2009 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises

Cloud Computing and the Lessons from the Past

Dr. Rao Mikkilineni and Vijay Sarathy
Kawa Objects, Inc.
Los Altos, CA
e-mail: rao@kawaobjects.com, vijay@kawaobjects.com

Abstract — **The skyrocketing demand for a new generation of cloud-based consumer and business applications is driving the need for next generation of datacenters that must be massively scalable, efficient, agile, reliable and secure.**

The authors see a parallel between the state of the datacenters today and the evolution of the Intelligent Network (IN) infrastructure in telecommunication. The telecommunications networks have for many years, demonstrated their ability to reliably enable network (voice) services creation, assurance and delivery on a massive scale. Based on an analysis of the Intelligent Networks in telecommunications to identify proven concepts and key lessons that can be applied to enable next generation IT datacenters experience this paper asserts that:

- **In order to scale cloud services reliably to millions of service developers and billions of end users the next generation cloud computing and datacenter infrastructure will have to follow an evolution similar to the one that led to the creation of scalable telecommunication networks.**

- **In the future network-based cloud service providers will leverage virtualization technologies to be able to allocate just the right levels of virtualized compute, network and storage resources to individual applications based on real-time business demand while also providing full service level assurance of availability, performance and security at a reasonable cost.**

- **A key component - identified in this paper as the Virtual Resource Mediation Layer (VRML), must be developed through industry collaboration to enable interoperability of various public and private clouds. This layer will form the basis for ensuring massive scalability of cloud infrastructure by enabling distributed service creation, service delivery and service assurance without any single vendor domination.**

- **The next generation virtualization technologies must allow applications to dynamically access CPU, memory, bandwidth and storage (capacity, I/O and throughput) in a manner similar to that of the telecommunications 800 Service Call Model[1] with one level of indirection and mediation.**

The authors believe that the next generation cloud evolution is a fundamental transformation – and not just an evolutionary stack of XaaS implementations, which will enable global service collaboration networks utilizing optimally distributed and managed computing, network and storage resources driven in real ‑ time by business priorities.

I. Introduction

The appetite for a new generation of network-based applications – both for consumers e.g. Twitter, Facebook, YouTube, Hulu, Animoto and for businesses e.g. Web Mail, Google Docs, Zoho, is driving the need to reorganize current datacenter infrastructure for massive scale. Another characteristic of the new web-based network applications is wildly fluctuating demand – especially in the mass consumer market. This need is stretching current IT architectures to their limits in terms of the ability to ensure on-demand service availability, reliability, performance and security at a reasonable cost.

While demand for network services is soaring, the economic pressure to do “more with less” is also rising. With virtualization technologies becoming more accepted, public and private cloud networks are emerging as an attractive means for sharing compute, storage and network resources amongst multiple service developers and service delivery applications. Such a sharing of resources immediately provides economies of scale through consolidation, energy savings and improved resource utilization. More importantly, the ability to dynamically reallocate resources using virtualization technologies can help mitigate the need for additional investment in infrastructure to meet sudden spikes in demand by temporarily diverting existing resources from low-priority business applications to high priority business applications.

While progress is certainly being made today with respect to resource consolidation and capacity scaling, a

[1] The 800 Service call model provides a level of indirection between the calling party (the application in this case) and the called party (computing, network or storage resource) based on their profile based demand

1524-4547/09 $25.00 © 2009 IEEE
DOI 10.1109/WETICE.2009.14

truly dynamic datacenter that ensures on-demand service creation, availability, reliability, performance and security is still a vision that is yet to be realized.

To that end, we make a case for an open standards approach - similar to the one that has proven successful in establishing Intelligent Networks (IN) for Telecommunications (through ITU) and for the Internet (through IETF), to enable massive scale and interoperability in all phases of service creation, delivery and assurance. Telecommunication's Intelligent Networks have already demonstrated how today hundreds of thousands of developers create millions of services that are consumed by billions of customers - who much like today's Web users, often create wild fluctuations in demand.

