International Journal of Enterprise Computing and Business Systems

ISSN (Online) : 2230-8849

http://www.ijecbs.com

Vol. 1 Issue 2 July 2011

EMERGING CLOUD COMPUTING

Abu Sarwar Zamani, Dr.Nassir S. Al-Araifi, Dr. Ashit K. Dutta

Saqra University, Saudi Arabia

ABSTRACT

Cloud computing is fundamentally altering the expectations for how and when computing, storage and networking resources should be allocated, managed and consumed. End-users are increasingly sensitive to the latency of services they consume. Service Developers want the Service Providers to ensure or provide the capability to dynamically allocate and manage resources in response to changing demand patterns in real-time. Ultimately, Service Providers are under pressure to architect their infrastructure to enable real-time end-to-end visibility and dynamic resource management with fine grained control to reduce total cost of ownership while also improving agility. What is needed is a rethinking of the underlying operating system and management infrastructure to accommodate the ongoing transformation of the datacenter from the traditional server-centric architecture model to a cloud or network-centric model. This paper proposes and describes a reference model for a network-centric datacenter infrastructure management stack that borrows and applies key concepts that have enabled dynamism, scalability, reliability and security in the telecom industry, to the computing industry. Finally, the paper will describe a proof-of-concept system that was implemented to demonstrate how dynamic resource management can be implemented to enable real-time service assurance for network centric datacenter architecture.

Keywords: Cloud Computing, Datacenter, Distributed Computing, Virtualization.

1 INTRODUCTION

The unpredictable demands of the Web 2.0 era in combination with the desire to better utilize IT resources are driving the need for a more dynamic IT infrastructure that can respond to rapidly changing requirements in real-time. This need for real-time dynamism is about to fundamentally alter the datacenter landscape and transform the IT infrastructure as we know it.

In the cloud computing era, the computer can no longer bethought of in terms of the physical enclosure – i.e. the server or box, which houses the processor, memory, storage and associated components that constitute the computer. Instead the “computer” in the cloud ideally comprises a pool of physical compute resources – i.e. processors, memory, network bandwidth and storage, potentially distributed physically across server and geographical boundaries which can be organized on demand into a dynamic logical entity i.e. a “cloud computer”, that can grow or shrink in real-time in order to assure the desired levels of latency sensitivity, performance, scalability, reliability and security to any application that runs in it. What is truly enabling this transformation today is virtualization technology – more specifically hardware assisted server virtualization. At a fundamental level, virtualization technology enables the abstraction or decoupling of the application payload from the underlying physical resource. What this typically means is that the physical resource can then be carved up into logical or virtual resources as needed. This is known as provisioning. By introducing a suitable management infrastructure on top of this virtualization functionality, the provisioning of these logical resources could be made dynamic i.e. the logical resource could be made bigger or smaller in accordance with demand. This is known as dynamic provisioning. To enable a true "cloud" computer, every single computing element or resource should be capable of being dynamically provisioned and managed in real-time. Presently, there are many holes and areas for improvement in today's datacenter infrastructure before we can achieve the above vision of a cloud computer.

