Demartek

Nutanix Delivers Intelligent Hyper-Converged Infrastructure that Simplifies the End-User Experience

Well-thought-out Distributed Storage Fabric implementation leverages HCI benefits for pain-point elimination.

Executive Summary

Hyper-converged infrastructure (HCI) is a software architecture that manages multiple clustered nodes of tightly-integrated server, networking and storage technology together as a whole unit. HCI has matured to offer standard enterprise features, however implementation differences between vendors can affect performance, limit efficacy of enterprise features, and complicate cluster management.

Hyper-converged systems offer a unique computing environment where a VM and its data can exist on the same node, forgoing the need for the SAN and separate storage management common in traditional data center environments, while at the same time having the ability to leverage an entire cluster of nodes for data protection, continuity, and compute-heavy tasks. The system designer has to put some thought into the implementation in order to take advantage of these divergent HCI characteristics of locality and data distribution, otherwise performance will suffer as data is perpetually pulled from a node across the network, or only one node is used to rebuild data after a failure. A well-thought-out solution like Nutanix will make intelligent decisions, taking advantage of the HCI environment instead of suffering from it.

A complete HCI solution like Nutanix will implement necessary enterprise features in a new, HCI-specific way and simplify management, allowing all nodes, compute, storage, and networking to be managed from one place. Instead of trying to force legacy data center implementations that do not work onto a new, software-defined platform, deduplication and compression need to be intelligently toggled, granular control of Replication Factor (RF) needs to be given, and hardware RAID should be replaced with HCI-specific erasure coding. Simplified management structures should be Web-scale, offer easy 1-click upgrades, conversions between different hypervisors, and native backup solutions that empower a single IT staff member to manage tens of thousands of VMs.

Nutanix commissioned Demartek to perform a third-party analysis of its hyper-converged solution on Cisco UCS and HPE ProLiant compared to three competitors: HyperFlex on Cisco UCS, HPE SimpliVity, and VMware VSAN. We found that in intelligent data placement, enterprise feature implementation, and simplified management, Nutanix offered a more thought-out solution for optimal performance, feature set, and management.

Key Findings

> Well-thought-out balance between data locality and distribution of data leverages HCI for customer benefit.

> Intelligent HCI-specific feature implementation.

> Simplified, 1-Click Management and 1-Click upgrades eliminates pain points.

> Innovation of Web-scale enterprise cloud in your own data center will continue with 100% software-defined continuous innovation.

> Simplified multiple hypervisor environments with VM mobility between hypervisors and license saving Nutanix AHV hypervisor.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

HCI Solution Feature Strength

	Data local to VM	Erasure Coding support	Control over deduplication and compression	Multiple generations of hardware in the same cluster	Replication Factor granularity	Multi-hypervisor support
Cisco Hyperflex	○	○	○	◉	○	○
HPE Simplivity	◉	○	○	○	○	○
VMware VSAN	○	●	◉	●	●	○
Nutanix Enterprise Cloud Platform	●	●	●	●	●	●

● - Rich support ◉ - Moderate support ○ - Weak support

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

HCI Definitions and Explanations

Because hyper-converged technology is new to many people, we provide the following list of definitions and explanations that will be used throughout this document.

Data Locality – the concept of keeping a complete copy of the data used by a virtual machine on the same node that is running the virtual machine.

Replication/Resiliency Factor (RF) – The number of data copies made when ensuring data redundancy and availability through making multiple replica copies.

Failure/Fault Tolerance (FT) level – The number of failures to be tolerated before a system goes offline. When employing RF for FT, typically an FT of 1 needs an RF of 2, and an FT of 2 needs an RF of 3. The term "failures to tolerate" (FTT) is also sometimes used.

Some HCI systems prefer to use RF terms and others prefer to use FT terms.

Erasure Coding/Encoding – The process that provides data resiliency, by spreading data blocks and parity information across multiple storage devices or systems. RAID is a form of Erasure Encoding where data is spread across local storage devices. For this paper, when we refer to Erasure Coding, we will be referring to the HCI-specific type that spans multiple nodes or systems. Erasure Encoding typically offers storage savings over employing RF for the same FT levels. For example, an FT 2 configuration with Erasure encoding will require 25% more space for the parity data, while an FT 2 configuration with RF 3 will require 200% more space for the replication data.

