High Availability Based Migration Analysis to Cloud Computing for High Growth Businesses

Dilip K. Prasad *dilipprasad@gmail.com*
School of Computer Engineering
Nanyang Technological University
Singapore, 639798

Abstract

High availability requirement of the network is becoming essential for high growth disruptive technology companies. For businesses which require migration to networks supporting scalability and high availability, it is important to analyze the various factors and the cost effectiveness for choosing the optimal solution for them. The current work considers this important problem and presents an analysis of the important factors influencing the decision. The high availability of network is discussed using internal and external risk factors of the network. A production network risk matrix is proposed and a scheme to compute the overall risk is presented. A case study is presented in which four possible network configurations are analyzed and the most suitable solution is recognized. This study provides a paradigm and a useful framework for analyzing cloud computing services.

Keywords: Cloud Computing, High Availability, Distributed Systems, Network Risk Matrix.

1 : INTRODUCTION

With the globalization of businesses more and more businesses are dependent on the internet for sales, marketing and their day to day activities. More and more companies are now moving towards software-as-a-services (SaaS) business model which is driven by four key goals: profitability, cash flow, growth and market share. Companies like Facebook, Amazon, Visa, Groupon, etc., can't afford their networks to be down for even a fraction of second. Demand for 99.99999…% uptime is increasing. Further, business computing requires an ever-increasing number of resources in order to deliver the results within a reasonable time frame for ever-growing problem sizes. In the world of disruptive companies like Facebook, Foursquare, Salesforce, etc., the next day requirement of which cannot be predicted in advance, if such companies depend on self-hosting servers then they will not be able to grow at rapid rate due to the inability to support its network load as demanded by its customers. Thus, affordable and quickly employable solutions providing scalability and high availability are crucial to such companies. In the last decade, while big companies like IBM were able to afford (the access to) expensive clusters and grids, many small scale high growth businesses were forced to opt for cheaper resources such as cloud computing servers.

Some of the important features of cloud based solutions are multi-tenancy, rapid deployment, pay-as-you-go model, high flexibility, On-demand solution, automatic upgrade and lower total cost of ownership. Cloud computing presents an alternative in which resources are no longer hosted by the companies' own hosting facilities, but are leased from big data centers only when needed. Despite the existence of several cloud computing offerings (products and services) by vendors such as Amazon [1] and GoGrid [2], the performance analysis of the production network availability based migration to clouds service providers for businesses remains largely unexplored. To address this issue, this paper presents a performance analysis of the production network availability goals based cost effective migration to the cloud computing services for high growth businesses from the legacy network system such as in-house hosting and maintenance.

The cloud computing paradigm holds great promise for the performance-hungry business computing community. Cloud servers can be a comparatively cheaper alternative to supercomputers and specialized clusters, while still being a much more reliable platform than the grids, and a far more scalable platform than the largest of the commodity clusters. Cloud based services also promise to "scale up by credit card," that is, to scale up instantly and temporarily within the limitations imposed only by the available financial resources, as opposed to the physical limitations of adding nodes to the clusters or supercomputers and to the administrative burden of over provisioning resources. Further, clouds provides good support for bags-of-tasks (BoTs), which are currently the most dominant application of grids [3]. However, clouds also raise important challenges in several aspects of business need, which include cost, performance, fast scalability, and high availability. These three aspects, the cost, performance, and high availability, are the focus of this work.

Here, we highlight that there are three main differences between business computing workloads and the scientific computing workloads of the clouds: 1. required system size, 2. performance demand, 3. job execution model. An example of the conflicting size requirements is that the top scientific computing facilities comprise of very large systems, with the top ten entries in the Top500 supercomputers list having a total of about one million cores, while cloud computing services are designed to replace the data centers of small-to-medium size enterprises. Performance wise, scientific workloads often require high performance computing (HPC) or high throughput computing (HTC) capabilities, many tasks computing (MTC) [4] (which is actually HPC of loosely coupled applications with possibly interrelated tasks. On the other hand, the focus of the business computing community is on high availability and throughput, and not so much on HPC or HTC. The job execution model of the scientific computing platforms is based on exclusive, space-shared usage of resources. In contrast, most business oriented clouds time-share their resources and use virtualization to abstract away from the actual hardware, thus increasing the concurrency of users but potentially lowering the attainable performance.

The clouds that support the workloads on fast growing businesses raise an important research question: Is the performance of the cloud sufficient for the computing requirements of the high growth business?, or, in other words, Can the current cloud execute the business computing workloads with similar performance (that is, for traditional performance metrics [5]) and at lower cost? What factors should be taken into the consideration while deciding to migrate to the cloud servers? Though there were early attempts to characterize clouds and other virtualized services [6], [7], [8], [9], [10], this particular question remains largely unexplored.

Much work has been put into the evaluation of novel supercomputers [11], [12], [13], [14], [15], [16] and nontraditional systems [17], [18], [19], [15], [20] for scientific and business computing. There has been a recent spur of research activity in assessing the performance of virtualized resources, in the cloud computing environments [7], [8], [9], [21], [22] and in other traditional networks [6], [23], [24], [25], [26], [27], [28]. Performance studies using general purpose benchmarks have shown that the overhead incurred by the virtualization can be below 5 percent for computation [23], [24] and below 15 percent for networking [23], [25]. Similarly, the performance loss due to the virtualization for parallel I/O and web server I/O has been shown to be below 30 percent [29] and 10 percent [30], [31], respectively. Recently, much interest for the use of virtualization has been shown by the HPC community, spurred by two seminal studies [6], [32] that find virtualization overhead to be negligible for computation intensive HPC kernels and applications such as the network attached storage (NAS) and NAS parallel benchmark (NPB) benchmarks. Other studies have investigated the performance of virtualization for specific HPC application domains [28], [33], or for mixtures of Web and HPC workloads running on virtualized (shared) resources [34].

A comparison of the performance and cost tradeoffs between a cloud and a grid was presented in [7]. A particular process workflow, Montage astronomical image mosaic application, was considered and its application client was located remotely from the HPC scientific community. However, in our opinion, the utility of [7] in the context of the current work is limited because it

considers that only a single unit on the cloud is available, and thus does not address the issue of scalability which is central to the theme of the current work. Another important study on the comparison of the performance of clouds was presented in the seminal work presented in [8], [35], which includes a performance evaluation of file transfers between Amazon EC2 and S3. Several small-scale performance studies of Amazon EC2 have been recently conducted, which include. The study of the performance of the Amazon EC2 using the NPB benchmark suite [9] and selected HPC benchmarks [36], the early comparative study of Eucalyptus and EC2 performance [21], , etc. An early comparative study of the Dawning Cloud and several operational models [10] extends the comparison method employed for Eucalyptus [21], but uses job emulation instead of job execution.

