Performance Challenges in Cloud Computing

Shailesh Paliwal

Cloud computing is important for the today's demanding business requirements. The cloud computing concept, with its salient features, and the three Cloud Service delivery models are explained here. The three cloud delivery models of Software as a Service (SaaS), Platform as a service (PaaS) and Infrastructure as a Service (IaaS) are explored with their Inter-dependencies and performance considerations. Cloud adoption in the business has performance obstacles, and suggestions to overcome these obstacles are provided while suggesting performance consideerations for the three cloud delivery models. Performance considerations are vital for the overall success of cloud computing, including the optimum cost of cloud services, reliability and scalability. They require a lot of attention and efforts by the cloud computing providers, integrators and service consumers.

WHY CLOUD COMPUTING?

It is simply difficult to manage today's complex businesses environments by traditional IT solutions. Some reasons are:

- **Explosive growth in applications:** Web 2.0 social networking, YouTube, Facebook, biomedical informatics, space exploration, and business analytics
- **Extreme scale content generation:** e-science and e-business data deluge
- **Extraordinary rate of digital content consumption:** digital gluttony: Apple iPhone, iPad, Amazon Kindle
- **Exponential growth in compute capabilities:** multi-core, storage, bandwidth, virtual machines (virtualization)
- **Very short cycle of obsolescence in technologies:** Windows Vista to Windows 7; Java versions; C to C#; Phython
- **Newer architectures:** web services, persistence models, distributed file systems/repositories (Google, Hadoop), multi-core, wireless and mobile.

What is cloud computing and why is it distinctive?

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

It Involves shifting the bulk of the costs from capital expenditures (CapEx), or buying and installing servers, storage, networking, and related infrastructure to an operating expense (OpEx) model, where you pay for usage of these types of resources.

Cloud computing is unique because of its distinct general characteristics:

- **Multi-tenancy:** Public cloud service providers often host the cloud services for multiple users within the same infrastructure
- **Elasticity and scalability:** Ability to expand and reduce resources according to your specific service requirement. e.g., you may need a large number of server resources for the duration of a specific task. You can then release these server resources after you complete your task.
- **Pay-per-use:** Pay for cloud services only when you use them, either for the short term (e.g., for CPU time) or for a longer duration (e.g., for cloud-based storage or vault).
- **On demand:** One can invoke cloud services on need basis, need not to be part of IT infrastructure—a significant advantage for cloud use as opposed to internal IT services.
- **Resiliency:** Cloud can completely isolate the failure of server and storage resources from cloud users. (Work can be migrated to a different physical resource in the cloud with or without user awareness and intervention.)
- **Workload movement:** It is important for resiliency and cost considerations, service providers can migrate workloads across servers — both inside the data center and across data centers (even in a different geographic area). Typical reasons for workload movement are due to a catastrophic event in a geographic region (say Hurricane Sandy in the US). Then the corkload can be moved to some other geographic region for the time being, or there can be some other business drivers for the workload movement to get these advantages:
    - **Less cost** - It is less expensive to run a workload in a data center in another area based on time of day or power requirements.
    - **Efficient** –Better resources / network bandwidth availability. For example, US nightly processing in India day time is less costly and more efficient.

- **Regulatory considerations** - For certain types of workloads, e.g. New York stock exchange processing from India.

CLOUD SERVICE DELIVERY MODELS AND THEIR PERFORMANCE CHALLENGES

Cloud service delivery models

Software as a Service (SaaS)

Enterprises will have software licenses to support the various applications used in their daily business. These applications could be in human resources, finance, or customer relationship management. The traditional option is to obtain the desktop and server licenses for the software products used.

Software as a Service (SaaS) allows the enterprise to obtain the same functions through a hosted service from a provider through a network connection. Consumer services include social platforms (e.g. Facebook) or online email services (e.g. Gmail). There are also increasing numbers of business services being delivered as-a-service (e.g. software package rendering through VDO / Citrix server to the mass public).

- Centralized services typically designed to cater for large numbers of end users over Internet.
- SaaS reduces the complexity of software installation, maintenance, upgrades, and patches for the IT team within the enterprise, because the software is now managed centrally at the SaaS provider's facilities.
- SaaS providers are responsible to monitor the application-delivery performance.

Platform as a Service (PaaS)

Unlike the fixed application functionality offered by SaaS, Platform as a Service (PaaS) provides a software platform on which users can build their own applications and host them on the PaaS provider's infrastructure (e.g. Google with its App-Engine or Force.com APIs).

- The software platform is used as a development framework to provide services for use by applications.
- PaaS is a true cloud model in that applications do not need to worry about the scalability of the underlying hardware and software platform.

- PaaS providers are responsible to monitor the application delivery performance elasticity and scalability.

