41] provides a visual way to map evolution. Cloud computing allows for different levels of network resources to be devoted to Pop!World based on the number of students using it [42].”
Cloud Providers	Amazon Web Services; Google Cloud Platform
Special Features	Adobe Flash
Dev. Environment	VMware
Use Regularity	Monthly
Cores Used Peak	100
Cores Steady State	100
Core Hours in a Year	200
Storage Accessed For	Reference
Preferred Storage	Object store
Accessed During Run	1TB
Short-Term Storage	1TB
Long-Term Storage	1TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Providing basic access
Data Accessed By	Department or institution
Software	Commercial; Adobe Flash
Additional Notes	“The project called *Pop!World* features three major levels: (i) the *Gateway* module for catering to K-12 students, (ii) the *Discovery* module for undergraduates, and (iii) the *Research* module for advanced learners and researchers. The *Discovery* module of *Pop!World* is currently in use in the introductory Biological Science course at UB (BIO 200). The project that began as a design of a prototype tool for learning and teaching soon faced two major issues: scalability and sustainability. Scalability in our case is about the ability to service thousands of users at a fairly reasonable quality of service. Sustainability is about accessibility and availability beyond the classroom. Learners are often introduced to useful tools and environments during their enrollment in a course. Yet, continued access to the tools beyond the duration of the course is critical for sustaining the learning that happened during the course and to enable experimentation, discovery and application of the knowledge they acquired. Therefore, we used cloud CI to address successfully the dual issues of scalability and sustainability [43].”
Cloud Funding	NSF
Research Funding	NSF



Scalability and sustainability of cloud enables engaging evolutionary biology learning tool

Biology: Macromolecular Modeling

Project	Atomic model of type III secretion system needle
Use Cases	Collaboration; commonly requested software; data sharing
Primary Researchers	Nikolaos Sgourakis and David Baker, University of Washington
Abstract	The ability of Gram-negative bacteria, such as the agents of plague, dysentery and typhoid fever to infect host cells is dependent on a syringe-like molecular machine known as the Type-III secretion system (T3SS). The core of T3SS consists of a hollow filament, the needle; composed of identical, symmetric repeats of an 80-residue protein, the needle forms a conduit for unfolded effector proteins to be delivered to the cytoplasm of the host cell at the early stages of infection. Determination of the three-dimensional structure of the needle by X-ray crystallography or solution NMR has been challenging thus far due to the inherent non-crystallinity and insolubility of the complex. Modeling based on docking of the known monomeric structure into EM reconstructions of isolated needle particles has been limited by the inability of such approaches to capture conformational change as a result of tertiary interactions. We have developed an alternative, hybrid approach through a combination of solid-state NMR data collected in the group of Prof. Adam Lange at the Max Planck Institute, previously published EM data and Rosetta modeling to determine a high-resolution model of in vitro reconstructed needle filaments. We show that the 80-residue subunits form a right-handed helical assembly with roughly 11 subunits per two turns of a 24A-pitch helix. While the more conserved C-terminus is forming key stabilizing towards the inside of the 25A needle pore, the more sequence variant N-terminus is positioned on the surface of the structure. The approach developed here presents a powerful way towards structure determination of large protein assemblies [44].
Cloud Providers	Windows Azure
Use Regularity	Weekly
Cores Used Peak	2000
Cores Steady State	2000
Core Hours in a Year	500000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	N/A
Accessed During Run	1TB
Short-Term Storage	1TB
Long-Term Storage	1TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 100Mb/sec
BW to Storage Within	N/A
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; outside collaborators
Software	Community developed; Rosetta
Additional Notes	"Sgourakis notes that in order to conduct this type of research before it would have taken an incredibly powerful system or would have required thousands of shared hours as a volunteer computing project. The researchers at Baker have already made use of a number of grid computing tools like Rosetta@Home, Foldit and others, but Sgourakis says that their time to solutions are happening far faster by tapping into the cloud [45]."
Cloud Funding	Microsoft Research
Research Funding	NIH



Complete atomic model of T3SS needle

Biology: Biotechnology Core Facility Support

Project	Support for Biotechnology Resource Center (BRC)
Cloud Use Cases	Burst resources; commonly requested software; computing and data analysis support for scientific workflows; data archiving; data management and analysis; data sharing; domain-specific computing environments
Primary Researchers	Jocelyn Rose, Jason Mezey, Adam Siepel and Haiyan Yu, Cornell University
Abstract	BRC provides an array of shared research resources and services to the Cornell University community and to outside investigators. The Center has seven biotechnology core laboratories, including genomics (DNA sequencing, genotyping, and microarrays), epigenomics, proteomics and mass spectrometry, microscopy and imaging, bio-IT, bioinformatics and computational biology, and advanced technology assessment. We use Red Cloud and associated GlusterFS storage at the Cornell Center for Advanced Computing to deliver storage and services such as fast file transfer (Globus Online), data archiving, and support software (MediaWiki, Redmine, etc.) to meet our customer's needs.
Cloud Providers	Amazon Web Services, Globus Online, Red Cloud
Dev. Environment	Eucalyptus
Use Regularity	Daily
Cores Used Peak	10
Cores Steady State	5
Core Hours in a Year	50000
Storage Accessed For	Analysis; archival
Preferred Storage	GlusterFS/NAS
Accessed During Run	0
Short-Term Storage	30TB
Long-Term Storage	180TB
Data Moved Into Cloud	180TB
Data Moved Out Cloud	50TB
BW In/Out of Cloud	10Gb/s
BW to Storage Within	10Gb/s
Type Data Moving	Providing basic access; research data sets or collections
Data Accessed By	Researcher; research group; department or institution
Software	Community developed; open source; commercial; Illumina Pipeline; genome analysis (open source and home-grown); Globus Online; Redmine; MediaWiki
Capabilities/Features	"Rapid prototyping. Affordable for constant use of small instances for things like MediaWiki and Redmine. Our use is generally data intensive and internal campus access to Red Cloud and GlusterFS avoids the data transfer dilemma."
Problems/Limitations	"Looking forward to the features in the new version of Eucalyptus (closer to AWS)."
Cloud Funding	Cost recovery, i.e., core facilities recover costs from researchers with a variety of public and private funding
Research Funding	Cost recovery

Core facilities can improve and expand their research services by leveraging the cloud

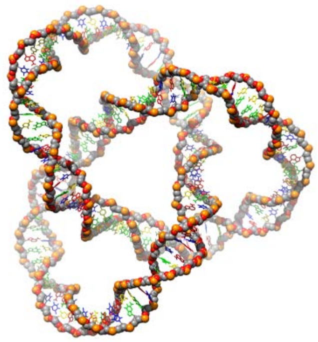
Biology: Computational Systems Biology

Project	VENUS-C – systems biology
Use Cases	Science gateways
Primary Researchers	Corrado Priami, COSBI
Abstract	COSBI's main goal in the Venus-C project was porting and deploying, over the Cloud infrastructure services, a dry experiment simulator for simulating and analyzing the dynamics of in-silico models of complex biological systems. These tools are of interest to the vast community of academic labs and companies doing research in medicine, biology and pharmacology [46].
Cloud Providers	Windows Azure
Special Features	Tables
Use Regularity	Monthly
Cores Used Peak	28
Cores Steady State	25
Core Hours in a Year	18000
Storage Accessed For	Analysis
Preferred Storage	Object store
Accessed During Run	2GB
Short-Term Storage	300GB
Long-Term Storage	300GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Department or institution
Software	Home-grown
Cloud Funding	European Union; institutional
Research Funding	European Union; public and private institutions
Additional Notes	“Database storage on the cloud is different from the one we are used to: databases are not relational and querying and paging is available on a limited number of fields….The cloud is less stable than a local server or HPC machines and may shut down unexpectedly because of upgrades or because of some unmanaged exception in other processes, plus non-relational DBs require architecting and coding effort to ensure transactional operations in order to preserve consistency – your code may be shut down at any minute….Scalability is a plus for Azure….no additional maintenance and administration costs are a plus for the cloud; often nodes’ software needs to be synchronized and local storage needs to be cleaned up due to dirty jobs’ trash….A theoretically infinite number of machines enable us to approach different scientific problems that require the simulation of large numbers of very similar input models. The stochastic nature of our simulator requires simulating the same input multiple times, so with ‘unlimited’ cloud resources, researchers can gather and analyze larger amounts of data and investigate new sets of problems that, with the usage of an HPC, were not possible. From a cost and scalability point of view, we would definitely consider requesting funding for cloud resources. Also, the cloud enables us to explore different classes of problems opening new doors to research [47].

Biochemistry: Molecular Dynamics Acceleration/MD-as-a-Service

Project	Collaborative Research SI2-SSE: Sustained innovation in acceleration of Molecular Dynamics on future computational environments
Cloud Use Cases	Commonly requested software; computing and data analysis support for scientific workflows; education, outreach, and training (EOT)
Primary Researchers	Ross Walker, University of California, San Diego
Additional Researchers	Adrian Roitberg, University of Florida
Abstract	Extending our 1-year pilot project funded by the 2010 SI2-SSE program we propose a continued collaborative project between the San Diego Supercomputer Center, the University of Florida and industrial partners at NVIDIA to continue development of an innovative, comprehensive, open-source software element library for the acceleration of all major computationally intensive aspects of condensed phase Molecular Dynamics (MD) simulations. We will extend our work to all hardware classes including Workstations, Supercomputers and Cloud resources. We have added as new industrial partners Intel and Amazon. Specifically we plan to extend our comprehensive GPU accelerated dynamics engine, developed as part of our SSE pilot, to support next generation accelerator technologies including Intel's Many Integrated Core (MIC) and future GPU technologies. We will extend the feature support to include all major MD techniques, develop accelerated analysis libraries and create open-source libraries of the software elements we develop. A priority will be enhanced sampling techniques including Thermodynamic Integration, constant pH algorithms, Multi-Dimensional Hamiltonian Replica Exchange and Metadynamics. We will further extend our affiliation with the SSI funded group of Prof. Todd Martinez developing direct connections between the elements we develop for accelerated MD and those they develop for accelerated Quantum Chemistry resulting in a sustainable community software framework. Finally, we will use these elements, in collaboration with Amazon to support MD as-a-service through easily accessible web front ends to cloud services, including Amazon's EC2 GPU hardware. Outreach and education will be provided through online and interactive workshops on MD simulation. Intellectual Merit: The work we propose is novel and timely. There is a clear need to develop software elements that are portable, scalable, fast, and accessible to all. Last year we were awarded a 1-year pilot project to begin our work. After great initial success, we propose to continue and expand our work attacking what is increasingly looking like a staple of future computing: the use of accelerator technologies, including GPUs [48] & MIC, in workstations, cloud resources and supercomputers. By comprehensively porting advanced MD techniques to such technologies and providing transparent portals to accelerated cloud resources such as Amazon's EC2 service, Microsoft Azure and/or Google AppEngine this project will enable users to: obtain substantial performance increases in their own local calculations; access elastically scalable on-demand cloud services and make effective use of accelerator technologies being deployed in NSF supercomputers. We will develop these software elements in close collaboration with NVIDIA, Intel and Amazon…Broad Impact:…With over 8,000 downloads of the latest AMBER Tools package from unique IPs and >500 sites using the AMBER MD engines it is clear that this work will benefit large communities of researchers. Additionally the libraries we release enabling the use of accelerators for all aspects of the MD workflow will be simple to implement in other packages providing both national and international impact across multiple domains. The development of a simple web-based front end for

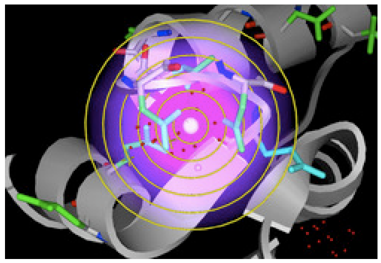
	use of elastically scalable cloud resources will also make simulations routine for all researchers. Our education and outreach efforts will train the next generation of scientists not just in how to use our MD acceleration libraries and advanced MD simulation but also get them thinking about how their approach can be transformed by the fact that performance that was previously restricted to large scale supercomputers is now available on individual desktops…[49].
Cloud Providers	Amazon Web Services; Windows Azure
Special Features	GPUs
Use Regularity	Daily
Cores Used Peak	1
Cores Steady State	1
Core Hours in a Year	1
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	500GB
Short-Term Storage	500GB
Long-Term Storage	0
Data Moved Into Cloud	1GB
Data Moved Out Cloud	100GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; outside collaborators
Software	Community developed; AMBER
Capabilities/Features	“We have developed a python command line and web front end to Amazon EC2. This makes it very easy to run jobs on EC2 instead of local or remote clusters. The script handles all upload and downloads and functions similar to how a queuing system works. We are also working on interactive analysis that allows deployment of calculations directly on cloud back end, i.e., to automate exploration of key areas of an energy surface.
Problems/Limitations	“There is no easy way to obtain time on cloud resources without spending real money. Currently it is much more beneficial to save real money for real people which means it is difficult to rationalize the use of cloud resources in academia. The virtual machine nature of cloud tends to be detrimental to performance.”
Cloud Funding	Gifts from Amazon and Microsoft
Research Funding	NSF



MD-as-a-service requires web front ends to the cloud

Biochemistry: Protein Research Acceleration

Project	FEATURE machine learning project accelerates protein research
Cloud Use Cases	Computer science research; computing and data analysis support for scientific workflows
Primary Researchers	Russ Altman, Stanford University; Dragutin Petkovic, San Francisco State University
Additional Researchers	Ljubomir Buturovic and Mike Wong, San Francisco State University
Abstract	FEATURE uses machine learning to predict functional sites in proteins and other three-dimensional (3D) molecular structures. Massively parallel optimization of machine learning involves the application of support vector machine (SVM) algorithms to thousands of training sets that are composed of hundreds of thousands of vectors. Optimal SVM parameters are found through brute-force parallelized grid searches with k-fold cross-validation. This optimization involves repeating similar operations many times independently [50].
Cloud Providers	Amazon Web Services
Special Features	Community datasets or collections; account management
Dev. Environment	MIT StarCluster [51]
Use Regularity	Monthly
Cores Used Peak	64
Cores Steady State	240
Core Hours in a Year	13189
Storage Accessed For	Analysis
Preferred Storage	Elastic Block Storage
Accessed During Run	1TB
Short-Term Storage	800GB
Long-Term Storage	600GB
Data Moved Into Cloud	474GB
Data Moved Out Cloud	1.116TB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Home grown; open source; libSVM
Capabilities/Features	“AWS offers a simple web-based UI, rapidly growing software features, and excellent support.”
Cloud Funding	Commercial
Research Funding	NIH



Students, researchers, educators use FEATURE software for protein functional classification

Biochemistry: Replica Exchange

Project	Asynchronous replica exchange molecular dynamics
Cloud Use Cases	Burst resources; computing and data analysis support for scientific workflows; data archiving; data sharing; domain-specific computing environments; event-driven real-time science
Primary Researchers	Manish Parashar, Moustafa AbdelBaky, Ivan Rodero, Aditya Devarakonda, Emilio Gallichio, and Ronald Levy, Rutgers University; and, Brian Claus, Rutgers University and Bristol-Myers Squibb
Abstract	Replica exchange is a powerful sampling algorithm that preserves canonical distributions and allows for efficient crossing of high-energy barriers that separate thermodynamically stable states. The replica exchange algorithm has several advantages over formulations based on constant temperature, and has the potential for significantly impacting the fields of structural biology and drug design. While these replica exchange simulations can definitely benefit from the potentially large numbers of processors available in clouds, general formulations of the replica exchange algorithm require complex coordination and communication patterns. We developed and validated an asynchronous replica exchange engine built on top of CometCloud and extended it to provide the abstractions and mechanisms required by asynchronous replica exchange, including mechanisms for dynamic and anonymous task distribution, task coordination and execution, decoupled communication and data exchange. It provides a virtual shared space abstraction that can be associatively accessed by all walkers without knowledge of the physical locations of the hosts over which the space is distributed [52].
Cloud Providers	Amazon Web Services; FutureGrid
Use Regularity	Weekly
Cores Used Peak	4000
Cores Steady State	256
Core Hours in a Year	100000
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	1GB
Short-Term Storage	1GB
Long-Term Storage	1GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	CometCloud
Cloud Funding	NSF; departmental; institutional
Research Funding	NSF

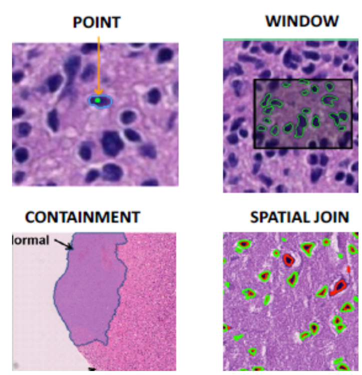
Biochemistry: Protein Data Bank Structure Mining

Project	Protein Data Bank
Cloud Use Cases	Burst resources; collaboration; commonly requested software; data archiving; data sharing; domain-specific computing environments; education, outreach, and training (EOT); event-driven real-time science; science gateways
Primary Researchers	Moustafa AbdelBaky, Rutgers University; Hyunjoo Kim, Xerox Research Center, Webster; Ivan Rodero; Rutgers University; Manish Parashar; Rutgers University
Abstract	Protein-ligand binding is the notion that a small molecule (a drug, a.k.a. the ligand) binds to a receptor or protein in the body. This binding event evokes a biological response, possibly the reduction of inflammation, pain relief, etc. Typically, there are a limited number of poses or configurations that this protein-ligand complex can assume. Identifying this bioactive pose is a tremendous challenge in drug discovery. There are many ways to generate these poses, as well as many ways to try to determine which ones are (or may be) correct. Some of these calculations are computationally inexpensive, while others may be extraordinarily expensive. One approach to this problem is to generate a large number of potential poses using a fairly inexpensive method and follow that up with a more expensive calculation to rank them in order of likelihood of being the bioactive pose. Another approach uses the Protein Data Bank; the Protein Data Bank (PDB) is a database of known crystal structures and Nuclear Magnetic Resonance (NMR) structures many of which are protein-ligand complexes. By mining the information contained in these structures, we are generating a scoring function based on known protein-ligand interactions [53]. We used the CometCloud framework to develop a protein data mining application operating on data from the Protein Data Bank. The application is deployed on a cluster at Rutgers University and/or Amazon EC2 based on deadline and budget constraints. The experimental results show that the MapReduce-CometCloud framework can effectively support applications operating on large numbers of small data files on a heterogeneous and distributed environment, and satisfy user objective autonomously using cloudbursts [54].
Cloud Providers	Amazon Web Services
Dev. Environment	CometCloud
Use Regularity	Monthly
Cores Used Peak	800
Cores Steady State	100
Core Hours in a Year	2000
Storage Accessed For	Analysis; reference
Preferred Storage	Object store
Accessed During Run	2GB
Short-Term Storage	2GB
Long-Term Storage	10GB
Data Moved Into Cloud	20GB
Data Moved Out Cloud	2GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Any users may access data collections and survey results
Software	Home-grown; community developed; open source
Cloud Funding	NSF; DOE; commercial; departmental
Research Funding	NSF; DOE; commercial

Accelerated mining of PDB data using MapReduce-CometCloud

Biomedical Imaging Informatics: Digital Pathology Imaging Analysis

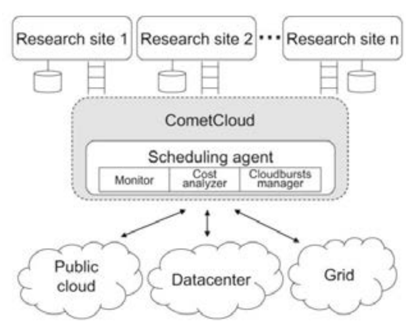
Project	Hadoop-GIS: A high performance query system for analytical medical imaging [55]
Cloud Use Cases	Data management and analysis
Primary Researcher	Fusheng Wang, Emory University
Abstract	Querying and analyzing large volumes of spatially oriented scientific data becomes increasingly important for many applications. For example, analyzing high-resolution digital pathology images through computer algorithms provides rich spatially derived information of micro-anatomic objects of human tissues. The spatial oriented information and queries at both cellular and sub-cellular scales share the common characteristics of a "Geographic Information System (GIS)," and provide an effective vehicle to support computer aided biomedical research and clinical diagnosis through digital pathology. The scale of data could reach a million derived spatial objects and a hundred million features for a single image. Managing and querying such spatially derived data to support complex queries such as image-wise spatial cross-matching queries poses two major challenges: the high complexity of geometric computation and the ``big data'' challenge. In this paper, we present a system Hadoop-GIS to support high performance declarative spatial queries with MapReduce. Hadoop-GIS provides an efficient real-time spatial query engine RESQUE with dynamically built indices to support on the fly spatial query processing. To support high performance queries with cost effective architecture, we develop a MapReduce-based framework for data partitioning and staging, parallel processing of spatial queries with RESQUE, and feature queries with Hive, running on commodity clusters. To provide a declarative query language and unified interface, we integrate spatial query processing into Hive to build an integrated query system. Hadoop-GIS demonstrates highly scalable performance to support our query cases [56].
Cloud Providers	FutureGrid
Special Features	Hive, MapReduce
Dev. Environment	Eucalyptus
Use Regularity	Monthly
Cores Used Peak	320
Cores Steady State	320
Core Hours in a Year	960
Storage Accessed For	Analysis
Preferred Storage	HDFS
Accessed During Run	1TB and 32GB
Short-Term Storage	0
Long-Term Storage	0
Data Moved Into Cloud	1TB and 32GB
Data Moved Out Cloud	2GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Cloud Funding	NSF
Research Funding	NIH



Example query spatial cases in analytical medical imaging

Biomedical Imaging Informatics: Medical Image Registration

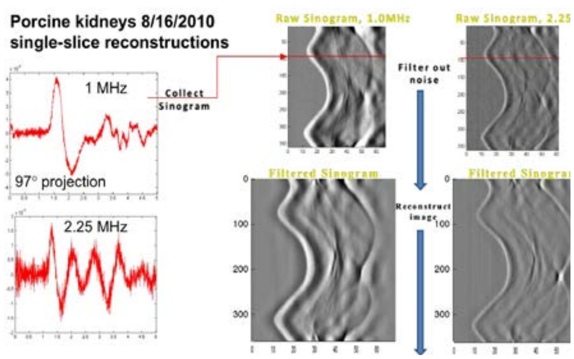
Project	Use of clouds and automatic cloud bursting to support medical image registration
Use Cases	Burst resources; collaboration; data archiving; data management and analysis; data sharing; domain-specific computing environments; education, outreach, and training (EOT); event-driven real-time science; science gateways
Primary Researchers	Manish Parashar, Moustafa AbdelBaky, Ivan Rodero, Xin Qi, Lin Yang and David Foran, Rutgers University
Abstract	Emerging cloud services represent a new paradigm for computing based on-demand access to computing utilities, an abstraction of unlimited computing resources, and a usage-based payment model. Furthermore, integrating these public cloud platforms (e.g., Amazon EC2) with existing computational Grids and HPC resources provides opportunities for on-demand scale-up and scale-down, i.e., cloudbursts. While such a paradigm can potentially have a significant impact on a wide range of application domains, various aspects of the existing applications and of current cloud infrastructure make the transition to clouds challenging. This work investigates the use of clouds and autonomic cloud-bursting to support a medical image registration application. The goal is to enable a virtual computational cloud that integrates local computational environments and public cloud services on-the-fly, and support image registration requests from different distributed researcher groups with varying computational requirements and QoS constraints. A policy-driven scheduling agent uses the QoS constraints along with performance history and the state of the resources to determine the appropriate size and mix of the public and private cloud resource that should be allocated to a specific request. The virtual cloud infrastructures and the cloud-based medical image registration were deployed on a combination of private clouds at Rutgers University, the Cancer Institute of New Jersey, and Amazon EC2 [57].
Cloud Providers	Amazon Web Services; FutureGrid
Special Features	GPUs
Use Regularity	Monthly
Cores Used Peak	1000
Cores Steady State	256
Core Hours in a Year	200000
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	1GB
Short-Term Storage	10GB
Long-Term Storage	10GB
Data Moved Into Cloud	2GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Department or institution
Software	CometCloud
Cloud Funding	NSF; departmental; institutional
Research Funding	NSF; NIH



Overview of medical image registration application scenario using CometCloud

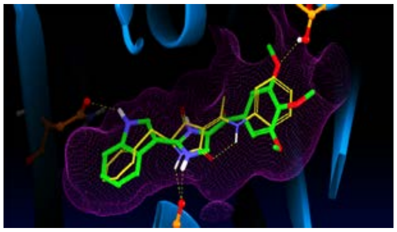
Biomedical Imaging Informatics: Thermoacoustic Computed Tomography

Project	Embarrassingly parallel backprojection of thermoacoustic tomography
Use Cases	Burst resources, event-driven real-time science
Primary Researchers	Sarah Patch, University of Wisconsin-Milwaukee
Abstract	We are reconstructing thermoacoustic tomography (TCT) data, which ideally represents a spherical radon transform. We reconstruct via filtered backprojection, and backprojection is a computationally costly and embarrassingly parallel operation. Our long-term goal is to quantify the robustness of TCT across different sizes, depths, and types of cancer. Ideally, TCT deposits electromagnetic (EM) energy impulsively in time and uniformly throughout the imaging object, causing thermal expansion. Cancerous masses preferentially absorb EM energy, heat and expand faster than neighboring healthy tissue, creating a pressure wave which is detected by ultrasound transducers at the edge of the object. We have developed an inversion formula for idealized TCT data and are now working to account for physical and experimental effects upon TCT data [58].
Cloud Providers	Red Cloud
Special Features	MATLAB Distributed Computing Server
Use Regularity	Weekly
Cores Used Peak	51
Cores Steady State	1
Core Hours in a Year	6000
Storage Accessed For	Analysis
Preferred Storage	Not specified
Accessed During Run	1GB
Short-Term Storage	1GB
Long-Term Storage	0
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Uncertain
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown
Cloud Funding	NIH
Research Funding	University of Wisconsin; NIH



