SDS vs hardware RAID vs software RAID
What is the future?

In this whitepaper, we look at the rapidly evolving capabilities of NVMe storage and its co-existence with SATA/SAS technologies which, together, offer an exciting array of advanced provisioning possibilities.

Determining the pace of change: storage insights for the here and now and beyond

For over a decade the phrase 'software-defined' has prefixed the virtualisation of a range of IT services traditionally managed as separate hardware instances.

It began with software-defined networking (SDN) and soon software-defined compute (SDC), software-defined storage (SDS) and even the entire data centre itself (SDDC) were virtualised to function as elements within a hyper-converged infrastructure (HCI).

The enabler of all of this connectivity is PCIe – the Peripheral Component Interconnect Express – the high-speed bus that has evolved from integrating CPU, GPU, storage and networking to support I/O virtualisation. PCIe also plays host to the Non-Volatile Memory Express (NVMe) interface, designed for optimal connectivity with a new breed of SSDs packaged in several different form factors.

Compared to SAS and SATA, NVMe SSDs are a considerable leap forward in terms of bandwidth, latency and power consumption. Since its standardisation in 2011, the NVMe storage interface has progressed to match the throughput gains that essentially double with each new PCIe generation, climbing from 5GT/s with PCIe 2.0 to 16GT/s with PCIe 4.0. While PCIe 5.0 has arrived, offering 32GT/s, it's still some way off from becoming mainstream.

The current roadmap has already established the PCIe 6.0 specification delivering speeds of 64GT/s with the PCIe 7.0 spec slated for release in 2025 and a goal of 128GT/s, whilst maintaining compatibility with previous generations. By association, NVMe storage will draw upon all these benefits.

Making the connections

These days, hardware supporting PCIe 3.0 and PCIe 4.0 is in abundance and the transition to NVMe SSDs does, on the surface, appear fairly straightforward. For instance, add-in cards (AICs) can be utilised to fit directly into PCIe slots. NVMe drives featuring the U.3 'tri-connector' interface are readily accommodated in server arrays alongside SATA and SAS storage. With drives bays situated at the front, allowing for easy serviceability, U.2 deployments are a practical path for many data centre operations compared to the accessibility drawbacks of AICs.

Alternatively, the M.2 interface offers connectivity for NVMe drives at their full PCIe x4 performance provided that the appropriate M-type connector keying is in place on the motherboard or AIC. B-type keying will only deliver SATA 3 or PCIe x2 speeds. And while you might see SSDs with B+M keying, they are almost certainly SATA 3 devices that offer compatibility between both socket types.

So, while these physical form factors allow NVMe storage to be installed in a system, it's what lies beneath that really matters and brings with it a complete rethink of how best to manage storage within an NVMe ecosystem.

For instance, attaching a single U.2 or M.2 NVMe SSD will, for optimal throughput, use four PCIe lanes (PCIe x4) per device, and a problem among older systems was that a CPU with say, 24 PCIe lanes could soon run short once a GPU taking up PCIe x16 lanes was in place. With more recent hardware, featuring multiple cores enabling up to 128 PCIe lanes, and PCIe switching that expands the lane count, this is less of an issue. Even so, PCIe lane provisioning must be factored into any infrastructure upgrade planning to ensure allocations deliver optimal results.

NVMe storage requires a different approach to deployment. Although familiar options exist to configure drives, such as software RAID and hardware RAID, their utilisation has evolved to embrace the gains that NVMe storage offers over SATA and SAS SSDs.

Software RAID

A pure and simple benefit of NVMe storage is that all the major operating systems feature NVMe drivers to support it. Add an NVMe SSD and, whether the host is Windows, Linux, macOS or Solaris, to name a few, the device will be accessible. VMware's virtualised environments have NVMe driver support enabling a broader range of options well-suited to software-defined storage applications.

This ready availability of NVMe storage devices complements software RAID applications that are featured as standard on all mainstream operating systems. Uncomplicated and effectively free, software RAID functions are available to all, in one form or another – from consumer end-users, gaming enthusiasts and content creators, to full-blown enterprise deployments – providing a convenient gateway to a fundamental set of robust storage management features.

