Nutanix Complete Cluster

A Technical Whitepaper

Table of Contents

Executive Summary

Storage is the biggest challenge in virtualized data centers today. The network storage architecture designed fifteen years ago for physical servers is too expensive and complex for virtual machines. Without an enterprise-class alternative, organizations are forced to use traditional solutions, which can't keep up with virtual machines that are dynamic, grow rapidly in number and continue to demand new levels of performance and capacity.

Nutanix Complete Cluster is a scale-out compute and storage infrastructure that allows organizations to virtualize their data centers without requiring network storage (SAN or NAS). Built from the ground up for virtual machines, it provides complete compute and storage capabilities along with enterprise-class performance, scalability, availability and data management features. It leverages industry-standard hardware components, solid-state drives and market-leading hypervisors to provide an out-of-the-box solution that makes virtualization extremely easy and cost effective.

Introduction

Limitations of Current Architecture

Organizations are building their virtualization infrastructure using the traditional servers-connected-to-storage-over-a-network architecture, which can't adapt to the ever-changing demands of virtualization. In addition to slow performance, network storage has become the single biggest source of cost and complexity in virtualized environments. The network storage-based architecture worked well for physical servers that served relatively static workloads. Virtualization, and now Cloud Computing, has made data centers extremely dynamic; virtual machines are created on the fly, move from server to server and depend heavily on shared resources. These characteristics make the management of virtual machines and their underlying physical infrastructure extremely complex.

Data volumes are growing at a rapid pace in the data center, thanks to the ease of creating new VMs. In the enterprise, new initiatives like desktop virtualization contribute to this trend. Service providers deal with an even larger number of VMs as they build data centers to serve customers who can't afford the cost and management overhead that virtualization requires. This growing pool of VMs is exerting tremendous cost, performance and manageability pressure on the traditional architecture that connects compute to storage over a multi-hop network.

The rise of solid-state drives is another trend that is rapidly widening the gap between compute and storage tiers. Use of SSDs that are 100X-1000X faster than traditional hard disks will make the existing network bottlenecks and network complexity even worse, if virtual machines need to access them over a network. Many SAN/NAS vendors are adding SSDs to their solutions, charging a hefty premium for these drives and requiring additional investments in network bandwidth to access an already expensive tier of storage.

The Google Approach

Google and other leading cloud-generation companies such as Amazon, Yahoo and Microsoft (Azure) realized that a network-storage based approach would

not work for their data centers. They built software technology (such as Google File System) that could glue a large number of commodity servers with local storage into a single cluster. This approach allowed Google to build a converged compute and storage infrastructure that used commodity servers with local storage as its building block. Google File System runs across a cluster of servers and creates a single pool of local storage that can be seamlessly accessed by applications running on any server in the cluster. It provides high availability to applications by masking failures of hard disks and even complete servers. Google File System allowed Google to build data centers with massively scalable compute and storage, without incurring the costs and performance limitations associated with network storage.

Nutanix has taken a similar scale-out approach to build an enterprise-ready compute and storage infrastructure that is designed from the ground up for virtual machines.

Nutanix Complete Cluster Architecture

Main Pillars

Nutanix Complete Cluster was designed from scratch to solve storage challenges for virtual machines by building a system that leverages the latest advances in system architecture, hardware and software technologies. There are three core pillars of the Nutanix architecture:

Distributed Computing

The Nutanix architecture is similar to Google's architecture in that it is a scale-out compute and storage infrastructure that eliminates the need for network storage. At the same time, Nutanix builds upon Google's architecture and provides an enterprise-class solution. While Google File System is a custom solution that works for Google's internal applications (search, Gmail, etc.), Nutanix provides a general-purpose solution for virtualized environments. In addition to its scale-out capabilities, it has the same or better enterprise-class data management features that are commonly provided by advanced network storage solutions, including high availability, backup, snapshots, and disaster recovery.

Virtualization

The Nutanix architecture was designed for virtual machines so it supports all hypervisor functions that are supported by the traditional network-storage based architecture, including live VM migration and high availability. In addition, because the Nutanix architecture is VM-aware, it overcomes limitations of traditional solutions that were optimized to work with physical servers. For example, while things are managed on a per VM basis on compute side, the unit of management on storage has traditionally been a LUN. When a LUN is shared by many VMs, it becomes more difficult to perform storage operations such as backup, recovery, and snapshots on a per-VM basis. It is also difficult to identify performance bottlenecks in a heavily-shared environment due to the chasm between computing and storage tiers. The Nutanix architecture overcomes these limitations.

