Datrium DVX Storage System Architectures Overview

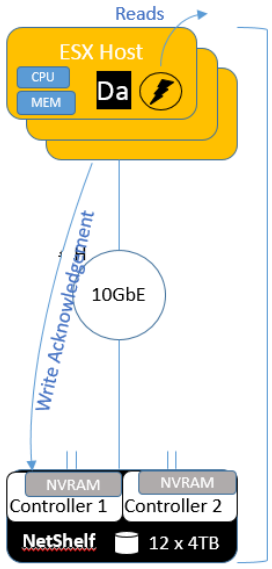
April 10, 2017
Jeremy Li

Datrium was named a Cool Vendor by Gartner in its annual Cool Vendors in Storage Technologies, 2016 dated April 22, 2016 (See ID:G00300354 for details.) It provides innovative easier storage management with great performance via a new architecture by using commodity, server-side solid-state-drive (SSD) technology. Since it came out from the stealth one year ago, it has gained just over 100 customers today.

Datrium DVX Storage System comprises of 1) DVX Hyperdriver software; and 2) a DVX NetShelf appliance with a single 10GbE connection. Its unique technology is to use a server powered storage, by installing a small agent – DVX Hyperdriver software (Da) on each host to utilize the local cache to deliver a high performance IOPS with deduped and compression. It is worth noting that you can manage VM application performance within vCenter using DVX plug-in, instead of managing storage. The main benefits are as follows:

- Simple
- Scalable
- Flexible
- Fast

Datrium's unique technology scaling up to 32 VMware hosts (ESXi) with effective storage capacity from 60TB to 180TB uses existing SSD drives (flash) in the host to cache read-only data via global post-dedupe technology, while using compressed copy stored in the host as a read cache, thus, eliminating the network latency occurring from the other traditional HCI architecture. Written access data to the DVX NetShelf is landed to the NVRAM via synchronous replication through one 10GbE connection (A dual-10Gbe connection will be supported in the future) after data is compressed at the host level, as illustrated in the screenshot left. Therefore, it eliminates a traditional host-to-host replication latency. The system is fast with near scaling because all READ IO will be fetched from the local flash cache for all VM data with scale for performance after a new host is added due to more server-side flash and CPU, thus more IOPS (Note: Datrium DVX technology cannot utilize any RAM



DVX NetShelf Appliance

available in a host). Datrium system have two modes: (1) Fast Mode, which uses up to 20% of host CPU; and (2) Insane Mode, which uses up to 40% of host CPU, if that CPU is available.

See ESG LAB REVIEW on Datrium DVX for video.

See how Infinio can utilize and allocate the server-side CPU, memory, and flash to speed up applications' performance by avoiding the network latency to gain a great performance – "**Infinio's Lightning fast I/O for any storage VSAN VVOLS SAN NAS DAS HCI** http://www.infinio.com/."

Hyper-convergence, sometimes refer to as Server SAN, or hyper-converged infrastructure provides cloud-like economics and scale by integrating compute resources, storage resources, software-defined virtualization, and software-defined storage onto standard x86 platforms. Hyper-converged solutions are made available through software-defined storage, as reference architectures, or dedicated appliances.

Watch a video of Datrium founders explaining Datrium Data Cloud software and RackScale systems to explain the importance of its design, feature and capabilities.

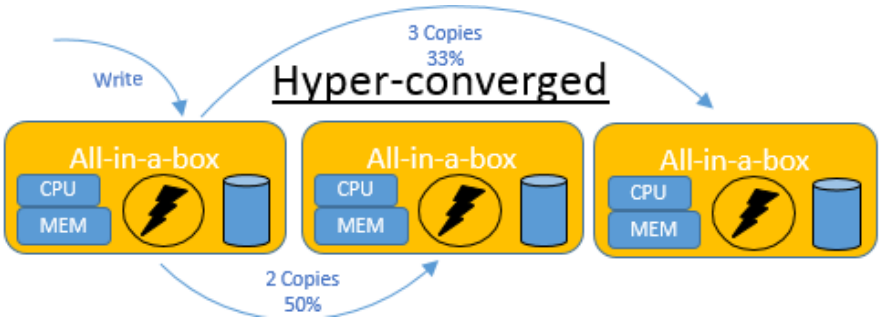
With one HPE G9 or its newer server or Cisco Blade M3 server, Datrium claims its server powered storage - NetShief with DVX Hyperdriver software Virtual Install Bundle (VIB) can achieve about 30,000 IOPS per host (ESXi) with 2 sockets with 12 cores each in September 2016, by using 10,000 IOPS per CPU core allocated to DVX performance. See Appendix below for details.