Basic software RAID applications may only offer RAID 0 (stripe) and RAID 1 (mirror) for performance and data safety, respectively. **Indeed, hardware RAID offers many more RAID levels than software alternatives.** Even so, applications such as mdraid, the default software RAID application on Linux, also provides RAID 4, 5, 6 and 10 – combinations that offer a balance of performance and data safety.

As SSDs have yet to match the capacity of individual hard drives, overall storage requirements for a given number of drives is also a major consideration when configuring a RAID array. Moreover, using software to manage data distribution and parity checking functions within RAID storage environments has an impact on the host CPU, which performs these routines. Algorithmic operations can vary in complexity – for instance, writes are more compute-intensive than reads – and if the volume of data throughput is substantial then, with a high level of redundancy in the RAID configuration, these tasks have the potential to impact on overall performance.

And when software licenses are charged on a per core basis, does it really make sense to burden a system with storage tasks? This has long since been the argument for hardware RAID but we're not in a SATA/SAS environment anymore. To a certain extent, the performance penalties inherent in software RAID have been offset by the latency and throughput gains within NVMe and its direct access to the PCIe bus.

Better by design

Moreover, the SATA interface was designed for hard drives and its use with SSDs has always been a compromise. The speed boost that SATA SSDs present over HDDs is extremely productive, but it is a fraction of what flash storage can actually deliver. The Advanced Host Controller Interface (AHCI) used by SATA, with all its legacy idiosyncrasies – over 120 commands built around the physical constraints of spinning disks – enables system upgrade compatibility with flash, but it is ultimately a bottleneck. By contrast, NVMe can function on a minimum of 13 commands – 10 admin and three I/O: read, write, flush.

And when it comes to command queues, AHCI/SATA technology has only one which can send 32 commands per queue. By contrast, NVMe has 64,000 I/O queues, with up to 64,000 commands per queue, which translates into a significantly lower use of CPU cycles.

The streamlined PCIe data path that NVMe storage uses, together with its huge throughput and efficiency, enables software RAID to be viewed in a different light within this domain. Rather than it being regarded as having limitations, software RAID is proving its efficacy in this space. Indeed, for many it has been the only choice as hardware RAID in the conventional sense has had to evolve to deliver features that enable the scaling out of NVMe storage.

Hardware RAID

A hardware RAID PCIe card has a dedicated controller chip that performs all the necessary compute functions to create and manage a RAID array from the targeted storage hardware. The processing is all off-loaded to the RAID card, consequently hardware RAID can offer a wide range of RAID levels with varying complexity, with no processing burden on the host platform.

As expensive host CPU resources are not involved in processing RAID algorithms, read and write speeds are optimised and drive hot swapping is supported. With software RAID, the lack of dedicated processing increases latency, throughput in high-capacity SAS/SATA environments. Unlike hardware RAID, drive replacements often necessitate RAID management procedures prior to removal, with reboots often required too.

Although it comes at a cost, the low latency, data protection and caching features of a PCIe hardware RAID card, along with its drive array expansion capabilities, earns it a place at the heart of enterprise storage management. And it has evolved too. While dedicated NVMe-only RAID cards are still relatively new to market, SATA, SAS and NVMe combined are supported in tri-mode PCIe Gen 4 RAID-on-chip (ROC) cards offered by vendors such as Broadcom, Marvell and Microchip, among others.

These hardware RAID cards provide a simplified way forward for NVMe SSDs to co-exist in mixed storage environments. By following basic cabling procedures, U.2 backplanes can be configured to use combinations of U.2 form factor SATA/SAS and NVMe SSDs.

The emergence of the U.3 standard takes this form factor a step forward, reducing complexity with its unified cabling accommodating a true tri-mode backplane. There is a catch though: the U.3 physical drive interface is the same, but the pin configurations have changed. Consequently, U.3 drives can be used in U.2 backplanes, but U.2 drives are not compatible with U.3 backplanes.

While the mix and match capabilities of U.3 may appear a worthy goal, how widespread such configurations are likely to become is another question.