After analyzing the evolution of telecommunications networks and comparing it to the current state-of-the-art with respect to IT datacenters, this paper identifies a key component called Virtual Resource Mediation Layer (VRML) that must be developed to support scalability and interoperability of various public and private clouds. By analogy with the telecommunications network, this layer will:

- Mediate between networked applications and virtualized computing, network and storage resources with dynamic provisioning;
- Enable development of end-to-end or application-to-spindle Fault, Configuration, Accounting, Performance and Security (FCAPS) management based on business priorities using dynamic monitoring of workloads on computing, network and storage resources and;
- Allow the development of next generation converged service creation, delivery and assurance infrastructure that is massively scalable and globally interoperable along with a new degree of agility.

The VRML essentially mediates the computing (CPU, and memory), Network (bandwidth) and storage (capacity, throughput and IO per second) resources between various applications that request them just as the telecommunications network allocates the switching, transmission and access resources to meet its IN service requirements. Using the VRML services, a Service Collaboration Network (SCN) Platform, can be developed and provided by multiple service providers.

The VRML approach proposed here offers a way to leverage emerging virtualization technologies in combination with COTS (Commercial off the Shelf) hardware to radically transform the next generation datacenters by moving more of the "intelligence" into the network. Additionally, deploying it does not require abandoning any current IT investments as it can accommodate a gradual migration from today's IT deployments with existing complex management systems to a more integrated and simplified virtualized computing, network and storage services platform that is massively scalable and interoperable.

II. THE CLOUD FORMATION

"You can see they've gone from 50 instances of EC2 usage up to 3,500 instances of EC2 usage. It's completely impractical in your own data center over the course of three days to scale from 50 servers to 3,500 servers. Don't try this at home."
- Jeff Bezos, CEO Amazon [1]

Animoto – a small startup with limited resources, created an online service that generates a unique custom video from photos and music uploaded by users. When they put the application on Facebook and it went viral and demand shot up through the roof. Astoundingly, they managed to scale from 50 servers to 3500 servers in three days – all without having to buy a single piece of hardware or having to create their own compute, network and storage infrastructure. This was all accomplished by renting compute infrastructure from cloud service provider – Amazon Elastic Cloud Computing (EC2) and complementary service management capabilities from management provider RightScale which enabled automated workload monitoring and Virtual Machine provisioning on Amazon's EC2 infrastructure.

The above example demonstrates how existing cloud infrastructure can be used to enable massive scale and agility at a very reasonable cost using:

1. Virtualization technology to dynamically provision virtualized software applications, load balancers and web application servers on-demand,
2. Innovative distributed computing technology that allows database distribution,
3. A managed Service Oriented Architecture for Web Service deployment and
4. A large number of commodity hardware devices (servers, storage and network elements)

Impressive as it is, this current state-of-the-art in cloud computing still is just a baby step when compared to what is expected in a fully functional cloud based service creation, delivery and assurance platform. Consider the following:

1. While the infrastructure services used by service developers are dynamically provisioned, and billed on usage, the system administration and management costs continue to increase with the number of servers used.
2. While service delivery is able to scale in the current cloud model to support spikes in demand, application availability, performance optimization and security management have to be implemented

separately. Today, a host of other companies are actively trying to fill this need [2,3,4,5,6] with additional services using customized point solutions.

3. Disaster Recovery (DR) and storage management (de-duplication, tiered storage) are mostly lacking and have to be individually implemented at additional cost and effort.

The above points highlight some of the reasons why the cloud is today divided into private and public instances. The rule of thumb that seems to have evolved is that if there is a need for developing and deploying services using more than 50 to 100 servers at near full utilization, then private clouds may prove economical. This is roughly the point at which the additional management cost and complexity required for service assurance – not just simple service delivery, makes private clouds viable. It is important to recognize that this number varies and depends on the extent of automation made available by the cloud infrastructure service providers to facilitate service creation, delivery and assurance. More the automation provided by public clouds, lesser the need for private clouds. History shows that economies of scale will favor public clouds if they can address availability, performance and security at all levels.

It is apparent that the datacenter infrastructure required to manage virtualized computing, network and storage resources in an integrated fashion has not yet evolved to take cloud computing to the next level. One of the reasons is that datacenters today are managed using a number of legacy management systems that invariably started with a server-centric management paradigm and have since evolved incrementally over the past couple of decades to accommodate the shift towards client-server and network-based computing paradigms. As a result, there is no single system today that provides truly integrated cross-domain management capabilities required for a service-oriented cloud infrastructure. At best each management offers specialized management of a particular infrastructure silo (i.e. servers, storage and networks) or partial management across more than one silo. It is also quite common for similar management functionality to be duplicated in solutions provided by multiple vendors specializing in different domains [7]. Further, the best practices promoted by each vendor may conflict when attempting end-to-end optimization across the datacenter.

