ComputerWeekly.com | TechTarget

E-guide

CW+ Content

Top 10 storage stories of 2018

In this e-guide:

The year 2018 in storage has witnessed a few notable trends.

Probably the most prominent has been cloud storage, which has palpably matured around hybrid and multicloud strategies, with key developments in file systems and service offerings that link on-premise and cloud capacity.

Alongside that, virtualisation continues to be a key trend, but in the form of containers – and specifically the requirement for persistent storage – and hyper-converged infrastructure (HCI) that marries storage and compute with hypervisor built in.

Flash storage also continues to mature, and has been pushed forward by the emergence of the super-fast, made-for-flash, NVMe protocol and the arrival of storage-class memory and a potential change to the IT architecture as we have known it.

Antony Adshead, storage editor

Contents

Hyper-convergence and containers key storage-related deployments in 2018

Antony Adshead, storage editor

Hyper-converged infrastructure and container management are key storage-related projects for UK organisations in 2018.

The 2018 ComputerWeekly/TechTarget IT priorities survey, which questioned 243 IT professionals in the UK, found the proportion of respondents that plan to deploy hyper-converged infrastructure (HCI) had nearly trebled year-on-year, from 9% to 22%.

Meanwhile, those that plan to work on container management this year – of which storage is a key component – saw a jump to just over double those that planned to in 2017, from 8% of respondents to 17%.

Hyper-converged infrastructure is a key trend in the datacentre. HCI appliances bring together compute and storage in one box, often with a hypervisor built in. These nodes can then be clustered in scale-out fashion to grow performance and capacity.

Initially starting as an option for small and medium-sized enterprises (SMEs), being well-suited to organisations with fewer IT resources, HCI tools have taken off across all sizes of organisation.

Containers, meanwhile, are a form of virtualisation that operates at the host operating system (OS) level, unlike virtual machines (VMs) that each have an OS of their own. Containers have been designed to be rapidly deployable, possibly short-lived building application blocks. Their benefits are they can quickly scale and are portable between operating environments and so allow an agile response to application demand.

Key to their use is the need to develop and manage persistent storage, with containers also used to provide a means to point to storage resources.

Other areas that saw a growth in the number of planned deployments in 2018 included flash arrays, with 9% planning to roll them out this year compared to 6% last year, and storage for virtual environments, which saw an increase in planned deployments from 11% in 2017 to 18% this year.

The deployment of storage for virtual environments also brings with it the need to back up VMs, and here there was growth too, with 29% planning a roll-out compared to 17% in 2017.

Finally, there was a noticeable increase in plans to deploy disk for backup, up from 8% of respondents to 13%. This could include deployments of hyper-

converged backup appliances, which have emerged in the past year as a means of building clusters of secondary storage for backups, with the advantages of HCI.

Next Article

Hyper-converged infrastructure 2018 technology and market survey

Chris Evans, guest contributor

Hyper-converged infrastructure (HCI) has been around for a number of years. HCI systems consolidate the traditionally separate functions of compute (server) and storage into a single scale-out hardware platform.

In this article, we review what hyper-converged infrastructure means today, the suppliers that sell HCI and where the technology is headed.

HCI systems are predicated on the concept of merging the separate physical components of server and storage into a single hardware appliance. Suppliers sell the whole thing as an appliance or users can choose to build their own using software and hardware components readily available in the market.

The benefits of implementing hyper-converged infrastructure are in the cost savings that derive from a simpler operational infrastructure.

The integration of storage features into the server platform, typically through scale-out file systems, allows the management of LUNs and volumes to be eliminated, or at least hidden from the administrator. As a result, HCI can be

operated by IT generalists, rather than needing the separate teams traditionally found in many IT organisations.

HCI implementations are typically scale-out, based on deployment of multiple servers or nodes in a cluster. Storage resources are distributed across the nodes to provide resilience against the failure of any component or node.

Distributing storage provides other advantages. Data can be closer to compute than with a storage area network, so it is possible to gain benefit from faster storage technology such as NVMe and NVDIMM.

The scale-out nature of HCI also provides financial advantages, as clusters can generally be built out in increments of a single node at a time. IT departments can buy nearer to the time the hardware is needed, rather than buying up-front and under-utilising equipment. As a new node is added to a cluster, resources are automatically rebalanced, so little additional work is needed other than rack, stack and connect to the network.

Shared core

Most HCI implementations have what is known as a “shared core” design. This means storage and compute (virtual machines) compete for the same processors and memory. In general, this could be seen as a benefit because it reduces wasted resources.

However, in the light of the recent Spectre/Meltdown vulnerabilities, I/O intensive applications (such as storage) will see a significant upswing in processor utilisation once patched. This could mean users having to buy more equipment simply to run the same workloads. Appliance suppliers claim that “closed arrays” don’t need patching and so won’t suffer the performance degradation.

But running servers and storage separately still has advantages for some customers. Storage resources can be shared with non-HCI platforms. And traditional processor-intensive functions such as data deduplication and compression can be offloaded to dedicated equipment, rather than being handled by the hypervisor.

Unfortunately, with the introduction of NVMe-based flash storage, the latency of the storage and storage networking software stack is starting to become more of an issue. But startups are beginning to develop solutions that could be classed as HCI 2.0 that disaggregate the capacity and performance aspects of storage, while continuing to exploit scale-out features. This allows these systems to gain full use of the throughput and latency capabilities of NVMe. NetApp has introduced an HCI platform based on SolidFire and an architecture that reverts to separating storage and compute, scaling each separately in a generic server platform. Other suppliers have started to introduce either software or appliances that deliver the benefits of NVMe performance in a scalable architecture that can be used as HCI.