In contrast to body of previous works aforementioned, ours is different in the scope: we perform critical analysis of production network availability, its cost and performance of using general purpose and high-performance computing to compare several clouds. Further, in most of the previously mentioned works, the evaluation is based upon one task or process. On the other hand, we study the actual network load and availability of network based analysis for really rapidly growing businesses (like Facebook, Google, Foursquare, etc.) whose network load increases each day in rapid rate. Our performance evaluation results are based on the availability of network and the analysis of the cost of migration to the cloud. It also gives more insights into the performance of other clouds. The main contribution of the present work is threefold:

1. A case study is used to demonstrate the cost analysis of hosting company owned servers vs. the cost of using virtualization and cloud computing based services for scalability.

2. The analysis of the cost of migration based on the criterion of high availability is studied for three different types of commercial cloud computing services suitable for high growth businesses.

3. The production network risk matrix is characterized based on the hardware, software, security, and backup issues related to the network stability and failure.

The remainder of the article is organized as follows: In Section 2, we give a general introduction to the use of cloud computing services for business computing, and selected three networks including clouds for use in our investigation. In Section 3, the production network availability goals for high growth business are discussed. In the same section, the production network risk matrix is categorized based on the software, hardware, security, backup issues, and risk involved with the network. In Section 4, we consider a case study for performing the cost analysis of the traditional network system with three commercial networks including clouds based on high availability. In Section 5, we compare the cost performance of the three clouds and the current configurations (self-hosted network attached storage system of the company) of the business computing environments. Lastly we present our conclusions and potential future research topics in Section 6.

2 : CLOUD COMPUTING SERVICES FOR BUSINES COMPUTING

In this section, a background for analyzing the performance of cloud computing services for business that require high availability has been provided.

We identify three categories of cloud computing services [37], [38]: Infrastructure-as-a-Service (IaaS), i.e., raw infrastructure and associated middleware, Platform-as-a-Service (PaaS), i.e., application programming interfaces (APIs) for developing applications on an abstract platform, and Software-as-a-Service (SaaS), i.e., support for running software services remotely. Many clouds already exist, but not all provide virtualization, or even computing services. Thus, in this study we focus only on IaaS providers. Further, we limit the study to only public clouds, i.e., clouds that are not restricted within an enterprise; such clouds can be used by our target audience, high growth businesses.

Based on the recent survey of the cloud computing providers [39], we have selected three IaaS clouds for this work. The reason for this selection is threefold. First, not all the clouds on the market are still accepting clients. For example, FlexiScale puts new customers on a waiting list for over two weeks due to system overload. Second, not all the clouds on the market are large enough to accommodate requests for even 16 or 32 co-allocated resources. Third, our selection already covers a wide range of quantitative and qualitative cloud characteristics, as summarized in cloud survey [39]. The three categorically different networks including virtualization, dedicated servers and public clouds system with high availability options are selected: Virtual Iron [40], GoGrid [2], and HostMySite [41].

3 : PRODUCTION NETWORK AVAILABILITY GOALS

3.1 Understanding Availability

To understand the requirement of high availability of a production network, it is required to understand the goals of the production network. What does availability mean for a network engineer of a company? While calculating the availability, the network engineer assumes that the time required for the following is zero: 1. Scheduled maintenance downtime of the hosting place servers, 2. the time spent in confirming an outage, 3. the time taken for resolving the issues (before they are reported by a client). While most networks have fixed scheduled downtime of the servers, which is a known and planned aspect of the network requirements, the other two are not planned and are the causes of unscheduled down time of the network.

Following the above assumption, the availability of 99% means that for the remaining time, the network is available for 99% of the time and the permissible maximum unscheduled downtime is 1% (3.65 days of a year or 14.4 minutes of a day). We present some examples of the network availability goals and the corresponding permissible unscheduled downtimes in TABLE 1.

Availability goal	Permissible Downtime/Year	Permissible Downtime/Week	Permissible Downtime/Day
90%	36.5 days	16.8 hours	2.4 hours
95%	18.25 days	8.4 hours	1.2 hours
98%	7.3 days	3.36 hours	28.8 minutes
99%	3.65 days	1.68 hours	14.4 minutes
99.5%	43.92 hours	50.4 minutes	7.2 minutes
99.8%	17.52 hours	20.16 minutes	2.9 minutes
99.9%	8.76 hours	10.1 minutes	1.4 minutes
99.95%	4.38 hours	5.04 minutes	42.3 seconds
99.99%	52.26 minutes	1.01 minutes	8.7 seconds
99.999%	5.26 minutes	6.05 seconds	0.86 seconds
99.9999%	31.5 seconds	0.605 seconds	0.086 seconds

TABLE 1: Examples of network availability goals and permissible down times corresponding to the goals.

It is reasonable to expect that as the availability goal of the network increases, the cost and complexity of the network management increases exponentially. And in order to increase the availability by a very small percentage, the costs, tremendous amount of money needs to be invested for the required service enhancement. Generally, 99.999% availability is reserved for emergency services (like 911 services) or for large financial organizations (like electronic trading sites) that require extreme levels of availability. This level of uptime may not be realistic, necessary, or cost effective for most companies to maintain. Since a company/business cannot eliminate all single points of failure (SPOF), risks must be identified, prioritized, and addressed in the order of their criticality for the business.

3.2 Risk Identification

The first step to managing risk is identifying what can fail, determining what effect the failure will have on the organization's ability to conduct business, the probability that the failure will occur, and what, if anything, can be done to minimize either the probability of the failure or its impact. There are a number of factors, internal and external to the network, which must be taken into account before the identification of the risk, some of which have been listed in TABLE 2.

External Factors	Internal Factors
• Co-location network or power failures • Connecting company/pass-through network failures • Unexpected customer configuration changes (firewalls, content filters etc.) • Offsite backup failure • Security issues	• Server hardware failure • Server misconfiguration • Web application failure/misconfiguration • Database failure • Backup Failure • LAN failure • Security issues

TABLE 2: Risk factors internal and external to the network.

External factors like co-location of the network (the location of data center) can have power failure or network failure and the hosting company might not have enough power backup for long power outage. The cloud computing data centers usually have three levels of power backup systems, like a battery backup for few hours, followed by a generator power backup for 3 days, and finally a third party power backup system via grid power. Unexpected changes made in the configuration of the network (by the customer or the hosting company) may lead to a network outage and the loss of business and reputation. Security issues always remain a concern, even for highly protected data centers.

Internal factors like failure of the hardware or software, the misconfiguration of the server, etc. should be considered while calculating the risks involved for meeting the high availability requirement of the focus network of a company. The TABLE 3 shows the production network risk matrix with some details on the most likely causes of production network outages.

