HCI Is Not Always As Simple As It Seems

Executive Summary

The major selling point around HCI is simplicity. Simplicity in design, simplicity in purchasing, simplicity in management, simplicity in scaling. The reality is often less than simple. HCI is designed around running mixed workloads on virtual machines (VMs). But in order to maintain endurance, and reliability, HCI administrators are often challenged by the number of decisions they need to make to simply provision and deploy a single VM.

In this Insight, Neuralytix explores a solution from Datrium that attempts to solve the complexity of HCI and truly deliver a high performance, scalable and simple approach to designing and deploying infrastructure.

Contents

Introduction

Hyperconverged Infrastructure (HCI) has brought simplification to the otherwise highly complex environment we call a datacenter.

Through the use of software, discreet resources of a datacenter, including compute, network, storage, management, power, cooling, and others, have been brought under a single piece of software that aggregates all these resources, and allows administrators to manage them collectively from a single pane of glass.

The software driving the datacenter (or software-defined data center [SDDC]) further simplifies the datacenter through the use of policies and automation to streamline provisioning and deployment of virtual servers, storage, and networks. Through clustering, datacenters can be scaled beyond a single server on demand, and through the use of industry standard components, HCI and SDDC has dramatically reduced the cost of owning and running the datacenter.

Hypervisors such as VMware ESXi, Microsoft Hyper-V, and KVM have allowed virtual servers and even virtual datacenters to be visually composed and deployed in a matter of minutes. These deployments can be made into templates that allow multiples virtual servers to be stamped out in a matter of seconds.

However, configurability, manageability, serviceability, performance and scalability is still complicated with HCI. The key reason for this are the numerous decisions that admins are forced to consider concerning storage. What type of durability should be used? Should it be mirroring, erasure coding, or RAID 5/6? What about data reduction, what is better? Compression, deduplication, or both? Should data be only stored locally, or can be distributed across multiple nodes? But if data is distributed, will cross-talk across nodes impact latency and performance? Should data be on solid state media, traditional hard disk drives, or a hybrid?

These and other questions need to be answered before a virtual server's storage subsystem is optimized. In many instances, administrators have to make many changes, which require the migration of data, to tune the storage subsystem for each distinct application.

In this Insight, we look at several key aspects to the storage dilemma, and what it takes to optimize each of these aspects:

- Data durability;
- Data locality;
- Media type;
- Data reduction;
- Encryption.

By no means is this list exhaustive, but for most applications and virtual servers created, they are major drivers. Furthermore, each of these aspects impact performance, the most important characteristic in technology in terms of achieving competitive advantage. High performance leads to market leadership, and can also increase the speed of failures, so that new hypotheses can be tested, and old hypotheses revised.

The Complexity of Storage

The complexity of storage in a HCI environment cannot be overstated. One example is the lock-step nature of scaling of compute and storage capacity.

Since storage is merged into compute nodes, scaling an HCI cluster by adding an additional node means scaling both compute and capacity. This restricts scaling flexibility and increases costs.Current HCI solutions, such as Dell EMC's VxRAIL, Nutanix's NX series, and HPE SimpliVity, all suffer from this and other challenges such as the ones we highlight below.

Data Durability

In any datacenter or any architecture for that matter, storage is the most critical component. There is no point having the fastest computing capacity, or broadest bandwidth if there is no data to process!

In HCI, since everything is software-defined, the durability of the data is in the hands of the hosts. When designing for durability, you essentially start designing for the management of failures.

In a HCI environment, the first thing one must solve for is what happens if a host goes down. One way to solve for this, is to overprovision – have extra copies of the data across the cluster. While disk capacity is relatively cheap, configuring for up to 2 nodes to fail requires three copies of data.

If you are using VMware vSAN their sizing guide requires that not only do you need to configure for failures, but you have to have enough capacity to rebuild.

Flash is often used as both a read and write cache. In the case of vSAN, they recommend that you cache for 10% of capacity (15% if using snapshots) and recommend a 70%/30% ratio for read/write, and for additional durability, that means mirroring the cache. In effect, you end up either buying twice the amount of cache, or lose half the cache for redundancy!

Ideally, the cache, and the capacity tier should be entirely flexible. But in vSAN, it has a highly-defined combination of one cache drive to 7 capacity drives. Furthermore, if the cache drive fails, the whole disk group needs to be rebuilt, not just the failed device. In effect, there is a need to do a storage vMotion to a new disk group.

Finally, under a vSAN environment, they recommend a homogeneous cluster – where every node has the same cache and same number of disk drives. So, mixing storage and compute intensive workloads on one cluster is not recommended.

Another consideration is the type of data protection that is done – erasure coding or mirroring. In the above example, we talked about triple mirroring which can be a significant investment. Whether mirroring or erasure coding is done, it requires a lot of CPU and memory.

With erasure coding, you are writing to a lot of nodes, and a lot more processing is done because of parity calculation and check-summing. This also results in increased cross-talk across the network.

Some of the leading HCI vendors were not originally designed for erasure coding, which means that while erasure coding may be beneficial, especially financially, it is not truly native to those environments and can come with unsustainable performance impacts.

Data Locality

Where data is located affects performance. Since most HCI vendors use either mirroring or erasure coding, the data is often spread across the cluster, that impacts on the performance of the workload and the overall cluster.

Ideally, the data should stay local to the server hosting the workload.