Infrastructure as a Service (IaaS)

An Infrastructure as a Service (IaaS) provider offers you raw computing, storage, and network infrastructure so that you can load your own software, including operating systems and applications, on to this infrastructure (e.g. Amazon's Elastic Computing Cloud (EC2) service).

- This scenario is equivalent to a hosting provider provisioning physical servers and storage, and letting you install your own OS, web services, and database applications over the provisioned machines.
- Greatest degree of control of the three models, resource requirement management, is required to exploit IaaS well.
- Scaling and elasticity are user's responsibility and not the provider's responsibility.

INTER-DEPENDENCIES OF DELIVERY MODELS AND THEIR PERFORMANCE MEASURES

The cloud can also be defined as the virtualized infrastructure found on the lowest level of the solution stack (i.e. IaaS layer as in Figure 1). The higher service layers depend on the underlying supporting service layers. Service providers can be service users as well:

- A SaaS provider may 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

Summary - Characteristics of Performance Measures

- SaaS performance measures are directly perceived by users as business transaction response times and throughput, technical service reliability and availability, and by scalability of the applications.
- PaaS performance measures are indirectly perceived by users and defined as technical transaction response times and throughput, technical service reliability and availability, and by scalability of the middleware.
- IaaS Performance Measures are defined as infrastructure performance, capacity, reliability, availability, and scalability.
- In general, characteristics of performance measures of the upper service layers depend on those characteristics in the underlying

layers, e.g. SaaS layer scalability depends on IaaS layer scalability.

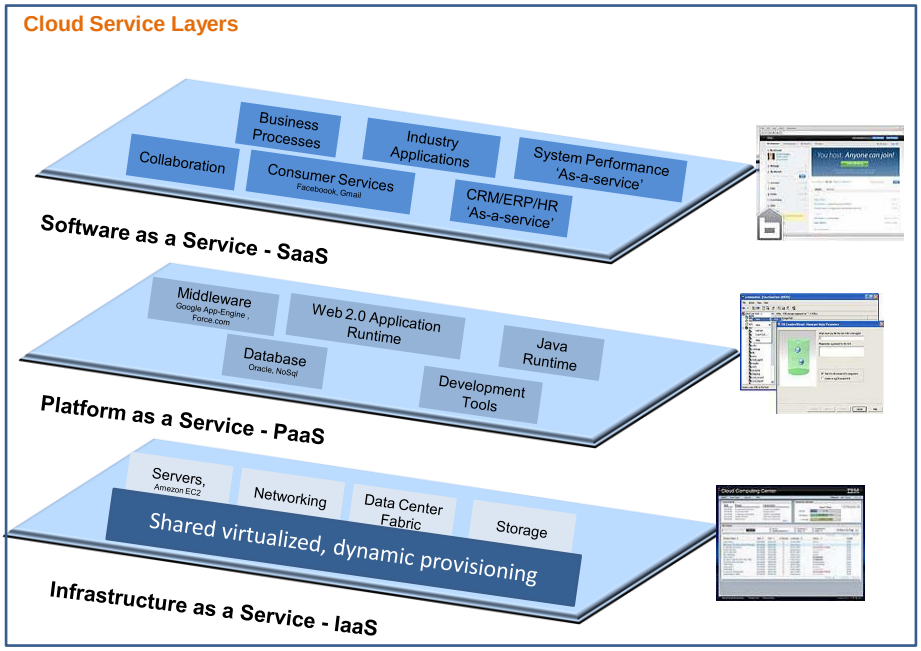

Figure 1 IBM's Cloud Service Layers Diagram

PERFORMANCE ASPECTS ARE MAJOR OBSTACLES IN CLOUD ADOPTION AND GROWTH

The success of Cloud deployments is highly dependent on practicing holistic performance engineering and capacity management techniques.

A majority of the obstacles for adoption and growth of cloud computing are related to the basic performance aspects, such as availability, performance, capacity, or scalability. Please refer below to Table 1 for the obstacles and opportunities details.

- Potential cloud solutions to overcome these obstacles need to be carefully assessed for their authenticity in real-life situations.
- Performance engineers need to get to the bottom of the technical transactions of underlying cloud services before advising cloud computing users and cloud computing providers for the cloud services.
- The degree to which cloud services can meet agreed service-level requirements for availability, performance, and scalability

can be estimated by using performance modeling techniques, so that potential performance anti-patterns can be detected before they happen.

- In the absence of sophisticated tooling for automated monitoring, the automatic provisioning and usage-based costing (metering) facilities, rely mainly on fine-grained capacity management. Until more data collection, analysis, and forecasting are in place, capacity management is more opportune than ever.
- Irrespective of sophisticated tooling for automated monitoring, cloud computing users need to analyze their demand for capacity and their requirements for performance. In their contract with cloud computing providers, users should always take a bottom-line approach to accurately formulate their service-level requirements.

