Article
Murray Oldfield · Nov 25, 2016 23m read

InterSystems Data Platforms and performance – Part 8 Hyper-Converged Infrastructure Capacity and Performance Planning

Hyper-Converged Infrastructure (HCI) solutions have been gaining traction for the last few years with the number of deployments now increasing rapidly. IT decision makers are considering HCI when scoping new deployments or hardware refreshes especially for applications already virtualised on VMware. Reasons for choosing HCI include; dealing with a single vendor, validated interoperability between all hardware and software components, high performance especially IO, simple scalability by addition of hosts, simplified deployment and simplified management.

I have written this post with an introduction for a reader who is new to HCI by looking at common features of HCI solutions. I then review configuration choices and recommendations for capacity planning and performance when deploying applications built on InterSystems data platform with specific examples for database applications. HCI solutions rely on flash storage for performance so I also include a section on characteristics and use cases of selected flash storage options.

Capacity planning and performance recommendations in this post are specific to *VMWare vSAN*. However vSAN is not alone in the growing HCI market, there are other HCI vendors, notably *Nutanix* which also has an increasing number of deployments. There is a lot of commonality between features no matter which HCI vendor you choose so I expect the recommendations in this post are broadly relevant. But the best advice in all cases is to discuss the recommendations from this post with HCI vendors taking into account your application specific requirements.

A list of other posts in the InterSystems Data Platforms and performance series is here.

What is HCI?

Strictly speaking converged solutions have been around for a long time, however in this post I am talking about current HCI solutions for example from Wikipedia: "Hyperconvergence moves away from multiple discrete systems that are packaged together and evolve into **software-defined** intelligent environments that all run in **commodity, off-the-shelf x86 rack servers**...."

So is HCI a single thing?

No. When talking to vendors you must remember HCI has many permutations; Converged and Hyper-converged are more a type of architecture not a specific blueprint or standard. Due to the commodity nature of HCI hardware the market has multiple vendors differentiating themselves at the software layer and/or other innovative ways of combining compute, network, storage and management.

Without going down too much of a rat hole here, as an example solutions labeled HCI can have storage inside the servers in a cluster or have more traditional configuration with a cluster of servers and separate SAN storage -- possibly from different vendors -- that has also been tested and validated for interoperability and managed from a single control plane. For capacity and performance planning you must consider solutions where storage is in an array connected over a SAN fabric (e.g. Fibre Channel or Ethernet) have a different performance profile and requirements to the case where the storage pool is software defined and located inside each of a cluster of server nodes with storage processing on the servers.

So what is HCI again?

For this post I am focusing on HCI and specifically *VMware vSAN* where *storage is physically inside the host servers*. In these solutions the HCI software layer enables the internal storage in each of multiple nodes in a cluster performing processing to act like one shared storage system. So another driver of HCI is even though there is a cost for HCI software there could also be significant savings using HCI when compared to solutions using enterprise storage arrays.

> For this post I am talking about solutions where HCI combines compute, memory, storage, network and management software into a cluster of virtualised x86 servers.

Common HCI characteristics

As mentioned above *VMWare vSAN* and *Nutanix* are examples of HCI solutions. Both have similar high level approaches to HCI and are good examples of the format:

- *VMware vSAN* requires VMware vSphere and is available on multiple vendors hardware. There are many hardware choices available but these are strictly dependent on VMware's vSAN Hardware Compatibility List (HCL). Solutions can be purchased prepackaged and preconfigured for example EMC VxRail or you can purchase components on the HCL and build-your-own.
- *Nutanix* can also be purchased and deployed as an all-in-one solution including hardware in preconfigured blocks with up to four nodes in a 2U appliance. Nutanix solution is also available as a build-your-own software solution validated on other vendors hardware.

