White Paper

FAS Hardware: Optimized for I/O, Expandability, and Reliability

Steven Miller, Philip Trautman, NetApp
September 2014 | WP-7193

Executive Summary

NetApp designs all the models in its FAS hardware line to deliver tremendous performance, scalability, and availability with superior manageability, ease of service, and ease of use. By using industry-standard components, we're able to deliver the same architecture and capabilities across the entire FAS product line, so you can move from one platform to the next with minimum complexity and no business risk.

This white paper describes how we optimize memory and I/O bandwidth to provide superior block and file performance while providing the network connectivity and capacity scaling needed to meet growing and changing data center requirements. At the same time, we deliver proven availability in excess of 99.999%.

TABLE OF CONTENTS

1 Introduction 3

1.1 Why NetApp? 3

2 The NetApp FAS Advantage 3

2.1 Maximizing Reliability and Availability 4

3 Optimizing I/O 4

3.1 Storage I/O Data Paths 4

3.2 Optimizing Memory Bandwidth 4

3.3 NVRAM 5

3.4 Maximizing I/O Bandwidth 5

3.5 Optimizing Driver Design 5

3.6 What It Means to You 5

4 Maximizing Expandability, Connectivity, and Scalability 6

4.1 PCIe Slots 7

4.2 Network Connectivity 7

4.3 Drive Subsystems and Flash Integration 8

4.4 What It Means to You 8

5 Reliability, Availability, and Serviceability 9

5.1 FAS Storage Design 9

5.2 Protecting Availability of Deployed Systems 10

5.3 What It Means to You 11

6 Conclusion 11

Appendix: Intelligent Caching of Write Requests 11

Journaling Write Requests 11

Optimizing Write Performance 12

LIST OF TABLES

Table 1) Comparison of published FAS versus Isilon SPECsfs benchmark results. 6

Table 2) NetApp entry, midrange, and enterprise expandability and connectivity versus storage and servers. 7

1 Introduction

Because modern servers have become so capable, it's not uncommon for even very industry-savvy people to conclude that a server platform can perform the same functions—at the same level of performance—as a dedicated storage system. However, data storage systems face a unique set of requirements and challenges that aren't easily addressed by standard server hardware.

Storage must provide applications and business processes with consistent high-speed access to data, but that's really just the beginning. Complete integration between hardware capabilities, storage software functionality, and the broader ecosystem of protocols, applications, hypervisors, and management tools is needed.

Focusing on only one component of the data storage solution without the others can result in poor performance, increased complexity, or higher cost. A tremendous amount of work goes into designing a platform capable of delivering the necessary throughput, scaling (capacity and connectivity), and overall ease of service and reliability that today's data centers need—at the right price point.

1.1 Why NetApp?

NetApp has been developing industry-leading storage platforms for 20 years and understands these challenges. A huge amount of engineering time and resources goes into designing advanced storage hardware and leading storage software and integrating them with our large portfolio of third-party products. NetApp has the broadest partner ecosystem in the industry. The result is storage platforms that combine tremendous performance, scalability, and availability with superior manageability, ease of service, and ease of use—with options priced to address almost any budget.

This paper discusses the hardware design philosophies that guide the way we build our NetApp® FAS storage line, as well as the factors you should consider when deciding which enterprise storage array to purchase.

2 The NetApp FAS Advantage

Before going into the specific considerations that drive the way a FAS storage system is built, it is important to understand the overriding design goal that differentiates FAS hardware from competing storage platforms and standard server designs: a unified design architecture. All FAS system controllers are fully compatible, enabling you to quickly upgrade to a new FAS system without migrating data. Other vendors have significant architectural differences between their entry-level, midrange, and enterprise storage systems, making it difficult or impossible to transition from one platform to the next as your business grows.

FAS systems might seem similar to other server and storage system designs, but the way components are integrated is significantly different. NetApp FAS systems:

- Are optimized to perform I/O
- Offer broad network connectivity
- Scale to support hundreds or thousands of drives
- Integrate flash technology to accelerate performance
- Are built to detect and stop errors as data moves through the system
- Eliminate single points of failure and routinely achieve availability beyond 99.999%

NetApp FAS hardware delivers the innovations that continue to move the storage industry forward. NetApp was the first to capitalize on SATA storage and the first to make the transition from Fibre Channel to SAS. It was the first major storage vendor to ship Gigabit Ethernet (GbE) and the first with 10-Gigabit Ethernet (10GbE) and Fibre Channel over Ethernet (FCoE) as well. NetApp was also one of the first to recognize the tremendous value of flash for accelerating storage performance in combination with existing SAS and SATA hard disk drives (HDDs).

2.1 Maximizing Reliability and Availability

Compute and network gear are essentially stateless; this equipment is relatively easy to shut down, replace, and bring back up. Because you can't replace storage without first moving the data elsewhere, the burden of availability for storage is much heavier. NetApp designs storage to deliver a level of reliability, availability, serviceability, and manageability that matches the critical importance of the data you store on our hardware. It's not uncommon for tens or hundreds of servers to get their data from a single highly available NetApp storage system.

3 Optimizing I/O

Any hardware platform has to be optimized to suit its intended use. A server platform, for instance, is built to host applications and other user processes that consume significant compute resources. Storage systems, on the other hand, are designed to store and serve data to network clients and/or SAN hosts as well as perform critical data protection tasks (for example, replication) to safeguard information. This difference is critical and has significant implications for how quickly your storage can feed information to applications.

A NetApp storage system does 10 to 20 times more I/O per core than a typical server. Because of the variety of data protection and replication tasks that are typical in today's data centers, it's not uncommon for us to do four to seven bytes of I/O out the back end for every byte that comes in.