HCI supplier roundup

Cisco Systems acquired Springpath in August 2017 and has used its technology in the HyperFlex series of hyper-converged platforms. HyperFlex is based on Cisco UCS and comes in three families: hybrid nodes, all-flash nodes and ROBO/edge nodes. Fifth generation platforms offer up to 3TB of DRAM and dual Intel Xeon processors per node. HX220c M5 systems deliver 9.6TB SAS HDD (hybrid), 30.4TB SSD (all-flash) while the HX240c M5 provides 27.6TB HDD and 1.6TB SSD cache (hybrid) or 87.4TB SSD (all-flash). ROBO/edge models use local network port speeds, whereas the hybrid and all-flash models are configured for 40Gb Ethernet. All systems support vSphere 6.0 and 6.5.

Dell EMC and **VMware** offer a range of technology based on VMware Virtual SAN. These are offered in five product families: G Series (general purpose), E Series (entry level/ROBO), V Series (VDI optimised), P Series (performance optimised) and S Series (Storage dense systems). Appliances are based on Dell's 14th generation PowerEdge servers, with E Series based on 1U hardware, while V, P and S systems use 2U servers. Systems scale from single-node, four-core processors with 96GB of DRAM to 56 cores (dual CPU) and 1536GB DRAM. Storage capacities scale from 400GB to 1,600GB SSD cache and either 1.2TB to 48TB HDD or 1.92TB to 76.8TB SSD. All models start at a minimum of three nodes and scale to a maximum of 64 nodes based on the requirements and limitations of Virtual SAN and vSphere.

NetApp has designed an HCI platform that allows storage and compute to be scaled separately, although each node type sits within the same chassis. A minimum configuration consists of two 2U chassis, with two compute and four storage nodes. This leaves two expansion slots. The four-node storage configuration is based on SolidFire scale-out all-flash storage and is available in three configurations. The H300S (small) deploys 6x 480GB SSDs for an effective capacity of 5.5TB to 11TB. The H500S (medium) has 6x 960GB drives (11TB to 22TB effective) and the H700S (large) uses 6x 1.92TB SSDs (22TB to 44TB effective). There are three compute module types: H300E (small) with 2x Intel E5-2620v4 and 384GB DRAM, H500E (2x Intel E5-2650v4, 512GB DRAM) and H700E (large) with 2x Intel E5-2695v4, 768GB DRAM. Currently the platform only supports VMware vSphere, but other hypervisors could be offered in the future.

Nutanix is seen as the leader in HCI, bringing its first products to market in 2011. The company floated on the Nasdaq in September 2016 and continues to evolve its offerings into a platform for private cloud. The Nutanix hardware products span four families (NX-1000, NX-3000, NX-6000, NX-8000) that start at the entry-level NX-1155-G5 with Dual Intel Broadwell E5-2620-v4 processors, 64GB DRAM and a hybrid (1.92TB SSD, up to 60TB HDD) or all-flash (23TB SSD) storage configuration. At the high end, the NX-8150-G5 has a highest specification Dual Intel Broadwell E5-2699-v4, 1.5TB DRAM and hybrid (7.68GB SSD, 40TB HDD) or all-flash (46TB SSD) configurations. In fact, customers can select from such a large range of configuration options that

almost any node specification is possible. Nutanix has developed a proprietary hypervisor called AHV, based on Linux KVM. This allows customers to implement systems and choose either AHV or VMware vSphere as the hypervisor.

Pivot3 was an earlier market entrant than even Nutanix, but had a different focus at that time (video surveillance). Today, Pivot3 offers a hardware platform (Acuity) and software solution (vSTAC). Acuity X-Series is offered in four node configurations, from the entry level X5-2000 (Dual Intel E5-2695-v4 up to 768GB of DRAM, 48TB HDD) to the X5-6500 (Dual Intel E5-2695-v4 up to 768GB of DRAM, 1.6TB NVMe SSD, 30.7TB SSD). Models X5-2500 and X5-6500 are “flash accelerated” as both a tier of storage and as a cache. Acuity supports the VMware vSphere hypervisor.

Scale Computing has had steady growth in the industry, initially focusing on SMB and gradually moving the value proposition of its HC3 platform higher by introducing all-flash and larger-capacity nodes. The HC3 series now has four product families (HC1000, HC2000, HC4000 and HC5000). These scale from the base model HC1100 (Single Intel E5-2603v4, 64GB DRAM, 4TB HDD) to the HC5150D (Dual Intel E5-2620v4, 128GB DRAM, 36TB HDD, 2.88TB SSD). There is also an all-flash model (HC1150DF) with Dual Intel E5-2620v4, 128GB DRAM, 36TB HDD and 38.4TB SSD. HC3 systems run the HyperCore hypervisor (based on KVM) for virtualisation and a proprietary file system called

Scribe. This allowed Scale to offer more competitive entry-level models for SMB customers.

Simplivity was acquired by HPE in January 2017. The platform has since been added to HPE's integrated systems portfolio. The Omnistack software that drives the Simplivity platform is essentially a distributed file system that integrates with the vSphere hypervisor. An accelerator card with dedicated FPGA is used to provide hardware-speed deduplication of new data into the platform. The HPE Simplivity 380 has three configuration options: Small Enterprise all-flash (Dual Intel Xeon Broadwell E-2600 v4 series, up to 1467GB DRAM and 12TB SSD); Medium Enterprise all-flash (Dual Intel Xeon Broadwell E2600-v4 series, up to 1428GB DRAM and 17.1TB SSD); and Large Enterprise all-flash (Dual Intel Xeon Broadwell E5-2600v4 series, up to 1422GB DRAM and 23TB SSD). Systems are scale-out and nodes can be mixed in a single configuration or spread over geographic locations.

Next Article

Storage in containers: The answer to persistent storage needs?

Chris Evans, guest contributor

For decades, the accepted wisdom in storage management has been the need for solid, persistent and (typically) hardware-based data storage systems. Over the past 20-plus years, this has meant a shared storage platform as the primary location for data.

