A Perspective on Scientific Cloud Computing

Craig A. Lee
Open Grid Forum, www.ogf.org
and
Computer Systems Research Department
The Aerospace Corporation
lee@aero.org

ABSTRACT

Cloud computing has the potential for tremendous benefits, but wide scale adoption has a range of challenges that must be met. We review these challenges and how they relate to scientific computing. To achieve the portability, interoperability, and economies of scale that clouds offer, it is clear that common design principles must be widely adopted in both the user community and marketplace. To this end, we argue that a private-to-public cloud deployment trajectory will be very common, if not dominant. This trajectory can be used to define a progression of needed common practices and standards which, in turn, can be used to define deployment, development and fundamental research agendas. We then survey the cloud standards landscape and how the standards process could be driven by major stakeholders, e.g., large user groups, vendors, and governments, to achieve scientific and national objectives. We conclude with a call to action for stakeholders to actively engage in driving this process to a successful conclusion.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: General

General Terms

Standardization

Keywords

Cloud computing, deployment trajectory, standardization

1. INTRODUCTION

Cloud computing is enjoying a tremendous level of interest across virtually all areas of computational use, including the scientific computing community [13, 17, 18]. While there are certainly many issues to resolve and pitfalls to avoid, the argument can be made that cloud computing may have a similar impact as that of cluster computing. The use of commodity processors and networks to build cluster computers fundamentally changed not only the composition of the Top500 List, but also the economic availability of supercomputing across all fields. The on-demand nature and economies of scale for cloud computing may do the same thing for science.

When cluster computing was gaining popularity, many people argued that the relatively slower commodity networks would hamper the performance of parallel codes, relative to that of parallel supercomputers with dedicated (and expensive) interconnects. While this may have been true for tightly coupled codes that were very bandwidth and latency-sensitive, for many codes the performance was quite adequate with an undeniable cost-effectiveness. Whatever performance issues may exist for cloud computing, there will be many scientific endeavors where the on-demand nature and cost-effectiveness will far outweigh any performance degradation. Across the scientific computing landscape, there is clearly a distribution of requirements for available compute power and data access that is necessary to facilitate progress. While there will always be a need for massive, parallel machines, there will also be a need for reasonable amounts of compute power, on-demand, at a reasonable cost.

While such performance-related arguments are very important to the scientific computing community, it is also important to understand that the science cloud concept is happening in the context of a much larger phase change in distributed computing environments. It is the mission of the Open Grid Forum (OGF) to understand, and to the best extent possible, manage this phase change on behalf of its members, stakeholders, and the wider distributed computing community. It is widely held that common best practices and standards will be needed to realize many of the benefits being touted for cloud computing. To that end, this paper argues for a perspective on how this phase change will occur and strategic efforts that could be taken to maximize its benefits.

2. A DISCUSSION OF CLOUDS

We begin by briefly discussing the expected benefits of cloud computing, in general, followed by outstanding issues and challenges. We also give a key example of the larger motivations to address the challenges and realize the benefits.

2.1 Benefits

The anticipated benefits of cloud computing can be broadly categorized into *infrastructure-oriented* benefits and *user-oriented* benefits. User-oriented benefits are those that an

individual user would be able to realize, while infrastructure-benefits are those that an infrastructure provider or data center operator would be able to realize across a large, aggregate set of users. Infrastructure-oriented benefits include:

- *Improved server utilization.* The use of virtual machines and virtual machine images provides flexibility in mapping work to physical servers, thereby allowing higher utilization to be achieved.
- *Improved reliability.* Likewise, the use of virtual machines can facilitate fail-over between physical servers.
- *Greener IT.* Energy consumption and costs can be reduced through improved utilization and moving work to where the cheaper energy is available.
- *Clear business models.* By providing resources through a simplified API that abstracts away many infrastructure details, clear consumer-provider business models are possible.

User-oriented benefits include:

- *Commodification of compute resources.* The commodification of any product means that it is no longer a specialized resource that must be uniquely designed, installed and maintained. It can be bought, sold, and replaced as needed, without costly re-engineering, etc.
- *Managing surge requirements with on-demand resources.* Since commodification allows resources to acquired and released on-demand, this allows users to more easily manage expected and unexpected surges in compute requirements.
- *Ease of deployment.* The use of virtual machine images can also ease the deployment of applications since the machine image may contain the exact OS, libraries, patches and application code necessary to execute.
- *Virtual ownership of resources.* Rather than having to deal with a shared resource, and the access contention that can go with it, users enjoy the ownership of a resource that is available at their beck and call, even if that ownership is virtual, as in a computing cloud.

We note that these benefits will become more significant when operating "at scale", especially for cloud providers. For cloud consumers, if only a small number of processes or servers are required for a given application, then existing in-house resources may be available and sufficient. However, as the number of required servers and storage increases, and surge requirements become more unpredictable, then the on-demand nature of commodity resources becomes more attractive. For cloud providers, the benefits realized must leverage the *economies of scale* made possible by providing cloud resources out of massive data centers.

