WHITE PAPER

Software-Defined Storage with the InfiniFlash™ System from SanDisk: Decouple Storage for Lower Costs, Higher Performance, and Improved Availability

SanDisk
a Western Digital brand

Western Digital Technologies, Inc.
951 SanDisk Drive, Milpitas, CA 95035

www.SanDisk.com

Table of Contents

Introduction 3

Flash in the Enterprise Data Center 3

- **All-Flash Arrays** 4
- **Just a Bunch of Flash (JBOF)** 5
- **Hyperconverged Appliances** 5

Software-Defined Storage with InfiniFlash 6

- **InfiniFlash IF150 Software-Defined Storage Architecture** 6
- **Case Study: Ceph Software-Defined Storage for OpenStack** 8
- **Case Study: SQL Database Storage** 8

Total Cost of Ownership (TCO) Analysis 9

Conclusion 11

Introduction

Organizations are constantly seeking new ways to address workload-specific storage demands in terms of performance and capacity while also meeting service level agreements (SLAs), response time objectives (RTO), and recovery point objectives (RPO). Many IT operations are inspired by successful hyperscale organizations such as Facebook, Google, and Amazon. However, most enterprises lack the scale and substantial development and operations commitment to deploy software-defined storage infrastructure in the same ways. Hyperscale economics also typically don't work out at smaller scale, resulting in poor utilization or unacceptable reliability issues. Another hurdle for enterprise IT is that hyperscale organizations typically have a very small, tightly controlled application environment that facilitates these economies of scale. In contrast, most large enterprises must deploy storage infrastructure to serve a heterogeneous application workload with diverse requirements.

The common thread for both hyperscale and enterprise IT is the compelling rise of software-defined storage. Though early adopters of software-defined storage technologies were cloud builders, the advantages of this approach are now quickly spreading to enterprise data centers with their increased focus on cost reduction, automation, and lock-in avoidance. Software-defined platforms like Ceph are enabling a wealth of diverse and scalable storage solutions based on industry-standard servers. Flash technology has emerged as a key component of these strategies, as organizations encounter the physical limitations of hard disk drives (HDDs) and even hybrid approaches. Unfortunately, aggressive storage vendor marketing and widely diverse architectural approaches often send conflicting messages about how flash should best be used in the enterprise.

InfiniFlash has emerged as a versatile, highly scalable, and cost-effective platform that is easily deployed in software-defined storage architectures. InfiniFlash offers 50x the performance, 5x the density, and 4x the reliability of traditional HDDs, while consuming 80% less energy.[1] Moreover, by decoupling the storage from the compute and networking elements of storage infrastructure, InfiniFlash offers a flexible and cost-effective new approach that eliminates many of the issues with how flash is deployed in the data center.

Flash in the Enterprise Data Center

Flash has long been seen as a solution to performance and reliability problems with HDD-based solutions, but there are significant issues with how flash has been deployed in the data center. As described in the sections that follow, two dominant approaches are prevalent in the marketplace.

- ***All-flash arrays*** and ***just a bunch of flash (JBOF)*** configurations are example of "pool and share" approaches that typically involve consolidating and managing all flash in a single system.
- ***Hyperconverged appliances*** go by many names but can be simply described as giving "every server a little flash." These systems typically have CPUs for applications, and both SSDs and HDDs within their enclosures.

[1] As compared to 200 IOPS HDDs, based on published specifications and internal benchmark tests.

All-Flash Arrays

All-flash arrays are dedicated systems with their own flash storage and software—usually integrating server form-factor SSDs into customized enclosures. While expensive, these systems can offer high performance, along with enterprise-grade features such as deduplication, compression, and encryption. There are numerous all-flash arrays available from multiple vendors.

Unfortunately, the high performance of the all-flash array approach brings with it all of the cost, complexity and monolithic attributes of traditional "scale-up" HDD arrays that many organizations are trying to escape. Storage software is often propriety and will need to be managed as a separate infrastructure element. Storage typically can't be added incrementally, once the capacity of the enclosure has been reached. While all-flash arrays can deliver extremely high performance, it typically comes at a high initial price tag with substantial ongoing costs. Vendor lock-in for both hardware and storage software is expensive over the long term, and costs are high when additional capacity or performance is required—whether overprovisioning to meet SLAs or adding all-flash array frames when additional capacity is needed. Due to software constraints, all-flash arrays have limited ability to integrate with organizations' infrastructure of choice, whether it be physical, virtual, or in the public cloud.