As storage form factors become more fragmented, some suppliers even offer persistent storage based on temporary entities, such as containers. Does this make sense and how can persistence be achieved with such an ethereal construct?

Shared storage arose from a need to reduce costs, consolidate and eliminate the management overhead of storage deployed in hundreds or even thousands of servers in the datacentre.

A shared storage array was a good solution. Fibre Channel and Ethernet networks offered the ability to connect servers over distance, without cabling issues. And servicing one or two (rather than hundreds) of physical devices reduced maintenance and costs for the customer and supplier.

We now live in a different world. Today, applications are mostly virtualised and container technology has started to gain ground. Shared storage is seen as difficult to use, because it focuses on connecting physical servers to physical storage.

But modern applications work on logical volumes, file systems and object stores. Public cloud computing extends this paradigm, and obfuscates the physical view of hardware altogether.

Persistence of data is still important, however. So how can we achieve this while meeting the needs of new application deployment methods? It's first worth looking at what the new requirements are.

Containers, storage array, I/O blender

Virtualisation brought us the problem of the I/O blender – an increasingly random workload created by many virtual machines that access the same LUN or file share. To overcome the issues of shared storage in virtualised environments, VMware (for example) introduced its own file system with specific additional commands to reduce contention and fragmentation. We also saw the introduction of features such as VMware Virtual Volumes (VVOLs), which specifically aim to eliminate the physical LUN and treat virtual machines as objects.

Issues in storage access seen with server virtualisation are exacerbated further with containers. In the container world, a single physical host may run hundreds of containers, each vying for storage resources. Having each container access a long and complex storage stack introduces the risk of contention and goes against the benefits of the lightweight nature of a container.

But this is what many suppliers are doing. Volume plugins for Docker, for example, provide automation to map LUNs and volumes on physical arrays to physical hosts and then onto an individual container.

With the increased adoption of public and hybrid cloud architectures, the idea of a central fixed storage array becomes something of a problem. Applications have become more portable, with the ability to spin up containers in seconds and in many different datacentre locations. This paradigm is in distinct contrast to that of physical servers, which typically would be installed and not moved for years, before eventual decommissioning.

As we can see, delivering storage for container environments brings a new set of requirements, including:

- **Data mobility** – Containers move around, so the data has to be able to do that too. Ideally, that means not just between hosts in one datacentre, but across geographic locations.
- **Data security** – Containers need to be secured at a logical or application level, rather than at the LUN level, because containers expect to be recycled regularly.

- **Performance** – Containers introduce the idea of hundreds of unrelated applications working on the same physical host. I/O must be efficient and easy to prioritise.

Delivering storage with containers

One solution to the problem of persistent storage and containers is to use containers themselves as the storage platform.

At first glance, this seems like a bad idea. A container is designed to be temporary, so can be discarded at any time. Also, an individual container's identity is not fixed against anything that traditional storage uses. And there is no concept of host WWNs or iSCSI IQNs, so how can persistent storage with containers be achieved and why is it worth doing?

Let's address the "why" question first.

As we have discussed, containers can be short-lived and were designed for efficiency. Eliminating the I/O stack as much as possible contributes to the overall performance of a container environment. If storage is delivered through a container, the communication path between application and storage is very lightweight – simply between processes on the same server. As an application moves, a container on the new host can provide access to the storage, including spinning up a dedicated storage container if one did not already exist.

Clearly, there is a lot of back-end work to be done to keep data protected and available across multiple hosts, but this is less of a challenge than with traditional storage arrays because for many applications, only one container accesses a data volume at any one time.
Disaggregating access to storage in this way eliminates one of the issues we will see as NVMe becomes adopted more widely – the problem of having data pass through a shared controller. NVMe has much greater performance than traditional SAS/SATA, making a shared controller the new bottleneck for storage. Disaggregation helps mitigate this issue, in the same way as hyper-converged systems distribute capacity and performance for storage across a scale-out server architecture.

The question of “how” can be answered by looking at the location for persistence.

The media offers the answer here, with either spinning-disk HDDs or flash drives providing that capability. Configurations, access, and so on can be distributed across multiple nodes and media, with consensus algorithms used to ensure data is protected across multiple nodes and devices. That way, if any host or container delivering storage were to die, another can be spun up or the workload rebalanced across the remaining nodes, including the application itself. By design, the data would move with the application.

Container storage suppliers

This is the kind of architecture that is being implemented by startup companies such as Portworx, OpenEBS, Red Hat and StorageOS. Each uses a distributed node-based scale-out architecture, with storage and applications that run on the same platform. Essentially, it is a hyper-converged model for containers.

Some suppliers, such as Scality (with RING) and Cisco HyperFlex (formerly Springpath), use containers within the architecture for scalability, even though the products are not just for container environments.

For all suppliers, integration with container orchestration platforms is essential. Kubernetes is leading this charge, with Kubernetes Volumes the most likely way for storage to be mapped in these environments.

Maturity issues

There are some issues that still need to be considered as the technology matures.

First is the question of data services. Clearly, compression and deduplication have an effect on performance. The efficiency of these features will be key in gaining adoption, as we saw with the all-flash market. End-users will expect data protection, such as snapshots, clones and replication.

Then there is the subject of integration with public cloud. Today's solutions are mostly focused on single-site implementations, but true mobility means being able to move data around in a hybrid environment, which is much more of a challenge.

Finally, we should highlight issues of security.

The recent Meltdown vulnerability has a specific risk for containers, with the ability to access data from one container to another on unpatched systems. This raises questions about data security and the use of techniques such as in-memory encryption that may be required to protect against the inadvertent leaking of data.

There is a particular challenge for the startups to solve here, which may have a direct impact on the uptake of container-based storage systems. It may also make some businesses think that the idea of physical isolation (shared storage) goes some way to mitigating against unforeseen risks as they are discovered and reported.