2.2 Issues

While these expected benefits are driving much of the interest in cloud computing, there are nonetheless a range of significant issues that both providers and consumers must address. These include:

- *Security* is widely stated as the primary issue facing cloud computing. A cloud-specific security issue is that of running arbitrary VM images. This is, however, only one aspect of the broader notion of *Information Assurance* [2], i.e., making sure the right data is available to the right user at the right time. In addition to the fundamental security operations of authentication, authorization, privacy, integrity and non-repudiation, information assurance also involves data reliability and availability. The Cloud Security Alliance [3] is directly pursuing many of these issues.
- *Deployment models.* The typical cloud deployment models, i.e,. public, private, hybrid and federated, strongly affect the security and information assurance issues of a given cloud. More on this later.
- Service level agreements. While the simplified API of current commercial cloud offerings is critical for providing a lower barrier of adoption and supporting clear business models, it complicates the notion of user control. User applications may have very specific performance or behavior requirements, in addition to regulatory policies that must be observed. To avoid exposing unnecessary infrastructure detail through the APIs, cloud providers must provide the right abstractions through service level agreements for effectively specifying the desired behavior or policy.
- *Governance.* Strongly related to the notion of service level agreements and policy, is that of governance – how to manage sets of virtual resources. Especially at the infrastructure level, applications may consist of many virtual machines, virtualize storage, and virtual networks. Managing these *virtual missions*, or *virtual data centers*, will require policy and enforcement from both the provider and consumer.
- *Cost* is an outstanding issue – especially for public and hybrid cloud use. Depending on an application's compute, storage, and communication requirements, public cloud resources could be more or less expensive than hosting the application in-house [13]. Even if it is cheaper to host an application in-house, if the application has variable surge requirements, there will be some break-even point where it makes sense to cloud-burst into a public cloud. Quantitatively evaluating such break-even points will require costing models that adequately capture an application's dynamic resource requirements [19].

2.3 A Key Example

While industry is vigorously pursuing cloud computing, both from the provider and consumer sides, for the reasons cited above, the key example we wish to give here is the US federal IT budget. As illustrated in Figure 1, the FY 2010 US federal IT budget is \$79B, of which ∼70% will be spent on maintenance [12]. The US federal CIO has publicly declared that cloud computing will be adopted to reduce this budget by eliminating redundant IT capacity across federal agencies [1]. To this end, the web sites `data.gov` and `apps.gov` have been stood-up whereby government data can be made more accessible, and government agencies can shop for software and also acquire cloud resources. The Eucalyptus-based

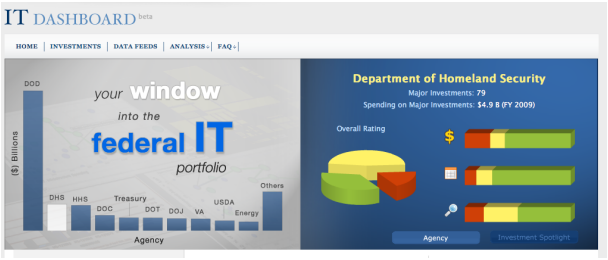

Figure 1: The US Federal IT Budget.

Nebula Cloud at NASA Ames is to be the first cloud backend for `apps.gov` to service federal computing requirements on-demand [16].

3. A DISCUSSION OF SCIENCE CLOUDS

Commercially available public clouds have been designed to satisfy general computing requirements, e.g., e-commerce and transactional communications, that are typically less sensitive to bandwidth and latency. As clouds become more mature, however, it is anticipated that clouds of different "flavors" will be deployed to meet the requirements of different user communities, e.g., scientific computing. Hence, while all of the potential benefits and issues of general cloud computing are relevant to the scientific computing community, the notion of *science clouds* will force an emphasis on specific benefits and issues.

3.1 Benefits

- *Identity and Federation management.* Currently available public clouds have relatively simple authentication mechanisms that are suitable for basic client-provider interaction. The established grid community, however, has extensive experience in identity management and federation management. That is to say, the scientific enterprise may involve the federation of distributed resources and the management of *Virtual Organizations* and the management of user roles and authorizations within a given *VO*. Given the wide industrial interest in cloud computing, it is anticipated that business-to-business interactions will involve, and eventually adopt, similar identity and federation management concepts and implementations.

- *Virtual ownership of resources.* This represents potentially the largest beneficial change for science clouds. The scientific computing community is very familiar dealing with batch schedulers, but nonetheless, the illusion of having "your own" resources or set of nodes is very attractive. In much the same way that private car ownership offers benefits (and trade-offs) with respect to public transportation, virtual ownership of cloud resources will reduce uncertainty concerning access to those resources when you need to use them.