There are some variations in implementation, but generally speaking HCI have common features that will inform your planning for performance and capacity:

- Virtual Machines (VMs) run on hypervisors such as VMware ESXi but also others including Hyper-V or Nutanix Acropolis Hypervisor (AHV). Nutanix can also be deployed using ESXi.
- Host servers are often combined into blocks of compute, storage and network. For example a 2U Appliance with four nodes.
- Multiple host servers are combined into a cluster for management and availability.
- Storage is tiered, either all-flash or a hybrid with a flash cache tier plus spinning disks as a capacity tier.
- Storage is presented as a pool which is software defined including data placement and policies for capacity, performance and availability.
- Capacity and IO performance are scaled by adding hosts to the cluster.
- Data is written to storage on multiple cluster nodes synchronously so the cluster can tolerate host or component failures without data loss.
- VM availability and load balancing is provided by the hypervisor for example vMotion, VMware HA, and DRS.

As I noted above there are also other HCI solutions with twists on this list such as support for external storage arrays, storage only nodes... the list is a long as the list of vendors.

HCI adoption is gathering pace and competition between the vendors is driving innovation and performance improvements. It is also worth noting that HCI is a basic building block for cloud deployment.

Are InterSystems' products supported on HCI?

It is InterSystems policy and procedure to verify and release InterSystems ’ products against processor types and operating systems including when operating systems are virtualised. Please note InterSystems Advisory: Software Defined Data Centers (SDDC) and Hyper-Converged Infrastructure (HCI).

For example: Caché 2016.1 running on Red Hat 7.2 operating system on vSAN on x86 hosts is supported.

Note: If you do not write your own applications you must also check your application vendors support policy.

vSAN Capacity Planning

This section highlights considerations and recommendations for deployment of *VMware vSAN* for database applications on InterSystems data platforms -- Caché, Ensemble and HealthShare. However you can also use these recommendations as a general list of configuration questions for reviewing with any HCI vendor.

VM vCPU and memory

As a starting point use the same capacity planning rules for your database VMs' vCPU and memory as you already use for deploying your applications on VMware ESXi with the same processors.

As a refresher for general CPU and memory sizing for Caché a list of other posts in this series is here: Capacity planning and performance series index.

One of the features of HCI systems is very low storage IO latency and high IOPS capability. You may remember from the 2nd post in this series the hardware food groups graphic showing CPU, memory, storage and network. I pointed out that these components are all related to each other and changes to one component can affect another, sometimes with unexpected consequences. For example I have seen a case of fixing a particularly bad IO bottleneck in a storage array caused CPU usage to jump to 100% resulting in even worse user experience as the system was suddenly free to do more work but did not have the CPU resources to service increased user activity and throughput. This effect is something to bear in mind when you are planning your new systems if your sizing model is based on performance metrics from less performant hardware. Even though you will be upgrading to newer servers with newer processors your database VM activity must be monitored closely in case you need to right-size due to lower latency IO on the new platform.

Also note, as detailed later you will also have to account for software defined storage IO processing when sizing *physical host* CPU and memory resources.

Storage capacity planning

To understand storage capacity planning and put database recommendations in context you must first understand some basic differences between vSAN and traditional ESXi storage. I will cover these first then break down all the best practice recommendations for Caché databases.

vSAN storage model

At the heart of vSAN and HCI in general is software defined storage (SDS). The way data is stored and managed is very different to using a cluster of ESXi servers and a shared storage array. One of the advantages of HCI is there are no LUNs, instead pool(s) of storage that are allocated to VMs as needed with policies describing capabilities for availability, capacity, and performance per-VMDK.

For example; imagine a traditional storage array consisting of shelves of physical disks configured together as various sized disk groups or disk pools with different numbers and/or types of disk depending on performance and availability requirements. Disk groups are then presented as a number of logical disks (storage array volumes or LUNs) which are in turn presented to ESXi hosts as datastores and are formatted as VMFS volumes. VMs are represented as files in the datastores. Database best practice for availability and performance recommends at minimum separate disk groups and LUNs for database (random access), journals (sequential), and any others (such as backups or non-production systems, etc).

vSAN is different; storage from the vSAN is allocated using storage policy-based management (SPBM). Policies can be created using combinations of capabilities, including the following (but there are more);

- Failures To Tolerate (FTT) which dictates the number of redundant copies of data.
- Erasure coding (RAID-5 or RAID-6) for space savings.
- Disk stripes for performance.
- Thick or thin disk provisioning (thin by default on vSAN).
- Others...

