the Availability Digest

www.availabilitydigest.com
@availabilitydig

Managing Your Private Cloud

April 2015

The Virtualized Data Center

Virtualization has enabled IT infrastructure to become a utility platform. Resources are pooled and are applied to computational loads only as needed. The result is an unparalleled efficiency in the use of compute, storage, and networking infrastructure in a data center. New applications can be easily spawned, and unneeded applications can be removed and the resources that they had been using returned to the common pool.

Shouldn’t you then just be able to build a cloud of virtualized computing resources and let it handle all of your processing needs with little management required?

Not hardly! As we shall see, a cloud needs constant managing to ensure that SLAs are met and that its infrastructure is properly provisioned. As they say, the only constant is change. This is certainly true of clouds as applications grow, as new applications are added with little notice, and as the computational needs of the various applications change with the business environment.

Public clouds are in the business of providing these services. They maintain a competent staff that can closely monitor all aspects of their data centers and respond effectively to any change in needs. Their pricing ensures that they can maintain the highest level of service.

Private enterprises do not always have this luxury. Their data centers are administered by a smaller staff of technicians that have a lot to do besides manage virtualization challenges. Even worse, companies are continually reducing their IT budgets and asking their IT staff to do more with less. In this article, we look at the challenges faced by companies that run their own private clouds.

Managing Virtual Resources

From an IT perspective, business continuity means that a company’s applications are meeting the company’s needs. Each application can handle the processing load sent to it with reasonable response time or batch-run time. The downtime of an application due to failure or planned maintenance must be limited to an acceptable value. Parameters such as these are usually set forth in a Service Level Agreement (SLA).

Different applications have different levels of business importance. Each application will have different performance and availability requirements as well as differences in other measures. Therefore, each application generally has its own SLA. It is the responsibility of the IT staff to ensure that all corporate applications (perhaps hundreds or thousands of them) meet their SLAs. It must ensure that there are sufficient resources to meet these needs without overprovisioning the underlying compute infrastructure. Overprovisioning results in unnecessary capital expenses and only leads to higher costs for running each application.

In one sense, virtualization simplifies business continuity management. It allows rapid changes of compute resources to ensure that application SLAs are met. Workloads can be easily moved to balance their resource needs. Data replication can be employed to guarantee rapid recovery from equipment failures or even data center failures. The provision of acceptable application services is no longer dependent upon the underlying IT infrastructure.

However, virtualization brings new challenges. With no control over how the virtualized infrastructure will be allocated to processing needs, how does the IT staff ensure that all of the SLAs are being met? How does it ensure that there is enough capacity to support its business continuity requirements? How does it benchmark its current infrastructure to mitigate the exposure of the organization's exposure to failures in the IT infrastructure? How does it plan for the future to accommodate application growth and new IT projects?

Furthermore, all of this monitoring and planning must be done within the technical and business constraints that govern how the infrastructure can be utilized to meet the company's business continuity requirements.

Redundancy

A virtualized data center is usually organized into compute clusters. Each cluster comprises a set of physical servers. Each physical server can host multiple virtual machines. The access of the virtual machines to the resources of their host system is controlled by a hypervisor. The cluster is designed to handle a specified workload, though the actual mix of applications running on the server is never known in advance.

Each cluster is designed with N + X redundancy. N + X redundancy means that X host servers in the cluster can fail, and the cluster can still provide the processing capacity required for its assigned workload. The cluster redundant design must take into account the policies set for the infrastructure.

For instance, the infrastructure policy may dictate that no server may carry a workload that exceeds 80% of its capacity. Consider a 5-node cluster with N + 1 redundancy. If a host server fails, the four remaining hosts cannot be loaded more than 80% each. Therefore, the cluster has a capacity to handle a 320% workload. The average workload on each server in a fully operational five-node cluster should not be greater than 64%.

Applications can be moved from cluster to cluster for load balancing. Additional applications may be added from time to time. Therefore, there must be a way to monitor the allocation of applications to clusters to ensure that the availability requirements of each are met.

Affinities

The data center's infrastructure policy usually includes affinity and anti-affinity rules. Affinity rules dictate that certain applications ideally would run on the same host server. This is often done for performance reasons for application sets that interact closely with each other. In these cases, it is desired to minimize the application intercommunication time.

Anti-infinity rules prevent interrelated applications from running on the same host platform so that a host server platform does not affect all of them. A common example is that multiple instances of the same application should never run on the same host server.

Affinity rules are generally configured in the hypervisor management system. They constrain the use of available resources as workloads are assigned or moved. The IT staff must be able to determine when the affinity rules are violated due to constraints on the allocation of available resources to applications.