Category	Component	Significance	Impact	Probability	Mitigation	Rating
Software	Web application failure	Loss of service	High	Medium	Testing, quality assurance, monitor counters	7
	Database failure	Loss of service/data	High	Low	Backups, auto-notifications	6
	Operating system failure	Loss of service/data	Critical	Low	Automatic failover	7
	Human error	Loss of service	High	Medium	Training, managing access control	7
Security	Application security exploits	Potential loss of service, defacing reputation	High	Minimal	Frequent Cenzic scans, code review	6

	Database security exploits	Loss of service, loss of data/ confidentiality	High	Minimal	Frequent Cenzic scans, code review	6
	OS-level exploit/compromise	Loss of service, loss of data/confidentiality	High	Minimal	Frequent security updates, Nessus scans	6
	Physical server compromise	Loss of service, loss of data/confidentiality	High	Non-issue	Secure co-location facility	4
Hardware	Web server: processor failure	Loss of service	Critical	Low	Network load balancing	7
	Hard disk failure	Required dispatch to replace the damaged part	Non-issue	Minimal	Hot swappable RAID enabled	1
	Power system unit (PSU) failure	None	Non-issue	Low	Redundant PSUs	2
	Fan failure	None	Non-issue	Minimal	System health monitor	1
	database server: processor failure	Loss of service	Critical	Low	Virtualization/ live failover	7
	Switch failure	Partial loss of service	High	Minimal	Connect switches in parallel, distribute ports	5
	Firewall failure	Complete server outage	Critical	Minimal	Redundancy or managed solution	6
	Other critical component	Loss of service	High	Low	Virtualization/ live failover	7
Backups	Local backup failure	Loss of recovery data	Medium	Low	Automated failure notifications	4
	Offsite backup failure	Temporary loss of recovery data	Minimal	Medium	Automated failure notifications	4
Legend:						

Impact	Probability*	Risk Rating=Impact + Probability
Critical=5	Extreme (97-99%)=5	
High=4	High (70-96%)=4	
Medium=3	Medium (34-69%)=3	*Probability of 100% means certainly risky, probability of 0% means no risk at all.
Low=2	Low (15-33%)=2	
Minimal=1	Minimal (2-14%)=1	
Non-issue=0	Non-issue(1%)=0	

TABLE 3: Production Network Risk Matrix

The risks identified above do not encompass every possible failure, but rather those most likely to cause, contribute to, or prolong the service interruptions. Risk identification is a continual process, and should re-evaluated on a regular basis, as well as when a change to the hardware, the software or a process is introduced. For example, implementing virtualization will lessen the impact of physical server failure. Thus a risk previously deemed "high" may be moved to a "medium" or "low" severity, after analyzing the risk aspects and production network availability goals.

Next we will analyze how and why migration to clouds will be highly beneficial for a high growth company which is looking for high scalability and cost effective solution for high availability network goals.

4 : CASE STUDY: HIGH AVAILABILITY MIGRATION COST ANALYSIS

If the success of a company depends on the availability of its web-based products, the company needs to evaluate its production infrastructure critically in order to ensure that high network availability goal and failover capabilities are realized even while minimizing the capital expense.

We consider the case of a company X and present the current expenses of the company on ensuring the production network availability. TABLE 4 details the current production network configuration and the network expenses of the company under analysis. The company was till now maintaining its own servers and hosting them in its own company space. The backup of the data (as a safety precaution) was also maintained at the highly secure location of GoGrid and the fees paid during 2009-2011 are shown as the co-location expenses (data backup) in FIGURE 4 excluding the initial setup cost of getting this service.

The business model of the Company X is SaaS (software-as-a-service) model. It has about 20000 customers and has to support 10+ millions of concurrent individual users to during the working hours. The network load is less during nights and during holidays. Thus, the company needs high availability during the day time. The customers are government and private educational institutes which insist upon data privacy from each other and from the world. In order to be competitive in the market, the company's business demands that less than 1 week should be taken to setup the software environment for their new customers. The growth of company is expected to be very high and current network configuration will not be able to support the load and the cost of scaling up the business due to hardware cost, cost of hosting space, getting skilled network engineers and maintenance cost of the network.

4.1 The Current Configuration

Each production server of the company utilizes redundant array of independent disks (RAID), redundant power supplies, and redundant dual port network interface controller (NICs). The company has three products, Product A, Product B and Product C, which reside on physically separate web servers and database servers. Only a few of the Product B databases are mirrored due to the inherent limitations of the SQL Server. Network Load Balancing (NLB) and internet information services (IIS) metabase replication have been configured on the web servers, but are not presently enabled since there is no means of synchronizing the applicant files quickly

enough1. All the websites and databases back up to the NAS, which in turn, backs up every night to GoGrid server image (GSI) offsite storage location. In the current configuration, a critical hardware failure on any of the web or database servers with the exception of the Product B and the NAS will result in several hours of lost availability.

Ref. #	Component(s)	Terms	Cost/Month (in USD)
1	24-Port GB Ethernet Switch	2 year lease (ends 01/2012)	$480.95
	PowerEdge 1950 (Principal DB Server)		
	PowerEdge 1950 (Mirror DB Server)		
2	PowerEdge 1950 (J7 Server)	3 year lease (ends 05/2012)	$503.29
	PowerEdge 1950 (Rollover Server)		
	PowerEdge 2950 (NAS)		
3	Co-location Services	Monthly Subscription	$580.00
	3Mbps/sec Bandwidth		$561.00
	100GB Backup Services		$220.00
4	SonicWALL 2040 maintenance agreement	Yearly Subscription($359.80/yr)	$29.98
5	Dell Service and Warranty (Production)	Yearly Subscription ($1865.54/yr)	$155.46
	Monthly Total:		$ 2,530.68
	Yearly Cost:		$30,368.16

TABLE 4: Production network monthly expenses of the company X (Current, 2012)

4.2 High Availability Options

After initial analysis, it is found that following three options are available, which can provide a better network support than the current self-hosted server system. It further gives an edge to the company X to meet the demand of its rapid growth, 'as-an-when needed' network load, and the desired increase in the availability goals. Maintaining self-hosted servers limits the prospects of scalability and rapid expansion because of a primary reason that new network systems usually take 3 months to get delivered after an order is placed. The three categorically different options are considered, Virtual Iron as virtualization solution, GoGrid as cloud computing solution and HostMySite as dedicated servers option managed by third party (HostMySite) and outsourcing the network maintenance and management to HostMySite rather than maintaining in-house.

4.2.1 Option 1: Virtual Iron's Virtualization Solution

Reducing the impact of hardware failure will require a combination of network load balancing (NLB), clustering, and virtualization.

Web Server Network Load Balancing

The files related to the website can be synchronized using netTransfer. A script scheduled to run several times daily will replicate the IIS metabase files. Once enabled, the Network Load Balancing (NLB) service will automatically distribute the load balancing requests using pre-configured ratios of load to be balanced.