Chemistry: Computational Chemistry

Project	Large scale utility supercomputing
Use Cases	Domain-specific computing environments; science gateways
Primary Researchers	James Watney, Schrodinger – Nimbus Discovery
Additional Researchers	James Stowe, Cycle Computing
Abstract	50,000-core compute across all 7 Amazon regions using on-demand and spot instances for a computational docking application, Glide, which performs high-throughput virtual screening of compound libraries for identification of drug discovery leads.
Cloud Providers	Amazon Web Services
Special Features	Community datasets or collections; GPUs; auto scale; spot instance management
Dev. Environment	Nimbus; OpenStack; VMware
Use Regularity	Daily
Cores Used Peak	50000
Cores Steady State	5000
Core Hours in a Year	1000000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Object store
Accessed During Run	500GB
Short-Term Storage	1TB
Long-Term Storage	3TB
Data Moved Into Cloud	20TB
Data Moved Out Cloud	15TB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Any users may access the data collection or survey results
Software	Home-grown; community developed; open source; commercial; meta-schedulers (CycleCloud, CycleServer)
Capabilities/Features	“Large elastic scalability; error handling; and, compute optimization.”
Additional Notes	“Following successful projects over the past 12 months to spin up a 10,000-core computer in the cloud with Genentech….the idea Schrödinger brought to Cycle was to conduct a virtual screen of 7 million compounds in multiple conformations—a total of 21 million ligand structures compared to a protein target using a docking application called Glide. The new run surpassed 50,000 cores distributed across seven AWS sites around the world—three in North America, and one each in Europe, Brazil, Singapore and Japan. About 80% of the workload was distributed across 5,000 servers at Amazon’s east coast facility in Virginia. The experiment—the equivalent of 12.5 processor-years—was conducted in a mere three hours. The final cost (2012) was $4,828/hour, or 9 cents/core/hour. Previously, it would take Schrödinger about 11 days to run a similar analysis on its in-house 400-core cluster—stopping all other work in the process [59].”
Cloud Funding	Commercial
Research Funding	Commercial



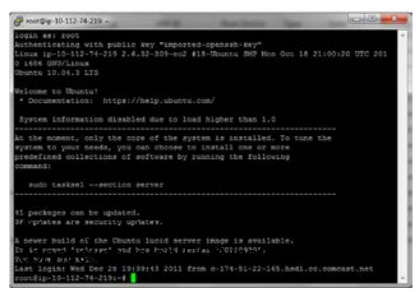
Screening chemical compounds and predicting binding modes

Chemistry: Predicting Chemical Properties

Project	VENUS-C
Use Cases	Burst resources
Primary Researchers	Jacek Cala, Newcastle University
Additional Researchers	James Stowe, Cycle Computing
Abstract	Under VENUS-C we were developing a drug discovery scenario which included a large number of statistical (QSAR) models to be built. The scenario is inherently bursty as the model generation is driven by the molecule data provided by external institutions, e.g. EBI updates their ChEMBL database twice a year. The target is to make all “good” models available to the public. Once this is done we expect some additional user traffic [60].
Cloud Providers	Windows Azure
Special Features	Tables
Use Regularity	Daily
Cores Used Peak	220
Cores Steady State	20
Core Hours in a Year	100000
Storage Accessed For	Analysis; archival
Preferred Storage	Object store
Accessed During Run	1GB
Short-Term Storage	20GB
Long-Term Storage	5GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	5GB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 100Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; any users may access the data collection and survey results
Software	Home-grown; community developed; open source; commercial; e-Science Central; JBOSS AS; postgreSQL; ClumsyLeaf (Table/Cloud); Xplorer; GNU wget; and, many more
Capabilities/Features	“Our scenario is very bursty and often we transfer 5GB in a few days after which almost no data is moved in or out.”
Additional Notes	“This cloud solution is primarily aimed at domain scientists who do not have advanced IT skills. Quantitative Structure Activity Relationship (QSAR) workflows have been built leveraging e-Science Central. Chemists use QSAR models to focus on the synthesis of new compounds, to design better, safer drugs, as well as more environmentally benign products. Being able to predict the activity of molecules reduces the need to test them in the laboratory, a costly and time-consuming process [61].
Cloud Funding	Commercial; European Union
Research Funding	European Union

CS: Education and Training – Computer/Network Security Labs

Project	Using Amazon EC2 in Computer/Network Security labs
Use Cases	Computer science research; education, outreach, and training (EOT)
Primary Researchers	Chuan Yue, University of Colorado at Colorado Springs
Additional Researchers	Weiying Zhu, Metropolitan State University of Denver; Greg Williams and Edward Chow, University of Colorado at Colorado Springs
Abstract	Cloud computing is a significant trend in computing. In this paper, we present our experience in using Amazon EC2 (Amazon Elastic Compute Cloud) as the platform to support the hands-on lab exercises of a computer and network security course. In this course, each student is required to perform four realistic lab exercises using Amazon EC2: an IDS (Intrusion Detection System) lab exercise, a Linux firewall lab exercise, a Web security lab exercise, and a software vulnerability exploitation lab exercise. Hosting these security lab exercises in the cloud brings us two main benefits. One is that we can better prepare our students for their future careers in a cloud computing world. The other is that we can effectively address the resource limitation of our existing lab environments and meanwhile ease the burden on our IT professionals who need to take care of the needs of many courses and maintain the existing infrastructure for college operations. Using Amazon EC2 in particular, we can take advantage of its reliability, availability, robustness, accessibility, security, and uniformity. Through the survey answered by our students, we found that the majority of our students are in favor of learning and using such a leading cloud computing platform, and a common opinion among students is that Amazon EC2 is easy to learn and convenient to use. We describe the setup of our EC2 environment and the design of those four lab exercises. We also detail the survey results and analyze the implications of those results. The experience presented in this paper [62] is valuable for our faculty members to move more lab exercises into the cloud. We believe our experience is also valuable to other educators who plan to use cloud computing services such as Amazon EC2 in their computer science and engineering courses. The link to our complete lab manuals and instructions [63] is listed at the end of the bibliographic section.
Cloud Providers	Amazon Web Services
Dev. Environment	Linux; Windows OS
Use Regularity	Weekly
Cores Used Peak	1
Cores Steady State	1
Core Hours in a Year	1600
Storage Accessed For	Analysis; archival
Preferred Storage	Elastic Block Storage
Accessed During Run	1TB
Short-Term Storage	1TB
Long-Term Storage	1TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Providing basic access; research data sets or collections
Accessed By	Researcher; research group; department or institution; outside collaborators; any users may access my data collections and survey results
Software	Community-developed; open source; Linux; Windows OS; Apache Web Server, MySQL; Nessus; etc.



Working on labs from an SSH terminal connecting to an EC2 instance

Capabilities/Features	“Amazon AWS Management Console and AWS Identity and Access Management (IAM) Web services are very convenient and easy to use. We love to use AWS.”
Problems/Limitations	“It would be great if the compute instances (e.g., EC2 instances) could be managed in a more flexible and fine-grained manner.”
Additional Notes	“In Fall 2011, our students used Amazon EC2 (Elastic Compute Cloud) as the platform to work on the four lab exercises of our computer and network security course. There were 28 students in the class. At the beginning of the semester, the instructor, Dr. Chuan Yue, was awarded an AWS Teaching Grant from Amazon to use the AWS cloud infrastructure in teaching. The total credit awarded to the instructor was $3,600. At the end of the semester when the students completed all four of the lab exercises, $3,311 remained on the instructor’s AWS account. Hence, 28 students, one instructor, and one teaching assistant amazingly used only $289 for four lab exercises in a semester, much less than the originally expected cost. At the end of each lab exercise, a survey was given to the students to obtain insight on the students’ perception of using Amazon EC2 for hands-on security lab exercises. According to the survey results, the average number of hours worked on each lab exercise varied between 7.0 hours and 14.5 hours….results indicated that Amazon EC2 can be used cost-effectively for hosting hands-on lab exercises [64].”
Cloud Funding	Amazon
Research Funding	Amazon

CS: Education and Training – Data Center Scale Computing Class

Project	Data center scale computing class
Use Cases	Data management and analysis; education, outreach, and training (EOT)
Primary Researchers	Dirk Grunwald, University of Colorado, Boulder
Abstract	I'm teaching a class on "data center scale computing." Students have been using FutureGrid to get experience with creating and managing cloud instances and storage, as well as with distributed systems software (ZooKeeper, RabbitMQ, etc.) and eventually Hadoop. We're using a combination of an Amazon Web Services donation and FutureGrid [65].
Cloud Providers	Amazon Web Services; FutureGrid
Special Features	MapReduce
Dev. Environment	Eucalyptus; OpenStack
Use Regularity	Weekly
Cores Used Peak	30
Cores Steady State	2
Core Hours in a Year	30
Storage Accessed For	Analysis; reference
Preferred Storage	Elastic Block Storage
Accessed During Run	40GB
Short-Term Storage	40GB
Long-Term Storage	40GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Department or institution
Software	Community developed; open source
Capabilities/Features	"When teaching a class about cloud/data center scale computing, it's useful to have a cheap/free service, because students spend a lot of time spinning up an instance only to tear it down again. Each up/down cycle would cost an hour's expense on AWS, and even though it's a few pennies, it adds up."
Problems/Limitations	"FutureGrid appears to have limited staff support and issues about upgrading to current software. However, these are mostly issues of timeliness (which is affected by budget I assume) rather than quality -- they do a good job and have a good infrastructure. That said, since it's free, it's better than what I would have cobbled together for my class of 30 students. I'm not certain what I will do in the future if e.g., FutureGrid is not available."
Cloud Funding	Amazon donation
Research Funding	None

CS: Education and Training – Cloud Programming

Project	Optimization cloud-friendly techniques
Use Cases	Education, outreach, and training (EOT)
Primary Researchers	Javid Taheri, The University of Sydney
Abstract	Providing student training on how to program in the cloud.
Cloud Providers	Windows Azure
Special Features	Community datasets or collections
Use Regularity	Monthly
Cores Used Peak	20
Cores Steady State	2
Core Hours in a Year	400
Storage Accessed For	Reference
Preferred Storage	HDFS
Accessed During Run	5GB
Short-Term Storage	10GB
Long-Term Storage	2GB
Data Moved Into Cloud	10GB
Data Moved Out Cloud	10GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Commercial
Cloud Funding	Institutional
Research Funding	Microsoft

CS: Education and Training – Data Management Labs

Project	University of Washington – CSE344
Use Cases	Education, outreach, and training (EOT)
Primary Researchers	Magdelena Balazinska, University of Washington
Additional Researchers	Dan Suciu, University of Washington
Abstract	Undergraduate database course – CSE344: Introduction to Data Management [66]
Cloud Providers	Amazon Web Services; Windows Azure
Special Features	MapReduce
Dev. Environment	Eucalyptus, OpenStack
Use Regularity	Annually
Cores Used Peak	1200 (60 students x 20 cores)
Cores Steady State	60
Core Hours in a Year	9600 (2 classes of 60 students)
Storage Accessed For	Analysis
Preferred Storage	Object store
Accessed During Run	500GB
Short-Term Storage	500GB
Long-Term Storage	500GB
Data Moved Into Cloud	500GB
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Uncertain
Type Data Moving	Not moving data, just programs
Data Accessed By	Department or institution
Software	Open source; commercial; SQL Azure; Amazon Elastic MapReduce with Pig; S3
Capabilities/Features	“We don't really move data in/out of the cloud. For the compute, we have about 120 students/year split into two classes. In Assignments 1 and 2, students use SQLite on their laptops with no cloud. In Assignment 3, students use SQL Azure with one instance for the whole class [67]. For Assignment 4, students work on XML on their laptops. Assignment 5 uses SQL Azure with one database for the class and one database/student. Students write Java programs that talk to the SQL Azure databases. In Assignment 6, students use approximately 20 cloud cores each [68].
Problems/Limitations	“Not really. The setup each quarter (two classes offered per year) is still a hassle but then the assignments work fine.”
Cloud Funding	Cloud providers
Research Funding	None

Students learn data management by doing homework assignments in the cloud

CS: Education and Training – Science Cloud Summer School 2012

Project	Science Cloud Summer School 2012
Use Cases	Burst resources; education, outreach, and training (EOT)
Primary Researchers	Gregor von Laszewski, Indiana University
Additional Researchers	Fugang Wang, Indiana University
Abstract	The Science Cloud Summer School targets education and training of graduate students and the fostering of a community around a topic that has increasing interest and relevance: the use of cloud computing technologies in science – including Infrastructure-as-a-Service and Platform-as-a-Service. Because cloud computing systems and technologies provide a considerable departure from traditional models and evolve at a rapid pace, this event would provide a basis for students to immerse in a focused, intensive curriculum to learn fundamentals and experiment with these technologies in practice. We will cover topics of interest to students with both an application and computer science focus [69].
Cloud Providers	FutureGrid
Special Features	MapReduce
Dev. Environments	Eucalyptus; Nimbus; OpenNebula; OpenStack
Use Regularity	Annually
Cores Used Peak	1000
Cores Steady State	1000
Core Hours in a Year	50
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Elastic Block Storage; HDFS; object store; parallel performance file system; Wide Area Files Systems
Accessed During Run	500GB
Short-Term Storage	500GB
Long-Term Storage	500GB
Data Moved Into Cloud	0
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Cloud Funding	NSF
Research Funding	NSF

CS: Education and Training – CCGrid 2011 Tutorial

Project	Tutorial: CCGrid2011
Use Cases	Education, outreach, and training (EOT)
Primary Researchers	Gregor von Laszewski, Indiana University
Additional Researchers	Andrew Younge, Indiana University
Abstract	The FutureGrid (FG) testbed provides computing capabilities that will enable researchers to tackle complex research challenges related to the use of Grids and Clouds. The FG testbed includes a geographically distributed set of heterogeneous computing systems, of about 5000 cores, a data management system that will hold both metadata and a growing library of software images necessary for Cloud computing, and a dedicated network allowing isolated, secure experiments. The testbed supports virtual machine-based environments, as well as operating systems on native hardware for experiments aimed at minimizing overhead and maximizing performance. The tutorial starts with an introduction and overview of the services offered by FutureGrid to the community [70].
Cloud Providers	FutureGrid
Dev. Environments	Eucalyptus; Nimbus
Use Regularity	Annually
Cores Used Peak	678
Cores Steady State	678
Core Hours in a Year	678
Storage Accessed For	Analysis
Preferred Storage	Elastic Block Storage; HDFS; object store; parallel performance file system; Wide Area Files Systems
Accessed During Run	0
Short-Term Storage	0
Long-Term Storage	0
Data Moved Into Cloud	0
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source; commercial
Cloud Funding	NSF
Research Funding	NSF

CS: Cloud Performance – Cloud Function and Performance Comparison

Project	Community comparison of cloud frameworks
Use Cases	Burst resources; computing and data analysis support for scientific workflows; science gateways
Primary Researchers	Yong Zhao, University of Electronic Science and Technology of China
Additional Researchers	Gregor von Laszewski, Indiana University
Abstract	We will conduct functionality and performance comparison of multiple clouds, and develop a set of benchmarks for clouds [71].
Cloud Providers	FutureGrid
Dev. Environment	Eucalyptus; Nimbus; OpenStack
Use Regularity	Weekly
Cores Used Peak	256
Cores Steady State	2
Core Hours in a Year	400
Storage Accessed For	Analysis
Preferred Storage	Parallel performance file system
Accessed During Run	10GB
Short-Term Storage	100GB
Long-Term Storage	100GB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	10GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Home-grown; open source
Cloud Funding	NSF
Research Funding	None

CS: Cloud Performance – Evaluating Clouds for Large Scale, Parallel Applications

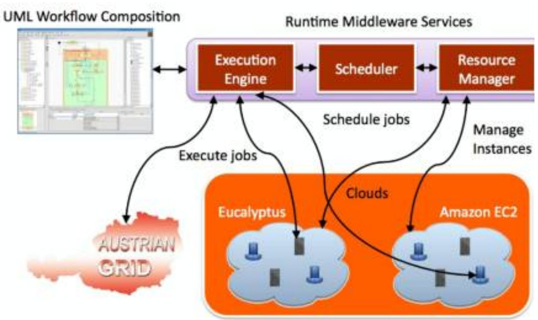
Project	GE Energy Multi-Area Reliability Simulation Software Program (MARS)
Use Cases	Collaboration; computing and data analysis support for scientific workflows
Primary Researchers	Ketan Maheshwari, Cornell University
Abstract	Performing large scale parallel runs. Mainly to evaluate clouds for parallel, large scale applications.
Cloud Providers	Amazon Web Services, FutureGrid, Red Cloud, Open Science Data Cloud
Dev. Environment	Eucalyptus; Nimbus; OpenStack
Use Regularity	Weekly
Cores Used Peak	200
Cores Steady State	180
Core Hours in a Year	2000
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	60GB
Short-Term Storage	50GB
Long-Term Storage	0GB
Data Moved Into Cloud	10GB
Data Moved Out Cloud	10GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Outside collaborators
Software	Open source; commercial
Cloud Funding	Commercial
Research Funding	European Union

CS: Cloud Performance – Scalable File Systems and Datastores

Project	Scalable file systems and datastores for cloud environments
Use Cases	Computer science research
Primary Researchers	Stergios Anastasiadis, University of Ioannina, Greece
Abstract	We are investigating the problem of storage scalability in the context of (i) file systems for virtual machines, and (ii) key-value stores.
Cloud Providers	Amazon Web Services
Special Features	Community datasets or collections; GPUs; MapReduce; queues; SQLaaS; tables
Dev. Environment	Xen
Use Regularity	Monthly
Cores Used Peak	32
Cores Steady State	4
Core Hours in a Year	5000
Storage Accessed For	Archival
Preferred Storage	Elastic Block Storage
Accessed During Run	10GB
Short-Term Storage	100GB
Long-Term Storage	100GB
Data Moved Into Cloud	10GB
Data Moved Out Cloud	10GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Research group
Software	Open source
Cloud Funding	Commercial
Research Funding	EU

CS: Cloud Programming/ Workflows – ASKALON

Project	ASKALON
Use Cases	Burst resources; computer science research; computing and data analysis support for scientific workflows; science gateways
Primary Researchers	Thomas Fahringer, Simon Ostermann, Kassian Plankensteiner, Hamid Mohammadi Fard, Malik Junaid, and Mathias Janetschek, University of Innsbruck, Austria
Abstract	The Cloud Computing paradigm holds good promise for the performance hungry scientific community. Clouds promise to be a cheap alternative to supercomputers and specialized clusters, a much more reliable platform than grids, and a much more scalable platform than the largest of commodity clusters or resource pools. Clouds also 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 even supercomputers or to the financial burden of over-provisioning resources. Our projects utilized this new resource to execute scientific workflow applications in a fast and cost efficient way [72].
Cloud Providers	Amazon Web Services; FutureGrid; Google Cloud Platform; Grid’5000
Dev. Environments	Eucalyptus
Use Regularity	Weekly
Cores Used Peak	160
Cores Steady State	12
Core Hours in a Year	34176
Storage Accessed For	Analysis
Preferred Storage	Object store
Accessed During Run	100GB
Short-Term Storage	10GB
Long-Term Storage	0
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher; research group; outside collaborators
Software	Home-grown; community developed; open source
Capabilities/Features	“Scalability.”
Problems/Limitations	“Not all features available that we would like to research (QoS, migration, dynamic scaling...).”
Cloud Funding	Standortargentur Tirol, Fonds zur Förderung der wissenschaftlichen Forschung
Research Funding	Standortargentur Tirol, Fonds zur Förderung der wissenschaftlichen Forschung



ASKALON cloud and grid application development and computing environment

CS: Cloud Programming/Workflows – CloudFlow Systems

Project	Context-oriented CloudFlow system and application in virtual screening
Use Cases	Computing and data analysis support for scientific workflows; data management and analysis; data sharing; education, outreach, and training (EOT)
Primary Researchers	Xiaoliang Fan, Lanzhou University, China
Abstract	Context-oriented CloudFlow system and application in virtual screening which makes context explicit during the lifecycle of scientific workflow (especially design and execution phase). A case study is about a classical data-intensive application: virtual screening.
Cloud Providers	Amazon Web Services; FutureGrid; Globus Online; Google Cloud Platform
Special Features	Community datasets or collections; GPUs; HBase; Hive; MapReduce; SQLaaS
Dev. Environments	Eucalyptus; Nimbus; OpenStack; VMware
Use Regularity	Annually
Cores Used Peak	2048
Cores Steady State	2048
Core Hours in a Year	300
Storage Accessed For	Analysis; archival
Preferred Storage	Object store
Accessed During Run	15TB
Short-Term Storage	3TB
Long-Term Storage	15TB
Data Moved Into Cloud	3TB
Data Moved Out Cloud	10TB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Open source
Cloud Funding	Departmental
Research Funding	NSFC

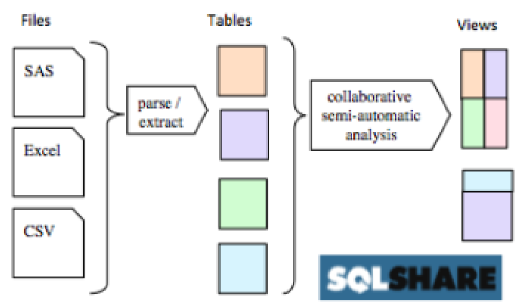
CS: Cloud Programming/Workflows – Cooperative Computing Tools

Project:	Bridging Cyberinfrastructure with the Cooperative Computing Tools
Use Cases	Burst resources; commonly requested software; computer science research; computing and data analysis support for scientific workflows; data management and analysis; domain-specific computing environments; data sharing; education, outreach, and training (EOT); science gateways
Primary Researchers	Douglas Thain, University of Notre Dame
Abstract	This project supports the maintenance and development of the Cooperative Computing Tools. This software package is designed to enable non-privileged users to harness hundreds to thousands of cores from multiple clusters, clouds, and grids simultaneously. The main components of the software package include Parrot, a virtual file system that interfaces with multiple distributed storage systems, and Makeflow, a workflow engine that interfaces with multiple computing systems. This project will develop, maintain, and support the software across a wide variety of operating systems and national scale cyberinfrastructure in support of high impact scientific applications in fields such as bioinformatics, biometrics, data mining, high energy physics, and molecular dynamics. Large scale computing systems such as cluster, clouds, and grids now make it easy for end users to purchase large amounts of computing power at the touch of a button. However, these computing systems are difficult to harness because they each present a different user interface, principle of operation, and programming model. This project addresses this problem by supporting the development of the Cooperative Computing Tools, a software package that makes it possible for ordinary computer applications to move seamlessly between different service providers. The software is primarily of interest to researchers in scientific domains that require large amounts of computation. It is currently used by researchers in the fields of bioinformatics, biometrics, data mining, high energy physics, and molecular dynamics [73].
Cloud Providers	Amazon Web Services; FutureGrid; Windows Azure
Dev. Environments	Cooperative Computing Tools
Use Regularity	Weekly
Cores Used Peak	2500
Cores Steady State	100
Core Hours in a Year	20000
Storage Accessed For	Analysis
Preferred Storage	Conventional file systems
Accessed During Run	10TB
Short-Term Storage	10TB
Long-Term Storage	10TB
Data Moved Into Cloud	10GB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Don't use cloud storage
Type Data Moving	Research data sets or collections; not moving data, just programs
Data Accessed By	Researcher; research group; department or institution; outside collaborators
Software	Home-grown; community developed; open source; commercial; distributed computing—cooperative computing tools; Hadoop; Condor; bioinformatics; BLAST; SSAHA
Capabilities/Features	"Not sure how to answer the cloud storage questions. We move data to each virtual machine for the duration of a run, but any important outputs

	are moved back to the home institution. We don't make use of cloud storage services, apart from the storage attached to each VM instance. So, that could be 10GB per VM instance, which might sum up to 10TB during a run, but is then discarded quickly once the important parts are saved."
Problems/Limitations	"Most of our collaborators have the following view of cloud resources: clouds are excellent at providing burst capacity and custom software environments for computation and data analytics. However, they are very wary of committing to any one cloud provider. On the storage side, they are very concerned about the high cost of long term storage, and the risk of data loss or extreme cost retrieval. Businesses place a much lower valuation on data than the researcher does. They are much more comfortable keeping their data at the local campus, where they can control and access it on demand. On the computing side, no one wants to get committed to a software framework (e.g., Google App Engine or Windows Azure) that would lock them into a single provider. Rather, they wish to be able to move codes to whatever service provides the most convenient/economical service today. So, it is much more desirable to construct the software framework independently of the cloud service, and then harness all cluster/clouds/grids that happen to be available at the moment."
Cloud Funding	NSF; personal; institutional; commercial
Research Funding	NSF

