Performance Implications of Cloud Computing

May 6, 2010

Lydia Duijvestijn, Avin Fernandes,
Pamela Isom, Dave Jewell, Martin Jowett,
Elisabeth Stahl, Todd R Stockslager

Table of Contents

Objectives & Scope

This paper is intended to be the first in a series of papers that consider the performance and capacity considerations of the evolving Cloud Computing model.

In this introductory paper we provide an overview of cloud computing and its potential benefits from both an IT and business perspective. We discuss the different cloud computing deployment models and introduce the roles involved in delivering Cloud Solutions.

We go on to highlight the key performance and capacity considerations when designing and delivering those solutions, discussing the performance engineering considerations as they pertain to those individual roles.

We conclude with a discussion of some of the Cloud Offerings available today, briefly discussing the performance considerations for each.

Future papers will then expand on the performance considerations in more detail, both from the perspective of those responsible for creating solutions for the cloud and those responsible for managing the solutions once deployed.

Introduction

Driven by trends in the consumer Internet, cloud computing is a new way to consume and deliver IT services. The cloud computing model builds on the maturation of the Web, combining rapid scalability, proliferation of the Internet and internet connected devices, unprecedented self-service and the emergence of elegant web-based applications. It allows users to execute complex computing tasks without the need to understand the underlying technology.

Cloud computing is emerging at a critical time for the IT industry. There is a growing realization that physical and IT assets, systems, and infrastructure are fast reaching a breaking point. As the pace of business and society in general continues to accelerate, the physical and digital foundations on which progress depends are straining to keep up.

- The explosion of data, transactions, and digitally-aware devices is straining existing IT infrastructure and operations.
- Exponential growth in communications subscribers and services is exposing limitations in network bandwidth and storage capacity.
- Supply inefficiencies as well as demand spikes are putting pressure on Today's energy and utility systems.

As a key part of the "smarter planet" vision articulated by IBM and embraced by others, the emerging cloud computing model leverages improved technologies and is characterized by innovative internet-driven economics that exploit the massive underlying ability of the technology to scale. Cloud computing offers the vision for improved service - not just high availability and quality of existing services, but also meeting expectations for real-time, dynamic access to innovative *new* services. The cloud computing model allows for reducing cost- not just containing it, but achieving *breakthrough* productivity gains through virtualization, optimization, energy stewardship, and flexible sourcing.

However this model is not without its challenges. The complexity of cloud computing, with its increased dependence on internet, virtualization and "on demand" enabling technologies requires close attention to the considerations of performance and capacity throughout the design, delivery and management of cloud based solutions. This is essential if we are to deliver the benefits alluded to above.

This paper starts by providing an overview of cloud computing and its potential benefits from both an IT and Business perspective, discusses the different cloud computing deployment models and introduces the different roles involved in delivering Cloud Solutions. It then goes on to highlight the key performance and capacity considerations when designing and delivering those solutions and discusses the performance engineering considerations as they pertain to those individual roles. It concludes with a discussion of some of the Cloud Offerings available today, briefly discussing the performance considerations for each. Usage scenarios are provided throughout for clarification and elaboration of real-life cases where performance considerations are part of the cloud solution strategy.

Cloud Computing Overview, Benefits and Challenges

Cloud computing is a flexible and cost-effective delivery platform for providing IT services over the Internet which has already proven useful for numerous applications.

Cloud resources can be rapidly deployed and easily scaled, with all business processes, applications, and services provisioned on demand, regardless of the user location or device. Cloud computing provides organizations the opportunity to increase service delivery efficiencies, streamline IT management, and better align IT services with dynamic business requirements. It provides support for core business functions along with the capacity to develop and deploy new and innovative services, making them more accessible than ever before.

Business Drivers for Cloud Computing

Why is cloud computing so interesting? There are a number of key business drivers that provide an incentive to examine cloud.

The number one driver is **cost**. Clouds offer a "pay as you go" model. This model allows a company to invest in resources as they are needed rather than in anticipation of the need. This fact is especially important when one considers that the investment is looked at by the business as an operating expense rather than a capital expenditure.

The next key driver is **speed to value** - "I can deploy resources in this environment quickly". This driver is especially important in environments that need to grow or shrink quickly.

Combine these two drivers and you get a service that grows and shrinks with users' needs and allows them to pay only for service usage. "Today it only cost me $10, yesterday when the demand was higher it cost me $20". Many look to the portability of services that ensure that the service will be available regardless of the operating platform. Overall, clouds present a way to dynamically offer a service to a community that will meet their needs from an availability and performance perspective, while keeping operating costs low and limited to expenses based on what was actually used rather than capital investment based on projections of what the user might need.

Advantages to IT Organizations Deploying Infrastructure Architecture

The cloud computing model offers many advantages to IT organizations deploying infrastructure architecture, i.e. infrastructure provisioning. It has the potential to expand and automates resource virtualization, datacenter resource billing and metering, and self-service catalogues and requests. Cloud computing enables *workload-optimized* solutions with efficiencies and innovations across the business in areas such as development, test, analytics, infrastructure, and storage that can be quantified as:

- Reduced capital expenditures and labor costs.
- Rapid provisioning and de-provisioning of services.
- Enhanced resource pooling as computing resources are pooled to provide multiple capabilities to consumers. Resource pooling allows virtual and physical resources to be dynamically configured and assigned based on service level agreements as well as *demand*.
- Superior service management with visibility, control and automation across IT and business services.
- New deployment choices over the cloud, behind the firewall or as an integrated service delivery platform.

Benefits to Business Organizations Deploying Business Solutions

Business solutions should always drive technology solutions, a long-acknowledged truism that is still and even more valid for cloud computing.

- *Today* up to 80% of data is unstructured content (email, video, images).
- Storage capacity shipments are growing at 54% CAGR. [9]
- Medical images will take up 30% of the world's storage.

  The field of medical imaging is driving breakthroughs in diagnosis and treatment in medicine – and that will only accelerate. As a result, there is exponential growth in the number and size of digital medical images. Medical images that used to be two-dimensional and 1MB in size a few years ago are now typically four-

dimensional and 1TB in size. By 2010, it's estimated that 30% of the world's storage will be taken up by these medical images.
[10]

Cloud computing as a technology is only viable because that technology offers many benefits to business organizations deploying business solutions:

- *Optimized systems resources* to keep developers productive
- Reduced capital and licensing expenses—as much as 50 to 75 percent—by on demand provisioning of virtualized test resources [2]
- Decreased operating and labor costs—as much as 30 to 50 percent—by automated provisioning and configuration of test environments [2]
- Facilitated innovation and shorter time to market by improving test provisioning from

  weeks to minutes and reducing test cycle time
- Improved quality by reducing defects that result from faulty configurations and poor modeling—as much as 15 to 30 percent [2]

Attacking Costs as a Driver

Just as it is true that business solutions should always drive technology solutions, it is equally true that cost drives both business ***and*** technology solutions. Cloud computing identifies and attacks costs as a driver for the IT advantages and business benefits listed above. Hardware virtualization reduces costs by boosting hardware utilization, stacking multiple virtual servers in a physical server. This virtualization reduces software licensing costs as charges for operating systems and other software are typically based on the number of physical servers instead of the number of instances. Hence fewer physical servers require fewer licenses.