VMDKs (individual VM disks) are created from the vSAN storage pool by selecting appropriate policies. So instead of creating disk groups and LUNs on the array with a set attributes, you define the capabilities of storage as policies in vSAN using SPBM; for example "Database" would be different to "Journal", or whatever others you need. You set the capacity and select the appropriate policy when you create disks for your VM.

Another key concept is a VM is no longer a set of files on a VMDK datastore but is stored as a set of *storage objects*. For example your database VM will be made up of multiple objects and components including the VMDKs, swap, snapshots, etc. vSAN SDS manages all the mechanics of object placement to meet the requirements of the policies you selected.

Storage tiers and IO performance planning

To ensure high performance there are two tiers of storage;

- Cache tier - Must be high endurance flash.
- Capacity tier - Flash or for hybrid uses spinning disks.

As shown in the graphic below storage is divided into tiers and disk groups. In vSAN 6.5 each disk group includes a single cache device and up to seven spinning disks or flash devices. There can be up to five disk groups so possibly up to 35 devices per host. The figure below shows an all-flash vSAN cluster with four hosts, each host has two disk groups each with one NVMe cache disk and three SATA capacity disks.

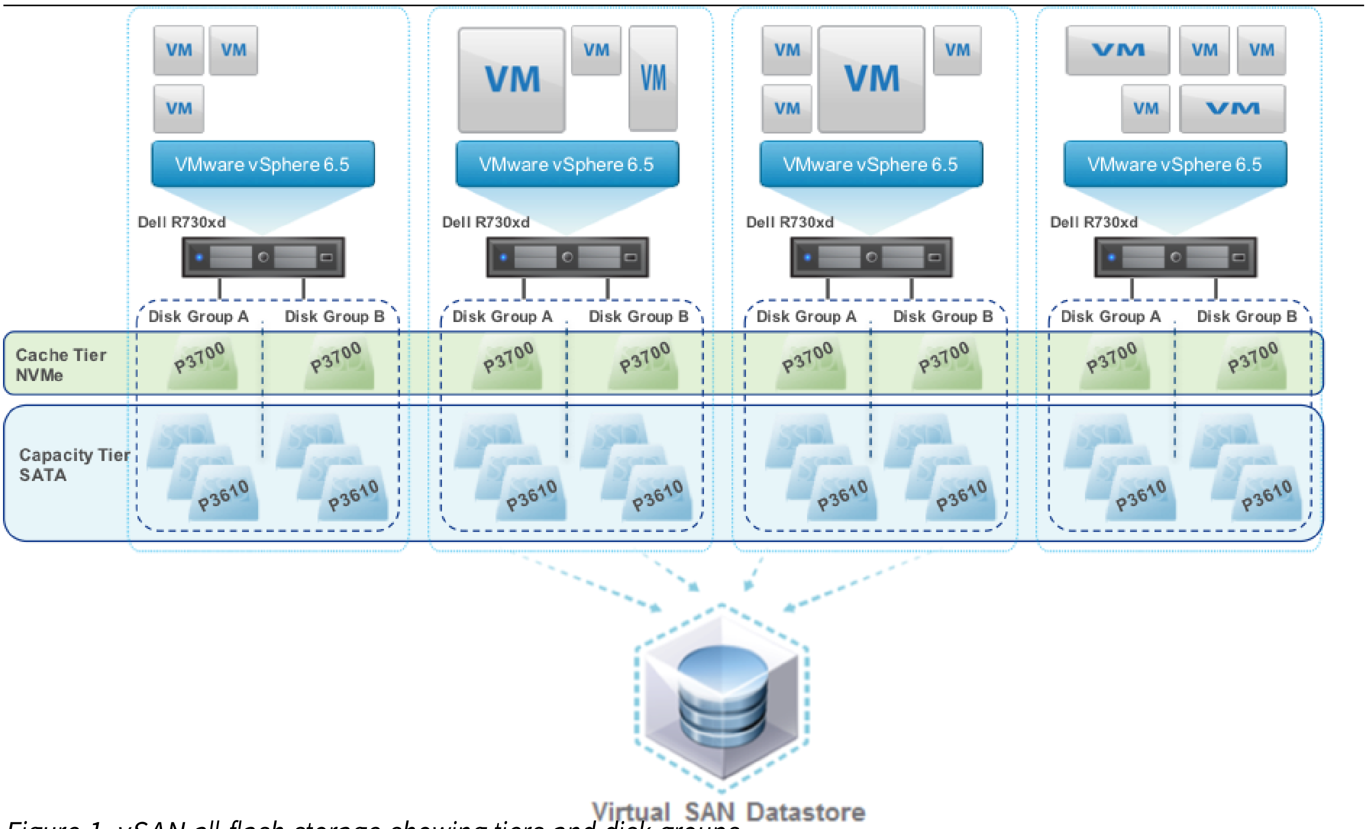

Figure 1. vSAN all-flash storage showing tiers and disk groups

When considering how to populate tiers and the *type* of flash for cache and capacity tiers you must consider the IO path; for the lowest latency and maximum performance writes go to the cache tier then software coalesces and de-stages the writes to the capacity tier. Cache use depends on deployment model, for example in vSAN hybrid configurations 30% of the cache tier is write cache, in the case of all-flash 100% of cache tier is write cache -- reads are from low latency flash capacity tier.

There will be a performance boost using all-flash. With larger capacity and durable flash drives available today the time has come where you should be considering whether you need spinning disks. The business case for flash over spinning disk has been made over recent years and includes much lower cost/IOPS, performance (lower latency), higher reliability (no moving parts to fail, less disks to fail for required IOPS), lower power and heat profile, smaller footprint, and so on. You will also benefit from additional HCI features, for example vSAN will only allow deduplication and compression on all-flash configurations.

- ***Recommendation:*** For best performance and lower TCO consider all-flash.

For best performance the cache tier should have the lowest latency, especially for vSAN as there is only a single cache device per disk group.

- ***Recommendation:*** If possible choose NVMe SSDs for the cache tier although SAS is still OK.
- ***Recommendation:*** Choose high endurance flash devices in the cache tier to handle high I/O.

