



---

# A Cloud Guide for HPC

## Top Drivers, Barriers, Use Cases, and Vendor Requirements for Private and Public HPC Cloud Computing

White Paper  
May 2009

**Author:**  
Ashar Baig  
Principal, Strategic Insight Consulting

### Summary of Contents

---

1	Executive Summary .....	4
2	Why Cloud? Top Drivers for Cloud Computing .....	6
3	Public versus Private Clouds .....	7
4	Top Barriers to Cloud Adoption .....	9
5	Vendor Requirements: What to Look For In a Cloud Vendor .....	13
6	Are All Clouds Created Equal? .....	14
7	Top Cloud Use Cases .....	17
8	Profile of a Cloud Enablement Vendor: Univa UD .....	19
9	Current Statistics and Future Predictions .....	21
10	Conclusion .....	22
11	Glossary of Terms .....	23

## About the Author

Ashar Baig is an independent grid consultant with 14 years of high-tech industry experience focused on data center technologies in practice today and future trends including grid computing, green computing, storage, virtualization, multi-core optimization, high-speed interconnects and cloud computing. Ashar has hosted numerous panel discussions and BOFs at trade shows, conducted seminars/webinars/podcasts and written numerous articles & white papers on various data center technologies. Ashar has held senior roles at some of North America's leading companies such as Platform Computing, Telus, Siemens and Intel. He frequently engages with customers at various levels i.e. C-level, VP-level, director-level and end-users. Most recently, Ashar Baig was the Industry Director focused on the EDA vertical representing Platform Computing, a large Cluster and Grid Computing vendor. Ashar also managed and marketed the LSF family of products at Platform Computing. Having worked extensively throughout North America, Ashar brings with him a seasoned business perspective, broad industry background, extensive customer contacts, and solid data center understanding. Ashar holds a Bachelor of Science from Rutgers University, New Jersey and an MBA from Iona College, New York. [ashar.baig@sympatico.ca](mailto:ashar.baig@sympatico.ca)

### Copyright

Copyright 2009 Strategic Insight Consulting.

### Document redistribution policy

This document is protected by copyright and you may not redistribute or translate it into another language, in part or in whole. You may only redistribute this document internally within your organization (for example, on an intranet).

### Trademarks

Other products or services mentioned in this document are identified by the trademarks or service marks of their respective owners.

## Table of Contents

<b>1</b>	<b>EXECUTIVE SUMMARY</b> .....	<b>4</b>
1.1	WHY CLOUD? .....	4
1.2	CLOUD ORIGINS AND DRIVERS.....	4
1.3	BARRIERS TO ADOPTION .....	4
1.4	TOP USE CASES .....	5
1.5	GETTING TO THE CLOUD.....	5
1.6	VENDOR PROFILE .....	5
<b>2</b>	<b>WHY CLOUD? TOP DRIVERS FOR CLOUD COMPUTING</b> .....	<b>6</b>
<b>3</b>	<b>PUBLIC VERSUS PRIVATE CLOUDS</b> .....	<b>7</b>
3.1	PUBLIC CLOUDS.....	7
3.2	PRIVATE CLOUDS.....	8
<b>4</b>	<b>TOP BARRIERS TO CLOUD ADOPTION</b> .....	<b>9</b>
4.1	SEAMLESS TRANSITION FROM PRIVATE TO PUBLIC CLOUDS .....	9
4.2	SECURITY .....	9
4.3	APPLICATION PERFORMANCE IN THE CLOUD.....	11
4.4	MEMORY LIMITS.....	11
4.5	VISIBILITY .....	11
4.6	SOFTWARE LICENSING .....	11
4.7	LEGACY OS SUPPORT .....	12
4.8	ADEQUATE INSURANCE COVERAGE.....	12
<b>5</b>	<b>VENDOR REQUIREMENTS: WHAT TO LOOK FOR IN A CLOUD VENDOR</b> .....	<b>13</b>
<b>6</b>	<b>ARE ALL CLOUDS CREATED EQUAL?</b> .....	<b>14</b>
6.1	COMPARISON OF CLOUD COMPUTING OFFERINGS .....	14
6.2	9 THINGS TO KNOW WHEN COMPARING CLOUD VENDORS .....	16
<b>7</b>	<b>TOP CLOUD USE CASES</b> .....	<b>17</b>
7.1	INTERNAL PRIVATE CLOUD .....	17
7.2	PERMANENT HYBRID CLOUD .....	17
7.3	SEASONAL HYBRID CLOUD .....	17
7.4	EXTERNAL PUBLIC CLOUD.....	18
<b>8</b>	<b>PROFILE OF A CLOUD ENABLEMENT VENDOR: UNIVA UD</b> .....	<b>19</b>
<b>9</b>	<b>CURRENT STATISTICS AND FUTURE PREDICTIONS</b> .....	<b>21</b>
<b>10</b>	<b>CONCLUSION</b> .....	<b>22</b>
<b>11</b>	<b>GLOSSARY OF TERMS</b> .....	<b>23</b>

# 1 Executive Summary

This white paper presents an evolutionary approach to cloud computing within a high-performance computing (HPC) context. The paper addresses the seamless and incremental process of moving from private/internal to public/external clouds, the realistic use cases, and the qualities needed in a vendor to get you there. This paper also compares several public cloud offerings and profiles a vendor who is enabling a range of cloud solutions for leading Fortune 500 companies.

## 1.1 Why Cloud?

Cloud computing is many things to many people. This 'data center on the internet' can facilitate flexible computing for enterprises of every size. Think of cloud computing as a cluster of virtual servers creating essentially an Infrastructure 2.0, enabling a new level of agility and nimbleness that is unfathomable with the traditional silo computing model.

But beyond this, cloud computing represents a paradigm shift in corporate expenditures: companies are graduating to the pay-per-use model offered by cloud computing with a focus on Operational Expenditures (OpEx) instead of building in-house data centers utilizing Capital Expenditures (CapEx). The cloud computing model is enticing more and more companies to explore expanding their computing resources to meet peak demand by tapping into the on-demand capacity offered by the cloud. It is an entirely new business model with its own set of value propositions for enterprise computing environments, including application scalability, improved economies of scale, reduced costs, resource efficiencies, resource elasticity / flexibility, faster deployment times, value-based pricing model, disaster recovery and an on-demand infrastructure enabling the truly dynamic data center.

## 1.2 Cloud Origins and Drivers

Just over a decade ago, grid/cluster computing introduced us to the concept of optimal utilization of enterprise hardware resources. Today, priorities like resource flexibility, cost efficiencies, reliability, fault tolerance, power consumption and environmental footprint of the data center top the list of IT manager must-haves. These concerns have given birth to cloud computing. Static compute farms are being replaced by an on-demand, dynamically provisioned infrastructure that provides the

flexibility to burst to the public cloud time and again by tapping into additional compute capacity made available by a public cloud.