Bay watch

Certainly, the arrival of the Universal Backplane Management (UBM) standard further enables mixed storage deployments and is compatible with both U.2 and U.3 designs. Championed by a consortium of over 20 leading storage hardware vendors, UBM enables host and controller devices to discover backplane capabilities and supports detection and monitoring of the different drive types (SATA, SAS and NVMe) even within a single drive bay. UBM also functions with SATA/SAS expanders and PCIe switches and provides a range of practical backplane management functions that further enhance U.2 and U.3 system architectures.

A tri-mode RAID or HBA (hardware bus adaptor) card will use x8 or x16 PCIe host lanes and feature PCIe switching to multiply the lane count and effectively increase bandwidth. Card specifications may quote support for say, up to 32 NVMe devices, but this isn't the same as supporting eight NVMe SSDs at full x4 speed, which would require 32 PCIe lanes. In theory, 32 physical NVMe drives at x1 speed could be accommodated and, even in a PCIe 3.0 environment, each one would run at 1000MB/s – two thirds faster than SATA's 600MB/s throughput. Even so, such a configuration would be a sub-optimal use of NVMe SSD storage, given how its superior performance capabilities are hugely increased through its PCIe lane parallelism. In a mixed-use scenario, the tri-mode controller may only dedicate x8 or x16 lanes available to the NVMe storage, which again, involves a choice of fewer drives or a reduced throughput.

Whether different drive types, seamlessly integrated within one backplane, will inspire boutique system builds capable of managing hot (NVMe), warm (SAS/SATA) and cold (SATA/HDD) storage demands within one chassis, remains to be seen.

After all, dividing up PCIe lane allocations to maintain compatibility with legacy storage devices is a compromise that, while enabling NVMe adoption, has its limits and its costs. Many operations, currently satisfied with existing SAS/SATA storage deployments, may only be concerned with drive refreshes to ensure reliability and enhanced capacity. While it's likely that U.2 storage will endure for some time to come, configurations that use one type of device lane are likely to be commonplace to utilise existing SAS/SATA storage assets and lower cost dedicated controllers and expanders. Likewise, to maximise performance gains and capacity, NVMe SSDs are best served exclusively.

The pace of NVMe storage adoption will largely depend on the intensity of the workloads, and how well it augments existing systems. Cloud service providers investing in substantial NVMe-only deployments are already realising the benefits, as the huge gains in bandwidth deliver the means to offer new services that are tiered to suit a wide range of customer needs.

Revising expectations

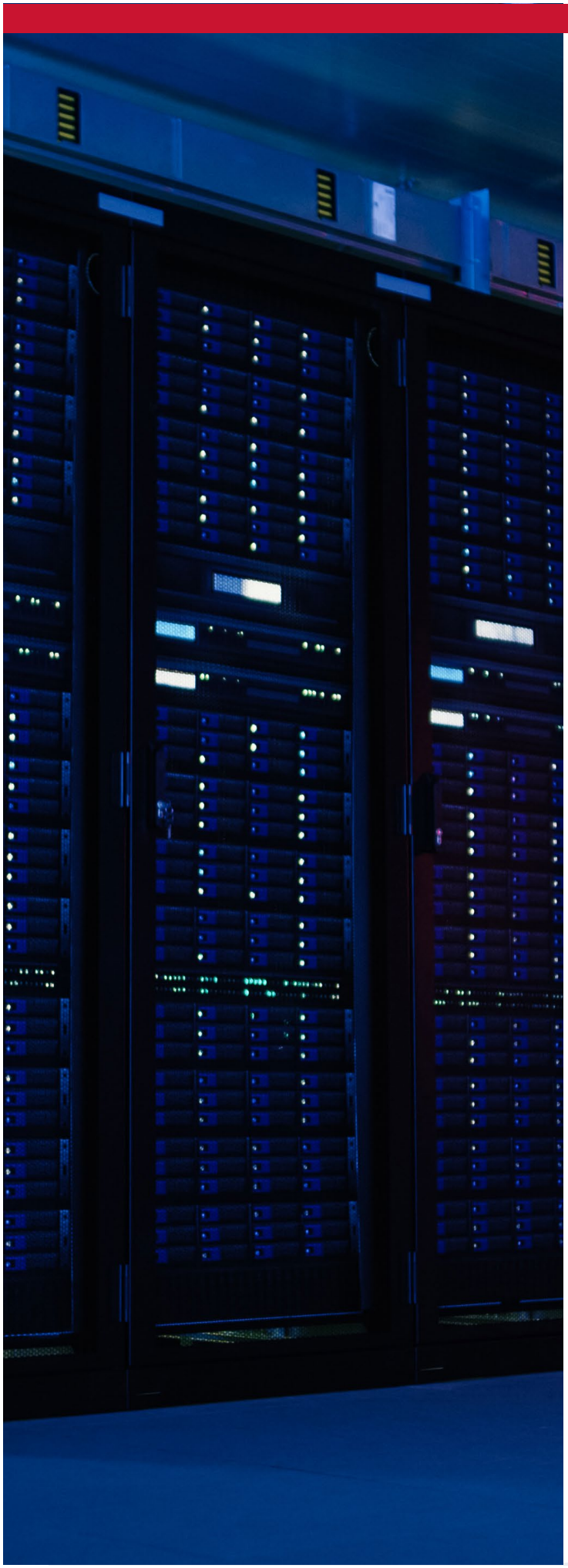
Somewhere in the middle of the extremes, from nimble NVMe-enriched cloud provider to the more traditional data centre, is the enterprise that demands more features, improved efficiency and scalability.