CS: Cloud Programming/Workflows – Database-as-a-Service

Project	SQLShare: Database-as-a-Service for long tail science
Use Cases	Collaboration; commonly requested software; data management and analysis; data sharing; domain-specific computing environments
Primary Researchers	Mike Cafarella, University of Michigan; Dan Suciu, University of Washington; David Maier, Portland State University
Abstract	Science is reducing to a database problem, but database technology is not keeping pace. This problem is especially acute in the long tail of science: the large number of relatively small labs and individual researchers who collectively produce the majority of scientific results. These researchers lack the IT staff and specialized skills to deploy technology at scale, but have begun to routinely access hundreds of files and potentially terabytes of data to answer a scientific question. This project develops the architecture for a database-as-a-service platform for science. It explores techniques to automate the remaining barriers to use: ingesting data from native sources and automatically bootstrapping an initial set of queries and visualizations, in part by aggressively mining a shared corpus of data, queries, and user activity. It investigates methods to extract global knowledge and patterns while offering the scientists access control over their data, and some formal privacy guarantees. The Intellectual Merit of this proposal consists of automating non-trivial cognitive tasks associated with data work: information extraction from unstructured data sources, data cleaning, logical schema design, privacy control, visualization, and application-building. As Broader Impacts, the project helps increase the productivity of scientists and researchers, by allowing them to focus on their problem at hand and relieving them of the need to perform tedious data management tasks [74].
Cloud Providers	Amazon Web Services; Windows Azure
Special Features	SQLaaS; tables
Use Regularity	Daily
Cores Used Peak	3
Cores Steady State	3
Core Hours in a Year	18000
Storage Accessed For	Analysis
Preferred Storage	Database
Accessed During Run	1GB
Short-Term Storage	50GB
Long-Term Storage	50GB
Data Moved Into Cloud	20GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections; survey data
Data Accessed By	Researcher; research group; outside collaborators; any users may access data collections and survey results
Software	Home-grown; open source; commercial
Additional Notes	"…several researchers we have surveyed informally have reported that the ratio of time they spend 'manipulating data' as opposed to 'doing science' is a staggering 9 to 1….spreadsheets and ASCII files remain the most popular tools for data management in the long tail. But as data volumes continue to explode, cut-and-paste manipulation of spreadsheets cannot scale, and the relatively cumbersome development cycle of scripts and workflows for ad hoc, iterative data manipulation



SQLShare simplifies collaborative, semi-automated data management in the cloud

	becomes the bottleneck to scientific discovery and a fundamental barrier to those without programming experience [75].”
Cloud Funding	Institutional; commercial
Research Funding	NSF; commercial; Gordon and Betty Moore Foundation

CS: Cloud Programming/Workflows – Interactive Multi-Tier Performance

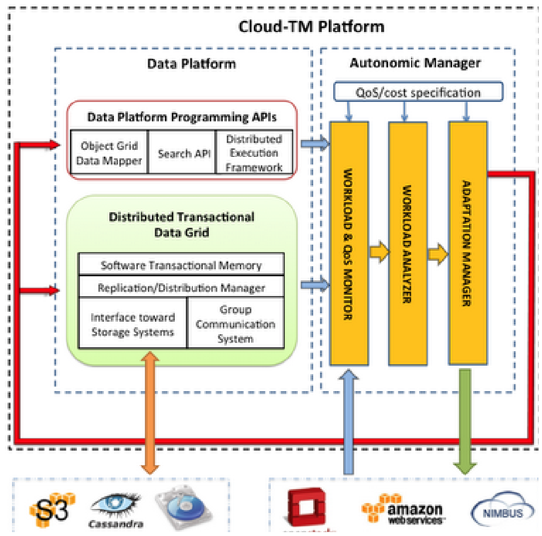
Project	Architecting latency sensitive applications for the cloud
Use Cases	Computer science research
Primary Researchers	Sanjay Rao, Mohammad Hajjat, and Shankar Narayanan, Purdue University
Abstract	Cloud computing offers IT organizations the ability to create geo-distributed, and highly scalable applications while providing attractive cost-saving advantages. Yet, architecting, configuring, and adapting cloud applications to meet their stringent performance requirements is a challenge given the rich set of configuration options, shared multi-tenant nature of cloud platforms, and dynamics resulting from activities such as planned maintenance. A unique area of focus of our research is interactive multi-tier applications (e.g., enterprise applications, web applications) which have received limited attention from the community. We are developing novel methodologies, and systems that can enable application architects to (1) judiciously architect their applications across multiple cloud data-centers while considering application performance requirements, cost saving objectives, and cloud pricing schemes guided by performance and cost models of cloud components such as key-value datastores; (2) create applications that can adapt to ongoing dynamics in cloud environments through transaction reassignment over shorter time-scales. Our research if successful can enable IT organizations to significantly reduce costs by optimally moving their operations to the cloud. We are also working on creating benchmarks based on operationally deployed applications and collecting workload traces which will be made available to the research community [76].
Cloud Providers	Amazon Web Services; Windows Azure
Special Features	GPUs; HBase; Hive; MapReduce; queues; SQLaaS; tables; Elastic cache/Azure cache
Dev. Environment	Azure SDK
Use Regularity	Daily
Cores Used Peak	100
Cores Steady State	25
Core Hours in a Year	200000
Storage Accessed For	Analysis
Preferred Storage	Object store
Accessed During Run	4GB
Short-Term Storage	20GB
Long-Term Storage	10GB
Data Moved Into Cloud	200GB
Data Moved Out Cloud	300GB
BW In/Out of Cloud	Up to 100Gb/s
BW to Storage Within	UP to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group; department or institution
Software	Community developed; open source
Cloud Funding	NSF; commercial
Research Funding	NSF

CS: Cloud Programming/Workflows – Data Enabled Science

Project	Programming environments and runtime for data enabled science
Use Cases	Burst resources; collaboration; data management and analysis; data sharing; education, outreach, and training (EOT)
Primary Researchers	Judy Qui, Indiana University
Abstract	Computational simulation and analysis were one of the keys to the future in data-intensive science but are facing a major challenge handling the incredible increases size and complexity in datasets. This requires attractive powerful programming models that address issues of portability with scaling performance and fault tolerance. Iterative computations are pervasive among data analysis applications, including web search, online social network analysis, image processing, and clustering as seen in Intel RMS Analysis. These applications typically involve data sets of massive scale. We intend to justify that extensions of Iterative MapReduce (as illustrated by Pregel and Twister) are a basis to address data intensive problems as they interpolate between the traditional tightly coupled MPI jobs typical of supercomputers, and the more loosely coupled information retrieval and pleasingly parallel (“map only”) applications typical of clouds and high throughput systems [77].
Cloud Providers	Amazon Web Services; FutureGrid; Window Azure
Special Features	GPUs; HBase; Hive; MapReduce; queues; tables;
Dev. Environments	Eucalyptus; Nimbus: OpenStack; VMware
Use Regularity	Weekly
Cores Used Peak	1344
Cores Steady State	800
Core Hours in a Year	1000000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	HDFS
Accessed During Run	100GB
Short-Term Storage	100GB
Long-Term Storage	40TB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	100GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group; outside collaborators
Software	Home-grown; community developed; open source; commercial
Cloud Funding	NSF; institutional; commercial
Research Funding	NSF; commercial

CS: Cloud Programming/Workflows – Transactional Memory Middleware

Project	Cloud-TM
Use Cases	Computer science research
Primary Researchers	Paolo Romano, INESC-ID, Portugal
Abstract	Cloud-TM aims at defining a novel programming paradigm to facilitate the development and administration of cloud applications. It will develop a Self-Optimizing Distributed Transactional Memory middleware that will spare programmers from the burden of coding for distribution, persistence and fault-tolerance, letting them focus on delivering differentiating business value. Further, the Cloud-TM platform aims at minimizing the operational costs of cloud applications, pursuing optimal efficiency via autonomic resource provisioning and pervasive self-tuning schemes [78].
Cloud Providers	Amazon Web Services; FutureGrid
Dev. Environment	Nimbus; OpenStack
Use Regularity	Daily
Cores Used Peak	1000
Cores Steady State	100
Core Hours in a Year	36500
Storage Accessed For	Analysis; archival
Preferred Storage	Elastic Block Storage
Accessed During Run	1GB
Short-Term Storage	100GB
Long-Term Storage	500GB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown
Cloud Funding	Institutional
Research Funding	European Commission



Cloud-TM Data Platform and Autonomic Manager

CS: Cloud Provisioning and Monitoring – Cloud Controllers

Project	Kriging-based controllers for the cloud
Use Cases	Burst resources; commonly requested software; computer science research; computing and data analysis support for scientific workflows; data management and analysis
Primary Researchers	Mauro Pezze, University of Lugano, Switzerland
Additional Researchers	Giovanni Toffetti and Alessio Gambi, University of Lugano
Abstract	Cloud infrastructures allow service providers to implement elastic applications. These can be scaled at runtime to dynamically adjust their resources allocation to maintain consistent quality of service in response to changing working conditions, like flash crowds or periodic peaks. Providers need models to predict the system performances of different resource allocations to fully exploit dynamic application scaling and implement sled-adaptive controllers. Traditional performance models such as linear models and queuing networks might be simplistic for real Cloud applications; moreover, they are not robust to change. This talk proposes a performance modeling approach based on Kriging surrogate models to approximate the performance profile of virtualized, multi-tier Web applications. The talk presents the Kriging based model, a self-adaptive controller and experimental data that show the validity of the approach. This group uses the RESERVOIR Framework [79].
Cloud Providers	Amazon Web Services; Google Cloud Platform; Windows Azure
Special Features	MapReduce; queues; SQLaaS
Dev. Environment	Eucalyptus; OpenNebula; OpenStack
Use Regularity	Daily
Cores Used Peak	128
Cores Steady State	32
Core Hours in a Year	10000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	N/A
Accessed During Run	5TB
Short-Term Storage	5TB
Long-Term Storage	2TB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	5GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Home-grown; community developed; open source, commercial
Capabilities/Features	“Fast (one minute or less) provisioning of VMs.”
Cloud Funding	Commercial
Research Funding	Commercial; European Union

CS: Cloud Provisioning and Monitoring – Developing an Information Service

Project	Development of an information service for FutureGrid
Use Cases	Computer science research
Primary Researchers	Hyungro Lee, Indiana University
Abstract	While using FutureGrid as a platform we will be designing, implementing and deploying an information service for FutureGrid that will collect detailed information about the actual utilization of FutureGrid in regards to provisioning, utilization of images, and distributed runtime frameworks (Hadoop, MPI ...). In addition, this information system can be used to implement application level monitoring for Grid and Cloud applications. This system is using a messaging system and a nonsql-based data service (most likely MongoDB and Apache QPID messaging system). We will demonstrate the usefulness of the system in two contexts: (a) observing utilization on the system level, (b) using the system to develop an application that is agnostic towards network faults. The application domain we chose for this project is bioinformatics while considering biological applications such as BLAST, R, and ClustalW [80].
Cloud Providers	FutureGrid
Dev. Environment	Eucalyptus; Nimbus; OpenStack
Use Regularity	Daily
Cores Used Peak	10
Cores Steady State	1
Core Hours in a Year	744
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	10TB
Short-Term Storage	10TB
Long-Term Storage	1TB
Data Moved Into Cloud	10TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections; survey data
Data Accessed By	Researcher; research group; outside collaborators
Software	Home-grown; community developed; open source
Cloud Funding	NSF
Research Funding	NSF

CS: Cloud Provisioning and Monitoring – Dynamic Cloud/HPC Provisioning

Project	Rain: FutureGrid dynamic provisioning framework
Use Cases	Burst resources; computer science research
Primary Researchers	Gregor von Laszewski, Indiana University
Additional Researchers	Javier Diaz Montes, Indiana University
Abstract	This project allows its users to use dynamic provisioning on the production cluster. It allows users to provision OS and software stacks not only in clouds, but also on bare metal. This allows unique performance comparisons [81].
Cloud Providers	FutureGrid
Special Features	Root access
Dev. Environment	Eucalyptus; HPC; Nimbus; OpenNebula; OpenStack
Use Regularity	Weekly
Cores Used Peak	678
Cores Steady State	678
Core Hours in a Year	50
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Self-written unified image registry for clouds
Accessed During Run	3TB
Short-Term Storage	3TB
Long-Term Storage	3TB
Data Moved Into Cloud	0
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Cloud Funding	NSF
Research Funding	NSF
Capabilities/Features	"We use the software on a regular basis to reprovision a machine we have in FutureGird. We use, therefore, all available cores and servers."
Problems/Limitations	"No. We would like to deploy RAIN on other resources."

CS: Cloud Provisioning and Monitoring – Provisioning for e-Science

Project	Resource provisioning for e-Science environments
Use Cases	Computers science research; computing and data analysis support for scientific workflow
Primary Researchers	Andrea Bosin, University of Cagliari, Italy
Abstract	Recent works have proposed a number of models and tools to address the growing needs and expectations in the field of e-Science. In particular, they have shown the advantages and the feasibility of modeling e-Science environments and infrastructures according to the Service-Oriented Architecture (SOA). At the same time, the availability and models of use of networked computing resources needed by e-Science are rapidly changing and see the coexistence of many disparate paradigms: high performance computing, grid and recently cloud, which brings very promising expectations due to its high flexibility. Unfortunately, none of these paradigms is recognized as the ultimate solution, and a convergence of all of them should be pursued. In this project we wish to test a model to promote the convergence and the integration of different computing paradigms and infrastructures for the dynamic on-demand provisioning of the resources needed by e-Science environments, especially those developed according to SOA. In addition, such a model aims at endorsing a flexible, modular, workflow-based computing model for e-Science. A working implementation used to validate the proposed approach will be developed and tested using FutureGrid resources [82].
Cloud Providers	FutureGrid
Dev. Environment	Eucalyptus; Nimbus; OpenNebula; OpenStack
Use Regularity	Monthly
Cores Used Peak	128
Cores Steady State	16
Core Hours in a Year	512
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	5GB
Short-Term Storage	0
Long-Term Storage	0
Data Moved Into Cloud	5GB
Data Moved Out Cloud	5GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets and collections
Data Accessed By	Researcher; research group; any users may access data collections and survey results
Software	Home-grown; community developed; open source
Cloud Funding	Institutional
Research Funding	Institutional

CS: Security/Highly Assured Clouds – Cataloging Cloud Security Issues

Project	Exploring and cataloging cloud computing security issues via FutureGrid
Use Cases	Collaboration; computer science research; computing and data analysis support for scientific workflows; data management and analysis; data sharing; domain-specific computing environments; education, outreach, and training (EOT); event-driven real-time science
Primary Researchers	Adetunji Adeleke, Indiana University-Purdue University Indianapolis; Bina Bhaskar, Indiana University
Additional Researchers	Gregor von Laszewski and Yangyi Chen, Indiana University
Abstract	A mention of the words "Cloud Computing" mostly comes with the question "How safe is Cloud Computing?," however the benefits that it offers in terms of improved costs and better performance via distributed computing resources in virtualized infrastructures and grid clusters makes it inevitable to use now and a lot more in the future. Over time a number of cloud service models have developed based on the kind of services and resources they provide, and a number of organizations are working actively to make the cloud safer for users. This project aims to develop a framework for classifying and determining the various risk factors and vulnerabilities affecting cloud computing deployments within various services models by harmonizing existing classifications by the Cloud Security Alliance (CSA) and the Open Web Application Security Project (OWASP) with other recommendations from industry experts in private, public and government sectors. Relevant tools for assessing vulnerabilities and risks will be used and other tools and utilities for the management and understanding of cloud security are expected to be developed over time. The project will start with cataloging and classifying a few security issues that currently exist in the domain from various users before being developed into a wider framework based on input from other researchers and interested parties [83].
Cloud Providers	Amazon Web Services; FutureGrid; Google Cloud Platform; Penguin Computing on Demand Indiana University; Red Cloud; Windows Azure
Special Features	Community datasets or collections; MapReduce; SQLaas
Dev. Environment	Eucalyptus; Nimbus; OpenStack; VMware
Use Regularity	Annually
Cores Used Peak	4
Cores Steady State	4
Core Hours in a Year	80
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Object store
Accessed During Run	1TB
Short-Term Storage	1TB
Long-Term Storage	2TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Providing basic access; research data sets or collections; survey data
Data Accessed By	Researcher; research group; department or institution; outside collaborators
Software	Community developed; open source; commercial
Problems/Limitations	"No problems yet, but this project is just about taking off fully and is being adapted for Health Informatics research. Some time is needed to gather new information, resources and data sets."
Cloud Funding	Institutional; commercial; personal and considering other sources
Research Funding	NSF; commercial; personal and considering other sources

CS: Security/Highly Assured Clouds – Co-Resident Watermarking

Project	Co-Resident Watermarking
Use Cases	Computer science research
Primary Researchers	Adam Bates, University of Oregon
Abstract	Virtualization is the cornerstone of cloud computing, allowing providers to instantiate multiple virtual machines on a single set of physical resources. Customers utilize cloud resources alongside unknown and untrusted parties, creating the co-resident threat: there is a possibility of unauthorized access to sensitive customer data through the exploitation of covert channels. Previous approaches to determining and exploiting co-residency require the ability to examine and manipulate internal hardware on these machines, behavior that can be patched or otherwise defended. We describe a new attack called co-resident watermarking that allows co-residents to inject a watermark into the network flow of a target instance. This watermark can be used to exfiltrate and broadcast co-residency data from the physical machine, compromising isolation without reliance on internal side channels. We evaluate co-resident watermarking under various network conditions and system configurations, showing co-residency can be determined in under 60 seconds and that a covert channel bitrate of 1.91 bps can be achieved. This work represents a first step in characterizing the co-resident watermarking threat [84].
Cloud Providers	FutureGrid
Special Features	Query physical node from within VM
Dev. Environment	Nimbus
Use Regularity	Weekly
Cores Used Peak	20
Cores Steady State	3
Core Hours in a Year	100
Storage Accessed For	Analysis
Preferred Storage	Elastic Block Storage
Accessed During Run	1GB
Short-Term Storage	1GB
Long-Term Storage	0
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Capabilities/Features	“One obstacle to performing research on science clouds is that the cloud abstraction can potentially mask important information, such as discovering the topography of our VMs within the datacenter. Fortunately, we were able to collaborate with Nimbus and the SDSC FutureGrid deployment to selectively expose this information. The change to the Nimbus codebase is available starting in cloud-client-21, and needs to be explicitly enabled by the cloud administrator.”
Cloud Funding	NSF
Research Funding	None

CS: Security/Highly Assured Clouds – Science of Cloud-Scale Computing

Project	CiC: Science of cloud-scale computing
Use Cases	Computer science research; domain-specific computing environments; event-driven real-time science
Primary Researchers	Kenneth Birman, Robbert van Renesse, and Hakim Weatherspoon, Cornell University
Abstract	Our use cases center on new platforms we are creating for highly assured cloud-scale computing in Cornell's Isis2 [85], GridControl [86], xCloud [87], and ShadowDB research.
Cloud Providers	Amazon Web Services; Red Cloud; Windows Azure; LLNL Computing Facilities (IaaS and Paas)
Special Features	Community datasets or collections
Dev. Environment	Eucalyptus; VMware
Use Regularity	Weekly
Cores Used Peak	25000
Cores Steady State	1000
Core Hours in a Year	100000
Storage Accessed For	Analysis
Preferred Storage	Object store
Accessed During Run	20GB
Short-Term Storage	500GB
Long-Term Storage	500GB
Data Moved Into Cloud	50GB
Data Moved Out Cloud	2GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	N/A
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Home-grown; open source; commercial
Capabilities/Features	"Our group does platform and infrastructure development. Your survey seems to focus much more on people who use existing platforms and infrastructure to curate and analyze data. But if the systems community is to create new and innovative cloud platforms, for example to address high assurance needs in the cloud, we need to be recognized more explicitly."
Problems/Limitations	"Yes, very much so. For our style of work we need really large numbers of cores for brief runs, to debug and test our solutions. Most cloud systems are optimized for long term use of resources in a steady-state but less ambitiously scaled manner. We would also find it desirable to have access to information about topology and node layouts, of the kind cloud providers use to build their own infrastructure solutions, but normally don't make accessible to their customers."
Cloud Funding	NSF; DOE; DOD; institutional; commercial
Research Funding	NSF; DOE; DOD

CS: Security/Highly Assured Clouds – Trusted Cloud Storage

Project	Compliance assurance services
Use Cases	Collaboration; data sharing, domain-specific computing environments
Primary Researchers	Shiping Chen, Commonwealth Scientific and Industrial Research Organization (CSIRO) ICT Centre, Australia
Abstract	This project aims to focus on both fundamental theories and practical technologies for services-based collaboration within and across organizations to ensure the collaborative services complying with the agreed business rules, SLA (Service Level Agreement) and/or government (domain) regulations.
Cloud Providers	Amazon Web Services; Windows Azure
Special Features	Secure data store and computation
Dev. Environments	VMware
Use Regularity	Weekly
Cores Used Peak	10
Cores Steady State	5
Core Hours in a Year	120
Storage Accessed For	Analysis
Preferred Storage	Database
Accessed During Run	5GB
Short-Term Storage	10GB
Long-Term Storage	100GB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	50GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Home-grown
Capabilities/Features	“Due to our application requirements, we'd like the cloud to provide secure data store and allow users to customize and copy their VM instances.”
Cloud Funding	Institutional
Research Funding	Internal

CS: Cloud Software Testing and Analysis – Bit Turner

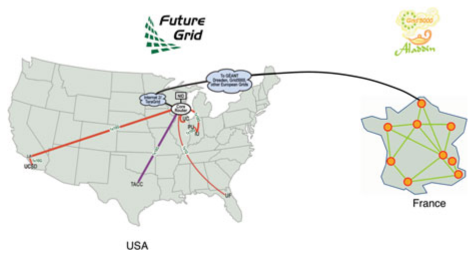
Project	Bit Turner
Use Cases	Computer science research
Primary Researchers	Pongsin Poosankam, University of California, Berkeley
Additional Researchers	Stephen McCamant, Dawn Song, Alex Bazhanyuk, Jimmy Su, and Dan Caselden, University of California, Berkeley
Abstract	Software testing is integral to the stability and security of software around the world. Among modern testing methods, fuzzing has withstood the test of time as an effective method of test generation and software analysis—especially for closed-source systems such as malware and off-the-shelf commercial software. However, existing fuzzing solutions have serious technical limitations and operational overhead. Analysts must train and monitor fuzzers, or the fuzzers will not properly exercise the system under test (SUT). Even with dedicated efforts to model the SUT, analysts may miss code paths such as paths to undocumented features or features that are only accessible due to execution fault. To obtain better coverage without the operational overhead required by common fuzzers, we integrated our symbolic execution tools BitFuzz and FuzzBALL into our distributed BitTurner solution. BitFuzz in essence uses dynamic traces to delve deep into the SUT's logic, and then queries a decision procedure with modified constraints to generate new inputs that cause the SUT to execute different branches. The process is iterative, where a generated input can be used instead of the seed input to further explore the SUT. FuzzBALL by contrast implements a symbolic interpreter, which can treat memory regions and registers as symbolic. As a result it does not rely on a seed input and can arbitrarily choose execution paths in the SUT. BitTurner uses BitFuzz, supplemented by features of FuzzBALL, on Amazon's cloud services to automatically explore, generate test cases for, and test for faults in software systems. Analysts simply upload their software via the BitTurner web portal, and BitTurner spawns EC2 instances that explore the uploaded software and pass test cases to concrete fuzzer instances. BitTurner periodically updates analysts with code coverage statistics and details of observed program faults [88].
Cloud Providers	Amazon Web Services
Use Regularity	Daily
Cores Used Peak	12
Cores Steady State	6
Core Hours in a Year	50400
Storage Accessed For	Analysis
Preferred Storage	Elastic Block Storage
Accessed During Run	3TB and 750GB
Short-Term Storage	3TB and 750GB
Long-Term Storage	3TB and 750GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	16GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Unknown
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Home-grown; open source
Cloud Funding	Commercial
Research Funding	Not specified

CS: Cloud Testing and Analysis – Developer Testing

Project	Developer testing of Azure cloud applications
Use Cases	Commonly requested software
Primary Researchers	Tao Xie, North Carolina State University
Abstract	Developer testing has been widely recognized as an important, valuable means of improving software reliability. However, manual developer testing is often tedious and not sufficient. Automated testing tools can be used to reduce manual testing efforts. This project develops a systematic framework for cooperative developer testing to enable effective, synergetic cooperation between developers and testing tools. This framework centers around test intentions (i.e., what testing goals to satisfy) and consists of four components: intention specification, test generation, test abstraction, and intention inference. The project also includes integrated research and educational plans [89].
Cloud Providers	Windows Azure
Use Regularity	Monthly
Cores Used Peak	5
Cores Steady State	2
Core Hours in a Year	100
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	4GB
Short-Term Storage	4GB
Long-Term Storage	10GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Open source; commercial
Cloud Funding	NSF
Research Funding	NSF

CS: Federated Clouds – FutureGrid and Grid’5000 Collaboration

Project	FutureGrid and Grid’5000 collaboration
Use Cases	Collaboration; computer science research
Primary Researchers	Mauricio Tsugawa, University of Florida
Abstract	This project investigates sky computing deployment across FutureGrid and Grid'5000 [90].
Cloud Providers	FutureGrid
Dev. Environment	Nimbus
Use Regularity	Daily
Cores Used Peak	1500
Cores Steady State	700
Core Hours in a Year	3000
Storage Accessed For	Analysis
Preferred Storage	Wide Area File Systems
Accessed During Run	100GB
Short-Term Storage	100GB
Long-Term Storage	100GB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	100GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Cloud Funding	NSF
Research Funding	NSF