To illustrate the above, take a look at any datacenter today and you are likely to find that they are paying thrice for a storage volume manager performing similar functionality in their servers, storage and network devices without even being aware of it. To ensure redundancy, clustering and multi-pathing may have been implemented in their servers, networks and their storage. Storage cache management is likely implemented in their virtual servers, physical servers and storage layers.

Figure 1 shows a typical datacenter with all its support systems demonstrating the incremental nature of its evolution and the resulting complexity and cost.

Clearly the inefficiencies incurred in terms of management complexity, sub-optimal performance and costs are untenable. Dynamic reconfiguration of all infrastructure i.e. compute, network and storage resources, based on an application's needs is a necessary condition for automating datacenter management.

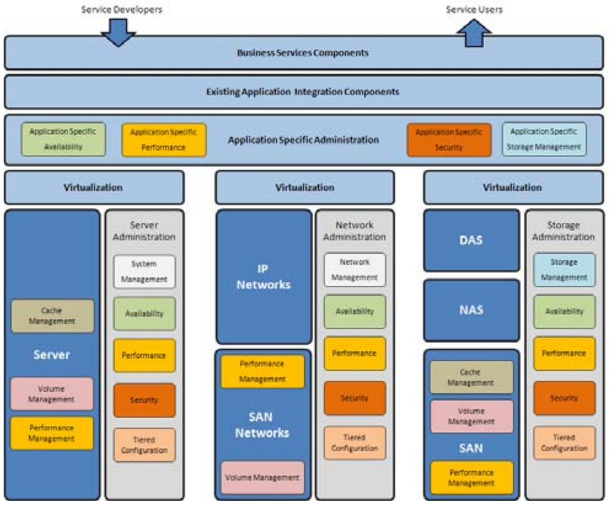

Figure 1. Datacenter complexity today with duplication of management functions

For this paper we analyzed the IN services in telecommunications and propose that a similar evolution that utilizes dynamic provisioning of computing, network and storage resources made possible by virtualization technologies will radically reduce the management complexity in next generation datacenters. By borrowing the FCAPS management and signaling abstractions from the telecommunications domain, a next generation virtualized intelligent service collaboration network infrastructure can be developed that will integrate both public and private clouds to offer massive scale and interoperability.

Management simplicity can be achieved by consolidating application, server, network and storage management intelligence into the SCN and enabling the brokering of compute, network and storage resources between the various applications that need them based on real-time demands, workload profiles and business priorities.

In section III, we review the IN implementation in telecommunications and propose a similar reference model for the intelligent service collaboration network with service creation, delivery and assurance platforms that are FCAPS managed. In Section IV, we discuss the evolution of virtualized datacenters and propose a next generation Virtual Resource Mediation Layer (VRML) that enables scalable and interoperable virtual clouds. The new architecture proposed will allow intelligent service collaboration network to be implemented using commercial off the-shelf (COTS) hardware and network intelligence. VRML also allows integration of current server, network and storage elements and facilitates easy migration to the new architecture using legacy virtualization interface units (LVIUs). Section V concludes with some suggestions for standards evolution and future direction.

III. THE CLOUD EVOLUTION: FAULT, CONFIGURATION, ACCOUNTING, PERFORMANCE AND SECURITY (FCAPS) MANAGEMENT AND THE INTELLIGENT SERVICE COLLABORATION NETWORK

"Although the root cause of this particular issue was a resource contention issue between instances, things like that are going to continue to happen. There may now be a fix for this particular edge case, but there are undoubtedly others that will crop up over time. The real failure here was a failure of monitoring, and a failure of transparency."

This quotation [8] from Oren Michels, the CEO of Mashery, regarding Amazon's EC2 outage, points out the need for application-specific Fault, Configuration, Accounting, Performance and Security (FCAPS) measurement, management and optimization.