Due to the huge potential and cost-saving promise of both public and private clouds, a vast majority of enterprises are seriously considering, if not already using, cloud computing for their computing needs.

## 1.3 Barriers to Adoption

We are currently in the *early adopter* stage<sup>1</sup> of the cloud computing lifecycle. For most companies, the cloud environment presents many challenges and unknowns that must be overcome before corporations can include cloud in their IT strategy.

The prospect of transferring data and Intellectual Property (IP) out of the data center and into the cloud environment strikes fear among many data center

*"Anyone with a web-based e-mail account or a profile on a social network is already taking advantage of cloud computing, whether they know it or not."*

managers. Hence, data security and accessibility are paramount when deciding what data remains inside the corporate firewall and what data can be transferred to the cloud. Data persistency is also very critical – one has to ensure that data is really deleted when it is deleted.

Cloud computing requires data center managers to make certain compromises with respect to

<sup>1</sup> "Rogers model of adoption and innovation". Geoffrey Moore of The Chasm Group also discusses this model in his book "Crossing the Chasm". The various stages of product adoption/maturity lifecycle include 'Innovators', 'Early Adopter', 'Early Majority', 'Late Majority' and 'Laggards'. Details at: [http://www.valuebasedmanagement.net/methods\\_rovers\\_innovation\\_adoption\\_curve.html](http://www.valuebasedmanagement.net/methods_rovers_innovation_adoption_curve.html)

control, visibility, and application performance. Since application performance and throughput are of the utmost importance to end users, corporations must set the right performance expectations for the cloud environment.

Companies must conduct complete due diligence comparing internal and public clouds. These include:

1. What are the performance benchmark results in the public cloud?
2. Building a comprehensive financial model to compare OpEx and CapEx.
3. How much insurance coverage is sufficient to host your IP in the cloud environment?

Additionally, configuration management, security and compliance issues raised by today's cloud solutions must be properly addressed.

## 1.4 Top Use Cases

Corporations that currently have data centers and server farms are evaluating cloud computing for the following four practical use-case scenarios, which will be discussed in detail in Section 7 of this paper:

1. Internal Private Cloud
2. *Permanent* Hybrid Cloud
3. *Seasonal* Hybrid Cloud
4. External Public Cloud

## 1.5 Getting to the Cloud

The most important operational requirement for utilizing public/external clouds is the ability for a seamless transition from the in-house data center and/or private cloud to the public cloud. Customers want simplified setup, configuration

and tear-down of virtual resources in public clouds, just like the machines in their internal data center. The winning formula for the utilization of public clouds encompasses providing transparent utilization of dynamic compute resources and facilitating seamless transition from private clouds in the physical (or mixed physical and virtual) environments to the public cloud in the virtual environments. Only then customers can tap into the full potential and promise of cloud computing, saving time and money.

## 1.6 Vendor Profile

Companies looking to extend their compute capacity utilizing cloud computing should look for vendors who can partner with them and provide the bridge between their internal data center resources and the virtual resources within the public cloud.

Univa UD presents an evolutionary approach to cloud computing. The company has developed a set of capabilities that simplify the setup, configuration and tear-down of HPC private as well as public clouds, providing transparent utilization of dynamic and virtual environments and thus allowing customers to tap into the full potential and promise of cloud computing.

Univa has experience in real-world customer environments for all four usage scenarios in Section 1.4. Univa HPC products have also been proven to run on Amazon Elastic Compute Cloud (EC2). These products are built on a proven, robust architecture and provide peace of mind with a built-in security features developed specifically to seamlessly bridge the gap between internal private clouds, public clouds and hybrid clouds.

## 2 Why Cloud? Top Drivers for Cloud Computing

Price is by far the biggest driver for corporations electing to utilize cloud computing. Application development and operational/maintenance budgets for most companies is 50% to 70% of total IT costs. By shifting the responsibility of maintaining Service Level Agreements (SLAs) to the cloud service provider, cloud computing not only saves costs with no upfront commitment but also drives competitive advantage by increasing IT resource efficiency and capacity that can be added in a matter of minutes<sup>2</sup>.

*“Application development and operational/maintenance budgets for most companies is 50% to 70% of the total IT costs.”*

In today’s ultra-competitive business environment, Time to Market (TTM) is one of the most critical components to business success, and even short delays in a product release can dramatically reduce profitability and competitive edge. Companies with static compute resources have to consistently grapple with the tradeoffs related to under and over provisioning of in-house compute capacity. Smart companies looking to eliminate cost inefficiencies rank IT infrastructure cost very high on the priority list. Even in the best of economic times these companies have explored various resource and energy efficiency tactics like follow-the-sun<sup>3</sup> or chase-the-moon<sup>4</sup>. In these economically turbulent times, cost efficiency, business efficiency, business agility and flexibility of IT as a Service (ITaaS) business model offered by cloud computing is hard to ignore.

*“In these economically turbulent times, the IT as a Service (ITaaS) business model of cloud computing is hard to ignore.”*

Univa UD solution architecture, described in Section 8 of this whitepaper, provides this infrastructure flexibility by transparently extending the corporate data center into the cloud for seasonal computing, bursting and incremental compute capacity.

Most large companies have vested corporate equity in their enterprise data centers and server farms. The horizon of cloud computing presents an attractive cost-effective way to increase their compute capacity while reducing their environmental footprint.

In summary, Cloud computing facilitates a number of benefits that are driving its investigation and adoption by organizations of all sizes:

- Business efficiency – resource, cost and energy efficiency
- Resource flexibility, efficiency, and elasticity – rapid boost to the available compute power
- Faster Time to Market (TTM)
- Ability to scale applications
- Disaster recovery and fault tolerance
- On-demand infrastructure enabling the truly dynamic data center
- Reduced environmental footprint
- Consumption and value-based model

---

<sup>2</sup> Why buy the cow when all you need is the milk.

<sup>3</sup> To maximize compute resource efficiency throughout the enterprise.

<sup>4</sup> To maximize the energy/power efficiency throughout the enterprise.

## 3 Public versus Private Clouds

### 3.1 Public Clouds

Public compute clouds consist of compute resources that are available to end users on a subscription basis. Virtualization is the engine that runs a public compute cloud. The virtual resources in a public cloud are similar to the physical resources in a data center, with simplified setup, configuration, tear-down ease of access and admin-free resource allocation. However, due to the virtualized nature of these resources, they are not visible to end-users.

Public clouds are a shared multi-tenant infrastructure with common SLAs, immediate deployment, low switching costs and a pay-per-use costing model (the customer pays only for what services they use). Public clouds have been characterized to be in the early stages of the cloud lifecycle, but anyone with a web-based e-mail account or a profile on a social network is already taking advantage of cloud computing<sup>5</sup>, whether they know it or not<sup>6</sup>.