FutureGrid and Grid’5000 test-beds used for Sky Computing research

CS: Federated Clouds – Scaling-Out CloudBLAST

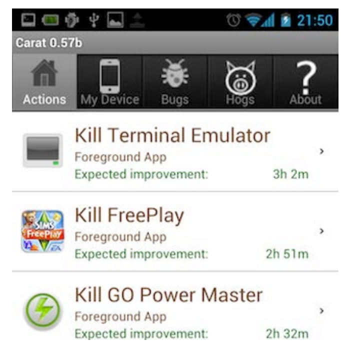
Project	Scaling-out CloudBLAST
Use Cases	Computer science research; computing and data analysis support for scientific workflows; data management and analysis; domain-specific computing environments
Primary Researchers	Andrea Matsunaga and Mauricio Tsugawa, University of Florida
Abstract	This project proposes and evaluates an approach to the parallelization, deployment and management of embarrassingly parallel bioinformatics applications (e.g., BLAST) that integrates several emerging technologies for distributed computing. In particular, it evaluates scaling-out applications on a geographically distributed system formed by resources from distinct cloud providers, which we refer to as sky-computing systems. Such environments are inherently disconnected and heterogeneous with respect to performance, requiring the combination and extension of several existing technologies to efficiently scale-out applications with respect to management and performance [91], [92], [93], [94].
Cloud Providers	FutureGrid; Grid'5000
Special Features	Virtual Networking (e.g., ViNe)
Use Regularity	Annually
Cores Used Peak	1500
Cores Steady State	1500
Core Hours in a Year	3000
Storage Accessed For	Analysis
Preferred Storage	HDFS
Accessed During Run	100GB
Short-Term Storage	1TB
Long-Term Storage	1TB
Data Moved Into Cloud	30GB
Data Moved Out Cloud	100GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; outside collaborators
Software	Home-grown; open source
Cloud Funding	NSF
Research Funding	NSF

CS: Federated Clouds – xCloud

Project	xCloud
Use Cases	Computer science research; educational, outreach, and training (EOT)
Primary Researchers	Hakim Weatherspoon, Cornell University
Abstract	Infrastructure-as-a-Service (IaaS) clouds are evolving from offering simple on-demand resources to providing diverse sets of tightly-coupled monolithic services. Like OS kernels of the 1980's and 1990's, these monolithic offerings, albeit rich in features, are significantly constraining users' freedom and control over the underlying---cloud---resources. For example, we are unaware of a true hybrid cloud, where its users can migrate virtual machines freely across clouds. In this research agenda, we investigate a new type of IaaS cloud, an xCloud that builds on ideas from extensible OSs to give users the flexibility to install custom cloud extensions, which can address the limitations outlined above. xClouds are very practical and can transform today's public clouds into xClouds [95].
Cloud Providers	Amazon Web Services; Red Cloud; Windows Azure
Special Features	Live migration
Dev. Environment	Eucalyptus
Use Regularity	Monthly
Cores Used Peak	100
Cores Steady State	1
Core Hours in a Year	50000
Storage Accessed For	Analysis
Preferred Storage	Object store
Accessed During Run	2TB
Short-Term Storage	2TB
Long-Term Storage	100GB
Data Moved Into Cloud	2TB
Data Moved Out Cloud	500GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections; not moving data, just programs
Data Accessed By	Researcher; research group
Software	Home-grown; open source
Capabilities/Features	"My research group mainly investigates the underlying systems for cloud computing as opposed to using cloud computing resources for other scientific research."
Problems/Limitations	"It is difficult to do systems research since a user does not have access to the underlying hypervisor. xCloud solves this problem with nested virtualization (i.e., adding another layer of virtualization)."
Cloud Funding	NSF
Research Funding	NSF; DOD

CS: Mobile Computing – Detecting/Diagnosing Energy Use in Mobile Devices

Project	Carat
Use Cases	Collaboration; computer science research; computing and data analysis support for scientific workflows; data management and analysis
Primary Researchers	Adam J. Oliner, Ion Stoica, and Anand P. Iyer, University of California, Berkeley; Eemil Lagerspetz and Sasu Tarkom, University of Helsinki
Abstract	We aim to detect and diagnose energy anomalies, abnormally heavy battery use. This paper describes a collaborative black-box method, and an implementation called Carat, for performing such diagnosis on mobile devices. A client app sends intermittent, coarse-grained measurements to a server, which identifies correlations between higher expected energy use and client properties like the running apps, device model, and operating system. The analysis quantifies the error and confidence associated with a diagnosis, suggests actions the user could take to improve battery life, and projects the amount of improvement. Carat detected all anomalies in a controlled experiment and, during a deployment to a community of more than 340,000 devices, identified thousands of energy anomalies in the wild. On average, a Carat user's battery life increased by 10% after 10 days [96].
Cloud Providers	Amazon Web Services
Use Regularity	Daily
Cores Used Peak	50
Cores Steady State	10
Core Hours in a Year	200000
Storage Accessed For	Analysis; reference
Preferred Storage	Object store
Accessed During Run	1TB
Short-Term Storage	1TB
Long-Term Storage	1TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Capabilities/Features	“Data resilience and resource scalability.”
Problems/Limitations	“Opacity of cost. We were occasionally surprised by how much we were spending on certain resources.”
Cloud Funding	NSF; commercial; DARPA; departmental
Research Funding	NSF; commercial; DARPA
Additional Notes	“The Carat server is a 1253-line Java application (excluding code auto-generated by Thrift) hosted on Amazon EC2, with mechanisms to scale by spawning new instances and to load-balance incoming connections. The data is stored in Amazon’s DynamoDB. The backend analysis is a 4K-line Scala program also running on EC2 [97].”



Screenshot of Carat in action

Engineering: Global Engineering from Supply Chains to High-Tech Design

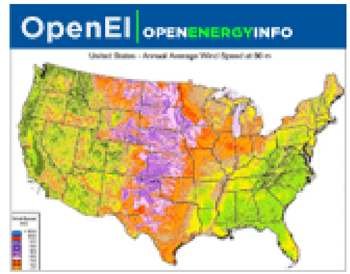
Project	Hardware accelerated clouds
Use Cases	Collaboration; computer science research; computing and data analysis support for scientific workflows; domain-specific computing environments
Primary Researchers	Theodore Omtzigt, Stillwater Supercomputing
Additional Researchers	Kitrick Sheets, KBS Software
Abstract	Middleware and workflow automation to leverage geographically dispersed supercomputer centers to maximize resource utilization and performance.
Cloud Providers	Amazon Web Services; Nimbix Hardware Accelerated Cloud
Special Features	Bare metal provisioning, community datasets or collections; GPUs; HBase; Hive; MapReduce; MPI; QDR InfiniBand; queues; SQLaaS; stateless blades; tables
Dev. Environment:	OpenStack
Use Regularity	Daily
Cores Used Peak	12000
Cores Steady State	128
Core Hours in a Year	999999
Storage Accessed For	Analysis
Preferred Storage	Parallel performance file system
Accessed During Run	50TB
Short-Term Storage	500TB
Long-Term Storage	200TB
Data Moved Into Cloud	50TB
Data Moved Out Cloud	10TB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 100Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Department or institution
Software	Home-grown; community developed; open source; commercial
Capabilities/Features	“FPGA accelerated servers.”
Problems/Limitations	“(1) latency of data movement between multiple sites, (2) clarity of the workflow, (3) collaboration between bare metal and virtual machine based clusters, (4) identity management, (5) storage management.”
Additional Notes	“The past 12 months, we have implemented a handful of global cloud platforms that connect US, EU, and APAC. The common impetus behind these projects is to connect brain trusts in these geographies. Whether they are supply chains in Asia program managed from the EU, healthcare cost improvements in the US by using radiologists in India, or high-tech design teams that are collaborating on a new car or smart phone design, all these efforts are trying to implement the IT platform to create the global village. The teachings provided by these implementations is that cloud computing is more or less a solved problem, but cloud collaboration is far from done. Cloud collaboration from an architecture point of view is similar to the constraints faced by mobile application platforms, so there is no doubt that in the next couple of years we'll see lots of nascent solutions to the fundamental problem of mobility and cloud collaboration: data movement. The data sets in our US-China project measured in the range from tens to hundreds of TBytes, but data expansion was modest at a couple of GBytes a day. For a medical cloud computing project, the data set was more modest at 35TBytes, but the data expansion of these data sets could be as high as 100GB per day, fueled by high volume instruments, such as MRI or NGS machines. In the US-China collaboration, the problem was network latency and packet loss, whereas in the medical cloud computing project,

	the problem was how to deal with multi-site high-volume data expansions. The cloud computing aspect of all these projects was literally less than a couple of man weeks' worth of work. The cloud collaboration aspect of these projects all required completely new technology developments [98]."
Cloud Funding	Departmental; institutional; commercial
Research Funding	Commercial

Energy Sciences: Energy Science Gateway

Project	OpenEI.org – Open Energy Information Initiative
Use Cases	Burst resources; collaboration; commonly requested software; computer science research; computing and data analysis support for scientific workflows; data archiving; data management and analysis; data sharing; education, outreach, and training (EOT); science gateways
Primary Researchers	Debbie Brodt-Giles, Jon Weers and Ryan McKeel, National Renewable Energy Laboratory
Abstract	Open Energy Information [99] is a platform designed to be the world's most comprehensive, open, and collaborative energy information network—supplying powerful data to decision makers and supporting a global energy transformation. The platform is developed by the National Renewable Energy Laboratory (NREL), but is intended for the world's contribution, collaboration, and participation. The platform provides a means for DOE and its laboratories to share energy data and information while addressing the White House directive to be open, participatory, and collaborative with open government data. Although much of the world's energy-related information and data are available as resources on the Internet, they are dispersed among innumerable individuals and organizations, available in widely disparate formats, and highly variable in quality and usefulness. This creates a major challenge for: (1) researchers, who need to share data to accelerate innovation; (2) consumers, who need to have timely, accessible data to make day-to-day decisions; (3) policy makers, who need to research effective solutions based on technology capabilities, resource availability, market needs and effective incentives; and, (4) entrepreneurs and application developers, who need to perform due diligence and market assessments based on real data. OpenEI provides a solution using its open-source Web platform—similar to the one used by Wikipedia. The platform provides large amounts of energy-related data, information, APIs, and other web services which can be easily searched, accessed, and used both by people and automated machine processes. NREL developed OpenEI using the standards and practices of the Linked Open Data community, which makes the platform much more robust and powerful than typical Web sites and databases. As an open platform, all users can search, edit, add, and access data in OpenEI — for free. The user community contributes the content and ensures its accuracy and relevance; as the community expands, so does the comprehensiveness and quality of the content. The data are structured and tagged with descriptors to enable cross-linking among related data sets, advanced search functionality, and consistent, usable formatting. Data input protocols and quality standards help ensure the content is structured and described properly and derived from a credible source. Although DOE/NREL is developing OpenEI and seeding it with initial data, it is designed to become a true community model with millions of users, a large core of active contributors, and numerous sponsors. The linked open data within OpenEI will have countless benefits because the platform links energy communities and decision makers (including policymakers, researchers, technology investors, venture capitalists, and market professionals) with valuable energy data, information, analyses, tools, images, maps, and other resources. By providing access to the best available data, OpenEI may help decision makers reduce missteps and save time and money. Through this improved sharing of energy information, we also can benefit from the acceleration of energy

	technology research and a transformation to a clean, secure energy future [100].
Cloud Providers	Amazon Web Services
Use Regularity	Daily
Cores Used Peak	137
Cores Steady State	125
Core Hours in a Year	1000000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Elastic Block Storage
Accessed During Run	0
Short-Term Storage	2TB
Long-Term Storage	5TB
Data Moved Into Cloud	530GB
Data Moved Out Cloud	2TB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Providing basic access; research data sets or collections; survey data
Data Accessed By	Any users may access the data collection and survey results
Software	Home-grown; community developed; open source; commercial
Additional Notes	“We have an international audience, and we need our system to be reliable and available to all our users on a 24/7 basis. As our platform grows, we anticipate very large datasets to be contributed, so being able to scale quickly is important….Key platform software for OpenEI includes Apache, Semantic MediaWiki, MySQL, and OpenLink Virtuoso. Customization to meet the specific needs of OpenEI has been performed primarily through PHP. Common deployment and operations for OpenEI have been automated using various AWS command-line tools [101].”
Cloud Funding	DOE
Research Funding	DOE



OpenEI uses the cloud to link the world’s energy information and data

Environmental Sciences: Hydrology Modeling

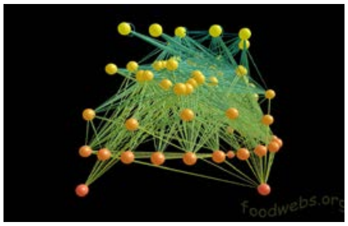
Project	Using the cloud to model and manage large watershed systems
Use Cases	Burst resources; collaboration; data sharing
Primary Researchers	Marty Humphrey, University of Virginia; Jon Goodall, University of South Carolina
Abstract	Understanding hydrologic systems at the scale of large watersheds is of critical importance to society when faced with extreme events such as floods and droughts, or with minimizing human impacts on water quality. Climate change and increasing population are further complicating watershed-scale prediction by placing additional stress and uncertainty on future hydrologic system conditions. New data collection and management approaches are allowing models to capture water flow through built and natural environments at an increasing level of detail. A significant barrier to advancing hydrologic science and water resource management is insufficient computational infrastructure to leverage these existing and future data resources within simulation models. We were awarded a National Science Foundation “Computing in the Cloud” grant to advance hydrologic science and water resource management by leveraging cloud computing for modeling large watershed systems. We use Windows Azure in three ways. First, we have created a cloud-enabled hydrologic model. Second, we are improving the process of hydrologic model parameterization by creating cloud-based data processing workflows. Third, in Windows Azure, we are applying the model and data processing tool to a large watershed in order to address a relevant hydrologic research question related to quantifying impacts of climate change on water resources.
Cloud Providers	Amazon Web Services; Windows Azure
Use Regularity	Weekly
Cores Used Peak	256
Cores Steady State	16
Core Hours in a Year	5000
Storage Accessed For	Analysis; reference
Preferred Storage	Windows Azure storage
Accessed During Run	5TB
Short-Term Storage	5TB
Long-Term Storage	1TB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	100GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; outside collaborators
Software	Home-grown; commercial
Capabilities/Features	“The usual – e.g., the ability to shut down everything at night.”
Problems/Limitations	“Justifying paying for them on NSF grants (in general).”
Additional Notes	“Next-generation hydrology modeling will be increasingly sophisticated, encompassing a wide range of natural phenomena. Furthermore, calibrating models will soon cease to be practically feasible on desktop computers….we have presented the design, implementation, and evaluation of a cloud-based system for watershed model calibration. With a representative watershed model whose calibration takes 11.4 hours on a commodity laptop, our cloud-based system (Windows Azure) calibrates the watershed model in 43.32 minutes using 16 cloud cores (15.78x speedup), 11.76 minutes using 64 cloud cores (58.13x speedup), and 5.03 minutes using 256 cloud cores (135.89x speedup).

Watershed modeling in the cloud enables more experiments

	We believe that such speed-ups we achieve in our cloud-based watershed model calibration system offer the potential toward real-time interactive model creation with continuous calibration, ushering in a new paradigm for watershed modeling [102]."
Cloud Funding	NSF
Research Funding	NSF

Environmental Sciences: Web Portal for Ecological Network Simulations and Analysis

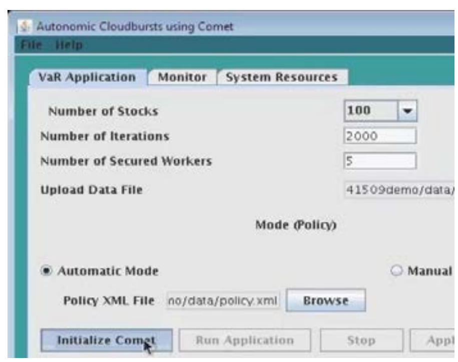
Project	Network3D – WoW (Webs on the Web)
Use Cases	Collaboration; data sharing; domain-specific computing environments; education, outreach, and training (EOT)
Primary Researchers	Jennifer Dunne, Sante Fe Institute; Neo Martinez, PEaCE Lab; Rich Williams, Microsoft Research
Additional Researchers	Paul Yoon, PEaCE Lab
Abstract	A project to develop an Internet knowledge base of food webs, which describe the network of who eats whom in ecological communities. Combined with integrated analytical, modeling, and 3D visualization tools, the Network3D knowledge base will increase the ability of scientists, policy makers, and students to exchange and analyze information about the structure, function, and dynamics of ecological networks. Network3D content will initially focus on trophic interactions among organisms as well as species bioenergetic parameters that enable the modeling of ecological dynamics. Webs on the Web ecoinformatics tools are designed to facilitate the research efforts of ecologists as well as an increasingly broad array of scientists across disciplines who are interested in network theory relating to biological and non-biological systems. In this spirit, the knowledge base design will be extensible to other types of ecological interactions and other types of networks. We are also developing biocomplexity educational modules for all levels of learning using Network3D tools.
Cloud Providers	Windows Azure
Special Features	Community datasets or collections; GPUs
Use Regularity	Daily
Cores Used Peak	14496
Cores Steady State	24
Core Hours in a Year	115968
Storage Accessed For	Analysis; reference; archival
Preferred Storage	N/A
Accessed During Run	1TB
Short-Term Storage	1TB
Long-Term Storage	5TB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections; not moving data, just programs
Data Accessed By	Researcher; research group; department or institution; outside collaborators; any users may access data collections and survey results
Software	Community developed; open source; commercial
Additional Notes	"We have yet to thoroughly evaluate the effectiveness of our implementation. It is clear that users with less computational expertise are more able to conduct computational research using our portal. However, it is not clear whether the amount of time developing and maintaining the portal is worth the additional functionality it provides….The Azure platform may provide the best platform for conducting our research but results are significantly delayed by initial development time [103]."
Cloud Funding	NSF
Research Funding	NSF



Visualization of coral reef food web at Virgin Islands shelf complex

Finance: Financial Mathematics

Project	Monte-Carlo Value-at-Risk computations
Use Cases	Burst resources; data sharing; domain-specific computing environments; event-driven real-time science; science gateways
Primary Researchers	Hyunjoo Kim, Xerox Research Center; Manish Parashar and Moustafa AbdelBaky, Rutgers University
Abstract	In today's turbulent market conditions, the ability to generate accurate and timely risk measures has become critical to operating successfully, and necessary for survival. Value-at-Risk (VaR) is a market standard risk measure used by senior management and regulators to quantify the risk level of a firm's holdings. However, the time-critical nature and dynamic computational workloads of VaR applications make it essential for computing infrastructures to handle bursts in computing and storage resources needs. This requires on-demand scalability, dynamic provisioning, and the integration of distributed resources. While emerging utility computing services and clouds have the potential for cost-effectively supporting such spikes in resource requirements, integrating clouds with computing platforms and data centers, as well as developing and managing applications to utilize the platform remains a challenge. In this work, we focused on two main goals: (1) to investigate the feasibility of using cloud computing services to support the dynamic requirements of online risk analytics, as well as (2) to demonstrate the ability of the CometCloud autonomic computing engine to provide programming and runtime infrastructure support to enable these applications to seamlessly and safely scale-out (and scale-in) from in-house private datacenters to Internet clouds such as the Amazon EC2, based on the dynamic computational load. We demonstrated how the CometCloud autonomic computing engine can support online multi-resolution VaR analytics using and integration of private and Internet cloud resources.
Cloud Providers	Amazon Web Services; Open Science Grid
Dev. Environment	CometCloud; Nimbus
Use Regularity	Monthly
Cores Used Peak	1000
Cores Steady State	100
Core Hours in a Year	3000
Storage Accessed For	Analysis; archival
Preferred Storage	Elastic Block Storage
Accessed During Run	1GB
Short-Term Storage	2GB
Long-Term Storage	20GB
Data Moved Into Cloud	0
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher; research group
Software	Home-grown; open source
Additional Notes	"The goal of autonomic cloud bursts is to seamlessly (and securely) integrate private enterprise clouds and datacenters with public utility clouds on-demand, to provide an abstraction of resizable computing capacity. It enables the dynamic deployment of application components, which typically run on internal organizational compute resources, onto a public cloud to address dynamic workloads, spikes in demands, and other extreme requirements. Furthermore, given the increasing application and infrastructure scales, as well as their cooling, operation



Getting VaR inputs and selecting cloudbursts

	and management costs, typical over-provisioning strategies are no longer feasible [104]. CometCloud supports policy-driven, robust autonomic cloud bridging and autonomic cloudbursts. CometPortal provides an interface for monitoring and controlling application deployment using CometCLoud, specifying and modifying policies controlling scale-out based on load dynamics, performance requirements, and/or economic constraints [105].”
Cloud Funding	NSF
Research Funding	NSF; DOE; commercial

Genetics and Bioinformatics: Bioinformatics Computing

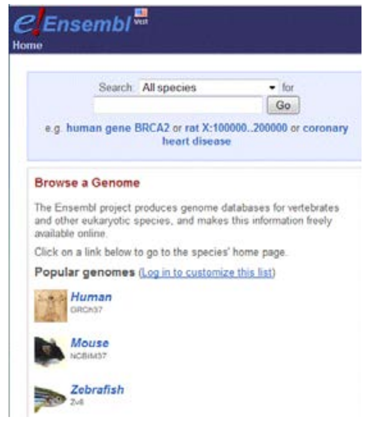
Project	Cloud BioLinux
Use Cases	Computing and data analysis support for scientific workflows
Primary Researchers	Konstantinos Krampis, J. Craig Venter Institute
Abstract	Cloud BioLinux is a publicly accessible Virtual Machine (VM) that enables scientists to quickly provision on-demand infrastructures for high-performance bioinformatics computing using cloud platforms. Users have instant access to a range of pre-configured command line and graphical software applications, including a full-featured desktop interface, documentation and over 135 bioinformatics packages for applications including sequence alignment, clustering, assembly, display, editing, and phylogeny. Each tool's functionality is fully described in the documentation directly accessible from the graphical interface of the VM. Besides the Amazon EC2 cloud, we have started instances of Cloud BioLinux on a private Eucalyptus cloud installed at the J. Craig Venter Institute, and demonstrated access to the bioinformatic tools interface through a remote connection to EC2 instances from a local desktop computer. Documentation for using Cloud BioLinux on EC2 is available from our project website, while a Eucalyptus cloud image and VirtualBox Appliance is also publicly available for download and use by researchers with access to private clouds [106].
Cloud Providers	Amazon Web Services
Use Features	Community datasets or collections; MapReduce
Dev. Environment	Eucalyptus; VirtualBox
Use Regularity	Daily
Cores Used Peak	64
Cores Steady State	8
Core Hours in a Year	1024
Storage Accessed For	Analysis
Preferred Storage	Elastic Block Storage
Accessed During Run	1TB
Short-Term Storage	2TB
Long-Term Storage	5TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 100Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Additional Notes	“A steep drop in the cost of next-generation sequencing during recent years has made the technology affordable to the majority of researchers, but downstream bioinformatic analysis still poses a resource bottleneck for small laboratories and institutes that do not have access to substantial computational resources. Sequencing instruments are typically bundled with only the minimal processing and storage capacity required for data capture during sequencing runs …. We can enable researchers without access to local computing clusters to perform large-scale data analysis, by tapping into a pool of on-demand Cloud BioLinux VMs that can be rented at low cost starting from \$0.085 per hour for a single core/1.7 GB memory RAM/160 GB of storage VM (early 2012 pricing) and up to \$2 for VMs with 8 cores/64 GB of RAM/1.68 TB of storage based on Amazon EC2 pricing, and are available worldwide and independently of institutional, economic or national boundaries …. Virtual

Cloud BioLinux on Amazon EC2 cloud console

	Machines (VMs) that run on cloud computing platforms are an alternative to in-house informatics infrastructures for bioinformatic data analysis, requiring minimal set-up and no up-front hardware costs. Renting servers on the cloud can work as a better model for smaller research laboratories, where the cost for hardware and data center maintenance, cannot be justified to support only a few experiments. Using VMs allows for snapshots of the computing server to be taken, including the operating system and software, input data files configured settings and analysis results. The VM snapshots can be shared among collaborating researchers using a commercial cloud platform such as Amazon EC2, open source clouds including Eucalyptus or OpenStack, or desktop virtualization software like VirtualBox. Snapshots are an ideal approach for reproducibility of in-silico analyses, given that bioinformatics research involves small but important configuration changes while working with the different tools and datasets. These include for example tuning algorithm parameters in software installations, or making ad-hoc modifications to software for specific data processing cases, which are otherwise difficult to capture and share among collaborators [107]."
Cloud Funding	NIH
Research Funding	NIH

Genetics and Bioinformatics: Distributing Genome Annotation Data

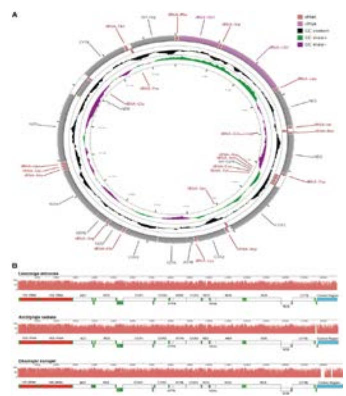
Project	Ensembl
Use Cases	Burst resources; collaboration; computing and data analysis support for scientific workflows; data sharing
Primary Researchers	Stephen Keenan, European Bioinformatics Institute
Abstract	Using the cloud to distribute Ensembl Genome annotation. The Ensembl project provides genome resources for chordate genomes with focus on human genome data and data for key model organisms such as mouse, rat, and zebrafish [108].
Cloud Providers	Amazon Web Services
Dev. Environment	Eucalyptus; Nimbus
Use Regularity	Daily
Cores Used Peak	400
Cores Steady State	30
Core Hours in a Year	262800
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Elastic Block Storage
Accessed During Run	0
Short-Term Storage	0
Long-Term Storage	20TB
Data Moved Into Cloud	6TB
Data Moved Out Cloud	500GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; any users may access my data collections and survey results
Software	Home-grown; open source
Additional Notes	"This year (2012) saw the release of a third mirror of the Ensembl website in the Asia-Pacific region located at http://asia.ensembl.org. As with our other mirrors at http://useast.ensembl.org and http://uswest.ensembl.org, the Asia mirror uses AWS to provide the infrastructure (the USWest mirror was migrated to AWS Northern CA data centre in 2011). By consolidating all of the supported Ensembl mirrors in AWS, we are able to provide consistent support and increased performance for users around the world. All users visiting the Ensembl website are automatically redirected to their nearest mirror, ensuring the best possible performance. For users accessing Ensembl data via our API or direct MySQL queries, we have also launched a second database server at useastdb.ensembl.org [109]....For each mirror, the architecture is identical and uses several Amazon Elastic Compute Cloud (Amazon EC2) technologies. The website sits behind Amazon Elastic Load Balancing (Amazon ELB) and has two load-balanced Apache Web Server instances, although this number can be increased using Auto Scaling, if necessary. The web server nodes talk to a MySQL database running on a separate AWS instance backed by a couple of Amazon Elastic Block Store (Amazon EBS) volumes. We have another MySQL instance that backs our Biomart tool. A separate instance is used for collecting log data from the other nodes and as an endpoint for our VPN. We use Amazon Simple Storage Service (Amazon S3) for backups and snapshotting and also to distribute the Ensembl data as part of the Amazon Public Data Sets initiative. Search functionality is handled by an instance that runs our Apache Lucene-based search server [110]."
Cloud Funding	Institutional
Research Funding	Not specified