The current definitions of cloud computing are just beginning to incorporate end-to-end management as a basic foundation for cloud IT. For example, Forrester Research Group now defines cloud computing [5,9] as "A pool of abstracted, highly scalable, and managed compute infrastructure capable of hosting end customer applications and billed by consumption." Meanwhile the ITU-T Telecommunication Management Network (TMN), already has a well articulated definition for managed infrastructure in the context of the telecommunications Intelligent Networks for voice services. In this layered model, each layer is responsible for different management functions, while interfacing with underlying and overlying layers, to provide a complete and comprehensive set of management capabilities:

1. The Network Element Layer (NEL) implements logical entities within a device
2. The Element Management Layer (EML), implements device level FCAPS management functions
3. The Network Management Layer (NML), implements path management, topology management and fault isolation
4. The Service Management Layer (SML), implements mechanisms to assure service level agreements and ensure Quality of Service (QoS)
5. The Business Management Layer (BML), implements strategic enterprise management functions, such as budgeting and billing

In this manner, the above TMN FCAPS framework enables:

1. Fault management, by detecting and correlating faults in network devices, isolating faults and initiating recovery actions
2. Configuration management, by providing change tracking, configuration, installation and distribution of software to all network devices
3. Accounting management capability through comprehensive network usage reports generated by collecting and parsing accounting data
4. Performance management by providing real-time access for the monitoring of network performance (QoS) and resource allocation data
5. Security management by providing granular access control for network resources

Applying the above framework, we propose a Cloud Computing Reference Model that explicitly incorporates FCAPS management and defines the various roles of infrastructure, service creation, delivery, and assurance platform providers. These roles can be assumed by a single provider or multiple providers depending on whether the solutions are proprietary or standards-based. However, history has consistently shown us that proprietary solutions may drive innovation initially but standards will ultimately be required to achieve massive scale by enabling the interoperability of competitive proprietary solutions.

Figure 2 shows the roles of various players (service operators, developers and end users) in order to realize massively scalable clouds where thousands of developers create millions of services that serve billions of customers.

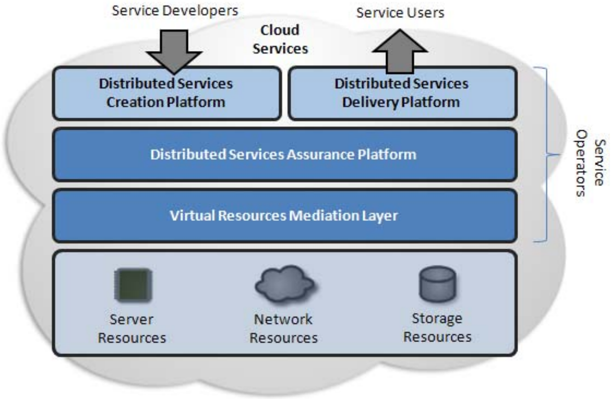
Service Developers
Service Users
Cloud Services
Distributed Services Creation Platform
Distributed Services Delivery Platform
Distributed Services Assurance Platform
Virtual Resources Mediation Layer
Service Operators
Server Resources
Network Resources
Storage Resources

Figure 2. Service Creation, Assurance and Delivery Model

A similar cloud model described by Frank Gillett [9] is shown in Figure 3. However, that model does not seem to address end-to-end management. Ultimately, the cloud service infrastructure must provide end-to-end service assurance (FCAPS management) to meet both service creation and service delivery platform user requirements. The service creators must be able to develop services rapidly using reusable and collaborating service components available globally. The infrastructure must also accommodate billions of users globally who will contribute to wildly fluctuating workloads.

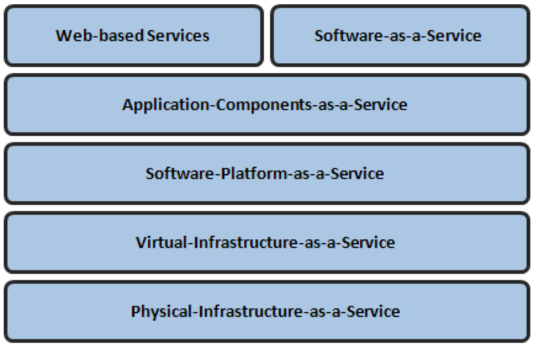
Web-based Services
Software-as-a-Service
Application-Components-as-a-Service
Software-Platform-as-a-Service
Virtual-Infrastructure-as-a-Service
Physical-Infrastructure-as-a-Service