It is true that Datrium is benefit from the Moore's law. On Sept. 27, 2016, Datrium claims 30,000+ IOPS per host with 12 cores per host. Today, it claims 60,000+ IOPS per host with 24 cores per host.

Coincidentally, the Moore's law has been broken by Field Programmable Gate Array (FPGA). See Microsoft Puts FPGAs to Work for Azure and Bing that has gained the leap forward by its nearest search competitor - Microsoft is augmenting the servers in its Azure cloud service with FPGAs to accelerate chores such as Bing searches.

The key differentiation of Datrium DVX architecture with other converged architecture is that Datrium does not have (1) a traditional dual-controller bottleneck architecture; (2) the scale-out or scale-up issue; (3) in comparison to other HCI architectures, to use a copy method from a host to a host with a 2X or 3X disk storage penalty, as illustrated in

the screenshot below; and (4) to go through multiple host-to-host network latency as well as eliminating a "noise neighbors" effect that often occurs when multiple VMs share NAS resources (a contention issue); and (5) to manage logical unit numbers (LUNs) or zones (a Fiber Channle use case). It can present a volume via a local NFS mount to a ESXi host, meaning a DVX Netshelf represents a separate NFS Datastore at a VM level, in a similar fashion as Nimble storage does, even though a traditional RAID 6 method is used in its NetShelf where a persistent (durable), higher availability (HA) and coordination data are stored. It does not care how many hosts are available or not since all data in the NetShelf will not be lost, if hosts are not available.



However, the NetShelf can only sustain two disk simultaneous failures, while Pivot 3 can sustain a 5-simultanious failure from which only one host will be allowed for a failure without losing data.

Below are a question and answer session:

With a 12 4TB HDDs RAID 6 computed-on-hosts NetShelf (Off-Host Durable 2U rackmount appliance), a usable disk space is 10 * 4TB = 40TB is available prior to data reduction applied. After data reduction applied, 60-180TB effective disk storage will be available.

Q1: Does the above 12 x 4TB NetShelf have a spare HDD?

A1: [Renken, Brian] There is not a dedicated spare or spares within the DVX Netshelf. All 12 drives are used actively to store the full write stripe (data + parity that is sequentialized, compressed, deduped and encrypted if encryption is enabled) which is created in memory on the ESX host using the DVX hyperdriver. The full write stripe delivers small chunks to each drive for an effective RAID 6 protection level. 2 drives can fail simultaneously with no data loss and only the actual data contained on each failed drive would need to be reconstructed when the drives are replaced. The RAW capacity of the DVX Netshelf is 48 TB out of which 30 TB will be useable after systems overhead

but before data efficiency (compression + dedupe). Effective capacity is 90 - 180 TB assuming a 3 - 6x data efficiency rate.

Q2: Do you think Datrium will have a big problem, meaning seeking a complex exercise referenced in Q1 is not even possible now when the data residing on the first NetShelf will be reaching a full capacity between 60TB to 180TB, depending on the data reduction ratio? Let us assume I do not need to add new additional VMs now, but keep adding more data to the existing first NetShelf!

Below is an excerpt from the Gartner Magic Quadrant for Solid-State Arrays report dated Aug. 22, 2016 (Emphasis added):

"There is no simple way to add capacity to Tintri's arrays other than buying a new array, which makes the cost of adding capacity and workload consolidation **a complex exercise.**"

BTW, Datrium use case will be very good if the data residing on the NetShelf (or back-end persistent storage by removing the controller bottleneck associated with traditional dual controller SAN architectures) will not or never reach to its full capacity (between 60TB to 180TB), because adding more hosts will bring more cached data in the cluster in the Host environment! For example, a virtual desktop infrastructure (VDI) or HVD, as well as test, development and staging environments will get benefit from Datrium's scenario.

On the other hand, if Datrium use case is for Cisco blade servers, it may defeat one or two purposes of HCI's simplicity listed below as well as one more important factor:

"SDS and hyperconverged infrastructure carry with them the promise for massively simplified data centers that eliminate existing points of failure and complexity, such as those routinely introduced in traditional SAN environments."