Load Balancing

A major capability of virtualization is the ability to move applications between compute resources with no interruption to the applications being moved. Load balancing is the basic method of ensuring that all compute resources in the data center are being optimally used. Load balancing enables the scalability of the virtualized data center.

Load balancing must obey the infrastructure policies of the organization and the SLAs of the application. The allowed load on any compute cluster cannot be exceeded. Affinity rules must be followed. Applications must be assigned only to those clusters that provide the specified N + X redundancy to assure application availability.

Manual monitoring of workload placement is virtually impractical because of the frequency of workload movements. A good monitoring tool is essential to ensuring that compute policies are met.

Recovery

Provisions can be made to back up a data center with processing resources maintained at a remote data center. Should a data center fail in its entirety, full-site recovery would allow at least the critical applications to be moved to the remote data center to continue operation. This capability requires that the application databases of the applications that are to be moved be continuously replicated to the remote data center so that the recovered applications have continuous access to their application states.

Site recovery can also be important in the partial failure of a data center. For instance, an improper recovery from a power failure or a failure of the cooling system may mean that some of the compute clusters in the affected data center will have to be shut down. Enough workload can be shifted to the remote data center to allow the surviving clusters to continue to perform according to policy.

Site recovery will usually require that less important applications be suspended at the remote site to make room for the recovered applications. For instance, development and test activities on some clusters at the remote site may be suspended so that these clusters can be used to support the recovered applications.

Recovery failover tests should be periodically run to ensure that site recovery will be successful.

Capacity

The available capacity in a virtualized data center must be continually monitored to ensure that sufficient capacity is available to handle the current and anticipated future workload and that the workload is evenly balanced across all compute clusters. This monitoring should also provide the data necessary to plan for future data center expansions without overprovisioning.

Regardless of the care in managing the capacity of the data center, there may be instances in which the capacity of the data center is about to be overwhelmed. For instance, there may be an unexpected significant increase in the workload imposed by one or more applications. The unanticipated positive response to a new TV ad during the Super Bowl is one example of such a workload spike.

The infrastructure policy of an organization should take this into account. Applications should be organized in priority order so that low-priority applications can be aborted to provide resources for the workload spike.

Future Planning

As we said earlier, the only constant is change. In a virtualized data center, workloads will always be growing due to the expanding needs of current applications and the addition of new applications. (The addition of new applications is often a function of employees running private applications that have not gone through the approval process, a problem common in clouds because it is so easy to spin up new applications.)

Therefore, there is a need to continually monitor the data center's infrastructure usage so that proper, effective, and efficient plans can be made for future upgrades. The data center's compliance with the organization's business continuity requirements must be continuously evaluated against projected future workloads.

There should be a capability to play what-if scenarios to see how the current and proposed future infrastructures will perform under a variety of projected workloads. Only in this way can future expansion planning be effectively pursued.

Data Center Management

Proper data center management will ensure adherence to the policies described above. Pockets of underutilized compute resources can be used to offload resources that are nearing their capacity limits. In the event of a failure, remaining resources can be enabled or resources can be reassigned in order of business priority so that critical applications can continue in service.

VMware and Hyper-V provide the dynamic migration of workloads between clusters of compute resources to meet the above needs. However, during critical overloads or failures, manual intervention may be required to maximize the effects of the rebalancing efforts.

It takes good data center management to deliver the business continuity that organizations need.

Management Tools

As can be seen, the manual monitoring of a virtualized data center can be particularly complex, time consuming, and error-prone. It requires the skills of highly experienced engineers on the IT staff. It is a 24x7 task.

What is needed are effective tools to aid in the continuous monitoring of the data center. Such a tool is VMTurbo, which is described in the companion article in this issue entitled "VMTurbo – Managing Virtualization."[1]

Summary

Virtualization has changed the way in which data centers are implemented. Virtualization allows fully efficient use of all processing resources. No longer is a data center populated with servers that are only 20% utilized. Workloads within the data center can be easily moved to a smaller set of servers whose resources are always being effectively used.

However, in order to continue to meet the business continuity needs of the business, a great deal of monitoring is required. Monitoring is complex and requires the 24x7 attention of experienced staff. There are, however, monitoring tools available to aid significantly in this task.

Acknowledgements

Our thanks to our subscriber, Terry Critchley, for pointing us to this topic. Information for this article was taken from "Can You Really Support Business Continuity Requirements?", a white paper from VMTurbo.

[1] VMTurbo - Managing Virtualization, *Availability Digest*; April 2015.
http://www.availabilitydigest.com/public_articles/1004/vmturbo.pdf