System administration and operation costs are also reduced in cloud infrastructures with fewer physical servers. Labor savings are realized in provisioning processes that automate and standardize service request management and fulfillment. These reduced administration and operations costs also extend to reduced development and test team costs; by providing ready-to-use systems quickly development and test team idle and waiting times are reduced, and their flexibility in planning, execution, and deployment is increased.

Cloud Computing Challenges

A number of challenges need to be overcome before the promises of cloud computing can be realized.

Deploying a set of cloud constructs requires proper planning. The areas noted here are all critically important as areas of focus. Each one incrementally can improve overall operations. But improvements in one area could cause strain in another. For example:

- Demand services provisioned on demand for a better enterprise wide information infrastructure could stress the issues of security and business resiliency.
- Creating highly virtualized resources demands a stronger, more integrated service management approach.
- Consolidation to optimize systems could drive up the density of systems, thereby putting more strain on the environment issues.
- Converging business and IT infrastructures can be a daunting task, if not handled in an integrated way.

So in today's world it is critical to look at all of it together. How do things interrelate so that improvements in one area are matched with tools and techniques to support them in another? Taking a view of each of these areas is important but more importantly is a plan to integrate them together, creating the backbone or "DNA" needed to thrive in a Smarter Planet.

Reference [8] mentions 10 obstacles for cloud computing, five of which are related to quality-of-service aspects, such as availability, performance and scalability.

Cloud Perspectives

Cloud is a new delivery and acquisition model of IT and IT-enabled services that is inspired by Internet consumer services. Cloud computing allows users to consume resources in a fundamentally different way. In the past, consumption was based on a limited set of resources that were well delineated and identified beforehand. With cloud computing, services can be rapidly provisioned and released with minimal human management effort or service provider interaction. The cloud may be considered from different perspectives. Clouds may be deployed via several different models and delivered as varied services. Roles for cloud consumption and delivery are important to understand in order to evaluate performance implications.

Figure 1 below illustrates the different perspectives that are discussed in this paper.

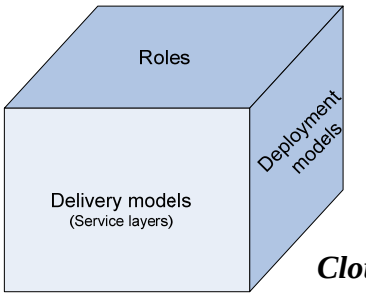

Figure 1. Cloud Perspectives

To introduce the different perspectives three fundamental questions should be examined as they pertain to the Cloud Service Offering:

Question 1: What types of service are provided?

Cloud Delivery Models

Different *service layers* can be distinguished in the cloud arena, depending on the level in the solution stack at which the service is being delivered:

- The top layer, corresponding to the business application viewpoint, hosts ***Software as a Service*** **(SaaS)**
- The middle layer, corresponding to the middleware services viewpoint, hosts ***Platform as a Service (PaaS)***
- The lowest layer, corresponding to the physical environment viewpoint, hosts ***Infrastructure as a Service (IaaS)***

In line with reference [8], the cloud is defined as the virtualized infrastructure that resides on the lowest level of the solution stack. The higher service layers are dependent upon the existence of the underlying, supporting service layers. Service providers may in turn be service users; e.g. a SaaS provider may or may not be a SaaS user; a SaaS provider may or may not be a PaaS user; SaaS and PaaS providers are directly or indirectly IaaS users.

The quality-of-service characteristics of the upper service layers are dependent upon the quality-of-service characteristics of the underlying layers.

Software as a Service (SaaS) contains:

- Business process support
- Enterprise applications
- Collaboration services

Software as a Service provides end user applications as a service. Although the SaaS classification preexisted the cloud computing terminology, it is generally accepted that massively scalable SaaS offerings can be considered cloud computing service models. The applications are generally accessible through the web. User configuration is limited to application configuration.

Examples of SaaS are Fidelity.com (business processes) or IBM LotusLive (collaboration services);

At the level of SaaS, quality of service is directly perceived by end users and defined by *business transaction response times* and - *throughput*, *business service reliability* and - *availability* and by *scalability* of the applications.

Platform as a Service (PaaS) contains:

- Databases
- Middleware
- Development tools

- Java and Web 2.0 runtimes

Platform as a Service provides a development and deployment environment for applications.

An example of PaaS is Google;

At the level of PaaS, quality of service is indirectly perceived by end users and defined by *technical transaction response times* and - *throughput, technical service reliability* and -*availability* and by *scalability* of the middleware.

Infrastructure as a Services (IaaS) contains :

- Server functionality
- Networking functionality
- Data center functionality
- Storage functionality
- An example of IaaS is Amazon Web Services

The quality of services, provided at the IaaS is defined by infrastructure *performance, capacity, reliability, availability* and *scalability.*

The three service layers are illustrated in Figure 2 below

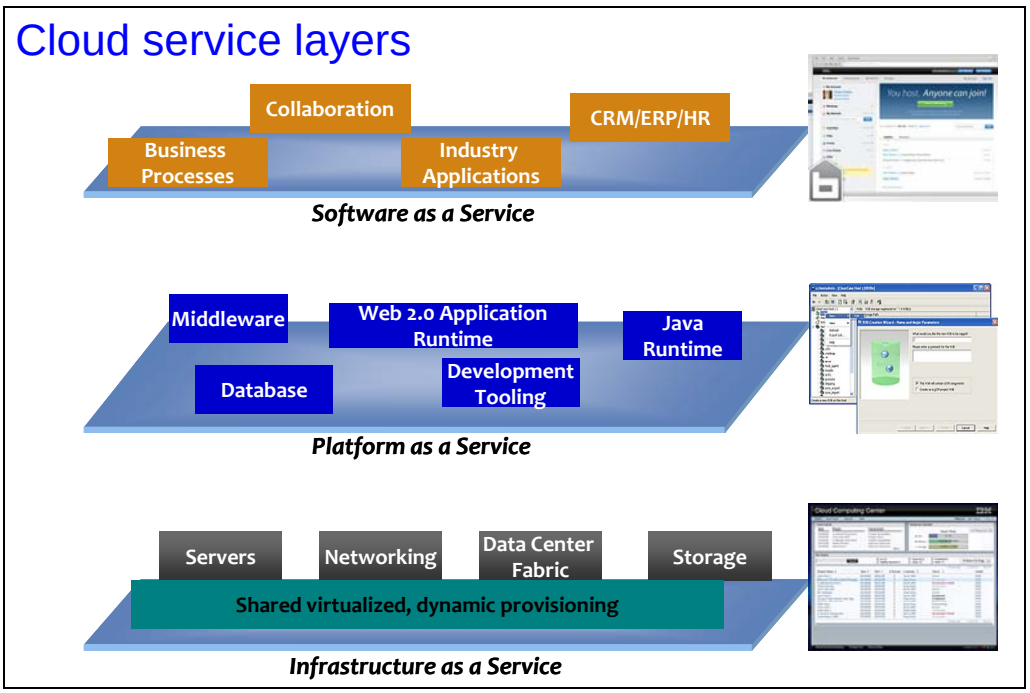

Figure 2. Cloud Service Layers

Question 2: How and where are the services deployed?

Cloud Deployment Models

Cloud deployments are based on workload characterization and non functional business requirements.

A **public cloud** is owned and managed by a third party service provider, and access is by subscription. A public cloud offers a set of standardized business process, application, and infrastructure services on a flexible price-per-use basis. Advantages of a public test cloud include standardization, capital preservation, flexibility, and a shorter time to deploy environments and as a consequence the applications under test.