- *Ease of deployment.* In traditional grid environments, where specific machine resources are directly exposed and available to users, the deployment of applications depends on explicitly managing the compatibility among binaries, operating systems, libraries, etc. The use of virtual machine images offers the ability to package the exact OS, libraries, patches, and application codes together for deployment. This doesn't make the configuration management problem completely trivial, but it does reduce the problem to guest OS-host OS compatibility. For scientific codes, however, this can have direct implications for numerical stability across different virtual platforms.

3.2 Issues

- *Performance Management: Abstraction vs. Control.* While scientific computing can be broadly categorized into *high performance computing* and *high throughput computing* that have very different workload characteristics, it is clear that scientific computing will have significant issues concerning *abstraction versus control*. There is a fundamental trade-off between the simplicity that abstraction can provide versus the ability to control application behavior by having visibility and control over the underlying resource infrastructure. Hence, the satisfaction of performance management issues for scientific applications will depend on what minimal abstractions can be exposed to users that enable the desired performance behaviors to be adequately controlled. Clearly it may be possible to expose such minimal abstractions through the use of service-level agreements whereby the necessary coupling of resources can be specified, e.g., tightly coupled computing clusters, compute-data locality, bandwidth and latency requirements, etc. This is a major outstanding issue for the deployment and use of effective science clouds.

- *Data Access and Interoperability* continues to be an outstanding issue for all inherently distributed applications and federated organizations. While this is not specific to cloud computing environments, the use of dynamically provisioned resources will underscore the need to easily integrate disparate data sources and repositories. In addition to supporting things like high-throughput computing, effective data access and interoperability will enable science that can't be done any other way. As an example, access to oceanic databases from different geographic regions of North America has enabled fundamental insights that were otherwise not possible. [14]

- *Execution Models, Frameworks and SaaS.* While dynamically provisioned resources at the infrastructure level have many advantages as previously mentioned, their effective management and utilization by the end-user will present challenges. Various execution models can be identified that provide useful abstractions and make scientific infrastructure resources easier to use. Map-Reduce is one popular tool for partitioning the processing that must be done across massive data stores. Data streaming applications, such as signal processing, could also be similarly supported. Parameter sweep applications have existing tool kits that could be adapted for cloud environments. All of the execution models could be supported by frameworks that make them easier to use. From the cloud perspective, this could be tantamount to providing *Software as a Service*. For scientific purposes, the SaaS concept can even be extended to the notion of *Models as a Service (MaaS)* whereby semantic annotations and

ontologies are used to compose computational models and execute them as a conceptual whole [5]. Clearly a distinguishing property of science clouds may be the availability of such high-level abstractions that can effectively use the infrastructure resources to provide top-to-bottom support for scientific computing goals.

3.3 A Key Example

While the economics of on-demand resources will drive and support a large segment of scientific users, the key example we wish to relate here is that of *operational hurricane forecasting*. Figure 2 from Bogden et al. [15] compares several hurricane track forecasting models applied to Hurricane Katrina in 2005, from one to five days in advance. The black line indicates the actual path. At five and four days out, there is essentially no agreement among these models. At three days out, the models are converging – but to a incorrect track. Finally at two days and one day out, the models tend to agree with ground truth.

The effective prediction of hurricane tracks, in addition to disaster response and mitigation, actually represents a *scientific* and *operational grand challenge problem*. This will require a fundamentally enhanced understanding of how the atmospheric and oceanic systems work, in addition to developing the computational models that accurately represent the physical systems. When a tropic depression becomes a hurricane, the results of these tracking models will have to be fed into precipitation models to determine where water will be deposited, which will have to be fed into flooding models to determine where lives and property will be at risk. To accomplish this will require an enormous amount of compute power that is just not economically possible to dedicate to this single purpose. Hence, shared resources will have to be used, but they must also be available *on-demand*, possibly from a national science cloud that can support coupled, HPC codes with strict processing deadlines.

4. THE CLOUD STANDARDS LANDSCAPE

It is clear that common best practices and standards will be needed to achieve the fundamental properties of portability and interoperability for cloud applications and environments. Portability will mean more than simply being able to run to completion without fatal errors – it will mean being able to preserve critical properties, such as performance, numerical stability, and monitoring. Interoperability will mean more than avoiding vendor lock-in – it will mean being able to avoid "cloud silos" that are non-interoperable since they are build on different APIs, protocols, and software stacks.