Inline data deduplication and compression – The data deduplication and compression performed before writing the deduplicated data to storage media. Typically, this is done in primary memory. Inline deduplication will require one I/O to write unique data and no I/Os for duplicate data.

There are differences with implementations of "inline" data deduplication. Some acknowledge the host write as soon as it arrives in the memory buffer and then perform the deduplication and/or compression in memory, while others do not return the acknowledgement until the deduplication and/or compression in memory has completed. In both cases, the deduplicated and/or compressed data is written (or de-staged) to storage media after the initial host write has been acknowledged.

Post-process data deduplication and compression – The data deduplication and compression performed after the data to be deduplicated has been initially written to disk. Post-process is typically undesirable because the data must be initially written and then re-read later so that a determination of the uniqueness of the data can be made. Then the data, if unique, must be re-written again to its final location. Each unique piece of data will require three sets of I/Os, and each duplicate piece of data will require two sets of I/Os. Some implementations use inline fingerprinting and post-process deduplication that may reduce the number of subsequent reads.

Pointer-based snapshot – A copy of data that is created using pointers to the original data plus the newest changes. A problem with this is that with higher snapshot depths, performance can be degraded as multiple pointers are followed in order to traverse the snapshot chain.

Block map update snapshot – Similar to pointer-based snapshots, except that instead of pointing back to older copies, a new block map is created for the snapshot. The block map points to the original blocks where there has been no change and to new blocks where there are changes. A chain of pointers does not need to be traversed and performance is not degraded.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

HyperFlex on Cisco UCS

Intelligent Data Placement and Recovery

Implementation

HyperFlex stripes all data for the original VM and replica copies across all nodes in the cluster. The idea is to avoid data locality and for the data to be evenly distributed in order to avoid network and storage hotspots. This also ensures all nodes in the cluster participate in rebuilds after node or drive failure.

Analysis

Due to the fully distributed striping across all nodes, excessive network traffic is frequently generated for normal VM operations. Read caches are also distributed, causing some read cache hits to be served over the network. This is an inefficient network design that will result in inconsistent performance, as some data will be local by accident while the rest will be remote. For all-flash implementations, the maximum cluster size is sixteen converged nodes and for hybrid deployments, the cluster size limit is eight converged nodes.

Data protection is accomplished through RF only and is set at cluster creation for all VMs in the cluster. If one VM needs RF 3, then the entire cluster will have to be RF 3. As with some other implementations, the replication factor is defined for an entire cluster and cannot be changed after the cluster is created. There is no erasure coding, so to tolerate two failures on a single VM, the entire cluster must have enough storage to efficiently hold three copies of the data.

Deduplication and Compression

Implementation

HyperFlex performs deduplication and compression when data is destaged from the SSD write cache or write buffer and placed in a persistent tier. The persistent tier is either HDD or SSD drives with larger capacity than the SSD cache devices. Deduplication and compression are always on for both all-flash and hybrid deployments.

Analysis

All customers will have access to standard enterprise features such as deduplication and compression.

No control over deduplication and compression is given to the customer, and no analysis of data streams is performed prior to deduplication and compression attempts to determine if the dataset is likely to be reducible. Processor will be wasted on irreducible data in the always on, non-configurable setup.

Upgrades and Hardware Replacement

HyperFlex advertises 1-click storage OS upgrades, however, they do not advertise simplified hypervisor upgrades. When adding new storage nodes, the new node must have the same number of disks, same capacity of disks, same Cisco HX Data Platform software version, and same ESX software version. Different CPUs, such as Intel v3 and v4, are allowed with VMware Enhanced vMotion Compatibility (EVC). Once again, when attempting to retire legacy hardware, the new replacement hardware cannot be added to the same cluster, and VMs must be manually migrated.

Snapshots and Backups

HyperFlex supports native replication and a Data Protection API, however no mention is made of data reduction techniques to be used. The bandwidth of data necessary on the WAN to support this new feature may prohibit backups and replication. There is also no mention of any cloud backup options.