Private clouds are owned and used by a single organization. They offer many of the same benefits as public clouds, and they give the owner organization greater flexibility and control. A private test cloud provides more ability to

customize, drives efficiency, and retains the ability to standardize and implement organizational best practices. Other advantages include availability, resiliency, security, and privacy. Private clouds can also provide lower latency than public clouds during peak periods, essential when guaranteeing performance as a QoS is key.

Many organizations embrace both public and private cloud computing by integrating the two models into **hybrid clouds** which are (according to National Institute of Standards and Technology (NIST), 2009) bound together by standardized or proprietary technology that enables data and application portability.[4] A hybrid cloud may also contain multiple services and any combination of providers and consumers (described further in the roles section below). In general, hybrids are designed to meet specific business and technology requirements, helping to optimize security and privacy with a minimum investment in fixed IT costs.

Question 3: Who is involved?

Cloud Roles

Three roles are considered for cloud service consumption and delivery.

- ***Consumer*** – User(s) of cloud services. In private clouds, the users are within the same enterprise boundary as the supplier. However, in public clouds the consumers can be either within the same enterprise as the provider or more commonly external to the provider.
- ***Provider*** – The party responsible for providing access to cloud services for registered consumers and for maintaining the QoS attributes for the individual service access requests, including performance. They own the physical assets which are required to produce and deliver cloud services to the consumer.
- ***Integrator*** – The party responsible for the design and construction of a composite cloud service offering. This role may be within the same enterprise as the provider or it could be separate. In the latter case the Service Integrator may well retain some level of accountability for the cloud services provided by the third party consumer.

Each of these roles actually breaks out into a number of sub-roles as described in Figure 3 below.

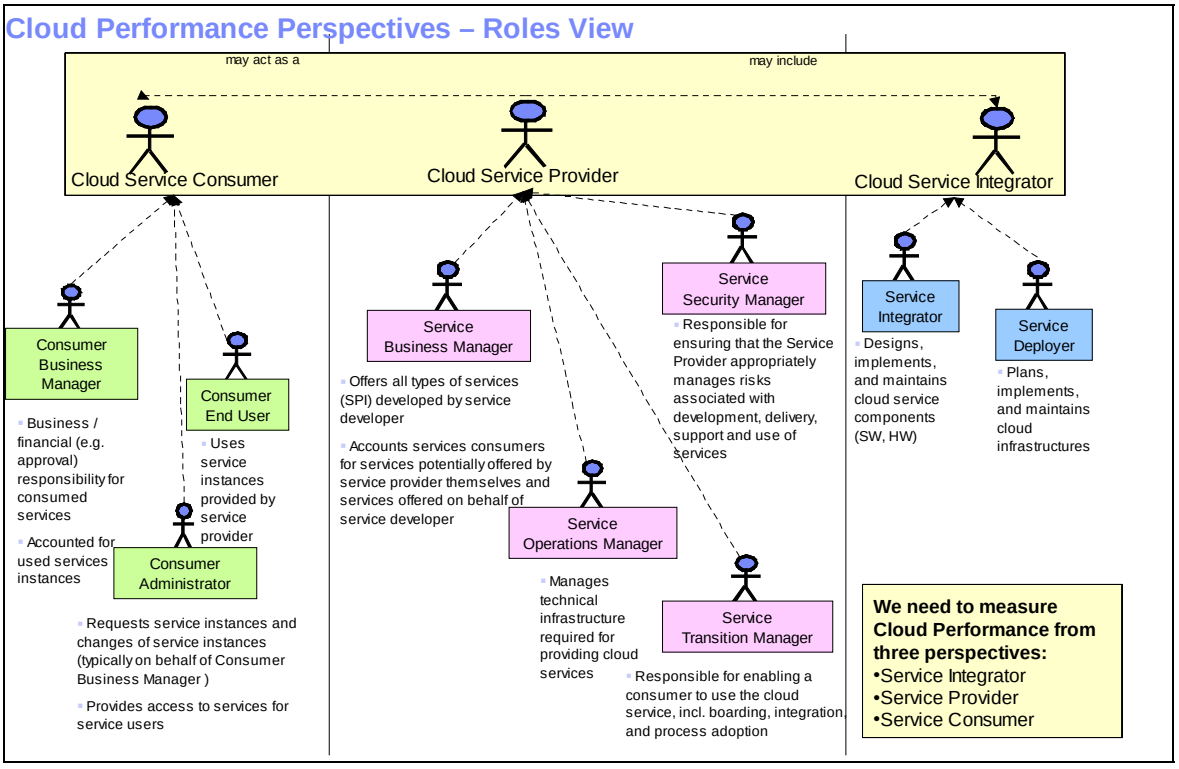

Figure 3. Cloud roles

The actual Deployment and delivery models for clouds as well as the perspective of the role involved, consumption and delivery, are integral to the performance analysis of these solutions.

Cloud Service Consumer Roles

Do not assume that cloud service consumers are merely end users, or that they are not interested in performance. Cloud services, by virtualizing and extending computing resources into the business, make cloud service consumers active and interested participants in the performance of cloud services.

The **consumer business manager** will be most interested in performance in terms of the benefit per dollar spent in obtaining cloud services. In addition, the business manager is keenly interested in ensuring that the cloud performs as expected so that the cloud services purchased are able to add to the profitability of the business.

The **consumer end user** may be the most interested in performance as the person most directly affected by cloud performance issues. The end user will also be responsible for integrating public cloud services into the consumer's processes, and private or hybrid cloud services into the consumer's computing services.

The person in the **consumer administrator** role may be an extremely technical staff member, responsible for integrating a private or hybrid cloud into existing data center services. They will be interested in performance as a component of the usability, manageability, and sustainability of the cloud service.

Cloud Service Provider Roles

The **cloud service business manager** will work with the consumer business manager to ensure that the cloud service meets the business needs for the consumer business manager. Note that for a private cloud, both business managers will work for different organizations within the same company.

The **service operations manager** will be focused on performance aspects of the technical infrastructure that supplies the cloud services. This is the role most directly focused on traditional system performance, in this case as it relates to delivering cloud services to integrators and consumers outside of the managed data center.

The **service security manager** will be concerned with the performance of security integration of private and hybrid cloud services with the consumer's user directory services. Ensuring overall security of the cloud services offerings without reducing performance is also a key component of the security manager's job description.

The **service transition manager** brings the services consumer on board, providing support and training, setting performance expectations and working as a liaison between the organizations

Cloud Service Integrator Roles

The increased technical complexity of integrating disparate components into deployable cloud services increases the performance risks and challenges for the cloud service integrator roles.

While the **service integrator** may do traditional "development" to design, implement, and maintain cloud services, the job description in a cloud computing environment expands to include the full range of hardware, software, or services configuration, integration or assembly of subcomponents to enable cloud services. Performance and scalability of the integration is a key concern of the service integrator.

The **service deployer** plans, implements and maintains cloud infrastructure and deploys and maintains cloud services to the infrastructure. The service deployer will focus on system performance and resource utilization of the cloud service.

Performance engineering and capacity management in cloud environments

The disciplines of performance engineering and capacity management will not become obsolete as a result of the introduction of the cloud and of services provided through the cloud. On the contrary, cloud deployments are more than ever dependent on professional performance engineering and capacity management being in place.

