Cloud Computing Task Force Report to the Information Technology Leadership Council

January 8, 2010

Table of Contents

Introduction

The time to start planning for the deployment of cloud-based services within the University of California's data center service portfolio is now.

- UC's current budget situation requires us to explore potentially cost-effective means for providing computing services.
- Cloud computing has the potential to advance the University's mission in a number of areas that have not previously been possible, because of cloud computing's highly dynamic, "pay as you go" service model. These include high-performance computing, research, instruction, administrative computing, large-scale storage, and disaster recovery.
- Cloud computing will leverage and enhance UC's current activities to establish regional data centers. In particular, it will enable greater agility in the use of data center space throughout UC.

The National Institute of Standards and Technology (NIST) defines cloud computing as a "...model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction." This model has the potential to create services that are very attractive for satisfying resource needs cost effectively and n short order.

Also, the days of users wanting to run their own hardware are ending. Coupled with the ease with which services can be purchased, cloud computing services are now very attractive, and UC people are using them. Unfortunately, none of this reduces the University's legal and policy requirements, and many of UC's cloud computing users are deploying systems without much guidance.

This report explores the issues surrounding cloud computing and makes a modest set of recommendations to enable UC to capitalize on this new computing paradigm. While these recommendations are specifically for UC, it should be noted that there are opportunities for broader collaboration that should be explored. These include:

- Current discussions of cloud computing within CENIC
- Potential cloud computing national and international partnerships via organizations like the RUCC, NLR, and Internet2
- Synergies with UC Grid, the Shared Research Computing System (SRCS), and the Triton cluster
- Existing campus virtualization activities

The UC Cloud Computing Task Force

The University of California's Cloud Computing Task Force (CCTF) was created in May 2009 by the Information Technology Leadership Council (ITLC) to assess cloud computing and its applicability within the University of California. The issues the CCTF was charged to address include:

- A definition and description of cloud computing
- The potential of cloud computing to address a variety of use cases, including high-performance computing, administrative computing, large-scale storage, and disaster recovery
- Considerations for the deployment of cloud computing, including security, capacity planning, and legal and contractual issues
- Recommendations for deployment and further study of cloud computing technologies, including an ongoing organizational structure

The full text of the CCTF's charge and the CCTF membership are available as appendices to this report.

The CCTF conducted its work between May and August of 2009 through a combination of face-to-face meetings, conference calls, and a wiki space. As part of this work, a number of white papers were written by members of the group addressing many issues related to cloud computing and are the basis for this report. They are included in the appendices.

Definition of Cloud Computing

The industry is currently grappling with an appropriate definition of cloud computing, as evidenced in the Definition of Cloud Computing white paper. For the purpose of this study, we have adopted the Draft NIST Working Definition of Cloud Computing, which defines cloud computing as a "...model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". According to this definition, the essential characteristics of cloud computing are:

- **On-demand self service**: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with the service providers.
- **Ubiquitous network access**: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
- **Location-independent resource pooling**: The customer generally has no control or knowledge over the exact location of the provided resources. The service provider assigns different physical and virtual resources dynamically according to consumer demand.
- **Rapid elasticity**: Capabilities can be rapidly and elastically provisioned to quickly scale up and released to quickly scale down.
- **Measured servicing**: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service.

In addition, the NIST definition defines three delivery models:

- **Cloud Software as a Service (SaaS)**: Delivery of an application that leverages the Cloud resources at the back-end, e.g. Google Mail, Facebook, etc.
- **Cloud Platform as a Service (PaaS)**: Delivery of a "platform" and/or solution stack as service using programming languages and tools supported by the service provider, e.g. the Google App Engine.
- **Cloud Infrastructure as a Service (IaaS)**: Delivery of processing, storage, networks, and other fundamental computing resource as a service, e.g. the Amazon Elastic Compute Cloud (EC2), the Nirvanix Storage Delivery Network (SDN), etc.

Finally, it also defines four deployment models:

- **Private Cloud**: The cloud infrastructure is operated solely for one organization. This does not imply that it is managed or located within the same organization - in fact, it can be managed by a 3rd party and located elsewhere.
- **Community Cloud**: The cloud infrastructure is shared by several organizations, which share common concerns.
- **Public Cloud**: The cloud infrastructure is made available to the general public, and is owned by an organization selling cloud services.

- **Hybrid Cloud**: The cloud infrastructure is a composition of two or more deployment models above.

Potential Use Cases

Based on the definition, below are examples and comments for Administrative, Instructional, and Research cases whereby Cloud Computing could be used. Many of these statements could be applied to all areas.

Research Computing

- Relieve administrative and operational challenges associated with high-performance computing (finding expert systems administrators, as well as adequate power and cooling).
- Remove long delays before codes can be run because Cloud Computing allows researchers to deploy computing resources quickly and pay for actual resource consumption.
- Assist in meeting deadline-driven computational work, or when rapid turn-around is required (for instance, when developing codes).
- Allow UC researchers to use cloud-computing services to deploy and 'undeploy' virtual clusters rapidly.
- Some domain experts are more interested in using higher-level computing services, such as Amazon's 'Elastic MapReduce' service and Microsoft's Azure.
- Able to build a 'private cloud' (local computational resource) so researchers can move computation work between their private cluster and the public cloud, without making any changes to their software.
- Existing cloud services are not currently well-suited for high-performance parallel processing, but many research problems don't require that kind of computation.

Instructional Computing

- Give students short-term access to computing resource at a scale that was previously off-limits to all but the largest organizations.
- Teaching of existing concepts such as experiencing firsthand what happens when a database tries to accommodate too many users.
- Potentially better technical support could be provided due to active developer community with its question boards, blogs, and documentation, which are typically far more comprehensive than what limited course staff could provide.
- Simplified courseware management - a course has natural peaks (midterm exam, final project, writing assignments, etc.) separated by relative lulls.
- Virtualization - multiple courses that share IT infrastructure often share a human administrator, who is overwhelmed by conflicting requests for installing and supporting different courseware for the various instructors. Many University computers are "locked down" so that only designated system administrators can install new software and perform maintenance. Could create a virtual machine image containing the required courseware and the VM image can be re-used for future offerings of the course and modified as needed.
- Student work product continuity could occur unlike typical current practices, where course-specific instructional accounts are deleted at the end of the course (forcing students to make local copies of work products if they wish to keep them). Potentially be used as the basis of future coursework or other projects.

Administrative Computing

- Build test and QA environments to mimic production capacity for performance testing of applications.
- Configure additional web farm servers to be used during peak student registration periods and then decommission them once the workload goes back down.
- Systems that are better hosted remotely, such as disaster recovery, for a campus portal or email services.

Potential Technical Issues and Challenges

Various issues and challenges will cross multiple areas shown below. Each of these areas has impact on one or more of the others no matter how large or small the implementation of Cloud services. Note that the recommendation section of this document has some suggestions as to how to resolve some of the items below. Over time, it is believed that more of these issues will become minimal or disappear but that the providers today still have many issues to iron out.

Legal/Contractual

- Much software hasn't yet moved to a "cloud friendly" licensing model - though recently IBM and Microsoft have announced specific steps in this direction
- Data Security, eDiscovery, data recovery, etc. issues must be resolved with Cloud providers (applies to outsourcing partners as well) via their written agreements.

Preparation for Cloud Computing at UC

- Existing financial and accounting models do not always support users who need to purchase cloud services
- Mechanisms for information exchange and dialog on the topic of cloud computing do not exist within UC

Service Continuity

- The cloud computing marketplace is evolving rapidly, so service providers' long-term viability is not generally assured
- Open, community-defined standards, and disclosure of facts about energy-efficiency of vendor services are still evolving

Capacity Planning

- Technical considerations such as horizontal scalability, the ability to exploit cost-associatively, data locality and data management, and the availability of specific hardware (such as low-latency interconnects or high-speed storage) determine whether a given research problem is a good fit for cloud computing.
- Network performance and throughput requirements must be considered in the design of any application using Cloud Computing.
- Current cloud computing tools, while improving rapidly, are relatively immature.

Security

- Authentication - public clouds may have their own, whereas campuses typically have engineered around their own auth system

- UC policy, particularly "IS-3: Electronic Information Security," applies to all applications, so the risks associated with outsourcing and cloud computing must be assessed on a case-by-case basis.
- Encryption - Key management is an important issue to ensure that UC always has access to its keys. This could be a road block to any data that was not public unless it was an internal UC cloud.

Environmental Sustainability

- Today, most power and facilities consumption costs are absorbed into general overheads. If users buy public cloud services with overhead-taxed dollars, they effectively pay twice for power and facilities. In general, financial policies should reward users for improving their power and physical-plant utilization and costs, rather than insulating them from these costs in a way that gives them no incentive to improve. The high power-densities of cloud computing installations may, in the long run, make those service a 'greener' choice, compared to maintaining local infrastructure. As with monetary cost, the 'green' impact calculation is not simple. To the extent that such a calculation is possible, though, UC policy should reward users for identified efficiency improvements.

Policy and Procedure

- Policies and procedures need to allow new pricing models to be created in order to sustain the overall University mission (e.g., how instructional computing is financed). New recharge models may be needed. "True pay-as-you-go" will clearly identify which courses use more IT infrastructure than others; a separate policy question is whether they should necessarily be paying more, or if this is a cost that should be buffered at department or University.

Recommendations and Conclusions

Develop a Policy and Legal Framework for Outsourcing IT Services

The lack of a well-understood and adopted policy and legal framework for outsourcing IT services for UC is the single largest barrier to the adoption of cloud-based services. It is also likely to lead to inappropriate use of cloud-based services in the future, putting the University at risk of privacy breaches and service losses.

While this is not specifically an issue of cloud computing, we believe it is essential for the University to provide guidance for all outsourced IT services before cloud-based services can be used widely within UC.

We propose that the ITLC designate a small group of people to work with the UC Information Technology Policy and Security (UCITPS) group, the Joint Data Center Managers' Group (JDCMG), IT Strategic Sourcing, and the Office of General Counsel to address the following:

- Risk assessment. Guidance for IT managers in assessing the alternative risks of outsourcing and insourcing, in particular privacy, security, regulatory compliance, and disaster recovery.
- Applicable policy and law. A guide to policy and law that are likely to impact outsourcing decisions.
- Potential risk mitigation. Approaches for addressing the risks associated with outsourcing, including:
    - Templates and sample language for inclusion in RFPs and vendor contracts
    - Business process measures to mitigate risk

    - Technologies that can help mitigate risk
- Environmental sustainability. Sample contract terms that encourage green computing practices and require verifiable disclosure of practices and environmental impacts by UC's IT service providers.
- Cloud-friendly software license terms. Sample contract terms that enable agility in how software is hosted, either on traditional server configurations or on public or private clouds, without undue financial or other contractual penalty.

