International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 3 Issue: 1 27-Jun-2014,ISSN_NO: 2321-3337

SCALABLE PERFORMANCE IN CLOUD COMPUTING

1B.JEEVARANI ,

[1]Research Scholar, Bharathiar University, Coimbatore 641046.

ABSTRACT *–Cloud computing is the latest evolution of Internet-based computing. Scalability is the ability of a system to increase the workload on its current hardware resources. Most applications experience spikes in traffic. Instead of over-buying your own equipment to accommodate these spikes, many cloud services can smoothly and efficiently scale to handle these spikes with a more cost effective pay-as-you-go model. Poor application performance causes companies to lose customers, reduce employee productivity, and reduce bottom line revenue. Because application performance can vary significantly based on delivery environment, businesses must make certain that application performance is optimized when written for deployment on the cloud or moved from a data center to a cloud computing infrastructure. However, planning cannot always cover sudden spikes in traffic, and manual provisioning might be required. A more cost-effective pursuit of greater scalability performance is the use of more efficient application development; this technique breaks code execution into silos serviced by more easily scaled and provisioned resources. Smart Technologies offer scalability features and options that aid application performance, including lightweight virtualization, flexible resource provisioning, dynamic load balancing and storage caching, and CPU bursting.*

Keyword: ***Virtual Scaling ,Horizontal Scaling , Smart Technology.***

1, INTRODUCTION

As companies move computing resources from premises-based data centers to private and public cloud computing facilities, they should make certain their applications and data make a safe and smooth transition to the cloud. In particular, businesses should ensure that cloud-based facilities will deliver necessary application and transaction performance now, and in the future. Much depends on this migration and preparation for the transition and final cutoff. Rather than simply moving applications from the traditional data center servers to a cloud computing environment and flick the “on” switch, companies should examine

performance issues, potential reprogramming of applications, and capacity planning for the new cloud target to completely optimize application performance.

Applications that performed one way in the data center may not perform identically on a cloud platform. Companies need to isolate the areas of an application or its deployment that may cause performance changes and address each separately to guarantee optimal transition. In many cases, however, the underlying infrastructure of the cloud platform may directly affect application performance.

Businesses should also thoroughly test applications developed and deployed specifically for cloud computing platforms. Ideally, businesses should test the scalability of the application under a variety of network and application conditions to make sure the new application handles not only the current business demands but also is able to seamlessly scale to handle planned or unplanned spikes in demand.

Besides the misperception that more hardware, either physical or virtual, effectively solves all performance issues, companies may have a fundamental misunderstanding of how performance, scalability, and network throughput are interdependent yet affect applications and data access in separate ways

2, RELATEDWORK

2.1 Horizontal and Vertical Scalability

When increasing resources on the cloud to restore or improve application performance, administrators can scale either horizontally (out) or vertically (up), depending on the nature of the resource onstraint. Vertical scaling (up) entails adding more resources to the same computing pool—for example, adding more RAM, disk, or virtual CPU to handle an increased application load. Horizontal scaling (out) requires the addition of more machines or devices to the computing platform to handle the increased demand. This is represented in the transition from Figure 1 to Figure 2, below

transition from Figure 1 to Figure 2, below.

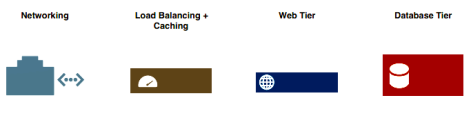

Figure 1: Basic, Single Silo, n-tier architecture

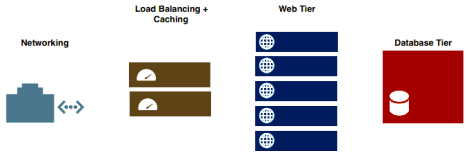

Figure 2: Horizontally scaled load balancing and web-tier. Vertically scaled database tier.

Vertical scaling can handle most sudden, temporary peaks in application demand on cloud infrastructures since they are not typically CPU intensive tasks. Sustained increases in demand, however, require horizontal scaling and load balancing to restore and maintain peak performance. Horizontal scaling is also manually intensive and time consuming, requiring a technician to add machinery to the customer's cloud configuration. Manually scaling to meet a sudden peak in traffic may not be productive traffic may settle to its pre-peak levels before new provisioning can come on line.

2.2 Administrative and Geographical Scalability

While adding computing components or virtual resources is a logical means to scale and improve performance, few companies realize that the increase in resources may also necessitate an increase in administration, particularly when deploying horizontal scaling. In essence, a scaled increase in hard or virtual resources often requires a corresponding increase in administrative time and expenses. This administrative increase may not be a one-time configuration demand as more resources require continual monitoring, backup, and maintenance. Companies with critical cloud applications may also consider geographical scaling as a means to more widely distribute application load demands or as a way to move application access closer to dispersed communities of users or customers. Geographical scaling of resources in conjunction with synchronous replication of data pools is another means of adding fault tolerance and disaster recovery to cloud based data and applications. Geographical scaling may also be necessary in environments where it is impractical to host all data or applications in one central location

2.3 Practical and Theoretical Limits of Scale

While scalability is the most effective strategy for solving performance issues in cloud infrastructures, practical and theoretical limits prevent it from ever becoming an exponential, infinite solution. Practically speaking, most companies cannot commit an infinite amount of money, people, or time to improving performance. Cloud vendors also may have a limited amount of experience, personnel, or bandwidth to address customer application performance. Every computing infrastructure is bound by a certain level of

complexity and scale, not the least of which is power, administration, and bandwidth, necessitating geographical dispersal.