For SSDs at the capacity tier there is negligible performance difference between SAS and SATA SSDs. You do not need to incur the cost of NVMe SSD at the capacity tier for database applications. However in all cases ensure you are using enterprise class SATA SSDs with features such as power failure protection.

- ***Recommendation:*** Choose high capacity SATA SSDs for capacity tier.
- ***Recommendation:*** Choose enterprise SSDs with power failure protection.

Depending on your timetable new technologies such as such as 3D Xpoint with higher IOPS, lower latency, higher capacity and higher durability may be available. There is a breakdown of flash storage at the end of this post.

- ***Recommendation:*** Watch for new technologies to include such as 3D Xpoint for cache AND capacity tier.

As I mentioned above you can have up to five disk groups per host and a disk group is made up of one flash device and up to seven devices at the capacity tier. You could have a single disk group with one flash device and as much capacity as you need, or multiple disk groups per host. There are advantages to having multiple disk groups per host:

- Performance: Having multiple flash devices at the tiers will increase the IOPS available per host.
- Failure domain: Failure of a cache disk impacts the entire disk group, although availability is maintained as vSAN rebuilds automatically.

You will have to balance availability, performance and capacity, but in general having multiple disk groups per host is a good balance.

- ***Recommendation:*** Review storage requirements, consider multiple disk groups per host.

What performance should I expect?

A key requirement for good application user experience is low storage latency; the usual recommendation is that database read IO latency should be below 10ms. Refer to the table from Part 6 of this series here for details.

For Caché database workloads tested using the default vSAN storage policy and Caché RANREAD utility I have observed sustained 100% random read IO over 30K IOPS with less than 1ms latency for all-flash vSAN using Intel S3610 SATA SSDs at the capacity tier. Considering that a basic rule of thumb for Caché databases is to size instances to use memory for as much database IO as possible all-flash latency and IOPS capability should provide ample headroom for most applications. Remember memory access times are still orders of magnitude lower than even NVMe flash storage.

As always remember your mileage will vary; storage policies, number of disk groups and number and type of disks etc will influence performance so you must validate on your own systems!

Capacity and performance planning

You can calculate the raw TB capacity of a vSAN storage pool roughly as the total size of disks in the capacity tier. In our example configuration in *figure 1* there are a total of 24 x INTEL S3610 1.6TB SSDs:

Raw capacity of cluster: 24 x 1.6TB = 38.4 TB

However *available* capacity is much different and where calculations get messy and is dependent on configuration choices; which policies are used (such as FTT which dictates how many copies of data) and also whether deduplication and compression have been enabled.