Beyond the issues of portability and interoperability from a user's perspective, we can also consider the notions of virtual organizations, virtual missions, or virtual data centers. As mentioned above, VOs address the issue of managing user roles and authorizations within a collaboration of organizations for a specific goal or purpose. VOs or enterprise-scale applications may require the deployment of many servers and functions that must be managed as a whole. That is to say, large applications may be deployed as sets of virtual machines, virtual storage, and virtual networks to support different functional components. Another perspective is that such large applications may be deployed as *virtual missions*

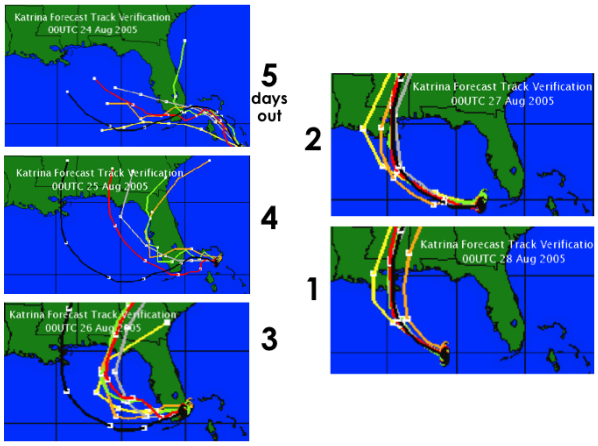

Figure 2: Predicting Hurricane Katrina.

that constitute their own *virtual data centers* allocated out of larger, physical data centers, i.e., clouds.

Such notions will be important to all large cloud providers and consumers. As a case in point, several *national cloud initiatives* have been announced. The US Cloud Storefront [11], the UK G-Cloud [10], and the Japanese Kasumigaseki [6] cloud initiatives will be major stakeholders in cloud standards since they will ultimately involve large cloud deployments. While these national cloud initiatives will support many routine IT business functions to reduce redundancy and improve utilization across government agencies, it is quite possible that these national clouds will support a range of application types and requirements. That is to say, while science clouds may or may not be deployed as part of such national clouds, they could nonetheless benefit from the ability of national cloud initiatives to drive the development of relevant best practices and standards.

To summarize, science clouds and national clouds may share the following similar requirements:

- Applications be transferable out and back into the same cloud, and between different clouds, and still retain desired properties, such as performance, numerical stability, monitoring for fault detection/resolution, etc.
- A registry or brokerage be available for the discovery of available resources and service level agreements for using them.
- Support for virtual organizations across different science clouds and national clouds.
- Support for virtual missions and virtual data centers allocated out of different science clouds and national clouds.

An outstanding issue that must be resolved is how to prioritize these requirements and structure their development for the benefit of national scientific goals.

4.1 Deployment Trajectories

These requirements must also be considered in the context of cloud deployment models. The relationship of private, public, hybrid and federated clouds is illustrated in Figure 3. Private clouds are typically deployed behind an organization's firewall where access and the user population is known and managed according to organizational goals. Public clouds, by contrast, have a very open admission policy, typically based on the ability to pay. Managing surge requirements, one of the key user-oriented benefits, is commonly called *cloudbursting* whereby an organization can acquire public cloud resources on-demand to form a *hybrid cloud*. We can also consider the case where two or more private clouds wish to interact for common goals and thereby form a *federated cloud* (also called a community cloud).

Organizations will adopt different deployment models based on their particular requirements. This casts the differences between private and public clouds into sharp relief. Public clouds offer wide availability and economies of scale that only very large data center operators can achieve, but are perceived to have a host of security and information assurance issues. When using a public cloud, any security or assurance requirements are essentially delegated to the cloud provider, where the user may or may not have sufficient insight and trust of the provider to satisfy the necessary requirements. Hence, many organizations with significant IT and security requirements that wish to realize the benefits of cloud computing will deploy their own private cloud.

To illustrate this issue, Figure 3 is labeled in the context of different government agencies that may have deployed their own private clouds, but may also wish to federate with other government agencies, or use a national public cloud. It would be easy to relabel this diagram for any sub-organizations that interact with peers, or a parent organization. Indeed, the primary distinction between private and public clouds may be more of a relative distinction concerning the ownership of resources and the ability to enforce security policies, etc., versus delegating those responsibilities to a second party.

This dichotomy between public and private clouds raises the issue of *deployment trajectory*. Will public cloud adoption be predominant in the community, or will private cloud deployment be predominant? In the context of governmental or national clouds, which will be predominant?

Despite the fact that national clouds are being designed and deployed, various government agencies are already deploying their own private clouds. This dynamic can be characterized as a *top-down* vs. a *bottom-up* approach. The top-down, national public cloud approach has the challenge of recruiting enough users whose security and assurance requirements can be met. The bottom-up, private cloud approach has the challenge of mitigating the risk of creating non-interoperable cloud silos that cannot federate or hybridize.

What will be the dominant deployment trajectory in the context of science clouds? While commercially available public clouds could be used for scientific computing, they were not designed with such applications in mind. Hence, different science clouds may be deployed to meet various

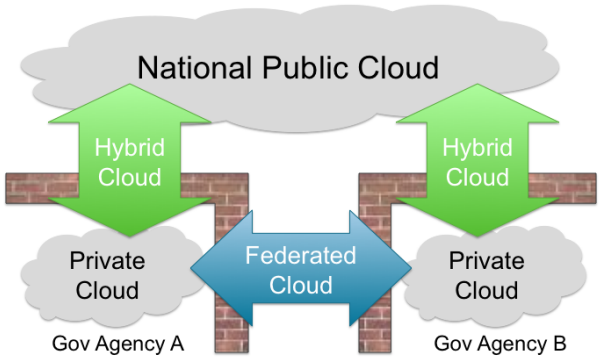

Figure 3: The Relationship of Cloud Deployment Models.

performance and control requirements for various scientific application domains. If this is the case, then science clouds may experience a similar deployment trajectory as national clouds.