High availability in all-flash arrays is traditionally provided by RAID. This approach can create bottlenecks when rebuilding failed drives, since many devices must feed into a single drive while rebuilding. Furthermore, heavy use of virtualization can cause additional issues via the "I/O blender" effect (Figure 1). This effect occurs when many virtual machines (VMs) running on a single physical server make otherwise sequential I/O traffic patterns appear random. Reads and writes intermixed across several VMs result in random I/O patterns, resulting in slower overall performance as storage array caching and prefetching algorithms become incapable of predicting what actions to take. Performance is impeded, because this results in a long I/O path, with a large number of intermediary devices in between that need to be traversed for an application or user to obtain or save the needed data.

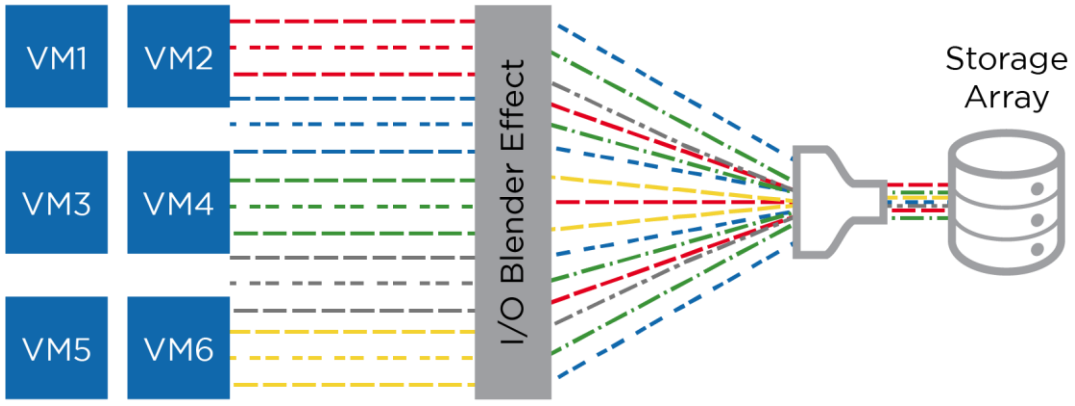

Figure 1. Like traditional HDD-based arrays, all-flash arrays are subject to the I/O Blender effect, which limits performance.

Just a Bunch of Flash (JBOF)

The JBOF approach uses a server platform with a large number of internal drive slots (24, 48, or more), and fills these slots with high-capacity SSDs. The processing cores available on the platform run the storage software that manages the individual SSDs. JBOFs can be constructed incrementally, adding SSDs as storage requirements dictate. This hardware is available from server suppliers such as Supermicro, Quanta, Lenovo and others. Individual SSDs are available from vendors like SanDisk, while the storage software is supplied independently from numerous ISVs such as Microsoft, Nexenta, DataCore, SoftNAS, and many others.

Typically, JBOFs can offer a very low cost of entry—beginning with a single server platform containing one SSD and storage software. Unfortunately, this initial attractiveness can be quickly overshadowed as more SSDs are added. Performance bottlenecks often appear as competing, heterogeneous workloads access flash. Management complexity increases rapidly, because storage is managed at the individual drive level. Additional costs incurred to address these issues quickly eliminate any benefit from the low initial acquisition cost.

Server choice for a JBOF configuration is an essential consideration in order to maintain sufficient performance. The server hosting the SSDs must be able to support the IOPS and throughput that a wide variety of unpredictable workloads will require. It must also provide adequate processing cores, DRAM, and I/O throughput to prevent the selected storage software from becoming the bottleneck. As an example, the I/O bus of a server using a single disk controller is typically limited to approximately 3Gbps. While this may be adequate for many smaller workloads, a 256TB InfiniFlash platform will provide 15Gbps read throughput and 8Gbps write throughput—offering substantial performance headroom.

Hyperconverged Appliances

Architecturally, hyperconverged appliances are usually deployed in a “one size fits all” approach—as a standard infrastructure element without individual customization or configuration. For a narrow range of uniform (non-changing) workloads and data sets, hyperconverged appliances can be configured to deliver very cost-effective performance at large scale. For selected hyperscale environments such as social media and search, these appliances may indeed offer the optimum choice. Because their architecture is simple and easily understood, they are ideal candidates for a pay-as-you-grow incremental investment for a highly homogenous software environment where workloads are known and predictable and the data traffic also remains predictable.