Remove Disincentives to "Do the Right Thing"

The cost of computing is shifting significantly toward the cost of space, power, and air conditioning, as opposed to the cost of acquiring and supporting the technology components themselves. This is infrastructure that the University does not generally track at a level where it can effectively educate design and deployment decisions for IT-based systems. For example, a locally-hosted system may appear to be less expensive on a department's budget, but the cost to the University might be higher than outsource, due to power costs.

Addressing this issue will not be easy:

- General cost comparisons of traditional and cloud computing paradigms should be approached with cautious skepticism.
- What is "right" in one situation may be "wrong" in another. Each use case must be taken on its own merits.

Also, each campus CIO will need to tailor solutions to the specifics of each campus's budget and management structures. However, a couple of methods for determining the power consumption of locally-managed hardware are recommended:

- Attach metering devices to measure the power consumption of server and storage installations. This is likely to be difficult and expensive, but does provide the most accurate information.
- Disseminate estimates of power consumption for canonical server configurations to be used to estimate consumption of actual data center installations. This is, of course, less accurate than the former method, but it is much more easily done and should provide sufficient accuracy to guide decisions about the use of local and cloud resources.

It should be noted that there may also be external policies that affect infrastructure decisions, particularly from grant agencies. The University should look for opportunities to encourage appropriate change in these policies, both directly with the agencies (probably via the campuses' offices of research) and through national IT organizations like EDUCAUSE and Internet2 that provide broad representation for higher education.

Identify Expertise to Advise Potential Cloud Users and Developers

Each of the campuses has staff in various parts of the organization that provide advice on the configuration and deployment of IT-based systems. Because cloud computing is a new phenomenon, however, these individuals often do not have expertise in the benefits and pitfalls of cloud computing as an alternative to locally-operated resources. This situation will result both in missed opportunities where cloud computing might have been a good fit for a particular application, as well as miss-application of cloud computing for other applications.

The CCTF recommends that each CIO identify cloud computing experts who can provide a second level of support for these staff via an electronic mail list and a web/wiki presence. This group would also address the following issues:

- Procurement. Work with Strategic Sourcing to:
  - Identify viable vendors and negotiate favorable agreements with them, leveraging UC's unique advantages, such as CENIC connectivity.
  - Document and facilitate the process of contracting for cloud resources.
- Application Design and Deployment. As described above, provide assistance to application designers and deployers in the following areas:
  - The appropriateness of cloud computing and its alternatives for an application, including consideration of network and platform issues, as well as policy, contractual, and legal issues
  - Design (or re-design) criteria for placing an application in a cloud
  - Appropriate test planning

Strategic Planning for Computing Service Provision

Cloud computing is part of the larger picture of the variety of ways IT-based services can be provided to the UC community. While there have been efforts of the past few years to address issues like computer room space and disaster recovery (and cloud computing), UC's IT community has not taken on that broader issue. We propose the creation of a task force, reporting to the ITLC that would work with the Joint Data Center Management Group (JDCMG), the Communications Planning Group (CPG), and the IT Architecture Group (ITAG) to address strategies for the design, acquisition, and support of computing equipment and services. Issues that should be considered include:

- Strategies to achieve an appropriate balance of public cloud services and private, UC-owned, resources to minimize cost
- Mechanisms to ensure that critical applications and sensitive are hosted within appropriate data center facilities and by appropriate service providers
- Financial support models that leverage resources throughout UC to sustain services while providing incentives for "doing the right thing."
- Guidelines to encourage environmental stability in the design, implementation, and operation of UC-owned IT resources.
- Disaster recovery strategies, particularly for departmental systems, that leverage cloud-based resources.
- Recommendations for pilot projects, using the information in this report, as well as new opportunities, such as
  - The acquisition of large computing resources
  - The creation of research projects with large computing needs
  - State-wide and national partnerships with peer institutions via CENIC, Internet2, NLR, and the RUCC.

Appendices

Appendix 1: Draft NIST Working Definition of Cloud Computing

Authors: Peter Mell and Tim Grance
6-1-09

National Institute of Standards and Technology, Information Technology Laboratory

Note 1: Cloud computing is still an evolving paradigm. Its definitions, use cases, underlying technologies, issues, risks, and benefits will be refined in a spirited debate by the public and private sectors. These definitions, attributes, and characteristics will evolve and change over time.

Note 2: The cloud computing industry represents a large ecosystem of many models, vendors, and market niches. This definition attempts to encompass all of the various cloud approaches.

Definition of Cloud Computing

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential **characteristics,** three **delivery models**, and four **deployment models**.

Essential Characteristics

On-demand self-service. A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.

Ubiquitous network access. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).

Location independent resource pooling. The provider's computing resources are pooled to serve all consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. The customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.

Rapid elasticity. Capabilities can be rapidly and elastically provisioned to quickly scale up and rapidly released to quickly scale down. To the consumer, the capabilities available for provisioning often appear to be infinite and can be purchased in any quantity at any time.

Measured Service. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported providing transparency for both the provider and consumer of the utilized service.

Note: Cloud software takes full advantage of the cloud paradigm by being service oriented with a focus on statelessness, low coupling, modularity, and semantic interoperability.

Delivery Models

Cloud Software as a Service (SaaS). The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure and accessible from various client devices through a thin client interface such as a Web browser (e.g., web-based email). The consumer does not manage or control the underlying cloud infrastructure, network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

Cloud Platform as a Service (PaaS). The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created applications using programming languages and tools supported by the provider (e.g., java, python, .Net). The consumer does not manage or control the underlying cloud infrastructure, network, servers, operating systems, or storage, but the consumer has control over the deployed applications and possibly application hosting environment configurations.

Cloud Infrastructure as a Service (IaaS). The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly select networking components (e.g., firewalls, load balancers).

Deployment Models

Private cloud. The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise.

Community cloud. The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on premise or off premise.

Public cloud. The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.

Hybrid cloud. The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting).

[The original of this document is available from the NIST Cloud Computing site at http://csrc.nist.gov/groups/SNS/cloud-computing/.]

Appendix 2: What is Cloud Computing?

Sriram Krishnan, UCSD

1. Background

Over the past decade, the field of Grid computing has seen a lot of hype of activity. The term "Grid computing" can be attributed to Ian Foster, who created a three-point checklist to define a "Grid" as follows [FOS02]. A Grid:

1. Coordinates resources that are not subject to centralized control. A grid integrates and coordinates resources and users that live within different control domains -- for example, different administrative units of the same company, or even different companies. A grid addresses the issues of security, policy, payment membership, and so forth that arise in these settings.
2. Uses standard, open, general-purpose protocols and interfaces. A grid is built from multi-purpose protocols and interfaces that address such fundamental issues as authentication, authorization, resource discovery, and resource access. It is important that these protocols and interfaces be standard and open. Otherwise, we are dealing with application, hardware, or OS -specific systems.
3. Delivers nontrivial qualities of service. A grid should be transparent to the end user, addressing issues of response time, throughput, availability, security, and/or co-allocation of multiple resource types to meet complex user demands. The goal is that the utility of the combined system is significantly greater than that of the sum of its parts.

Many real-world grids exhibit one or more of the above properties – in practice, it can be often observed that none of the so-called grid systems satisfy all of the above requirements to qualify as a true Grid system. For instance, the TeraGrid (http://www.teragrid.org) integrates high-performance computers, data resources and tools, and high-end experimental facilities at 11 partner sites around the country. The TeraGrid satisfies requirements 1 and 2 above, but it is debatable how much it satisfies requirement 3, if at all. The TeraGrid coordinates resources across the individual partner sites, which define the local policies and administrative setup. And with the help of the Open Grid Services Architecture (OGSA) [OGSA], Web service concepts and technologies are being used to satisfy the second requirement. However, transparency, co-allocation of multiple resources across various administrative domains and meta-scheduling remains a pipe dream – for all practical purposes, users typically choose and are fully aware of the resources being used for their applications.

In the industry, the term Grid computing is often used more loosely. In fact, most so-called industry grids in the past and present (e.g. Oracle Grid, Sun Grid, etc.) enable access to resources that are subject to centralized administrative control, and do not use any standard, open, general-purpose protocols. Most industry Grids relied heavily on virtualization to create a pool of assets to distribute workloads [IBM06]. In many ways, this looser definition of Grid computing in the industry and the resulting technologies to support the same have led to the evolution of Cloud Computing.

2. Cloud Computing

2.1 Definitions & Classification

Several definitions for Cloud Computing can be found on the Internet. McKinsey and Company [McK09] define Clouds as hardware-based services that offer computer, network and storage capacity, where hardware management is highly abstracted from the buyer, buyers incur infrastructure cost as variable Operational Expenditure (OPEX), and where infrastructure cost is highly elastic (up or down). They define the following characteristics of clouds:

1. Enterprises incur no infrastructure capital costs, just operational costs on a pay-per-use basis

2. Architecture specifics are abstracted
3. Capacity can be scaled up or down dynamically, and immediately
4. The underlying hardware can be anywhere geographically

In [AMBR09], the authors define Cloud Computing as both applications delivered as services over the Internet, and the hardware and software in the datacenters that provide those services. They view Cloud Computing as a sum of Software as a Service (SaaS) and Utility Computing, which is defined as when such a service is sold (in possibly, a pay-as-you-go manner). They emphasize three aspects of Cloud Computing from a hardware point of view:

1. The illusion of infinite computing resources available on demand
2. The elimination of up-front commitment by Cloud users
3. The ability to pay for use of computing resources on a short-term basis as needed.

The difference between the two definitions above is that the abstraction of infrastructure is explicitly emphasized in [McK09], whereas it is implied in [AMBR09]. Additionally, [McK09] differentiates "Cloud services" from Clouds as a service where the underlying infrastructure is abstracted and can scale elastically – in other words, it views Clouds as mostly abstractions for hardware.