Solid-State Drives

The Nutanix architecture was designed to take advantage of enterprise-grade solid-state drives (SSDs). It is important to note that the traditional storage systems were designed for spinning media and it is hard for them to leverage SSDs efficiently due to the entirely different access patterns that SSDs provide. While hard disks have to deal with the rotation and seek latencies, SSDs do not have such mechanical limitations. This difference between the two media requires the software to be optimized differently for performance. One cannot simply take software written for hard disk-based systems and hope to use it efficiently on solid-state drives. The Nutanix architecture uses SSDs to store a variety of frequently-accessed data, from VM metadata to primary data storage, both in a distributed cache for high-performance and in persistent storage for quick retrieval. To maximize the performance benefits of using SSDs, the Nutanix architecture:

- Reserves SSDs for I/O-intensive functions
- Includes space-saving techniques that allow large amounts of logical data to be stored in a small physical space
- Migrates "cold" or infrequently-used data to hard disk drives automatically, allows administrators to bypass SSDs for low-priority VMs

Architecture Overview

Nutanix Complete Cluster is a scale-out cluster of high-performance nodes, or servers, each running a standard hypervisor and that contains processors, memory and local storage, including SSDs) and hard disk drives. Each node runs virtual machines just like a standard virtual machine host. In addition, local storage from all nodes is virtualized into a unified pool by Nutanix Scale-out Converged Storage (SOCS) (Figure 1). In effect, SOCS acts like an advanced SAN that uses local SSDs and disks from all nodes to store virtual machine data. Virtual machines running on the cluster write data to SOCS as if they were writing to a SAN. SOCS is VM-aware and provides advanced data management features. It brings data closer to virtual machines by storing the data locally on the system, resulting in higher performance at a lower cost. Nutanix Complete Cluster can horizontally scale from a few nodes to a large number of nodes, enabling organizations to scale their infrastructure as their needs grow.

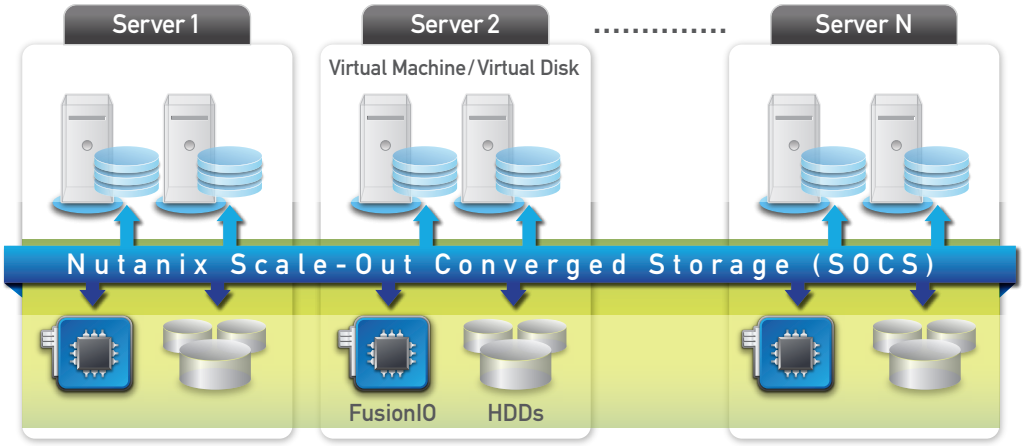

Figure 1: Nutanix Scale-Out Converged Storage Architecture

With Nutanix Complete Cluster, all virtualization features, including high availability and live VM migration, continue to work seamlessly. Administrators create virtual machines on Nutanix Complete Cluster using their standard processes. Nutanix SOCS provides storage for these virtual machines in the form of virtual disks, or vDisks, which are standard iSCSI devices.

Nutanix Scale-Out Converged Storage – Key Components

The key to the Nutanix architecture is SOCS – a patent-pending scale-out converged storage layer that has the following unique set of capabilities:

- It is converged with the compute layer. VMs and SOCS co-exist on the same cluster.
- It is VM-aware. SOCS provisions storage on a per-VM basis and can identify I/O coming from each VM.
- It can scale out from a few nodes to a large number of nodes.
- It has ground-up integration with solid-state drives.
- It provides high availability against disk or node failures.
- It provides high performance by making I/O access local, leveraging solid-state drives and employing a series of patent-pending performance optimizations.
- It provides unique capacity optimization capabilities.

SOCS is enabled by the following components:

n-Way Controller Cluster

While traditional SAN solutions typically have 1, 2, 4 or 8 controllers, an n-node Nutanix Complete Cluster has n controllers (Figure 2). Every node on Nutanix Complete Cluster runs a special virtual machine, called a Controller VM. This virtual machine acts as a virtual controller for SOCS. All Controller VMs in the cluster communicate with each other to form a single distributed system. Unlike traditional SAN/NAS solutions that are limited to a small number of fixed controllers, this architecture continues to scale as more nodes are added.