International Journal of Enterprise Computing and Business Systems

ISSN (Online) : 2230-8849

http://www.ijecbs.com

Vol. 1 Issue 2 July 2011

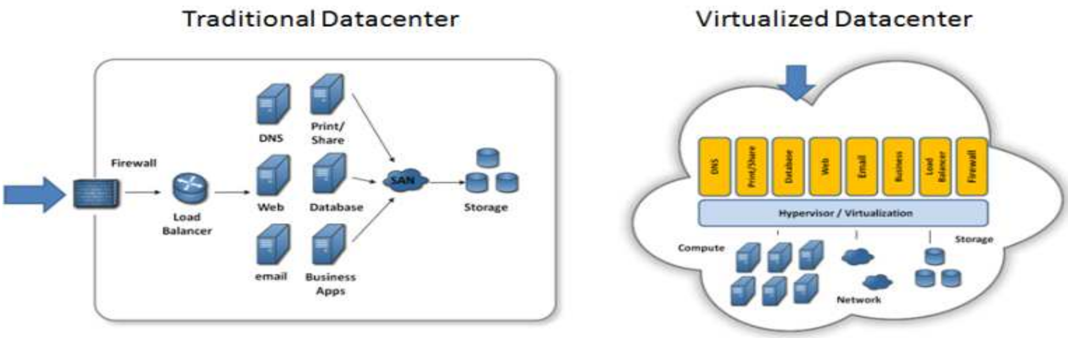

Figure 1: Transformation of the Traditional Datacenter

1.1 SERVER OPERATING SYSTEMS AND VIRTUALIZATION

Whereas networks and storage resources - thanks to advances in network services management and SANs, have already been capable of being virtualized for a while, only now with the wider adoption of server virtualization do we have the complete basic foundation for cloud computing i.e. all computing resources can now be virtualized. Consequently, server virtualization is the spark that is now driving the transformation of the IT infrastructure from the traditional server-centric computing architecture to a network-centric, cloud computing architecture. With server virtualization, we now have the ability to create complete logical (virtual) servers that are independent of the underlying physical infrastructure or their physical location. We can specify the computing, network and storage resources for each logical server (virtual machine) and even move workloads from one virtual machine to another in real-time (live migration). All of this has helped to radically transform the cost structure and efficiency of the datacenter. Despite the numerous benefits that virtualization has enabled we are yet to realize the full potential of virtualization in terms of cloud computing. This is because:

1.1.1 TRADITIONAL SERVER-CENTRIC OPERATING SYSTEMS WERE NOT DESIGNED TO MANAGE SHARED DISTRIBUTED RESOURCES:

The Cloud computing paradigm is all about optimally sharing a set of distributed computing resources whereas the server-centric computing paradigm is about dedicating resources to a particular application. The server-centric paradigm of computing inherently ties the application to the server. The job of the server operating system is to dedicate and ensure availability of all available computing resources on the server to the application. If another application is installed on the same server, the operating system will once again manage all of the server resources, to ensure that each application continues to be serviced as if it has access to all available resources on that server. This model was not designed to allow for the “dial-up” or “dial down” of resource allocated to an application in response to changing workload demands or business priorities. This is why load-balancing and clustering was introduced.

1.1.2 CURRENT HYPERVISORS DO NOT PROVIDE ADEQUATE SEPARATION BETWEEN APPLICATION MANAGEMENT AND PHYSICAL RESOURCE MANAGEMENT:

Today's hypervisors have just interposed themselves one level down below the operating system to enable multiple "virtual" servers to be hosted on one physical server. While this is great for consolidation, once again there is no way for applications to manage how, what and when resources are allocated to themselves without having to worry about the management of physical resources. It is our observation that the current generation of hypervisors which were also born from the era of server-centric computing does not delineate hardware management from application management much like the server operating systems themselves.

1.1.3 SERVER VIRTUALIZATION DOES NOT YET ENABLE SHARING OF DISTRIBUTED RESOURCES:

Server virtualization presently allows a single physical server to be organized into multiple logical servers. However, there is no way for example to create a logical or virtual server from resources that may be physically located in separate servers. It is true that by virtue of the live migration capabilities that server virtualization technology enables, we are able to move application workloads from one physical server to another potentially even geographically distant physical server. However, moving is not the same as sharing. It is our contention that to enable a truly distributed cloud computer, we must be able to efficiently share resources no matter where they reside purely based on the latency constraints of applications or services that consume the resources. .

1.2 STORAGE NETWORKS & VIRTUALIZATION

Before the proliferation of server virtualization, storage networking and storage virtualization enabled many improvements in the datacenter. The key driver was the introduction of the Fibre Channel (FC) protocol and Fibre Channel-based Storage Area Networks (SAN) which provided high speed storage connectivity and specialized storage solutions to enable such benefits as server-less backup, point to point replication, HA/DR and performance optimization outside of the servers that run applications. However, these benefits have come with increased management complexity and costs.

1.3 NETWORK VIRTUALIZATION

The virtual networks now implemented inside the physical server to switch between all the virtual servers provide an alternative to the multiplexed, multi-pathed network channels by trunking them directly to WAN transport thereby simplifying the physical network infrastructure.

1.4 APPLICATION CREATION AND PACKAGING

The current method of using Virtual Machine images that include the application, OS and storage disk images is once again born of a server-centric computing paradigm and does not lend itself to enable distribution across shared resources. In a cloud computing paradigm, applications should ideally be constructed as a collection of services which can be composed, decomposed and distributed on the fly. Each of the services could be considered to be individual processes of a larger workflow that constitutes the application. In this way, individual services can be orchestrated and provisioned to optimize the overall performance and latency requirements for the application.