It's embracing NVMe but with a more targeted approach – staggering adoption as costs, benefits, integration and optimisation are studied.

System shortcomings, such as inefficiently coded applications throttling expected latency and throughput gains, quickly come to light when NVMe storage is used in caching. Other bottlenecks will reveal themselves and will need to be addressed to realise the performance benefits of a PCIe/NVMe ecosystem.

This is no like-for-like swap-out but more like a bicycle being replaced by a bullet train. In this respect, SSD specifications need to be reappraised too, as service level agreements may insist on operational measures that fail to recognise how much more you can do with NVMe storage.

One example is the Drive Writes Per Day (DWPD) figure, used to determine the endurance of flash storage over its warranty lifetime period. Flash drives suffer from what is called write amplification, which increases wear on the SSD due to the methodology employed to store data in the memory cells. In essence, cells do not store data directly, but have to be erased first before they can be overwritten and, over time, this convoluted procedure contributes to storage degradation. Over-provisioning, a kind of reserve tank of SSD capacity, is utilised to overcome these issues and perform drive housekeeping routines, such as garbage collection. It's a process of reallocating data to free up storage blocks (that are then erased in preparation for writes) and is a major cause of write amplification.



Getting zoned

An addition to the recent NVMe 2.0 specification is Zoned Namespaces (ZNS), which offers a new approach to NVMe SSD read/write procedures. A zoned block management interface sits between the host and the NVMe SSD. The zoning has some similarities to disk partitioning but at a host application level. ZNS enables the SSD to communicate with the host, describing or 'hinting' at performance capabilities, for instance, providing details of the best patterns and layouts for data placement, as ZNS write, and erase actions are performed sequentially.

This cooperative interaction offloads some of the storage management functions to the host application, with the advantage that it reduces the need for over-provisioning with the potential to expose up to 20 per cent more storage capacity. Implementing ZNS offers improved I/O latency and a reduction in drive amplification of between 4x to 5x. Also, different zones can be allocated specific workloads or data types to enable more predictable performance patterns.

Zoned Namespaces uptake is in its infancy, but ZNS is already a feature of the Linux kernel 5.9. Also, research into ZNS was sponsored by Microsoft, Alibaba, and NetApp – with an eye on large hyper-scaling operations – which suggests that ZNS adoption on an industrial scale is only a matter of time.

Applications will need to be updated to fully utilise this feature set as it evolves and, as a growing number of NVMe drivers now feature ZNS, implementation with existing NVMe SSDs may only need a firmware update in some cases.