Multi-Hypervisor Support

HyperFlex currently runs on vSphere 5.5 and 6.0 U1. There are plans to incorporate other hypervisors in future, but nothing at present. For now, the HyperFlex user is locked into VMware vSphere.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

HPE SimpliVity

Intelligent Data Placement and Recovery

Implementation

Simplivity works in conjunction with DRS within VMware to create host pairs for VM placement. This allows VMs to be spread across nodes within the cluster and attempts are made to evenly balance the distribution, along with achieving RF 2. However, there is no automatic rebalancing available. The local data copies are stored on RAID protected SSD and HDD.

Analysis

In this setup, legacy local hardware RAID is the primary means of data protection and capacity optimization. For a modern software-defined system, we would expect to find erasure coding. When there is a drive failure, the RAID group goes into a degraded state and the failed drive must be manually replaced before the node can start a rebuild. The node with the replaced drive is solely responsible for the rebuild, as a local RAID hardware rebuild cannot be shared between nodes. This adds length to the rebuild time. Using RAID 6, failures on two drives in a disk group can be tolerated, but during a RAID rebuild, the system is vulnerable to another drive failure. In this environment, the user cannot choose to increase to an RF 3 if they desire to avoid this problem, because Simplivity does not have the ability to keep three separate copies of data.

Furthermore, there is the problem of uneven distribution. While the totality of the cluster VM data will start out distributed evenly, all VM data must grow at the same rate or the cluster will become unbalanced. Over time, different VMs will grow at different rates, leading to a need for rebalancing so that these VMs don't run out of room simply because the space consuming VMs all happened to land on one node while the rest of the cluster has free space. Moving a VM to a different node may not achieve an even distribution of data. In this case, rebalancing must usually be done manually by SimpliVity support, will take a while to complete, and while rebalancing is running, cluster performance will be degraded.

Deduplication and Compression

Implementation

SimpliVity offloads compression and deduplication to a PCIe OmniStack Accelerator Card (OAC) so that CPU is not impacted. The OAC card has processor, NVRAM, and SimpliVity's deduplication, compression, and optimization algorithms on the hardware. For entry level systems, the card uses 57 GB of memory, and for larger systems, it can use 100 GB of memory, which can be more than other HCI solutions. Deduplication and compression are always on.

Analysis

Data is staged in the write buffer of the OAC card before being written to capacity HDDs. In this example, we eliminate the extra IOPs to our cache tier by sending data directly to a hardware write buffer. SimpliVity does not suffer from reduced processor performance due to compression and deduplication, however, the trade-off is reduced memory available on each node due to the OAC card.

Again, no control over deduplication and compression is given to the customer, and no analysis of data streams is performed prior to deduplication and compression attempts to determine if the dataset is likely to be reducible. Memory will be wasted in irreducible data in the always on, non-configurable setup.

Upgrades and Hardware Replacement

SimpliVity advertises a version-specific Upgrade Manager software to facilitate upgrading SimpliVity storage management for an entire federation, or group of clusters, at one time in parallel with "a few clicks." SimpliVity does not advertise simplified hypervisor upgrades, which is a separate management task. In addition, upgrades to the storage O.S. will also require upgrades to the Upgrade Manager, Witness software and the vCenter management plugin.

As for hardware upgrades, when new nodes are added, only nodes of the same product family, with equal socket count and the same CPU model, can be added to the same data center.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

Snapshots and Backups

SimpliVity creates "integrated" or full logical backups that are comprised of metadata only. These backups are similar to the snapshots discussed before, as no second copy of the actual data is made. This can be misleading to new customers who think an actual data copy has been made instead of a pointer-based backup, or snapshot.

SimpliVity also has SimpliVity Rapid DR to automate disaster recovery as well as cloud backup to AWS only, offering some other options for full logical backups.

When performing these backups, SimpliVity employs metadata analysis as a data reduction technique. SimpliVity likes to call their metadata analysis "WAN Optimization," to remind the customer of how this could replace a WAN Optimization appliance, further simplifying the data center. However, according to Gartner, WAN optimizers do a lot more than just data reduction, but rather include protocol and application-specific optimizations, employ traffic identification prioritization, policing, and shaping features, provide traffic monitoring, and provide WAN path control across internet and MPLS.