Media Type

The type of media used can also affect performance. In some HCI configurations, such as vSAN, they only recommend SAS drives over SATA drives, increasing cost.

vSAN also advises not to mix all-flash and hybrid disk groups on the same node, further reducing flexibility.

Data Reduction

Most HCI vendors suffer from weaknesses in two other key data services, again because they were "bolted on" after the original storage layer was designed – these data services are deduplication and compression.

Since they are not native, most vendors do not enable these services by default, and in fact, for certain workloads, they even advise customers to turn off either or both completely!

That increases the cost of storage, and reduces flexibility and performance.

Encryption

With more and more customers who are using HCI operating in regulated industries, the concept of encryption is no longer a secondary concern.

Some HCI vendors recommend the use of self-encrypting drives, such as Nutanix, which is only good for data at rest. Self-encrypting drives also create silos that are costly and reduces flexibility in design.

Solving the dilemma – Datrium DVX

Datrium's approach to convergence and especially storage solves all the problems listed above with HCI.

Datrium, a privately held storage company, founded in 2012, has a new approach to solving the storage problem associated with HCI. Its solution is what Datrium calls *Open Convergence*.

It claims that its approach is "a radical new way to enable a simpler journey to hybrid clouds, but without the rigidity of traditional convergence, and without the lock-in and scaling unpredictability of HCI."

Datrium brings a unique approach to convergence through its DVX System. DVX separates data that is in process on compute nodes with durable persistent data on data nodes.

Data Durability

Referring to Figure 1, a safe copy of all data is stored on the data node, which is a highly redundant dual controller, dual NVRAM array-like appliance. This means that the compute node stands independent of data availability and are stateless. This simplifies host maintenance and failovers significantly as hosts can fail or be put into service without affecting data durability or performance. No

rebuilds, no sizing for rebuilds or disk groups to manage.

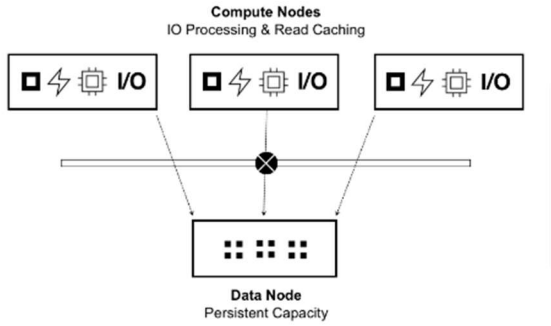

Figure 1: Datrium Block Diagram (Datrium 2017)

DVX also allows compute and storage to truly scale independently. No longer do enterprises have to buy more compute in order to get more storage, or *vice versa*.

A key to DVX success is that it is VM-centric. DVX's I/O and fault isolation technology simplifies workload deployments, and its management software provides an in-depth real-time view of all hosts and VMs.

Data Locality

All data local to the host is cached for reads in local flash media. 100% of flash is available for this (with no protection overheads) and data is deduplicated or compressed before being written to local flash so effective capacity is large. A single terabyte of flash can support 2TB to 6TB of user data, making it affordable to keep as much flash in the server as needed for active VM data.

Consequently, data is always local to its VM, and results in the ability to predictably meet service level agreements (SLAs) on a per-VM/application basis.

Under normal operation, all I/O in DVX is between the hosts and data node (in the north–south). The hosts don't cross-talk eliminating any noisy-neighbor issues. Through I/O isolation between hosts, running mixed workloads is greatly improved.

Media Type

The media type used for persistence is highly cost optimized. Low cost commodity read-intensive SATA drives with compression and global deduplication provide an extremely cost effective secondary storage repository not just for primary data, but for all secondary copies used for backup/restore, disaster recovery and copy data management. There is full media freedom – SATA, SAS or PCIe SSDs can be used depending on workload performance and latency requirements.

Data Reduction

With Datrium, compression, deduplication, and erasure coding is always on, simplifying user choices (and related performance impacts) about what data reduction to use where.

There is no need to add extra server memory and make workload-based choices to turn on these services, and the performance overhead is negligible because these data services were engineered into the architecture from day 1.

Encryption

Because the architecture extends from the server through to shared capacity, Datrium can offer a level of encryption not possible with previous convergence approaches. Its encryption spans compute nodes, network and data nodes for in-use, in-flight and at-rest protection while providing full data reduction.

Since the encryption is software-based, there is no need for special hardware. Encryption resources are provided by all hosts in the cluster, which allows the capability to scale as more hosts are added. Datrium also uses the AES-NI instruction set in modern processors to offload the encryption work, so performance overhead is negligible.

Guidance and Conclusion

Datrium, DVX, and Open Convergence brings a unique solution to the data storage problems commonly associated with HCI.

As easy as HCI seems, there are enterprises are still confounded with too many choices, and too many "knobs" to optimize on a per VM basis.

Datrium solves this problem by being VM-centric, and through its I/O and fault isolation; compute and data persistence separation; and the high performance resulting from dedicating flash to processing.

Neuralytix believes that Datrium brings a very compelling story to convergence. The background of the founders of Datrium is in both data storage and hypervisors, and their understanding of data movement and data services has allowed them to deliver a solution that focuses on optimizing data to the application.

They have also recognized the limitations of HCI – such as simplicity and multi-workload environments, and have designed an elegant solution that enhances the converged system experience without sacrificing performance, scaling or reliability.

Notes