For system architects, mindful of exacting specifications, it's time to re-write the rule book on what DWPD actually means. With ZNS implemented, significantly lower write amplification amounts to massive gains in drive endurance. And how many drives do you need too? With hugely reduced over-provisioning, drive capacity is increased considerably. Looking to the future of data management, with NVMe SSDs and ZNS interfacing, you really do get more with less.

Software-defined storage

NVMe brings with it a veritable mix of pathways to adoption, from M.2 drives and PCIe add-in cards to U.2 or U.3 storage. The emerging Enterprise and Data Centre SSD Form Factor (EDSFF) is yet another storage format designed for the NVMe ecosystem featuring drives of two widths (E.1 and E.3) in long and short (L and S) configurations. E.1L drives enable high storage density in a 1U chassis, with the more flexible E.1S size having thermal efficiency benefits that suit scalability. Packaged as a replacement for U.2 2.5-inch SSDs, E.3 drives fit into more conventional 2U server and drive array chassis and are designed to accommodate more flash memory chips per drive to increase storage density.

Certainly, having NVMe as one common standard, with driver support on all mainstream operating systems, makes any of the above options less troublesome to implement. The choices will depend on storage characteristics and configurations that best match the workloads and redundancy requirements. This could involve integrating NVMe storage on edge servers, with SAS/SATA hardware facilitating less intensive operations. Hard disk drive or even tape back-ups could form part of the storage infrastructure too. With no shortage of proprietary platforms in enterprise storage management, orchestrating these disparate storage systems can very rapidly escalate in complexity. This is where software-defined storage (SDS) comes into play, providing the means to harmonise the operations of a mixed storage estate and optimise its utilisation.

In the realm of software-defined storage, the available storage resources are abstracted from the storage hardware and virtualised. Using industry standard protocols, even proprietary hardware can be accessed through SDS virtualisation, with monolithic storage appliances unbound to become part of a larger pool that might also feature new, low-cost scalable storage built with commodity servers. This uncoupling also avoids disruption when storage hardware is replaced, upgraded or expanded.

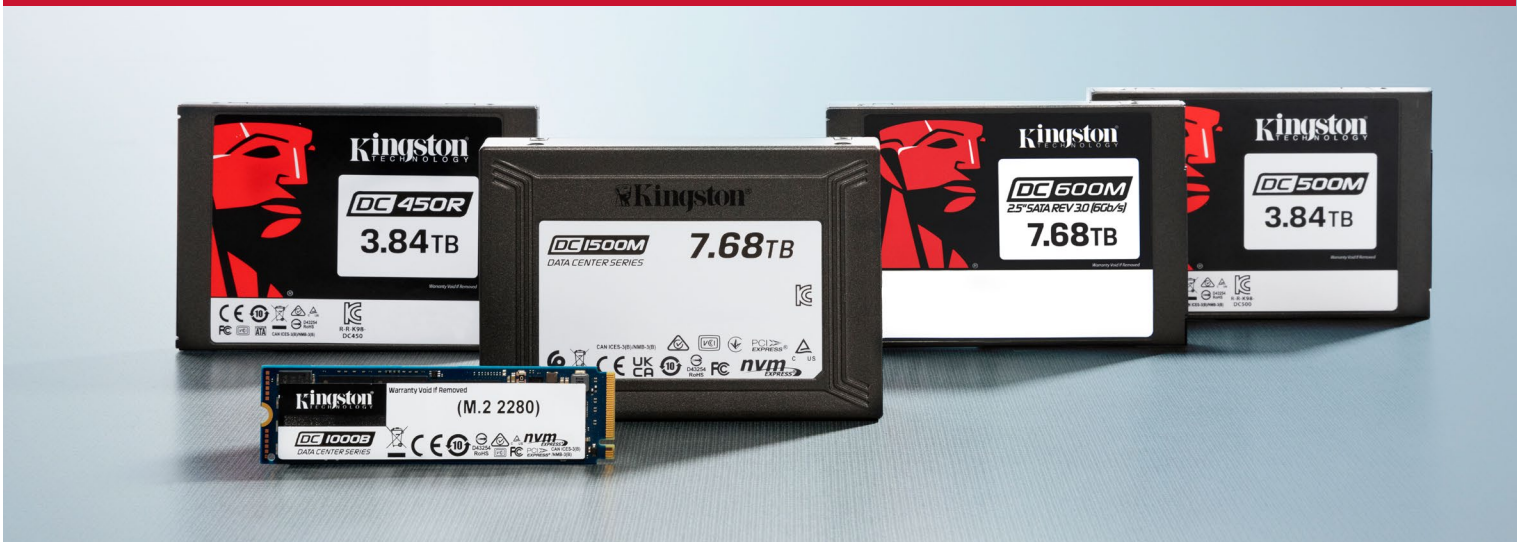
With all the available storage consolidated into virtual pools, decisions on provisioning will need to be made and a wide range of features exist to assist in these assignments, including automation. In the SDS dashboard - hot, warm and cold storage is identified, based on the hardware profiles of the storage in the various pools. And by using scripts, tasks can be executed to allocate and distribute data loads that best match these repositories.