Without simplicity, often, many organizations will find out later that the recovery time will be much longer and an active/active fail-over capability planned for many years will not work when a disaster really occurs because it always requires highly trained and skilled IT subject matter experts (SMBs) to figure out the issue first, often, it involves many groups to coordinate together with many conference calls prior to making a sound decision how to proceed the action. Thus, it often makes the resolution **a complex exercise.**"

A2: [Renken, Brian] Currently each DVX Netshelf represents a separate NFS Datastore and can scale up to 32 host per Netshelf. In the near future (Q4FY2017) Datrium will introduce a scale-out clustering capability of up to four Netshelves in a single NFS datastore/namespace. As part of this scale-out clustering there will be a layer of data protection across the up to 4 Netshelves insuring data reconstruction even if an entire Netshelf were to fail. Details will be announced and provided in the future. This represents a capacity scale of 360 - 720 TB of effective capacity in a single NFS datastore/namespace.

Q3: Can you tell me more details how Datrium Erasure Coding really works via the combination the DVX driver and Host-level data services via a postprocess dedupe, host-level-compression and cloning? What's the efficient level via erasure coding?

Note: A compressed copy is stored in the host cache for quick access, and the cache is deduplicated in-line at the host (Source: Gartner)

A3: [Renken, Brian] Please see my response to question #1. All RAID calculations are performed in memory (after write acknowledgement to NVRAM) on the ESX host by the hyperdriver. The full write stripe with data efficiency and encryption (if enabled) is broken in to chunks and distributed across the 12 x 4 TB containers in the Netshelf.

Q4: How many snapshots can Datrium support?

A4: [Renken, Brian] In the current release of the Datrium OS there is a 200,000 limit on snapshots. An administrator will be alerted as you reach this limit and we will stop taking additional snapshot until older snapshots have been deleted or expired by the Protection Group (PG) policies.

Q5: How can the DVX NetShelf provide both 2.4GB/s Throughput and $0.50/GB?

SAS-3: 12.0 Gbit/s, available since March 2013
SAS-4: 22.5 Gbit/s,[4] under development and expected in 2017[3]

Source: https://en.wikipedia.org/wiki/Serial_Attached_SCSI

A5: No response from Datrium

Q6: Nimble storage InfoSight Predictive Analytics capability can often phone its customers that a potential issue is coming prior to a real outage occurs, unlike many other legacy vendors' expensive products, which often miss the above feature and rely on the outage occurring, thus, leads to lose many work productive among hundreds or thousands of people in enterprises. Do you have a similar feature?

A6:

Summary:

Datrium DVX storage system delivers simple, scalable, flexible, fast, and linear scaling storage with greater efficiency and lower TCO. The more hosts added, the better the performance by eliminating many network latency (or avoiding the host-to-host copy latency or noise neighbor), in addition to avoiding 2X or 3X storage penalty from a traditional RAIN technology. It also stores apps and files you use most on the SSDs within the host for read-cache to speed up the application performance, meaning closer to the application without going through network latency in comparison to traditional converged infrastructure. See Intel Optane Memory – "It has also said that Optane could be up to **10 times faster** than traditional SATA-based SSDs." Therefore, the time is in favor of Datrium DVX NetShelf architecture, when Intel Optane memory is popular for the server side in the near future.

It can scale up to 32 VMware hosts (ESXi) with effective storage capacity from 60TB to 180TB by using existing SSD drives (flash) in the host to cache read-only data via global post-dedupe technology. Written access data is landed to the NVRAM within the NetShelf via synchronous replication through two 10GbE connections after data is compressed at the host level.

For a VDI deployment use case, Datrium might be a good fit if the gold images will not be used more that 60-180TB disk space from the NetShelf, which eliminates the host-to-host copies by avoiding 2X or 3X redundant data, whose method is used ubiquitously among most HCI vendors, as illustrated in the above screenshot, except from Pivot 3, which uses a Patented Erasure Coding technology from which it is better than Datruim's RAID-6 technology within the NetShelf appliance.

However, Datrium lacks Nimble storage InfoSight Predictive Analytics capability, which can often phone its customers' a potential issue prior to the real outage occurs.

Challenge:

1. Since Datrium just came out from its stealth one year ago, many big enterprises will choose a wait and see approach in order to avoid a costly and repeated famous scenario after Cisco acquired a Whiptail all-flash technology. All Cisco customers that acquired the Whiptail technology were impacted tremendously.

2. Cisco and other leading HCI vendors are also in this crowded HCI market place – See appendix B for details.

3. Datrium DVX NetShelf does not support a clustering technology as of this writing. Thus, a complex exercise referenced in the Gartner report above cannot even be considered when the NetShelf storage capacity will be reached full, while the computing resource is far more enough in a environment.