I will step through selected policies and discuss their implications for capacity and performance and recommendations for a *database workload.*

All ESXi deployments I see are made up of multiple VMs; for example, TrakCare a unified healthcare information system built on InterSystems ’ health informatics platform, HealthShare is at its heart at least one large (monster) database server VM which is absolutely fits the description "tier-1 business critical application". However a deployment also includes combinations of other single purpose VMs such as production web servers, print servers, etc. As well as test, training and other non-production VMs. Usually all deployed in a single ESXi cluster. While I focus on database VM requirements remember that SPBM can be tailored per VMDK for all your VMs.

Deduplication and Compression

For vSAN deduplication and compression is a cluster-wide on/off setting. Deduplication and compression can only be enabled when you are using an all-flash configuration. Both features are enabled together.

At first glance deduplication and compression seems to be a good idea - you want to save space, especially if you are using (more expensive) flash devices at the capacity tier. While there are space savings with deduplication and compression my recommendation is that you do not enable this feature for clusters with large production databases or where data is constantly being overwritten.

Deduplication and compression does add some processing overhead on the host, maybe in the range of single digit %CPU utilization, but this is not the primary reason not recommending for databases.

In summary vSAN attempts to deduplicate data as it is written to the capacity tier within the scope of a single disk group using 4K blocks. So in our example at *figure 1* data objects to be deduplicated would have to exists in the capacity tier of the same disk group. I am not convinced we will see much savings on Caché database files which are basically very large files filled with 8K database blocks with unique pointers, contents, etc. Secondly vSAN will only attempt to compress duplicated blocks, and will only consider blocks compressed if compression reaches 50% or more. If the deduplicated block does not compress to 2K it is written uncompressed. While there may be some duplication of operating system or other files *the real benefit of deduplication and compression would be for clusters deployed for VDI.*

Another caveat is the impact of a (albeit rare) failure of one device in a disk group on the whole group when deduplication and compression is on. The whole disk group is marked "unhealthy" which has a cluster wide impact: because the group is marked unhealthy all the data on a disk group will be evacuated off that group to other places, then the device must be replaced and vSAN will resynchronise the objects to rebalance.

- ***Recommendation:*** For database deployments do not enable compression and deduplication.

Sidebar: InterSystems database mirroring.

For mission critical tier-1 Caché database application instances requiring the highest availability I recommend InterSystems synchronous database mirroring, even when virtualised. Virtualised solutions have HA built in; for example VMWare HA, however additional advantages of also using mirroring include:

- Separate copies of up-to-date data.
- Failover in seconds (faster than restarting a VM then operating System then recovering Caché).
- Failover in case of application/Caché failure (not detected by VMware).

I am guessing you have spotted the flaw in enabling deduplication when you have mirrored databases on the same cluster? You will be attempting to deduplicate your mirror data. Generally not sensible and also a processing overhead.

Another consideration when deciding whether to mirror databases on HCI is the total storage capacity required. vSAN will be making multiple copies of data for availability, this data storage will be doubled again by mirroring. You will need to weigh the small incremental increase in uptime over what VMware HA provides against the additional cost of storage.

For maximum uptime you can create two clusters so that each node of the database mirror is in a completely independent failure domain. However take note of the total servers and storage capacity to provide this level of uptime.

Encryption

Another consideration is where you choose to encrypt data at rest. You have several choices in the IO stack including;

- Using Caché database encryption (encrypts database only).
- At Storage (e.g. hardware disk encryption at SSD).

Encryption will have a very small impact on performance, but can have a big impact on capacity if you choose to enable deduplication or compression in HCI. If you do choose deduplication and/or compression you would not want to be using Caché database encryption because it would negate any gains as encrypted data is random by design and does not compress well. Consider the protection point or risk they are trying to protect from, for example theft of file vs. theft of device.

- ***Recommendation:*** Encrypt at the lowest layer as possible in the IO stack for a minimal level of encryption. However the more risk you want to protect move higher up the stack.