The National Institute of Standards and Technology (NIST) [NIST09] defines Cloud Computing as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. It stresses five essential characteristics:

1. On-demand self service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with the service providers.
2. Ubiquitous network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
3. Location-independent resource pooling: The customer generally has no control or knowledge over the exact location of the provided resources. The service provider assigns different physical and virtual resources dynamically according to consumer demand.
4. Rapid elasticity: Capabilities can be rapidly and elastically provisioned to quickly scale up and rapidly released to quickly scale down.
5. Measured servicing: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service.

This definition is very similar to the two definitions above, except that it emphasizes one more key aspect of Cloud Computing, which is an ability for a consumer to unilaterally provision computing capabilities, as needed automatically without requiring human interaction with each service's provider.

Finally, Gartner defines Cloud Computing as a style of computing where massively scalable IT-related capabilities are provided "as a service" using Internet technologies to multiple external customers [GART08].

In general, several delivery models exist for Cloud Computing, as defined in [NIST09]:

1. Software as a Service (SaaS): Delivery of an application that leverages the Cloud resources at the back-end, e.g. Google Mail, Facebook, etc.

2. Platform as a Service (PaaS): Delivery of a "platform" and/or solution stack as service using programming languages and tools supported by the service provider, e.g. the Google AppEngine.
3. Infrastructure as a Service (IaaS): Delivery of processing, storage, networks, and other fundamental computing resource as a service, e.g. the Amazon Elastic Compute Cloud (EC2), the Nirvanix Storage Delivery Network (SDN), etc.

The Clouds also have several deployment models [NIST09]:

1. Private: The cloud infrastructure is operated solely for one organization. This does not imply that it is managed or located within the same organization – in fact, it can be managed by a 3rd party and located elsewhere.
2. Community: The cloud infrastructure is shared by several organizations, which share common concerns.
3. Public: The cloud infrastructure is made available to the general public, and is owned by an organization selling cloud services.
4. Hybrid: The cloud infrastructure is a composition of two or more deployment models above.

2.2 Discussion

Irrespective of however Cloud computing is defined; the consensus is that Clouds enable a utility or pay-as-you-go model without an upfront commitment to infrastructure costs or human intervention on the part of the service provider. The use of virtualization is also accepted as a de facto norm for providing Cloud-based infrastructure and services. Finally the illusion of elasticity where resources are available on demand, and can be scaled up or down, eliminates the need for Cloud Computing users to plan ahead for peak loads.

Many have said that Cloud computing is just Grid computing by another name. In a lot of ways, it delivers on the promise of Grid computing addressing the requirement for non-trivial qualities of service. However, it is important to note that the problem domains addressed by Grid and Cloud computing are significantly different, at least at the time of writing this document. Grid computing is mostly designed for a smaller number of users in the high performance computing community, who need exclusive access to a large number of resources at once. On the other hand, Cloud computing supports a large number of users concurrently, each of whom has access to a small portion of the resources. The above requirement is often manifested in the way people access these resources. For instance, Grid users typically use a batch queuing system to submit jobs, and may wait for their jobs for an unspecified amount of time. Cloud users require and gain access to resources on-demand, leveraging the illusion of infinite elastic resources.

There is also a concern that Cloud resources may not be appropriate for high performance computing applications due the heavily virtualized nature of the resources. In [WALK08], the authors concluded that a performance gap exists between performing HPC computations on a traditional scientific cluster and on an EC2 provisioned cluster. The performance gap is seen not only in the MPI performance of distributed memory parallel programs, but also in the single node OpenMP performance for shared-memory parallel programs.

Some of the other obstacles in the growth of Cloud Computing listed in [AMBR09] are:

1. Availability and service up-time
2. Data lock-in, because of which consumers can't easily transfer their data from one site to another
3. Data confidentially and auditability on the Cloud
4. Data transfer bottlenecks on the edges of the Cloud
5. Unpredictability of performance
6. Scalable storage – no upfront cost, and infinite capacity and elasticity on-demand
7. Debugging large-scale distributed systems

8. Rapid scaling, up and down
9. Reputation fate sharing between a rogue user and a Cloud provider, and legal liability
10. Software licensing

3. Conclusions

Cloud Computing can be thought of as an evolution of Grid computing, delivering on its promise of non-trivial qualities of service. Although it may not yet be suitable for all classes of applications, it provides an illusion of infinite computing resources that can scale up and down, where users can pay for use of resources as needed (pay-as-you-go), thus eliminating the up-front infrastructure capital costs. For the purposes of this document, we will adopt the NIST view of Cloud Computing as the definitive definition.

Various scenarios for the use of Cloud Computing at the University of California may be possible. For instance, UC may itself build its own "Community Cloud", where it may limit access to its resources to UC staff and students. Or UC may use an enterprise Cloud such as Amazon EC2 as an overflow service.

4. References

[AMBR09] M. Armbrust et al. "Above the Clouds: A Berkeley View of Cloud Computing". Technical Report No. UCB/EECS-2009-28. http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html.

[FOS02] I. Foster. "What is the Grid: A Three Point Checklist". http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf.

[GART08] Gartner Newsroom. "Gartner Says Cloud Computing Will Be As Influential As E-business". http://www.gartner.com/it/page.jsp?id=707508.

[IBM06] IBM Corporation. "Grid Computing: Past, Present and Future. An Innovation Perspective". http://www-03.ibm.com/grid/pdf/innovperspective.pdf.

[McK09] McKinsey & Company. "Clearing the air on Cloud Computing". http://uptimeinstitute.org/images/stories/McKinsey_Report_Cloud_Computing/mckinsey_clearing_the%20clouds_final_04142009.ppt.pdf.

[NIST09] NIST Working Definition of Cloud Computing v14, http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v14.doc

[OGSA] The Open Grid Services Architecture. http://www.globus.org/ogsa/.

[WALK08] E. Walker. "Benchmarking Amazon EC2 for high-performance scientific computing". In ;login: online. http://www.usenix.org/publications/login/2008-10/openpdfs/walker.pdf

[WIKI] Wikipedia. "Cloud Computing". http://en.wikipedia.org/wiki/Cloud_computing

Appendix 3: Cloud Computing in Research

Armando Fox, UCB; Greg Bell, LBNL; Bill Labate, UCLA

Key Points for Research:

- Benefit: Unlike other use cases, lower cost may not be the driving factor for adopting cloud computing in research. Instead, researchers may be attracted to scalability, flexibility, and near-instantaneous service delivery, which enable research progress at a pace and scale previously beyond the reach of most academics. In some cases these benefits are so compelling that individual researchers and labs will adopt cloud computing whether or not UC-wide action is taken to support this.
- Caveat: A significant obstacle is the relative immaturity of tools enabling "off the shelf" use of cloud computing for science, requiring early adopters to apply significant IT expertise. In addition, funding and chargeback models have not fully caught up with the cloud computing "pay-as-you-go service" model.
- What UCOP can do: It is premature for UC to launch a centralized effort to coordinate access to public cloud services, or to build and maintain a private cloud. However, UCOP can act as a catalyst by:
    - supporting researchers who identify models for productively harnessing public or private clouds,
    - developing financial and accounting models that do not penalize researchers who purchase cloud services when those services are cost-effective,
    - encouraging and rewarding researchers who "productize" their cloud computing tools to facilitate uptake by others.

Who can benefit and what are the opportunities?

Cloud computing has the potential to benefit many researchers in the sciences, social sciences, and humanities. In the sciences, for example, large-scale computation has become vitally important in a range of disciplines, from astrophysics to machine learning. Traditionally, scientists in need of computational resources have bought dedicated clusters, or have competed for CPU cycles on large, shared machines. Both approaches have disadvantages. Scientists who purchase a cluster usually encounter new administrative and operational challenges (finding expert systems administrators, as well as adequate power and cooling). Scientists who compete for time on leadership-class facilities may experience long delays before their codes can run. Because cloud computing allows researchers to pay for actual resource consumption, even if the consumption occurs at a nonuniform rate, new opportunities for research arise. Cloud services may be especially attractive when scientists have deadline-driven computational work, or when they require rapid turn-around (for instance, when developing codes).

UC researchers already use cloud-computing services, and we expect more of them to do so in the future as cloud offerings evolve and mature, and as barriers to acceptance are addressed. A major opportunity for scientists performing research on large-scale computing itself is the ability to rapidly deploy and 'undeploy' virtual clusters. For example, the Reliable Adaptive Distributed Systems Lab (RAD Lab) at UC Berkeley is conducting research on improving the throughput of large-scale batch computing jobs, and is using Amazon EC2 to pilot approaches on different configurations of virtual clusters, without purchasing a physical cluster. In addition, researchers at LBNL have provisioned virtual clusters in Amazon EC2 in order to measure and understand the performance of scientific codes in a cloud-services environment. (Their results suggest that the performance of virtual clusters varies greatly, according to the scientific code under test. Codes which are sensitive to network latency do not yet run efficiently in the EC2 environment.)

Some domain experts may be less interested in provisioning virtual clusters, and more interested in using higher-level computing services which are now available in public clouds. For example, Amazon offers

an 'Elastic MapReduce' service that lets researchers submit parallel batch jobs structured around a MapReduce algorithm and a large data set. Elastic MapReduce automatically schedules and manages batch jobs, without exposing researchers to the underlying virtual machine infrastructure. Other cloud vendors have also introduced higher-level services that may be useful for researchers. For example, researchers at the Berkeley Water Center are using Microsoft's Azure service to process satellite images and to make evapotranspiration calculations.

Finally, some researchers may be interested in building a 'private cloud' - that is, a local computational resource offering features available in a large, public cloud. UC Berkeley's RAD Lab has configured a 50-node private cluster using the open-source Eucalyptus software developed by Rich Wolski et al. at UC Santa Barbara (and now being commercialized by Eucalyptus Systems Inc.). Eucalyptus allows a private cluster to be managed using tools compatible with Amazon EC2. As a result, RAD Lab researchers can move computation between their private cluster and the public cloud without changes to their software. The RAD Lab is actively exploring the challenges and benefits of hybrid 'surge computing,' in which work is usually done on a private cluster, but can overflow to the public cloud when additional capacity is needed.

What are the potential benefits?