Virtualization

The model presented in FIGURE 1 aims to provide the best possible deployment and maintenance flexibility while reducing the impact of SPOFs to a manageable level. In FIGURE 1, NLB is enabled on the web servers, and their files and IIS metabase will be replicated across two or more web servers. A second NAS will be added and configured as redundant iSCSI target using the Microsoft Clustering Service (MSCS). The VM (virtualization machine) images will reside on a

[1] It will require files to be converted to base64 for allowing the storage in a database

clustered iSCSI drive in an active/passive model. Snap shots will be taken on a scheduled basis and backed up to an alternate iSCSI drive. The resources of the 1950's will be pooled to allow redundant VM's to reside on both servers, granting live migration and automatic failover capability. As with current model of the company, all servers will continue backup to local shared storage, and then to GSI's offsite location.

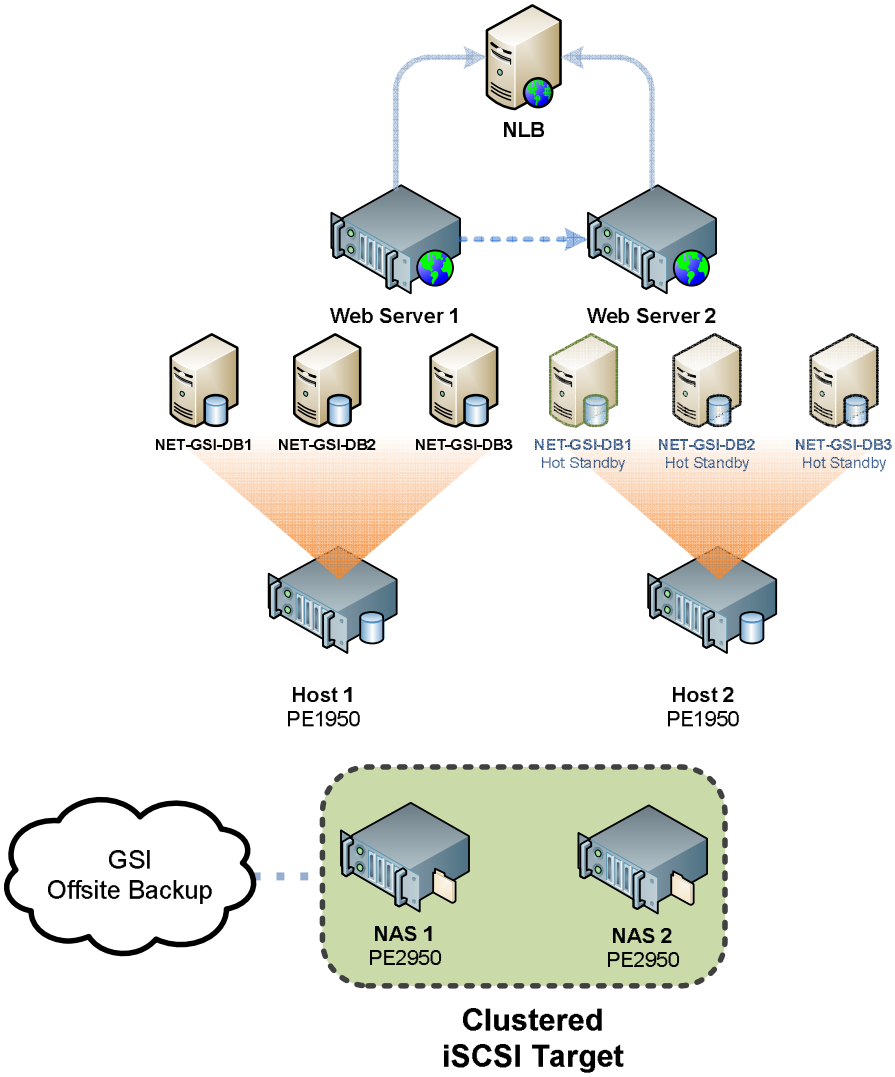

FIGURE 1: Virtual Iron logical topology [40]

4.2.2 Option 2: GoGrid's Cloud Computing Service

Moving to a fully-managed solution will transfer the risk and reduce the cost involved with maintaining own servers. High availability (HA) at the hardware level is achieved, backed by a 100% availability guarantee with a 10,000% remedy clause; i.e., 100 times the client's entire service level fees will be reimbursed in the event of a failure. In this model (shown in FIGURE 2), company will lease persistent VM images from GoGrid.

Servers will be provisioned using a web interface console. Once VM image will be created, clients can RDP into their VM, install additional software and/or make configuration changes. VM image creation may take about 15 minutes, allowing for rapid scalability. Billing will be based upon the total RAM (GB)/hour and the total outbound traffic. Individual VM resources vary by RAM allocation.

In this scenario, a combination of virtual and managed services is used to achieve both flexibility and high availability. Web servers are virtualized in GoGrid's cloud environment with a cross connect two physical SQL Servers in a high availability model using MSCS. In FIGURE 2 the quorum disk is 500GB of SAN storage accessed via 1Gbps iSCSI.

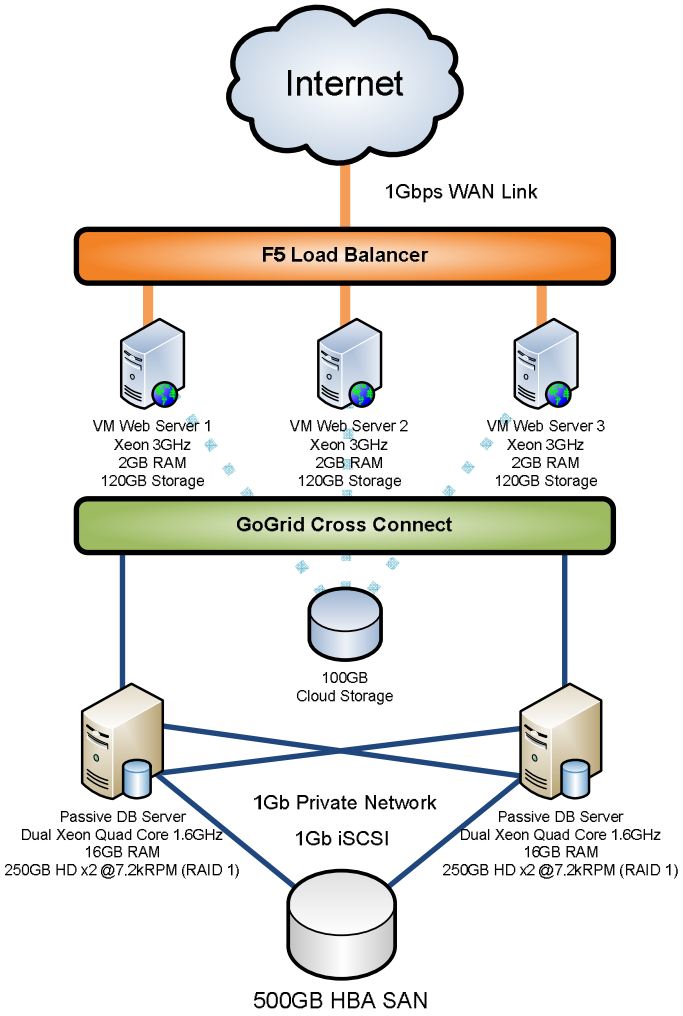

FIGURE 2: GoGrid Advanced Cloud Service /w Managed Database Server Cross Connect [2].