Table 1: Top Ten Obstacles and Opportunities for Adoption and Growth of Cloud Computing – Performance view;
Reference : A Berkeley View of Cloud Computing

	Obstacle	**Opportunity**
1	Availability of Service	Ensure Business Continuity by defining firm SLAs and Fault tolerance through multiple cloud providers in place. Use Elasticity (scale resources both up and down as needed) to Defend Against Distributed Denial of Service (DDoS) attacks
2	Data Lock-In	Standardize APIs; Make compatible software available to enable Surge Computing
3	Data Confidentiality and Audit-ability	Deploy Encryption, VLANs, and Firewalls; Accommodate National Laws via Geographical Data Storage
4	Data Transfer Bottlenecks	FedExing Disks; Lower WAN Router Costs; Higher Bandwidth LAN Switches Tools which can deliver high-speed transfers in, out and across the cloud with scale-out transfer capacity on web, mobile or embedded in application itself.
5	Performance Unpredictability	Improved Virtual Machine Support; Flash Memory; Gang Scheduling VMs for HPC apps Performance A number of as-a-service performance predictability solutions are in place and growing for Holistic Cloud performance engineering and capacity management
6	Scalable Storage	Invent Scalable Store Cloud solutions for high-performance, scalable storage-virtualization to facilitate growth and innovation at lower operational costs.
7	Bugs in Large-Scale Distributed Systems	Invent Debugger that relies on Distributed VMs
8	Scaling Quickly	Invent Auto-Scalar that relies on Machine Learning; Snapshots to encourage Cloud Computing Conservationism A number of as-a-service solutions are in this research focus area to address the dynamic scaling (scale up, out and down) needs
9	Reputation Fate Sharing	Offer reputation-guarding services like those for email
10	Software Licensing	Pay-for-use licenses; Bulk use sales

Leveraging as-a-service technologies for cloud performance considerations

Performance Monitoring can itself be as-a-service

With new technologies and middleware platforms such distributed file systems, MongoDB databases, Search platforms as well as for very heavy systems processing Big Data, there is a constant need for performance monitoring and analysis techniques to be developed. These monitoring and analysis techniques need to ensure that performance metrics can be obtained, analyzed and understood in the context of these new technologies.

- New centralized monitoring techniques will be required specifically for these new technologies and middleware platforms.
- The solution is to devise new as-a-service for monitoring and management of the cloud.
- This means that tool providers will centrally store monitoring data from large numbers of customers systems along with new opportunities in terms of data analytics.

Automated Resource Utilization Monitoring and Analysis

A new requirement coming in with cloud technologies and middleware platforms is the availability of automated utilization metrics. These metrics should be efficiently collected and properly understood.

- A major challenge for cloud providers is to centrally monitor the hardware is being utilized with the changing load on the system. This is required to harness the power of existing hardware and maximize the efficiency of cloud infrastructures.
- Lots of research is being conducted in this area of utilization analysis in the context of different software workloads. Such analysis can be applied say to maximize the system utilization by workload relocation or to increase energy efficiency.
- Analysis and monitoring can dramatically reduce the costs of services by providing more cost-optimized cloud platforms and services.

Horizontal Scaling

Horizontal scaling, or scale out, usually refers to clustering multiple independent computers together to provide more processing power. This type of scaling typically implies multiple instances of operating systems, residing on separate servers.

- SaaS requires very dynamic horizontal scaling, i.e. the ability to quickly scale out and down during times of different workloads.
- Performance considerations such as scalability and reliability are an important area for SaaS systems. It can even be more challenging on large scale SaaS systems with large numbers of components.
- For dynamic scale out, hardware resources may be immediately available, but with a huge financial cost associated, and may not be an effective and elegant solution in case of inefficient design.
- The financial cost implications become even worse in SaaS systems, as in non-SaaS systems the cost of running inefficient hardware was capped by the available hardware resources in-house (generally the CapEx expenses). In the cloud this is no longer the case, and developers and designers are now closer to the financial costs associated with running their software. Thus responsible design of software with respect to performance is required so that efficient usage of the cloud is attained.
- Autonomic management of systems has been a growing area of research over the past decade. Automatic scaling based on alerting and user defined thresholds is something available today from as-a-service providers so that system will scale on demand.
- Pre-defined performance non-functional requirements (NFRs) and service level agreements (SLAs), workload modeling, user load planning for minimum, average and maximum users, and scalability testing are the best way to proactively handle the performance issues. Table 2 below shows typical Cloud SLAs and KPIs that are used to assess SLA attainment.