Popular models of server reliability in a data center environment are based on hardware and network visibility that become opaque in a cloud environment. If you don't see the data center environment the failure is invisible to you. Hence, you would not know the failure point and will not be able to pinpoint the single point of failure.

*"When the operator of the hardware upgrades or changes things – you would not have visibility into that."*

Public compute clouds have been referred to by Nicholas Carr, former executive editor of Harvard Business Review and author of the book 'Does IT Matter?' as the new Infrastructure-level power grid<sup>7</sup>, where businesses can buy raw processing power and storage capacity from IT giants like Google, Amazon, SUN, HP, and others.

Carr's 2008 book 'The Big Switch' compared cloud computing with how electricity was generated a century ago. "Computing power has become so cheap today, you can take existing hardware, servers and storage, and turn them into pure software and run it on other computers", Carr said. "This is the essence of virtualization".

"The price of computing will go way, way down and accessibility of computing will go way, way up," added Carr. "That will force companies to re-think how they build their products and connect with customers."

Eric Schmidt, CEO of Google, who was CTO of Sun Microsystems back in 1993, said "When the network becomes as fast as the processor, the computer hollows out and spreads across the network."

*"When the network becomes as fast as the processor, the computer hollows out and spreads across the network."*

"I don't think companies have realized what this is going to mean," said Carr. "Not only what they can do quickly and cheaply without having to make a big investment, but the IT department won't be the bottleneck for big computing jobs within the company."

*"IT department won't be the bottleneck for big computing jobs within the company."*

According to Marketspace Chairman Jeffrey Rayport "One reason the cloud grew is that Google and Amazon and other companies had excess capacity after building these data centers for their own core businesses – Google for search, Amazon for e-commerce – and they had these server networks that could then be used for other purposes."

<sup>5</sup> Result of a study of 2,251 adults on cloud computing released by the Pew Internet & American Life Project, which found that 69 per cent of online users have been using cloud computing in one form or another.

<sup>6</sup> By Kenneth Corbin "Study: Cloud Needs Universal, Open Networks" Internetnews.com March 2009

<sup>7</sup> IDC Directions '09 conference March 2009 in San Jose, California and Boston, Massachusetts and also in his most recent book, *The Big Switch: Rewiring the World, from Edison to Google*.

The fact remains that turning a classical data center into a dynamic data center – by combining Cluster and Cloud technology – requires a much lower investment compared to the cost of new hardware and data center coupled with ongoing maintenance costs. The cloud-enabled dynamic data center yields high ROI by improving utilization, right-sizing resource allocation and cost-effective provisioning for peak demand.

### 3.2 Private Clouds

A private HPC cloud (also known as an internal cloud or enterprise cloud) is Cluster Computing coupled with server virtualization. A private compute cloud<sup>8</sup> consists of IT infrastructure that is confined to the enterprise and is constrained from the outside world via a firewall. It is built to offer a Service Oriented Infrastructure (SOI)<sup>9</sup> to provide capacity-on-demand to corporate departments, lines-of-business, divisions, and applications, with desirable quality of service (QoS) on a firewall-constrained private network. Only corporate users within the firewall can subscribe to services such as compute-horsepower and data storage on this enterprise-wide private network.

The private compute cloud appeals to organizations that want more control over their data and have the capability to provide custom SLAs to their stakeholders.

Many large Fortune 500 companies have successfully built and deployed private clouds, but due to the cost, technical know-how and human capital required to build and maintain private clouds, this approach is not recommended for Small and Medium Sized Businesses (SMBs) on their own.

A key advantage of the private cloud is that it is more secure than the public cloud because the IT department can control access to this network and can easily pinpoint offenders. Additionally, users are aware that they can be tracked and disciplined for any security breaches. Hence, they are more restrained in over-stepping their entitlements.

*"Turning a classical data center into a dynamic data center – by combining Cluster and Cloud technology – requires a much lower investment compared to the cost of new hardware and data center coupled with on-going maintenance costs."*

While a HPC private cloud is more secure and provides better visibility, control and performance than a public cloud, companies relying on private clouds exclusively may be missing out on the business agility and benefits associated with infinite compute capacity, dynamism, flexibility and pay-per-use business benefits ascribed by public cloud computing. Additionally, the cost associated with the long-term energy and operational costs may not make economic sense for companies of all sizes.

---

<sup>8</sup> A hosted cloud is also a private cloud.

<sup>9</sup> In practice, most corporate HPC Clouds are rarely designed as a Service Oriented Infrastructure (SOI). The most common design is a Batch Oriented Infrastructure (BOI).

## 4 Top Barriers to Cloud Adoption

My numerous in-depth conversations with real-world customers, both large and small, have revealed some key concerns towards the adoption of cloud computing. According to all industry predictions, cloud is how business computing will be conducted in the future. In order for this phenomenon to become mainstream, it needs to tackle the barriers listed in this section. This is an exhaustive list that covers technical and business barriers including the most obvious (like security) and the not so obvious (like insurance coverage). Some barriers are technical while others are more subjective and/or based on perception (and in some cases misperception).

*"Cloud is how business computing will be conducted in the future."*

Section 5 of this white paper lists recommendations to minimize these barriers.

### 4.1 Seamless Transition from Private to Public Clouds

Seamless transition from the machines in the corporate data center to the machines in the public cloud environment is the most important operational requirement for utilizing public/external clouds. When extending compute capacity to a public cloud, customers are demanding same simplified setup, configuration and tear-down of the virtual compute resources within the public cloud that they are used to in their corporate data centers. To ease the migration jitters, customers require transparent utilization of dynamic compute resources and seamless transition from physical/virtual resources within private clouds to the virtual resources in public clouds.

### 4.2 Security

#### 4.2.1 User Authentication and Authorization

A major obstacle to cloud computing is user authentication and authorization. It is trivial to note that the machines in the public clouds are actually on the internet which is swarming with hackers. The consequences of un-authorized access within a cloud environment are downright scary. One has to place enough safeguards on these machines to ensure proper authentication and authorization. On the other hand, within the corporate environment one can track, pinpoint, control and manage users who try to access machines with improper credentials.

*"Machines in the public clouds are actually on the internet which is swarming with hackers."*

Additionally, due to the lack of visibility in a virtual environment, it is impossible to know who else is using the machine your jobs are running on. It is your responsibility to ensure the machines are configured properly for security and authentication.

*"I recommend that companies should use multiple levels of security, making it difficult for hackers to break."*

#### 4.2.2 Data Security

Cloud Computing moves the application software and databases to the large data centers, where the management of these corporate assets raises security concerns. Security of the corporate data and Intellectual Property (IP) in the public cloud environment is ranks very high on the concerns relating to cloud computing. Users are at the mercy of their cloud service providers for the availability and integrity of their data. Data center administrators need to ensure that there are appropriate and enough security safeguards in place to protect the privacy, integrity, and availability of their IP.