Global distribution of genome data

Genetics and Bioinformatics: Sharing a Data Center

Project	Collaborative research: North East Cyberinfrastructure Consortium
Use Cases	Burst resources; collaboration; commonly requested software; data archiving; data sharing; education, outreach, and training (EOT)
Primary Researchers	James Vincent, University of Vermont
Abstract	Under the North East Cyberinfrastructure Consortium, the bioinformatics cores of the five partner states have formed a virtual organization, the North East Bioinformatics Collaborative (NEBC), to develop collaborative activities such as shared workflows and promote the development of protocols for a new Shared Data Center for the movement, life cycle management, storage and recovery of data that are simultaneously viewed/analyzed/worked on by multiple users across the region [111}. We have implemented the Shared Data Center in a cloud infrastructure (Amazon) and have begun developing on-demand, cloud enabled workflows. We would like to extend this work to encompass directly NSF resources such as FutureGrid.
Cloud Providers	Amazon Web Services; FutureGrid; Globus Online
Special Features	Community datasets or collections
Dev. Environment	Eucalyptus
Use Regularity	Weekly
Cores Used Peak	100
Cores Steady State	8
Core Hours in a Year	20000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Elastic Block Storage
Accessed During Run	1TB
Short-Term Storage	5TB
Long-Term Storage	20TB
Data Moved Into Cloud	5TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Capabilities/Features	"The ability to instantiate clusters on demand with software/environments specific to the analysis at hand enhances research productivity."
Problems/Limitations	"Overall cost and charging to a grant that does not have such a cost model built in are challenges."
Cloud Funding	NSF; NIH
Research Funding	NSF; NIH



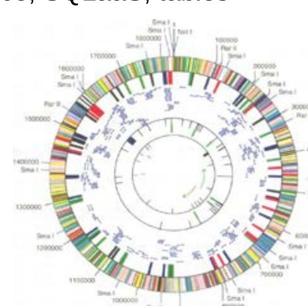
Shared data centers in the cloud may increase collaboration

Genomics and Bioinformatics: Streaming Next-Generation Sequences

Project	Streaming and compression approaches to next-generation sequences
Use Cases	Burst resources; collaboration; data archiving; data sharing; education, outreach, and training (EOT)
Primary Researchers	Titus Brown, Michigan State University
Abstract	In recent years, next-generation DNA sequencing capacity has completely outstripped our ability to computationally digest the resulting volume of data. Driven by the need to actually analyze the data, our lab has developed a suite of novel data structures and algorithms for graph compression and data reduction [112]; in addition to being very efficient on their own, our approaches make use of probabilistic data structures that enable substantially lower memory usage than the best possible exact approach. Using these approaches we have been able to scale de novo data assembly approaches down to cloud computing infrastructure, and we have also completed some of the largest de novo assemblies of metagenomes ever done [113]. Last but not least, these approaches show the way to essentially infinite de novo assembly of environmental microbial data.
Cloud Providers	Amazon Web Services
Use Regularity	Weekly
Cores Used Peak	10
Cores Steady State	1
Core Hours in a Year	5000
Storage Accessed For	Analysis; archival
Preferred Storage	Object store
Accessed During Run	10TB
Short-Term Storage	10TB
Long-Term Storage	5TB
Data Moved Into Cloud	2TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; outside collaborators; any users may access my data collections and survey results
Software	Home-grown; open source
Problems/Limitations	“RAM limitations -- I need more than the maximum provided by Amazon (and most cloud providers). 300GB+ needed.”
Additional Notes	See also the benefits of teaching a next-generation sequence analysis course using the cloud [114].
Cloud Funding	NSF; NIH; personal; departmental
Research Funding	NSF; NIH; DOE

Genomics and Bioinformatics: Predicting Transaction Factor Binding Sites

Project	Large scale prediction of transcription factor binding sites
Use Cases	Data archiving; data management and analysis; domain-specific computing environments; education, outreach, and training (EOT)
Primary Researchers	Zhengchang Su, Ehsan Tabari, Afshan Jalali, Sirinvas Akella and Vikas Gandham, The University of North Carolina at Charlotte
Abstract	Although tremendous advances have been made in identifying the gene-coding DNA sequences in bacterial genomes using computational methods, our understanding of regulatory DNA sequences is very limited due to the lack of efficient computational methods for predicting them. Regulatory sequences specify when, how much, and where the genes should be expressed in the cell through their interactions with small proteins called transcription factors (TFs). Therefore, identifying these sequences, also called TF binding sites (TFBS), in a genome is as important as identifying gene-coding sequences for understanding the biology of the cell. Rapid recent advances in genome sequencing technology are dramatically reducing the time and cost of sequencing a genome. Over 1,500 bacterial genomes have been sequenced and this number is rising exponentially. Our very limited understanding of the gene regulatory systems in sequenced prokaryotic genomes has largely hindered our understanding of their biology and applications in renewable energy production and environment protection as well as the prevention of the diseases they cause. To fill in this gap, we have recently developed an efficient and accurate algorithm for predicting TFBSs in a group of related genomes, and have parallelized it on an in-house cluster using MPI. Although this algorithm can potentially predict TFBSs in a few thousand genomes, its capability will soon be dwarfed by the sequencing of hundreds of thousands genomes as a result of the on-going world-wide efforts to sample various microbiomes using new sequencing technologies. Cloud computing holds promise to overcome the computational and storage challenges for predicting TFBSs in all sequenced genomes in the future. I will present our preliminary results to port our algorithm on the Microsoft Azure Cloud Platform as an attempt to achieve such a goal.
Cloud Providers	Window Azure
Special Features	Community datasets or collections; MapReduce; SQLaaS; tables
Use Regularity	Weekly
Cores Used Peak	500
Cores Steady State	100
Core Hours in a Year	200000
Storage Accessed For	Analysis
Preferred Storage	Elastic Block Storage
Accessed During Run	20TB
Short-Term Storage	2TB
Long-Term Storage	10TB
Data Moved Into Cloud	2TB and 500GB
Data Moved Out Cloud	2TB and 500GB
BW In/Out of Cloud	Up to 100Gb/s
BW to Storage Within	Up to 100Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group; any users may access my data collections and survey results
Software	Home-grown; community developed; open source
Additional Notes	“Like many other large scientific problems, our problem could also be solved on a large supercomputer; however, we currently do not have

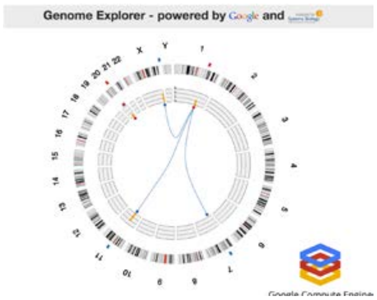


Cloud holds promise for meeting growing TFBS prediction needs

	access to a larger supercomputer. Furthermore, a supercomputer can be used by hundreds of users, and sometimes the priority to run large problems can be low, meaning a long waiting time. In addition, based on our experiences of programming on Hadoop, MPI and Azure frameworks, it is much easier to develop our solution on Azure. Although Hadoop or MPI provides plenty of APIs to encapsulate network operations and job scheduling, those APIs are not uniform and reliable compared with Azure's very high level of abstraction. Moreover, Azure also provides an efficient interaction framework (web roles) with web users, which large supercomputer solutions are hard to achieve. Without easy-to-use interfaces for users, the system can only be used by experts who know implementation details. Especially for our project, we intend to provide this system to biological researchers who may not be familiar with programming implementations. Compared with supercomputer solutions, Azure is more suitable for our project. Azure also seems to be more cost-effective as the median amount of cloud resource consumed by a group is about $25,000 for 2011. If we had a way to request this amount from our funding agency for cloud resources in the future, we will very glad to do that, because in a foreseeable future with $25,000 hardware, we cannot conduct the scale of computation that we are currently doing. However, we do face challenges of working with the Azure cloud resource. In particular, in our computational pipeline, we rely on several third-party programs; and we have difficulty to port some of them on the Windows platform, so we have to seek alternative solutions [115].
Cloud Funding	NSF; Microsoft Research
Research Funding	NSF

Genomics and Bioinformatics: The Cancer Genome Atlas (TCGA) Cloud Demonstration

Project	TCGA cloud compute engine demonstration
Use Cases	Burst resources; collaboration; data management and analysis; event-driven real-time science
Primary Researchers	Ilya Shmulevich and Hector Rovira, Institute for Systems Biology (ISB)
Abstract	The Institute for Systems Biology explores the latest in web and enterprise technologies for use within research collaborations in computational biology. Cloud technologies are often used to scale the computational and data resources available to a project. We also have a number of large-scale family genome projects that require use of cloud resources to securely distribute data to researchers. For the TCGA project we have developed web applications and visualizations using the latest HTML5 standards [116]. These software applications are integrated with cloud technologies to provide users with the ability to explore the data in a rich, interactive, and dynamic environment [117].
Cloud Providers	Amazon Web Services; Google Cloud Platform (Google Compute Engine)
Special Features	SQLaaS; tables; compute instances; web application platform
Regularity	Weekly
Cores Used Peak	1000
Cores Steady State	20
Core Hours in a Year	10000
Storage Accessed For	Analysis; archival; reference
Preferred Storage	Object store
Accessed During Run	100GB
Short-Term Storage	10TB
Long-Term Storage	50TB
Data Moved Into Cloud	5TB
Data Moved Out Cloud	200GB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 100Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group; department or institution; outside collaborators; any users may access data collections and survey results
Software	Home-grown; community developed; open source
Feature/Capabilities	“Security through open standards like OpenID and OAuth.”
Problems/Limitations	“Limitations in terms of patient identifiable data.”
Additional Notes:	“…Rovira and Shmulevich…started analyzing their data (on Google Compute Engine) in February 2012 with help from Google’s Computational Discovery Department. The institute sends the Google team data sets containing publically available clinical information and genomic measurement from the project’s patient population-for example, information on DNA mutations in a cancer cell. Google then loaded data into Compute Engine, and the analysis helps guide the institute’s research. The Google team also analyzed ISB’s data using Exacycle, an experimental Google system that also offers researcher fast, large-scale data analysis…The system has analyzed a cancer data set in two hours, compared with 15 hours on the institute’s internal system [119].
Cloud Funding	Institutional
Research Funding	NIH; ITMI



Genome Explorer application scaled to 600,000 cores (“Google Compute Engine [118]”)

Genomics and Bioinformatics: 1000 Genomes

Project	Bioinformatics and cyberinfrastructure project
Use Cases	Computing and data analysis support for scientific workflows; data archiving; data management and analysis; education, outreach and training (EOT)
Primary Researchers	Andrew Younge, Indiana University
Abstract	Recent improvements in sequencing technology ("next-gen" sequencing platforms) have sharply reduced the cost of sequencing. The 1000 Genomes Project [120] is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation. The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied. While recent work has been conducted towards sequence alignment and nucleotide matching, there is a large need for protein sequencing and comparison between the 697 currently sequenced datasets. This project will look at the protein synthesis at a low level order to identify differences between members of the population, which can hopefully lead to a better understanding of how proteins differ between individuals.
Cloud Providers	FutureGrid
Special Features	Community datasets or collections
Dev. Environment	Eucalyptus
Regularity	Daily
Cores Used Peak	100
Cores Steady State	10
Core Hours in a Year	1000
Storage Accessed For	Analysis
Preferred Storage	Elastic Block Storage
Accessed During Run	50GB
Short-Term Storage	1TB
Long-Term Storage	1TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	2GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher, research group
Software	Home-grown; community developed; open source
Problems/Limitations	"I wish EBS volumes would work more like Lustre file systems, i.e., high performance, high availability, and the ability for multiple VMs to read/write to one EBS volume."
Cloud Funding	NSF
Research Funding	None

Genomics and Bioinformatics: Transcriptomic Assembly of Algae

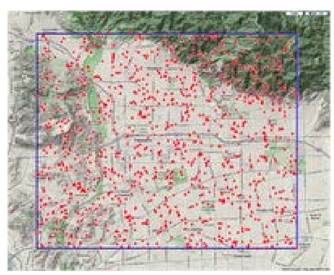
Project	Transcriptomic assembly of diverse green algae
Use Cases	Burst resources; data archiving; data management and analysis; data sharing
Primary Researchers	Charles F. Delwiche, University of Maryland
Additional Researchers	Edymion Cooper; Bastian Bentlage and Theodore Gibbons, University of Maryland
Abstract	We are engaged in the assembly and analysis of deep transcriptomes obtained by Illumina sequencing of mRNA from organisms for which reference genomes are not available. In some cases we are assembling metatranscriptomes (i.e., transcriptomes that are derived from more than one organism). The DeBruijn graph assemblers that are currently available require large (100GB – 1TB) memory spaces to run, and scale with the complexity of the dataset (such that metatranscriptomes require substantially more memory than single transcriptomes). We believe that cloud resources would be the best way of completing the more difficult analyses, but we have not yet identified a really appropriate resource [121].
Cloud Providers	Amazon Web Services; Google Cloud Platform
Special Features	GPUs, queues
Dev. Environment	VMware
Regularity	Annually
Cores Used Peak	48
Cores Steady State	4
Core Hours in a Year	100
Storage Accessed For	Analysis; reference; archival
Preferred Storage	HDFS
Accessed During Run	4TB
Short-Term Storage	8TB
Long-Term Storage	2TB
Data Moved Into Cloud	2TB
Data Moved Out Cloud	2TB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Providing basic access; research data sets or collections
Data Accessed By	Researcher, research group; outside collaborators
Software	Community-developed
Feature/Capabilities	"Because we have only intermittent need for high performance computing, it would be highly beneficial if we could move our more intense computation to the cloud, because it would minimize the in-house computing resources we need to maintain. We are still in search of the ideal solution."
Problems/Limitations	"Our use of the cloud is still developmental. Right now we have two major problems: (1) most cloud computing resources do not have sufficiently large-memory resources for our uses (100GB and larger), and (2) pricing has not been favorable for our applications. We are still searching for a really suitable cloud solution.
Cloud Funding	NSF
Research Funding	NSF

Geographic Information Science: GIS Analysis

Project	Crayons: A cloud based parallel framework for GIS overlay operations
Use Cases	Domain-specific computing environments
Primary Researchers	Dinesh Agarwal, Georgia State University
Additional Researchers	Sushil Prasad, Georgia State University
Abstract	GIS vector-based spatial data overlay processing through cloud computing is much more complex and challenging than raster data processing because raster data is based on regular grid-based fixed-size pixels, while vector features have irregular geometric shapes represented by a list of large number of vertices. The GIS data files can be huge and their overlay processing is computationally intensive. The emerging Cloud platforms such as Azure, with their potential for large scale computing and storage capabilities, easy accessibility by common users and scientists, availability on demand, easy maintenance, sustainability, and portability, has the promise to be the platform of choice for such GIS applications. We propose to discover distributed algorithms and their scalable implementations for GIS overlay processing on Azure platform. Meager amount of work has been done on large volume of vector geospatial data processing through parallel/distributed computing (as opposed to for raster data processing), and none on cloud platforms. The existing parallel approaches mostly developed in the 1990s are not scalable and/or limited to small set of polygons on the traditional cluster and other platforms. Better algorithms are clearly needed. The discovery and implementation of new methods for the analysis of geospatial data on cloud platform will dramatically improve the efficiency of disaster modeling and consequently enable the relevant agencies (such as FEMA) to implement emergency mitigation, preparedness and response plans more effectively. We envisage that the geospatial analytical methods derived in this research will contribute to mainstream GIS software, and spatial applications employing cloud computing in allied disciplines [122].
Cloud Providers	Windows Azure
Special Features	Queues, tables
Regularity	Weekly
Cores Used Peak	100
Cores Steady State	100
Core Hours in a Year	100000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Elastic block storage
Accessed During Run	1TB
Short-Term Storage	1TB
Long-Term Storage	1TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	N/A
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed
Cloud Funding	Commercial
Research Funding	NSF

Geosciences: Seismic Network

Project	Community Seismic Network
Use Cases	Burst resources; collaboration; data archiving; data management and analysis; data sharing; event-driven real-time science
Researcher	Michael Olson, California Institute of Technology
Abstract	Community Seismic Network (CSN) is a new earthquake monitoring system based on a dense array of low-cost acceleration sensors. A primary goal of the system is to produce block-by-block measurements of strong shaking during an earthquake. Such "shake maps" can then be used by first responder agencies (e.g., fire department, utilities) to prioritize dispatch to areas of greatest likely damage. Effective emergency response can occur despite damaged telephone services that prevent civilian calls for help from succeeding. Volunteers from greater Pasadena, CA host a small seismometer in their homes or offices for several years. Volunteers connect the sensor to their own computer, download an application from the CSN website, and are immediately part of the data collection network. CSN data is used during an earthquake to provide very high resolution data on actual ground shaking in real time to first responders. Longer term, it enables scientific construction of 3D geologic models of the ground underneath the sensors, which will influence land use policy and construction codes. CSN will add one thousand community-based sensors with automatic reporting of shaking data to a remote computing service, and produce shake maps within minutes of the onset of an event. CSN will keep producing updated maps through the lifetime of the event and beyond, until it receives no additional data. The diagram shows an example random distribution of one thousand stations across greater Pasadena. Even if the reference stations were at each corner of the diagram, there would be no detailed knowledge of the actual shaking occurring in the rectangle itself: existing shake map tools would apply a summary of the limited knowledge that exists today about the subsurface geology of Pasadena, and make crude estimates of the behavior between stations. In contrast, with hundreds of times the number of accelerometer stations deployed across Pasadena, the average distance between stations drops from about 10 miles to a quarter mile. This is a crucial change in density, as significant (unexpected) variations in a 10 mile distance have been seen in recent California earthquakes [123].
Cloud Providers	Google Cloud Platform
Special Features	MapReduce; queues
Use Regularity	Daily
Cores Used Peak	1
Cores Steady State	1
Core Hours in a Year	1
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Object store
Accessed During Run	100GB
Short-Term Storage	100GB
Long-Term Storage	100GB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	100GB
BW In/Out of Cloud	Up to 100Gb/s
BW to Storage Within	N/A
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; department or institution; outside collaborators



Sample random distribution of 1000 in-home seismic sensors

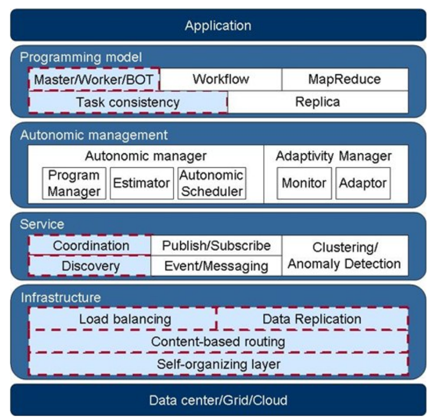
Software	Home-grown
Additional Notes	"CSN's pursuit of high sensor densities leads to one of its key design characteristics: scalability. Google App Engine is used because of its ability to scale in small amounts of time from using minimal resources to consuming large amounts of resources. During quiescent periods the only data sent on the network is control traffic, which amounts to very little; however, the data sent during seismic events is substantial….The biggest impact on system performance is the occurrence of loading requests, which occur when a request causes a new instance to be created to serve it.…. Error rates are another important factor. The most common type of error caused by App Engine's environment is deadline exceeded errors. These occur when requests are terminated for exceeding the processing deadline imposed by App Engine. The extreme variability in the processing of a request cannot be reasonably attributed to developer code, but rather to conditions within the cloud system. This is one side effect of sharing servers….In conclusion, we find that while PaaS applications in general and Google App Engine in particular can fulfill the needs of cyber-physical systems, it's important to pay close attention to the design characteristics of the chosen platform. The performance implications of even seemingly small or obvious choices can make substantial differences in how applications behave in the long term [124]."
Cloud Funding	NSF; commercial
Research Funding	NSF

Geosciences: Reservoir Characterization Application Workflow

Project	EnKF based history-matching for oil reservoir characterization application workflow
Use Cases	Burst resources; domain-specific computing environments; event-driven real-time science
Primary Researchers	Hyunjoo Kim, Xerox Research Center; Yakoub el Khamra, University of Texas at Austin; Shantenu Jha, Rutgers University; Manish Parashar, Rutgers University
Abstract	Clouds are rapidly joining high-performance Grids as viable computational platforms for scientific exploration and discovery, and it is clear that production computational infrastructures will integrate both these paradigms in the near future. As a result, understanding usage modes that are meaningful in such a hybrid infrastructure is critical. We used CometCloud to explore meaningful usage modes for a hybrid HPC plus Cloud infrastructure. In particular, we used a reservoir characterization application workflow, which uses the EnKF for history matching as the driving application, and we complemented TeraGrid resources with Amazon EC2 public Cloud instances. We explored 5 different usage modes: (1) acceleration – using Clouds as accelerators to reduce the application time to completion, for example given budget constraints, (2) conservation – using Clouds to conserve HPC allocations, within the appropriate runtime and budget constraints,(3) resilience – using Clouds to handle unexpected situations such as an unanticipated HPC downtime, inadequate allocations, unanticipated queue delays or failures of working nodes, while meeting user objectives, (4) Cloud bursting – using Clouds to perform the actual computation in the Cloud if this is more effective than moving the data to HPC resources, and (5) analytics/visualization – using Clouds to perform data analytics or visualization at the same time that complex simulations are run in HPC resources [125], [126].
Cloud Providers	Amazon Web Services
Special Features	CometCloud
Use Regularity	Monthly
Cores Used Peak	100
Cores Steady State	10
Core Hours in a Year	1000
Storage Accessed For	Analysis; archival
Preferred Storage	Object store
Accessed During Run	1GB
Short-Term Storage	1GB
Long-Term Storage	10GB
Data Moved Into Cloud	0
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Outside collaborators
Software	Home-grown; open source
Cloud Funding	NSF
Research Funding	NSF; DOE; commercial

Geosciences: Scalable Oil Reservoir Simulations

Project	Scalable ensemble-based oil reservoir simulations
Use Cases	Burst resources; collaboration; commonly requested software; domain-specific computing environments; event-driven real-time science; science gateways
Primary Researchers	Moustafa AbdelBaky and Manish Parashar, Rutgers University
Abstract	In an early experiment we explored how a Cloud abstraction can be effectively used to provide a simple interface for current HPC resources and support real-world applications. In particular, we experimentally validated the benefits of the Cloud paradigm, such as ease of use and dynamic allocation, and their application to supercomputers, specifically, on an IBM Blue Gene/P system. The CometCloud-based framework essentially transformed Blue Gene/P into an elastic Cloud, bridged multiple Blue Gene/P systems to create a larger HPC federated Cloud, and supported dynamic provisioning. We used the framework for an oil-reservoir data assimilation and history matching application, which consisted of the EnKF workflow with multiple reservoir instances. The exercise demonstrated the ease-of-use of the elastic as-a-service Cloud abstraction, and its effectiveness in improving utilization. This experiment was demonstrated at the 4th IEEE SCALE Challenge, and was awarded first place. During the experiment, Blue Gene/P resources varied from 640 to 22,016 processors, spanning across two Blue Gene systems in two different continents [127], [128].
Provider	IBM Blue Gene P
Dev. Environment	CometCloud
Use Regularity	Monthly
Cores Used Peak	22000
Cores Steady State	1000
Core Hours in a Year	80000
Storage Accessed For	Analysis; archival
Preferred Storage	Parallel performance file system
Accessed During Run	10GB
Short-Term Storage	500GB
Long-Term Storage	1TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	500GB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 100Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; department or institution; outside collaborators
Software	Home-grown; open source; commercial
Cloud Funding	None
Research Funding	NSF; DOE; commercial



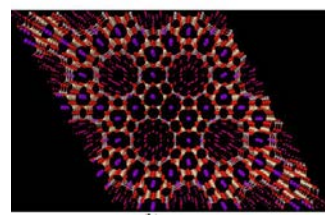
CometCloud framework for federated multi-clouds (clouds, HPC-grids & clusters)

Industrial Engineering: Modeling Supply Chains

Project	Supply chain network simulator using cloud computing
Use Cases	Collaboration; commonly requested software; computing and data analysis support for scientific workflows; education, outreach, and training (EOT)
Primary Researchers	Manuel Rossetti, University of Arkansas
Additional Researchers	Yaohua Chen, University of Arkansas
Abstract	Large-scale supply chains usually consist of thousands of stock keep units (SKUs) stocked at different locations within the supply chain. The purpose of this project is to develop a prototype software program that can allow the simulation of large-scale multi-echelon, multi-item supply networks using cloud-computing resources. These simulations are essentially compute-intensive Monte-Carlo experiments requiring multiple replications. Replications will be distributed across virtual machines within cloud architecture.
Cloud Providers	FutureGrid
Dev. Environment	Nimbus; VMware
Use Regularity	Monthly
Cores Used Peak	1
Cores Steady State	1
Core Hours in a Year	50
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	0
Short-Term Storage	1GB
Long-Term Storage	1GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher
Software	Home-grown; open source
Capabilities/Features	“Less command line oriented. Will make educating students easier.”
Cloud Funding	NSF; institutional
Research Funding	NSF; Center for Excellence in Logistics and Distribution (CELDi), an NSF I/UCRC

