True
Clouds and Scientific Computing

Enabling On-Demand Science via Cloud Computing

Kate Keahey
Argonne National Laboratory and University of Chicago

Manish Parashar
Rutgers University

//AUTHOR: *IEEE Cloud Computing* uses 25 word abstracts on the title page. Please feel free to suggest changes to the version below, if necessary. A longer abstract (see the end of this file) may appear in the Computer Society's Digital Library.//

The advantages of on-demand resource availability are making cloud computing a viable platform option for research and education that may enable new practices in science and engineering.

Infrastructure cloud computing has emerged as a new, revolutionary resource procurement paradigm that has been widely adopted by enterprises. Clouds provide on-demand access to computing utilities, user control over the computing environment, and an abstraction of unlimited computing resources—overall, a fundamental building block for on-demand scale up, scale out, and scale down. Furthermore, a diverse and dynamically federated marketplace of "cloud-of-clouds" can accommodate heterogeneous and highly dynamic application requirements by composing appropriate (public and/or private) cloud services and capabilities best suited to the needs of a given application. Clouds are also rapidly joining high-performance computing (HPC) systems, clusters, and grids as viable platforms for scientific exploration and discovery.[1]

Analogous to their role in enterprise IT, clouds can enable the outsourcing of many of the potentially distracting aspects of research and education, such as procuring, building, housing, and operating infrastructures, thus enabling research institutions to make science their primary focus. Furthermore, the ability to represent a computational environment as an appliance that different researchers can publish and then easily share enables the reproducibility of associated computations and thus facilitates sharing not only data but also new algorithms and methods. Additionally, similar to their enterprise role, clouds in the research and education context can democratize access to computational and data resources because institutions and individual researchers can lease powerful resources for a short time at relatively little cost.

It's important, however, to look beyond these benefits and understand application formulations and usage modes that are meaningful in a cloud-centric cyberinfrastructure (CI). We must also determine how a hybrid CI can enable new practices in science and engineering. Not all application patterns or usage patterns common in the scientific community lend themselves to the cloud computing platform, and likewise, not all the potential created by infrastructure clouds is currently being leveraged. As we look to the future needs of scientific platforms, it becomes clear that their characteristics will evolve to place additional requirements on computational support, in particular as extreme data and computing scales continue to transform and drive science and engineering research.

This article discusses the current and future needs of data-driven scientific exploration based on traditional as well as emergent scientific instruments and experiments. We explain how on-demand resource availability provided by cloud computing can become a vital part of such an instrument and discuss both opportunities and obstacles to cloud adoption in science. We then articulate challenges in research and current practices that need to be overcome to leverage those opportunities and overcome obstacles before developing cloud computing into a viable scientific platform. We conclude with recommendations for catalyzing integration of cloud computing opportunities into the current scientific landscape and discuss the significance of such an integration.

Science On Demand

Large-scale experiments, such as the Large Hadron Collider (LHC), equipped with millions of sensors and capable of producing up to petabytes of data per second, have highlighted the importance—or, rather, the criticality—of

computational support as an extension of scientific instruments. Data produced in such large quantities often must first be reduced by orders of magnitude in real time to a volume that can be stored at an acceptable cost. Data may even have to be analyzed in real time so that it can provide feedback during the experiment. Additionally, raw data must be processed into derived products that give actual insight into the observed phenomena and can be analyzed by groups of diverse scientists contributing their expertise and generating new scientific insight. Such processing must happen within the context of an experiment—that is, in real or near-real time. Thus, science is performed in "bursty" cycles, akin to the uptick of shopping during the Christmas season relative to other times of the year. This exploration pattern increasingly places a premium on the on-demand availability of resources, a demand that traditional batch-oriented computational centers can't always satisfy.

The advent of infrastructure cloud computing has had a tremendous, disruptive force in this space; it enables **the ability to lease** resources on demand with a preconfigured environment that guarantees correct and consistent execution. The transformation and the potential that this capability has opened up are exemplified by the **STAR //If this is an acronym, please spell out.//** nuclear physics experiment.[2] Using a traditional approach, a local cluster's computational capacity would have throttled the speed at which experimental results could be processed, and as a result, STAR scientists would have had to wait almost a year to assess the results of the experiment. With cloud computing resources, the STAR scientists were able to reduce this time to just three months—a significant difference in a competitive field. Furthermore, the scientists were able to run the data calibration component of processing concurrently with data collection, opening up the possibility of adaptively tuning the experimental parameters, a highly desirable capability.