Many applications for cloud computing are evaluated largely in terms of cost savings. In the case of research, however, other considerations - including scalability, flexibility, and near-instantaneous service delivery - may be just as important. Cloud computing may allow researchers to solve problems which exceed the capacity of local computing resources. Local capacity constraints include computing and storage resources, but also power, space, and cooling. For example, the RAD Lab has published research results [link?] demonstrating improved batch-processing performance on problem sizes requiring over 1,000 machines, even though no such cluster is currently available at UC Berkeley for dedicated experimentation. This research could not have been conducted without cloud computing services.

Though raw capacity is important, there is often value in obtaining research results sooner. If computation is "cloud-friendly", a short-term burst of resource allocation can mean getting results in hours instead of weeks. A UC Berkeley database research project involved running experiments to simulate not one, but several existing large-scale systems simultaneously. Using EC2, the researcher completed all the experiments for about $3,500 - about the same cost as one fully-loaded server - in the few days before an important publication deadline. Doing the same work using UCB-only resources would have taken weeks.

Depending on local costs and the nature of the computation being performed, there may or may not be cost advantages associated with "cloudsourcing" (for further discussion, see the *Cloud Computing Economics* section of this report). However, notwithstanding the outcome of economic analyses, we emphasize that in some research scenarios, cost considerations may be secondary compared to the opportunity for doing more productive work.

What are the technical issues and challenges?

Technical considerations such as horizontal scalability, the ability to exploit cost-associativity, data locality and data management, and the availability of specific hardware (such as low-latency interconnects or high-speed storage) determine whether a given research problem is a good fit for cloud computing. Today, these determinations must generally be made by domain experts in consultation with IT experts, as the "off the shelf" cloud computing tools are immature. Research environments that have traditionally 'rolled their own' IT support infrastructure will have to determine how to move that infrastructure to the cloud, or alternatively how to phase it out in favor of cloud infrastructure standards which we believe will soon emerge. In addition, there may be considerations around integrating local resource management with cloud resource management: authentication, billing/metering of usage,

monitoring, scheduling, security. The process of integration is likely to be disruptive for the research IT establishment in similar ways that the adoption of open Internet technologies was disruptive in the enterprise. Because of the immaturity of cloud-services tools and models, and because the service space continues to change rapidly, this is unlikely to be the moment for UC to make a centralized cloud 'play' - either to coordinate access to public cloud services, or to build and maintain a private cloud. Nevertheless UC can play a productive role as cloud services continue to mature.

Where can UCOP action make a difference?

To the extent that there are potential advantages (in terms of cost and quality of science) for moving some research to the cloud, UC policy can directly influence whether those opportunities are seized. For example, most researchers' power and facilities consumption costs are absorbed into general overheads. If researchers buy public cloud services with overhead-taxed dollars, they effectively pay twice for power and facilities. In general, financial policies should reward researchers for improving their power and physical-plant utilization and costs, rather than insulating them from these costs in a way that gives them no incentive to improve. The high power-densities of cloud computing installations may, in the long run, make those service a 'greener' choice for researchers who can use them, compared to maintaining local infrastructure. As with monetary cost, the 'green' impact calculation is not simple. To the extent that such a calculation is possible, though, UC policy should reward researchers for identified efficiency improvements.

In short, UC can encourage the appropriate use of cloud services by:

- supporting researchers who identify models for productively harnessing public or private clouds
- developing financial and accounting models that do not penalize researchers who purchase cloud services
- building mechanisms for information exchange and dialog on the topic of cloud computing
- encouraging cloud services providers to work towards open, community-defined standards, and to disclose facts about the energy-efficiency of their services.

Selected References

1. Labate and Korambath, UCLA white paper
2. Walker et al. ;login article [shows poor performance for latency-sensitive apps]
3. nimbus + EC2 for STAR experiment at Argonne [using EC2 to cloud-burst for STAR experiment]

Appendix 4: Cloud Computing in Instructional Computing

Armando Fox, UCB; Russ Hobby, UCD

Key takeaways for Instructional Computing:

- Benefit: Instructional computing is a potential area of major early UC-wide opportunity for benefiting from Cloud Computing, with potential savings administration costs and simplification of IT resource management for both EECS and non-EECS courses.
- Caveat: using CC in courses will *enable* finer-grained metering of a course's IT resource usage; UC will need to speak out on the policy question of whether this necessarily means each course should somehow "pay for itself" or else how those costs will be aggregated and by whom they will be borne.
- Caveat: Courses requiring specific software whose licensing is still stuck in a "per-seat" or other Cloud-unfriendly model may find it difficult to move even if other benefits could be gained.
- What UCOP can do: Be prepared to make policy regarding how to reward courses that can improve their IT resource efficiency using Cloud Computing; be prepared to negotiate with software vendors for Cloud-friendly licensing arrangements where necessary for proprietary courseware.
- Our direct experience converting an upper-division Software-as-a-Service (SaaS) course to Cloud Computing in Fall 2008 confirms that two key aspects of cloud computing—elasticity and virtualization—show promise in making instructional computing both more cost-effective and less labor-intensive. The specific benefits may be different, however, for IT-intensive vs. non-IT-intensive courses.

What are the potential benefits for IT-related courses?

Computing courses face IT challenges beyond those outside the computing discipline. We want to provide our students with instruction that reflects the state of the art in industry and research. Our direct experience converting an upper-division Software-as-a-Service (SaaS) course to Cloud Computing in Fall 2008 confirms that the *elasticity* of Cloud Computing provides both cost benefits and pedagogical benefits.

Enables teaching new concepts. With IT companies like Google running datacenters of tens of thousands of computers, and the major sea change in computer architecture whereby future performance increases will have to come from parallelism rather than faster single-core performance, teaching students about how to express and control massive parallelism is more important than ever. While the concepts are already taught in entry-level and mezzanine-level courses, Cloud Computing gives students short-term access to computing resource at a scale that was previously off-limits to all but the largest organizations. (Even the largest university clusters are trivial compared to a 1000-node "virtual cluster" on a public cloud.) Indeed, the "democratization of big computing" is evident in contests such as Apps for America 2.0, in which contestants write parallel applications that operate on huge publicly-available datasets and exercise hundreds or thousands of machines in parallel.

Enables better teaching of existing concepts. In a UC Berkeley Web 2.0 software engineering course, students got to experience firsthand what happens when a database tries to accommodate too many users. Each student team needed about 10 servers in order to generate enough workload to see this effect. If we had used UC resources, we would have needed over 200 servers to accommodate all the students, even though this lab exercise only lasted two weeks. With EC2, we were able to buy time on 200 servers for a few dollars, releasing the servers after the lab deadline. Similarly, with EC2 we can give every student "superuser" (root) access on her own cloud computing machine—something we could never do for technical and administrative reasons on shared UC Berkeley-owned hardware.

Potentially better technical support. When we moved our SaaS course infrastructure to Amazon Web Services, the students reported that AWS was no harder to use than Berkeley-owned equipment, and since AWS has an active developer community, its question boards, blogs, and documentation are far more comprehensive than what limited course staff could provide.

What are the general benefits to other courses?

Simplified courseware management. Elasticity is a benefit whenever the IT needs of a course have natural peaks (midterm exam, final project, writing assignments, etc.) separated by relative lulls. However, independently of elasticity, virtualization also simplifies courseware management. Today, multiple courses that share IT infrastructure often share a human administrator, who is overwhelmed by often conflicting requests for installing and supporting different courseware for the various instructors. Even if the instructors have savvy teaching assistants, for administrative reasons many University computers are "locked down" so that only designated system administrators can install new software and perform maintenance. With Cloud Computing, each course can create a virtual machine image containing the required courseware; this can be done by a savvy TA or an administrator. This VM image can be re-used for future offerings of the course and modified as needed. If course students "damage" their virtual machine for some reason, it is easy to restore it from the VM image.

Student work product continuity**.** Unlike typical current practices, where course-specific instructional accounts are deleted at the end of the course (forcing students to make local copies of work products if they wish to keep them), cloud computing would allow the work product to outlive the course and potentially be used as the basis of future coursework or other projects. For example, WeJoinIn.com, which coordinates teams of volunteers to staff an activity and was used to organize the ASUC's voter registration drive in 2008, began as a project in our SaaS course. Cloud computing positions projects perfectly for such a transition: the most popular projects can scale up on demand. If this seems an unlikely scenario, remember that the initial prototype of eBay was created over a long weekend by founder Pierre Omidyar.

What are the technical issues and challenges?

- Authentication: some course resources may require authentication by the student or instructor. Public clouds tend to have their own authentication infrastructure; many UC campuses have engineered their own (e.g. CalNet at Berkeley). Bridging these may not be easy, though authentication schemes such as OpenID may help in the future.
- Some courseware hasn't yet moved to a "cloud friendly" licensing model. Although there is an immediate opportunity for a "Save to Cloud" feature in programs such as Microsoft Office, scientific software such as MATLAB and Mathematica have traditionally negotiated per-seat licenses for a specific number of simultaneous users. Even if such software could be made to take advantage of the cloud's parallelism, until the licensing is "cloud friendly" deployment may be problematic.
- Similarly, new recharge models may be needed to support cloud-based courseware deployment. "True pay-as-you-go" will clearly identify which courses use more IT infrastructure than others, and the policy question may arise of whether such courses should necessarily be paying more, or if this is a cost that should be buffered at Dept or University level (as, for example, power costs are today).

Selected References

[1] Armando Fox. Cloud Computing in Education. UC Berkeley iNews, March 2009

Appendix 5: Administrative Computing

Charlotte Klock, UCSD

Summary: Below is feedback from various campus Admin. Computing groups and their current status regarding Cloud Computing. These questions were posed to them via email and the responses have been consolidated. At the time this was written, the task force had not determined a concise definition of Cloud Computing they wanted to present as the over arching meaning so the information below is reflecting a definition that had no bounds around it (i.e. Cloud Computing could mean many things to many people).