Multi-Hypervisor Support

Previously, Simplivity supported multiple hypervisors. However, since Simplivity was acquired by HPE, they currently only support VMware vSphere.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

VMware VSAN

Intelligent Data Placement and Recovery

Implementation

In VMware, VSAN nodes employ disk groups. Each disk group is composed of two tiers: a capacity tier and a cache tier. There are up to seven devices for each capacity tier and one flash device for each cache tier. In the case of hybrid systems with HDD and SSD, the SSD device is the cache tier, and the HDD devices are in the capacity tier. The HDDs use the cache tier as a write buffer and read cache. In an all SSD system, a higher endurance SSD may be used as the cache tier device while lower endurance SSDs may be used in the capacity tier. The lower endurance flash will use the cache tier as a write buffer. The default disk group configuration is a single disk group for all storage devices in the node.

Storage policies define the striping and RF. A storage policy is then assigned to a VM, and VMware distributes the VM data accordingly.

In all-flash clusters with the advanced license, storage policies may employ FT1, FT2 or RAID 5/6 with erasure codding. This is not a local hardware RAID but is a RAID across disks hosted on multiple nodes across the cluster, containing data for a single VM.

Analysis

The disk group forms a storage failure domain, where the failure of a cache tier flash device will fail the entire disk group. In the default disk group configuration (where all node storage is in one disk group), the failure of a cache tier device will fail all the storage on a node. Rather than just rebuilding one drive, the entire disk group will have to be rebuilt by the cluster, leading to long rebuild times during which the system will be vulnerable if there is another failure, depending on the replication factor that was used.

VMware does allow RF to be changed on the fly (providing there are enough resources). Their erasure coding spreads individual VM data across all nodes participating in the stripe. Consequently, there will be high-bandwidth network traffic due to data requests. This will increase network latency onto the low-latency cache tier.

An additional license is necessary in order to employ erasure coding for space savings. This additional license also provides support for data deduplication and compression. Besides, the feature is not available for hybrid clusters. Many hybrid customers with FT 2 storage policies will have a 175% increase in storage needed due to being limited to data protection through RF only.

Deduplication and Compression

Implementation

VMware vSAN performs deduplication and compression after write acknowledgement. Deduplication occurs when data is de-staged from the cache tier to the capacity tier of an all-flash Virtual SAN datastore. The compression algorithm is applied after deduplication has occurred just before the data is written to the capacity tier. Deduplication and compression are enabled or disabled together in a single cluster-wide setting and are implemented at a disk group level. Data is deduplicated within a disk group but not across multiple disk groups, so it is possible to have the same redundant data in multiple disk groups on the same node.

Analysis

Deduplication and compression, which are standard enterprise features, are not available to the hybrid VSAN customer, only all-flash configurations.

Granularity of control is not given to the customer. If the customer knows that one VM will have mostly irreducible data, they cannot exclude that data from going through compression and deduplication attempts if the deduplication and compression is configured at the cluster level. No analysis of data streams is performed to determine if the dataset is likely to be reducible before attempting to deduplicate or compress. Processor will be wasted on irreducible data in this all on, all off approach.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

Upgrades and Hardware Replacement

VMware vSAN is implemented in the kernel, and both hypervisor and storage management upgrades require a reboot. Before reboot, the VMs must be manually migrated to another node and the node must then be put into maintenance mode. After upgrade, the node must be taken out of maintenance mode and the VMs migrated back. This process must be repeated for each node in the cluster. Even for small clusters, this is a lengthy upgrade process. When new nodes are added, VMs must be manually migrated in order to rebalance VM and storage load, and when a legacy node is retired, the VMs must be manually migrated to a new node before the legacy node is removed.

Snapshots and Backups

VMware snapshots are pointer-based, which can affect performance as the snapshot depth increases.

VSAN has native vSphere Replication or vSphere Data Protection to replicate data to a remote Disaster Recovery (DR) location. To reduce bandwidth, vSphere Replication has metadata analysis, and vSphere Data Protection has deduplication and compression. A VM can use either Replication or Data Protection as their data reduction, but not both.