Five of the 10 obstacles and opportunities for cloud computing as mentioned in reference [8] are related to quality-of-service aspects such as availability, performance, capacity or scalability.

- Obstacle # 1 "Availability of service" discusses availability risks for cloud computing as a result of e.g. programming errors, overload of common services or Distributed Denial of Service (DDoS) attacks.
- Obstacle # 4 "Data transfer bottlenecks" discusses the growing data intensity of applications and how this impacts data transfer rates and costs in the cloud.
- Obstacle # 5 "Performance unpredictability" discusses performance risks caused by e.g. inefficiencies in I/O sharing and by high performance computing.
- Obstacle # 6 "Scalable storage" discusses the difficulties of applying cloud computing to solutions requiring highly scalable persistent storage.

- Obstacle # 8 “Scaling quickly” discusses the difficulties of quickly scaling up and down in response to load without violating service level agreements.

Potential solutions to overcome these obstacles have to be assessed for their feasibility in real life situations.

Performance engineers who are advising cloud computing users and cloud computing providers will need to gain a deep understanding of the technical transactions underlying cloud services. The degree to which cloud services can meet agreed service level requirements for availability, performance and scalability can be estimated using performance modelling techniques. Potential performance anti-patterns can be detected before they happen.

The automatic provisioning and usage based costing facilities, two of the major features of cloud computing, rely heavily on fine-grained capacity management. Until more sophisticated tooling for automated monitoring, data collection, analysis and forecasting are in place, capacity management will be more opportune than ever.

But even in an ideal world, where capacity management is fully automated, cloud computing users will still have to analyse their *demand for capacity* and their *requirements for quality of service*. In their negotiations with cloud computing providers, they have to accurately formulate their service level requirements.

It is to be expected that in future the work of the capacity manager will shift towards translating the capacity demands of cloud computing users into capacity plans on which service level agreements with cloud computing providers can be based.

Service Level Agreements in the Cloud

A service level agreement (SLA) is a reciprocal agreement between the provider of an IT service and the consumer of that service about the level of service, or Quality of Service (QoS), to be delivered. Examples of QoS attributes are performance which could cover both response time and throughput, availability, security, etc. The notion of reciprocity is important here; while it is true that SLAs may address the manner in which the service is supplied (e.g. achieving specified response time or availability targets), a complete SLA will also address expected demands for services (e.g. transaction mix and volumes).

In the classic data center environment, SLAs would be formalized once the service provider organization had sufficient experience with the application in question to understand its performance and capacity characteristics relative to the provider’s delivery capabilities. Cloud delivery models may create multiple levels of SLA dependencies, as illustrated by the example in Figure 4 below.

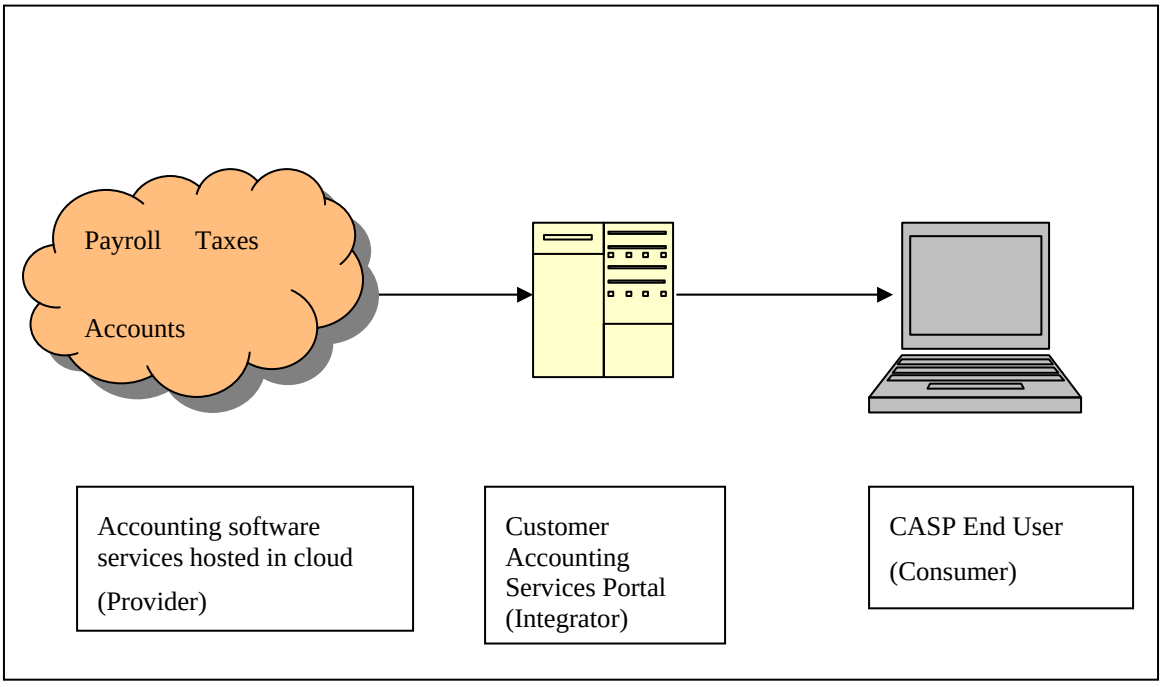

Figure 4. Cloud SLAs

In this example, the Integrator has put together a commercial offering allowing small businesses to obtain accounting services from the cloud through a portal. Consumers may choose to subscribe to one or many accounting services through the portal. The portal in turn will funnel requests from multiple Consumers to the Provider's cloud-hosted accounting services. This scenario creates the opportunity for at least two levels of SLAs.

1. The level of service to be provided by the Integrator to the Consumer;
2. The level of service to be provided by the Provider to the Integrator.

To guarantee a given level of service, the Provider must consider the demands which will be made upon the cloud by this Integrator (and possibly others) from a volume and processing perspective, as well as the architecture, design and the infrastructure used to implement the cloud solution. Moreover, during the assembly of the cloud solution, estimation, testing and measurement activities will be required on the part of the Integrator and/or Provider to finalize cloud SLAs, and ongoing monitoring will be required to ensure that the cloud SLAs continue to be met after deployment.

Typically the Integrator would also be responsible for guaranteeing a given level of service to one or more Consumer groups. However, except for the end user connectivity and portal technologies used, in this example the Integrator is dependent on the cloud SLAs being met by the Provider to meet the portal SLAs guaranteed to the Consumer. This concept of dependent SLAs is not new to the IT industry having first emerged when enterprises started outsourcing management of their networks to third parties.

While the service level responsibilities flow left to right in Figure 4, requirements information must flow right to left. The portal Integrator must collect or estimate usage information (volumes, transaction mixes, usage patterns and required responsiveness) from the Consumer community. Likewise, the cloud Provider must understand the demands on the cloud which may be coming from multiple Integrators and/or Consumers.

Once the SLAs are defined, the interactions between cloud components and their effects on performance must be considered both during solution integration (to confirm that SLAs can be met) and after solution deployment (to confirm that SLAs are being met). Using our previous example, if some of the SLAs involved response time, interactions such as the following would need to be considered both by the Integrator when devising performance budgets for the "round trip" and by the Provider when setting up performance monitoring.