Current Situation: state and work that has already been done

1. UCLA: UCLA Administrative Computing has not pursued this as an option yet. Having just finished having the auditors here, there are a lot of questions as to how to deal with security and audit requirements using cloud computing for administrative applications.
2. UCSC: In my operations group, the only thing we are looking at that is somewhat related is "cloud storage" services - like the new offering from Amazon. We are evaluating its use as a cheaper way to do off site backup data replication for DR.
3. UCI: I attended the UC Grid Summit at UCLA last month where cloud computing for research was discussed. This came up in the discussion of 'Eucalyptus' the UCSB ICS project to replicate the Google cloud computing service in freeware. That was interesting to attendees there since it might help with 'checkpointing' and 'job migration' from server to server used in research clusters. As regards using Google for this effort there was little interest since the data transfer and storage needs for large scale research computing would make the Google service too expensive. But with Eucalyptus or such it might make sense for one or more UCs or allied places to offer a 'cloud model' for other UCs to use.
4. UCB: We have been getting customers go to the cloud, usually Slice Host or EC2, most have come back. But the word on the street is how great the cloud is, magic bullet and cheap. But it is not, but it has uses. So my start was really a position statement of when we think people should use it, when they should do it in house. True cost of the cloud and gotchas. How we will help in supporting people on eh cloud etc. I'll get you our write-up next week.
5. UCD: We as a campus have only begun to think of ways of using Cloud Computing for administrative computing needs other than for PPS, which the campus began doing long before the term "cloud computing" existed. We are just beginning the process of using Virtual computing which can be a precursor to cloud computing. We are also looking at the possibility of moving some administrative services off campus to other sites such as SDSC. But as with UCLA, the issues with security and audit requirements will probably play a role in any decisions made to move administrative services to the "cloud".
6. UCSD: various forms are already being used for different applications. These include: Admin Applications-- Sciquest, Skillsoft- UCSD/UCI, Kuali, Mobile Solutions---iUCSD (coming soon), Disaster Recovery- UCSD/UCOP (SaaS type apps). We are looking at what would be possible on a high transaction application and how that could or should be implemented in a cloud environment.
7. UCOP: Cloud computing to me just means using computing services outside of my firewall. It can be from a vendor, from a partner (in UC's case - another part of the system), can be "free", can be something you pay for. From my customer's point of view, for them to do their work they need access to a range of IT services and they probably could care less who provides them or where they geographically reside. An example, some number of people at UCSD might rate PPS as their most important service (I'm not saying best or they like it!!) - but UCSD doesn't run it. Some would say

Google, some might say CDL (California digital library) accessed services. So - we do a lot of cloud computing at UC already.

Reliability/availability: a lot of folks fall back on that as reason to not change our models. The notion that if I control it I can then manage it and deliver the right quality of service and it will be better than if that service comes from somewhere else. Do not buy this reasoning for a few reasons: Most of our users want mobility and flexibility as their top priority; they want to access services from anywhere and anytime. Networks have evolved a lot in the last 15 years and are far more reliable. Wireless works and is no longer some whiz bang trick. Yes - if my users only sat at their desks and the services were only in my network in my data center - then maybe full control makes sense. But, we do not control where users go anymore - so it is just complicated and networks are the system so to speak. To me, where the services are in that network is just a fact of life we have to deal with. For some services, delivering really high availability is expensive and takes sheer size to make it solid. Not sure UC has that capability at any one location to compete with the big shots (Google, et.al.), so controlling may falsely lead us to think it will be better.

Capability and quality: leveraging what other service providers do and the huge investments they make is a way to end up delivering better services to our customers then doing it ourselves in many cases. There are a number of services that UC users need/want that are the same ones everyone else wants (email being one example). There is a security question, but it should not be shot down on first thought and in fact email outsourcing is catching on as I type. What UC should focus much of our IT might on are services unique to our institution, not the utility ones and there are cloud services for some of those now.

Potential Use Cases

1. What UC should do for administrative services relative to cloud or not is understand total cost of ownership by providing it at each location vs. providing it in common somewhere w/in UC vs. providing it from an outside service provider. If costs are close then decisions of security and control might prevail. If costs are not close, security and control might be a factor but maybe decision would be different. Another angle is to use shared infrastructure but still retain local control of the application (i.e. - have a VMware farm and leverage an economy of scale but have each location still administer their own virtual servers). Once you emotionally let go of where the gear is geographically, you could then decide to share amongst UC or to just get the service from somewhere else.
2. Off site backup data replication for DR. Could be used for smaller departmental systems. Network issues (IP addressing) would need to be resolved. Database structures within applications sometimes do not lend themselves to just adding capacity to the application. There would need to be some architecting review to ensure data integrity was maintained.
3. Administrative computing needs other than for PPS
4. Configure additional web farm servers to be used during peak student registration periods and then decommission them once the workload goes back down.
5. Build test and QA environments to mimic production capacity for performance testing of applications.
6. Any application that uses non-restricted data (the challenge is knowing you only have this type of data)
7. Systems that you do not want locally like a Disaster Recovery system for a campus portal or email services.

Low-hanging fruit and things that are not ready for cloud computing (or for which cloud computing is not yet ready)

In a sense, everyone does cloud computing already as many things our users do is go to the internet to research something, find some info - and that processing is being done in the cloud (i.e., you don't really know where it is coming from as an end user and don't care and I as the IT service provider am not providing it w/in my firewall boundary that I control). Many of our users use their own email account from work, they use Facebook, they go to Google, they store their music online somewhere, have their spreadsheets/word docs at Google apps, have photos stored somewhere, mix and match where data is kept for work and for personal - intermingled on their PC/MAC, iPHONE-like gadget, on home devices. I would bet if you polled our users and had them rate what services they most value from most to least, a number of the top 10 would NOT be IT services the local IT shop provides.

Costs

Recommend that the way to rationally decide on how/where/when cloud computing would play for UC is via cost analysis - true and honest costs that are normalized and holistic, without fund boundaries clouding things and without worrying about who pays - but get down to the bottom line. Then line that up with control, quality of service and security issues and attempt to make an objective and informed decision. We should not hide behind policies and the way we did it in the past.

Sustainability

One massive challenge I see is to use jigsaw puzzle as an analogy. Each piece interacts with others and removing one piece may be difficult. But if you had a piece in the puzzle that only intersected with one other - you might be able to remove it and provide it from somewhere else with single interface. But some pieces intersect with too many others and it isn't worth separating out. In IT, there are many interactions between things. So looking at which ones you can remove and provide in another manner might possibly be matter of finding those that are most stand alone. A potential obstacle for UC to really leverage cloud concept is finding the pieces with the fewest interfaces, or finding a group of services that are all related and take them as a whole. One campus recommended to UCOP that they could provide VMware based windows servers. My problem with that is our windows team spends most of its time interacting with network and apps teams, our help desk, and our change mgt process - not setting up virtual servers itself. Removing that one piece, to me, does not deliver enough value as the overhead of redeveloping multiple interfaces is not worth it regardless of whatever savings we may have derived.

Solution - look for integrated suites of services from the cloud or discrete enough ones that the interface is simple enough to address.

Security

1. There are a lot of questions as to how to deal with security and audit requirements using cloud computing for administrative applications.
2. The issues with security and audit requirements will probably play a role in any decisions made to move administrative services to the "cloud".
3. Given that freely-available internet applications with real potential to improve productivity are proliferating, university employees now have a great degree of choice in how they meet their work-related computing needs. Yet the exercise of such choice can cause a university to suffer a near complete loss of control over confidential university information, student and employee personal information and other work-related information in which it has an ownership interest. Consider the example of an administrative staff member who uses an externally-hosted and freely-available internet-based "wiki" to encourage collaboration on her department's work. She signs up for an account and invites colleagues to do the same. She and her colleagues agree to terms of service that grant them some control over the web pages they create, but the service provider also reserves a

very broad right of use over content. Most significantly, the university itself is an apparent third-party to the contractual relationships entered into by its employees, and has no effective right of control. The wiki works well, and after a year of use, it contains a significant amount of useful and sensitive content, all of which is related to the department's business and some of which includes student personal information. The wiki is password protected, but the security-related promises in its terms of service pale in relation to those the university requires in its own outsourcing contracts.

Here are the main risks and costs associated with this arrangement:

- The information is not necessarily secure from loss, theft and misuse.
- Regarding student personal information, the university may be in breach of the safeguarding duty imposed by the Freedom of Information and Protection of Privacy Act.
- It is relatively easy for the participants, should they depart from employment, to take the information. The university may have a legal right to control it based on its right as employer, but the service provider will not likely recognize this right absent a court order.
- The university does not know that information is there. If there is a legal claim to which the information relates, it may be overlooked in the university's e-discovery process. The university could lose the benefit of the information if it is helpful evidence or, if it is not, may face production related sanctions. In either case, it has now become more costly to search for, retrieve, process and produce electronically stored information in the course of the litigation.

Conclusion from an Administrative Perspective

Based on feedback from some of the Administrative Computing groups across UC, the below are highlights of where Cloud Computing may or may not fit into the near term strategy for these units. This is based on a general definition of the following; any IT resource used outside of the local UC campus. This can be Amazon, Google, IBM, EC2, or another UC campus. In general, the different campuses view there are many "potential" uses but the security aspects of implementation and ongoing support impact/prevent many benefits from this type of service.

Examples of CC already being used across UC Administrative systems include: Sciquest (e-commerce at UCSD), PPS (9 campuses use UCOP resources), gMail (UCD students), Recruit (UCSD/UCI recruitment application hosted at UCI), APOL (UCSD/UCI academic personnel hosted at UCSD), iUCSD (iphone app. Hosted/built by vendor), data replication (UCSD, UCOP, UCSB, UCR, UCB, etc...), Disaster Recovery (UCSD/UCOP, UCB/UCLA, UCSB/UCOP, etc...), other benefit/retirement apps (UCOP hosted), etc... The "Google Apps & gMail" suite, is being looked at across multiple campuses for applicability and cost saving measures but is still not being used widely due to privacy and other legal issues.

In order to take advantage of the Cloud both inside and outside of UC, the following would need to be in place:

- Non-restricted data (the challenge is knowing you only have this type of data)
- Not a high activity OLTP system. Performance is a potential show stopper across the network. The application architecture plays a large role in this in that depending on the database structure and how the application is built to do updates, performance, licensing, ongoing support, etc. could be factors on whether or not an application could be moved to the cloud.
- Has available a sophisticated user who can manage their own System Administration work.

AND at least one of the following:

- "Elastic" workloads like applications that are much more heavily used at start or end of term.

- Systems that you don't want locally like a Disaster Recovery system.

The Pros include:

- Can spin up a new instance in 15 minutes
- Only charged for resources you do use

The Cons include:

- Getting support can be difficult
- SLAs say they are best effort only
- database structure could be important factor on performance

Gotchas:

- Usually support of OS, patching and middleware are not thought about.
- Cost is usually more than thought since with many models you are charged for bandwidth or transfer fees which can really add up.