2 PROPOSED REFERENCE ARCHITECTURE MODEL

If we were to distil the above observations from the previous section, we can see a couple of key themes emerging. That is:

2.1 The next generation architecture for cloud computing must completely decouple physical resources management from virtual resource management; and

2.2 PROVIDE THE CAPABILITY TO *MEDIATE* BETWEEN APPLICATIONS AND RESOURCES IN REAL-TIME.

As we highlighted in the previous section, we are yet to achieve perfect decoupling of physical resources management from virtual resource management but the introduction and increased adoption of hardware assisted virtualization (HAV) as an important and necessary step towards this goal. Thanks to HAV, a next generation hypervisor will be able to manage and truly ensure the same level of access to the underlying physical resources. Additionally, this hypervisor should be capable of managing both the resources located locally within a server as well as any resources in other servers that may be located elsewhere physically and connected by a network. Once the management of physical resources is decoupled from the virtual resource management the need for a mediation layer that arbitrates the allocation of resources between multiple applications and the shared distributed physical resources becomes apparent.

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
http://www.ijecbs.com
Vol. 1 Issue 2 July 2011

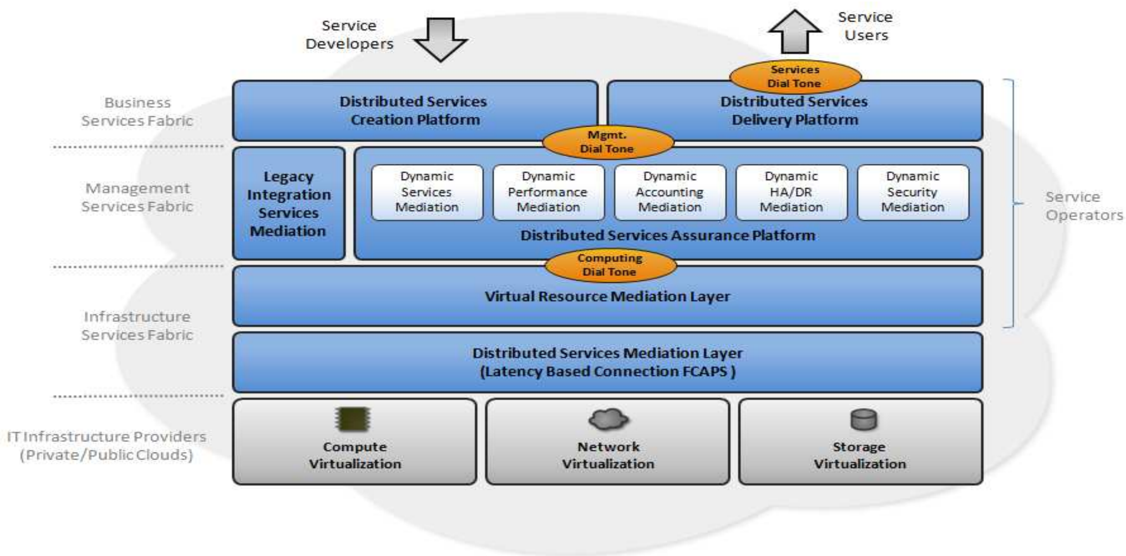

Figure 2: Reference Architecture Model for Next Generation Cloud Computing Infrastructure

3 INFRASTRUCTURE SERVICE FABRIC

This layer comprises two pieces. Together the two components enable a computing resource “dial-tone” that provides the basis for provisioning resource equitably to all applications in the cloud.

3.1 DISTRIBUTED SERVICES MEDIATION

This is a FCAPS based (Fault, Configuration, Accounting, Performance and Security) abstraction layer that enables autonomous self-management of every individual resource in a network of resources that may be distributed geographically, and

3.2 VIRTUAL RESOURCE MEDIATION LAYER

This provides the ability to compose logical virtual servers with a level of service assurance that guarantees resources such as number of CPUs, memory, bandwidth, latency, IOPS (I/O operations per second), storage throughput and capacity.

4 DISTRIBUTED SERVICES ASSURANCE PLATFORM

This layer will allow for creation of FCAPS-managed virtual servers that load and host the desired choice of OS to allow the loading and execution of applications. Since the virtual servers implement FCAPS-management, they can provide automated mediation services to natively ensure fault management and reliability (HA/DR), performance optimization, accounting and security. This defines the management dial-tone in our reference architecture model.

5 DISTRIBUTED SERVICES DELIVERY PLATFORM

This is essentially a workflow engine that executes the application which - as we described in the previous section, is ideally composed as business workflow that orchestrates a number of distributable workflow elements. This defines the services dial tone in our reference architecture model.

6 DISTRIBUTED SERVICES CREATION PLATFORM

This layer provides the tools that developers will use to create applications defined as collection of services which can be composed, decomposed and distributed on the fly to virtual servers that are automatically created and managed by the distributed services assurance platform.

7 LEGACY INTEGRATION SERVICES MEDIATION

This is a layer that provides integration and support for existing or legacy application in our reference architecture model.