Multi-Hypervisor Support

Because of in-kernel integration, VSAN only runs on VMware vSphere and is not expected to support other hypervisors.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

Nutanix Enterprise Cloud Platform on Cisco UCS and HPE ProLiant

Intelligent Data Placement and Recovery

Implementation
Nutanix allows RF 2, RF 3 or RF3 plus erasure coding to be chosen.

When employing RF, Nutanix makes sure that one of the copies of the VM data is on the local node for the VM, while the other copy or copies are distributed across other nodes in the cluster.

Erasure coding is usually employed on RF 3 data. After the data is initially written, data from multiple VMs are gathered together as a stripe, parity is calculated and placed on new cluster nodes, and non-local replica copies are erased.

Nutanix does automatic rebalancing to make sure that even on heterogenous clusters, data is distributed evenly.

Analysis
All data nodes participate in rebuild upon disk or node failure, so rebuilds take less time and the cluster is in a failure-vulnerable state for a shorter time. This is the only solution that uses multiple nodes for rebuild while still providing a local data copy for every VM running. Other solutions that leverage multi-node rebuilds require VMs to regularly go across the network for their everyday traffic.

Due to data locality, most storage data will not need to go across the network to other nodes; in most cases, data can be pulled from the local flash devices. In the rare case that local access fails, the ILM transparently moves frequently accessed remote data to the local controller, correcting the issue. Data locality allows the VM to fully take advantage of low-latency flash, instead of adding a high-latency network access on top of a low-latency flash access as it goes to other nodes for its data.

RF can be changed on the fly for individual VMs without compromising data locality.

Erasure coding is available to all types of systems, providing data protection without a large increase in storage capacity required. Erasure coding with RF 3 requires five nodes.

Max cluster size is unlimited, and cluster data is automatically rebalanced by the Nutanix system. Due to data locality reducing network chattiness significantly, data re-localization and rebalancing can be done efficiently.

The Nutanix system is self-healing, never failing an entire node of storage due to a problem with one disk. If a disk fails to the point where it needs to be replaced, the parity information is used to rebuild the data from that drive elsewhere in the cluster. The failed drive can be replaced during the next maintenance cycle, adding new capacity to the storage pool.

Deduplication and Compression

Implementation
The Nutanix Capacity Optimization Engine (COE) offers two levels of compression:

> *inline compression* - Nutanix analyses the data to see if it is likely to be reducible. Sequential streams of data or data with large I/O sizes are compressed in memory before being written to the extent store. Random I/O's are written to the flash OpLog and coalesced before compression is performed and they are written to the extent store.
>
> *offline compression* - after data has become cold and written in the HDD tier, a post-process compression will be performed.

Nutanix does fingerprinting for deduplication at ingest when the data has large I/O size and stores the fingerprint on flash. For smaller I/O sizes fingerprinting is done as a background process so that no additional I/O is required to identify data that is eligible for deduplication. The Nutanix Elastic Dedup engine has two functions:

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

> Post-process deduplicaton on the cooled capacity tiers using the fingerprint already stored on performance media. When the process is complete, data is not only deduped on the node, but is deduped across the cluster as well.

> Inline deduplication on read in the performance tier's unified cache.

Nutanix makes recommendations to the customer as to which type of deduplication and compression will be most valuable to individual customer scenarios.

Analysis
Nutanix attempts to categorize the data stream based on I/O size and randomness in order to only attempt compression and deduplication on sequential or pre-coalesced data instead of wasting resources attempting to reduce random small blocks. For data that is most likely to benefit from deduplication, Nutanix can fingerprint the data at ingest, so as to not have to re-read the data later to determine if it is a duplicate. This saves IOPs. The post-process type of duplication is performed on the cooled capacity tiers instead of the flash tiers, extending the life of the flash.

Multiple options and flavors of reduction are available to the customer, making this platform fully "tunable" to the kind of workloads the end user is running. This is an intelligently implemented deduplication and compression for the HCI environment.