If we assume that the predominant cloud deployment trajectory will start with private clouds, and then progress to federated, hybrid, and finally public clouds, then we can consider the progression of deployment issues and capabilities that must be addressed in sequence. Such a progression is summarized in Figure 4. Within a private cloud, the cloud operator has explicit knowledge of all users, can set policy and can use traditional security mechanisms with a secure firewall perimeter. When two private clouds federate, there can be explicit, out-of-band knowledge about the joint users and agreement about policy and security mechanisms. When a private cloud acquires some public cloud resources to become a hybrid cloud, the private cloud operator must deal with resources that may have an unknown number of unknown tenants, but at least the private cloud operator has a secure perimeter whereby they can make a decision about which data and workloads are stored and processing inside and outside of the secure perimeter. Finally, if all data and operations are hosted in a public cloud, the user has no explicit knowledge of the other tenants and has delegated all management and security requirements to the public cloud operator.

This progression of issues and concerns that accumulate from left to right can be further partitioned into necessary technical capabilities, legal and organizational issues, and for lack of a better term, "landscape" issues concerning the forging of agreements across user groups, vendors, and major stakeholders. The technical issues progress from basic capabilities such as managing different job types and mixes thereof, workload management, and governance, through capabilities that would be needed for managing federated clouds, such as virtual organizations, to highly theoretical topics in public clouds, such as practical ways to operate on encrypted data which currently do not exist. With regards to legal and organizational issues, costing models to help organizations evaluate the potential cost benefits for their own computing requirements would be useful. This progresses through joint organizations to manage distributed infrastructures, such as the International Grid Trust Feder-

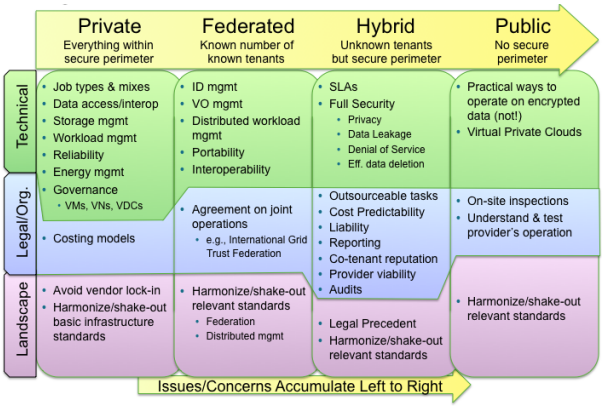

Figure 4: A Trajectory of Deployment Issues.

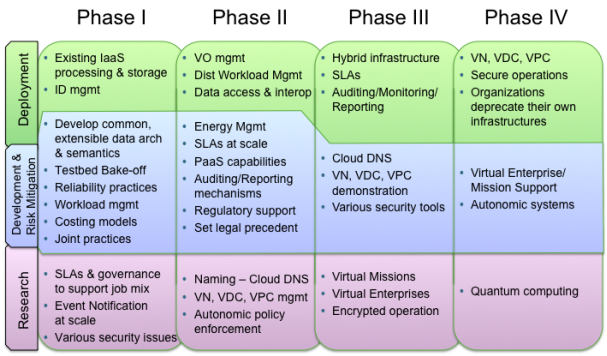

Figure 5: A Draft Deployment, Development, and Research Roadmap.

ation, and then arrives at the many issues concerning client management of their service and security requirements when dealing with a public cloud provider. The landscape issues also progress across this deployment trajectory where users, vendors, and standards organizations will need to collaborate on the harmonization and shake-out of various efforts such that there is an emergent dominant best practice in the user community ideally codified in a standard.

While this progression of issues and necessary capabilities could be driven down into much more detail, this structuring approach can nonetheless be used to derive deployment, development, and research agendas or roadmaps, as shown in Figure 5. Here we somewhat arbitrarily partition the roadmap into four phases. Each of these phases are also partitioned into deployment, development & risk mitigation, and research issues. Phase I deployment can essentially proceed immediately, while development and risk mitigation efforts must be done first before they can be deployed in subsequent phases. Fundamental research issues must also be identified whereby experimental work can be done across the design space for a particular capability, prior to subsequent development phases and eventual deployment. Hence, there is a general promotion of concepts and capabilities from lower left to upper right as they mature. In the bottom right, we can consider longer-term research issues, such as quantum computing, which might have an impact across all areas of computing, including that of clouds.