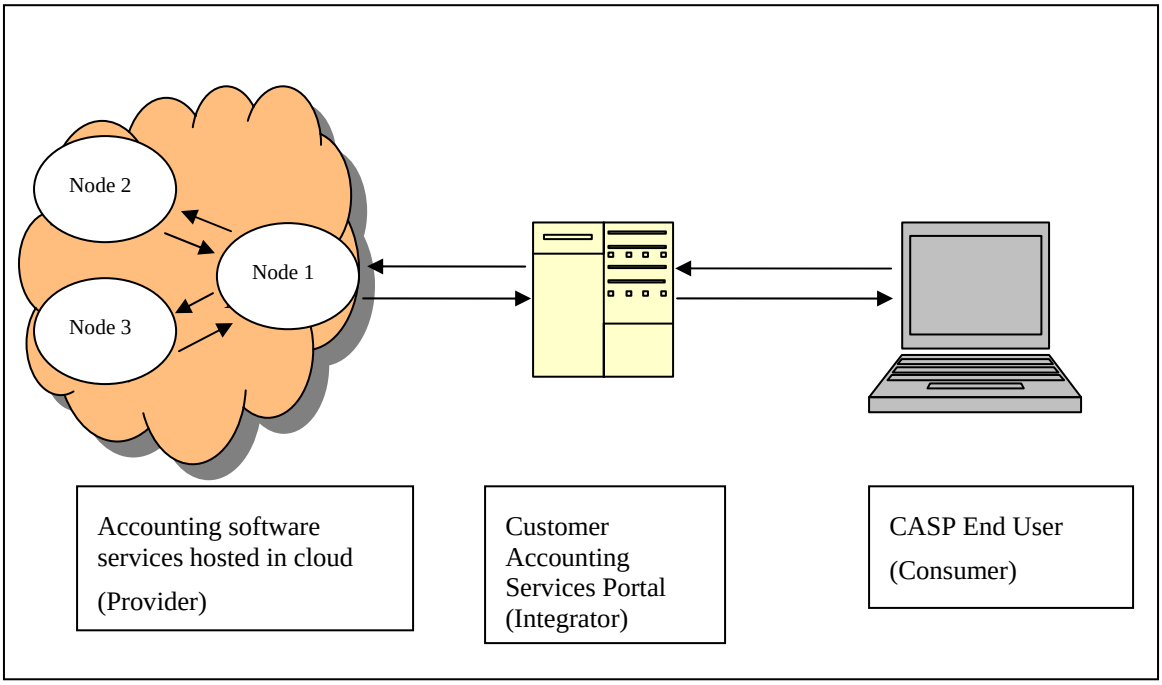

Figure 5 - Cloud interactions affecting SLAs

In short, the classic principles of performance engineering apply to the end-to-end cloud solution, but the responsibility for applying performance engineering methods to the solution must be divided based on the respective SLA responsibilities of the Provider and the Integrator.

Service level requirements for **performance** usually address

- Acceptable response times or elapsed times
- Requirements for throughput
- Capacity requirements, e.g. for network bandwidth or storage

Service level requirements for **availability** are related to performance, as lack of availability may be considered as the worst possible performance.

- Service window
- Required availability within the service window

To ensure stable performance in the long term, requirements for **scalability** and **reliability** will also have to be captured in the SLA.

Table 1 below shows each of these service level categories along with the key performance indicators (KPIs) that would typically be used to assess SLA attainment.

Service Level category	**KPI**	**Definition**	**Unit of measurement**
Availability	Service window	Time window within which KPIs are measured	Time range
	Service / System availability	Percentage of time that service or system is available	%
	MTBF	Mean time between failure	Time units
	MTTR	Mean time to repair	Time units
Performance	Response time	Response time for composite or atomic service	Seconds
	Elapsed time	Completion time for a batch or background task	Time units
	Throughput	Number of transactions or requests processed per specified unit of time	Transaction or request count
Capacity	Bandwidth	Bandwidth of the connection supporting a service	bps
	Processor speed	Clockspeed of a processor (CPU)	MHz
	Storage capacity	Capacity of a temporary or persistent storage medium, such as RAM, SAN, disk, tape	GB
Reliability	Service / System reliability	Probability that service or system is working flawlessly over time	%
Scalability	Service / System scalability	Degree to which the service or system is capable of supporting a defined growth scenario	Yes / No, or description of scalability upper limit

Table 1. Cloud Service Level Categories and Key Performance Indicators (KPI's)

The cloud delivery model determines what performance indicators must be covered in an SLA.

- If the delivery model is ***Business Process as a Service***, the service consumer is likely to have requirements regarding response times, throughput, availability, reliability and scalability of ***business processes.*** They will leave the derived capacity requirements to the provider.
- If the delivery model is ***Software as a Service (SaaS)*** or ***Platform as a Service (PaaS),*** the service consumer is likely to have requirements regarding response times, throughput, availability, reliability and scalability of ***transactions, supported by the software.*** They will leave the derived capacity requirements to the provider.
- If the delivery model is ***Infrastructure as a Service (IaaS)***, however, the service consumer is likely to have requirements regarding throughput, capacity, availability, reliability and scalability of ***infrastructure components.*** In this case the services to be provided on top of the cloud infrastructure are not in scope of the contract. Therefore, service response times or elapsed times cannot be addressed in the SLA.

The cloud deployment model determines how the SLA is offered to the consumer.

- If the cloud deployment model is ***public,*** the cloud provider may choose to offer either a fixed service level or a very simple set of service levels, ranging from e.g. bronze to gold, where better service is offered to consumers who are willing to pay more.
- If the cloud deployment model is ***private,*** the cloud consumer and provider may choose to agree upon a custom SLA.
- If the cloud deployment model is ***hybrid***, the SLA offering may be determined by a combination of the two methods.

In addition to making arrangements on the performance indicators that will be monitored, the SLA will have to address the financial implications of cloud usage. Naturally, cloud capacity is simultaneously used by multiple consumers. All consumers will be asked to pay for their own usage. As a result the cloud has to be equipped with usage based costing capabilities. The Consumer, Provider, and Integrator must provide data to ensure that these SLAs are in compliance. The activities are outlined below.

The Role of the Service Catalog in the Cloud

The service catalog documents the services available for delivery through the cloud. The service catalog consists of a list of standard services that can be requested from and provisioned by the provider, along with a standard price list for the services offered, typically based on QoS or usage metrics such as level of response, number of accesses per time unit, units of system processing or application usage time, or number of named or concurrent end users, and any other terms and conditions for usage of the services in the catalog.

The service catalog is part of a cloud management service that provides the user interface, business support systems providing the metering, billing and reporting, and operational support systems providing the system provisioning, monitoring, and support. For example IBM's Tivoli product family has tools such as Tivoli Service Automation Manager (TSAM), Tivoli Provisioning Manager (TPM), and IBM Tivoli Monitoring (ITM) which provide service catalog features as a central part of cloud management.

By documenting exactly what services the cloud has available for an individual consumer, the level of each QoS attribute available and requested, the price at which the service maybe obtained, and the terms and conditions of using the services on the cloud, the service catalog holds information uniquely defining the performance and capacity commitments within the contractual SLA for each consumer. Ensuring that these contractual SLA commitments can be met from a performance and capacity perspective by the compute cloud infrastructure providing the services requires a structured method of performance engineering and capacity management that we will introduce in the next section.

As shown in Figure 5, the service catalog is built by the cloud administrator and integrates the infrastructure, platform, and software services that exist within the cloud.