Upgrades and Hardware Replacement
Nutanix advertises 1-click storage upgrades, 1-click hypervisor upgrades and 1-click firmware upgrades. The user simply logs in and accesses the management screen. From there, a single click by IT staff can kick off an upgrade process for the entire cluster.

Nutanix uses a 100% software-defined implementation, where a Controller Virtual Machine or CVM is deployed on each node to handle storage traffic. Storage upgrades use CVM autopathing so that VMs do not have to be migrated during upgrades. Instead, the VMs stay on the node being updated, and I/O is redirected to a CVM on another node while the current node CVM is updating. The VM and data do not have to be migrated during CVM upgrade. This implementation allows upgrades to be performed during normal business hours.

Hypervisor upgrades are applied each node one at a time, automatically, in the form of a rolling cluster upgrade. However, hypervisor updates usually do require a system restart.

Nutanix can have multiple generations of hardware in the same cluster. When new nodes are added, the storage and CPU configurations do not need to match. Nutanix clusters can have heterogeneous configurations and will automatically rebalance between nodes with different hardware generation and/or storage capacity. Once newer nodes are added, legacy hardware can be removed from the cluster and Nutanix will automatically rebalance. No manual VM migration is required.

Snapshots and Backups
Nutanix does snapshots and clones with block map updates, eliminating the performance penalty from legacy pointer-based snapshots.

Nutanix offers snapshots and VSS backups natively in addition to the ability to interface with multiple third-party backup tools. Cloud backup to AWS and Azure is also available. In short, Nutanix offers the most flexibility in backup options.

When replicating, Nutanix also employs metadata analysis as a data reduction technique, calling it "Global Data Awareness."

Multi-Hypervisor Support
Nutanix runs on vSphere, Hyper-V, Xen Server, and its own AHV hypervisor (which is their own Nutanix version of KVM provided free of charge.) Each cluster has a single hypervisor, and Nutanix offers the ability to transfer a VM between these clusters from one hypervisor to another. Moreover, the AHV hypervisor can run native on storage nodes. AHV is built into the platform for free and can save customers significant money in VMware hypervisor licensing. In addition, backups and disaster recovery procedures can be the

same regardless of hypervisor type. Nutanix Prism Central can manage multiple clusters, each running a different hypervisor, from a single console.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

Demartek
August 2017

Summary and Conclusion

Nutanix provides customers a wealth of features that use the same system management procedures regardless of hypervisor or generation of underlying server and storage hardware. The design of the Nutanix hyper-converged platform lends itself well for future growth. These benefits provide improved productivity for customers who use or manage Nutanix platforms.

Nutanix has a well-thought-out balance between locality and distribution of data which provides superior performance due to preservation of data locality. At the same time, replicated copies of data are distributed across the cluster, enabling rapid self-healing in the event of failure. Intelligent implementation of compression and deduplication features attempts data reduction on ingest only when it determines a likely benefit. Erasure coding provides RF 3 while requiring as low as 25% more space and preserving data locality for performance benefit.

Nutanix offers simplified management with a fast-fail, scale-out, self-healing Enterprise Cloud environment. The enhanced end-user experience with more flexibility and granularity of control, automatic cluster rebalancing, 1-click automation, native replication and multi-hypervisor support is the result of the continuous innovation available from a 100% software-defined solution. This Web-scale enterprise cloud in your own data center will continue to evolve as more features are added and refined rapidly in the future.

The most current version of this report is available at http://www.demartek.com/Demartek_Nutanix_Hyperconverged_Infrastructure_2017-08.html on the Demartek website.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

APPENDIX – General Concepts

Intelligent Data Placement and Recovery

VM performance is best when there is data locality, eliminating the need to regularly pull data across the network from a different node. Without data locality, networks can be flooded with storage network traffic, and low-latency flash benefits from the cache or performance tier are lost when network latency is added on. As new ultra-low latency flash and other emerging non-volatile memory technologies are deployed for storage, reducing storage latencies to below 100 microseconds (< 100µs), data locality becomes increasingly important for taking advantage of future performance benefits.