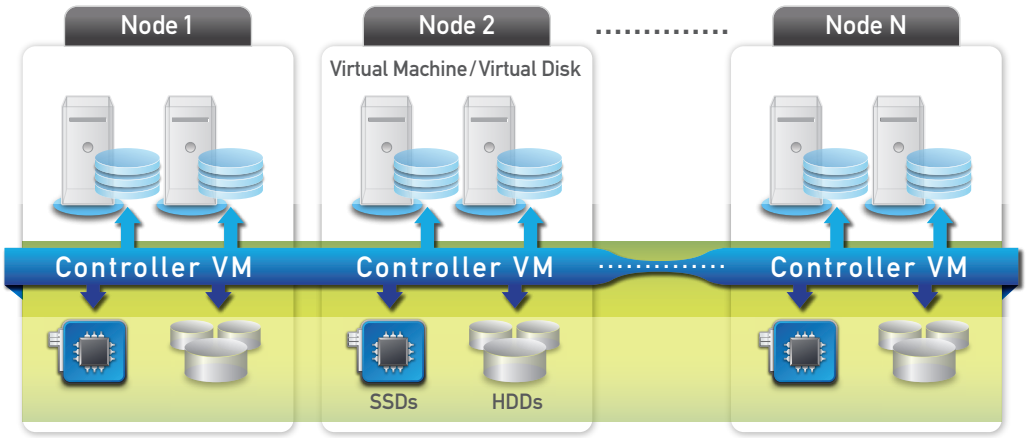

Figure 2: Nutanix Scale-Out Controller VM Architecture

Heat-Optimized Tiering Cache (HOTcache)

HOTcache is a high performance cache backed by SSDs from each node in a cluster. When guest VMs write data, that data is first written to HOTcache and then, in the background, it is flushed to SOCS. HOTcache uses a sequential data layout to provide high performance even if workloads from VMs get mixed into a random workload. HOTcache keeps one data copy on a local SSD and another copy on a different node so that there is no data loss, even in the case of a node failure.

Distributed Metadata Service (Medusa)

In traditional storage systems, controllers often become a bottleneck as more storage is added. One of the main reasons for this issue is that the storage metadata is stored on the controllers. Traditional systems scale to a small number of controllers, so as the number of VMs increases, so does the I/O load on each controller. Nutanix's distributed metadata service, Medusa, distributes

the cluster metadata across the cluster for scalability and replicates the data on multiple nodes for fault tolerance. The service is highly available and can tolerate multiple module failures. In comparison, traditional solutions that depend on a few storage controllers cannot tolerate multiple controller failures.

Distributed Data Maintenance Service (Curator)

Nutanix's distributed data maintenance service, Curator, is a MapReduce[1] -based framework for executing background data management operations in a massively parallel manner. Such operations include:

- Migration of cold data to lower tiers (for Heat-Optimized Tiering)
- Garbage collection of data that has been deleted
- Data consistency through routine checksums
- Replication of data in case of node or disk failures
- Re-balancing of data when nodes are added or removed
- Migration of data to maximize local access when a VM moves from one node to another

FlashStore

FlashStore is the persistent, flash-based storage provided by the pooling of SSDs from all nodes in the cluster. Data is first written to FlashStore and then moved off to DiskStore, as it becomes cold. As cold data becomes hot again, it is brought back into FlashStore. For vDisks that serve low priority VMs, administrators have an option to skip FlashStore.

DiskStore

DiskStore is the high-capacity SATA storage tier spread across the cluster. DiskStore provides large storage capacity for cold data. Nutanix's Heat-Optimized Tiering (HOT) combines FlashStore and DiskStore to provide high performance as well as high capacity at a lower cost.

1 MapReduce is a technology originally built at Google for massively parallel analysis of data in a cluster

Key Features

Nutanix Complete Cluster not only eliminates the need for a SAN, but it also provides top of the line availability, performance and data management features. Some of the key features are described below.

Capacity Optimization

Nutanix QuickClone

Nutanix enables administrators to rapidly deploy new virtual machines by using its QuickClone feature. QuickClones are writeable snapshots that behave just like standard vDisks - administrators can attach them to a VM, write data on them, and even snapshot them further. This is useful for deploying new virtual desktops, creating test and development copies of a production database and any other scenario requiring clones without duplicating the data. The system also supports read-only snapshots for backup purposes.

Nutanix Thin Provisioning

Storage for virtual machines is thinly provisioned in the system. Administrators can set the capacity of a vDisk but physical storage is allocated only when required. Administrators can also set a minimum reservation parameter that guarantees the specified amount of storage for a collection of vDisks.

Nutanix Converged Backup

The Converged Backup feature provides instant backup and recovery capabilities for vDisks. Several months' worth of backups can be kept inside the appliance without requiring external backup storage. When recovery is necessary, administrators can instantaneously restore a vDisk to any of its past backups. The appliance also supports offsite backups using standard third-party tools.