Next Article

ESG survey: Storage spending habits become ‘hybrid-cloud-defined’

Antony Adshead, storage editor

Use of cloud storage is on the rise as a way of gaining capacity without affecting on-premise spend. But at the same time, organisations have pulled some workloads back from the cloud and are saving money by using software-defined storage.

These are some of the key findings of a survey by analysts Enterprise Strategy Group (ESG) which involved 412 respondents in mid-size (33%) and enterprise (67%) organisations across the UK and Europe.

The findings have led ESG to characterise such customer behaviour as “hybrid-cloud-defined”. This is where IT professionals try to find a balance between storage types – neither predominantly on-premise or cloud – whereas in the past, storage was all about steady progress in essentially similar storage systems.

Key among the responses to the survey that seem to have fed that conclusion include the use of cloud storage as a way of gaining storage capacity without additional on-premise spend (cited by 36%, equal with big data analytics).

Meanwhile, the vast majority of respondents (94%) reported their spend on on-premise data storage to be accelerating (49%) or remaining constant (45%).

The survey also asked those whose storage spending was flat or falling what they thought was responsible. Top answer here (39%) was that respondents believe their organisations are storing data more efficiently. Meanwhile, and possibly elaborating on that, 37% are using more cloud applications while 34% are using more cloud infrastructure services.

But although cloud figured heavily as a spending target, the survey also found that 57% had moved at least one workload from a cloud software or infrastructure service back to on-premise resources.

And 23% of those with flat/falling storage spend cited the use of software-defined storage with commodity server hardware, which is likely to form part of in-house budget spending.

When asked whether storage was a strategic consideration for their organisation that could lead to competitive advantage, the UK led the way, with 59% of those questioned agreeing, with 36% considering it tactical. France was next on 56%/35%, then Germany (51%/42%) and Italy (50%/43%).

When asked about their biggest storage challenge, the largest proportion said data protection (35%). Then came rapid data growth (28%), hardware costs (27%) and management of data placement (27%).

Disaggregated systems try to unlock the NVMe conundrum

Chris Evans, guest contributor

The shared storage array has seen amazing success as a key component of IT infrastructure over the past 30 years. Consolidation of storage from many servers into one appliance has provided the ability to deliver more efficient services, increase availability and reduce costs.

But as storage media moves towards the use of flash NVMe, shared arrays are showing their age and are being superseded by a new wave of disaggregated storage products.

To understand the root cause of the impending issue with shared storage, we need to look at the media in use.

Recently, hard drives have given way to flash (Nand) storage that is many orders of magnitude faster than spinning media, but that wasn't the case for many years.

The performance profile of a single hard drive was such that centralising input/output (I/O) through one or more controllers didn't impact on performance and, in most cases, improved it.

In other words, spinning disk drive hardware actually was the bottleneck in I/O. The controller provided much-needed functionality with no further effect on performance.

Performance-wise, an HDD-based array might, for example, deliver latency of 5ms to 10ms. Flash set the bar at less than one 1ms, with suppliers looking to achieve ever lower numbers. The first all-flash systems were based on SAS/Sata drives and connectivity.

The next transition in media is towards NVMe drives, where the I/O protocol is much more efficiently implemented. As a result, traditional array designs that funnel I/O through two or more centralised controllers simply can't exploit the aggregate performance of a shelf of NVMe media. In other words, the controller is now the bottleneck in I/O.

Removing the controller bottleneck

The answer so far to the NVMe performance conundrum has been to remove the bottleneck completely.

Rather than have all I/O pass through a central controller, why not have the client system access the drives directly?

With a fast, low-latency network and direct connectivity to each drive in a system, the overhead of going through shared controllers is eliminated and the full value of NVMe can be realised.

This is exactly what new products coming to the market aim to do. Client servers running application code talk directly to a shelf of NVMe drives, with the result that much lower latency and much higher performance numbers are achieved than with traditional shared systems.

NVMe implementation

Creating a disaggregated system requires separation of data and control planes. Centralised storage implements the control and data path in the controller. Disaggregated systems move control to separate elements of the infrastructure and/or to the client itself.

The splitting of functionality has the benefit of removing controller overhead from I/O. But there is also a negative effect on management, as the functions that were performed centrally still have to be done somewhere.

To understand what we mean by this, imagine the I/O that occurs to and from a single logical LUN on shared storage mapped to a client server. I/O to that LUN is done using logical block address (LBA).

The client writes from block 0 to the highest block number available, based on block size and capacity of the LUN. The controllers in the storage take the responsibility of mapping that logical address to a physical location on storage media.

Then as data passes through a shared controller, the data block will be deduplicated, compressed, protected (by Raid or erasure coding, for example) and assigned one or more physical locations on storage. If a drive fails, the controller rebuilds the lost data. If a new LUN is created, the controller reserves out space in metadata and physically on disk/flash as the LUN is used.

In disaggregated systems, these functions still need to be done and are, in general, passed out to the client to perform. The client servers need to have visibility of the metadata and data and have a way to coordinate between each other to ensure things go smoothly and no data corruption occurs.

Why disaggregate?

The introduction of NVMe offers great performance improvements.

In certain applications low latency is essential, but without disaggregation the only real way to implement a low-latency application is to deploy storage into the client server itself. NVMe flash drives can deliver super low latency, with NVMe Optane drives from Intel giving even better performance.

Unfortunately, putting storage back into servers isn't scalable or cost effective and was the original reason shared storage was first implemented. Disaggregation provides a middle ground that takes the benefit of media consolidation and (seemingly) local storage to get the highest performance possible from new media.

The type of applications that need low latency include financial trading, real-time analytics processing and large databases where transaction performance is a direct function of individual I/O latency times.