4.2.3 Option 3: HostMySite's Fully-Managed Solution With Hyper-V

Like the option 2, all network and server equipment will be leased. However, in this instance, all the hardware would be dedicated exclusively to the leased company. HostMySite would provide the company with the configuration in FIGURE 3.

This configuration reflects Microsoft's ideal HA topology. The web servers are load balanced with an F5 Hardware Load Balancer. The database servers are clustered using MSCS, the central storage is a 500GB Storage Area Network (SAN) partition. Communication between the SAN and database servers is achieved with 4GB Fibre Channel Host Bus Adapters (HBAs). Internet connectivity is provided at 100Mbps with a soft-cap of 2TB a month. It is notable that HostMySite does not charge overages for exceeding bandwidth, but may request a service upgrade if cap is consistently exceeded.

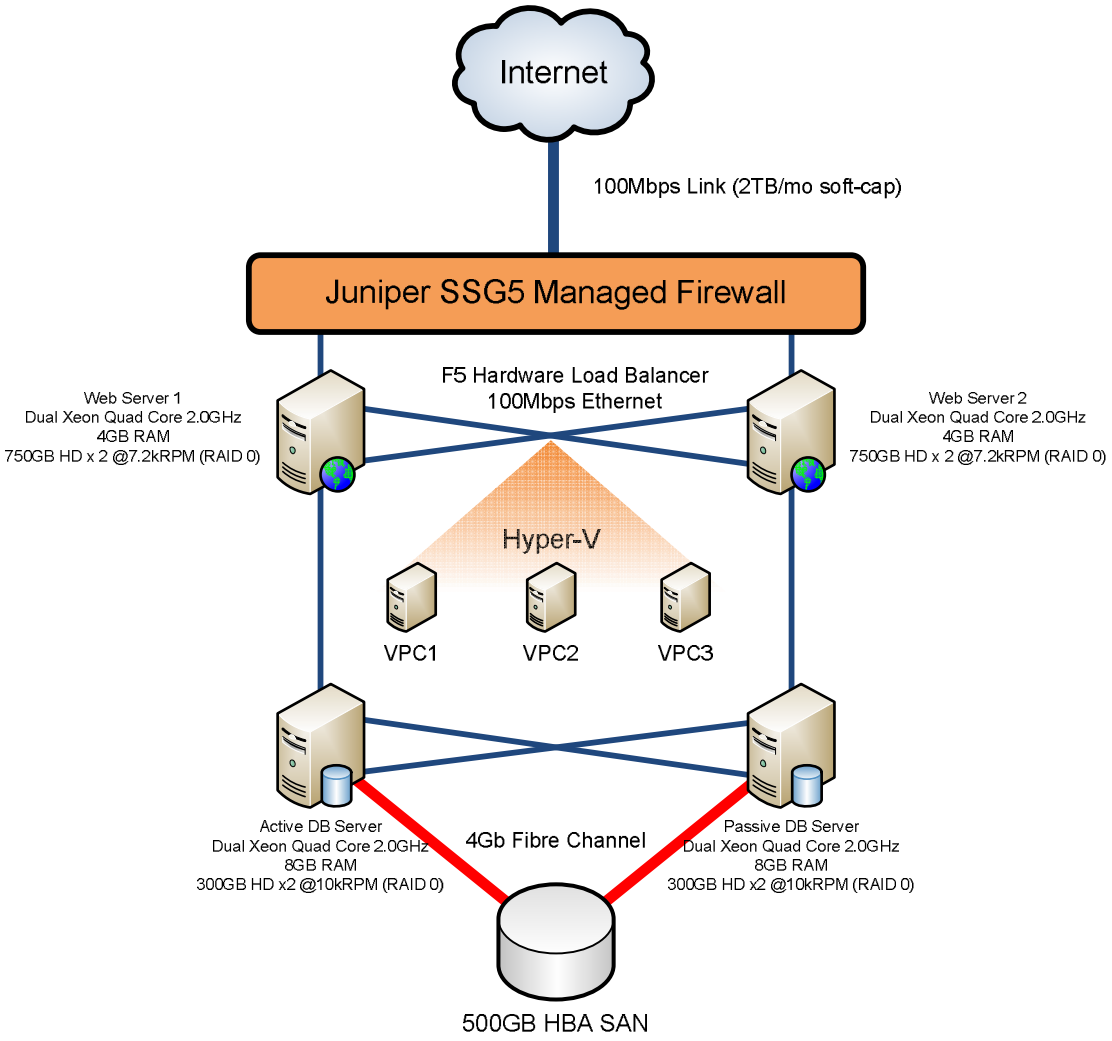

FIGURE 3: HostMySite topology [41].

5 : COMPARISON OF VARIOUS NETWORK OPTIONS

5.1 Cost Comparison of Various Options

In this section discussion on cost analysis of various network solution options available for the company has been done. While comparison of the cost current and future requirement of network load and storage space, scalability, ease of scaling up (when high business session and network load is high) and scaling down (when low business session), security of customer data and bandwidth available. The configurations of various options are chosen by keeping in account the optimal performance, cost effectiveness and closest required configuration offered by the particular service providers.

Current: FIGURE 5 shows a projection of the costs for a period of two years (2011 to 2013) for the current configuration (section 4.1). It is assumed that the network configuration remains the same and the company adds 25% more resources every six months. It can be estimated from the FIGURE 4 that there might be a steady increase in the co-location costs incurred by the company. Despite the increasing costs, the benefits of high availability will not be achieved since the

essential configuration remains the same. This drawback further does not support the scalability and robustness requirement of company's business goals.

Virtual Iron
While we can expect some improvement in the terms of high availability, the co-location fees and future hardware lease charges are expected to increase continuously. Over a time period of two years, this option will be the costliest due to the additional hardware lease charges in comparison to the current configuration.

GoGrid
This solution requires an initial setup fee, which increases the cost of entry. In this solution, the company will continue to incur the monthly co-location fees until the end of the contract period in June, 2011. Company will also incur the cost of the remaining hardware leases until January and May 2012 respectively. After May 2012 however, the total monthly cost will be expected to fall below the cost of the estimated co-location service costs at that time. After the first year, the contract can be made renewable on a month to month basis, allowing the company to explore more sophisticated solutions as they become available (e.g. Microsoft Azure).

HostMySite
While less expensive than pursuing Virtual Iron, this option comes at greater cost than GoGrid without the flexibility of being able to add additional web servers on the fly(dynamically as and when required).