### 4.2.3 Data Transport

In order to take advantage of public clouds, corporations must transfer the data and applications from their private cloud to the public cloud in order for jobs to run successfully.

Indeed, a key obstacle to cloud computing is the security of the data while being transported to the public cloud<sup>10</sup> and while inside the cloud. Most public cloud providers, including Amazon EC2, do not guarantee such security.

Some companies do not mind transferring their corporate data over the internet; others would not even be able to obtain approval from their legal departments to do so, or are forbidden by the government to transport data over an unsecure public entity like the internet<sup>11</sup>.

Most companies have the following data transport concerns with extending their cluster into a public cloud:

1. Network – speed, latency, isolation and quality
2. Bandwidth charges for data transfer – the amount of data moving back and forth from the customer's machines to the public cloud is a huge concern since the public cloud providers charge for this data transfer
3. NFS mounting<sup>12</sup>

In practice, data transport between private cloud and public cloud is slow<sup>13</sup> – once data is in the public cloud, the data access times are fast.

Due to the infancy of public clouds, most customers exploring cloud computing are not clear on how data is transferred:

1. To the public cloud environment – if storage services such as those provided by Amazon's Simple Storage Service (S3) cloud are in Atlanta and the public cloud is in Chicago? How to move data to where the machines are physically located? How fast is the network link between Atlanta and Chicago?
2. Between public and private clouds
3. Within public clouds

### 4.2.4 Data Persistency

Another critical factor with respect to data security is data persistency, as we all remember from the issues associated with the Facebook announcement (that they own the data on their site and the data uploaded to their site will remain there forever, even when the account is deleted). Companies need to contractually ensure that their corporate data is really deleted when it is deleted.<sup>14</sup>

---

<sup>10</sup> The security of data while being transferred to the cloud can be enhanced by using both a VPN and SSH. That makes it practically impossible for a hacker to compromise this security the only way is if they get the keys – and that can only happen if the customer is sending the keys in email.

<sup>11</sup> Businesses conducting business with the US federal government, and governments in most of the western countries, are mandated to ensure that their data should not be accessed across a public system.

<sup>12</sup> Mounting NFS volumes across the Internet is done all the time, even today, However there are authentication and encryption security concerns. NFSv3 provides simple UNIX style authentication which is not sufficient for mounting filesystems across the internet. NFS v4 provides Kerberos Authentication and DES/TripleDES encryption thus it is possible to use NFSv4 over the internet. Even with NFSv4.0 some customers may not be comfortable sending their data across a NFS mount to the public cloud, in those cases creating multiple layers of security is the best approach.

<sup>13</sup> Typical data transfer speed between private and public cloud is 6 Mb/sec which is pretty fast but slow compared to disk.

<sup>14</sup> With Amazon EC2 if you do not use S3 or EBS once you shut down your virtual machines the data on those machines is gone.

### 4.3 Application Performance in the Cloud

There are technical factors limiting application performance in a public cloud environment. The most obvious performance-limiting factor is the virtualization penalty. Users do not care if their compute services are provided by internal or external resources. However, IT managers have to set the right performance expectations.

Most applications and workloads in a cloud computing environment run 30 – 35% slower than the physical compute infrastructure<sup>15</sup>. This performance toll varies depending on application type, type of workload being executed<sup>16</sup> (processor-bound, I/O bound, data-intensive, etc.), cloud provider (virtualized infrastructure or similar infrastructure to your internal data center), and suitable hardware (Intel's Nehalem processor offsets the performance penalty with significant performance gains over previous generation processors). This performance penalty can also be offset by pre-planning peak need against time value.

### 4.4 Memory Limits

Today, public cloud providers have a 15.5 GB limit on available memory on the machines within their environment. This limit is the by-product of virtualization. For customers used to running specific workload on large memory machines, this limit could pose a problem. It is up to the virtualization vendors e.g. VMware and Citrix to increase this memory limit in the near future, as end users begin to demand it.

*"Companies concerned about the performance penalty associated with virtualization should look into Server Beach, SliceHost, Rackable, etc. which are positioning themselves as an alternative to Amazon EC2 – you get a physical machine that delivers higher performance than VM, but you need to provision it yourself."*

### 4.5 Visibility

With most public cloud providers, users request Linux Virtual Machine (VM) instances that are created on the fly and billed based on actual usage. The users of cloud infrastructure always knows how many virtual machines they have, what their individual IP addresses are, and the "sizes" for each instance. However, clients don't know where the machines are located geographically<sup>17</sup>, what kind of hardware is being used or what the connectivity is. Some cite this lack of visibility as a concern, while others have no issue with it. Again, this is the by-product of virtualization and this is what makes the service cloud-like.

### 4.6 Software Licensing

There are two key issues sometimes cited with respect to software licensing in the cloud computing environment:

1. Customers cannot finalize software licensing agreements with large software vendors for use of their software in a cloud computing environment. The fact remains that many ISVs still license their software on a per-node, per-CPU or per-core basis and are still wrestling with how to license their software in a virtual environment.
2. System integration issues where various services require different information and translation.

These same issues hindered some implementations in the early days of cluster and grid computing, and will certainly need to be addressed by ISVs at some point, because end users will demand it.

---

<sup>15</sup> In para-virtualization the performance penalty is less than 10%.

<sup>16</sup> Amazon EC2 currently runs all types of workloads today from batch jobs to regression workload.

<sup>17</sup> Amazon does allow you to start a machine in an "Availability Zone" so you can start your machine in east coast, central, etc...

## 4.7 Legacy OS Support

Most public cloud providers offer support for the latest version of Red Hat Enterprise Linux (RHEL) only<sup>18</sup>. If customers require a legacy OS support they need to work with third-party vendors, like Univa UD, to obtain support for an older version of the OS e.g. RHEL 4.0 in an AMI on Amazon EC2.

## 4.8 Adequate Insurance Coverage

An important business factor that concerns most IT executives exploring public clouds is the amount of insurance that may be necessary on the corporate data in the public cloud environment<sup>19</sup>. For large organizations these calculations may involve complex financial models and for others, mostly smaller organizations, the insurance limit criteria is based purely on customer's own gut-feel.

---

<sup>18</sup> Amazon EC2 officially only supports 5.1+.

<sup>19</sup> Today, the insurance companies have no idea how to value this risk.

## 5 Vendor Requirements: What to Look For In a Cloud Vendor

Section 4 of this white paper outlined a list of concerns identified during my conversations with current and potential users of cloud computing. Partnering with the right vendor can help you navigate through the obstacles and barriers to cloud computing.

The option to utilize public clouds is not a black and white choice. It is very much a case-by-case evaluation based on application, tradeoffs and opportunity cost.