Failures To Tolerate (FTT)

FTT sets a requirement on the storage object to tolerate at least *n* number of concurrent host, network, or disk failures in the cluster and still ensure the availability of the object. The default is *1* (RAID-1); the VM ’ s storage objects (e.g. VMDK) are mirrored across ESXi hosts.

> So vSAN configuration must contain at least n + 1 replicas (copies of the data) which also means there are 2n + 1 hosts in the cluster.

For example to comply with a number of failures to tolerate = 1 policy, you need three hosts at a minimum at all times -- even if one host fails. So to account for maintenance or other times when a host is taken off-line you need four hosts.

- ***Recommendation:*** A vSAN cluster must have a minimum four hosts for availability.

Note there is also exceptions; a Remote Office Branch Office (ROBO) configuration that is designed for two hosts and a remote witness VM.

Erasure Coding

The default storage method on vSAN is RAID-1 -- data replication or mirroring. Erasure coding is RAID-5 or RAID-6 with storage objects/components distributed across storage nodes in the cluster. The main benefit of erasure coding is better space efficiency for the same level of data protection.

Using the calculation for FTT in the previous section as an example; for a VM to tolerate *two* failures using a RAID-1 there must be three copies of storage objects meaning a VMDK will consume 300% of the base VMDK size. RAID-6 also allows a VM to tolerate two failures and only consumes 150% the size of the VMDK.

The choice here is between performance and capacity. While the space saving is welcome you should consider your database IO patterns before enabling erasure coding. Space efficiency benefits come at the price of the amplification of I/O operations which is higher again during times of component failure so for best database performance use RAID-1.

- ***Recommendation:*** For production databases do not enable erasure coding. Enable for non-production.

Erasure coding also impacts the number of hosts required in your cluster. For for example for RAID-5 you need a minimum of four nodes in the cluster, for RAID-6, you need a minimum of six nodes.

- ***Recommendation:*** Consider the cost of additional hosts before planning to configure erasure coding.

Striping

Striping offers opportunity for performance improvements but will likely only help with hybrid configurations.

- ***Recommendation:*** For production databases do not enable striping.

Object Space Reservation (thin or thick provisioning)

The name for this setting comes from vSAN using objects to store components of your VMs (VMDKs etc). By default all VMs provisioned to a VSAN datastore have object space reservation of 0% (thin provisioned) which leads to space savings and also enables vSAN more freedom for placement of data. However for your production databases best practice is to use 100% reservation(thick provisioned) where space is allocated at creation. For vSAN this will be Lazy Zeroed – where 0 ’ s are written as each block is first written to. There are a few reasons for choosing 100% reservation for production databases; there will be less delay when database expansions occur, and you are guaranteeing that storage will be available when you need it.

- ***Recommendation:*** For production database disks use 100% reservation.
- ***Recommendation:*** For non-production instances leave storage thin provisioned.

When should I turn on features?

You can generally enable availability and space saving features after using the systems for some time, that is; when there are active VMs and users on the system. However there will be performance and capacity impact. Additional replicas of data in addition to the original are needed so additional space is required while data is synchronised. My experience is that enabling these type of features on clusters with large databases can take a very long time and expose the possibility of reduced availability.

- ***Recommendation:*** Spend time up front to understand and configure storage features and functionality such as deduplication and compression before go-live and definitely before large databases are loaded.

There are other considerations such as leaving free space for disk balancing, failure etc. The point is you will have to take into account the recommendations in this post with vendor specific choices to understand your raw disk requirements.

- ***Recommendation:*** There are many features and permutations. Work out your total GB capacity requirements as a starting point, review recommendations in this post [and with your application vendor] then talk to your HCI vendor.

Storage processing overhead

You must consider the overhead of storage processing on the hosts. Storage processing otherwise handled by the processors on an enterprise storage array is now being computed on each host in the cluster.

The amount of overhead *per host* will be dependent on workload and what storage features are enabled. My observations with basic testing I have done with Caché on vSAN shows that processing requirements are not excessive, especially when you consider the number of cores available on current servers. VMware recommends planning for 5-10% host CPU usage

The above can be a starting point for sizing but *remember your mileage will vary* and you will need to confirm.