Such advances are particularly interesting as we consider the types of experiments we are likely to conduct in the future. Inexpensive and increasingly sophisticated sensor devices now allow scientists to instrument ecological systems (such as oceans and rivers) or cities, turning our planet and everything in it into an "instrument at large"—dynamic, customized, and often self-organizing groups of sensors with outputs that we can aggregate and correlate to support experiments organized around specific questions. For example, structured deployments, such as the global network of flux towers, are being augmented by the innovative use of personal mobile devices (such as using cell phones to detect earthquakes), data from social networks, and even citizen science. Many of those sensors allow for adaptive feedback or can be combined with actuators that can control the experiment's environment. Driven by the proliferation of personal sensors marketed at large scales and technological progress (battery life, alternative energy sources, miniaturization, and so on) as well as economic factors (price), this trend is likely to continue accelerating and offering unprecedented opportunities to science.

The online analysis needs of such instruments at large are more challenging than those of traditional instruments. Although traditional instruments present demanding—but roughly known and finite—requirements for online processing, an instrument at large consists of a dynamic set of sensors that can become active or inactive at different times, for example, as their batteries run out, darkness prevents taking pictures, or social network sources become inaccessible. Furthermore, using carefully selected streams of spatial data from a variety of sources, scientists can uniquely customize experiments to answer specific questions.

Such levels of dynamicity and customization, however, will result in unpredictable and highly volatile requirements for the instruments that provide computational support. Moreover, based on the online inspection of data streams, a user may want to modify an experiment by accessing additional data streams or moving mobile sensors to different locations. Doing so requires a responsive infrastructure, capable of performing such an inspection within the required time constraints. Finally, whereas an experiment on a traditional instrument often has a well-defined beginning and end, experiments supported by instruments at large can—and often do—go on indefinitely. Such experiments require always-on, highly available computational support as incoming data is filtered, reduced, correlated, processed, and stored.

Given these requirements, computational support for this kind of experiment is clearly no longer an option; rather, it constitutes an inherent and indispensable component of an instrument at large. The groundbreaking possibilities created by such instruments will make them widely useful and a focus of activity over the coming years. The on-demand availability provided by cloud computing will be a fundamental building block for the support of experimental science, but further development is necessary to combine it with the additional infrastructure that satisfies the timeliness, scalability, and reliability requirements of experiment-driven processing. This situation places new emphasis and urgency on investigating the applicability of infrastructure clouds in science and shaping their capabilities and ecosystem into a viable and responsive scientific tool.

Clouds as Enablers of On-Demand Science

The opportunities offered by on-demand and data-driven science are compelling and could dramatically impact science, engineering, and society. Clouds can help make this vision a reality in multiple ways. They can provide resources for running applications on demand when local infrastructures are unavailable. They can also supplement existing systems by providing additional capacity or complementary capabilities to meet heterogeneous or dynamic needs. For example, clouds can serve as accelerators or provide resilience to scientific workflows by moving the workflow execution to alternative resources when a failure occurs.

Current cloud installations can also provide effective platforms for certain classes of computational and data-enabled science and engineering (CDS&E) applications—for example, high-throughput computing (HTC) applications. The cloud abstraction's simplicity can also alleviate some of the problems that scientific applications face in current HPC environments. For example, the increasingly important and growing class of many task computing (MTC) applications can benefit from the ease of use, the abstraction of elastic and readily accessible resources, and the ability to easily scale up, down, or out. Finally, the abstractions provided by the cloud model will allow scientists to address their problems more effectively and can even enable them to formulate their applications and algorithms in new ways.

Several early projects have reported successful deployments of such applications on existing clouds.[3,4] Running these applications typically involves using virtualized commodity-based hardware, which is provisioned on demand at commercial cloud providers such as the Amazon Elastic Computer Cloud (EC2) or Microsoft Azure. A recent technical report by Geoffrey Fox and Dennis Gannon provided an extensive study of HPC applications in the cloud.[5] According to that study, commodity clouds work effectively only for specific classes of HPC applications. Application examples include *embarrassingly parallel applications* (those that are efficiently distributed and can withstand latency across networked systems **//definition okay? I'm not sure readers will be familiar with this phrase.//**) that analyze independent data or spawn independent simulations that integrate distributed sensor data, science gateways and portals, or data analytics that can use MapReduce-like formulations.

However, although cloud computing has been around for more than a decade, has developed rapidly in that time, and has been enthusiastically adopted by many branches of industry, its broad adoption in science has been slow. Several factors play a part in the applicability of cloud computing to science.