Conclusions

We as an entity need to rationally decide on how/where/when cloud computing would play for UC via cost analysis. This would need to be true costs that are normalized and holistic, without fund boundaries clouding things and without worrying about who pays. Cutting the bottom line and providing good service are the goals. Today there are too many policies and procedures that are in the way to effectively make this happen in any way other than one-off exceptions. Next would be to line that up with control, quality of service and security issues and attempt to make an objective and informed decision.

Finally, the issues with security and audit requirements will play a role in any decisions made to move administrative services to the "cloud". Under the advisement of UC legal counsel, UC Management MUST make a decision as to viability or not so resources can either move forward or drop these conversations and work on other priorities and solutions until such time an answer can be had that will stand up across all campuses and administrative systems.

Appendix 6: Legal and Contractual Issues

Michael Mundrane, UCB; David Walker, UCD

Introduction

In the current technology climate, there is a growing suggestion that cloud services can be utilized to provide flexible capacity and a lower price point compared to their local equivalents. These represent tangible benefits to institutions that can successfully exploit these emerging services offering to provide flexible and high quality options to their respective communities. While the potential to leverage these new systems or services is exciting, there are also potential pitfalls with respect to data covered by a variety of laws or other controlling policies.

Discussion

Universities are complex institutions with varied constituencies and with real needs to hold and manage data. The data itself is potentially sensitive in nature and much of it is protected by one of a number of laws as well as various policies and guidelines. Institutions are obligated to meet all legal requirements, either locally within their own systems and operations, or remotely through any service providers. Cloud computing, regardless of current excitement, is still a service delivered by an external provider. The nature of these services may be new and there may be fresh potential, but the relationship between institution and cloud is not fundamentally different than those we are already familiar with. More importantly, information covered by law or policy that moves off site, is still covered by these same laws or policies. Unfortunately, care and disclosure obligations may be much stricter than standard service provider service agreements.

The Health Insurance Portability and Accountability Act (HIPAA) provides federal protections for personal health information held by covered entities and gives patients an array of rights with respect to that information. The privacy rule is balances to that it permits the disclosure of personal health information needed for patient care and other important purposes. This comprehensive rule fully regulates the use and disclosure of "individually identifiable health information." So, while there may be information that would appear to be innocuous (Hepatitis vaccination for incoming students), it is all still covered by HIPAA. This means that there is a due diligence obligation to protect this data and to control its' use and release. In fact, before HIPAA controlled information may be transferred to a service provider, the institution and the service provider must enter into a "business associate agreement." This data may not be sufficiently protected based on basic cloud provider service level agreements.

The Violence Against Women Act (VAWA) is a United States Federal Law intended to enhance the investigation and prosecution of violent crime perpetrated against women. It prohibits all nonconsensual disclosures of data unless compelled by a court order. This act applies to domestic violence service providers. University health providers or other organizations that provide services to victims of domestic violence (for example) may come under this requirement. Data and records from these activities would likely be controlled by this act and may not be sufficiently protected based on basic cloud provider service level agreements.

Typically, states have laws that govern the control of personal information (usually name and one other piece of data such as SSN, drivers license number, bank account number, credit card number, etc. While these laws frequently suggest only reasonable care and protection of this data, they generally have rather explicit notification requirements in the event that this data is compromised. Cloud service providers holding this data would need to know what data was compromised, but the institutions would generally be responsible for the notification. These notification costs dominate the expense of any data compromise and provider service agreements usually indemnify them from these exposures. A provider service agreement may very well describe reasonable security, but any breach of private information would be still be managed by the institution.

In addition to legal obligations, there are potentially other obligations that, while not legal, have a similar impact on institutions. The Payment Card Industry (PCI) requirements may not be legal, but they are onerous and they apply to any credit card data that an institution may have. If this data is preserved locally, the institution must comply with a broad array of practices in order to meet the PCI standard. If this data is moved into the cloud, the provider would have to meet the PCI standard. It is highly unlikely that would be the case and the provider service agreement will not allow for a lower standard of security. The institution will not be relieved of their PCI obligations by moving this information to a cloud provider and they will still be responsible for any breach associated with this data.

The sharing of legally privileged information varies based on the privilege and the state. Some examples of this type of information would be between doctor and patient or lawyer and client. The impact on privilege when moving data to a cloud provider is not necessarily clear. If the provider, through agreement, does not have the right to view the data, then privilege may still hold. If the provider, through agreement, has the right to view the data, then it is much more difficult to argue that the privilege holds. This also holds for any secrecy obligations. When information is required to be held secret based on a privileged relationship, then a service provider agreement may constitute a breach of professional obligation.

In addition to obligations on data, cloud providers may be compelled to disclose information and may, or may not, indicate this to their customers. The Electronic Communications Privacy Act (ECPA) protects electronic communication both in transit and while stored. It specifies requirements for search warrants. The ECPA is significantly out of date, and there is legal precedent that the data held can be obtained by warrant without the institutions knowledge. It is difficult to ascertain how ECPA applies and its' interpretation could be by judge at the time a warrant is written. Regardless, the USA PATRIOT Act may simply over rule the ECPA through its' expanded use of National Security Letters. Although a court order may be required, it seems clear that the USA PATRIOT Act extends the FBI's authority to records maintained by an external cloud provider.

Data may physically reside in locations that are not in either the state or the country of the institution. Under these circumstances, there is a notion of virtual protection and then there are local laws to which the local holder of data is obligated to comply. These laws could compel the data holder to release information under circumstances that are not known to the data owner. A service agreement cannot be written that would allow the data holder to circumvent local law. This particular risk may be greater when data is held by international cloud providers. They may not be able to specify within the service agreement where the data might physically reside. Data held in a foreign country without the comparable privacy protection laws cannot be assumed secure.

These concerns aside, there is also some question regarding the destruction of data. A service agreement may not be explicit with respect to disposition of data after contract termination. Regardless, even a specific clause in a service agreement may not be sufficient if a provider files for bankruptcy. Under these circumstances, there may be few options to compel the provider to destroy and dispose of data. This is not to say that data would not be destroyed, but bankruptcy laws provide very limited protection to make this happen. The disposition of data in the event of a failed cloud provider may not be known and agreements may not be binding if there are insufficient resources or liquidated assets to meet these obligations.

Moving data out to a cloud provider does not release the data owner from legal obligations to protect and preserve information requested via any discovery mechanism. The provider service agreement may not have a provision to facilitate this activity, but this would not relieve the data owner of this responsibility. If access to data from the provider is sufficient in terms of throughput, then even large discovery requests can be met by moving copies of the requested data out of the cloud and into local storage. This mitigation strategy does require the provisioning of reasonable local storage in order to meet potential discovery

requests. To facilitate this, it is important that the institution and the provider both adhere to common standards for storage or have robust strategy for dealing effectively with inconsistencies.

Conclusion

It is certainly possible to employ cloud services or infrastructure as part of an overall technology strategy. There are a variety of potential advantages and potentially some disadvantages. If an institution is not clear about the content of their data, have no business moving this into the cloud. However, being aware of the content also requires that an institution be aware of any laws or policies govern the management and disposition of the data. Moving data that is protected or controlled by one or more laws or policies into the cloud does not absolve an institution of any obligations or requirements established by these laws or policies. Further, laws and policies that apply to the data owner are not always the same as those that apply to the data holder. Basic service agreements between the institution and cloud providers are unlikely to be sufficient and disposition of data during either warrant or end of life may not be known or controllable by the institution.

An institution can meet discovery and provide additional protection by maintaining either temporary or permanent local copies of selected data. Regardless, exposure can be further reduced by either scrubbing the data so that information moved to the cloud will not be controlled by any of the laws or policies cited here or they can have a service agreement written that specifically meets all requirements and obligations. The latter is not likely at this time and this leaves the former as the best current alternative.

Appendix 7: Cloud Computing Continuity

Russ Hobby, UCD

In the past where the hardware was owned and operated by the same organization that was responsible for applications that ran on the hardware, changes and upgrades to the infrastructure were coordinated between those that operated the hardware and those that used it. For large applications, major upgrades required much planning on all sides. Upgrades and changes were also performed to best serve the applications.

Cloud Computing may separate the infrastructure and applications into different organizations whose goals may not always be aligned. Thought and planning are needed to consider actions that can be taken when they diverge to the point where it is no longer a workable solution. Situations that could lead to this problem include:

Cloud Provider

- Goes out of business
- Costs increase beyond what is affordable
- Upgrades system too incompatible with client applications

Applications

- Upgrade to require resources that are not available from the provider
- Have increased security requirements
- The application outgrows a private cloud

For these and other reasons, part of planning of the use of cloud computing needs to be an exit or migration strategy. This can include emergency migration that could be part of normal disaster planning. As normal software upgrades need to occur situations will change and exit and migration strategies need to be reevaluated. There is also end of life migration issues when it comes time to replace the application with a new one.

Factors that can ease migration are adherence to standards, though many aspects of cloud computing have no standards set yet. A set of attributes and parameters should be established to enable comparison of cloud services. This list can be used to determine if an application can be easily moved to another system and identify potential problems in such a move.

Another consideration in evaluating compatibility between cloud services is different cloud resources working together to create an overall applications system. How easily can tasks be implemented on different clouds and maintain a smooth workflow? Where is data stored and how is it accessed or shared?

Appendix 8: Capacity Issues for Cloud Services

David Walker, UCD

Because of the dynamic nature of the allocation of resources to applications within a cloud, where processors and storage can be allocated in large quantities at a moment's notice, the perception is that the cloud's resources are infinite, able to accommodate any demand. Of course, this is not true in reality. Cloud service providers must project likely demand and deploy enough capacity to avoid having to deny resource requests because of a lack of physical resources.

The amount of spare capacity maintained by a cloud service provider, then, becomes an issue of service level and cost. Too much spare capacity results in higher costs than needed and too little results in frustrated resource requests. The spare capacity required is a function of the demand rate and the time required to increase physical capacity. Cost is, in part, a function of the spare (unsold) capacity. The following example illustrates these interrelationships.