There's an analogy here to the early days of flash storage, where all-flash arrays were deployed in the enterprise onto applications that would be expensive to rewrite or simply couldn't be speeded up by any other method than delivering lower latency.

In the first implementations it's likely we will see disaggregated systems deployed on only those applications that will benefit most, as there are some disadvantages to the architecture.

Compromises

As highlighted already, depending on the implementation, client servers in disaggregated systems have a lot more work to do to maintain metadata and perform calculations such as Raid/erasure coding, compression and deduplication.

Support is limited to specific operating systems and may require the deployment of kernel-level drivers or other code that creates dependencies on the OS and/or application. Most systems use high-performance networks such as InfiniBand or 40Gb Ethernet with custom NICs.

This increases the cost of systems and will introduce support challenges if this technology is new to the enterprise organisation. As with any technology, the enterprise will have to decide whether the benefits of disaggregation outweigh the support and cost issues.

One other area not yet fully determined are the standards by which systems will operate. NVMe over a network or NVMe over Fabrics (NVMeF) is defined by the NVM Express organisation, and covers the use of physical transports such as Ethernet and Infiniband with access protocols such as RDMA over Converged Ethernet (RoCE) and Internet Wide-Area RDMA Protocol (iWarp), which provide remote direct memory access (RDMA) from client server to individual drives.

Some suppliers in our roundup have pushed ahead with their own implementations in advance of any standards being ratified.

NVMe supplier systems

DVX is a disaggregated storage system from startup Datrium. The company defines its offering as open convergence and has a model that uses shared storage and DRAM or flash cache in each client server. The company claims some impressive performance figures, achieving an IOmark-VM score of 8,000 using 10 Datrium data nodes and 60 client servers.

E8 Storage offers dual or single appliance models. The E8-D24 dual controller appliance offers Raid-6 protection across 24 drives, whereas the E8-S10

implements Raid-5 across 10 drives. Both systems use up to 100GbE with RoCE and can deliver up to 10 million IOPS with 40GBps throughput. E8 also offers software-only systems for customers that want to use their own hardware. Note that the dual controller implementation is to provide metadata redundancy.

Apeiron Data Systems offers a scale-out system based on 24-drive NVMe disk shelves. Client servers are connected using 40Gb Ethernet. Apeiron claims performance figures of 18 million IOPS per shelf/array and an aggregate of 142 million with eight shelves. Latency figures are as low as 100µs with MLC flash and 12µs with Intel Optane drives.

Excelero offers a platform called NVMesh that is deployed as a software system across multiple client servers. Each client server can contribute and consume storage in a mesh architecture that uses Ethernet or Infinband and a proprietary protocol called RDDA. Systems can be deployed in disaggregated mode with dedicated storage or as a converged system. Performance is rated as low as 200µs, with 5 million IOPS and 24GB/s of bandwidth.

Next Article

Storage class memory set to upset hardware architectures

Chris Evans, guest contributor

There was a time when the hierarchy of persistent storage products consisted of just disk and NAND flash storage. But today a new layer has become possible, between disk/SSD and memory.

It is often called storage class memory, or persistent memory, and comprises hardware products made with flash and other persistent media to allow a number of permutations of cost, performance, capacity and endurance.

They are being used to extend the performance of existing architectures and allow storage performance that approaches 200µs and five million IOPS.

So how will storage class memory/persistent memory affect the enterprise, and which suppliers are starting to build them into products?

NAND flash has long ruled the solid-state media market because of the decreasing cost profile of products. With each successive generation of flash architecture, manufacturers have been able to reduce cost by increasing bit density, implementing multiple bits per cell and layering cells on top of each other.

But the drive for better performance is as relentless as efforts to reduce cost, and this has provided an opportunity for new products to come to market.

Samsung recently announced Z-NAND and Z-SSD, a new range of products based on modified NAND technology. Performance from the first released product, the SZ985 (800GB), shows 3.2GBps read/write throughput, excellent read IOPS capability (750K) and good write IOPS (170K). Read latency is between 12µs and 20µs, with 16µs typical for writes.

Much of the performance of Z-SSD could be the result of a high ratio of DRAM to NAND on the hardware (1.5GB of DDR4 memory), plus a high-performance controller. But Samsung has not been forthcoming with architecture details. It is rumoured that SLC NAND is being used, so the actual implementation could be some kind of hybrid device of various NAND formats.

Meanwhile, 3D-XPoint is a technology developed jointly by Intel and Micron. Again, although no specifics have been released, 3D-Xpoint is believed to be based on Resistive RAM (ReRAM) technology.

Flash has outstripped Moore's Law. If software development caught up, the solid-state datacentre would be a cost-effective reality, bringing a step-change in agility and performance.

The interest in persistent memory continues to grow, but a few things – including support from server suppliers – must happen before widespread adoption.

ReRAM uses electrical resistance to store the state of bits of data, compared with NAND technology that stores bits using an electrical charge. In contrast to NAND flash, which has to be written and read in blocks, 3D-Xpoint is byte-addressable, making it more like DRAM but of course, non-volatile.

To date, only Intel has released 3D-Xpoint, under the Optane brand name. Micron uses the brand name QuantX, but has yet to release any products.

Datacentre products such as the Intel DC P4800X delivers 2.5GBps write and 2.2GBps read performance with 550,000 IOPS (read and write) and latencies of 10µs (read/write). Endurance is around 30 DWPD (drive writes per day), much higher than any traditional NAND flash products and comparable to Z-NAND from Samsung.

Between the speeds of DRAM and Z-SSD/Optane lies magneto-resistive RAM or MRAM, specifically STT-MRAM (Spin-transfer Torque MRAM) from Everspin.

Devices such as the Everspin nvNITRO NVMe accelerator card offer effectively unlimited endurance, 6GBps read/write performance with 1.5 million IOPS at average latencies of 6µs (read) and 7µs (write). Unfortunately, capacity is the compromise here at only 1GB per device.