8 DEPLOYMENT OF THE REFERENCE MODEL

Any generic cloud service platform requirements must address the needs of four categories of stake holders (1) Infrastructure Providers, (2) Service Providers. (3) Service Developers, and (4) End Users. Below we describe how the reference model we described will affect, benefit and are deployed by each of the above stakeholders.

8.1 INFRASTRUCTURE PROVIDERS

These are vendors who provide the underlying computing, network and storage resources that can be carved up into logical cloud computers which will be dynamically controlled to deliver massively scalable and globally interoperable service network infrastructure. The infrastructure will be used by both service creators who develop the services and also the end users who utilize these services.

8.2 SERVICE PROVIDERS

With the deployment of our new reference architecture, service providers will be able to assure both service developers and service users that resources will be available on demand. They will be able to effectively measure and meter resource utilization end-to-end usage to enable a dial-tone for computing service while managing Service Levels to meet the availability, performance and security requirements for each service. The service provider will now manage the application's connection to computing, network and storage resource with appropriate SLAs.

8.3 SERVICE DEVELOPERS

They will be able to develop cloud based services using the management services API to configure, monitor and manage service resource allocation, availability, utilization, performance and security of their applications in real-time. Service management and service delivery will now be integrated into application development to allow application developers to be able to specify run time SLAs.

8.4 END USERS

Their demand for choice, mobility and interactivity with intuitive user interfaces will continue to grow. The managed resources in our reference architecture will now not only allow the service developers to create and deliver services using logical servers that end users can dynamically provision in real-time to respond to changing demands, but also provide service providers the capability to charge the end-user by metering exact resource usage for the desired SLA.

9 CONCLUSION

In this paper, we have described the requirements for implementing a truly dynamic cloud computing infrastructure. Such an infrastructure comprises a pool of physical computing resources – i.e. processors, memory, network bandwidth and storage, potentially distributed physically across server and geographical boundaries which can be organized on demand into a dynamic logical entity i.e. “cloud computer”, that can grow or shrink in real-time in order to assure the desired levels of latency sensitivity, performance, scalability, reliability and security to any application that runs in it. We identified some key areas of deficiency with current virtualization and management technologies. In particular we detailed the importance of separating physical resource management from virtual resource management and why current operating systems and hypervisors – which were born of the server-computing era, are not designed and hence ill suited to provide this capability for the distributed shared resources typical of cloud deployment. We also highlighted the need for FCAPS-based (Fault, Configuration, Accounting, Performance and Security) service "mediation" to provide global management functionality for all networked physical resources that comprise a cloud – irrespective of their distribution across many physical servers in different geographical locations.

We then proposed a reference architecture model for a distributed cloud computing mediation (management) platform which will form the basis for enabling next generation cloud computing infrastructure. We showed how this infrastructure will affect as well as benefit key stakeholders such as the Infrastructure providers, service providers, service developers and end-users.

We believe that what this paper has described is significantly different from most current cloud computing solutions that are nothing more than hosted infrastructure or applications accessed over the Internet. The proposed architecture described in this paper will dramatically change the current landscape by enabling cloud computing service providers to provide a next generation infrastructure platform which will offer service developers and end-users unprecedented control and dynamism in real-time to help assure SLAs for service latency, availability, performance and security.

REFERENCES

[1] Adriana Barnoschi, " Backup and Disaster Recovery For Modern Enterprise", 5th International Scientific Conference, Business and Management' 2008, Vilnius, Lithuvania. - http://www.vgtu.lt/leidiniai/leidykla/BUS_AND_MANA_2008/infcommunication/ 630-635-G-Art-Barnoschi00.pdf

[2] David Chisnall, "guide to the xen hypervisor", First edition, Prentice Hall, Press, NJ, 2009.

[3] Gartner's 2008 Data Center Conference Instant Polling Results: Virtualization Summary – March 2, 2009.

[4] Graham Chen, Qinzheng Kong, Jaon Etheridge and Paul Foster, "Integrated TMN Service Management", Journal of Network and Systems Management, Springer New York, Volume 7, 1999, p469-493.

[5] Jason A. Kappel, Anthony T. Velte, Toby J. Welte, "Microsoft Virtualization with Hyper-V", McGraw Hill, New York, 2009.

[6] Rao Mikkilineni, Vijay Sarathy "Cloud Computing and Lessons from the Past", Proceedings of IEEE WETICE 2009, First International Workshop on Collaboration & Cloud Computing, June 2009.

[7] Rajkumar Buyyaa, Chee Shin Yeoa, , Srikumar Venugopala, James Broberga, and Ivona Brandicc, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility", Future Generation Computer Systems, Volume 25, Issue 6, June 2009, Pages 599-616.