Table 2 Typical Cloud SLAs and KPIs; Reference:
IBM's Cloud service-level categories and key performance Indicators table

Service-level category	KPIs	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	Meantime between failure	Time units
	MTTR	Meantime 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 that supports a service	Bps
	Processor speed	Clock speed of a processor	MHz
	Storage capacity	Capacity of a temporary or persistent storage medium, such as RAM, SAN, disk, or 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 can support a defined growth scenario	Yes/No, or description of scalability upper limit

Advanced Data Analytics for better performance

An individual enterprise may produce terabytes of log data per month which can contain millions of events per second. The techniques for gathering monitoring data have become a lot better through the development of performance tools for in-house enterprise systems, however the analysis of the large volume of data collected has been still a major challenge.

- There is a burning need of having efficient and preferred log analytics systems in cloud environment, with the surfacing of new cloud technologies on these challenges such as log management as-a-service.

- A log management as-a-service technology handling log analysis for large numbers of enterprises must be able to manage millions of events per second, performing visualization, analysis and alerting in real time to allow for autonomic management of the system.
- Cloud has produced new challenges due to the larger scale of systems and the much larger volumes of data produced by these systems. Real time analytics is a growing area and provides challenges in the analysis of upwards of millions of events per second with real time constraints.
- Real time analytics can be a BIG aid for performance monitoring; this is another emerging rich area of research. Real time analytics with time constraints will certainly enhance the performance management of cloud based systems.

SUMMARY

This paper has emphasized the importance of cloud computing to fulfill computing needs of today's complex business scenarios, its benefits such as shift from CapEx to OpEx, elasticity and scalability, pay-per-use, on-demand, resiliency, workload movement and multi-tenancy. Also presented were the different cloud computing perspectives, highlighting performance engineering with various cloud models, and detailed performance considerations for the cloud. The standard cloud service delivery models, namely SaaS, PaaS and IaaS, have their own performance challenges as well as inter-dependencies for implementation, and to performance of one model to another.

Performance aspects are highlighted that are major obstacles in cloud adoption and growth. Business opportunities for new enterprise or cloud services exist, for example having centralized performance monitoring and tooling for the automation of cloud performance and capacity management (for example a central monitor showing cloud end-to-end enviroment performance statistics). Holistic performance engineering practices are vital in enterprise (non cloud) environment and in cloud environments.

The core performance engineering and capacity management practices, workload modeling, NFRs and SLA definitions and Performance modeling must be done in the cloud environment. The cloud service providers, integrators and consumers are the stakeholders for cloud deployments with their varying interest and goals. In summary, the challenge posed for performance is substantial and discipline needs to continue to innovate to meet that challenge. Industry watchers (such as Gartner) are predicting that Performance Engineering becomes even more critical to the success of the Cloud based IT industry.

References

1. Michael Armbrust: Above the clouds: A Berkeley View of Cloud Computing - http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf
2. Borthakur, D. et. Al.: Apache Hadoop Goes Realtime at Facebook. In Proceedings of the International Conference on Management of Data (2011)
3. Tate, B., Clarke, M., Lee, B., Linskey, P.: Bitter EJB. Manning (2003)
4. Parsons, T. and Murphy, J.: Detecting Performance Antipatterns in Component Based Enterprise Systems, Journal of Object Technology, 7 (3) pp. 55-90 (2008)
5. Dobson, S., Sterritt, R., Nixon, P., Hinchey, M.: Fulfilling the Vision of Autonomic Computing. Computer 43 (1) pp. 35-41 (2010)
6. John Murphy : Performance Engineering for Cloud Computing : http://www.csi.ucd.ie/staff/jmurphy/publications/1731.pdf
7. Parsons, T., Mos, A., Trofin, M., Gschwind, T. and Murphy, J.: Extracting Interactions in Component Based Systems. IEEE Transactions on Software Engineering, 34 (6) pp. 783-799 (2008)
8. Kozioleka, H.: Performance evaluation of component-based software systems: A survey.
9. Elisabeth S., Lydia D., Avin F., Pamela K., Dave J., Martin J., Todd R.: Performance Implications of Cloud Computing http://www.redbooks.ibm.com/redpapers/pdfs/redp4875.pdf
10. Performance Evaluation 67 (8) pp. 634-658 (2010) Chef, Systems integration framework, http://www.opscode.com/chef/
11. Apache Lucene, http://lucene.apache.org/java/docs/index.html
12. Oltsik, J.: The invisible log data explosion" Cnet.com (2007)
13. Log Math, http://chuvakin.blogspot.com/2010/08/log-math.html
14. Swrve, Real time gaming analytics, http://swrve.com/