> A storage provider has a service level agreement that says that 99% of requests for storage up to 1 TB can be fulfilled within a minute. The storage provider has projected that the 99th percentile of demand will be an aggregate of 1 TB in requests per day and is able to deploy additional disk drives within two weeks in quantities of 100 TB. With these assumptions:
>
> - The service level will be met, as long as the storage provider orders an additional 100 TB whenever spare capacity drops below 14 TB.
> - The pricing model must recover the cost of unsold spare capacity that varies between 14 TB and 114 TB over time.
>
> If, on the other hand, the 99th percentile of demand were 20 TB/day, then an additional 100 TB would be required whenever the spare capacity dropped below 280 TB, and the pricing model would need to accommodate unsold capacity varying between 280 TB and 380 TB.

Note that the impact of spare capacity on pricing may be large or small, depending on total physical capacity. If a storage provider's total capacity were 2000 TB, then maintaining 14 TB to 114 TB (up to, say, 6%) of spare capacity might be acceptable, but 280 TB to 380 TB (up to 19%) might not.

Conclusion

In order to ensure that service expectations are met, cloud service providers must balance the following factors in designing their services:

- Projected demand
- Service levels
- Time to increase physical capacity
- Price

Cloud service consumers should consider these factors when evaluating potential providers.

Appendix 9: Capacity Issues for Network Infrastructure

Michael Van Norman, UCLA; Ken Lindahl, UCB

Overview of Bandwidth and Latency Issues

Bandwidth and latency issues arise from the need to move data in and out of the cloud. We assume (for now) that the Cloud is located outside any campus (i.e. either "Public Clouds" or a "Private Cloud" located at a shared data center/collocation facility). We assume that researchers would move data in/out of the cloud from/to a campus-resident storage facility, but note that there will likely be cases where this is not true: where researchers will want to transfer data from storage outside the campus directly to the cloud. There are additional bandwidth and latency issues that arise from the possibility that applications may be "pulled apart" and run on different clouds; these are not addressed here.

EECS paper reported average bandwidths of 5 to 18 Mb/s writing to Amazon's S3 cloud. Based on those results, they postulated a "best case" scenario in which a researcher writes 1TB of data to S3 at 20Mb/s. On average: it would take approximately 45 days to complete the data transfer. We consider this scenario unacceptable and identify the issues that need to be addressed in order to resolve this problem.

Network capacity/capabilities

Campus

At present, most UC campuses have multiple Gigabit Ethernet (GE) connections to the CalREN networks, CalREN-HPR and CalREN-DC. A few campuses have upgraded one (or more) connections to CalREN-HPR to 10Gigabit Ethernet (10GE); no campus (to our knowledge) has plans to upgrade its CalREN-DC connections to 10GE. At the larger campuses, these GE connections are largely consumed with normal day-to-day campus traffic demands; a significant increase in bandwidth due to cloud computing would require funding additional connectivity between the campus and the CalREN backbone(s).

Many campuses operate border firewalls or packet filters that restrict various network protocols (SNMP, SMB, NFS, etc.). It may not be possible to utilize or manage some cloud computing services, especially storage-based services, if these restrictions are not relaxed or removed. In the case of computing services that demand very high levels of performance, these devices might introduce performance penalties that compromise the cloud computing service.

Similar issues were discussed and addressed by the ITGC's Advanced Network Services Work Group recommendations; in particular, recommendation #3, Enhance Network Connectivity, directly speaks to the need to upgrade campus connectivity to CENIC's networks.

CENIC

The CENIC backbone is believed to have sufficient capacity to support the initial stages of a cloud computing roll-out, but would likely require augmentation to support large scale use of cloud computing (either internal or external). If increasing capacity can be done within the existing footprint of the CalREN networks (e.g. by adding line cards and transponders), this work could proceed relatively quickly given available funding. However, if increasing capacity to meet the needs of cloud computing were to require additional rack space and/or power at the CENIC POPs, significant delays could occur as space and power are not readily available at all collocation facilities.

UC campuses have connections to both the CalREN-DC and CalREN-HPR networks. In general, higher capacity and performance is available via the CalREN-HPR network, and it is assumed that most cloud computing connectivity should be provided via that network. In the case of internal clouds, this is

relatively easy to ensure. However, in the case of externally provided clouds, it is most likely that the external provider will be connected to the CalREN-DC network and not the CalREN-HPR network. Heavy utilization of external cloud computing services might require significant re-engineering to match traffic patterns to the available network topology.

Exchange/Peering Points

If cloud computing services are provided by institutions not directly connected to CalREN networks, the traffic to and from those clouds will need to pass in and out of the CalREN networks through established peering and transit points. CENIC maintains a large number of peering relationships; however, these connections are sized to cover existing traffic loads. Assuming the large scale use of external cloud facilities, the bandwidth provided at these peering facilities will need to be increased to avoid negatively impacting existing use of the network. Additionally, the geographic/topological placement of these facilities might need to be reviewed to address latency or other network performance related issue. Both increasing bandwidth and changing peering locations involve, often significant, costs to CENIC, and by extension the CENIC membership. Settlement free peering (via the CalREN-DC network; see above) is in place for connectivity to the largest cloud computing providers (Google, Amazon, Microsoft); the use of providers for which settlement free peering is not available will require the payment of ISP bandwidth fees.

L2/L3 Issues/Concerns

The preceding sections address capacity on the campus and CENIC Layer 3 (routed IP) networks. Applications requiring very high bandwidth (i.e. approaching 10Gbps) or very low latency/jitter might be better served by a Layer 2 connection: either a dedicated wave on an optical network with Ethernet presented at both ends, or a VLAN configured on a switched Ethernet network running on top of an optical network. CENIC offers both types to campuses connected to the HPR network. (The switched Ethernet network is presently being built; it is expected to be available before the end of calendar year 2009.) Thus, an L2 connection between any UC campus and a cloud that is directly connected to CENIC's optical network will be relatively easy and inexpensive.

Additionally, L2 connection capabilities are present (or soon will be) in both national R&E networks, the Internet2 Network and National Lambda Rail (NLR). An L2 connection between any UC campus and a cloud connected to either the Internet2 Network or NLR will be relatively easy to set up, but will involve additional costs that may be significant.

L2 connections to a cloud not directly connected to the CENIC, Internet2 or NLR optical networks will likely be challenging and very expensive.

However, campus security concerns arising from these kinds of connections remain largely unresolved, since in most cases they will bypass existing campus firewalls and intrusion detection/prevention systems. Addressing security concerns on such "bypass networks" will require additional campus resources, both human and machine. These concerns already exist in the larger context of research computing, of course; they are not confined to cloud computing.

It should be noted that we are unaware that any existing cloud provider has been asked if it would support an L3 connection. The large public clouds have clearly made large investments in their L3 connectivity and might be understandably reluctant to consider alternatives. That seems likely to translate to a requirement that the organization requesting an L2 connection pay the entire cost of building and operating the connection. Even so, use of an L2 connection over the CENIC operated

portion(s) of the path, and possibly over any NLR or Internet2 portions, could provide sufficiently improved performance to make it worthwhile.

Appendix 10: Environmental Impact of Cloud Computing

Greg Bell, LBNL

Executive Summary

- Lack of published data makes it difficult to assess the environmental impact of moving a local IT workload into a cloud-services environment.
- The impact can be estimated, though, based on vendor reports and on reasonable assumptions.
- Bottom line: power consumption per IT service unit is likely to be lower (perhaps by 25% or more) in a cloud-services environment.
- This assumes that campus data centers have achieved server utilization rates comparable to those of cloud providers; most have not, which means that electrical savings may be even greater.
- Cloud vendors should be encouraged to develop consistent, verifiable disclosure practices about the environmental impacts of their services.

Introduction

It is difficult to assess the net environmental impact of migrating a local IT workload to a cloud-services environment. Even if UC campuses understand the impact of hosting local IT services (which is not always the case), they probably cannot obtain comparable information about cloud-service providers. Without accurate local data and without vendor disclosure, it's necessary to rely on the incomplete information currently in the public domain, and on reasonable assumptions about the operational practices and economic incentives of cloud vendors.

Discussion

Despite the difficulty of making a precise comparison, rough estimates are possible. The huge scale of cloud-service data centers implies correspondingly-large operating costs, as well as strong incentives to conserve power and water on the part of data center owners. Some cloud vendors have released information about the design innovations of their large-scale data centers. These include elimination of conventional UPS systems, elimination of chassis fans, novel air-flow management techniques, water conservation and recycling measures, DC power distribution, and use of high-density modular enclosures. Google, Yahoo, and Microsoft all claim energy efficiency metrics that far exceed the values commonly associated with smaller, lower-density data centers (see LBNL benchmarking study).

In order to roughly assess the environmental the impact of 'cloud-sourcing', we make a few simplifying assumptions. First, we ignore the issues of water consumption, carbon emissions, and waste disposal in order to focus on the more tractable question of electrical consumption. Second, we assume that server hardware in both environments is roughly comparable and that the degree of server utilization in each environment is roughly similar (in other words, we do not compare a non-virtualized local environment with a virtualized cloud environment). Not that this final assumption almost certainly has the affect of underestimating the electrical savings associated with cloud-sourcing, because deployment of virtualization technologies at UC data centers is spotty.

Given those assumptions, the relative electrical consumption for each environment can be modeled using a simple formula:

ElectricalUsage[cloud] = ElectricalUsage[local] * speed factor * energy efficiency factor

'Speed factor' is a ratio between the average time to complete a compute job locally, and the average time to complete the same job in the cloud, based on the workload under consideration. Tests by the IT Division at LBNL, as well as published reports, have shown that speed factor can vary by a factor of 10 or more in the Amazon EC2 environment, depending on the nature of the computation. High values are associated with scientific codes that are extremely sensitive to network latency, and that normally run on clusters with high-speed switching interconnects; high-speed interconnects are not currently available in

the EC2 environment. For this reason, there is no generic speed factor. Any IT organization that wants to consider cloud-sourcing should carefully measure the speed-factor of relevant jobs and processes. 'Energy efficiency factor' is the ratio between the Data Center Index of Efficiency (DCiE) of the local data center and that of the cloud-based data center. The energy efficiency factor will probably require estimation, unless a site has accurate information about local DCiE and can obtain accurate DCiE from a cloud-services vendor. If we assume that the average local data center has a DCiE of about .5 (which is consistent with benchmarking studies), and that a cloud-based data center has a DCiE of .8 (which is consistent with published reports), then the energy efficiency factor becomes .5/.8, or .625.