Clearly this roadmap could be driven into much more detail, based on the progression of requirements arising from a private to public cloud deployment trajectory. National cloud initiatives will certainly have their own roadmap activities, but it would benefit all involved to compare notes and coordinate efforts on common, necessary items on their development and research agendas. Science clouds should definitely be part of this wider national discussion such that scientific computing requirements are directly addressed.

4.2 Standardization Trajectories

In addition to considering the effect of deployment trajectories on necessary cloud standards, we must also consider where in the software stack that standardization might be most useful and effective. In addition to coordinating major stakeholders, we must also try to understand the market dynamics that may also drive the trajectory of standards. Cloud computing is often categorized as *Infrastructure as a Service (IaaS)*, *Platform as a Service (PaaS)*, and *Software as a Service (SaaS)*, ascending from the lower to higher levels in the system software stack. Some argue that standardization at the IaaS level will progress faster since this entails the commodification of compute resources, e.g., servers, storage, and networks, and that providers will start to compete on price and the quality of their product, as determined by service level agreements. This same argument says that innovation and product differentiation will predominantly occur at the SaaS level where providers will target specific market segments that are supported by commodity resources.

A contrary argument is that standardization will primarily occur at the SaaS level since customers in specific market segments, such as customer relations management, will demand portability and interoperability, resulting in a common look-and-feel. SaaS providers will then be free to implement their services in any way possible "on the back-end" in their data center.

We argue that standardization will primarily occur at the infrastructure level since the commoditization of basic resources will have the farthest reaching impact across all market segments; industry, commerce, government, and science. To this end, we describe a currently developing set of standards for cloud computing at the infrastructure level.

The Open Cloud Computing Interface (OCCI) [7] from OGF is a simple, RESTful API for managing the lifecyle of compute, storage, and network resources, as illustrated in Figure 6. Here the basic create, read, update, and delete (CRUD) operations can be applied to URLs that identify the specific providers and resources in question. Attributes are maintained for each resource instance, along with resource links that identify sets of related resources that are being managed as a whole.

OCCI can be used with the Open Virtualization Format (OVF) [4] from DMTF. As illustrated in Figure 7, OVF is essentially a representation format for virtual machines that can be used as the "coin of the realm" for defining virtual applications and moving them among IaaS providers. An OVF Package consists of exactly one Descriptor file, an XML document commonly called the *OVF envelope* that defines the

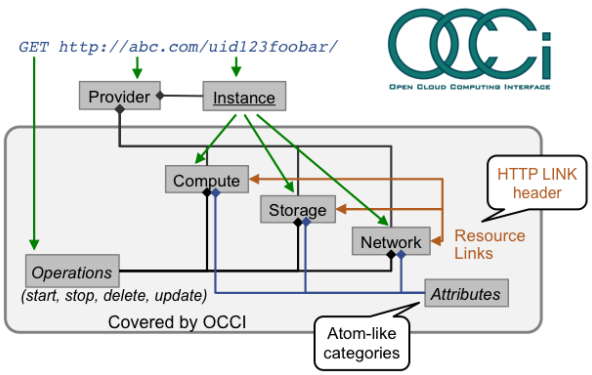

Figure 6: The OGF Open Cloud Computing Interface.

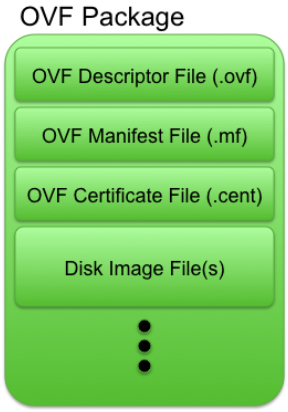

Figure 7: The DMTF Open Virtualization Format.

content and requirements of a virtual appliance. If not omitted, there is one Manifest file that contains SHA-1 digests to provide data integrity for the package. If not omitted, there is one Certificate file that is used to sign the package and provide package authenticity. Finally there are zero or more Disk Image files that represent the virtual disks supporting the virtual appliance.

OCCI can also be used with the Cloud Data Management Interface (CDMI) [9] from SNIA, as shown in Figure 8. CDMI can manage the provisioning of block-oriented, file-oriented, and object-oriented storage. Hard or soft containers are allocated out of physical storage, depending on requirements. A separate Storage Management Client can also be used to for the direct management of the physical storage.

5. DRIVING CLOUD STANDARDS

Regardless of the trajectory that cloud standards may actually take, there exists a fundamental open loop between the consumers and producers of standards. Consumers typically "just want something that works" and consider the development of standards as being outside their charter and budget. On the other hand, vendors and other producers of standards typically focus on establishing market share and only consider standards when demanded by customers.