In contrast, most IT organizations are required to support a wide range of applications that are frequently changing. When these diverse workloads emerge, a one-size-fits-all approach using hyperconverged appliances will begin to show severe inefficiencies:

- Workloads requiring additional processor cores will need more hyperconverged appliances assigned to them. To meet this need, the available flash storage will be underutilized.
- Conversely, when additional storage capacity or storage performance is required, additional appliances will be assigned—leading to underutilizing processor cores and software licenses.

- These problems are exacerbated in a highly virtualized environment, where VMs may be created and taken down frequently and workloads are unpredictable. This scenario tends to be CPU or memory limited, resulting in underutilization of storage.

To avoid these inefficiencies, the one-size-fits-all approach must now be modified to a many-sizes-fit-all hardware approach. This in turn results in different appliances configured for different workloads, with these workloads restricted to the "right sized" appliance. These challenges can greatly impact the expected cost and management savings of an otherwise uniform hyperconverged appliance approach.

Software-Defined Storage with InfiniFlash

InfiniFlash takes an entirely different approach to delivering the benefits of software-defined storage with the latest flash technology. By design, InfiniFlash is not engineered with computational CPU cores inside the storage unit. Storage compute is thus separated and decoupled from storage capacity, so that the right amount of compute power can be matched to the right amount of storage capacity for a given workload–no more, no less.

InfiniFlash IF150 Software-Defined Storage Architecture

As a long-standing leader in flash memory development and manufacturing, SanDisk has the unique perspective to bring operational knowledge and innovative flash storage solutions to the enterprise. The InfiniFlash System IF150 in particular is an ideal building block for all-flash software-defined storage infrastructures. The IF150 system, combined with a server platform and a software-defined storage stack, is primarily targeted at Big Data, virtualization, and database environments.

From a price/performance perspective, InfiniFlash is positioned between high-performance, high-cost conventional all-flash arrays and lower-performance, lower-cost JBOF systems. The solution provides scalable capacity and breakthrough economics in both capital expenses (CapEx) and operational expenses (OpEx)–easily surpassing economics for high-performance HDD-based storage systems, while providing additional benefits:

- **Reliability and availability**. Flash memory is fundamentally more reliable than HDDs. The failure rate of flash is approximately 1/20th that of HDDs.[2] Better reliability results in reduced maintenance costs and lower spares inventories required, while contributing strongly to more availability for applications.
- **TCO savings**. A fully populated 512TB InfiniFlash IF150 consumes approximately 450 watts of power. When combined in operation with servers, it will use up to 80% less power than a comparable HDD array requires. The TCO savings improve dramatically if HDD overprovisioning is required to achieve the spindle count needed for IOPS-intensive workloads. For high-density configurations, InfiniFlash can provide up to 6PB of storage in a single rack. This density, combined with lower power, cooling, floor space, server, and software

[2] Flash reliability calculations based on projections from internal testing and comparative data from HDDs.

licensing costs will deliver a significant reduction in total cost of ownership (TCO) over alternative high-performance HDD offerings.

- **Software flexibility**. As a building block for a software-defined storage solution, InfiniFlash allows organizations to choose the ideal storage software for their needs from a wide range of industry-leading ISVs and server vendors. More information can be found through the SanDisk Technology Partner program (www.sandisk.com/business/partners/manufacturers).

The IF150 system delivers from 64TB to 512TB of resilient flash storage in a highly dense form factor for petabyte-scale capacity, high density, and high-performance storage environments. Each IF150 system can be configured with up to sixty-four 8TB hot-swappable solid state cards—delivering up to half a petabyte (512TB) of raw flash storage in a 3 rack unit (3U) enclosure and up to 6PB in a single rack. The system scales easily, and each IF150 system can connect up to eight individual servers with 12Gbps SAS connections.

Unlike many all-flash array systems, InfiniFlash does not use off-the-shelf SSDs, which can impose some limitations on flash storage. Instead, InfiniFlash uses innovative custom SanDisk 8TB flash cards. These cards deliver excellent flash performance, reliability, and storage density with low latency. The individual state of each flash cell is visible to the system's firmware. This property enables the system to better manage I/O, more efficiently schedule housekeeping operations such as free space management (i.e., garbage collection) and, in general, deliver consistent performance in the face of changing workloads and rigid SLA requirements.

Figure 2 illustrates InfiniFlash deployed in a typical software-defined storage architecture. In this example, servers running a choice of software-defined storage (SDS) are each connected to a pair of InfiniFlash IF150 systems via redundant SAS connections. The number of InfiniFlash systems connected to a given server (or servers connected to a given InfiniFlash system) can vary endlessly as dictated by the characteristics of the storage workload. Diverse workloads running on separate servers then access the software-defined storage platforms over a choice of network interconnects.