*"The option to utilize public clouds is not a black and white choice. It is very much a case-by-case evaluation based on application, tradeoffs and opportunity cost."*

Companies looking to tap into the full potential and promise of cloud computing should partner with a trusted vendor who can devise a solution that meets their specific needs and who possess the following attributes and capabilities:

1. Have a good solution for HPC private clouds and can also provide seamless transition from there to public clouds.
2. Allows them to manage their compute capacity and facilitates similar setup, configuration and tear-down of their compute resources in both private and public HPC cloud environments.
3. Provide iron-clad user authentication and authorization in public clouds.
4. Guarantee the security of data while being transported to the public cloud or while inside the cloud.
5. Provide a complete cloud computing software stack (see Figure 1 on the next page).
6. Have experience deploying public clouds and have best practices that they can share with you of their current cloud deployments.
7. Have conducted extensive security audits for real-world customers.
8. Can provide ironclad guarantees in the contract and or insurance about the persistency of your corporate data within the public cloud environment.
9. Have an enterprise architecture that can address any system integration concerns.
10. Can provide support for older versions of the OS in the public cloud environment.
11. Provide solutions that are open-stands based and portable to prevent vendor lock-in, ideally supported by the open source community<sup>20</sup>.
12. Are there when you need them – you should be absolutely clear on whom to call when there is a problem<sup>21</sup>.
13. Is not a one-trick pony: Doesn't do only one thing, like support for just private HPC clouds. Customers should have the choice to decide which type of cloud solution is right for them: private, public or hybrid.

I will highly recommend that customers should work with their public HPC cloud provider to improve the contractual language that reflects their corporate interest.

*"Demand an accurate and detailed audit report from the cloud provider."*

I was impressed by the product and services offerings of Univa UD (described in Section 8). The company has experience in real-world environments and deployments for most<sup>22</sup> of the above scenarios.

*"Customers of the cloud must have the access to define and modify their own security policies and controls."*

<sup>20</sup> Clayton M. Christensen in his book "The Innovator's Dilemma" details how a standards-based approach will win in a mature market.

<sup>21</sup> Many cloud vendors now offer online services and professional services.

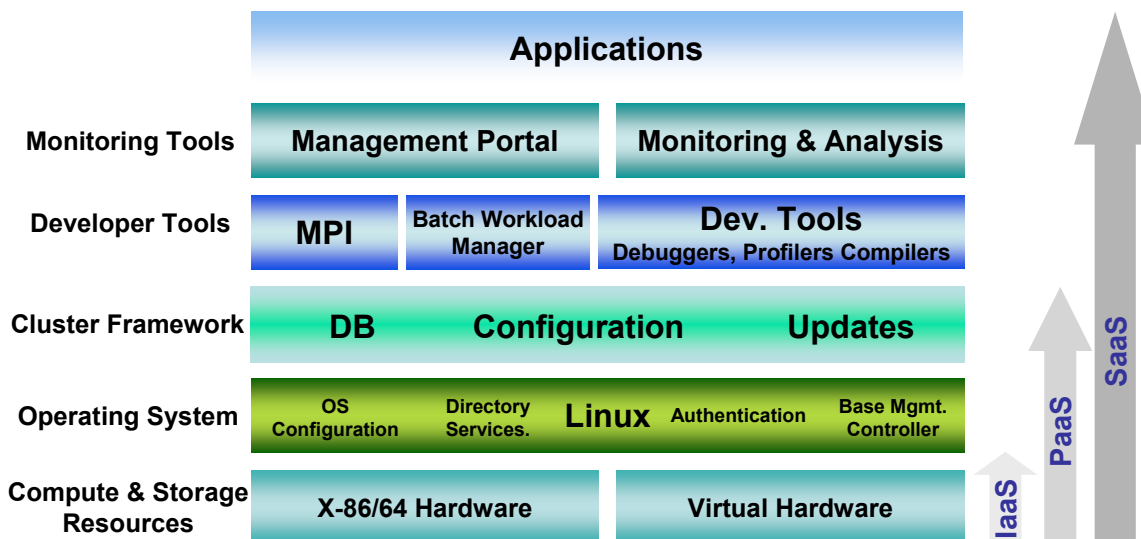
<sup>22</sup> Univa does not provide insurance on your data on Amazon EC2.

## 6 Are All Clouds Created Equal?

### 6.1 Comparison of Cloud Computing Offerings

Cloud computing encompasses many concepts like Application Service Provider (ASP), Infrastructure as a Service (IaaS), Integration as a Service (IaaS), Platform as a Service (PaaS), Hardware as a Service (HaaS) and Software as a Service (SaaS).

**Figure 1: Cloud Computing Software Stack**



Following the recent cloud rush there has been a myriad of providers offering public cloud infrastructure services including Microsoft's Azure Services Platform<sup>23</sup>, Google File System (GFS)<sup>24</sup>, Google MapReduce<sup>25</sup>, Google BigTable<sup>26</sup>, Amazon Elastic Compute Cloud (EC2)<sup>27</sup>, Amazon's Simple Storage Service (S3)<sup>28</sup>, Sun Cloud<sup>29</sup>, IBM's Cloud services<sup>30</sup>, Salesforce.com<sup>31</sup>, GoGrid<sup>32</sup>, Flexiscale<sup>33</sup>,

<sup>23</sup> PaaS offering that provides a cloud computing and services platform hosted in Microsoft data centers. The Azure Services Platform provides a range of functionality to build applications that span from consumer web to enterprise scenarios and includes a cloud operating system and a set of developer services.

<sup>24</sup> Google File System (GFS) is a distributed file system developed by Google for its own use. It is also a PaaS offering designed to provide efficient, reliable access to data using large clusters of commodity hardware. [http://en.wikipedia.org/wiki/Google\\_File\\_System](http://en.wikipedia.org/wiki/Google_File_System)

<sup>25</sup> MapReduce is a software framework introduced by Google to support distributed computing on large data sets on clusters of computers. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

<sup>26</sup> BigTable is a compressed, high performance, and proprietary database/storage system built on Google File System (GFS), for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers.

<sup>27</sup> EC2 is a HaaS offering that provides raw compute power for Linux-based clouds. <http://aws.amazon.com/ec2/>

<sup>28</sup> S3 is also a HaaS offering that provides raw storage capacity in the cloud. Amazon initially built this data storage infrastructure to run its own global network of web sites. <http://aws.amazon.com/s3/>

<sup>29</sup> Sun Cloud, a public compute and storage cloud, which is due out later this year. Sun's cloud services will be able to import Amazon cloud services to allow customers to move from Amazon to Sun. <http://www.sun.com/solutions/cloudcomputing/index.jsp>

Rackspace<sup>34</sup>, Nirvanix<sup>35</sup>, Hosting.com<sup>36</sup>, Server Beach<sup>37</sup>, RightScale<sup>38</sup>, Rackable<sup>39</sup>, vmware<sup>40</sup>, etc. providing various cloud or cloud-related offerings.