After node or drive failure, cluster performance often degrades as resources are devoted to rebuild. Moreover, during rebuild, the system is in a compromised state and is vulnerable should another failure occur. The best systems will have mechanisms in place to reduce rebuild times. This can be achieved by having more nodes participate in the rebuild process and by limiting storage failure domains so less data has to be rebuilt.

When taking these factors into consideration, it follows that an erasure coding configuration with data and parity fragments distributed across the cluster provides the data protection and space savings best suited to newer, innovative software-defined HCI environments. Other methods of data protection, such as local hardware RAID, require manual drive replacement and only utilize one node for the local rebuild.

HCI-Specific Implementation of Enterprise Features: Deduplication and Compression

When considering data deduplication and compression, several factors must be considered. The type of application data to be stored affects the efficiency of the deduplication and compression processes. Some systems always deduplicate and compress data, for others, this is configurable. Some systems perform these functions upon initial ingest of the data, others perform these later in the processing. Some systems offload some of these processes onto separate hardware adapters and others use the main host CPU. The specific algorithms used for deduplication and compression must also be taken into account because of the various block sizes and fingerprinting used. The scope of these functions also varies from local node to global across the cluster.

The processor resources required for inline data reduction can impact cluster performance, and as a result, many HCI vendors implement a combination of inline and post-process data reduction or have offloaded deduplication and compression to hardware.

We must look at how each solution manages the trade-off between IOPs consumed in post-process and performance hits from inline. An intelligent, HCI-Specific implementation of compression and deduplication will allow the customer to turn features on and off, assuming the customer knows how dedupable and compressible their data is and if they want to consume processor cycles on these tasks. The solution will also attempt to analyze data and not waste processor resources attempting to compress or deduplicate a dataset that is not reducible. Ultimately, the solution should attempt to somehow limit the extra IOPs used in post-process, especially when data is being written to flash.

Simplification of Management

The management structure of an HCI environment should be Web-scale. In short, routine maintenance should not create pain-points for the IT staff but should be streamlined, automated, and simplified as much as possible, enabling IT staff to manage many VMs in less time.

Nutanix Delivers Intelligent Hyper-converged Infrastructure that Simplifies the End User Experience

Upgrades and Hardware Replacement

Hyper-converged environments require many different upgrades: hypervisor upgrades, storage management upgrades, firmware upgrades, and hardware upgrades. Simplified, streamlined upgrade processes that do not require a lot of time from IT staff are necessary for each of these scenarios. Staff should not be repeating the same lengthy upgrade process for every node in a Web-scale environment. In short, hypervisor updates should be automated and not require manual migration of VMs. Storage system updates should not require VM migration or host reboot, and clusters should automatically rebalance load and storage when nodes are added and removed in a hardware upgrade process.

The solution should also allow for heterogeneous clusters for flexibility in management and hardware upgrades.

Snapshots and Backups

For customers evaluating HCI, data protection capabilities such as snapshots and backups are important, including data and virtual machines. Similarly, all hyper-converged systems need some kind of full logical backup or replication. Snapshots (or pointer-based backups) are useful for creating local recover points but do not offer a true off-cluster copy of the data that can be used for recovery in the event the primary cluster is lost. For this use case, customers may replicate snapshots to an alternate site or use VSS based backup solutions that store backups on a secondary storage device or public cloud. HCI solutions should include a number of different data protection capabilities to meet different customer requirements and not depend solely on pointer-based backups.

In addition, data reduction techniques are usually implemented to reduce bandwidth as this sort of traffic is most likely going over the WAN. Common types of data reduction are deduplication, compression, and metadata analysis to eliminate sending data over the WAN that is already present at the other end.

Multi-Hypervisor Support

According to some industry analyst data, as many as half or more of IT shops deploy multiple hypervisors because some applications require particular hypervisors, some groups have expertise with particular hypervisors or because of possible pricing or licensing advantages. This type of multi-hypervisor environment requires more staff, training, and time to maintain, however, due to the large array of different applications necessary to run a business, it has become unavoidable.

Any hyper-converged solution that can help eliminate the pain points, extra staffing, time, and training requirements necessary to manage these environments will help the end-user. A helpful hyper-converged solution will run on multiple hypervisors and will offer tools and utilities to help deal with multi-hypervisor environments.