- ***Recommendation:*** Plan for worst case of 10% CPU utilisation and then monitor your real workload.

Network

Review vendor requirements -- assume minimum 10GbE NICs -- multiple NICs for storage traffic, management (e.g. vMotion), etc. I can tell you from painful experience that an enterprise class network switch is required for optimal operation of the cluster -- after all - all writes are sent synchronously over the network for availability.

- ***Recommendation:*** Minimum 10GbE switched network bandwidth for storage traffic. Multiple NICs per host as per best practice.

Flash Storage Overview

Flash storage is a requirement of HCI so it is good to review where flash storage is today and where its going in the near future.

The short story is whether you use HCI or not if you are not deploying your applications using storage with flash today it is likely that your next storage purchase will include flash.

Storage today and tomorrow

Let us review the capabilities of commonly deployed storage solutions and be sure we are clear with the terminology.

Spinning disk

- Old faithful. 7.2, 10K or 15K HDD spinning disks with SAS or SATA interface. Low IOPS per disk. Can be high capacity but that means the IOPS per GB are decreasing. For performance typically data is striped across multiple disks to achieve 'just enough' IOPS with high capacity.

SSD disk - SATA and SAS

- Today flash is usually deployed as SAS or SATA interface SSDs using NAND flash. There is also some DRAM in the SSD as a write buffer. Enterprise SSDs include power loss protection - in event of power failure contents of DRAM are flushed to NAND.

SSD disk - NVMe

- Similar to SSD disk but uses NVMe protocol (not SAS or SATA) with NAND flash. NVMe media attach via PCI Express (PCIe) bus allowing the system to talk directly without the overhead of host bus adapters and storage fabrics resulting in much lower latency.

Storage Array

- Enterprise Arrays provide protection and the ability to scale. It is more common today that storage is either a hybrid array or all-flash. Hybrid arrays have a cache tier of NAND flash plus one or more capacity tiers using 7.2, 10K or 15K spinning disks. NVMe arrays are also becoming available.

Block-Mode NVDIMM

- These devices are shipping today and are used when extremely low latencies are required. NVDIMMs sit in a DDR memory socket and provide latencies around 30ns. Today they ship in 8GB modules so are not likely to be used for legacy database applications, but new scale-out applications may take advantage of this performance.

3D XPoint

This is a future technology - not available in November 2016.

- Developed by Micron and Intel. Also known as **Optane** (Intel) and **QuantX** (Micron).
- Will not be available until at least 2017 but compared to NAND promises higher capacity, >10x more IOPS, >10x lower latency with extremely high Endurance and consistent performance.
- First availability will use NVMe protocol.

SSD device Endurance

SSD device *endurance* is an important consideration when choosing drives for cache and capacity tiers. The short story is that flash storage has a finite life. Flash cells in an SSD can only be deleted and rewritten a certain number of times (no restrictions apply to reads). Firmware in the device manages spreading writes around the drive to maximise the life of the SSD. Enterprise SSDs also typically have more real flash capacity than visible to achieve longer life (over-provisioned), for example an 800GB drive may have more than 1TB of flash.

The metric to look for and discuss with your storage vendor is full Drive Writes Per Day (DWPD) guaranteed for a certain number of years. For example; An 800GB SSD at 1 DWPD for 5 years can have 800GB per day written for 5 years. So the higher the DWPD (and years) the higher the endurance. Another metric simply switches the calculation to show SSD devices specified in Terabytes Written (TBW); The same example has TBW of 1,460 TB (800GB * 365 days * 5 years). Either way you get an idea of the life of the SSD based on your expected IO.

Summary

This post covers the most important features to consider when deploying HCI and specifically VMWare vSAN version 6.5. There are vSAN features I have not not covered, if I have not mentioned a feature assume you should use the defaults. However if you have any questions or observations I am happy to discuss via the comments section.

I expect to return to HCI in future posts, this certainly is an architecture that is on the upswing so I expect to see more InterSystems customers deploying on HCI.

#Best Practices #InterSystems Business Solutions and Architectures #Performance #System Administration #Caché #InterSystems IRIS

Source URL:https://community.intersystems.com/post/intersystems-data-platforms-and-performance-%E2%80%93-part-8-hyper-converged-infrastructure-capacity