I predict that due to its enormous resource infrastructure and the potential business benefit, eBay will also announce its cloud offerings later this year.

The above is by no means an exhaustive list. It lists companies that the customers I engage with frequently have been talking to or are interested in talking to. Today, there is a proliferation of cloud related tools and services available which certainly helps consumers and prospective buyers. I predict that in the near future we will see industry consolidation with larger companies absorbing smaller ones.

The most important thing that customers should be looking for is how far deep into the Cloud Computing Software Stack” (Figure 1) does the vendor go and do they have the expertise in-house to fill the gaps. The business model that has worked for the most successful customers is to shred everything that is not strategic and get someone else to do it for you.

Storage services such as those provided by Amazon’s S3, GoGrid, Flexiscale, Nirvanix, etc. allow organizations to store data and documents without adding even a single on-site server.

Cloud-related technologies like Hadoop<sup>41</sup> could be important in supporting the growing importance of data-intensive science in particle physics, biology, earth/environmental science among other disciplines. Workflows will evolve to support the data intensive model of Hadoop and Dryad and access both classic clouds and MPI engines.

---

<sup>30</sup> Launched in October of 2008, this combination of software, services and technical resources extends IBM’s traditional software delivery model toward a mix of on-premise and cloud computing applications.

<sup>31</sup> SaaS/PaaS offering from Salesforce.com, Force.com is a PaaS offering that delivers CRM services, so clients can manage their customer information without installing specialized software.

<sup>32</sup> IaaS offering that provides compute horsepower and storage capacity for Windows-based clouds. <http://www.gogrid.com/>

<sup>33</sup> HaaS offering that provides raw compute power and storage capacity in the cloud. <http://flexiscale.com/>

<sup>34</sup> Rackspace (Mosso, The Rackspace Cloud) provides on-demand scalable website, application and storage hosting as well as raw compute capacity. <http://www.mosso.com/> Rackspace Hosting, Inc. recently **acquired SliceHost** another HaaS provider.

<sup>35</sup> Cloud storage platform provider. <http://www.nirvanix.com/>

<sup>36</sup> Hosting.com’s CloudNine, a cloud hosting platform is a PaaS/IaaS offering that provides server collocation, managed hosting for cloud computing environments. <http://www.hosting.com/>

<sup>37</sup> ServerBeach delivers powerful dedicated servers Web hosting on Linux and Windows for cloud computing environments. <http://www.serverbeach.com/>

<sup>38</sup> RightScale is a web based cloud computing **management platform** for managing cloud infrastructure from multiple providers. <http://www.rightscale.com/>

<sup>39</sup> Offers products to HaaS/IaaS providers tailored for the Cloud Computing environments as well as dedicated servers. <http://www.rackable.com/>

<sup>40</sup> Vmware recently introduced vSphere 4.0, a cloud Operating System, Details at [www.vmware.com/](http://www.vmware.com/)

<sup>41</sup> Apache Hadoop is a free Java software framework that supports data intensive distributed applications. It enables applications to work with thousands of nodes and petabytes of data. Hadoop was inspired by Google’s MapReduce and Google File System (GFS) papers. Hadoop is a top level Apache project, being built and used by a community of contributors from all over the world. Yahoo! has been the largest contributor to the project and uses Hadoop extensively in its Web Search and Advertising businesses. IBM and Google have announced a major initiative to use Hadoop to support University courses in Distributed Computer Programming. Hadoop was created by Doug Cutting (now a Yahoo! employee), who named it after his child’s stuffed elephant. It was originally developed to support distribution for the Nutch search engine project.

## 6.2 9 Things to Know When Comparing Cloud Vendors

1. Most of the HaaS and IaaS providers offer four nines (99.99%) of SLAs on uptime<sup>42</sup>. Due to customer demand many cloud providers are seriously considering providing “five nines” SLAs later this year.
2. Amazon EC2 charges for incoming bandwidth, whereas GoGrid does not.
3. Amazon uses Xen virtualization.
4. Amazon officially only supports RHEL 5.1+.
5. Amazon charges \$.10/VM per hour for compute capacity \$.15/Gb per month for data storage<sup>43</sup>.
6. Amazon EC2 provides its own firewall and networking configuration. Standard HPC networking configurations do not translate well to EC2.
7. Google App Engine datastore has to be BigTable format which is quite different from the relational database format.
8. Typical admin to user ratio in an enterprise environment is 1 to 100. Conversely, the typical admin to user ratio in the public cloud environment is 1 to 20,000<sup>44</sup>.
9. Some vendors provide you with a single bill for all your cloud computing services instead of separate bills from various cloud vendors. This centralized/unified billing model can result in economies of scale that lowers per-unit costs.

---

<sup>42</sup> In 2009, Amazon upped their SLA to 99.95% from 99.9%. Four nines (99.99%) availability translates into 52 minutes of downtime per year.

<sup>43</sup> These are list prices and do not factor in volume discounts or any special pricing.

<sup>44</sup> Centralized delivery of applications gives service providers the cost advantage.

## 7 Top Cloud Use Cases

The most common applications deployed in the public clouds today are Test and Development, Batch, and Web Services. Additionally, the public cloud is also used for seasonal computing, bursting and storage. The following are typical use cases of cloud computing. Most users fall into one of the following four categories.

### 7.1 Internal Private Cloud

*Definition: Enterprise-wide mixed physical<sup>45</sup> and virtual environment for on-demand compute capacity.*

With internal private clouds, companies can consolidate their computing by running virtual machines (VMs) on permanent hardware to get more virtual machines on finite physical machines. Incremental compute capacity can be obtained by virtue of virtualization and by borrowing from other groups or departments. The virtual machines allow for additional compute capacity on demand, extending the current physical environment as needed to handle spikes in workloads.

Previously, one could run 2 to 3 VMs, with decent performance, from the x86 hardware. Intel's Nehalem architecture accelerates the consolidation process by facilitating more than double the number of VMs.

*"Intel's Nehalem architecture accelerates the consolidation process by facilitating more than double the number of VMs."*

Ideally, companies are looking for automated or relatively quick (less than 5 minutes) setup, management and tear down of the virtual environment for incremental compute capacity.

### 7.2 Permanent Hybrid Cloud

*Definition: Enterprise-wide private cloud with the ability to burst to the public cloud for compute capacity.*

All nodes in the permanent hybrid cloud run planned jobs as part of a project or process (e.g. running Quality Assurance (QA) tests, building a temporary virtual development/test cluster, running weekly tests, software regression or benchmarking). In these cases, customers rely on the cloud for incremental compute capacity rather than going through the hassle of purchasing additional compute horsepower and going through setup, management and teardown at the end of the project.