Option	Pros	Cons	Estimated monthly expenses[2]
Current Configuration	• Direct control of hardware • Offsite backup service	• No failover capability • Little redundancy • High operational cost • Inability to react quickly to change • Limited physical capacity (rack space) • Extremely limited bandwidth and severe overage penalties	$1,361.00/mo
Virtualization (Virtual Iron customized integration of virtualization of server management s/w)	• Direct control of hardware • Improved failover capability • Improved availability • Easy to manage	• High operational cost • Must procure more hardware to scale up • Must manage hardware and co-location services • Limited physical capacity (rack space) • Extremely limited bandwidth and severe overage penalties	$1,682.08/mo
GoGrid (Customized solution on top of Pay-as-you-go plan with 16GB RAM+16 core systems)	• Highly scalable • Can easily downscale during slow seasons • Easy to manage • Increased bandwidth (1Gbps) • Cost effective at	• Expensive to pilot • SQL backend must be achieved through a physical cross connect rather than virtualization (performance issues)	$2,714.90/mo

[2] Monthly cost excludes hardware leases and maintenance fees

	higher tiers • 10,000% SLA remedy clause		
HostMySite (Customized solution + Dedicated application server plan with 16 GB RAM)	• Best performance • Microsoft recommended configuration • Increased bandwidth (100Mbps) • Fibre Channel backend	• Expensive monthly fee • Expensive to expand • Cannot scale quickly	$3,633.60/mo

TABLE 5: Comparison of the current configuration, the Virtual Iron solution, the Go-Grid solution, and the HostMySite solution in terms of their pros and cons and estimated monthly expenses.

5.2 Comparison of Network Risk

From TABLE 3, software, security and hardware failure have the highest risk ratings. So, we consider these issues for comparing the three options proposed in section 4.2.

The web application failure depends on software developer of the company X, the availability of network, and its bandwidth. The failure related to the availability can be reduced by moving to better availability options like Virtual Iron or GoGrid network. HostMySite does provide better availability than the current configuration but is still unable to handle the scalability aspect.

Database failure risk rating is high and its impact can be reduced by auto-backups and auto notification when the failure occurs. Auto backup service is provided by all the available three options. GoGrid has a very high level of disaster recovery support as they provide three layer backup facilities at three remote locations other than the data center. On the other hand, the other options have two levels of back up facilities only, which is better than current configuration, but not as good as GoGrid.

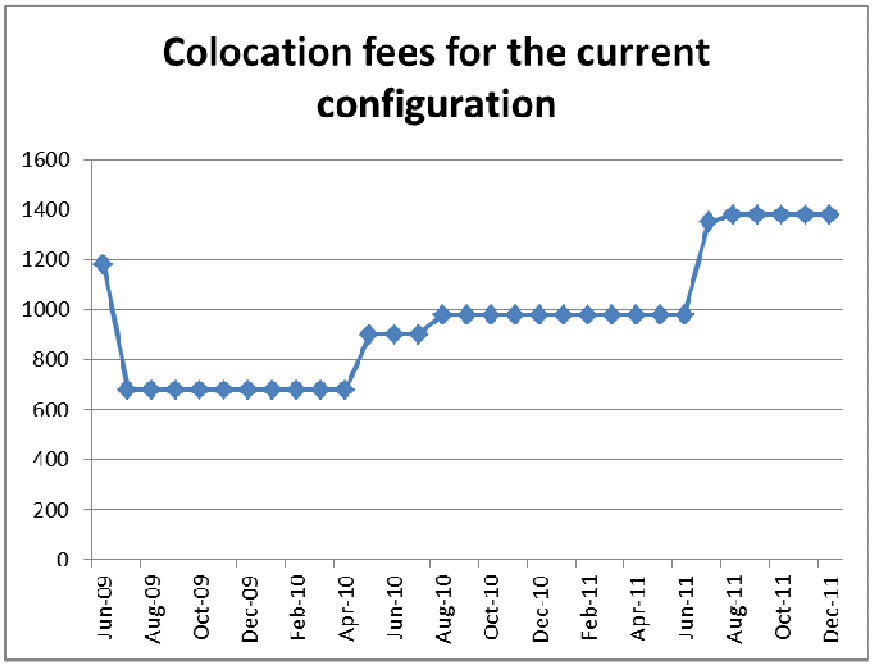

FIGURE 4: With the exception of the initial setup cost, co-location hosting fees had increased incrementally at rate of approximately 25% every 6 months.

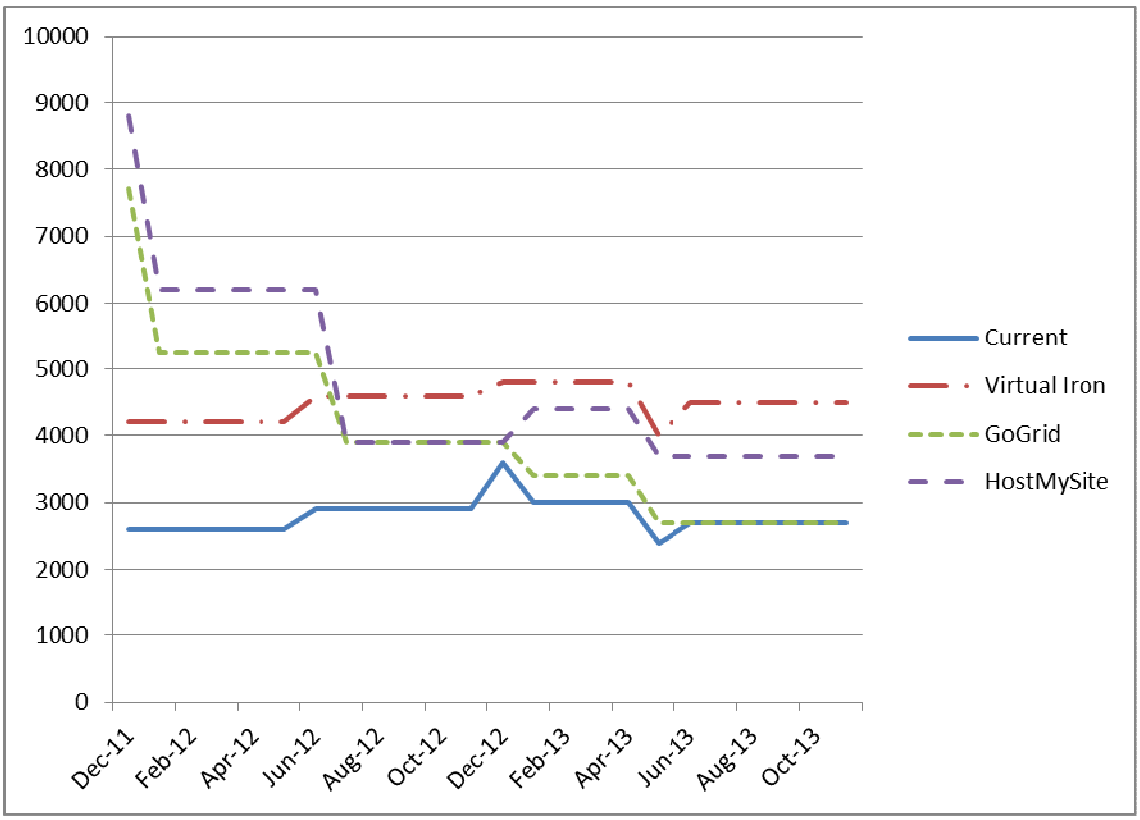

FIGURE 5: The total expenses for each option including the setup fees, the current co-location fees, and the hardware leases. The 'current' line assumes that company will continue to add hardware and increase the capacity at rate of 25% (over the previous period of 6 months) every 6 months.