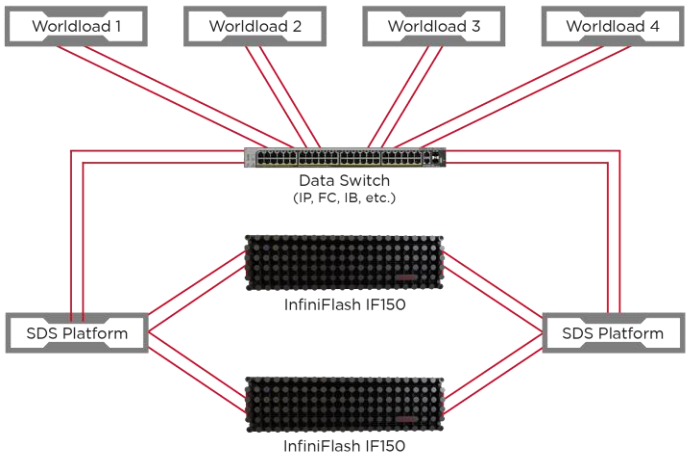

Figure 2. The InfiniFlash System IF150 is decoupled from compute in a software-defined storage architecture.

Case Study: Ceph Software-Defined Storage for OpenStack

Ceph is the leading storage platform for OpenStack, according to semi-annual OpenStack user surveys. Figure 3 illustrates a typical Ceph configuration based on the InfiniFlash system. Importantly, even when using Ceph for OpenStack, the ability to independently tune CPU and storage capabilities is paramount. For example, more CPU cores are required for high performance small-block Cinder workloads while bandwidth-driven large object Swift workloads can get by with fewer CPU cores.

InfiniFlash's high density enables fewer Ceph OSDs per unit of capacity and thereby requires much less CPU compared to sparser SSDs or HDDs. In addition, as users scale the cluster, they can scale the storage compute and storage capacity independently, making for a perfectly balanced cluster for the workload, and delivering a greater cost advantage for large-scale cluster deployments. The superior reliability of an InfiniFlash storage node requires less hardware and provides 10x higher reliability (1.5 million hours of MTBF) than a typical HDD node. Three full copies (3x replication, as is typically used on HDD nodes) are no longer required due to the higher reliability of InfiniFlash.

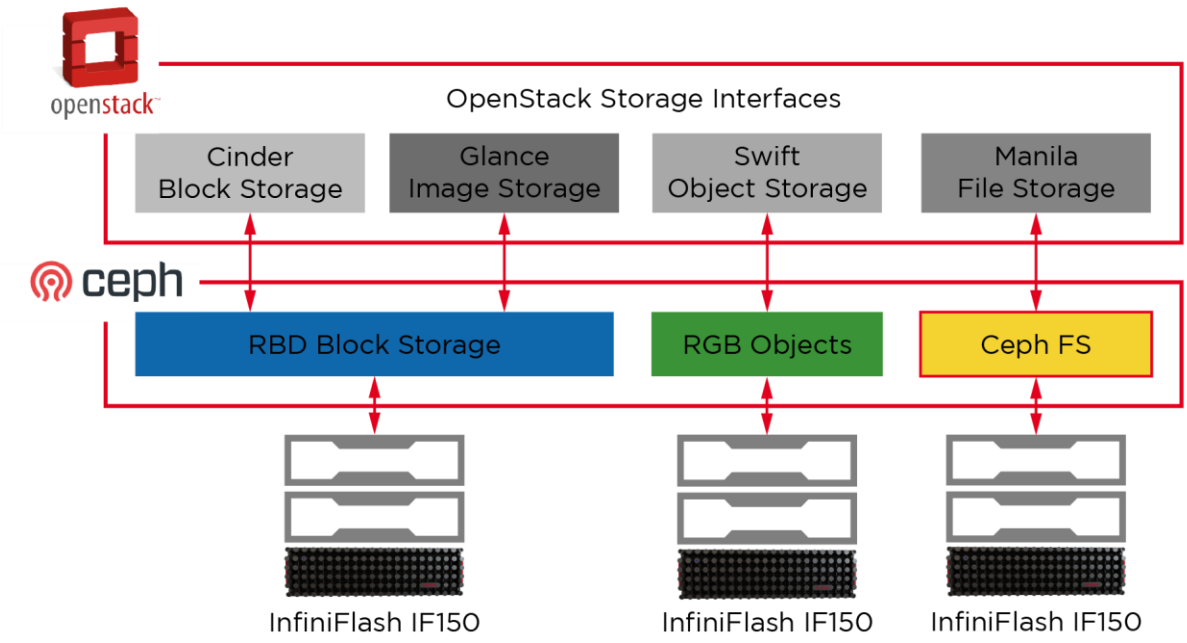

Figure 3. Utilizing InfiniFlash with Ceph storage for OpenStack applications.