There are other companies also working on ReRAM and STT-MRAM products, including Crossbar and Spin Transfer Technologies, although no details of products from these companies have been made public.

Storage class memory applications

With a plethora of media to choose from, users and appliance supplier now have a range of deployment options.

The more expensive and lower capacity devices offer the capability to be used as a host or array cache that add the benefit of persistence compared with simply using DRAM. The extended endurance of these products compared with NAND flash also makes them more suited for write caching or as an active data tier.

Hyper-converged infrastructure (HCI) solutions can take advantage of low latency persistent memory deployed into each host. Placing the persistent storage on the PCIe or even the memory bus will significantly reduce I/O latency. But this also risks exposing inefficiencies in the storage stack, so suppliers will want to be quick to identify and fix any issues.

Disaggregated HCI solutions, such as those from Datrium and NetApp, should see large performance improvements. In both cases, the architecture is built on shared storage with local cache in each host. Performance is already high with NAND flash, but offers more resiliency (and less cache warm-up) with persistent

caching using products such as Optane. There are also other disaggregated storage solutions starting to use faster media, such as Optane. We'll look at these in the product roundup below.

We're unlikely to see widespread adoption of these faster storage products in traditional all-flash arrays. Storage networks introduce too much overhead and the shared controller model means throughput and latency can't be fully exploited. Price is also a factor here.

We can, however, expect to see suppliers use these fast storage class memory/persistent memory systems as another tier of storage that add an extra bump to performance. It becomes a cost/benefit game where some extra high-speed storage could deliver significant performance improvements.

Storage class memory supplier roundup

Who's using these new systems? In December 2016, HPE demonstrated a 3PAR array using Intel Optane as a cache (a feature called 3D Cache). At the time, HPE were claiming latency figures that were half of those seen without 3D Cache, at around 250µs and an increase in IOPS by around 80%.

Tegile, now a Western Digital company, is believed to use Intel Optane NVDIMM products in its NVMe IntelliFlash arrays. This allows the company to offer products such as the N-Series with latencies as low as 200µs.

Storage startup Vexata has developed a new architecture that uses either all-flash or Intel Optane technology. Optane-based systems claim a typical latency level of around 40µs with seven million IOPS and minimum latencies of 25µs.

In August 2017, E8 Storage, a startup supplier of disaggregated solutions announced the E8-X24, a system based on Intel Optane drives. Although no performance figures have yet been released, latency figures are expected to be well under the 100µs (read) and 40µs (write) figures quoted for E8's existing NVMe flash system.

Apeiron Data Systems, another disaggregated all-flash storage startup has a system based on 24 NVMe drives in a shared storage chassis. Performance figures are quoted as low as 12µs when using Optane drives, with more than 18 million IOPS.

Scale Computing has announced the results of testing its HC3 platform with Intel Optane, seeing figures of around 20µs response time in guest virtual machines. The capability is made possible by the SCRIBE file system that is integrated into the Linux kernel of the platform.

VMware announced the results of testing with Virtual SAN and Intel Optane in March 2017. This showed a 2.5x improvement in IOPS and a reduction in latency by a factor of 2.5. In the virtual SAN architecture, it's expected that Optane would act as the cache tier, with traditional NAND flash for capacity in an all-flash configuration.

NetApp has demonstrated the use of Samsung Z-SSD technology in ONTAP storage arrays. Figures claim a three times greater number of IOPS using Z-SSD as a non-volatile read-only cache. NetApp acquired Plexistor in May 2017 and used the technology to demonstrate a disaggregated architecture at the 2017 Insight event in Berlin. This delivered approximately 3.4 million IOPS at around 4µs of host latency.

We can see from these suppliers that storage class memory/persistent memory is being used to extend the performance of existing architectures.

A few short years ago, we were looking at 1ms latency and one million IOPS as the benchmark for storage arrays. Today those numbers are starting to creep towards 200µs and perhaps five million IOPS.

New architectures (HCI and disaggregated) are being used to minimise SAN overheads even further. Faster storage is available, but at the top levels of performance users may well have to rethink their application architectures to gain the most performance, as the media becomes less of an overhead when storing data.

Looking at the future of media, 3D-Xpoint appears to have just got started, with performance and capacity improvements expected over the coming years. Other technologies (Z-NAND and STT-MRAM) will need to reduce costs and improve capacities to compete. Expect to see companies such as Huawei

develop persistent memory products to complement the SSDs they already produce.

Next Article

NHS trust dumps creaking EMC SAN for Nutanix hyper-converged

Antony Adshead, storage editor

Chesterfield Royal Hospital NHS Trust has expanded its hyper-converged deployment to around 400 virtual machines (VMs) on 24 Nutanix nodes.

The move saw Nutanix replace traditional server and storage architecture, including an EMC SAN, with huge reductions in server room space and cooling as well as snag-free patching of the infrastructure. The project also saw multiple backup products replaced with CommVault.

The Trust is centred on Chesterfield Royal Hospital, which serves a population of 350,000 and has 550 beds and an IT estate that runs around 260 applications.

At the start of the process, in which it was helped by Chesterfield-based integrator Coolspirit, the trust had about 70 Dell servers each dedicated to an application, with 12 more running two VMware ESX clusters.

Storage was provided by an EMC VNXe SAN and it was here that the key bottlenecks arose as it reached capacity.

“The biggest headache was storage,” said IT technical delivery lead, David Sawyer. “When we first got it it had been purchased for requirements at the time. No-one could know the explosion of servers there would be.”

“We had only about 50 VMs then,” said Sawyer. “In seven years that has gone to 400. As that number has increased storage has been battered, with the app guys wanting more and more.”

At the same time, said Sawyer, the compute side was running out of resources, including RAM on occasions.