How can we close this loop and make the standards production process more responsive and effective? The Open

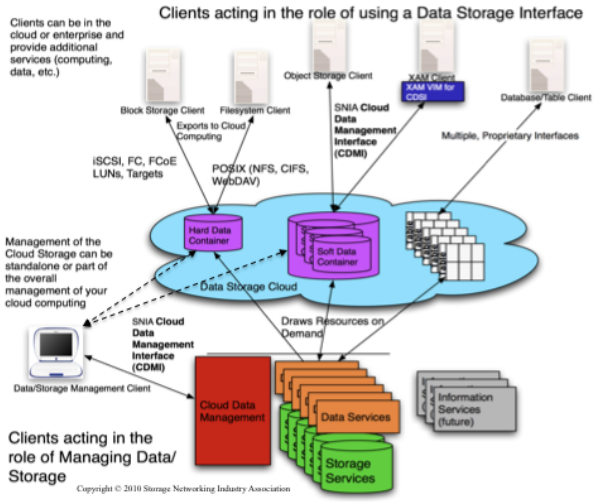

Figure 8: The SNIA Cloud Data Management Interface. (Used by permission.)

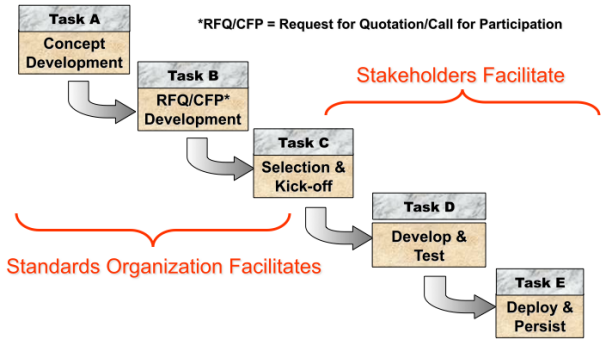

Figure 9: Driving Cloud Standards.

Geospatial Consortium [8] has developed a roughly annual process whereby major stakeholders can drive the development of geospatial standards. This general process could be applied in many other areas where standards are required and is illustrated in Figure 9. In the first half of the process, the Standards Organization facilitates the collection and prioritization of requirements across all major stakeholders into specific, short-term project plans. This information is then used to issue a Request for Quotation or Call for Participation. Major stakeholders and participants respond to the RFQ/CFP and begin contractual execution on specific projects to achieve near-term goals, e.g., demonstrating implementation interoperability, end-to-end integration of capabilities, specific application scenarios of interest, etc. Ultimately, after development and test, the results are ideally deployed in persistent operations.

This process has worked quite well for government agencies that need geospatial standards. As larger organizations that have stable, long-term needs for standards, there is a clear business case for them to engage more directly in the standards process. Since they are collaborating with other stakeholders, the need for common standards with wider applicability and potential adoption are identified earlier. By working on common goals within a collaboration, the stake-

holders and participants realize a significant *return on investment* since the entire cost of a project is not borne by one organization.

Given the current thrust of national cloud initiatives, it is clearly possible that this general process could work for cloud standards as well. As the size and number of organizations involved increases, the more likely that vendors will also engage. In fact, if there is a "critical mass" of participants, the stakeholders many only need to define what type of standard is needed without specifying any of the technical details. The technical specifics could be left to the developers during the development and test phase.

To facilitate the coordination of producers and consumers of cloud standards in this manner, standards organizations working in the area of cloud computing recognized the need to coordinate themselves. This led to the creation of *Cloud-Standards.org*, an informal collaboration of standards organizations, including OGF, DMTF, SNIA, OCC, CSA, TMF, OMG and OASIS. Through regular telecons and common workspaces, these organizations keep each other appraised of work on cloud standards and opportunities to collaboration on development, demonstrations, and outreach. Engagement with large user communities are actively encouraged, including national clouds *and also science clouds*.

6. SUMMARY

We have discussed the benefits and issues of cloud computing, in general, and then considered the specific benefits and issues surrounding the use of dynamically provisioned resources for scientific computing requirements. Like most users, scientific users of computation will appreciate the "ownership" of virtual resources, since this reduces the uncertainly concerning access when needed. This difference between batch job queues and allocation of virtual resources is a fundamental difference between the grid and cloud experiences for the user. Likewise, virtual machines can simplify application deployment by reducing the possible compatibility issues between the application and the hosting environment. This is another fundamental difference and is particularly important for scientific users where numerical accuracy and stability may be an issue.

We also argue that while public clouds will be deployed and used, the fact that enterprises can deploy and use private clouds while managing their security and information assurance requirements using existing tools and approaches, means that in the near-term, private cloud deployment will take on greater importance for enterprise requirements (even if the private clouds are deployed virtually from a cloud provider's data center). Furthermore, user groups that have specific computing requirements may not want to acquire resources from a public cloud provider if those resources cannot meet requirements. As a case in point, scientific computing will demand closer coupling of servers, access to data, and the ability to manage performance through proper abstractions exposed through service level agreements. This argues for the separate deployment of science clouds that have these properties. Hence, even science clouds may follow the deployment trajectory from private, to federated, hybrid, and finally public clouds. Given this deployment trajectory, we derived a deployment, development and research roadmap to achieve the necessary capabilities.