Case Study: SQL Database Storage

It is common to deploy storage infrastructure—and hyperconverged appliances in particular—sized for an initial target workload. Unfortunately, when workloads emerge that don't match the initial assumptions, the system can suffer from unused storage and/or performance capacity. InfiniFlash eliminates the need to trade off either poor utilization of CPU, IOPS, or flash in each system to support a diverse set of workloads.

Consider a storage consolidation effort for several SQL databases, each with workload requirements from 2K to 100K IOPS. A traditional all-flash Array with inline deduplication and data compression normally can only achieve around 250K IOPS for a limited capacity point (e.g., around 100 TB effective). Assuming the data is not compressed and can be deduplicated, it might be possible to achieve close to 100TB effective storage–if all of these conditions are consistently met. Alternatively, using a hyperconverged appliance, an organization could configure 1-4TB of flash per server and get plentiful IOPS. However, those SQL workloads that don't require high IOPS or don't require the capacity would result in unused capacity and/or performance in each server.

In contrast, InfiniFlash allows storage compute and storage capacity to be varied independently for individual workloads. This flexibility effectively avoids the unused and underutilized capacity problems of hyperconverged appliances. From a performance standpoint, each InfiniFlash system can typically achieve over 800K IOPS and provide 64-384TB of usable capacity (utilizing erasure coding or RAID 6) in only 3U of rack space. Moreover, the system provides a true usable capacity not impacted by compressibility or dedupability of data. It also provides enough IOPS to handle most SQL and virtual desktop infrastructure (VDI) workloads and Big Data throughput.

InfiniFlash can also eliminate read-while-writing performance impacts which can be key to database performance. Also known as performance read scaling, this effect occurs most frequently in OLTP environments. Routine report generation tasks can cause system performance to suffer when a dataset is analyzed while being simultaneously updated. InfiniFlash can provide an authoritative copy /duplicate using products such as Oracle's Data Guard, GoldenGate, or other software stacks that can be used for high performance access. This model allows the primary enterprise-class storage array to focus on taking writes and on information life management. The copies on InfiniFlash can then be used as a local (metro area) copy or a report/analytics feed without causing any impacts to the primary storage array—thus allowing more write I/O operations to happen without the added cost of replacing it with an all-flash array. With this configuration, no performance impact will occur on the OLTP system, and reports no longer need to be scheduled for after-hours for fear of system degradation.

Total Cost of Ownership (TCO) Analysis

Instead of weighing storage decisions by simplistic metrics such as dollars per gigabyte, most modern organizations are already looking at the longer-term value and costs of their storage infrastructure. With storage needs escalating so dramatically, the cost and availability of sufficient power and cooling is a significant consideration, as is the reality of finite available rack space. Cost is certainly a primary driver for software-defined storage, since it allows organizations to take advantage of savings over their legacy appliance vendors by avoiding lock-in and giving them the ability to select the exact server hardware that meets their needs.

Coupled with software-defined storage, InfiniFlash introduces many opportunities for cost reduction, including:

- **Performance**. HDD-based and hybrid systems often need to be overprovisioned to achieve desired levels of throughput or IOPS. The superior performance of InfiniFlash requires stocking far fewer spare devices and provides all-flash performance at the cost of HDD-based systems. The performance impact of constant HDD rebuilds is also eliminated.

- **Density and efficiency**. Density is a desirable characteristic, but not if it drives unacceptable levels of heat and power consumption. InfiniFlash allows up to 2PB in 16 rack units utilizing only 2,500 Watts, significantly reducing energy consumption as compared to HDD-based systems.
- **Reliability**. With HDDs in a large data center environment suffering up to a 4% failure rate, there is clearly room for improvement. InfiniFlash offers significantly higher reliability than HDD and hybrid systems. Fewer failing drives means less performance degradation due to HDD rebuild times, and less likelihood of pulling the wrong drive. Drive replacement can shift from a monthly event to an annual event.