4. Datrium does not support a synchronous VM-to-VM replication, meaning a third-party disaster recovery technology must be used if any enterprises require to have one due to the requirement of both RPO and RTO.

5. Datrium NetShelf must be purchased with the full 12 x 4TB capacity up front, regardless whether you are in a small environment or not, but cannot extend its current full capacity due to lacking support of a cluster. Therefore, it limits to its use cases.

6. Pivot3, a Distributed Scale-Out Architecture vendor, will be a big challenge to Datrium since it not only uses its Patented Erasure Coding technology at the HCI storage level with 94% storage capacity efficiency at scale with capability of sustaining 5- simultaneous drive failures (or 2 drives and 1 node failures) with six 9s availability, while consuming less than 8% of system resources, but also can support Cisco blade servers by just loading its vSTAC OS that creates a lean operating environment with an extremely low overhead, in addition to some features such as dynamic QoS Datrium does not have.

7. Nutanix, a leader in the HCI platform for many years with a turnkey infrastructure platform that converges compute, storage, and virtualization through intelligent software to create flexible building blocks that replace legacy infrastructure consisting of separate servers, storage networks, storage arrays and virtualization

solutions. As a result, it can run any application at any scale and overcomes the HCI management headache by delivering a comprehensive management solution, refers to Nutanix Prism solution that may dramatically simplify and reduce management and operational complexity in datacenter (OPEX) with vMotion/Distributed Resource Scheduler (DRS) or Live Migration, meaning being able to deliver a Seamless VM Migration, in comparison with some incumbent vendors whose storage products are transitioned from legacy technology to a modern technology, which often uses a bolted on or workaround solution as well as through a new acquisition, which often results from different internal designs between two products. Thus, often a much longer troubleshooting time or longer downtime occurs because many different groups (e.g., a VM group, a storage group, a network group or even a VDI/HVD group or via a third-party tech support group) must be involved for a final solution. Further, a root cause of the issue will never be found in many cases.

Below is a quote from Gartner's report titled "**Large Established Storage Vendors, Brands and Products Are No Longer Risk-Free**" dated Nov. 16, 2016:

"*For example, mixing SSDs and HDDs within a hybrid array causes storage area network (SAN) or Ethernet flow control issues, as SSDs require smaller queue depths with quickly serviced flow control and HDDs require deeper queue depths with slower flow control servicing (see "The Future of Storage Protocols").*"

8. Datrium does not support KVM as well as Hyper-V.

9. Datrium does not offer any HyperGuarantee like the SimpliVity's HyperGuarantee - the Industry's Most Complete Guarantee!

10. The employees turnover rate is very high in the IT industry. Therefore, the public sector usually looks for a more reliable vendor for a long-term post support.

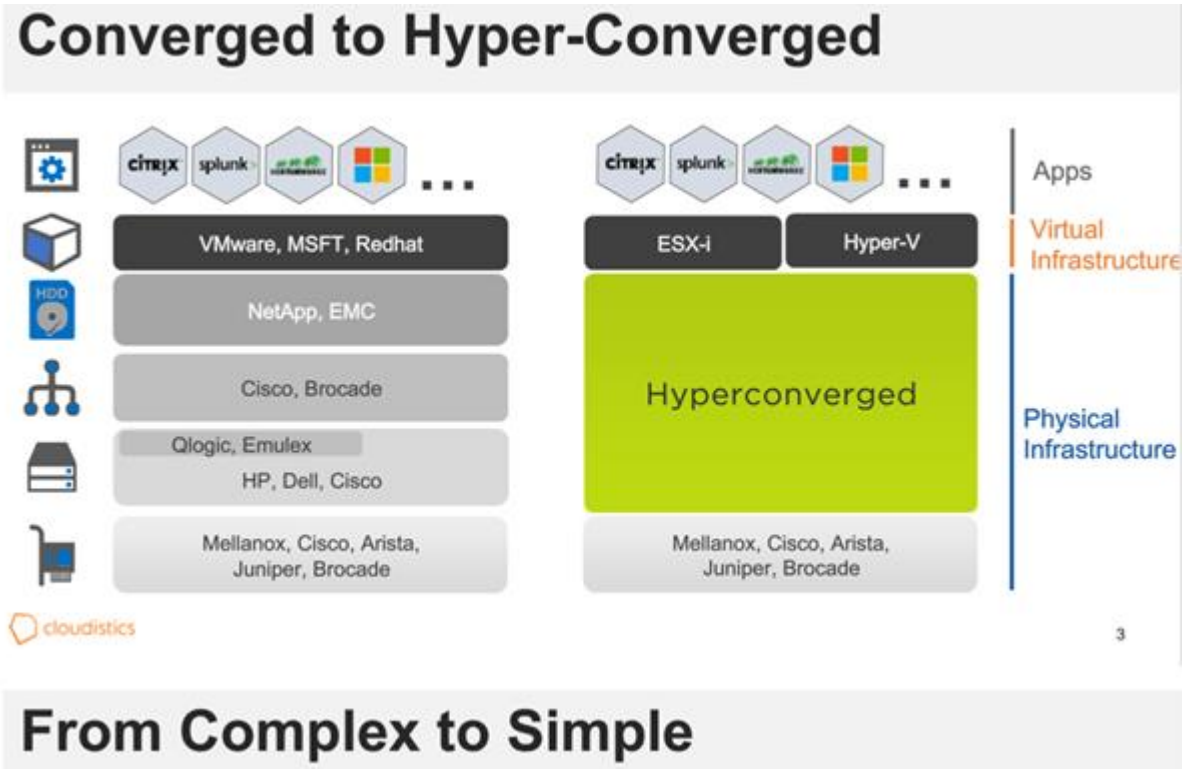
Recommended Reading:

- Gartner Magic Quadrant for Integrated Systems dated Oct. 10, 2016
- Gartner Critical Capabilities for Solid-State Arrays dated Aug. 31, 2016
- Magic Quadrant for Solid-State Arrays dated Aug. 22, 2016

Success – A project delivers expected business value such as measurable improvement to revenue, profits or net income, automation to improve productivity, new product release, reduce inventory costs or some other targeted outcome.

Failure – A project that did not meet or exceed expected business value.

Source: Paul Dandurand, CEO of PieMatrix and Lawrence Dillon, Practice Leader of ENKI LLC>



Source: Courtesy of Cloudistics

Appendix A

[Renken, Brian] From benchmarks we have run internally, we know that most G9 or newer servers can run roughly 10,000 IOPS per CPU core allocated to DVX performance. As we roll out to older pre-G9 servers, we typically see 5,000 - 8,000 IOPS per CPU core depending on clock speed. With the conservative assumption of 5,000 IOPS per G8 CPU core the math looks as follows.

2 x 12-Core CPU's = 24 total cores.

- Fast Mode (20%) = 24 X .20 = 4.8 cores - 1.5 cores (DVX system overhead) = 3.3 cores allocated for DVX performance X **5,000/core** = 16,500 IOPS

- Fast Mode (20%) = 24 x .20 = 4.8 cores - 1.5 cores (DVX system overhead) = 3.3 cores allocated for DVX performance x **8,000/core** = 26,400 IOPS
- Insane Mode (40%) = 24 x .40 = 9.6 cores - 1.5 cores (DVX system overhead) = 8.1 cores allocated for DVX performance x **5,000/core** = 40,500 IOPS
- Insane Mode (40%) = 24 x .40 = 9.6 cores - 1.5 cores (DVX system overhead) = 8.1 cores allocated for DVX performance x **8,000/core** = 64,800 IOPS

Appendix B

Cisco introduced its first Hyperconverged product based on "RAIN" technology, also known as "Log Structured File System" and named as "HyperFlex Systems" in March 2016. Cisco markets its HCI as "2nd Gen Hyperconverged" that includes:

- Unified Fabric Networking, as illustrated in a picture below or see a video for details.
- Pre-loaded HX Data Platform, a core software, which is designed for distributed storage to offer Data Services and Storage Optimization
- Dynamic Data Distribution - Elastic

Note: Cisco B200 M4 blades (or servers) are the #1 market share in the U.S. in 2016, while its UCS platform has 48,000 customers in March 2016.

- Independently scale-up and scale-out
- Security

- Call Home and Onsite 24x7 support
- Pointer-based snapshot
- Near Instant Clones
- Inline dedupe and compression
- Self-healing
- Single pane of glass for management

Special note: *Cisco acknowledges that HyperFlex is not designed for low latency apps such as databases and operational and mission critical applications. It is designed for operational simplicity – see https://www.youtube.com/watch?v=BVMpcitCQcw for reference!*

Acknowledgement

Thanks Brian Renken, Sr. Engineer at Datrium for providing a presentation and a Q&A session on April 10, 2017.