“We got to the point where we had to ask, ‘Do we throw loads of money into this and keep expanding it, buying shelves and drives?’ In the end we decided to see what was out there,” said Sawyer.

His team considered a setup from NetApp in the traditional three-tier architecture. “They wanted to come along with a pre-configured cabinet, but we simply didn’t have space.”

The trust eventually plumped for Nutanix hyper-converged infrastructure and now has 400 virtual machines running on 24 nodes.

Did Sawyer have any worries about opting for what was a new alternative to traditional IT architectures? “Yes, we had concerns. It was something

completely new but with some research we felt we knew where we were going," he said.

"The attraction was that we could easily add to it and not create bottlenecks. We had been able to add to the SAN but that created a bottleneck between servers and storage. We decided to take the risk and go down the hyper-converged route."

Key benefits of Nutanix

Key benefits of the Nutanix deployment for Sawyer have been space saved in the trust's server rooms and lack of disruption during patching and upgrades.

Previously, patching and upgrades had to be restricted to weekends because of the likelihood of causing unplanned downtime.

"We always expected that something would break," said Sawyer." Often it would be ESX. It may have been something we'd done. We'd patch the server and we'd lose a node. Now we know that upgrades will happen without a blip."

"That means upgrades can happen in office hours without needing to pay for resources at weekend rates and in the knowledge we won't have anything fall down."

Meanwhile, the reduction in total hardware has emptied seven racks of servers and Sawyer is looking to reduce the size of the server room. Alongside that around half of the air conditioning units often switch themselves off.

The big benefit for Sawyer is, however, that compute and storage is now less costly and more reliable, and easily added to.

“The number of VMs has gone through the roof, and traditional storage would have struggled, with extra chassis and disks needed,” he said. “With Nutanix, you put a node in, go back to your desk and turn it on; it’s not a project.”

Next Article

San Francisco 49ers add pace to the team with NVMe flash from Datrium

Antony Adshead, storage editor

US gridiron football team the San Francisco 49ers has opted to deploy NVMe-based flash storage from Datrium to provide the performance needed to support in-game coaching and scouting decision-making.

The move was driven by the expansion of the organisation following a move to a new stadium, along with increased use of technology to support statistics and video-driven decision-making.

Video is used during a game by coaches to review, for example, the opposition's defensive setup and to counter it.

Meanwhile, statistical data – of college players, for example – is used during games, but is also vital to the 49ers scouting operations during the draft, where clubs get to pick new players.

Here, decisions about individuals need to be made by coaches very quickly, with five minutes per draft pick allowed by league regulations.

In all these cases, information in video or data form has to be delivered very rapidly and reliably to team staff.

“These are multimillion-dollar decisions,” said 49ers corporate partnerships vice-president Brent Schoeb.

Datrium will replace a legacy storage estate comprising four different suppliers’ products that “didn’t speak to each other as well as they could”, according to Schoeb, and which had begun to lag in terms of performance and were nearing capacity.

“We needed to make a decision. We were bursting at the seams in terms of performance. We were at full capacity for the entire organisation,” he said.

“It was either add more to the storage we had, or scrap the old system and consolidate on something new,” said Schoeb.

So, the 49ers decided to opt for NVMe-based flash storage from Silicon Valley neighbour Datrium. The team will deploy two DVX systems, each comprising four data nodes and four compute nodes.

Each compute node comprises server hardware with NVMe flash and Datrium software, while data nodes each have 12 drives of 4TB capacity to make a total of just under 400TB.

Schoeb said all critical data would be migrated to Datrium.

Datrium is one of a set of emerging flash storage makers focused on NVMe. NVMe is a subset of PCIe. It comes in card form factor and offers hugely increased I/O performance and lower latency than existing flash products that use the SCSI protocol, a spinning disk-era connectivity method.

NVMe boosts flash performance exponentially by doing away with SCSI. That can be done by slotting it into servers, but a key stumbling block to achieving NVMe's potential in a shared storage environment is the ability to handle controller functionality at speeds that don't bottleneck I/O.

Datrium's answer to this is to put NVMe cards and controller functionality in host servers, with a claimed performance premium of two to four times over SCSI-connected flash.

"We'll be leaning on technology to get competitive advantage. There are constraints such as the salary cap, so we have to use technology and statistics to get the advantage," said Schoeb.

"It's all about making the right decision on draft day. If you don't draft well, you're not a good NFL team," he said. "And on the coaching side it's all about the use of video to call the right play."

Next Article

Public Health England builds Ceph and Lustre cloud for analytics

Antony Adshead, storage editor

Public Health England (PHE) has put open source storage at the centre of its IT strategy, with a private cloud built on Red Hat's Ceph object storage distribution and the DDN-supported Lustre parallel file system for high-performance computing (HPC) analytics.

The move has allowed it to avoid supplier lock-in and build a petabyte-scale storage environment across three sites in southern England to support PHE's healthcare analytics.

PHE was created in 2013 from more than 70 agencies to provide the NHS and other public and private organisations with scientific support in the sphere of public health.

It employs around 5,500 people and core to its activities are data analysis across three key areas.

The first is bioinformatics, which analyses DNA for surveillance of infectious diseases and requires high-throughput computing with many small jobs and lots of CPU and disk input/output (I/O).

The second is modelling and economics, which runs real-time simulations to predict expected disease dynamics.

Third is emergency response, which runs multiple models to predict risks posed by infectious disease threats.

The latter two need mainly traditional HPC, but with lots of CPU and low to moderate disk I/O.

When PHE was formed, it needed to build a shared IT infrastructure for the newly created body.

Windows and VMware were in use for business-as-usual workloads, but choices had to be made over what file system environment would be used for its bioinformatics HPC environment, said Francesco Giannoccaro, head of high-performance computing and infrastructure for PHE.