Operating system failure can critically impact the business though its mitigation is simply automatic recovery. The recovery does take some time and so the business is affected. The cloud based services provide better OS failure support because at least two levels of back up are available for every kind of support at the data center. Thus, during the recovery process of the operating system, another standby operating system can serve the business with minimizing the loss.

Human error impact is very high on the business when the company is lacking in expert network engineer. Other than the expertize, if the resource and network engineer is overloaded with the long working hours, then moving to a third party managed network service can reduce the probability of human error. Further, the expertise of data centers is in managing the network services and high level of automation reduced the chances of human errors, thus their services are expected to be highly efficient and better managed than the legacy self-managed network system.

Security risks always haunt the network engineers in any organization. The probability of unauthorized access to the hosting place is higher when the physical servers are managed by a third party [42]. However, most third party solution providers take serious measures for reducing the probability of data compromise and unauthorized access. At the cloud center of GoGrid, the security level is very high. Two level biometric security is in place in addition to physical security and access card based security. The server at the data center can be accessed by a thin client (with access to own data only) after passing the entire security authentication. HostMySite and Virtual Iron facilities have relatively lesser measures in place.

Public cloud is shared resources space and recent spurs in data theft due to various software and hardware issues at the data center put a red flag for companies looking for cloud as network option when their data privacy is highly required [42]. To mitigate this aspect, GoGrid has an option to take the services with dedicated (not shared) hardware, of course at a higher price (customized solution charges are mentioned in TABLE 5). For less sensitive and low cost applications, the companies may still consider the usual option in which the data may be hardware shared, but software separated from other companies. Further, GoGrid uses hardware virtualization layer Xen [2] and F5 hardware load balancer for managing various types of hardware failures. HostMysite provide dedicated server without virtualization and Virtual Iron have performance similar to GoGrid in this regards. With very high provision of real time network load balancing makes the cloud services highly suitable for high growth disruptive technologies based companies like company X.

5.3 Recommendation

Based on the above analysis from the considered options, GoGrid offers the best combination of both flexibility and performance. Over a period of two years, the GoGrid solution is projected to be less costly in comparison to the case where the company keeps the current configuration and adds 25% more resources every six months (see FIGURE 5). Between now and the end of the GSI contract agreement, all production network servers would be brought in-house (that is within the company premise), and would server and our internal testing and development environment.

Cloud computing solutions are fast, user-friendly and cost-effective compared to traditional IT solutions. So, should we assume that traditional system has outlived its utility and cloud based solutions are going to replace them? This depends on the companies and their network requirements.

There are certain downsides also with cloud computing solutions. The biggest threat is that as an organization's crucial data is stored in the cloud and if unauthorized access is obtained to these data, it would pose a big question mark to the cloud based CRM reliability. As these cloud service providers have tie-ups with various third-party vendors, it makes the system more susceptible to data theft. In the above case study, part of this issue has been addressed by using dedicated servers at the cloud computing data center which are separate from general public pool. However, 100% security is almost impossible in practice. Future research and development in making the cloud computing more secure will reduce the susceptibility of cloud centers from security issues.

6 : CONCLUSION AND FUTURE WORK

This paper targets the problem of analyzing the cost of migration to the cloud servers for the purpose of scalability and high network availability. The risks challenging the high network availability goals are identified. A detailed classification of risks and their sources are identified and a risk matrix is proposed to evaluate the risks based on their impacts and probability of occurrence. Mitigation techniques are also suggested. In addition, an explicit case study of a hypothetical company is presented. The various options for improving the scalability and the network availability are discussed and the costs of each of the options are evaluated. In the current case study, migration to a GoGrid cloud solution is the best option. However, the suitability of the options may vary from case to case. The case study presented here shall serve as example for calculating the suitability of moving to any cloud service providers for scalability and high availability seeking companies/organization/application [43]. We will extend this work with additional analysis of the other services offered by clouds, in particular storage and network services, and try to address the questions like how do various cloud computing solutions respond to the combined stress of workloads with different characteristics and the requirements that the diverse populations of cloud users are supposed to incur in the future?

REFERENCES

[1] Amazon Inc. "Amazon Elastic Compute Cloud (Amazon EC2)," http://aws.amazon.com/ec2/, 2011, [Jan 15, 2012].

[2] GoGrid. "GoGrid Cloud-Server Hosting," http://www.gogrid.com, 2011, [Jan 15, 2012].

[3] A. Losup, O. Sonmez, S. Anoep *et al.*, "The performance of bags-of-tasks in large-scale distributed systems," in Proceedings of the 17th International Symposium on High Performance Distributed Computing 2008, HPDC'08, 2008, pp. 97-108.

[4] I. Raicu, Z. Zhang, M. Wilde *et al.*, "Toward loosely coupled programming on petascale systems," in Proc. ACM Conf. Supercomputing (SC), 2008, pp. 22.

[5] D. G. Feitelson, L. Rudolph, U. Schwiegelshohn *et al.*, "Theory and practice in parallel job scheduling," *Job Scheduling Strategies for Parallel Processing,* vol. 1291, pp. 1-34, 1997.

[6] L. Youseff, R. Wolski, B. Gorda *et al.*, "Paravirtualization for HPC Systems," *Lecture Notes in Computer Science*, pp. 474-486, 2006.

[7] E. Deelman, G. Singh, M. Livny *et al.*, "The cost of doing science on the cloud: The montage example," in SC '08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing, 2008, pp. 1-12.

[8] M. Palankar, A. Lamnitchi, M. Ripeanu *et al.*, "Amazon S3 for science grids: A viable solution?," *International Symposium on High Performance Distributed Computing, HPDC 2008 - Proceedings of the 2008 International Workshop on Data-aware Distributed Computing 2008, DADC'08*, pp. 55-64, 2008.

[9] E. Walker, "Benchmarking amazon EC2 for high-performance scientific computing," *Login,* vol. 33, no. 5, pp. 18-23, 2008.

[10] L. Wang, J. Zhan, W. Shi *et al.*, "In cloud, do mtc or htc service providers benefit from the economies of scale?," in Proc. Second Workshop Many-Task Computing on Grids and Supercomputers (SC-MTAGS), 2009.

[11] J. S. Vetter, S. R. Alam, T. H. D Jr *et al.*, "Early evaluation of the cray XT3," in Proc. 20th Int'l Conf. Parallel and Distributed Processing Symp. (IPDPS), 2006.