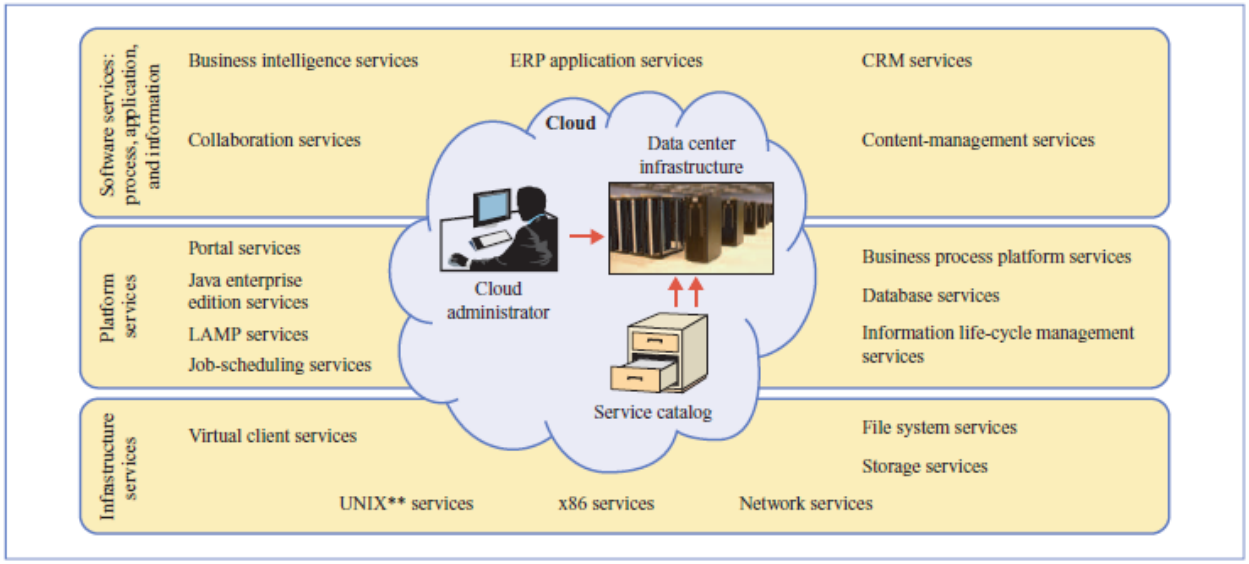

Figure 6. Service Catalog in the cloud computing infrastructure

Once services are added to the catalog, they are published and made available to the services consumers. This puts the service catalog at the core of the metering, billing, reporting, and operational support systems of the compute cloud. [7]

Activities to ensure SLA Compliance

Specific profile information is required from the Consumer on how the cloud is intended to be used in order to ensure SLA compliance. These include:

- A profile of the user population that will be using the services within the cloud such as the number of users, user types, and the number of concurrent users and hence service requests expected at the peak times.
- The number of Business Service executions or transactions expected to be performed at the peak times. This can either be expressed as a projection or actual numbers aggregated daily/monthly/annually which can be used in calculating the peak hour volumes.
- If requesting a Test Cloud, a profile of the test environments that details the required system components, how many environments are required, the number of people that would be accessing them and the times when these test environments would be required.
- Security information as it relates to systems hardware components, data and application system components including requirements on shared versus dedicated access required.

The Provider needs to assess the requirements of the consumer and develop a solution that will meet those requirements over the course of the project lifecycle. From a performance and capacity perspective this solution will include:

- The creation of the appropriate test environments from the virtualized pool that meets the required hours of operations.
- Execution of performance tests to establish the behavior of the application and the footprint it would have in the cloud production environment.
- An estimate of planned resources required within the Cloud along with Performance Objectives which would be confirmed during the life cycle process in the Performance Model initially and ultimately in the Deployment Model.
- The system resources required in production based on performance modeling. Testing should be identified in terms of a graduated scale based on a baseline resource unit and the required number of additional resource units that coincides with horizontal increments of resources at each tier. The Deployment Model contains sizing and configuration information that is required when the application is migrated to the

production cloud and provides confirmation on the resource and capacity selections initially from the Services Catalog offerings.

The Integrator has to confirm that all the system components required for the solution are included and verify that the solution as proposed by the provider works and will meet the performance and service level objectives. This confirmation includes:

- Confirming that the proposed test systems environments will satisfy the needs of the requester.
- Validating the contents of the Deployment Model and the proposed solution for Production.
- Coordinating the activities for the production migration/cutover.

The next section of the paper outlines several use case scenarios to clarify real business situations and highlights where performance to consider as a part of the cloud solution strategy.

Cloud Computing Scenarios and their Performance Implications

Use case scenarios help to clarify real business situations and highlight where performance considerations are part of the cloud solution strategy. Scenarios involving private, public, and hybrid cloud deployment models are described below.

Scenario #1 - Private Cloud Deployment

An internet company must implement a strategic new business model without increasing operational costs. The current manual processes have led to unexpected downtimes and system outages. This internet company is anticipating storage capacity increases from hundreds to thousands of servers in order to accommodate anticipated business growth, volume increases, and automation.

Solution Strategy

The solution strategy was to introduce cloud gradually and maintain an open and adaptable cloud management platform. The guiding principles were to:

- Start with internal/private cloud first then grow into a public cloud.
- Generate an open & adaptable cloud management platform with support of heterogeneous hardware.

Solution Details

- Implement Tivoli Service Automation Manager (TSAM) out of the box to provide an end to end service management layer, and automated deployment & management along with a virtualized infrastructure.
- Deploy virtual servers with LAMP (Linux, Apache) Middleware to a specified hardware, operating system (OS), network and storage configuration.
- Provide virtualization hypervisor on bare-metal hardware.
- Manage/scale infrastructure/platform services (virtual servers/LAMP).

Performance Considerations

- Consider service level agreements and work load optimization in order to manage capacity and prevent bottlenecks from occurring within the private cloud.
- Organize work loads so that they can scale to support anticipated and unanticipated growth using techniques that will optimize use of storage such as virtualization.
- Consider the skill-set of teams as well as the systems management capabilities to help drive service level performance.
- Consider security implications.

Scenario #2 - Public Cloud Deployment

Desktop Cloud (Refer to "Smart Business Development & Test on the IBM Cloud" in the next section)

A sales organization wants to reduce the management and support costs associated with remote mobile laptop users. New sales team members need to be quickly implemented with a standard suite of office productivity and sales tools.

Solution Strategy

The solution strategy was to contract with the Desktop Cloud provider to provide

- Standard Desktop image
- Standard tools plus custom software built into the image
- Desktop data storage options

The solution is a pre-priced, pre-packaged subscription service, with some limited customization to keep costs down.

Total cost is based on metered usage, so the sales organization will only pay for the time they use.

Solution Details

- The Desktop image is accessible from any remote device over Remote Desktop Protocol (RDP) or Independent Computing Architecture (ICA) protocol
- The Desktop image is accessible from corporate network or public internet.
- The public internet access connects to the Desktop Cloud through a dedicated security appliance to ensure data security for corporate data.
- The Desktop Cloud service, infrastructure, and availability is managed by the provider.
- The Standard Desktop image can be modified by the consumer.