*"In the near future, most QA departments will have one machine that they log on to and then will run all their tests on the machines in the public clouds."*

Because of the cost efficiency and flexibility of public clouds, I envision that in the near future most QA departments will have one machine that they log onto and then will run all their tests on the machines in the public clouds.

### 7.3 Seasonal/Hybrid Cloud

*Definition: Enterprise-wide private cloud with the ability to burst to the public cloud for a short period of time for seasonal or incremental compute capacity (e.g. movie rendering, customer/partner/employee training, odd combination of hardware/OS/software, building a temporary virtual development/test cluster, etc. periodically e.g. once a quarter or once a month) to seasonally create that environment for a set of users for a limited duration.*

---

<sup>45</sup> Physical machines are required in a private HPC cloud because some applications do not run well on virtual hardware.

This scenario is particularly difficult to accomplish in a physical environment due to time and effort required. If at the very last minute this seasonal demand is not required due to a change in priorities, these additional machines in the public cloud can be shut down without incurring additional charges for those machines.

## **7.4 External Public Cloud**

*Definition: The typical public perception of “cloud computing” where all compute capacity resides in the public cloud*

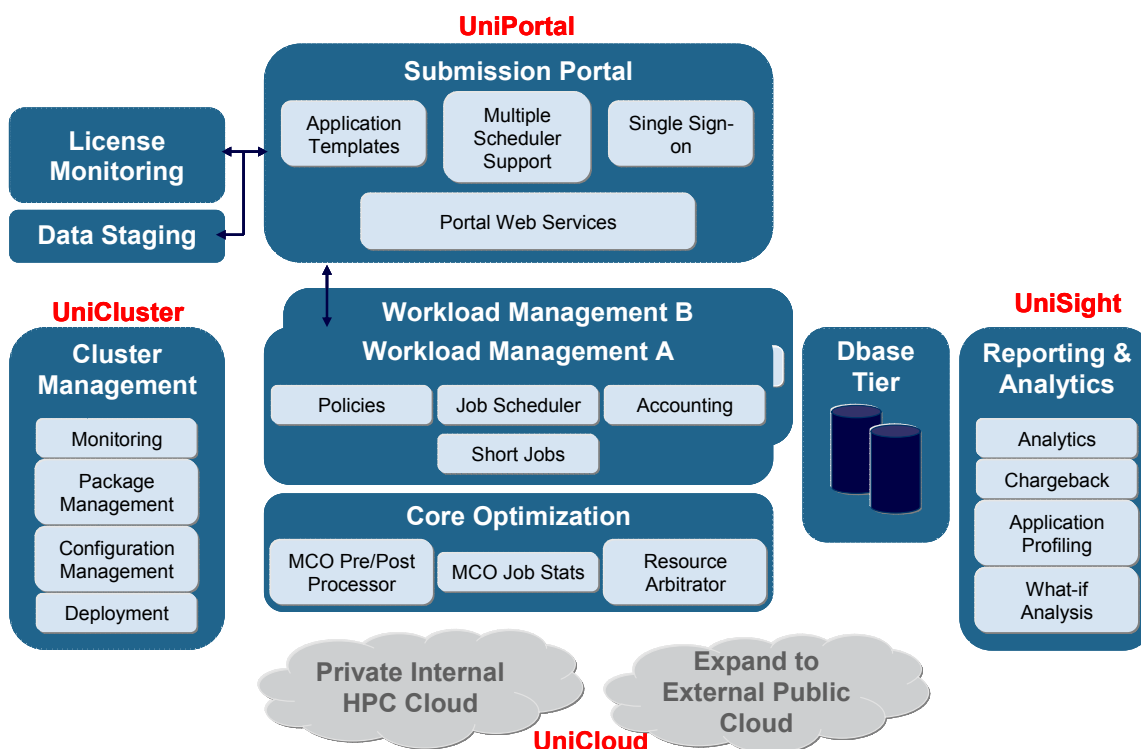
In an external public cloud, compute capacity is available to customers on a subscription basis. The machines in the external public cloud are virtual and hence are not visible to the end users. The challenges associated with virtualization are performance degradation, lack of visibility – impairing IT troubleshooting, memory limits and loose security standards. However, in most cases customers are willing to overlook these obstacles by partnering with a vendor who can help them mitigate these barriers in the interest of business agility, cost savings and resource flexibility.

## 8 Profile of a Cloud Enablement Vendor: Univa UD

Univa UD provides application and infrastructure enablement software for dynamic IT environments. The company's products (Figure 2 below) are designed to work in the following four types (within a specific type and any combination of these four types) of private and public HPC cloud scenarios:

1. In-house physical nodes (Cluster Computing)
2. In-house physical and virtual nodes (Private HPC Cloud)
3. Hosted physical nodes (Public or private HPC Cloud<sup>46</sup>)
4. Hosted virtual nodes (Public HPC Cloud)

**Figure 2: Univa UD HPC Solution Architecture**



Univa supports HPC cloud computing via UniCloud, an extension to Univa's UniCluster and Grid MP products, that provides policy-driven automated dynamic provisioning and business load leveling capabilities. It enables transparency between private and public clouds (such as Amazon EC2) with respect to capacity management and secures the data 'in flight' between the customer and the public cloud by creating a VPN from the customer's network to the public cloud provider. Once the data is in the cloud an encrypted filesystem on top of an EBS<sup>47</sup> device ensures that the customer's data is encrypted within the public cloud.

<sup>46</sup> If a company has a private dedicated link to the hosted nodes it is a private cloud. If a company has to go over the internet to access the hosted nodes, then it is a public cloud.

<sup>47</sup> Amazon Elastic Block Storage device. A virtual storage device that you can format then mount on a machine.

Univa's infrastructure management products and contention resolution capabilities enable companies to establish internal cloud environments. Such an environment includes resource management and dynamic allocation, chargeback and support for virtualization.

Univa's products are capable of provisioning nodes in the cloud using the same tools, services and framework as it would for creating local physical or virtual hosts.

They can provide support for older versions of the OS in the public cloud environment e.g. RHEL 4.0 in an AMI on Amazon EC2 or any other public cloud environment.

An excellent customer benefit that Univa UD provides is that, due to the open source nature of their products, they provide customers an opportunity to learn and test their applications and tools on the cloud for as long as is deemed necessary by the customer, without incurring additional license costs on the stack. Not many grid and cloud computing vendors can claim to offer this business benefit.

Additionally, Univa provides the tangible business benefit of unified billing. Customers receive a single bill from Univa instead of separate bills from the public cloud provider as well as other vendors in the ecosystem. This business model benefits the customer because Univa can provide you with much lower costs as a result of their bulk purchase from the public cloud provider, like Amazon, achieving significant economies of scale and passing the saving on to you.

Univa has real-world experience building and deploying internal and external HPC clouds for Fortune 500 companies for all four types of private and public cloud scenarios listed above and have conducted security audits for large global customers.