Through its virtual storage layer, SDS delivers both flexibility and scalability; managing the creation and deployment of storage environments suited to enterprise demands and the variable needs of clients, from caching and provisioning virtual machines (VM) to mirroring and replication.

When it comes to NVMe SSDs, SDS platforms can access the storage directly through the PCIe bus using a feature called NVMe passthrough. For example, VMware has its own NVMe storage driver for its ESXi/vSAN SDS platform, which enables the direct assignment of NVMe storage to virtual machines using a feature called VMDirectPath I/O. Depending on the host CPU configuration, a maximum of 16 passthrough devices are supported per VM.

Overall, enabling NVMe passthrough minimises interference from the host, improves performance and simplifies configuring NVMe SSDs for VM instances and other services. To this end, whether a third-party software or hardware RAID controller supports NVMe RAID functions becomes less of an issue with SDS, as it can configure an NVMe software RAID directly.

And while SDS has the potential to be the cure-all for data management, its cost and initial configuration complexity can give pause for thought for some businesses that may have more straightforward requirements. But like the storage itself, these costs are scalable and different versions are available to suit smaller hardware deployments.



Pacing change

Storage is evolving but overnight change rarely occurs, as existing assets are likely to figure in a strategy of planned obsolescence. Hence, storage development continues with technologies such as hard disks and SATA SSDs. They have their place and continue to deliver a useful service in storage arrays. For example, Kingston's DC600M 2.5-inch mixed-use enterprise SATA SSD with a doubling of capacity to 7.68TB.

Hardware RAID and host bus adaptors remain dominant in data centres the world over and vendors continue to innovate to meet the demands of the ever-expanding IT industry.

Through its partnerships with Broadcom and Microchip, Kingston SSDs undergo rigorous trials to ensure they meet the exacting demands of today's data-driven technologies.

Using these leading vendors' storage adaptors, testing programmes involving punishing workloads and challenging configurations ensure that Kingston enterprise SSDs are qualified to deliver on performance, endurance and reliability. Needless to say, Kingston's U.2 DC1500M PCIe NVMe Gen3x4 enterprise SSD has been through it all. With a capacity of up to 7.68TB, combined with 1 DWPD, it's more than qualified to feature in the latest-generation servers and storage arrays.

Even though a software-defined storage estate can be built around commodity hardware, the choice of SSD becomes even more critical at scale. Consumer-grade SSDs may appeal in terms of cost but it's a false economy when compared to enterprise SSDs that are built for endurance and sustained high bandwidth loads. Delivering services from a hyper-converged infrastructure requires performance predictability, so that workloads are managed efficiently and fulfil client expectations. Kingston enterprise SSDs are qualified to work with VMWare storage applications, ensuring that even in the virtual world of software-defined storage, real world goals are met.

Storage provisioning is changing but the pace of change will vary to suit different business models. At every level there are improvements from legacy interface continuity to NVMe innovation. And if upgrading seems like a daunting task, Kingston's Ask an Expert service can help. It offers free assistance in making those vital decisions to suit your business and your budget. So, wherever you happen to be on this journey, Kingston is with you.

#KingstonIsWithYou