Figure 3. Current Cloud Computing Model

Amazon has successfully demonstrated that virtualization, distributed computing and service oriented software environment can be combined with commodity hardware to both develop and deliver massively scalable services. It has created a virtual server environment that can be successfully used to create a certain class of applications (web based service delivery). Where it falls short is in the scalability of system administration. The end-user is left to worry about various datacenter functions such as load balancers, firewalls, replication, disaster recovery (DR), and storage and security management. This has opened the opportunity for a host of startups to attempt to fill this gap [2].**Error! Reference source not found.**

Current cloud evolution is limited to the following three areas:

1. The Virtualization of servers, load balancers, and some server IP address management services
2. The replacement of SAN/NAS infrastructure with large commodity server farms that support virtual applications using Direct Attached Storage (DAS) or File Systems (distributed or otherwise)[2]
3. Application of distributed computing innovations through Web Services and Service Oriented Architecture (SOA)

It is apparent from above that the datacenter is evolving incrementally from the bottom up without the top down end-to-end architectural framework that is required to enable scalability, performance, availability and security for cloud services. It is only a matter of time before we see the IT industry recognizing the need to move beyond server virtualization and incorporate virtualized network and storage resources[3] to enable dynamic provisioning of resources end-to-end. At this point, the cloud IT industry would do well to adopt a telecommunications-style IN model[4] and implement application FCAPS management and a Virtual Resource Mediation Layer (VRML) to enable a 800 Service Call Model that can provision CPU/memory, bandwidth and storage resources dynamically based on application requirements. Using this model, application resource optimization based on application workload needs and business constraints becomes as simple as making a phone call. Service creation, delivery and assurance will become very similar in reliability and performance to those offered by the Telecommunications IN Services platform.

Current IT emerged from a server-centric architecture that later evolved into a client-server architecture to accommodate network-based computing. Optimization in these architectures centered primarily on server resources. With the shift to network-based services, a next generation network-centric mediation layer is required to optimize the services platform for massive scaling and interoperability. By providing the mediation between virtualized computing, networking and storage resources, the VRML

[2] The Current approach to storage replication and storage based application management using multi-vendor SAN/NAS solutions is being made obsolete by the adoption of virtualization technologies. Next generation virtualization technologies will allow the network based IN services platform to utilize COTS storage elements which will be virtualized and dynamically allocated to provide the right throughput, IOPs and capacity to the right application based on business priorities. This also will simplify HA/DR, performance & security optimization and scaling by using network services such as "broadcast", "call forwarding" and "call waiting" with "dumb" end devices. This strategy eliminates duplication of intelligence at the edge with multi-vendor resource management solutions and complex storage subsystems.

[3] We include application FCAPS monitoring in real-time and dynamic provisioning of resources in virtualization technologies of the future. Today such provisioning is done through a number of management systems or is often manually executed. We also envision a change from current hypervisor based virtualization to a 800 service model dynamic provisioning of CPU and memory in place of application and OS image switching from server to server.

[4] The similarity between Telecommunication network (which provides voice service by connecting right switching, transmission and access resources based on user profiles) to Next generation virtualized application network becomes clear if we look at all applications essentially being switched to the right computing, network and storage resources based on application and resource profiles by a switching platform. Virtualization offers dynamic configuration and reconfiguration of the computing (CPU and memory), network (bandwidth) and storage (capacity, throughput and IOPs) based on application's need and business priorities. In a globally scalable and interoperable switched network, the intelligence resides in the network and not in multiple computing, network and storage devices to reduce overall CAPEX and OPEX.

will become a network Operating System (OS) and its domination by a single vendor can create monopoly that may not be in the best interest of cloud computing.

IV. Conclusion

In this paper, current trend in cloud computing have been analyzed and compared with the evolution of the telecommunications Intelligent Network (IN). A new reference model for the next generation datacenters that will enable both public and private clouds to be massively scalable and interoperable has been proposed.