[12] S. Saini, D. Talcott, D. C. Jespersen *et al.*, "Scientific application-based performance comparison of SGI altix 4700, IBM POWER5+, and SGI ICE 8200 supercomputers," in Proc. IEEE/ACM Conf. Supercomputing (SC), 2008, pp. 7.

[13] T. H. Dunigan Jr, "Early Evaluation of the Cray X1," in Proc. ACM/IEEE SC2003 Conf. (SC 03), 2003, pp. 18.

[14] S. R. Alam, R. F. Barrett, M. Bast *et al.*, "Early evaluation of IBM bluegene/P," in Proc. ACM Conf. Supercomputing (SC), 2008, pp. 23.

[15] F. Petrini, D. J. Kerbyson, and S. Pakin, "The case of the missing supercomputer performance: Achieving optimal performance on the 8,192 processors of ASCI Q," in SC '03: Proceedings of the 2003 ACM/IEEE Conference on Supercomputing, 2003, pp. 55-55.

[16] D. J. Kerbyson, A. Hoisie, and H. J. Wasserman, "A performance comparison between the Earth Simulator and other terascale systems on a characteristic ASCI workload," *Concurrency Computation Practice and Experience,* vol. 17, no. 10, pp. 1219-1238, 2005.

[17] A. Iosup, C. Dumitrescu, D. Epema *et al.*, "How are real Grids used? The analysis of four Grid traces and its implications," in Proceedings - IEEE/ACM International Workshop on Grid Computing, 2006, pp. 262-269.

[18] R. Biswas, M. J. Djomehri, R. Hood *et al.*, "An application-based performance characterization of the columbia supercluster," in Proc. IEEE Conf. Supercomputing (SC), 2005, pp. 26.

[19] A. Iosup, and D. Epema, "GRENCHMARK: A framework for analyzing, testing, and comparing grids," in Sixth IEEE International Symposium on Cluster Computing and the Grid, 2006. CCGRID 06, 2006, pp. 313-320.

[20] S. Williams, J. Shalf, L. Oliker *et al.*, "The potential of the cell processor for scientific computing," in Proceedings of the 3rd Conference on Computing Frontiers 2006, CF '06, 2006, pp. 9-20.

[21] D. Nurmi, R. Wolski, C. Grzegorczyk *et al.*, *The Eucalyptus Open-source Cloud-computing System*, vol. 2011, 2008.

[22] B. Quétier, V. Neri, and F. Cappello, "Scalability comparison of four host virtualization tools," *Journal of Grid Computing,* vol. 5, no. 1, pp. 83-98, 2007.

[23] P. Barham, B. Dragovic, K. Fraser *et al.*, "Xen and the art of virtualization," *Operating Systems Review (ACM),* vol. 37, no. 5, pp. 164-177, 2003.

[24] B. Clark, T. Deshane, E. Dow *et al.*, "Xen and the art of repeated research," in USENIX Annual Technical Conference, FREENIX Track, 2004, pp. 135-144.

[25] A. Menon, J. R. Santos, Y. Turner *et al.*, "Diagnosing performance overheads in the xen virtual machine environment," in Proceedings of the First ACM/USENIX International Conference on Virual Execution Environments, VEE 05, 2005, pp. 13-23.

[26] N. Sotomayor, K. Keahey, and I. Foster, "Overhead matters: A model for virtual resource management," in Proc. IEEE First Int'l Workshop Virtualization Technology in Distributed Technology (VTDC), 2006, pp. 4-11.

[27] A. B. Nagarajan, F. Mueller, C. Engelmann *et al.*, "Proactive fault tolerance for HPC with Xen virtualization," in Proceedings of the International Conference on Supercomputing, 2007, pp. 23-32.

[28] L. Youseff, K. Seymour, H. You *et al.*, "The impact of paravirtualized memory hierarchy on linear algebra computational kernels and software," in Proceedings of the 17th International Symposium on High Performance Distributed Computing 2008, HPDC'08, 2008, pp. 141-152.

[29] W. Yu, and J. S. Vetter, "Xen-based HPC: A parallel I/O perspective," in Proceedings CCGRID 2008 - 8th IEEE International Symposium on Cluster Computing and the Grid, 2008, pp. 154-161.

[30] L. Cherkasova, and R. Gardner, "Measuring CPU Overhead for I/O Processing in the Xen Virtual Machine Monitor," in Proceedings of the USENIX Annual Technical Conference, 2005, pp. 387-390.

[31] U. F. Minhas, J. Yadav, A. Aboulnaga *et al*., “Database systems on virtual machines: How much do you lose?,” in Proceedings - International Conference on Data Engineering, 2008, pp. 35-41.

[32] W. Huang, J. Liu, B. Abali *et al*., “A case for high performance computing with virtual machines,” in Proceedings of the International Conference on Supercomputing, 2006, pp. 125-134.

[33] L. Gilbert, J. Tseng, R. Newman *et al*., “Performance implications of virtualization and hyper-threading on high energy physics applications in a grid environment,” in Proc. IEEE 19th Int'l Parallel and Distributed Processing Symp. (IPDPS), 2005.

[34] J. Zhan, L. Wang, B. Tu *et al*., “Phoenix cloud: Consolidating different computing loads on shared cluster system for large organization,” in Proc. First Workshop Cloud Computing and Its Application (CCA '08) Posters, 2008, pp. 7-11.

[35] M.-E. Bgin, B. Jones, J. Casey *et al*., *An EGEE comparative study: Grids and Clouds - Evolution or revolution?*, 2008.

[36] C. Evangelinos, and C. Hill, “Cloud Computing for Parallel Scientific HPC Applications: Feasibility of running Coupled Atmosphere-Ocean Climate Models on Amazon EC2,” *Ratio,* vol. 2, pp. 2-34, 2008.

[37] L. Youseff, M. Butrico, and D. Da Silva, “Toward a unified ontology of cloud computing,” *Grid Computing Environments Workshop*, pp. 1-10, 2008.

[38] M. Armbrust, *Above the clouds: A berkeley view of cloud computing*, 2009.

[39] R. Prodan, and S. Ostermann, “A survey and taxonomy of infrastructure as a service and web hosting cloud providers,” in Proc. Int'l Conf. Grid Computing, 2009, pp. 1-10.

[40] Oracle Inc. "Oracle Cloud Computing," http://www.oracle.com/us/technologies/cloud/index.html, 2011, [Dec 11, 2011].

[41] HostMySite. "Hosting Cloud Hosting," http://www.hostmysite.com/cloud/, 2011, [Jan 15, 2012].

[42] S. O. Kuyoro, F. Ibikunle, and O. Awodele, “Cloud Computing Security Issues and Challenges,” *International Journal of Computer Networks (IJCN),* vol. 3, no. 5, pp. 247-255, 2011.

[43] D. K. Prasad, “Adaptive traffic signal control system with cloud computing based online learning,” in Eighth International Conference on Information, Communications, and Signal Processing (ICICS 2011), Singapore, 2011.