Materials Science: Computational Materials Science

Project	SAMP: Structure-Adaptive Materials Prediction
Use Cases	Computing and data analysis support for scientific workflows; data management and analysis
Primary Researchers	Estela Blaisten-Barojas, George Mason University
Additional Researchers	Qi Xing, George Mason University
Abstract	Cloud computing is attracting the attention of the scientific community. In this paper, we develop a new cloud-based computing system in the Windows Azure platform that allows users to use the Zeolite Structure Predictor (ZSP) model through a Web browser. The ZSP is a novel machine learning approach for classifying zeolite crystals according to their framework types. The ZSP can categorize entries from the Inorganic Crystal Structure Database into 41 framework types. The novel automated system permits a user to calculate the vector of descriptors used by ZSP and to apply the model using the Random Forest™ algorithm for classifying the input zeolite entries. The workflow presented here integrates executables in Fortran and Python for number crunching with packages such a Weka for data analytics and Jmol for Web-based atomistic visualization in an interactive compute system accessed through the Web. The compute system is robust and easy to use. Communities of scientists, engineers, and students knowledgeable in Windows-based computing should find this new workflow attractive and easy to be implemented in scientific scenarios in which the developer needs to combine heterogeneous components.
Cloud Providers	Windows Azure
Special Features	Queues, tables
Dev. Environment	Nimbus; VMware
Use Regularity	Daily
Cores Used Peak	20
Cores Steady State	20
Core Hours in a Year	7920
Storage Accessed For	Analysis
Preferred Storage	Windows Azure drive
Accessed During Run	50GB
Short-Term Storage	202GB
Long-Term Storage	202GB
Data Moved Into Cloud	3GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Home-grown; open source; commercial
Problems/Limitations	"Third party runtime libraries for java and Visual C++….Our biggest challenge is the need to create a *compute system* for each number crunching project. In turn, this implies to have a full time person creating such system so that computations could be carried on. This is a very expensive investment in human resources, investment that we cannot price in dollar amount, and that we cannot expect NSF to be able to afford for every future project. For this specific project, in addition of the time paid to an employee, there are uncountable PIs (principal investigator) hours invested, that will be fully lost when a new project would need to be implemented again in the Azure platform. While it is very simple to port our codes to a Linux cluster that has 20 cores, use of the Azure cloud requires the complete creation of a system for only then

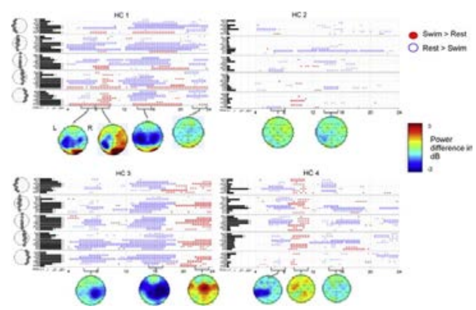


Atomic rendering of the zeolite supercell

	be able to start production. Any of our students can port our codes to a cluster. However, for the programming paradigm in Azure we need a special person, with knowledge of the Windows/Windows Server/ASP.net, etc. environments….Lack of secure access (secure shell, ssh) to load/access your own space is not efficient. Lack of *any* of the tools regularly accessible in a supercomputing center (compilers, libraries, environments, etc.) is a drawback. Lack of advanced compilers compatible with Windows makes our applications slower and more bulky …. Technical support is basically inexistent [129].
Additional Notes	"The absence of science and engineering consumers using public clouds is recognized by organizations such as the US National Science Foundation. This organization funds fundamental research and can adopt a pay-per-use funding mechanism, if the sciences, engineering, and mathematics communities embrace the cloud computing paradigm, a large number of educational and small-to-medium research laboratories would benefit. Cloud computing differentiates from grid computing because instead of batch job queues, the user receives virtual resources. Of particular importance for scientific research where numerical accuracy is important is that cloud computing offers deployment and control of applications, thus reducing compatibility issues between the application and the hosting environment. However, for cloud computing to become efficient for a given science application, a specific-to-problem computer system workflow needs to be created to link the user application with the IT cloud resources … (Our) cloud-based compute system developed is easily generalizable. It suffices to change the services (codes to be executed) and other science and engineering applications that manipulate data can use this cloud compute system. Such applications are usually data intensive and computationally intensive. Both will benefit by the parallelization scheme that supports the SAMP compute system. In particular, the researcher community that employs classical and quantum scientific open source packages for atomistic simulations (LAMMPS, NAMD, SIESTA, CPMD, among others) will find that the SAMP compute system allows access to resources in the WA cloud that otherwise might be difficult to procure with local hardware [130].
Cloud Funding	Microsoft Research
Research Funding	NSF

Neuroscience: Electroencephalography (EEG) Data Analysis

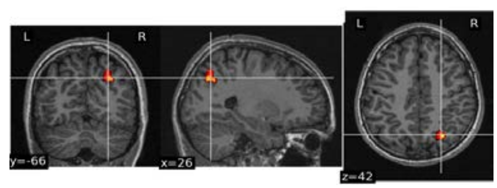
Project	EEG for determination of consciousness
Use Cases	Burst resources
Primary Researchers	Andrew M. Goldfine and Nicholas D. Schiff, Weill Cornell Medical College
Abstract	Andrew Goldfine is a neurorehabilatation neurologist interested in enhancing recovery after brain injury through modulation of the brain's arousal and arousal regulation networks. He is currently working in the labs of Nicholas Schiff [131] and Jonathan Victor at Weill Cornell Medical College, studying the pathophysiology of disorders of consciousness and the use of neurophysiological tools to track recovery of movement and large-scale cerebral networks. His current work is on using EEG (electroencephalography) to determine the presence of consciousness in patients who are unable to communicate, as well as using EEG and wireless accelerometer to understand the role of arousal regulation in motor performance in patients with diffuse brain injury [132]. He needed to run an analysis (permutation test) on a large dataset a large number of times. It would have taken 2 weeks on his laptop but took only two hours on Red Cloud. He hasn't used the system other than for that one big analysis, though he might in the future.
Cloud Providers	Globus Online; Red Cloud
Use Regularity	Annually
Cores Used Peak	52
Cores Steady State	1
Core Hours in a Year	1000
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	0
Short-Term Storage	0
Long-Term Storage	1TB
Data Moved Into Cloud	0
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher, research group
Software	Home-grown; commercial
Capabilities/Features	"It was nice to have so many cores to run my code in parallel. Excellent support staff (though limited on the weekends)."
Problems/Limitations	"It took a lot of effort to get it set up, though the help was good. The great majority of my code runs pretty fast so rarely do I need to do such a large project."
Cloud Funding	NIH
Research Funding	NIH



Power spectral analysis of EEG (1 run of motor imagery task)

Neuroscience: Neuroimaging and Genetic Data Analysis

Project	A-brain
Use Cases	Burst resources; science gateways
Primary Researchers	Radu Tudoran, INRIA Rennes, France
Additional Researchers	Gabreil Antoniu, INRIA Rennes; Bertrand Thiron, INRIA Saclay; Goetz Brasche, EMIC
Abstract	Joint genetic and neuroimaging data analysis on large cohorts of subjects is a new approach used to assess and understand the variability that exists between individuals. This approach has remained poorly understood so far and brings forward very significant challenges, as progress in this field can open pioneering directions in biology and medicine. As both neuroimaging- and genetic-domain observations represent a huge amount of variables (of the order of millions), performing statistically rigorous analyses on such amounts of data represents a computational challenge that cannot be addressed with conventional computational techniques. In this project, we explore cloud computing techniques to address the above computational challenge. The project relies on Microsoft's Azure cloud platform and leverages the complementary expertise of KerData in the area of scalable cloud data management and Parietal team (Saclay) in the field of neuroimaging [133].
Cloud Providers	Grid'5000; Windows Azure
Special Features	MapReduce; Map-IterativeReduce; queues
Dev. Environment	Azure
Use Regularity	Daily
Cores Used Peak	1000
Cores Steady State	350
Core Hours in a Year	70000
Storage Accessed For	Analysis
Preferred Storage	TomusBlobs [134]
Accessed During Run	10TB
Short-Term Storage	10GB
Long-Term Storage	100GB
Data Moved Into Cloud	10GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Research group
Software	Home-grown; open source; TomusMapReduce; TomusBlobs; MapIterative Reduce; Venus-C
Capabilities/Features	"Scalability."
Problems/Limitations	"Network bandwidth bottlenecks."
Additional Notes	"As both neuroimaging- and genetic-domain observations represent a huge amount of variables (of the order of 106), performing statistically rigorous analyses on such amounts of data represents a computational challenge that cannot be addressed with conventional computational techniques. On one hand, sophisticated regression techniques need to be used in order to perform sensitive analysis on these large datasets; on the other hand, the cost entailed by parameter optimization and statistical validation procedures (e.g. permutation tests). However, the computational framework can easily by run in parallel [135]."
Cloud Funding	Microsoft
Research Funding	Microsoft; INRIA-Microsoft



A-Brain explores gene and brain characteristic relationships

Operations Research: Simulation Optimization

Project	Decision-theoretic methods in simulation optimization
Use Cases	Burst resources; commonly request software
Primary Researchers	Peter I. Frazier and Jing Xie, Cornell University
Additional Researchers	Stephen E. Chick, INSEAD
Abstract	The research objective of the proposed work is to provide new algorithms for simulation optimization and related problems with good average-case performance. Simulation optimization is the practice of optimizing or calibrating a stochastic simulator, and is of critical importance in many simulation applications. Existing algorithms can be difficult to use in a way that provides consistently high-quality solutions. The algorithms and analysis resulting from this proposed research will improve our ability to optimize and calibrate a variety of simulations from within operations research, but also simulations from other engineering fields and in the natural sciences. Examples are calibrating a model of climate change, accurately reconstructing a collection of whole genomes from fragmented genetic data, or setting the right schedule or staffing level for a large organization such as a hospital [136].
Cloud Providers	Red Cloud
Use Regularity	Weekly
Cores Used Peak	96
Cores Steady State	12
Core Hours in a Year	120000
Storage Accessed For	Analysis; reference
Preferred Storage	N/A
Accessed During Run	25GB
Short-Term Storage	25GB
Long-Term Storage	25GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	N/A
Type Data Moving	Research data sets or collections
Accessed By	Researcher; research group; outside collaborators
Software	Home-grown; community developed; open source; commercial; dacefit; MATLAB; software research group developed
Capabilities/Features	“Allows me to not have to worry about maintaining a cluster; accelerate the design and testing of algorithms by using a compute/analysis resource that “bursts” on demand; and, run Parallel Computing Toolbox codes on an optimal number of cores in the cloud (using MATLAB Distributed Computing Server) rather than procure dedicated hardware/software for only periodic use.”
Additional Note	“The on demand convenience of Red Cloud with MATLAB is ideal for our work in sequential decision-making and optimal methods for collecting information. It provides the software we need when we need it, enabling us to develop simulation optimization and feasibility determination algorithms faster and more efficiently. We are not burdened with procuring and maintaining our own computational resources and can share the cloud resource with other researchers, providing economies of scale for all. We look forward to continuing to improve our use of Bayesian statistics and dynamic programming using Red Cloud with MATLAB [137].”
Cloud Funding	DOD
Research Funding	DOD

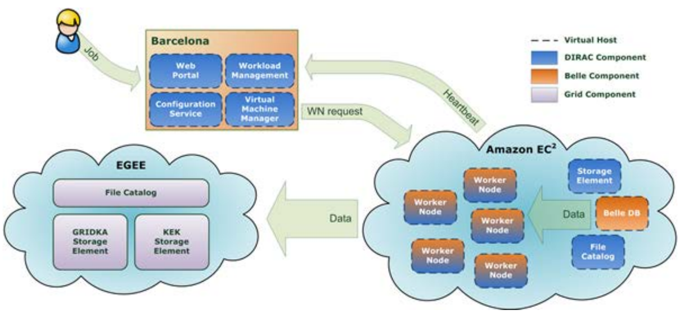
Plant Pathology: Citrus Greening Science Gateway

Project	Citrus greening community resource
Use Cases	Data sharing; science gateways
Primary Researchers	Surya Saha and Magdalen Lindeberg, Cornell University
Abstract	Citrus greening (also known as Huanglongbing or HLB) is a devastating agricultural disease threatening citrus production. Genome sequence data provide a critically important foundation for characterization of candidate virulence factors, understanding the nutritional requirements and basic physiology, and identification of regions in the sequence suitable for diagnostic probe development. The citrus greening website [138] is a community resource designed for dissemination of genome sequence data and related analyses of organisms associated with citrus greening (HLB) with emphasis on the genus Ca. Liberibacter. Analyses provided here are currently derived from publically available sequence data and can be accessed via the GBrowse genome viewer, with additional data and links found at "other genome resources."
Cloud Providers	Red Cloud
Dev. Environment	Eucalyptus
Use Regularity	Daily
Cores Used Peak	1
Cores Steady State	1
Core Hours in a Year	8760
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	50GB
Short-Term Storage	50GB
Long-Term Storage	50GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Uncertain
Type Data Moving	Research data sets or collections
Data Accessed By	Any users may access the data collection and survey results
Software	Community developed; open source
Additional Notes	"Leading plant pathologists selected Red Cloud to host the community genome assembly and analysis resource for Ca. Liberibacter asiaticus (Las). Las is an alpha-proteobacteria vectored by psyllid insects and believed to be the causal agent of citrus greening, a devastating agricultural disease threatening citrus production in Florida and other regions throughout the world. Red Cloud was selected to enable the scientific community to access this genome resource quickly without researchers having to procure, deploy, and maintain their own data server. Cornell CAC consultants helped deploy the application [139].
Cloud Funding	Citrus Research and Development Foundation
Research Funding	Citrus Research and Development Foundation

Citrus disease gateway was launched quickly in the cloud

Physics: Particle Physics – Belle Experiment

Project	Belle MC production on Amazon Web Services
Use Cases	Burst resources
Primary Researchers	Martin Sevior, University of Melbourne
Abstract	The joint Barcelona-Melbourne team is using the DIRAC distributed computing software framework to define and steer the execution of a sizable part of Belle Experiment simulation needs for their data reprocessing using computing resources on Amazon Elastic Compute Cloud (EC2). The team is using Amazon EC2 as a supplement to its existing large-scale grid computing infrastructure [140].
Cloud Providers	Amazon Web Services
Use Regularity	Annually
Cores Used Peak	800
Cores Steady State	1
Core Hours in a Year	1600
Storage Accessed For	Analysis
Preferred Storage	Object store
Accessed During Run	1TB
Short-Term Storage	1TB
Long-Term Storage	1TB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; department or institution; outside collaborators
Software	Open source
Cloud Funding	Not specified
Research Funding	Not specified
Additional Notes	“We particularly liked the flexibility to build exactly the virtual machines we needed and to transfer data to and from every instance we created. This flexibility and openness allowed us to rapidly deploy the sophisticated collection of programs needed for the Belle experiment and to integrate the results into the world-wide grid. Consequently we were able to accelerate our joint effort with researchers across the world to build exactly the application we needed. We’re very interested in seeing how far EC2 scales ….The advantage of using cloud is that we also find that our load CPU demand over a year isn’t constant. There are peaks and there are troughs. If we priced our purchase to satisfy our peak needs, we’d find that our system would lay idle for some fraction of the year [141].”



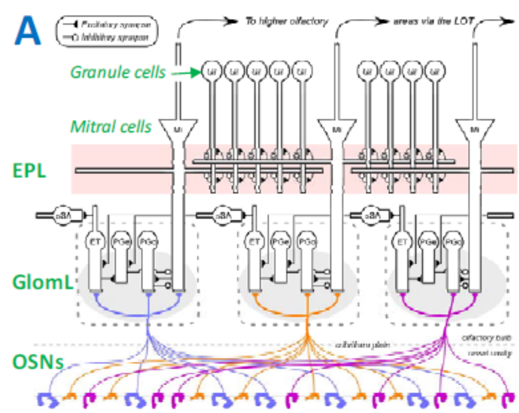
Belle Experiment scientific workflow

Physics: Particle Physics – ATLAS LHC

Project	Particle physics data analysis cluster for ATLAS LHC experiment
Use Cases	Computing and data analysis support for scientific workflows; data management and analysis
Primary Researchers	Doug Benjamin, Duke University
Abstract	This activity will study the ability to establish, configure and run as small analysis cluster for particle physics data analysis from the ATLAS experiment at the Large Hadron Collider (LHC). Such a cluster includes interactive part for data analysis visualization and batch component for larger scale throughput prior to visualization.
Cloud Providers	FutureGrid
Dev. Environment	Nimbus, OpenStack
Use Regularity	Weekly
Cores Used Peak	10
Cores Steady State	4
Core Hours in a Year	1000
Storage Accessed For	Analysis
Preferred Storage	Wide Area File System
Accessed During Run	100GB
Short-Term Storage	1TB and 25GB
Long-Term Storage	10TB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	10GB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Cloud Funding:	DOE
Research Funding	DOE
Additional Notes	“Intend to investigate the use of cloud resources to emulate the behavior of particle physics analysis cluster found at Universities. Data taken from the ATLAS experiment at the LHC will be analyzed [142].”

Physiology and Biophysics: Modeling the Olfactory System

Project	Neural computation
Use Cases	Collaboration; commonly requested software; data archiving; data management and analysis
Primary Researchers	Thomas A. Cleland, Cornell University
Abstract	To run large-scale models of biological neural networks based on the membrane and circuit properties of neurons in the brain. The long-term goal is to understand the complex interactions in the brain that underlie cognitive processes. For example, the olfactory bulb is a physically segregated region of the cerebral cortex that acquires and processes sensory information about odor, and also learns from experience. It also is a convenient microcosm of the larger brain for understanding the mechanisms of learning and memory in the brain, and how they adapt to the statistics of personal experience. We are studying the neural circuitry of the olfactory bulb, focusing on how these "wetware" circuits construct representations of odors, how these learned representations adapt according to experience, and how cellular and circuit mechanisms determine the form and longevity of the resulting memory.
Cloud Providers	Red Cloud
Special Features	Parallel MATLAB
Use Regularity	Daily
Cores Used Peak	12
Cores Steady State	12
Core Hours in a Year	4000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	N/A
Accessed During Run	50GB
Short-Term Storage	50GB
Long-Term Storage	50GB
Data Moved Into Cloud	50GB
Data Moved Out Cloud	50GB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Providing basic access; research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; open source; commercial; MATLAB; Python; NEURON
Additional Notes	"Reverse-engineering neural circuitry requires a great deal of exploratory modeling, for which interactivity and ease of use are priorities. Although occasionally it is necessary to set up large parameter searches, it is more typical that many simulations of moderate complexity need to be performed interactively. These simulations often are too large to execute effectively on desktop workstations (requiring hours to days to weeks to complete), but can be completed in an interactive timeframe (minutes to hours) on Red Cloud with MATLAB. The results from these moderately complex simulations then often guide the construction of larger-scale simulations for which efficient parallelization and high-end resources are absolute necessities. 'Our need for computational power is substantial but uneven. Computing in the cloud with Red Cloud with MATLAB and leveraging other CAC computational resources when we need them is an ideal solution for us and enables us to work effectively without assuming the complex burden of cluster hosting and maintenance [143].'"
Cloud Funding	NIH
Research Funding	NIH



Olfactory bulb network with sensory neurons

Physiology and Biophysics: Simulating Muscle Dynamics

Project	Computational simulations of muscle dynamics
Use Cases	Burst resources; data management and analysis
Primary Researchers	C. David Williams (now at Harvard) and Tom Daniel, University of Washington
Abstract	We seek to discover and understand how the spatial arrangement of molecular motors in muscle controls, or fails to control, the force which muscle generates. In our work, we treat the proteins that constitute muscle as a series of springs, arranged in a three-dimensional network, whose connection pattern is governed by protein interactions. This allows us to change the spatial configuration of muscle's proteins in the same ways which occur in vivo (within the cell) and observe the changes in generated force. To find the force generated by our muscle simulation, we have to ensure that it reaches a steady state at each time step. Such a steady state exists when all the interior points of the model (those which don't connect it to the outside world) have no net force upon them. Finding this steady state is a large non-linear root-finding problem that requires substantial computation.
Cloud Providers	Amazon Web Services
Special Features	MapReduce
Use Regularity	Monthly
Cores Used Peak	1200
Cores Steady State	1
Core Hours in a Year	300000
Storage Accessed For	Analysis; reference
Preferred Storage	Object store
Accessed During Run	4GB
Short-Term Storage	0
Long-Term Storage	2TB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	3GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher; research group
Software	Home-grown; open source
Additional Notes	"…(This) research on muscle contraction simulation involved hundreds of thousands of independent calculations that are not dependent on each other.…Williams described the challenge of cloud computing as being the time and expertise needed to configure the 'cloud cluster.' Every time you create what is called a 'machine image,' it's comparable to a new cluster that must be reconfigured and made to talk to each other or the local computer. The start-up, programming, and configuration are more challenging than an in-house local cluster, However, Williams claims it isn't difficult to learn [144]."
Cloud Funding	NSF; commercial
Research Funding	NSF; NIH

Systems Engineering: Instructional Website

Project	Instructional website
Use Cases	Education, outreach, and training (EOT)
Primary Researchers	Peter Jackson, Cornell University
Abstract	Professor Jackson is active in educational curriculum development for operations research and systems engineering. He is the recipient of several awards for curriculum innovation. He is now using the cloud to research and develop web-based educational experiences for *Model-Based Systems Engineering*, an online textbook that will guide students in the use of the Systems Modeling Language (SYsML).
Cloud Providers	Red Cloud
Dev. Environment	Eucalyptus
Use Regularity	Weekly
Cores Used Peak	1
Cores Steady State	1
Core Hours in a Year	800
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	1GB
Short-Term Storage	1GB
Long-Term Storage	1GB
Data Moved Into Cloud	1GB
Data Moved Out Cloud	1GB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher
Software	Open source
Additional Notes	"I used Red Cloud because I wanted to have a personal Linux 'machine' that would act as a web server so that I could learn web development for the Linux platform. Red Cloud provided an inexpensive way to do it. I would not have attempted using MediaWiki had I stuck with my PC platform. I learned about Apache, PHP, MediaWiki, and more. I continue to use Red Cloud. I am assembling a collection of open source tools to support further educational development: Calliope for optimization formulations, Octave for MATLAB-type programming and more. My research may lead me to use Red Cloud with MATLAB for parallel processing but I am still developing the basic MATLAB code [145]."
Cloud Funding	School of Operations Research and Engineering budget
Research Funding	None

MODEL-BASED SYSTEMS ENGINEERING

Using the cloud to develop an online engineering textbook

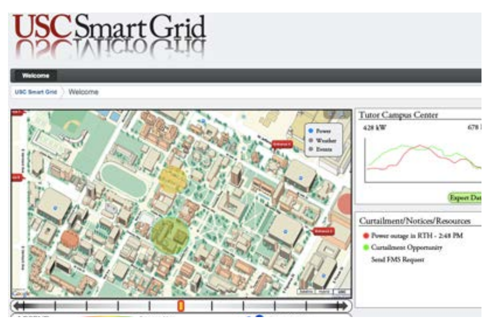
Humanities, Arts, and Social Sciences Cloud Projects (HASS) Surveyed: Complete Data

Cross-HASS: Data Repository

Project	Shared digital repository
Use Cases	Collaboration; data management and analysis; data sharing
Primary Researchers	Patrick Burns, Colorado State University
Abstract	This is a shared digital repository serving seven institutions of higher education in Colorado. All types of data, programs, protocols, from all institutions. We use Google Apps for Education email and unified messaging cloud.
Cloud Providers	Google Cloud Platform
Special Features	Data management
Dev. Environment	Standard digital repository services
Use Regularity	Weekly
Cores Used Peak	2048
Cores Steady State	383
Cores Hours in a Year	10000
Storage Accessed For	Analysis; archival
Preferred Storage	Parallel performance file system
Accessed During Run	1GB
Short-Term Storage	1TB
Long-Term Storage	3TB
Data Moved Into Cloud	2TB
Data Moved Out Cloud	1TB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group; department of institution; outside collaborators
Software	Home-grown
Problems/Limitations	“Need a preservation infrastructure.”
Cloud Funding	Institutional
Research Funding	NSF; NIH; DOE; DOD; commercial; institutional