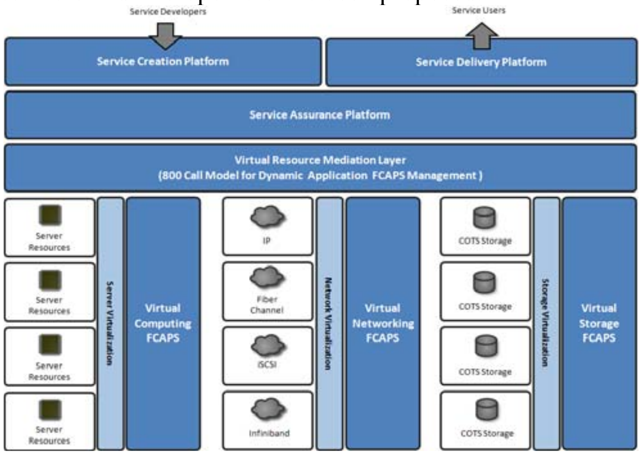

Figure 4. Next Generation Cloud Infrastructure with VRML

Learning from the lessons of the past, the paper proposes a next generation Virtualization Mediation Layer that goes beyond current server virtualization and integrates network and storage virtualization to enable seamlessly unified management. The VRML layer allows the creation of next generation virtualized computing, network and storage devices using "dumb" COTS components, while integrating into current generation architectures with plug-in adapters. This will allow gradual migration[5] from current generation applications to SCN, providing the next generation services architecture in massively scalable and globally interoperable cloud platforms.

The proposed platform can help transform IT infrastructure to bring it on par with telecommunications and Internet platforms that can scale massively while delivering reliable, and optimal performance along with fine grain security controls. The authors envision transforming the datacenter into a "central office" for enabling application connection, (in this case, the connections of multiple applications with computing, network and storage resources) much like the telecommunications "central office" connects billions of people anywhere in the world and assures those connections even in the case of disasters.

Developing the proposed VRML platform will require implementing a new distributed computing model, the 800 service call model, dynamic end-to-end FCAPS management, and signaling for business priority based resource allocation - all borrowed heavily from the telecommunications domain.

The paper also proposes that to be successful, VRML must be defined through a standards based RFI process with leadership driven by global standards bodies such as the IETF or ITU. The evolution of the telecommunications network and the Internet has demonstrated the success of this approach. While ITU provided top down standards development, IETF followed bottom up request for comment RFC process. For clouds to be massively scalable and for both public and private clouds to become globally interoperable, the role of the VRML is critical and must be vendor agnostic.

The authors believe that the next generation cloud evolution is a fundamental transformation – and not just an evolutionary stack of XaaS implementations, which will enable global service collaboration networks utilizing optimally distributed and managed computing, network and storage resources driven in real-time by business priorities.

V. References

[1] http://blog.animoto.com/2008/04/21/amazon-ceo-jeff-bezos-on-animoto/

[2] "Let it rise – A special report on corporate IT", The Economist, October 25th, 2008

[3] Jeff Cogswell, "RightScale eases developing on Amazon EC2" e-week.com, October 21st, 2008

[4] Peter Wayner, "Cloud versus cloud – A guided tour of Amazon, Google, AppNexus and GoGrid", InfoWorld, July 21, 2008

[5] James Staten, "Is Cloud Computing Ready for the Enterprise?", Forrester Report, March 7, 2009,

[6] http://searchitchannel.techtarget.com/generic/0,295582,sid96_gci1336995,00.html

[7] "Our survey confirms that businesses are indeed challenged most by the need to effectively manage the increased complexity in today's data centers, while at the same time keeping networks running smoothly, and power consumption costs down," said Ben Grimes, Avocent CTO and vice president of corporate strategy. (http://www.cio.in/news/viewArticle/ARTICLEID=5636520)

[8] http://oren.blogs.com/praxis/2008/04/the-amazon-ec2.html, blog entry from Oren Michels, CEO of Mashery talking about Amazon's E2C outage

[9] Frank E. Gillett, "Future View: The new technology ecosystems of cloud, cloud services and cloud computing" Forrester Report, August 2008.

[5] Migration of a large base of current applications to the cloud without interrupting the services they are currently providing will be an essential requirement for the SCN infrastructure. Virtualization and resulting dynamic provisioning capabilities and the 800 service call model will allow SCN to build cloud based mediation applications that interface with current generation storage systems through their management systems. Similar approaches have been adopted in the past in migrating legacy telecommunications system to IN platforms using plug-in adopters and mediation and conversion functions.