Performance Considerations

- The performance of remote desktops is dependent on the software installed and used on the remote desktop.
- The performance of remote desktops is highly dependent on the types of data remote users will be working with on the remote desktop.
- The performance of remote desktops is dependent on the network connection bandwidth and latency between the consumer point of entry and the location of the desktop cloud.
- The Customer will need to consider the type of device their users will use to access the Remote Desktop – desktop, laptop, netbook, thin client device. Each will have different performance, security, and usability impacts.
- Service level agreements may be difficult to define and enforce for a public cloud scenario. Customers with particularly high application service requirements for performance and capacity may best be directed to either the private or hybrid cloud scenario.

Scenario #3 - Hybrid Cloud Deployment

Hybrid Development Cloud (see solution offering: Smart Business Development & Test on the IBM Cloud - http://www.ibm.com/ibm/cloud/smart_business)

An application development company is investigating new development tools, and wants to "try before they buy" to see if the new tools fit well in their existing tool set. They need quick turnaround, and don't want to buy servers and software licenses that may never be used in their actual development environment.

Solution Strategy

The solution strategy was to contract with a development cloud service provider.

The development cloud is hosted by the service provider, and includes a suite of images with development and test tools preinstalled and configured that the consumer can access

The consumer also selects the VPN option so that the image instances that they create have IP addresses, host names, and access security rules that allow them to access the instances as if they were in their own data center.

Solution Details

- The consumer requests instances using the self-service portal, then configures the development tools to integrate with their existing tool set.
- Total cost is based on metered usage of the development tool license, so the sales organization will only pay for the time they use.
- If the integration is successful, the consumer can choose either to continue using the development tool through the development cloud instance, or purchase their own license and install the tool in their own data center infrastructure.

Performance Considerations

- The performance will be highly dependent on the bandwidth and latency between the service provider data center and the consumer data center.
- Authentication access integration between the development cloud instances and the consumer tool set will impact performance. If the development cloud instance must be authenticated by an enterprise authentication directory service in the consumer data center, this will impact performance.
- Some development tools will require access to shared databases for code sharing, for example. If these databases are remote from the development tool, this will impact performance.
- If the consumer is considering moving the development tool from the development cloud into its standard toolset, they will need to verify that the performance is adequate for daily usage by a larger user base than the evaluation user generated.
- Service level agreements need to consider the level of service between the cloud provider and the application integrator, and between the application integrator and the end-user Consumer of the cloud service.

Solution Offering Considerations for the Cloud

A large variety of cloud computing solution offerings are available on the market from the major vendors today. A common theme across all those offerings is that the type of service being offered highlights the products and services the cloud computing vendor specializes in. For example:

- Oracle/Sun and HP's cloud offerings focus on hardware and services to enable cloud construction
- Google's cloud computing offering is Google Apps, a pure software-as-a-service offering
- HP's cloud offering highlights its system management strengths as a software-as-a-service offering
- IBM and Amazon offer a mixture of platform-as-a-service cloud offerings and more complete turnkey public, private, or hybrid cloud computing packages.

The technical components, business focus, and marketing direction that result from these cloud computing vendor decisions directly impact the performance implications of the underlying cloud computing solution and need to be analyzed and evaluated appropriately.

Cloud offerings from IBM

The offerings discussed in this section are part of the IBM Smart Business cloud services overview. Offerings include:

CloudBurst

IBM CloudBurst V1.2 is a prepackaged and self-contained service delivery platform that can be easily and quickly implemented in a data center environment. IBM CloudBurst V1.2 is positioned for enterprise customers looking to get started with a private cloud computing model. It is ideal for those customers looking for both a solution to complement the customers existing IT infrastructure and assistance in efforts to realize a payback from their cloud investment in a challenging economic environment.

Source: IBM United States Software Announcement 209-402 November 3, 2009

Deployment Model	Private Cloud
Roles Focus	Provider
Performance Considerations	• Work load organization to support anticipated and unanticipated growth • Consider the skill-set of teams to support the CloudBurst platform • Plan systems management capabilities and integration with other infrastructure • There may be security integration implications.

Smart Business Development & Test on the IBM Cloud

The IBM Smart Business Development and Test on the IBM Cloud is a dynamically provisioned and scaled runtime environment that provides everything needed to develop and test application code highlighting IBM's software offerings. This includes tools to configure and manage the dynamic execution environment, an IDE that facilitates the direct use of the execution environment and build and test tools that can exploit the execution environment.

Source: Smart Business Development & Test on the IBM Cloud--About

Deployment Model	Public/Hybrid Cloud
Roles Focus	Provider
Performance Considerations	• Performance will be dependent on bandwidth/latency between the data centers.

- Authentication between cloud and customer tools will impact performance. .
- Development tools may need access to shared databases
- Customer will need to verify performance for daily usage by developers.

Smart Business Desktop Cloud

IBM Smart Business Desktop Cloud provides access to applications, information and resources through thin clients or any other Internet-connected device. The Smart Business Desktop Cloud delivers a resilient, efficient, standards-based IT infrastructure for almost any traditional desktop application, employing portal, thin client, messaging, and security technologies, delivered through a single, consistent framework. All that's needed is a machine—a thin client or PC—capable of running an Internet browser and JavaTM.

Source: Smart Business Desktop Cloud

Deployment Model	Public Cloud
Roles Focus	Consumer
Performance Considerations	• Performance is dependent on the software on the remote desktop. • Performance is dependent on data users will be using on the remote desktop. • Performance is dependent on the network connection bandwidth and latency. • The access device (PC/laptop/netbook) will have different performance impacts.

Tivoli Service Automation Manager

Tivoli Service Automation Manager (TSAM) provides the capability to request, deliver, and manage IT services. It is a strategic cross-IBM solution for the operational support systems necessary to help enterprise data centers benefit from cloud computing. Tivoli Service Automation Manager is designed to enable faster IT response and delivery capabilities, and help lower IT operational costs.

Source: IBM Tivoli Service Automation Manager—Fact sheet

Deployment Model	Private/Hybrid Cloud
Roles Focus	Developer / Integrator
Performance Considerations	• TSAM is a building block for private clouds. • TSAM workflows will be a key component of performance. • Disk and networks will need to be optimized for performance for the workloads.

Cloud offerings from Amazon

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers. Amazon EC2 allows users to obtain and configure capacity with minimal friction. It provides complete control of resources running on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes. Amazon EC2 allows users to pay only for capacity actually used. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

Source: Amazon Elastic Compute Cloud (Amazon EC2)

Deployment Model	Public Cloud
Roles Focus	Provider / Integrator
Performance Considerations	• Performance will be dependent on bandwidth/latency between the data centers. • Authentication between cloud and consumer tools will impact performance. • Development tools may need access to shared databases • Consumer will need to verify performance for daily usage by developers

Cloud offerings from Oracle/Sun

Sun combines open systems, software, and architectural expertise to build clouds and maximize their capabilities. Responding to the demand for interoperability, Sun is pursuing its vision of open, interoperable clouds in four key areas: open standards-based software, open systems hardware for interoperability, microelectronics with multi-threading and multi-core computing to enable higher compute densities, and professional services and systems integration to leverage the benefits of cloud computing.

Source: Sun Cloud Computing

Deployment Model	Private Cloud
Roles Focus	Provider / Integrator
Performance considerations	• Sun offerings are building blocks for private clouds • Performance will need to be engineered into the cloud • Open hardware/software increase flexibility, not necessarily performance

Cloud offerings from Google

Google's web-based messaging and collaboration applications require no hardware or software and need minimal administration, creating time and cost savings for businesses. End users can use mobile email, calendar and Information Management access, with several options for accessing their information while on the go. Google guarantees that Google Apps will be available at least 99.9% of the time, and each employee gets 25 GB for email storage, so they can keep important messages and find them instantly with built-in Google search.