Economics: Energy Informatics

Project	Software architecture for demand response optimization
Use Cases	Burst resources; collaboration; data sharing; event-driven real-time science; science gateways
Primary Researchers	Yogesh Simmhan, University of Southern California
Additional Researchers	Viktor Prasanna, University of Southern California
Abstract	The Smart Grid group is conducting research into informatics-driven scalable software architecture on Cloud infrastructure to address real time power management in the domain of Smart Power Grids. Demand Response Optimization focuses on enabling electricity customers to conserve their energy consumption during peak demand periods and relieve stress on the power grid to ensure its resilience. As part of the Los Angeles Smart Grid Demonstration project, we are investigating different curtailment strategies for the USC Campus Microgrid that will inform the wider city's service area. In particular, we are investigating the use of enhanced data collection capabilities from sensors and smart meters on the USC campus to offer deeper visibility into the real time power usage patterns and intelligent selection of voluntary and direct control strategies. Our informatics approach uses advanced forecasting and data analytics for performing Dynamic Demand Response (D2R) in the USC Campus Microgrid that can scale to the city. Energy informatics lies at the cusp of information technology, power systems, and social behavior domains, and is an emerging area of critical importance to global sustainability. This cyberphysical system (CPS) offers unique challenges to existing computer science algorithms, approaches and frameworks due to the data complexity, application dynamism, massive scale, and need for real time and resilient response. Some of the research topics being explored in this project include semantic information integration, complex event and stream processing, data mining and machine learning, data security and privacy, and public and private Cloud computing platforms, with scalability being a central theme [146], [147].
Cloud Providers	Amazon Web Services; FutureGrid; Windows Azure
Special Features	MapReduce; queues; tables
Dev. Environment	Eucalyptus; OpenStack
Use Regularity	Daily
Cores Used Peak	256
Cores Steady State	32
Core Hours in a Year	320000
Storage Accessed For	Analysis; reference; archival
Preferred Storage	Object store
Accessed During Run	500GB
Short-Term Storage	1TB
Long-Term Storage	500GB
Data Moved Into Cloud	1TB
Data Moved Out Cloud	250GB
BW In/Out of Cloud	Up to 1Gb/s
BW to Storage Within	Up to 1Gb/s
Type Data Moving	Research data sets or collections; survey data
Data Accessed By	Researcher; research group; outside collaborators
Software	Home-grown; community developed; open source
Additional Notes	“The decision on whether to acquire Cloud resources or physical hardware through agency funding will be a function of the applications



Researching scalable software in a cloud infrastructure for real-time power management

	that are being run and the users they serve. For highly performance driven application that operate on a tightly coupled model, purchasing and managing a rack with ~50 cores is a better model than Cloud resources…..However, much of the research in our group deals with large scale problems rather than high performance problems. In such a scenario, on-demand access to a large number of virtual machine is more useful than round the cloud availability of a captive cluster. In addition, the overhead for installing and maintaining a local cluster is non-trivial unless strong system administration resources are available at the local institution. Availability of platform services such as storage and programming abstractions such as .NET or MapReduce reduces the overhead of installing, monitoring and managing such services locally [148].”
Cloud Funding	NSF; commercial; personal
Research Funding	NSF; DOE; commercial

Linguistics: Calculating Similarity Scores and Large-Scale Data Mining

Project	Data transformation
Use Cases	Burst resources; collaboration; commonly requested software; computer science research; computing and data analysis support for scientific workflows; data archiving; data management and analysis; data sharing; domain-specific computing environments
Primary Researchers	Gavin La Rowe and Bruce Herr, Chalklabs
Abstract	Most recently we used AWS for calculating similarity scores and large-scale data modeling. Using multiple AWS HPC instances in parallel for a large-scale data mining algorithm, we achieved 95% run-time optimizations in both processes.
Cloud Providers	Amazon Web Services; Google Cloud Platform
Special Features	Community datasets or collections; GPUs; Hive; MapReduce; queues; tables
Dev. Environment	Eucalyptus; Nimbus; OpenNebula; OpenStack; VMware
Use Regularity	Monthly
Cores Used Peak	17920
Cores Steady State	17920
Core Hours in a Year	504
Storage Accessed For	Reference
Preferred Storage	Elastic Block Storage
Accessed During Run	300GB
Short-Term Storage	4TB
Long-Term Storage	3TB
Data Moved Into Cloud	5TB
Data Moved Out Cloud	3TB
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Up to 10Gb/s
Type Data Moving	Providing basic access; research data sets or collections
Data Accessed By	Researcher; research group; department of institution; outside collaborators
Software	Home-grown; community developed; open source; commercial
Additional Notes	"We used AWS for calculating similarity scores and large-scale data modeling. Using multiple AWS HPC instances in parallel for a large-scale data mining algorithm, we achieved 95% run-time optimizations in both processes. The first job involved calculating similarity scores for a total of 8.6 trillion data pairs. The second job … we optimized the data modeling application to best use the memory and cores available for the AWS high-memory instance and reduced our run-time processing for a job of 3.8 million ScienceDirect articles from 100 days on our infrastructure down to just 5 days of processing time on AWS [149].
Cloud Funding	NIH; commercial
Research Funding	NIH; commercial

Data mining 3.8 million ScienceDirect articles

Linguistics: Predicate-Argument Structure Analysis

Project	Predicate-argument structure analysis of huge web
Use Cases	Domain-specific computing environments
Primary Researchers	Daisuke Kawahara and Sadao Kurohashi, Kyoto University
Abstract	We have been developing a search engine infrastructure, TSUBAKI, which is based on deep Natural Language Processing. While most conventional search engines register only words to their indices, TSUBAKI provides a framework that indexes synonym relations, hypernym-hyponym relations, dependency/case/ellipsis relations and so forth. These indices enable TSUBAKI to capture the semantic matching between a given query and documents more precisely and flexibly. Case/ellipsis relations have not been indexed in a large scale because the speed of these analyses is not fast enough due to the necessity of referring to a large database of predicate-argument patterns (case frames). To apply case/ellipsis analysis to millions of Web pages of TSUBAKI in a practical time, it is necessary to use 10,000 CPU cores. Because of limits on the Azure fabric controller, it was necessary to divide this into 29 hosted services of 350 CPUs each. This was the largest experiment of any of the research engagement projects.
Cloud Providers	Windows Azure
Special Features	Community datasets or collections
Use Regularity	Annually
Cores Used Peak	10000
Cores Steady State	10000
Core Hours in a Year	1000000
Storage Accessed For	Analysis
Preferred Storage	Parallel performance file system
Accessed During Run	4GB
Short-Term Storage	3TB
Long-Term Storage	3TB
Data Moved Into Cloud	300GB
Data Moved Out Cloud	3TB
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	N/A
Type Data Moving	Research data sets or collections
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source; commercial
Additional Notes	"… since our case/ellipsis analysis system has been developed using the C language on Linux, it was necessary to port our system to Windows in order to execute on Azure. To do this, we employed a 118 Unix-like environment, Cygwin. We implemented the above framework and tested our analysis on 1x350, 2x350 and 8x350 step by step. Once we confirmed that we could obtain 29x350 CPU cores, we executed our analysis on these CPU cores. A remaining problem at this moment was the high cost of manually managing 29 hosted services. We then kept developing a manager of 29 hosted services based on the Windows Azure Service Management API [150].
Cloud Funding	Microsoft Research
Research Funding	Japan Society for the Promotion of Science

Social Sciences: Disseminating Confidential Data

Project	Exploring new methods of protecting and distributing confidential research data
Use Cases	Collaboration; computing and data analysis support for scientific workflows; data management and analysis; data sharing
Primary Researchers	Felicia LeClere, NORC at the University of Chicago
Additional Researchers	Bryan Beecher, Inter University Consortium for Political and Social Research (ICPSR)
Abstract	The sharp increase in the sophistication of social science data systems that accompanied computer-assisted data collection methods created a concomitant increase in the risk of disclosing individual respondent's identities when the data are shared more broadly. Public use data files, which substantially reduce the risk of disclosure through statistical and technical methods often also reduce the analytic utility of these data. Data producers have increasingly chosen to retain the original analytic potential of the data by releasing the files under a modified data use agreement or legal contract with analysts. Large data collection programs, both inside and outside the Federal Statistical System, increasingly issue a substantial number of these contracts annually. The contracts often place a large burden on the end user to provision and secure computing platforms that are designed to protect the electronic security of the data files. Different data systems also will often require separate machinery for each data use contract. This ad hoc system for securing and disseminating confidential data has limited both the availability and the security of the data. In this project, the Inter University Consortium for Political and Social Research [151] and partners at the Rand Corporation and the Survey Research Center at the University of Michigan will build and test a data storage and dissemination system for confidential data, which obviates the need for users to build and secure their own computing environments. Recent advances in public utility (or "cloud") computing now makes it feasible to provision powerful, secure data analysis platforms on-demand. We will leverage these advances to build a system which collects "system configuration" information from analysts using a simple web interface, and then produces a custom computing environment for each confidential data contract holder. Each custom system will secure the data storage and usage environment in accordance with the confidentiality requirements of each data file. When the analysis has been completed, this custom system will be fed into a "virtual shredder" before final disposal. This prototype data dissemination system will be tested for (1) system functionality (i.e., does it remove the usual barriers to data access?); (2) storage and computing security (i.e., does it keep the data secure?); and (3) usability (i.e., is the entire system easier to use?). Contract holders of two major data systems (the Panel Study of Income Dynamics and the Los Angeles Family and Neighborhood Study) will be recruited to assess both the user interface and the analytic flexibility of the new customized computing environments.
Cloud Providers	Amazon Web Services
Special Features	Community datasets or collections
Use Regularity	Annually
Cores Used Peak	1
Cores Steady State	1
Core Hours in a Year	1
Storage Accessed For	Analysis

ICPSR is testing prototype confidential data dissemination systems in the cloud

Preferred Storage	Elastic Block Storage
Accessed During Run	10GB
Short-Term Storage	10GB
Long-Term Storage	10GB
Data Moved Into Cloud	10GB
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 100Mb/s
BW to Storage Within	Up to 100Mb/s
Type Data Moving	Research data sets or collections; survey data
Data Accessed By	Researcher; research group; outside collaborators
Software	Home-grown; community developed; open source; commercial
Additional Notes	"ICPSR has been utilizing the capabilities of several public cloud providers since 2009, generally for fail over and replication of select functions such as DNS, our search service, and encrypted storage of copies of non-confidential archived data. Based on this experience we concluded early in the project that an optimal cloud-based computing environment for our use case would be very similar to traditional solutions, and that we would be able to leverage existing expertise, skills, tools and management infrastructure. The question became would the threat model also be similar in nature or would cloud introduce unique additional risk vectors….The issues (our ethical hacker) found were almost entirely challenges we would face and issues we would have had to protect against whether this was locally hosted, using our on premise physical infrastructure, or remotely hosted at a public cloud provider. Admittedly more effort will be needed to ensure each use case's regulatory and compliance concerns are or can be addressed. And while we believe we satisfied the concerns presented by co-location with other clients of a public cloud provider, it is reasonable to assume further efforts may be needed if a higher level of isolation is demanded for specific confidential data. However, our results affirmed our belief that institutions such as our own can responsibly utilize cloud and public cloud providers….We encourage educational institutions to assess the value proposition of computing in the cloud [152]."
Cloud Funding	NIH
Research Funding	NIH

Discipline Unspecified Cloud Projects Surveyed: Complete Data

Cloud Investigation by Research Computing Services: Columbia University

Project	HPC cloud investigation
Use Cases	Education, outreach, and training (EOT)
Primary Researchers	Rob Lane, Columbia University
Abstract	No actual research. Just investigating service for possible future use.
Cloud Providers	Red Cloud
Dev. Environment	Eucalyptus
Use Regularity	Monthly
Cores Used Peak	1
Cores Steady State	1
Cores Hours in a Year	3000
Storage Accessed For	Analysis
Preferred Storage	N/A
Accessed During Run	0
Short-Term Storage	0
Long-Term Storage	0
Data Moved Into Cloud	0
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	Uncertain
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher
Software	None
Capabilities/Features	“We hope to eventually use cloud resources to support research for which they are particularly suited.”
Cloud Funding	Departmental
Research Funding	None

Cloud Investigation by Research Computing Services: University of Colorado at Boulder

Project	Investigation of cloud technologies for the advancement of campus research
Use Cases	Burst resources; collaboration; commonly requested software; computing and data analysis support for scientific workflows; data archiving; data management and analysis; data sharing; domain-specific computing environments; education, outreach, and training (EOT); science gateways
Primary Researchers	Jazcek Braden and Thomas Hauser, University of Colorado at Boulder
Abstract	This is a shared digital repository serving seven institutions of higher education in Colorado. All types of data, programs, protocols, from all institutions. We use Google Apps for Education email and unified messaging cloud.
Cloud Providers	Amazon Web Services; FutureGrid; Globus Online
Special Features	Community datasets or collections
Dev. Environments	Eucalyptus; Nimbus; OpenNebula; OpenStack
Use Regularity	Weekly
Cores Used Peak	10
Cores Steady State	1
Cores Hours in a Year	100
Storage Accessed For	Reference; archival
Preferred Storage	Elastic Block Strorage
Accessed During Run	10GB
Short-Term Storage	10GB
Long-Term Storage	100GB
Data Moved Into Cloud	100GB
Data Moved Out Cloud	0
BW In/Out of Cloud	Up to 10Gb/s
BW to Storage Within	N/A
Type Data Moving	Not moving data, just programs
Data Accessed By	Researcher; research group
Software	Home-grown; community developed; open source
Capabilities/Features	“Ability to provide custom environments for those researchers whose research platforms advance faster or slower than the commonly provided environments can support.”
Problems/Limitations	“Somewhat time consuming to learn the utilities and nuances/bugs with trying to deploy, admin and monitor base resources.”
Cloud Funding	None
Research Funding	None

Appendix

Acronyms

ACI	Division of Advanced Cyberinfrastructure (NSF)
ATLAS	A Toroidal LHC Apparatus (particle physics experiment at Large Hadron Collider)
AWS	Amazon Web Services
CI	Cyberinfrastructure
CISE	Directorate of Computer & Information Science & Engineering (NSF)
CSA	Citizen Science Alliance
DARPA	Defense Advanced Research Projects Agency
DOD	Department of Defense
DOE	Department of Energy
EBI	European Bioinformatics Institute
EBS	Elastic Block Storage (Amazon)
EC2	Elastic Compute Cloud (Amazon)
EEG	Electroencephalography
EnKF	Ensemble Kalman filter
EOT	Education, Outreach, and Training
ESA	European Space Agency
EU	European Union
FEMA	Federal Emergency Management Agency
FPGA	Field-Programmable Gate Array
GPGPU	General-Purpose Graphics Processing Unit
HASS	Humanities, Arts, and Social Sciences
HPC	High Performance Computing
HIPAA	Health Insurance Portability and Accountability Act
IaaS	Infrastructure as a Service (user deploys/controls operating system, apps, storage, etc.).
Jmol	Java viewer for chemical structures in 3D
LHC	Large Hadron Collider
LLNL	Lawrence Livermore National Laboratory
MC	Monte Carlo
MD	Molecular Dynamics
MIC	Many Integrated Core Architecture (Intel)
MOOC	Massive Open Online Course
MPI	Message Passing Interface
mRNA	Messenger RNA molecules
NAS	Network-Attached Storage
NGS	Next-Generation Sequencing
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NMR	Nuclear Magnetic Resonance
NREL	National Renewable Energy Laboratory
NSF	National Science Foundation
OEM	Original Equipment Manufacturer
OS	Operating System
PaaS	Platform as a Service (languages and/or tools provided)
QoS	Quality of Service
RDMS	Relational Database Management System
S3	Simple Storage Service (Amazon)
SaaS	Software as a Service (applications provided)
SLA	Service Level Agreement
STEM	Science, Technology, Engineering, and Mathematics
VM	Virtual Machine

Terminology*

BigQuery	Web service for interactive analysis of massive datasets (Google)
BLAST	Basic Local Alignment Search Tool used to compare biological sequences
ChEMBL	Database of bioactive drug-like small molecules
ClumsyLeaf	CloudXplorer UI client used to browse Windows Azure storage
Cloud-TM	Transactional Memory for the cloud
ClustalW	Command line multiple sequence alignment
CometCloud	Autonomic framework to enable applications on hybrid infrastructure (Rutgers)
CPMD	Car Parrinello Molecular Dynamics
CycleCloud	Utility computing software to create HPC clusters in the cloud (Cycle Computing)
CycleServer	Management and submission tool (Cycle Computing)
DynamoDB	Fully managed NoSQL database service (Amazon)
e-Science Central	Cloud-based platform for data analysis
Eucalyptus	Open-source software for building AWS-compatible private and hybrid clouds
Glacier	Low cost, archival storage (Amazon)
GlusterFS	Network/cluster file system written in user space
Hadoop	Open-source framework that supports data-intensive applications (Apache)
HBase	Radom, real-time read/write access to Big Data
Hive	Data warehouse system for Hadoop for ad-hoc queries and data analysis
Hybrid cloud	Combination of two or more clouds (public, private or community)
Hypervisor	Software or hardware that runs virtual machines
LAMMPS	Molecular Dynamics Simulator (Sandia)
Makeflow	Workflow engine for executing large complex workflows on clouds (Notre Dame)
MapReduce	Programming model for processing large data sets with parallel algorithm
MediaWiki	Wiki implementation that uses PHP to process/display data stored in a database
MongoDB	Open-source NoSQL document database
MySQL	Open-source database that enables cloud applications to scale-out
Multi-clouds	Running applications across different clouds (private and public cloud portability)
NAMD	Parallel molecular dynamics code for large biomolecular systems
Nimbus	EC2/S3-compatible IaaS implementation
noSQL	Non-relational, distributed, open-source database
Object store	Object storage device, e.g., Amazon S3, OpenStack Swift
Octave	High-level language for numerical computations (GNU)
OpenFlow	A way for researchers to run experimental protocols in every day networks
OpenNebula	Used to build cloud infrastructures on Xen, KVM, and VMware deployments
Open Science Grid	A consortium that administers worldwide resources for distributed computing
OpenStack	Open-source IaaS cloud operating system
Parrot	Tool to attach existing programs to remote I/O systems
RabbitMQ	Open-source message broker
PostgreSQL	Open-source Object-Relational DBMS supporting almost all SQL constructs
Private cloud	Cloud accessed by only one organization
Public cloud	Cloud accessed by the general public, e.g., AWS, Google Platform Services, etc.
Red Cloud	Public IaaS with exclusive access to CPU cores; MATLAB SaaS (Cornell)
Redmine	Project management web application
RESEVOIR	Reservoir software suite (Baker Hughes)
Rosetta	Software suite for modeling macromolecular structures
Scala	Object-functional programming and scripting language
SSAHA	Sequence Search and Alignment by Hashing Algorithm
SQLaaS	SQL Server as a Service
StarCluster	Open source cluster-computing toolkit for Amazon EC2 (MIT)
VirtualBox	x86 virtualization software
VMware	Virtualization software (also VMware vCloud Suite for integrated cloud)
Xen	Hypervisor that allows multiple OSs to run at same time on same hardware

* NIST has developed comprehensive cloud terms and definitions [153]

Service Providers Mentioned in this Report

Amazon Web Services	http://aws.amazon.com/
Cloudera	http://www.cloudera.com
CloudSigma	http://www.cloudsigma.com/
Connectria	https://www.connectria.com/
Cycle Computing	http://www.cyclecomputing.com/
CSC	http://www.csc.com/cloud
Dell	http://www.dell.com/learn/us/en/19/dell-cloud-computing
FutureGrid	https://portal.futuregrid.org/
Globus Online	https://www.globusonline.org/
Google Cloud Platform	https://cloud.google.com/
Grid'5000	https://www.grid5000.fr/
HP	https://www.hpcloud.com/
IBM	http://www.ibm.com/cloud-computing/us/en/
Nimbex	http://www.nimbix.net/
Open Science Data Cloud	https://www.opensciencedatacloud.org/
Open Science Grid	https://www.opensciencegrid.org/bin/view
Penguin On Demand	http://www.penguincomputing.com/services/hpc-cloud/pod
Rackspace	http://www.rackspace.com/
Red Cloud	htttp://www.cac.cornell.edu/redcloud/
SDSC Cloud Storage	https://www.cloud.sdsc.edu/
SGI	http://www.sgi.com/products/
Windows Azure	http://www.windowsazure.com/en-us/

Other Service Providers*

Bit Refinery	http://bitrefinery.com/
BlueLock	http://www.bluelock.com/
Claris Networks	http://clarisnetworks.com/
eApps	http://www.eapps.com/
ElasticHosts	http://www.elastichosts.com/
extreme factory	http://www.extremefactory.com/
GoGrid	http://www.gogrid.com/
GMO Cloud	https://us.gmocloud.com/
Green House Data	http://www.greenhousedata.com/
Joyent	http://joyent.com/
Layered Tech	http://www.layeredtech.com/
PhonenixNAP	http://www.phoenixnap.com/
Qube	http://www.qubemanagedhosting.com/
ScaleMatrix	http://www.scalematrix.com/
SoftLayer	http://www.softlayer.com/
TekLinks	http://teklinks.com/
ZeroLag	http://www.zerolag.com/

*Intel Cloud Finder is an online tool for identifying service providers: http://www.intelcloudfinder.com/.

References

[1] Pop!World (n.d.). *University at Buffalo*. Retrieved from http://popfrontpage.appspot.com/.

[2] Yue, C., Zhu, W., Williams, G., & Chow, E. (2012). Using Amazon EC2 in computer and network security lab exercises: design, results, and analysis. *American Society for Engineering Education*. Retrieved from http://www.cs.uccs.edu/~cyue/papers/ASEE12.pdf.

[3] Cloud BioLinux (n.d.). Retrieved from http://cloudbiolinux.org/.

[4] Zooniverse (2012). *Zooniverse: Real Science Online*. Retrieved from https://www.zooniverse.org/.

[5] Community Seismic Network (n.d.). California Institute of Technology. Retrieved from http://csn.caltech.edu/index.html.

[6] National Science Foundation. *Cyberinfrastructure for 21st Century Science and Engineering Advanced Computing Infrastructure: Vision and Strategic Plan.* February 2012. NSF 12-051. Retrieved from http://www.nsf.gov/pubs/2012/nsf12051/nsf12051.pdf.

[7] Distributed and Parallel Systems Group: Cloud Computing (n.d.). *University of Innsbruck*. Retrieved from http://www.dps.uibk.ac.at/en/projects/cloud/.

[8] Birman, K. "Items that need cloud computing research investment." E-mail to P. Redfern, April 12, 2013.

[9] Neovise. *Enterprise Cloud Essentials: Multiple Clouds, Hybrid Environments and Support for Mission-Critical Applications.* March 2013. Retrieved from http://bit.ly/10Zf9Py.

[10] XSEDE Cloud Use Survey (2013). Retrieved from http://xsede.org/CloudSurvey/.

[11] Forrester Q3 2012 Global Cloud Developer Online Survey Summary Results (2012). Retrieved from http://www.slideshare.net/johnrrymer/summary-of-forrester-q3-2012-global-cloud-developer-survey.

[12] Intel Cloud Finder: Service provider quick search (n.d.). Retrieved from http://www.intelcloudfinder.com/quicksearch.

[13] Roosakos, P. (2013, April 18). [Web log message]. Retrieved from http://blog.foghornconsulting.com/2013/04/18/capacity-management-cloud-style/?goback=%2Egmr_61513%2Egde_61513_member_236921386.

[14] Foster, I. (2013, May 21). CERN, Google, and the future of global science initiatives. *HPC in the Cloud*. Retrieved from http://www.hpcinthecloud.com/hpccloud/2013-05-21/cern_google_and_the_future_of_global_science_initiatives.html.

[15] Amazon EC2 Spot Instances (n.d.). Retrieved from http://aws.amazon.com/ec2/spot-instances/.

[16] Cornell University: Red Cloud with MATLAB (n.d.). Retrieved from http://www.cac.cornell.edu/wiki/index.php?title=Red_Cloud_with_MATLAB.

[17] Johnson, S. (2011, March 12). [Web log message]. RStudio and Amazon Web Services. Retrieved from: http://community.mis.temple.edu/stevenljohnson/2011/03/12/rstudio-and-amazon-web-services-ec2/.

[18] Kar, S. (2012, Aug. 21). Gartner hype cycle for cloud computing: SaaS most promising technology. *CloudTimes*. Retrieved from http://cloudtimes.org/2012/08/21/gartner-hype-cycle-cloud-computing-saas/.

[19] Allen, B., Bresnahan, J., Childers, L., Foster, I., Kanadaswamy, G., Kettimuthu, R., Kordas, J., Link, M., Martin, S., Pickett, K., & Tuecke, S. (2011, July). Globus Online: Radical Simplification of Data Movement via SaaS. Retrieved from https://www.globusonline.org/files/2011/07/Globus-Online-SaaS-Simplification-of-Data-Movement.pdf.

[20] Pew Research Center. *The Digital Revolution and Higher Education*. By K. Parker, A. Lenhart, & K. Moore. August 2011. Retrieved from http://www.pewinternet.org/~/media/Files/Reports/2011/PIP-Online-Learning.pdf.

[21] Hardy, Q (2012, Aug. 28). Amazon quietly harnesses the coming cloud-computing future. *The New York Times* via *The Tech Online*. Retrieved from http://tech.mit.edu/V132/N32/long3.html.

[22] National Science Foundation. *NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging.* By C. Stewart, G. Almes, V. Welch, M. Lingwall, et al. March 2011. Retrieved from https://www.nsf.gov/cise/aci/taskforces/TaskForceReport_CampusBridging.pdf.

[23] Gens, F. (2013). IDC Predictions 2013: Competing on the 3rd Platform. *IDC Top 10 Predictions.* Retrieved from http://www.idc.com/research/Predictions13/downloadable/238044.pdf.