Ease of Management

Ease of Deployment

Nutanix Complete Cluster is a plug-and-play solution that includes all hardware and software necessary to run a large number of virtual servers or virtual desktops. Administrators can set it up and start creating VMs in a matter of minutes.

Next-Generation User Interface

Nutanix Command Center is a highly intuitive Flex-based user interface that provides administrator complete visibility across compute and storage resources in the cluster. It enables them to troubleshoot issues related to a virtual machines easily by mapping each VM to the physical resources in the system. Nutanix Complete Cluster also provides a command-line interface for management.

Nutanix Scale-out Converged Storage (SOCS)

Nutanix SOCS eliminates the need for managing a complex network-based storage infrastructure, making it easy to manage virtual environments at any scale.

Conformance to IT Standards

While Nutanix Complete Cluster enables a converged architecture, it continues to support standard tools and interfaces that IT departments already use. For example, the nodes run an industry-standard hypervisor (VMware ESXi) and all IT processes and software tools that work with this hypervisor continue to work with Nutanix. Similarly, vDisks are standard iSCSI devices that are connected to VMs using a standard iSCSI initiator in the hypervisor. By leveraging such standard interfaces, Nutanix Complete Cluster can seamlessly fit into an existing IT ecosystem.

Performance and Scalability

Solid-State Drives

Nutanix Complete Cluster was designed with SSDs in mind. The included server-based SSDs provide higher performance than SAN-based SSDs because they avoid the network bottleneck. Traditional systems have often been limited by the amount of metadata they can keep in a controller's cache. In Nutanix Complete Cluster, SSDs are used not only for VM data, but also for storing SOCS metadata for fast access. Keeping rich metadata on SSDs enables SOCS to provide advanced data management capabilities. Given the scale-out architecture, SSD capacity in the system grows as more nodes are added to the cluster.

Nutanix Heat-Optimize Tiering (HOT)

In the Nutanix cluster, a vDisk can be allocated a mix of SSD and HDD capacity. To ensure that only high-value data stays on SSDs, SOCS moves cold data to high-capacity SATA drives in the background using its HOT feature.

Scalability

Nutanix Complete Cluster is designed to scale from a few nodes to a large number of nodes. Every aspect of the system was designed with the scalability requirements of today's virtualized data centers in mind. For example, there is no centralized metadata master in the system. The metadata layer itself is distributed across the cluster, eliminating a common bottleneck found in most scale-out systems. Also, with every module running a SOCS Controller, the number of "controllers" in the system can be much higher than a typical network-based storage solution with only a few controllers. Such design innovations enable the system to start small and scale massively.

High Availability

Nutanix Cluster RAID

Nutanix Complete Cluster is a highly available scale-out system with no single point of failure. Using Nutanix Cluster RAID, data is striped across disks within a node for high performance and replicated across the cluster for high availability. This provides high availability for virtual machines even if disks or complete nodes fail.

Nutanix Distributed Metadata Service (Medusa)

Nutanix's distributed metadata service, Medusa, distributes the cluster metadata across the cluster for scalability and replicates the data on multiple nodes for fault tolerance. The service is highly available and can tolerate multiple node failures. In comparison, traditional solutions that depend on a few storage controllers cannot tolerate multiple failures.

Nutanix Instant vDisk Motion

The ability to migrate a live VM from one host to another is a very powerful feature provided by the industry-standard hypervisors. So far, organizations have been able to leverage such live migration capabilities only with network

storage. In fact, many organizations were forced to use network storage only to achieve live migration and high availability. In Nutanix Complete Cluster, live VM migration is supported even with the converged architecture that uses local storage. vDisks in the appliance are logical entities that are fully mobile. This is enabled by Nutanix Instant vDisk Motion feature that can quickly move a vDisk from one node to another, when necessary.

Backup and Recovery

The ability to perform off-site data backup and recovery is key to an organization's data protection strategy. Nutanix Complete Cluster is fully compatible with the VMware vStorage API for Data Protection (VADP), and provides fully functional backup and recovery of virtual machines, through integration with VADP compatible backup and recovery products.

Disaster Recovery

Disaster Recovery is key to the business continuity needs of an enterprise. Nutanix Complete Cluster provides failover and failback capabilities through integration with VADP compatible Disaster Recovery products.

Conclusion

Nutanix Complete Cluster is an innovative system that eliminates the need for network storage without compromising the requirements of enterprise IT. It is built using a next-generation scale-out architecture that has been proven at some of the most innovative cloud-generation companies. With Nutanix Complete Cluster, organizations can build a compute and storage infrastructure for virtual machines that is highly available, fast, rich in data management features and can grow as their needs grow.