2.4. Addressing Application Scalability

For a cloud computing platform to effectively host business data and applications, however, it must accommodate a wide range of performance characteristics and network demands. Storage, CPU, memory, and network bandwidth all come into play at various times during typical application use. Application switching, for example,places demands on the CPU as one application is closed, flushed from the registers, and another application is loaded. If these applications are large and complex, they put a greater demand on the CPU.

Serving files from the cloud to connected users stresses a number of resources, including disk drives, drive controllers, and network connections when transferring the data from the cloud to the user. File storage itself consumes resources not only in the form of physical disk space, but also disk directories and metafile systems that consume RAM and CPU cycles when users either access or upload files into the storage system.

As these examples illustrate, applications can benefit from both horizontal and vertical scaling of resources on demand, yet truly dynamic scaling is not possible on most cloud computing infrastructures. Therefore, one of the most common and costly responses to scaling issues by vendors is to over-provision customer installations to accommodate a wide range of performance issues.

2.5Application Development to Improve Scalability

One practical means for addressing application scalability and to reduce performance bottlenecks is to segment applications into separate silos. Web-based applications are theoretically stateless, and therefore theoretically easy to scale—all that is needed is more memory, CPU,storage, and bandwidth to accommodate them, as was depicted in Figure 2. However, in practice Web-based applications are not stateless. They are accessed through a network connection(s) that requires an IP address(es) that is fixed and therefore stateful, and they connect to data storage (either disk or database) which maintains logical state as

well as requiring hardware resources to execute. Balancing the interaction between stateless and stateful elements of a Web application requires careful architectural consideration and the use of tiers and silos to allow some form of horizontal resource scaling. To leverage the most from resources, application developers can break applications into discrete tiers—state or stateless processes—that are executed in various resource silos. Figure 3 depicts breaking an application into two silos identified by their DNS name. By segregating state and stateless operations and provisioning accordingly, applications and systems can run more efficiently and with higher resource utilization than under a
more common scenario.

Figure 3: Multi-silo, n-tier, scaled architecture

3, Smart machine improve the scaling performance

3.1 SmartMachine lightweight virtualization. SmartMachines have been designed to provide best possible performance with limited overhead. The SmartOS operating system combines the operating system and virtualization to eliminate redundancy and maximize available RAM for applications.

3.2 SmartCache. Joyent makes use of all of the unused DDR3 memory in the cloud by providing a large ARC Cache pool delivering unparalleled Disk I/O. Both reads and writes are greatly improved as content that would traditionally be served from disk are cached in high speed memory without any customer interaction.

3.3 CPU bursting. The Joyent implementation of its CPU engine allows on-demand processing cycles from a resource pool of available CPUs, enabling instantaneous vertical scaling to meet bursts of application demand without costly

3.4 Choice of virtualization. While SmartMachines provide ideal performance, it recognizes that legacy operating systems and development environments are required for many applications, and SmartOS provides XVM virtualization technology as an integral component of the OS allowing for other operating systems such as Windows and Linux. These operating systems are still able to take advantage of SmartOS capabilities such as SmartCache to achieve improved performance, and management by SmartDataCenter.

3.5 Build clouds on architecture not rent-a-machine. The SmartDataCenter architecture is built with performance and scale of applications in mind versus a simplistic concept of adding more and more virtual machines to solve application performance issues. It understands that application architecture is supported by several tiers of servers that need low latency interconnect. Our patent pending Honeycomb design ensures that servers (Web, App, DB, Cache) while completely distributed and fully redundant, are provisioned in the highest performance and lowest latency manner possible. Rent-a-machine cloud solutions merely move physical data center inefficiencies to virtual, cloud-based inefficiencies.

4. CONCLUSION

Scalability is the best solution to increasing and maintaining application performance in cloud computing environments. Cloud computing vendors often resort to brute-force horizontal scaling by adding more physical or virtual machines, but this approach may not only waste resources but also not entirely solve performance issues, especially those related to disk and network I/O. In addition, customers and vendors alike have practical limits on their ability to scale, primarily constrained by costs and human resources. Smart application development can alleviate performance issues in many cases by isolating resource-intensive processes and assigning the appropriate assets to handle the load. However, for the most part scaling to meet performance demands remains a manual process and requires vigilant monitoring and real-time response.

5. REFERENCES

[1] Michael A, Armando F, et al., "A View of Cloud Computing," Communications of the ACM, Vol.53, April 2010, pp.50-58.

[2] George Pallis, "Cloud Computing: The New Frontier of Internet Computing," IEEE Internet Computing, September-October 2010, pp.70-73.

[3] Rimal, B., et al., "A Taxonomy, Survey, and Issues of Cloud Computing Ecosystems," Springer, London, 2010, pp.21-46.

[4] Mouline, Imad. "Why Assumptions About Cloud Performance Can Be Dangerous."
Cloud Computing Journal. May, 2009.