Source: Google Apps Benefits

Deployment Model	Public Cloud
Roles Focus	Consumer
Performance Considerations	• Performance is dependent on data users will be using in the Google Apps. • Performance is dependent on the network connection bandwidth and latency. • The access device (PC/laptop/netbook) will have different performance impacts.

Cloud offerings from Hewlett-Packard

HP Cloud Assure consists of HP services and software, including HP Application Security Center, HP Performance Center and HP Business Availability Center. HP will also provide engineers to perform security scans, execute performance tests and deploy availability monitoring. HP Cloud Assure helps customers validate security by scanning networks, operating systems, middleware layers and web applications and performing penetration testing,

performance by making sure cloud services meet end-user bandwidth and connectivity requirements and satisfy service-level agreements, and availability by monitoring cloud-based applications to isolate potential problems and identify root causes with end-user environments and business processes and to analyze performance issues.

Source: HP Unveils "Cloud Assure" to Drive Business Adoption of Cloud Services

Deployment Model	Private Cloud
Roles Focus	Provider/ Integrator
Performance Considerations	• HP offerings are building blocks for private clouds • Performance will need to be engineered into the cloud • HP performance and systems management tools are strong part of offering

Conclusion

Cloud computing provides an efficient, scalable, and cost-effective way for today's organizations to deliver business and consumer IT services over the Internet. A variety of different cloud computing models are available, providing both solid support for core business functions and the flexibility to deliver new services. Performance considerations must be addressed throughout the lifecycle of the cloud based solutions if they are to deliver Quality-of-Service levels that are typical of well constructed enterprise applications, the responsibility for undertaking those roles splitting across the roles of Consumer, Provider and Integrator.

There are clear advantages to be gained today by exploiting cloud technology for specific purposes and many organizations are already realizing benefits. For example offerings such as Test Cloud enable test environments to be provisioned more quickly thereby reducing the timeframe for deploying new applications into production. Whilst those environments may be fine for supporting functional testing and even some level of non functional testing the Cloud is yet to mature sufficiently to be proposed as the solution for all non functional testing needs.

This paper has provided an overview of cloud computing and benefits, discussed the different cloud computing perspectives, highlighted performance engineering interfaces with cloud solutions, and detailed performance considerations for the cloud. Usage scenarios were shown for clarification and elaboration of real-life cases where performance considerations are part of the cloud solution strategy. Examples of offerings were provided to assist in the development of performing cloud systems.

Large enterprises deploying their mission critical applications over the cloud are still in the future. However, the current and emergent trends within our industry are fueling the demand for truly global business services that are agile and responsive to the needs of the marketplace and can be invoked from anywhere at anytime. The consequence is that the potential demand for successful services will take a quantum leap beyond that of more traditional commercial IT systems. This scenario has significant implications for how this new breed of services must be built and deployed. Quality of Service is already becoming an increasingly important differentiator in today's global market and the performance and scalability of the IT systems underpinning those business services will be vital in enabling organizations to meet those demands.

The challenge posed for Performance Engineering is substantial and the discipline will need to continue to innovate in order to meet that challenge. Industry watchers are predicting that Performance Engineering will become even more critical to the success of the IT industry in the future.

Acknowledgements

The authors would like to thank Pirooz Joodi, Damian Towler, Lewis Grizzle, Stefan Pappe and Ann Dowling for reading draft versions of this paper and providing many valuable comments.

References

[1] Cloud Security Guidance - IBM Recommendations for the Implementation of Cloud Security
http://www.redbooks.ibm.com/redpapers/pdfs/redp4614.pdf

[2] Cloud Computing: Save Time, Money, and Resources with a Private Test Cloud
http://www.redbooks.ibm.com/redpapers/pdfs/redp4553.pdf

[3] Cloud Computing
http://www.ibm.com/ibm/cloud

[4] National Institute of Standards and Technology – Notational Definition of Cloud Computing
http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc

[5] Forrester Research “Major Hurdles Remain In Enterprise Cloud Services”
http://www.forrester.com/rb/Research/major_hurdles_remain_in_enterprise_cloud_services/q/id/54780/t/2

[6] IBM Smart Business Cloud Computing
http://www.ibm.com/ibm/cloud/smart_business/

[7] “Life cycle and characteristics of services in the world of cloud computing” by G. Breiter and M. Behrendt, IBM Journal of Research and Development, V. 53, Number 4, Paper 3 2009.

[8] “Above the clouds: A Berkeley View of Cloud Computing”
Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, Daviod Patterson, Ariel Rabkin, Ion Stoica and Matei Zaharia
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf

[9] IDC, "Worldwide Disk Storage Systems 2008-2012 Forecast: Content- Centric Customers- Reshaping Market Demand," Doc # 212177, May 2008. Funda has permission from "Suzanne Hopkins" <SHopkins@idc.com> 07/15/2008 02:23 PM

]10] IBM Global Technology Outlook for 2005

IBM Corporation
Systems and Technology Group
Route 100
Somers, New York 10589

Produced in the United States of America
May 2010

This document was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this document in other countries.

The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only.

IBM, the IBM logo, ibm.com, POWER, POWER5, POWER6, Power Systems, and PowerVM are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml

Other company, product, and service names may be trademarks or service marks of others.

IBM hardware products are manufactured from new parts, or new and used parts. In some cases, the hardware product may not be new and may have been previously installed. Regardless, our warranty terms apply.

Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM.

This equipment is subject to FCC rules. It will comply with the appropriate FCC rules before final delivery to the buyer.

Information concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of the non-IBM products should be addressed with those suppliers.

All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM. Buyers should consult other sources of information, including system benchmarks, to evaluate the performance of a system they are considering buying.

When referring to storage capacity, 1 TB equals total GB divided by 1000; accessible capacity may be less.

The IBM home page on the Internet can be found at: http://www.ibm.com.

The IBM Power Systems home page on the Internet can be found at: http://www.ibm.com/systems/power/

The Power Architecture and Power.org wordmarks and the Power and Power.org logos and related marks are trademarks and service marks licensed by Power.org.

UNIX is a registered trademark of The Open Group in the United States, other countries or both.

Linux is a trademark of Linus Torvalds in the United States, other countries or both.

Microsoft, Windows, Windows NT and the Windows logo are registered trademarks of the Microsoft Corporation.

Intel, Itanium and Xeon are registered trademarks and MMX and Pentium are trademarks of Intel Corporation in the United States and/or other countries.

AMD Opteron is a trademark of Advanced Micro Devices, Inc.

Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. In the United States and/or other countries.

TPC-C and TPC-H are trademarks of the Transaction Performance Processing Council (TPPC).

SPECint, SPECfp, SPECjbb, SPECweb, SPECjAppServer, SPEC OMP, SPECviewperf, SPECapc, SPEChpc, SPECjvm, SPECmail, SPECimap and SPECsfs are trademarks of the Standard Performance Evaluation Corporation (SPEC).

SAP, mySAP and other SAP product and service names mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries all over the world.

SPC Benchmark-1 and SPC Benchmark-2 are trademarks of the Storage Performance Council.