More details on Univa UD can be found at:  
<http://www.univaud.com>

*"... due to the open source nature of Univa UD's products, Univa provides their customers an opportunity to learn and test their applications and tools on the cloud for as long as is deemed necessary by the customer, without incurring additional license costs..."*

*"Univa UD's customers receive a single bill from Univa instead of separate bills from the public cloud provider as well as other vendors."*

## 9 Current Statistics and Future Predictions

According to all industry predictions, cloud computing is how business computing will be conducted in the future. This technology is poised to become ubiquitous and cheap quickly. Over the longer term most businesses will adopt the hybrid model of a mix of public and private clouds. Analysts are already claiming that cloud computing is anywhere from a **\$40 billion to \$80 billion per year business**.

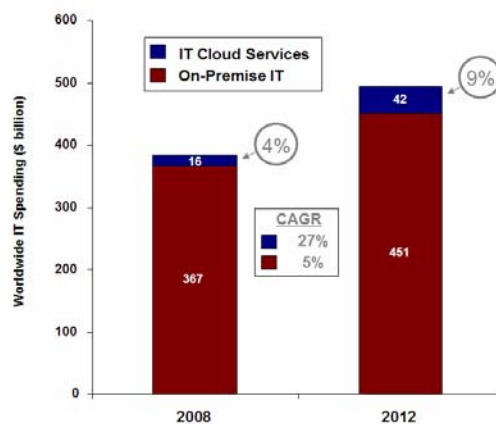
*"Over the longer term most businesses will adopt the hybrid model of a mix of public and private clouds."*

A few choice findings and predictions are presented below:

- **Amazon** claims to have 500,000 paying customers for its cloud infrastructure services.
- **Goldman Sachs'** recent survey of IT spending intent predicted that spending will fall by 1% in 2009 compared with a 6% growth in 2008. This reduction in CapEx spending spells a huge opportunity for public clouds which rely on OpEx spending.
- **IDC** also predicts that of the \$383 billion that customers will spend this year on IT services, \$16.2 billion - or 4% - will be consumed as cloud services.
- Over the next five years, **IDC** expects **spending on IT cloud services to grow almost threefold**, reaching \$42 billion by 2012 and accounting for 9% of customer spending. More importantly, spending on cloud computing will accelerate throughout the forecast period, capturing 25% of IT spending growth in 2012 and **nearly a third of growth the following year**.
- In his presentation on April 21<sup>st</sup> 2009 at CloudSlam09, *The Sky's the Limit; How cloud computing is changing the rules?*, William Fellows, analyst from **The 451 Group** shared that there were 60% more users of public clouds at the end of 2008 than there were at the beginning of 2008.
- According to **IDC**, at 27% the Cloud Computing CAGR is over **five times the growth rate** of the traditional, on-premise IT delivery/consumption model.

**Figure 3: Worldwide IT Spending**

**Worldwide IT Spending\* by Consumption Model  
2008, 2012**



\* Includes enterprise IT spending on Business Applications, Systems Infrastructure Software, Application Development & Deployment Software, Servers and Storage

Source: IDC, October 2008

## 10 Conclusion

In today's turbulent economy, the pay-for-only-what-you-use model of cloud computing is very attractive for most companies whose IT budgets are constantly under the microscope. Cost advantages associated with foregoing CapEx make cloud technology very attractive for many organizations. However, issues such as privacy, the regulatory environment, performance (What is the throughput for a cloud?), latency (What is the latency between nodes?), security and visibility (what's happening "behind the scenes" in the public cloud?) and liability must be tackled before choosing public clouds.

Cloud computing allows organizations to deliver highly scalable applications quickly and cost effectively. It is especially advantageous for companies that have bursty or seasonal demand for compute capacity. Cloud computing enables these companies to be more agile in delivering products rapidly to their customers. These companies are not constrained for scale and can dramatically reduce the complexity of their IT infrastructure, achieving greater ROI. Such companies are finding the on-demand computing model, whereby compute capacity is mapped directly to demand and they pay only for what they use, exponentially more attractive compared to maintaining often under-utilized compute capacity in the form of data centers and more hardware.

Cloud computing is a game changer technology that offers enterprises a new level of scalability, agility, automation and a much reduced TCO. But without a cloud platform that provides transparency between private and public clouds, as well as an experienced partner, enterprises cannot reap the full benefit of cloud computing. The right cloud-enabled platform can bridge the gap between physical and virtual nodes in public as well as private cloud, allowing users to take full advantage of the scalability, agility and reliability benefits of cloud.

Opponents of cloud computing may use ownership and control to rationalize avoiding cloud, but the fact is that factors like risk of changing, current vendor limitation, and job security are usually the real reasons behind refusal to explore cloud computing.

A private cloud alone lacks the scale, automation, ease of access, admin-free resource allocation, and self-service properties that have become the trademark of public clouds.

Some traditional HPC or grid computing vendors will have you believe that "private clouds" are exclusively the way to go, but the private cloud approach allows you only to share in-house resources and usually does not allow for running multiple applications simultaneously or providing an easily accessible, on-demand dynamic environment at a small fraction of the cost. The reality remains that in today's economic turbulence companies looking to "build" private clouds have to purchase, deploy and manage the private clouds, which requires investment in form of capital expenditures that could be better spent on business innovation and hence can be deemed fiscally irresponsible in this day and age.

In spite of the advantages of private HPC clouds, companies who ignore public HPC clouds altogether will be at a competitive disadvantage for not leveraging the flexibility and dynamism of public clouds. What makes better sense is a combined approach that employs both internal and external HPC cloud approaches to attain flexible capacity by optimal utilization of the private cloud while tapping into public clouds for incremental capacity, thus ensuring that any new capital expenditures are spent towards driving business innovation, continuity, and driving competitive advantage.

## 11 Glossary of Terms

ASP	Application Service Provider
BOI	Batch Oriented Infrastructure
COTS	Common off-the-shelf
CPU	Central Processing Unit
EBS	(Amazon's) Elastic Block Storage
EC2	Elastic Compute Cloud
HaaS	Hardware as a Service
HPC	High Performance Computing
DRM	Distributed Resource Manager
IaaS	Infrastructure as a Service
IaaS	Integration as a Service
ISV	Independent Software Vendor
ITaaS	IT as a Service
OS	Operating System
OSS	Open Source Software
PaaS	Platform as a Service
QoS	Quality of Service
RHEL	Redhat Enterprise Linux
ROI	Return on Investment
SaaS	Software as a Service
SLA	Service Level Agreements
SMBs	Small and Medium Sized Businesses
SOA	Service Oriented Architecture
SOI	Service Oriented Infrastructure
TCO	Total Cost of Ownership
TTR	Time to Result
VM	Virtual Machine
VPN	Virtual Private Network
WLM	Workload Management