Moving the Mountain

//The *Cloud Computing* EIC was leery of the original religious reference here. I reworded slightly to maintain your analogy, which still works even if we don't refer to Mohammad.//

When we consider whether clouds can provide a suitable platform for HPC, we tend to think in terms of whether the mountain will come to **us**—in other words, will cloud computing evolve to suit the needs of traditional HPC? Although clouds can accommodate ever-larger computations (the largest known virtual cluster currently stands at 156,000 cores), the characteristics of such "supercomputers" differ from those of traditional HPC machines. The most frequently cited reasons for the lack of adoption of cloud computing in science are technical factors: inadequate performance and management options, size, and a lack of understanding of reliability concerns of large computations.

The impact of these limitations can be significant.[6] As the underlying technology that allows cloud providers to enable users to project their images onto their infrastructure securely, quickly, and reliably to support on-demand availability, virtualization imposes a performance penalty. Furthermore, research is still ongoing regarding how best to integrate cloud computing with hardware accelerators and fast communication hardware such as InfiniBand. Work on lightweight hypervisors holds out promise that, in the future, virtualized supercomputers can offer performance close to those of HPC resources.[7] However, a host of features, ranging from HPC-specific resource management to reliability, make computing at large scales using infrastructure clouds a challenge.

Other ways of looking at this question may exist, however—**we** may come to the mountain. As HPC resources grow in scale to include ever-larger numbers of processing elements, sustaining a model when all of them must reliably move in lockstep is increasingly hard. This situation opens the possibility of turning to more loosely coupled and asynchronous computational models. Key research challenges here include exploring application formulations that can effectively utilize clouds and addressing, at the algorithmic level, implications of cloud characteristics such as elasticity, on-demand provisioning, virtualization, multitenancy, and failure. For example, the asynchronous replica exchange[8] formulation is a novel, decentralized, asynchronous, and resilient formulation of the "replica exchange" algorithm for simulating protein structure, folding, and dynamics.

Another key research challenge is developing appropriate programming models and systems that can enable CDS&E applications to take advantage of clouds. These include developing programming abstractions and tools to support the federation of clouds and CIs to provide elastic access to cloud services and extending existing cloud programming models and platforms (such as MapReduce and BigTable) to support scientific computing. Additionally, entire applications will need to be exported, facilitating a need for new application patterns and kernels, optimized libraries, and/or specialized middleware as a service. New tools will thus be necessary for application debugging, validation management, and performance engineering.

Wholesale to Retail

A more insidious problem with cloud computing adoption within the scientific community is social rather than technological. Perhaps the most important aspect of the cloud computing disruption is that it has revolutionized our idea of resource procurement. Instead of buying a system wholesale to run a certain class of computations—an investment that can cost millions of dollars to buy, house, build, and operate—we can now shop retail and spend only a few thousand dollars on a per-computation basis as the need arises. This capability makes the time-capacity product more flexible. For example, instead of buying a small cluster and waiting a year for a computation to complete, researcher can now "rent" a large cluster for a short time and complete the computation using all the available resources.

Making this equation work, however, requires admitting that there is a premium on time, in other words, acknowledging that the instantaneous, on-demand availability of a resource is worth more than batch cycles and that this should be reflected in the market price of time on said resources. Currently, established funding, procurement, and allocation systems aren't equipped to deal with such a nuanced and multifaceted concept of worth, even if it can bring substantial benefits.

And now that we have computing power "on tap," turning the tap on proves to be a nontrivial operation. Previously, maintenance and user support were provided as part of the wholesale purchase; a traditional cluster user would expect it to be configured and upgraded as needed and to include all the standard software. In contrast, the cloud currently provides some features, such as resource availability, but doesn't provide other features, such as virtual machine configuration. Moreover, choosing the optimal configuration among the myriad cloud offerings—including diverse services, instance types, billing models, storage options, and providers—requires special expertise and a significant time commitment.

The Case of the Missing Infrastructure

Many of the challenges we outline here are, arguably, merely the growing pains of a deceptively simple but deeply disruptive innovation. Certainly, many of them can be resolved with a research and development investment in critical ecosystem components and by creating new support relationships that provide the necessary layer between users/applications and cloud services.