At this point, we note that as soon as clouds federate, many of the necessary capabilities that have been fundamental to the grid concept, such as identity management and virtual organizations, will become directly relevant to cloud environments. This includes concepts and tooling such as identity management, virtual organizations, the GLUE schema for describing computing resources, data access and transfer standards such as OGSA-DAI, GridFTP, and SRM, etc. As concepts concerning distributed infrastructure management and dynamic provisioning, grid and cloud are not in competition, but are rather quite complimentary.

Another argument we make is that standardization will be most effective and useful at the infrastructure level, since the commodification of basic resources, such as servers, storage, and networks, will have the widest impact across application domains. With this in mind, we described the developing OCCI, OVF, and CDMI standards that could be used together to deploy and manage infrastructure clouds. While standardization at the PaaS and SaaS levels could also occur, it will mostly likely happen after standardization at the infrastructure level.

Finally we considered how to drive cloud standards. Rather than just letting the marketplace do a "random walk", or be driven by corporate interest in maximizing market share, we describe a collaborative project process whereby major stakeholders and participants can define short-term goals to make clear progress concerning implementations, demonstrating interoperability, and the integration of end-to-end capabilities By engaging in a collaborative process, stakeholders can realize a substantial return on interest. This process has worked well for government agencies that are direct consumers of geospatial standards. Hence, this same process could work well for the development of standards for national clouds, as well as science clouds for national objectives.

While some may argue that it is too early for cloud standards, we argue that from a technical perspective, it is often quite easy to see where a common practice would be beneficial by reducing incompatibility and increasing reuse. From a market perspective, however, the wide adoption of a common practice is problematic since it can involve market "timing", development schedules, and other competing non-technical issues. Hence, in many cases, the best that can be done is to put a "stake in the sand". This allows a technical solution to be defined and tested in the "marketplace of ideas" wherein it might gain adoption if the time is right.

The final message of this paper is a *call to action* for all stakeholders to engage in the best practices and standards processes described above. This paper was motivated by the need to organize not only the technical development issues for scientific and national clouds, but also to organize the development roadmap goals of science clouds, national clouds and the wider distributed computing community. This can only be done if stakeholders engage and help drive the process to a successful conclusion in community-based organizations such as the Open Grid Forum.

7. ACKNOWLEDGMENTS

The author wishes to thank Geoffrey Fox, Steven Newhouse, and Martin Walker for valuable comments on earlier drafts of this paper.

8. REFERENCES

[1] FY 2010 U.S. Federal Budget. http://www.gpoaccess.gov/usbudget.

[2] Information Assurance. http://en.wikipedia.org/wiki/Information_assurance.

[3] The Cloud Security Alliance. http://www.cloudsecurityalliance.org.

[4] The DMTF Open Virtualization Format. www.dmtf.org/standards/published_documents/DSP0243_1.0.0.pdf.

[5] The Envision Project. http://www.envision-project.eu.

[6] The Kasumigaseki Cloud Concept. http://www.cloudbook.net/japancloud-gov.

[7] The OGF Open Cloud Computing Interface. http://www.occi-wg.org/doku.php.

[8] The Open Geospatial Consortium. http://www.opengeospatial.org.

[9] The SNIA Cloud Data Management Interface. http://www.snia.org/cloud.

[10] The UK G-Cloud. http://johnsuffolk.typepad.com/john-suffolk---government-cio/2009/06/government-cloud.html.

[11] The US Cloud Storefront. http://www.gsa.gov/Portal/gsa/ep/contentView.do?contentType=GSA_BASIC&contentId=28477.

[12] U.S. Federal IT Dashboard. http://it.usaspending.gov.

[13] M. Armbrust et al. Above the Clouds: A Berkeley View of Cloud Computing. Technical Report EECS-2009-28, UC Berkeley, 2009. www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf.

[14] P. Bogden. Personal Communication., June 2009. Former project director, SURA Coastal Ocean Observing and Prediction program (SCOOP).

[15] P. Bogden et al. Architecture of a Community Infrastructure for Predicting and Analyzing Coastal Inundation. *Marine Technical Society Journal*, 41(1):53–61, June 2007.

[16] Chris Kemp. Standards, Nebula, and Interoperability. www.omg.org/news/meetings/tc/ca/special-events/ci/NASA.pdf.

[17] I. Foster and others. Cloud Computing and Grid Computing 360-Degree Compared. In *IEEE Grid Computing Environments (GCE08)*, pages 1–10, 2008.

[18] R. Buyya and others. Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. *Future Generation Computer Systems*, 25(6):599–616, June 2009.

[19] J. Weinman. Mathematical Proof of the Inevitability of Cloud Computing. cloudonomics.wordpress.com/2009/11/30/mathematical-proof-of-the-inevitability-of-cloud-computing, Nov. 30 2009.