“[Dell EMC] Isilon was high performance, but quite expensive and not built for high rates of parallel I/O, more for virtual machines. [IBM’s] GPFS was in use with about 40% of the world’s top 500 supercomputer users, but the skillsets weren’t necessarily there,” said Giannoccaro.

PHE went with the Lustre file system on DDN SFA10K hardware for the bioinformatics HPC cluster. Giannoccaro was later asked to begin work on an HPC environment for the whole organisation.

A key consideration was that the three main HPC clusters would not run at full utilisation 24/7/365. For this reason, among others, the organisation settled on a Red Hat virtualisation environment with the KVM hypervisor and an HPC cloud based on OpenStack. This cloud architecture allowed each HPC cluster to burst to additional processing capacity when needed.

In this cloud environment, the PHE scientific data catalogue was developed using iRODS, an open source technology that helps organise, catalogue, aggregate and share large sets of data.

Storage-wise, the bioinformatics 1,500-core Lustre-based cluster is on DDN ES7K hardware with 400TB of storage capacity (plus 500TB of archives).

Meanwhile, the emergency response and statistics and modelling clusters – with 3,000 cores – are on Red Hat Ceph in the HPC cloud that runs on Lenovo blade servers with 500TB of storage and 8PB of archiving capacity.

Flash is used for metadata for Lustre while the Ceph environment comprises hybrid flash tiers of storage.

“Lustre is the technology that’s most widely used for I/O-bounded large amounts of data,” said Giannoccaro.

He said some of the advantages of going down the open source route are that it “reduces vendor lock-in and is open in a way that other organisations can access and share information”

For its cloud infrastructure, Ceph is the storage behind the OpenStack environment.

“It’s the most widely diffused for OpenStack,” said Giannoccaro. “We wanted to find the right balance between the right technology and something that would be supported in a robust way by the right partner.”

The Ceph platform is a software-only product based on multiple storage nodes and a technology called Rados (reliable autonomic distributed object store) that lays out and manages data across multiple clusters.

Red Hat packages and sells a commercial version of Ceph, marketed as Red Hat Ceph Storage.

Ceph supports S3, Swift and native object protocols, as well as providing file and block storage offerings.

Lustre is an open source parallel distributed file system built for for large-scale cluster computing. Lustre’s open licensing and high performance mean it is commonly used in supercomputer applications.

Giannoccaro said PHE’s next moves would be towards use of containers to deliver and rapidly scale content to its web front end using Red Hat’s OpenShift container management platform.

ITV goes on air with SpectraLogic LTFS tape backup and archive

Antony Adshead, storage editor

Broadcaster ITV has deployed more than 2PB of SpectraLogic disk and tape, in a move that has integrated backup and archiving with production apps and freed up capacity on Dell EMC Isilon clustered network-attached storage (NAS) that had struggled to cope with multiple tiers of data.

ITV is the largest commercial television network in the UK. It operates numerous TV channels and delivers content via mobile devices, video on demand and third-party platforms.

The company had been reliant solely on its Isilon clustered NAS infrastructure – comprising about 8PB in total – and other local NAS boxes, to which all parts of the organisation stored production and archive data. But that setup was nearing capacity and had become difficult to manage.

“It was getting full,” said technology director Peter Russell. “And it was a single tier of storage that was being used for everything from fast-turnaround data to longer-term storage.”

When it became apparent a new system was needed, ITV looked at “all the big players in the marketplace”, said senior project manager Marcel Mester.

“But we chose SpectraLogic because we were looking for a platform that could operate with automated tools in the front end but also manual drag-and-drop. We also needed data replication across multiple datacentres that is highly available 24/7,” said Mester.

ITV has deployed SpectraLogic hardware at two separate datacentres. The Greenwich datacentre deployed a BlackPearl and T950 with IBM TS1150 drives, and the Leeds datacentre deployed a BlackPearl and T950 with LTO-7 tape drives.

At each location, the Black Pearl disk cache is 65TB, with around 2PB of tape capacity behind it so far at each site. Production storage is still handled by Isilon, but data is pushed to Black Pearl at the end of those processes. Spectra Logic’s Black Pearl acts as a disk cache and an object storage gateway to Spectra’s tape libraries, which store data on tape using the linear tape file system (LTFS) standard.

LTFS is a file system-style record of files on tapes and so gives NAS-like access to data. This means applications and users can copy to and from tape using drag-and-drop or other standard methods of data movement.

ITV's SpectraLogic setup makes copies that go to the two datacentres on two media formats. "It's risk mitigation in terms of geographical location and against industry failure with regard to media types," said Russell. "We need to keep data for 30 years, so we need to be assured."

ITV organisations using SpectraLogic

Three ITV organisations use the platform: ITV Broadcast, which archives all programmes produced on UK channels; ITV studios, which archives all media produced in making programmes and is the biggest producer by volume; and ITV regional news, which archives programmes and important sequences from 10 regional sites.

A key benefit is that moving secondary data to SpectraLogic has removed the need for its Isilon estate to house multiple tiers of data, said Russell.

"It has very much lightened the load on the Isilon clusters," he said. "It allows them to act as one tier of storage instead of several. The idea was to get the older media off them and now we can use it as a work-in-progress store as originally intended."

Russell also said a key benefit is that applications are fully integrated into SpectraLogic. "Some are fully automated, some semi-automated via drag-and-drop. It is one platform that can accept user-driven ways of operating though to fully automated without middleware."

Getting more CW+ exclusive content

As a CW+ member, you have access to TechTarget's entire portfolio of 140+ websites. CW+ access directs you to previously unavailable "platinum members-only resources" that are guaranteed to save you the time and effort of having to track such premium content down on your own, ultimately helping you to solve your toughest IT challenges more effectively—and faster—than ever before.

Take full advantage of your membership by visiting www.computerweekly.com/eproducts

Images; stock.adobe.com