A relatively short-term challenge is establishing a cloud ecosystem that can enable and drive research and can address issues related to deployment and transition to practice. Research issues include the definition of community standards, development of community testbeds and benchmarks, documentation of experiences and best practices, and development of curricula and training modules. Providing appliances that can be automatically rendered as consistent sets of images working across virtual machine image formats (such as Xen and the Kernel-based Virtual Machine, or KVM) and cloud providers (including Amazon Web Services and Microsoft Azure) will provide a solution to the image problem. Similarly, higher-level abstractions for science at the platform-as-a-service (PaaS) and software-as-a-service (SaaS) levels can make clouds more accessible to scientists. Defining a community administrator in charge of the computing environment for a given community will also help. Services and informative benchmarks for service comparison, in terms of performance, consistency, and reliability, together with a better understanding and automated characterization of the load, can all help resolve the new complexity problem.

A platform layer that can take all this information and automatically build and maintain a user- or application-specific platform, clearly indicating the trade-offs and their implications, will make that tuning exercise simple. Middleware stacks and services are essential for supporting CDS&E application formulations and hybrid usage modes targeted to cloud and CI environments, including support for dynamic cloud bursting and infrastructure federation. A key research issue is provisioning, scheduling, management, and optimization of these hybrid infrastructures with respect to multiple objectives including performance, energy, cost, and reliability. Data management research challenges exploring the different types of cloud storage solutions and the nature of cloud connectivity are also important; specific issues include support for selecting from among the diverse storage options with varying service levels, networks architectures to support data transport needs and their interaction with cloud

storage offerings, and the colocation of computing and data. Combining **those two areas** of exploration into support for cyberphysical systems will ultimately provide a viable platform for instruments at large.

Lack of understanding of security and privacy issues as they relate to clouds is a critical barrier to adoption, especially in areas dealing with private data such as biomedical applications. Clouds renegotiate the security space with new types of attacks proposed all the time, emphasizing the need for high-quality security mechanisms because of the sharing of storage and computing.

In addition to crosscutting cloud security challenges, specific issues related to cloud and CI integration with CDS&E include the interoperability with broader CI security mechanisms and policies, such as single-sign-on, federated identify management (such as inCommon, cilogin, and SCIM), and security policies and mechanisms for specific applications (including differential privacy and data anonymization requirements for bio/medical informatics applications). Investments in homomorphic or partial homomorphic encryption are driven largely by the needs of those applications.

The Path Forward

What can we do to overcome obstacles to adoption of a promising innovation and catalyze its impact? We propose several ideas that can accelerate the development of cloud computing capabilities relevant to science and promote an understanding of the impact of its power.

Throw Down a Challenge

A well-defined challenge that captures the gradient of missing capability can be an effective vehicle of progress. Successive milestones in responding to such a challenge can be a good yardstick for judging the state of the art of a promising technology, as projects such as the Top 500 list have successfully demonstrated. Such challenges are also effective in nucleating a community. On the cloud computing frontier, such challenges to date have been driven more by what we can answer than by what we'd like to know. This approach highlights the strengths of a technology, but it doesn't fully relate it to the context of surrounding requirements. Offering specific problem formulations as well as benchmarks and metrics in collaboration with the scientific community will help address this shortcoming and highlight areas in which additional work is necessary.

Construct Experimental Testbeds

An open, reconfigurable experimental testbed—large enough to reflect the scale appropriate to handle the big data and big compute challenges we face—is as critical to the advancement of computer science as large instruments such as LHC are to the advancement of physical sciences. A testbed alone is insufficient, however. Data that can lead to specific problem formulations, such as cloud utilization data, is of critical importance as well. This data is often available only from commercial providers, and thus collaboration between academia and industry emerges as a critical ecosystem element of such a testbed.

Another critical enabler is the deep familiarity with specific usage patterns that can be obtained only by working directly with application scientists. Open access to such resources, problems, and data will create a community operating within the same collaborative context and thus capable of creating research that is more than the sum of its parts. A viable experimental testbed should therefore place emphasis on building such a community.

Standards, Policies, and Practices

Many important standard activities exist, from those specifying the basic virtual machine structure to higher-level standards defining the PaaS/SaaS environment. Although these standards, such as the Open Grid Forum's Open Cloud Computing Interface (OCCI) in OpenNebula and OpenStack, have some support, this area is still under development, with the US National Institute of Standards and Technology (NIST) and IEEE playing leadership roles. In addition, substantial progress is needed to enable the procurement of services through capped purchase orders or subcontracts; subaccount administration; resource and authority delegation; and monitoring, managing, and reporting. Furthermore, the development and modification of codes adapted to the cloud environment require a unique skill set that necessitates appropriate educational and training structures.