Figure 4 illustrates the result of SanDisk TCO modeling comparing the three-year TCO of InfiniFlash as compared to that of an all-flash array competitor providing 500TB of raw storage for a database application.[3] When all factors are taken into account, InfiniFlash requires around half the TCO of the more expensive all-flash array.

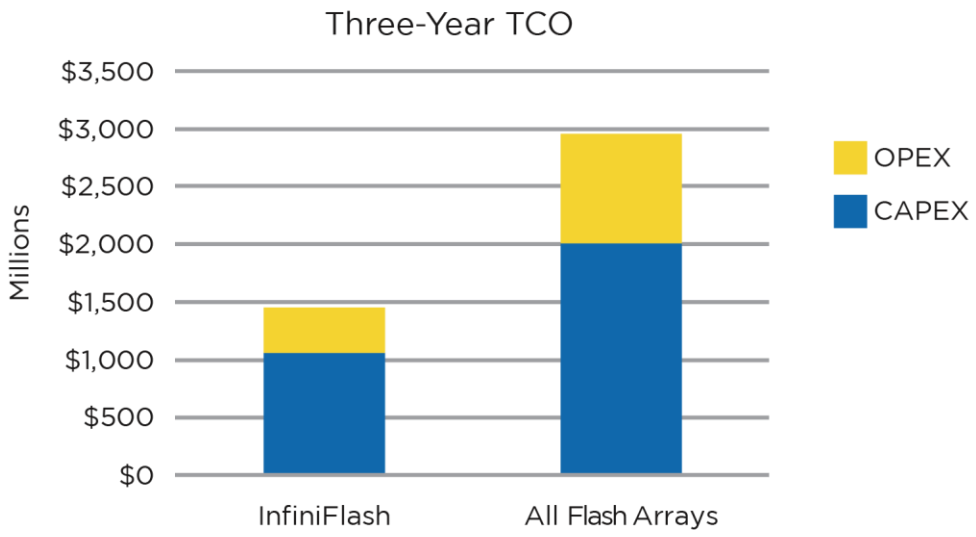

Figure 4. Three-year TCO comparing InfiniFlash with an All Flash Array.

[3] The TCO calculation assumed two replicas of data and a power cost of $0.12 per kilowatt hour. TCO analysis details are available at www.sandisk.com/infiniflash.

Conclusion

Over time, flash has been replacing HDDs in the enterprise. This shift has recently become more dramatic, driven by cost and capacity improvements in flash technology. As a leader in flash memory development and manufacturing for close to three decades, SanDisk has been at the forefront of these trends. SanDisk data center flash now powers the top 12 worldwide Internet properties and cloud providers. By deploying InfiniFlash in software-defined storage environments, enterprise IT and Cloud architects can benefit from the architectural approach chosen by other industry leaders.

When architected properly, software-defined storage can greatly reduce costs while providing the performance and availability equivalent to the best-of-breed traditional storage environments and new hyperconverged storage solutions. With its innovative design and high performance, the InfiniFlash System IF150 from SanDisk delivers a unique capability to provide all-flash software-defined storage for heterogeneous, scale-out, and scale-up storage environments. InfiniFlash delivers consistent performance without overprovisioning or actively balancing workloads across compute nodes.

By decoupling storage capacity and performance from storage server platforms, support for heterogeneous workloads is greatly improved with dramatically better utilization. No storage is overprovisioned or wasted when additional processing cores are needed, and no cores or DRAM are underutilized when additional storage capacity is required. The storage is managed as a single entity, not as individual cards or drives. By using the appropriate software-defined storage technology InfinFlash blends seamlessly with existing IT management infrastructure.

Authors:

Roark Hilomen–Data Center Architect, Data Center Systems, SanDisk. Roark is a senior member of the Technology and Advanced Development group and is responsible identifying new areas to enable flash adoption. He is a 20-year veteran in development and operations in companies ranging from Quest Software (Now Dell), Pillar Data (Now Oracle), Network Associates to Apple computer. Previously Roark was an eBay Storage Architect/Infrastructure Technologist. He enabled eBay's Flash adoption strategy for DBs, NoSQL, Cloud and Big Data. While at eBay, Roark's team received the Flash Memory Summit Flash Innovation award in 2014.

Dave Montgomery—Marketing and Strategy Manager for SanDisk's InfiniFlash all-flash storage system. Dave has over 20 years of experience in the enterprise storage, computing, and software industry. He has a deep understanding of storage and ISV ecosystems required for business-critical IT infrastructures and has extensive experience in complete life-cycle software product management and product marketing at companies such as NetApp and HP.