It is necessary to choose a specific outsourcing scenario in order to use the formula. Assuming the workload is computational genomics, the measured speed factor 1.2, and the estimated energy efficiency factor .625, then electrical usage in the cloud would be approximately 1.2 * .625, or .75 local usage, and the energy savings approximately 25%. If the same site wanted to migrate an already-virtualized web-services environment into a cloud-services environment, and measured the speed factor in this case to be 1, total energy savings might exceed 35%. On the other hand, for latency-sensitive scientific codes with high speed factors in EC2, energy consumption is likely to be higher in the cloud.

Conclusion

By making estimates about the efficiency of data centers in local environments and in the cloud, and by measuring the performance of target applications in each environment, we can roughly calculate the energy savings, if any, associated with cloud-sourcing. In many cases, cloud-sourcing is likely to produce significant energy savings per unit of IT service. However, reliance on unverifiable efficiency claims from vendors presents a source of potential error. Because consumers of cloud services may be motivated by a desire to decrease their environmental footprint, the authors of this report urge that cloud providers disclose the environmental impacts of their services. It is encouraging that some vendors have released limited information. However, we urge that vendors work together to develop consistent, verifiable disclosure practice (encompassing electricity, water, carbon, and equipment lifecycle management), so that customers can make more informed choices about the environmental impact of their purchases.

Appendix 11: Cloud Computing Economics

Armando Fox, UCB; Bill Labate, UCLA

Key points for Cloud Economics:

- **Benefit:** Usage-based pricing—true pay-as-you-go with a pricing model that scales up and down with your *actual* resource usage—is a key to Cloud Computing and to enabling its "elastic" properties. Properly exploited, it can result in significant cost savings in a variety of variable-demand scenarios.
- **Caveat**: True apples-to-apples comparisons between private and public clouds may be very difficult because hidden, bundled, shared, or aggregated costs implicit in the private scenario distort the comparison to public-cloud pricing.
- **Caveat**: The low price (and attendant potential savings) associated with public Cloud Computing derive largely from the unprecedented economies of scale enjoyed by the cloud operator when building the mega-datacenters housing the cloud facilities. In addition, the capital expenses for those datacenters had already been justified by a prior business need. Consequently, "converting" a private cluster to cloud computing, or building and deploying a private cloud, is unlikely to result in the same low usage costs that public Cloud Computing providers offer now.
- What UCOP can do: Approach cost-based comparisons with cautious skepticism.

Usage-based Pricing is Key to Cloud Computing

The key economic property of Cloud Computing is the ability to pay exactly for what you use, even if your usage is nonuniform over longer time periods. This *usage-based pricing*, as it has long been called in the networking community, is not the same as renting. Renting a resource involves paying a negotiated cost to have the resource over some time period, whether you use it or not. (Renting a car costs $50.00 a day whether you're driving it or leaving it in a parking lot.) Pay-as-you-go involves metering usage and charging based on actual use, independently of the time period over which the usage occurs. Note that the metering itself is critical: pay-as-you-go implies the ability to measure usage at a fine grain, and this capability is not easy to retrofit into existing systems.

Three particularly compelling use cases favor usage-based pricing:
(1) Demand for a service varies with time. Provisioning for peak load means resources are wasted at nonpeak times. Note that even if the hourly rate to rent a machine from a cloud provider is higher than the rate to own one, money may still be saved. For example, suppose that for 4 hours each day, a service experiences peak demand requiring 500 servers; the other 20 hours each day, it requires only 300 servers to meet average demand. If the service operator owns their own servers, it would have to purchase enough to handle the peak: 24 hours x 500 servers = 12,000 server-hours, even though the actual requirements over one day are (4 x 500) + (20 x 300) = 8,000 server-hours. The "buy your own" scenario effectively costs 1.5 times as much per day, so as long as the usage-based pricing is less than 1.5 times the purchase price (under suitable depreciation assumptions), cloud computing is still cheaper. One example of this in the UC system is courses that have periodic assignment deadlines (when there will be high demand for computing) interspersed with long periods of lower demand.
(2) Demand for the service is unknown in advance. In this case, the risk of making a wrong prediction for provisioning is shifted from the service provider to the cloud vendor. A UC example would be email or information dissemination during an emergency: demand for computing might temporarily rise to an unknown level.
(3) A corollary of usage-based pricing is *cost associativity:* using 1,000 computers for an hour costs the same as using 1 computer for 1,000 hours. Batch computation that parallelizes well can exploit this property to do computations much faster. A UC example that has actually occurred is getting research results in hours instead of days, enabling much faster research progress.

Cloud Economics: Caveats
The potential cost advantages of cloud computing may seem overwhelming. Indeed they may be, but several caveats are in order when trying to make an "apples to apples" cost comparison.

Caveat: Due to the sizes of their datacenters, public cloud providers such as Amazon AWS were able to realize economies of scale, building and operating their datacenters 5 to 7 times cheaper than if the datacenters were medium-sized. In addition, these datacenters already had to be built to satisfy an existing business need; Cloud Computing began as an *additional* incremental revenue stream exploiting them.
Implication: Constructing a "private cloud", or converting an existing private cluster to cloud computing by enabling metering and moving to usage-based pricing, may not yield a cost-per-resource that is necessarily competitive with public clouds. The CACM version of the Above the Clouds paper (http://abovetheclouds.cs.berkeley.edu) further explores which potential Cloud Computing benefits apply to all clouds vs. which ones derive directly from economies of scale and therefore might not apply to smaller-sized private clouds.

Caveat: Due to high volumes of usage and a "low touch" service model, public cloud providers can very predictably factor their operational overhead costs into their usage-based pricing.
Implication: In existing private IT facilities, many costs are often absorbed into general overheads (power consumption, physical plant upkeep, physical security) or are shared among business units or labs (system administration, technical support). "Unbundling" these costs may be essential to facilitating a direct cost comparison between insourcing vs. outsourcing. In addition, there may be benefits such as simplified software management that may be difficult to quantify in advance.

Caveat: Even if a direct cost comparison is possible, in some cases other factors may dominate the decision.
Implication: The economic benefits of Cloud Computing have been widely touted, so it is easy to over-focus on cost as a factor in the decision whether to adopt cloud computing in some particular scenario. Any decision process should completely expose all the factors involved in the decision, including unbundling costs that are currently aggregated ***and*** examining non-cost-related benefits and drawbacks, whose overall economic impact may be hard to quantify. (How do you quantify the benefit of breakthrough research? How do you quantify the penalty of having sensitive student data compromised with no clear legal strategy for handling a sensitive situation?)

Example: Comparing SDSC/TAPP with Amazon Web Services
The San Diego Supercomputing Center's Triton Affiliates & Partners Program (TAPP) is a new charging model for providing high performance computing resources to UC researchers in a cloud like manner. TAPP allows researchers to purchase computing time on one of three different hardware resources: the Triton Compute Cluster, PDAF-256 GB and PDAF-512 GB. Compute power usage is based on "SU" or core hour with a sliding scale depending on the hardware used; the Triton Compute Cluster at $0.06/SU (also called the Base SU), PDAF-256 GB at $0.12/SU and PDAF-512 GB at $0.24/SU. One caveat is that TAPP requires a minimum of 200,000 Base SU's or an initial outlay of $12,000, although this outlay can be used to purchase any combination of quantity or hardware type for actual usage. In contrast, public clouds including Amazon EC2 and Microsoft Azure require no up-front financial commitment.

TAPP also provides some key services that are either absent or less developed in the case of other cloud providers. The two key areas are (a) user support from basic system help to training programs, and (b) the availability of several popular scientific software packages such as SAS, Star-P, Fluent and Techplot. Another possible benefit for TAPP users is that there is no need to create and deploy their own system images; even though most cloud providers have pre-built images available, customizing them still presents extra complexity that TAPP does not require. On the other hand, TAPP currently lacks any

large-scale user storage or archival storage, although scratch space is provided. SDSC does intend to add additional storage options in the future through their Data Oasis service.

Appendix 12: Charge Letter

(3/19/2009)

The Information Technology Leadership Council (ITLC) has created an *ad hoc* Cloud Computing Task Force (CCTF) to assess cloud computing and its applicability within the University of California. Membership of the CCTF is drawn from the following standing subcommittees of the ITLC:

- UC Research Computing Group (UCRCG)
- Joint Data Center Managers Group (JDCMG)
- Communications Planning Group (CPG)
- UC Grid developers
- UC cloud computing researchers

The CCTF will address the following issues:

- A definition and description of cloud computing, including its relationship to other architectural components, such as clusters and grids.
- The potential of cloud computing to address the following:
    - High-performance computing
    - Administrative computing
    - Large-scale storage
    - Disaster recovery
    - Non-stop / fail-soft services
    - Green computing
- Considerations for the deployment of cloud computing, including:
    - Security
    - Capacity planning (network, processing, storage)
    - "End to end" system performance
    - Portability
    - Legal and contractual issues
    - Data center planning
    - Build *vs*. buy *vs*. sell (commercial, government, and internal UC providers)
- Recommendations to the ITLC for deployment and further study of cloud computing technologies, including an ongoing organizational structure.
    - Consider whether these activities might become incorporated into the missions of existing ITLC groups, or a new group (perhaps transitioned from the UC Cloud Computing Task Force into a Research Technology Group or an Internet Services Technology Group)

The CCTF's final report will be delivered to the ITLC by July 31, 2009 after review by the groups represented in the CCTF's membership, as well as the Information Technology Architecture Group (ITAG). Once approved by the ITLC, the report will be incorporated into the ITAG's architecture repository.

Appendix 13: CCTF Membership

- Greg Bell, LBNL
- Armando Fox, UCB
- Bob Grant, UCR
- Russ Hobby, UCD, Co-convener
- Charlotte Klock, UCSD
- Sriram Krishnan, UCSD
- Bill Labate, UCLA
- Ken Lindahl, UCB
- Michael Mundrane, UCB
- Philip Papadopoulos, UCSD
- Mike Van Norman, UCLA
- David Walker, UCD, Co-convener

These members were drawn from the following UC communities:

- UC Research Computing Group (UCRCG)
- Joint Data Center Managers Group (JDCMG)
- Communications Planning Group (CPG)
- UC Grid developers
- UC cloud computing researchers