[24] Agee, A., Rowe, T., & Woo, M. (2010, Sept. 22). Building research cyberinfrastructure at small/medium research institutions. *Educause Review*. Retrieved from http://www.educause.edu/ero/article/building-research-cyberinfrastructure-smallmedium-research-institutions.
[25] The Uber-Cloud Experiment (n.d.). Retrieved from http://www.hpcexperiment.com/.
[26] Council on Competitiveness. *Make: An American Manufacturing Movement*. Dec. 2011 (pg. 39). Retrieved from: http://www.compete.org/images/uploads/File/PDF%20Files/USMCI_Make.pdf.
[27] FutureGrid Portal (n.d.). Retrieved from https://portal.futuregrid.org/.
[28] Roberts, D. (2013, Feb. 5). Business agility: how does cloud computing and big data help? Leverhawk. Retrieved from http://leverhawk.com/business-agility-how-cloud-computing-and-big-data-help-2013020526.
[29] CycloneCenter (n.d.). Retrieved from http://www.cyclonecenter.org/#/home.
[30] National Institutes of Health News (2012, March 29). *1000 Genomes Project available on Amazon Cloud*. Retrieved from http://www.nih.gov/news/health/mar2012/nhgri-29.htm.
[31] National Institute of Standards and Technology. *Cloud Computing Synopsis and Recommendations.* By L. Badger, T. Grance, R. Patt-Comer, J. Voas. May 2012. NIST Publ. 800-146. Retrieved from http://csrc.nist.gov/publications/nistpubs/800-146/sp800-146.pdf.
[32] Duffy, J. (2013, April 2). "Tsunami" of bandwidth demand pushes IEEE 400G Ethernet standards process. *NetworkWorld.* Retrieved from http://www.networkworld.com/news/2013/040113-ieee-400g-ethernet-268261.html.
[33] Virtustream Neovise Research Report: Use of public, private and hybrid cloud computing (n.d.) Retrieved from http://www.virtustream.com/company/buzz/press_releases/neovise_research_report.
[34] Intel Solution Brief: Fast, low-overhead encryption for Apache Hadoop (2013). Retrieved from https://hadoop.intel.com/pdfs/IntelEncryptionforHadoopSolutionBrief.pdf.
[35] Zooniverse (2012). *Zooniverse: Real Science Online*. Retrieved from https://www.zooniverse.org/.
[36] Schwamb, M.E., Orosz, J.A., et al. (2012). *Planet hunters: a transiting circumbinary planet in a quadruple star system.* Informally published manuscript, Available from arXiv:1210.32 [astro-ph.EP]. Retrieved from http://arxiv.org/abs/1210.3612.
[37] Smith, A. (2011, April 8). [Web log message]. Retrieved from http://arfon.org/getting-started-with-elastic-mapreduce-and-hadoop-streaming.
[38] Parsons, P. (2009). *Cloud science or astrometric data processing in Amazon EC2.* Retrieved from http://www.slideshare.net/pparsonsesp/cloud-science.
[39] The Server Labs (n.d.). *Amazon Web Services Case Studies.* Retrieved from http://aws.amazon.com/solutions/case-studies/the-server-labs/.
[40] O'Mullane, W, Olias, A., Parsons, P. et al. (2011). *Using EC2 for the Gaia astromteric solution.* Submitted for publication. Retrieved from http://www.astro.lu.se/~david/papers/womullan_20112009.pdf.
[41] Pop!World (n.d.). *University at Buffalo*. Retrieved from http://popfrontpage.appspot.com/.
[42] Rae, T. (2011, January 28). Biology professors use cloud computing to reach students. *Chronicle of Higher Education.* Retrieved from http://chronicle.com/blogs/wiredcampus/biology-professors-use-cloud-computing-to-reach-students/29330.
[43] Ramamurthy, B., Pulin, J., & Dittmar, K. (2012). Cloud-enabling biological simulations for scalable and sustainable access: an experience report. *XSEDE '12 Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond ACM*, doi: 10.1145/2335755.2335852. Retrieved from http://dl.acm.org/citation.cfm?doid=2335755.2335852.
[44] Loquet, A. & Sgourakis, N. (2012). Atomic model of the type iii secretion system needle. *Nature*, *486*(7402), 276-279. doi: 10.1038/nature11079. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22699623.
[45] Trader, T. (2011, June 16). Researchers tap Azure for salmonella clues. *HPC in the Cloud*, Retrieved from http://www.hpcinthecloud.com/hpccloud/2011-06-16/researchers_tap_azure_for_salmonella_clues.html.
[46] COSBI (n.d.). Retrieved from http://www.cosbi.eu/.

[47] Di Cosmo, M. & Sanger, A. (2012). BetaSim in the cloud. In D. Gannon & J. Vargas (Eds.), *The Cloud Computing Engagement Research Program* (pg. 71). Seattle, WA: Microsoft.
[48] Amber 12 NVIDIA GPU acceleration support (2012). Retrieved from http://ambermd.org/gpus/.
[49] Amber home page (n.d.). Retrieved from http://ambermd.org/.
[50] FEATURE Project Overview (n.d.). *Simtk.org*. Retrieved from https://simtk.org/home/feature.
[51] StarCluster (n.d.). Retrieved from http://star.mit.edu/cluster/.
[52] Asynchronous replica exchange for grid-based molecular dynamics applications (n.d.). *Rutgers: The Cloud and Automatic Computing Center*. Retrieved from http://nsfcac.rutgers.edu/CometCloud/CometCloud/applications/repex.
[53] PDB: Protein Data Bank (n.d.). Retrieved from http://www.rcsb.org/pdb/home/home.do.
[54] AbdelBaky, M., Kim, H., Rodero, I., & Parashar, M. (2012). *Accelerating MapReduce analytics using CometCloud*. In *CLOUD Pro*ceedings (pp. 447-454). Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6253537.
[55] Wang, F. (2011, August). High performance spatial data warehouse over MapReduce. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/141.
[56] Wang, F., Aji, A., Liu, Q., & Saltz, J. (2011). *Hadoop-GIS: A high performance spatial query system for analytical medical imaging with MapReduce*. Informally published manuscript, Center for Comprehensive Informatics, Emory University, Atlanta, GA. Retrieved from http://confluence.cci.emory.edu:8090/download/attachments/4033450/CCI-TR-2011-3.pdf.
[57] High-throughput medical image registration on Comet framework (n.d.). *Rutgers: The Cloud and Automatic Computing Center*. Retrieved from http://nsfcac.rutgers.edu/CometCloud/CometCloud/applications/imagereg.
[58] Patch, S. (2012). The Research of Sarah Patch. *University of Wisconsin-Milwaukee*. Retrieved from http://www4.uwm.edu/letsci/physics/research/patchs/.
[59] Davies, K. (2012, April 19). Cycles launches 50,000-core utility supercomputer in the cloud. *Bio-IT World*. Retrieved from http://www.bio-itworld.com/2012/04/19/going-up-cycle-launches-50000-core-utility-supercomputer-in-cloud.html.
[60] Cala, J., Hiden, H., Woodman, S., & Watson, P. Fast Exploration of the QSAR Model Space with e-Science Central and Windows Azure. *In: Microsoft Cloud Futures*, 7-8 May 2012, Berkeley, California. Retrieved from http://eprint.ncl.ac.uk/file_store/production/185854/80966CE5-F8B7-45BD-9FE8-D3D1755EBB5D.pdf.
[61] Chemical property prediction (2012, September 5). *VENUS-C*. Retrieved from http://www.venus-c.eu/Content/UserScenarios.aspx?id=fc57df9e-26a8-4793-af5b-72208d4b43f4.
[62] Yue, C., Zhu, W., Williams, G., & Chow, E. (2012). Using Amazon EC2 in computer and network security lab exercises: design, results, and analysis. *American Society for Engineering Education*. Retrieved from http://www.cs.uccs.edu/~cyue/papers/ASEE12.pdf.
[63] CS 5910 Fundamentals of Computer/Network Security (2011). *University of Colorado at Colorado Springs*. Retrieved from http://www.cs.uccs.edu/~cyue/teaching/CS5910LabMaterial/.
[64] Yue, C., Zhu, W., Williams, G., & Chow, E. (2012). Using Amazon EC2 in computer and network security lab exercises: design, results, and analysis. *American Society for Engineering Education*. Retrieved from http://www.cs.uccs.edu/~cyue/papers/ASEE12.pdf.
[65] Grunwald, D. (2012, August). Course: Data center scale computing. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/244.
[66] CS344 Introduction to Data Management (2012). *University of Washington: Computer Science & Engineering*. Retrieved from http://www.cs.washington.edu/education/courses/cse344/.
[67] CSE344 Homework 3: SQL (2012). *University of Washington: Computer Science & Engineering*. Retrieved from http://www.cs.washington.edu/education/courses/cse344/11au/hw/hw3/hw3.html.
[68] CSE344 Homework 6: Hadoop and Pig (2012). *University of Washington: Computer Science & Engineering*. Retrieved from http://www.cs.washington.edu/education/courses/cse344/11au/hw/hw6/hw6.html.
[69] von Laszewski, G. (2012, July). Course: Science Cloud Summer School 2012. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/241.
[70] von Laszewski, G. (2011, May). Course: CCGrid2011 Tutorial. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/124.
[71] von Laszewski, G. (2012, July). Community comparison of cloud frameworks. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/239.

[72] Distributed and Parallel Systems Group: Cloud Computing (n.d.). *University of Innsbruck*. Retrieved from http://www.dps.uibk.ac.at/en/projects/cloud/.
[73] The Cooperative Computing Lab (n.d.). *University of Notre Dame*. Retrieved from: http://www3.nd.edu/~ccl/.
[74] SciFlex: Data-as-a-Service for long tail science (2011). Retrieved from http://sciflex.cs.washington.edu/.
[75] Howe, B., Cole, G., Key, A., Khoussainova, N., & Battle, L. (2012). SQLShare: Database-as-a-Service for long tail science. In D. Gannon & J. Vargas (Eds.), *The Cloud Computing Engagement Research Program* (pgs. 52-56). Seattle, WA: Microsoft.
[76] Internet Systems Lab (n.d.). *Purdue University*. Retrieved from https://engineering.purdue.edu/~isl/index.htm.
[77] SALSA (n.d.). *Indiana University at Bloomington*. Retrieved from http://salsahpc.indiana.edu/.
[78] CloudTM: A novel programming paradigm for the cloud. Retrieved from https://sites.google.com/site/cloudtmproject/.
[79] RESERVOIR Framework (n.d.). Retrieved from http://www.reservoir-fp7.eu/.
[80] Lee, H. (2010, December). Development of an information service for FutureGrid. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/20.
[81] Rain: Cloud/HPC provisioning (n.d.). *FutureGrid*. Retrieved from http://futuregrid.github.com/rain/index.html.
[82] Bosin, A. (2011, September). Resource provisioning for e-Science environments. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/157.
[83] Adeleke, A.A. (2012, April). Exploring and cataloging cloud computing security issues via FutureGrid. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/210.
[84] Bates, A., Mood, B., & Pletcher, J. (2012). Detecting co-residency with active traffic analysis techniques. *4th ACM Cloud Computing Security Workshop (CCSW 2012)*, Raleigh, NC, October 2012. Retrieved from http://ix.cs.uoregon.edu/~butler/pubs/ccsw12.pdf.
[85] Isis2 cloud computing library (n.d.). *CodePlex*. Retrieved from http://isis2.codeplex.com/.
[86] GridControl: A software platform to support the smart grid (n.d.). *Cornell University Computer Science*. Retrieved from http://www.cs.cornell.edu/projects/gridcontrol/.
[87] Xen-Blanket (2012). *Cornell University*. Retrieved from http://xcloud.cs.cornell.edu/.
[88] BitBlaze: Binary analysis for computer security (n.d.). *University of California, Berkeley*. Retrieved from http://bitblaze.cs.berkeley.edu/.
[89] Zhang, L., Ma, X., Lu, J., Tillmann, N., & de Halleux, P. (2012). Environmental modeling for automated cloud application testing. *IEEE Software*, *29*(2), 30-35. doi:10.1109/MS.2011.158. Retrieved from http://www.computer.org/csdl/mags/so/2012/02/mso2012020030-abs.html.
[90] Tsugawa, M. (2011, March). FutureGrid and Grid'5000 collaboration. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/97.
[91] Matsunaga, A. (2012, May). Scaling-out CloudBLAST: Deploying Elastic MapReduce across geographically distributed virtualized resources for BLAST. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/216.
[92] Matsunaga, A. & Fortes, J. (2010, May). On the use of machine learning to predict the time and resources consumed by applications. *10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid)*. Melbourne, Australia. doi: 10.1109/CCGRID.2010.98. Retrieved from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5493447.
[93] Keahey, K., Tsugawa, M., Matsunaga, A., & Fortes, F. (2009). Sky computing. *IEEE Internet Computing*, *13*(5), 43-51. Retrieved from http://www.computer.org/csdl/mags/ic/2009/05/mic2009050043-abs.html.
[94] Tsugawa, M., Matsunaga, A., & Fortes, J. (2009). User-level virtual network support for sky computing. *E-Science*, 72-9. doi: 10.1109/e-Scienc.2009.19. Retrieved from http://dl.acm.org/citation.cfm?id=1724805.
[95] Xen-Blanket (2012). *Cornell University*. Retrieved from http://xcloud.cs.cornell.edu/.
[96] Carat (n.d.). University of California, Berkeley, Retrieved from http://carat.cs.berkeley.edu/.
[97] Oliner, A., Iyer, A., Lagerspetz, E., Tarkoma, S., & Stoica, I (2012). Collaborative energy debugging for mobile devices. *HotDep'12 Proceedings of the Eighth USENIX Conference on Hot Topics in System Dependability*, 2012, Hollywood, CA. Retrieved from http://adam.oliner.net/files/oliner_hotdep_2012.pdf.

[98] Omtzigt, T. (2012, October 3). [Web log message]. Retrieved from http://stillwater-cse.blogspot.com/.

[99] OpenEI: Open Energy Info (n.d.). Retrieved from http://en.openei.org/wiki/Main_Page.

[100] Brodt-Giles, D. (2012). Open Energy Information video [Web]. *The White House: Open Government Initiative.* Retrieved from http://www.whitehouse.gov/open/innovations/OpenEnergyInformation.

[101] National Renewable Energy Laboratory's OpenEI.org (n.d.). *Amazon Web Services Case Studies.* Retrieved from http://aws.amazon.com/solutions/case-studies/openei/.

[102] Humphrey, M., Beekwilder, N., & Goodall, J. (2012). Calibration of watershed models using cloud computing. *IEEE International Conference on eScience 2012*. Retrieved from http://www.cs.virginia.edu/~humphrey/papers/HumphreyWatershed_eScience2012.pdf.

[103] Dunne, J., Yoon, S., & Martinez, N. (2012). Network3D on Windows Azure: Web portal for ecological network simulations & analysis. (pgs. 16-19). *The Cloud Computing Engagement Program*. Seattle, WA: Microsoft.

[104] Kim, H., Chaudhari, S., Parashar, M., & Marty, C. (2009, May). Online risk analytics on the cloud. *International Workshop on Cloud Computing*, in conjunction with the *9th IEEE Symposium on Cluster Computing and the Grid (CCGRID 2009)*, Shanghai, China. Retrieved from http://nsfcac.rutgers.edu/CometCloud/sites/nsfcac.rutgers.edu.CometCloud/files/pub/cloud2009.pdf.

[105] Kim, H., Abdelbaky, M., & Parashar, M. (2009, October). CometPortal: A portal for online risk analytics using CometCloud. *17th International Conference on Computing Theory and Applications (ICCTA2009)*, Alexandria, Egypt. Retrieved from http://nsfcac.rutgers.edu/CometCloud/sites/nsfcac.rutgers.edu.CometCloud/files/pub/ICCTA2009.pdf.

[106] Cloud BioLinux (n.d.). Retrieved from http://cloudbiolinux.org/.

[107] Krampis, K., Booth, T., Chapman, B., Tiwari, B., Bicak, M., Field, D., & Nelso, K. (2012). Cloud BioLinux: pre-configured and on-demand bioinformatic computing for the genomics community. *BMC Bioinform*, *13*(42), doi: 10.1186/1471-2105-13-42. Retrieved from http://www.biomedcentral.com/1471-2105/13/42.

[108] Ensembl (n.d.). Retrieved from http://www.ensembl.org/.

[109] Flicek, P., Amode, M.R., Barrell, D. et al. (2012). Ensembl 2012. *Nucleic Acids Research, 40,* D84-D90. doi: 10.1093/nar/gkr991. Retrieved from http://nar.oxfordjournals.org/content/40/D1/D84.full.

[110] Ensembl (2011). Amazon Web Services Case Studies. Retrieved from https://aws.amazon.com/solutions/case-studies/ensembl/.

[111] SkateBase web portal (n.d.). *North East Bioinformatics Collaborative*. Retrieved from http://skatebase.org/.

[112] Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A., & Brom, T. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. Unpublished manuscript, Michigan State University, USDA, East Lansing, MI, Retrieved from http://arxiv.org/pdf/1203.4802v2.pdf.

[113] Pell, J., Hintze, A., Casino-Koning, R., Howe, A., Tiedje, J., & Brown, C. (2012). Scaling metagenome sequence assembly with probabilistic de bruijn graphs. *Proceedings of the National Academy of Science*, doi: 10.1073/pnas.1121464109. Retrieved from http://www.pnas.org/content/early/2012/07/25/1121464109.full.pdf+html.

[114] Running a next-gen sequence analysis course using Amazon Web Services (2010, June 8). [Web log message]. *Living in an Ivory Basement: C. Titus Brown*. Retrieved from http://ivory.idyll.org/blog/ngs-course-with-aws.html.

[115] Su, Z., Arkela, S. and Zhou, Y. (2012). Large scale annotation of gene transcription regulatory sequences in bacterial genomes using cloud computing. In D. Gannon & J. Vargas (Eds.), *The Cloud Computing Engagement Research Program* (pgs. 34-35). Seattle, WA: Microsoft..

[116] TCGA center for systems analysis of the cancer regulome. (n.d.). *Institute for Systems Biology and MD Anderson Center*. Retrieved from http://www.cancerregulome.org/.

[117] Thomas, U. G. (2012, July 20). Google works with ISB to evaluate life sciences as application area for new cloud infrastructure. *BioInform*, Retrieved from http://www.genomeweb.com/informatics/google-works-isb-evaluate-life-sciences-application-area-new-cloud-infrastructur.

[118] Google Compute Engine: Behind the Compute Engine demo at Google I/O 2012. *Google Developers*. Retrieved from https://developers.google.com/compute/io.

[119] Cancer investigators use Google Compute Engine to accelerate life-saving research. (2012). *Google Case Study: Google Compute Engine*. Retrieved from https://cloud.google.com/files/ComputeISBCaseStudy.pdf.

[120] 1000 Genomes: A deep catalog of human genetic variation (n.d.). Retrieved from http://www.1000genomes.org/.

[121] Delwiche Lab (n.d.). *University of Maryland*. Retrieved from http://www.life.umd.edu/labs/delwiche/index.html

[122] Crayons (n.d.). Georgia State University. Retrieved from http://www.cs.gsu.edu/dimos/?q=content/gis-vector-data-overlay-processing-azure-platform.html.

[123] Community Seismic Network (n.d.). California Institute of Technology. Retrieved from http://csn.caltech.edu/index.html.

[124] Olson, M & Chandy, K.M . (2011). Performance issues in cloud computing for cyber-physical applications. *Proceedings of the IEEE International Conference on Cloud Computing*, 2011, Washington, DC. Retrieved from http://www.infospheres.caltech.edu/sites/default/files/cloud11-wip.pdf.

[125] Ensemble Kalman Filter. *Rutgers University: The Cloud and Automatic Computing Center*. Retrieved from http://nsfcac.rutgers.edu/CometCloud/applications/EnKF.

[126] Kim, H., el-Khamra, Y., Jha, S. & Parashar, M. Exploring adaptation to support dynamic applications on hybrid grids-clouds infrastructure, *1st Workshop on Scientific Cloud Computing*, in conjunction with the *ACM International Symposium on High Performance Distributed Computing (HPDC)*, Chicago, Illinois, June 20-25, 2010. Retrieved for http://nsfcac.rutgers.edu/CometCloud/sites/nsfcac.rutgers.edu.CometCloud/files/pub/final-version050.pdf.

[127] Parashar, M. (2011, June 27). Blue gene sniffs for black gold in the clouds. *HPC in the cloud*, Retrieved from http://www.hpcinthecloud.com/hpccloud/2011-06-27/blue_gene_sniffs_for_black_gold_in_the_cloud.html.

[128] AbdelBaky et al. (2011). *Scalable ensemble-based oil-reservoir simulations using Blue Gene/P as-a-service*. Unpublished manuscript, Rutgers University, IBM T.J. Watson Research, and The University of Texas at Austin. Retrieved from http://nsfcac.rutgers.edu/icode/scale/docs/scale.white.paper.final.pdf.

[129] Blaisten-Barojas, E. (2012). SAMP-computing in the cloud using the Windows Azure platform. D. Gannon & J. Vargas (Eds.), *The Cloud Computing Engagement Research Program* (pgs. 36-43). Seattle, WA: Microsoft.

[130] Blaisten-Barjas, E., & Xing, Q. (2012). A cloud computing system in Windows Azure platform for data analysis of crystalline material. *Concurrency and Computation: Practice and Experience*, doi: 10.1002/cpe.2912. Retrieved from http://www.cmasc.gmu.edu/publications_blaisten/108.pdf.

[131] Schiff, N.D. (Presenter) (2011). *The inside story: Breakthrough research and treatments in brain health* [Web]. Retrieved from http://www.youtube.com/watch?v=gitn8jEhSVs.

[132} Goldfine, A.M. & Schiff, N.D. (2011). What is the role of brain mechanisms underlying arousal in recovery of motor function after structural brain injuries? *Current Opinion Neurology, 6*, 564-569. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22002078.

[133] Tudoran, R. (2012). *A-Brain: Using the cloud to understand the impact of genetic variability on the brain* [Web PPT]. Retrieved from http://research.microsoft.com/en-us/um/redmond/events/cloudfutures2012/monday/Life_Sciences_A-Brain_Radu_Tudoran.pdf.

[134] Tudoran, R., Costan, A., Antoniu, G., & Soncu, H. (2012). *TomusBlobs: Towards communication-efficient storage for MapReduce applications in Azure*. Unpublished manuscript, INRIA Rennes, France. Retrieved from http://www.irisa.fr/kerdata/lib/exe/fetch.php?media=a-brain-ccgrid2012.pdf.

[135] A-brain. (n.d.). *Microsoft Research-INRIA Joint Centre Project*. Retrieved from http://www.msr-inria.fr/projects/a-brain/.

[136] Frazier, P., Xie, J., & Chick, S. (2011). *Value of information methods for pairwise sampling with correlations*. In Jain, S., Creasey, R.R., Immelspach, J., White, K.P, Fu, M. (Eds.), *Proceedings of*

the 2011 Winter Simulation Conference. Retrieved from http://people.orie.cornell.edu/pfrazier/pub/2011_FrazierXieChick_Correlated.pdf.
[137] Peter Frazier – Bayesian optimization via simulation with correlated sampling and correlated prior beliefs; Red Cloud with MATLAB case study (2012). *Cornell University Center for Advanced Computing*. Retrieved from http://www.cac.cornell.edu/about/studies/Frazier.pdf.
[138] Citrus Greening-HLB genome resources (n.d.). Retrieved from http://citrusgreening.org/index.html.
[139] Cloud computing. (2012). *Cornell University Center for Advanced Computing*. Retrieved from http://www.cac.cornell.edu/technologies/cloud.aspx.
[140] Diaz, R., Ramo, A., Aguero, A., Fifield, T., & Sevior, M. (2011). Belle-DIRAC setup for using Amazon Elastic Compute Cloud. *Journal of Grid Computing*, *9*, 65-79. doi: 10.1007/s10723-010-9175-7. Retrieved from http://www.springerlink.com/content/m28347312480n00j/fulltext.pdf.
[141] University of Melbourne/University of Barcelona (n.d.). *Amazon Web Services Case Studies*. Retrieved from http://aws.amazon.com/solutions/case-studies/university-melbourne-barcelona/.
[142] Douglas, B. (2012, September). Particle physics data analysis cluster for Atlas LHC experiment. *FutureGrid Portal*. Retrieved from https://portal.futuregrid.org/projects/257.
[143] Thomas Cleland – Understanding the complex computations of brain circuitry: Red Cloud with MATLAB case study (2012). *Cornell University Center for Advanced Computing*. Retrieved from http://www.cac.cornell.edu/about/studies/Cleland.pdf.
[144] Simulating muscle dynamics in the cloud. (n.d.). *University of Washington eScience Institute*. Retrieved from http://escience.washington.edu/get-help-now/dave-williams-simulating-muscle-dynamics-cloud.
[145] Jackson, P. (2012, October 23). Interview by P. Redfern [Personal Interview]. Cloud development of model-based systems engineering online textbook.
[146] Prassana, V. (2012). The P. Group Wiki: Smart Grid. University of Southern California. Retrieved from http://ganges.usc.edu/wiki/Smart_Grid.
[147] USC smart grid (n.d.). Retrieved from http://smartgrid.usc.edu/.
[148] Simmhan, Y., Deng, L., Kumbhare, A., Redekopp, M, & Prasanna, V. Scalable, secure analysis of social sciences data on the Azure platform, *IEEE International Scalable Computing Challenge (SCALE),* 2012 *(First Place)*.
[149] Chalklabs (n.d.) *Amazon Web Services Case Studies.* Retrieved from http://aws.amazon.com/solutions/case-studies/the-server-labs/.
[150] Kawahara, D. & Kurohashi, S. (2012). Argument structure analysis of huge web corpora for improving a search engine infrastructure. In D. Gannon & J. Vargas (Eds.), *The Cloud Computing Engagement Research Program* (pgs. 34-35). Seattle, WA: Microsoft.
[151] ICPSR: Find & analyze data (n.d.). Retrieved from http://www.icpsr.umich.edu/icpsrweb/ICPSR/.
[152] Hutchings, S., LeClere, F., & Beecher, B. (2012). Cloud computing and confidential data: an analysis of security implications and mitigation strategies. *Proceedings of 3rd International Conference on Society and Information Technologies (ICSIT 2012)*, Retrieved from http://www.iiis.org/CDs2012/CD2012IMC/ICSIT_2012/PapersPdf/HB806FZ.pdf.
[153] NIST cloud specific terms and definitions (2011, March 31). Retrieved from http://collaborate.nist.gov/twiki-cloud-computing/pub/CloudComputing/ReferenceArchitectureTaxonomy/Taxonomy_Terms_and_Definitions_version_1.pdf