Practice Makes Perfect

Some problems can be fully understood and resolved only by facilitating the use of clouds in practice, in the context

of specific applications or application groups, and by experiencing and solving problems on the fly. Encouraging cloud-based application platforms will lead to solutions that offer practical solutions and thereby generate more confidence, familiarity, and expertise related to this emergent platform.

Although industry has enthusiastically embraced cloud computing, and it has demonstrated enticing possibilities for various branches of science—particularly those that place a premium on on-demand availability such as the experimental sciences—cloud computing currently runs the risk of getting stuck crossing the chasm between potential and reality in its application to scientific problems. This impasse is due to the computationally demanding nature of scientific applications, both in terms of performance and infrastructure support, as well as the lack of economic flexibility in the scientific environment. Catalyzing progress in this space is an important activity.

As we look to the future and ponder the needs of technologies underlying future experimental instruments integrating computation as their inherent component, we can see this will become all the more important. Such computations will rely on the on-demand availability and control over the environment provided by infrastructure clouds. They will also require support for the big compute applications currently running in HPC centers. Finding ways to overcome the performance, use model, and infrastructure barriers currently dividing both models is therefore of primary importance.

This work was supported by the U. S. Department of Energy, Office of Science, under Contract No. DE-AC02-06CH11357.

References

1. M. Parashar et al., "Cloud Paradigms and Practices for Computational and Data-Enabled Science and Engineering," *Computing in Science & Eng.*, vol. 15, no. 4, 2013, pp. 10–18.
2. J. Balewski et al., "Offloading Peak Processing to Virtual Farm by STAR Experiment at RHIC," *J. Physics Conf. Series*, 2012, p. 368.
3. E. Deelman et al., "The Cost of Doing Science on the Cloud: The Montage Example," *Proc. 2008 ACM/IEEE Conf. Supercomputing*, 2008, pp. 1–12.
4. K. Keahey and T. Freeman, "Science Clouds: Early Experiences in Cloud Computing for Scientific Applications," *Proc. Cloud Computing and Its Applications*, 2008, pp. 825–830.
5. G. Fox and D. Gannon, "Cloud Programming Paradigms for Technical Computing Applications," *Proc. Cloud Futures Workshop*, 2012, //**page range?**//.
6. K. Yelick et al., *The Magellan Report on Cloud Computing for Science*, Office of Science and Office of Advanced Scientific Computing Research (ASCR), US Dept. of Energy, 2011.
7. J. Lange et al., "Minimal Overhead Virtualization of a Large Scale Supercomputer," *Proc. 2011 ACM SIGPLAN/SIGOPS Int'l Conf. Virtual Execution Environments* (VEE), 2011, pp. 169–180.
8. Z. Li and M. Parashar, "Grid-Based Asynchronous Replica Exchange," *Proc. 8th IEEE/ACM Int'l Conf. Grid Computing*, 2007, pp. 201–208.

Kate Keahey *is a scientist in the Mathematics and Computer Science Division at Argonne National Laboratory and a Computation Institute fellow at the University of Chicago. She is also the creator and leader of the Nimbus Project. Her research interests include virtualization, resource management, and cloud computing. Keahey has a* ***//highest degree?//*** *in* ***//what subject?//*** *from* ***//university?//****. Contact her at keahey@mcs.anl.gov.*

Manish Parashar *is professor of electrical and computer engineering at Rutgers University. He is also the founding director of the Rutgers Discovery Informatics Institute (RDI2), the NSF Cloud and Autonomic Computing Center (CAC) at Rutgers (CAC@Rutgers), and the Applied Software Systems Laboratory (TASSL), as well as the associate director of the Rutgers Center for Information Assurance (RUCIA). His research interests are in parallel and distributed computing, with a focus on computational and data-enabled science and engineering. Manish has a PhD in computer engineering from Syracuse University. Contact him at parashar@rutgers.edu.*

//digital library abstract & keywords//

This article discusses the current and future needs of data-driven scientific exploration based on traditional as well as emergent scientific instruments and experiments. The authors explain how the on-demand resource availability provided by cloud computing can become a vital part of scientific research and discuss both the opportunities and obstacles to cloud adoption in science. Still, several challenges in research and current practices must be overcome to leverage those opportunities and overcome obstacles to developing cloud computing into a viable scientific platform. The authors offer recommendations for catalyzing integration of cloud computing opportunities into the current scientific landscape and discuss the significance of such integration.

cloud, cloud computing, scientific computing, orig-research

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.