NetApp hardware is optimized to accommodate this tremendous I/O load. In addition, we make full use of the compute horsepower provided by modern processors to drive the many capabilities that NetApp offers, including compression, encryption, and deduplication.

3.1 Storage I/O Data Paths

Storage system I/O comes down to two basic types of requests:

- Requests that read data
- Requests that write data

(Of course, there are many types of requests that may be received by a storage system. This is especially true with NAS protocols, but at the most fundamental level all the various types of requests result in data—or metadata—being read and/or written.)

When a read request comes in from a NAS client or SAN host, it traverses the client network or SAN and arrives at an I/O port on the storage system. At that point, the request—which could be one of thousands of requests received every second on a busy system—has to be efficiently processed, and the data being read has to be identified and returned to the client or host.

Write requests also traverse the client network or SAN to arrive at an I/O port on the storage system. A write request has to commit data to permanent storage as quickly as possible and acknowledge the writer, so this is a major area for I/O optimization.

In both cases, opportunities exist to optimize I/O along the entire path.

The NetApp FAS architecture is tuned for I/O in two respects:

- Optimizing memory bandwidth for both system memory and nonvolatile RAM (NVRAM)
- Maximizing I/O bandwidth on interface cards, buses, storage interconnects, and storage devices

3.2 Optimizing Memory Bandwidth

Every time a request is sent to a storage system to read or write data, blocks of data have to be passed in and out of the storage system's memory. To handle the I/O requests coming from dozens or even

hundreds of physical or virtual servers, a storage system needs to be architected to have both sufficient memory and memory bandwidth to accommodate the volume of requests.

All FAS controllers are designed to deliver the maximum available memory bandwidth from each processor to deliver maximum throughput for I/O requests. NetApp accomplishes this by using a single large-capacity, high-performance memory module on each channel from each Intel® processor. Unlike server designs, which heavily load memory buses, NetApp FAS designs minimize memory loading to deliver maximum memory speed. Some FAS models achieve twice the memory bandwidth. This approach not only improves I/O performance; it also allows us to drive a better TCO, because less memory is needed to achieve specific performance goals. Since memory is on the data path, memory bandwidth is critical to achieve the lowest latency for each I/O operation.

3.3 NVRAM

When data is written to a storage system, it has to be written to permanent storage as quickly as possible. Storage protocols such as NFS require that this occurs before the write is acknowledged as a success. FAS controllers use nonvolatile memory (NVRAM or NVMEM) as a journal for incoming writes, allowing the system to commit write requests to nonvolatile memory and respond to writing hosts without delay. This is much different from the typical approach. Typically, write caching is far down in the stack. This means that you have to accelerate the entire stack to accelerate writes, increasing latency. It also means that you have to decide immediately where a particular data block is going to be written, making it more difficult to optimize writes. By comparison, FAS storage allows I/O to be optimally scheduled to minimize disk head movement for HDDs and to optimize data placement on both HDDs and SSDs.

3.4 Maximizing I/O Bandwidth

For maximum I/O performance, it is equally important to optimize the I/O pathways in and out of the storage system in order to provide much greater bandwidth for both client networks and back-end drive shelves. The FAS8080 EX offers 80 PCI Express (PCIe) Gen3 lanes with over 100GB/sec of bandwidth. In comparison, a standard server design usually offers only 20 to 30 lanes.

3.5 Optimizing Driver Design

The interface cards connected to PCIe—and how efficiently they are utilized—also affect I/O bandwidth. NetApp works closely with chip manufacturers and I/O card designers so that the interface cards it offers with FAS systems deliver maximum I/O bandwidth to front-end networks and back-end storage.

All drivers for interface cards are designed by NetApp to deliver maximum I/O efficiency and minimum latency.

3.6 What It Means to You

Does maximizing memory and I/O bandwidth in this fashion make a difference? We believe it does. If you look at other scale-out architectures, you will notice that they need many more nodes—with up to six times as many cores and up to five times as much memory—to approach the performance of FAS storage. As Table 1 illustrates, our storage controllers, or the combination of our controllers, our software, and our customized hardware drivers, appear to be significantly more efficient, delivering six times more IOPS per core.

Table 1) Comparison of published FAS versus Isilon SPECsfs benchmark results.

	FAS Cluster*	Isilon S200**	FAS8020***	Isilon S210****
Year Published	2011	2011	2014	2014
# Nodes	24	140	2	14
CPUs/Cores	**48/192**	**280/1,192**	**2/12**	**28/168**
Total Memory	1248GB	6790GB	56GB	3612GB
Flash	12TB (Flash Cache™)	25TB (140, 200GB SSDs)	1TB	10TB (14 x 800GB)
10GbE Networks	72	140	6	14
Number of Disks	1,728	3,360	144	322
Exported Capacity	288TB	864TB	56TB	145TB
SPECsfs2008_nfs.v3 Ops/Sec	**1,512,784**	**1,112,705**	**110,281**	**253,357**
Ops per Core	**7,878**	**933**	**9,190**	**1,508**

* http://www.spec.org/sfs2008/results/res2011q4/sfs2008-20111003-00198.html
** http://www.spec.org/sfs2008/results/res2011q2/sfs2008-20110527-00186.html
*** http://spec.org/sfs2008/results/res2014q1/sfs2008-20140120-00235.html
**** http://spec.org/sfs2008/results/res2014q3/sfs2008-20140609-00249.html

4 Maximizing Expandability, Connectivity, and Scalability

Storage systems, especially unified storage systems, require much greater expandability and connectivity than a typical server. On the front end, there's a need for multiple types of connectivity to support SAN hosts and NAS clients. This typically includes Fibre Channel ports as well as Ethernet connections supporting communication protocols such as FCoE, iSCSI, NFS, and CIFS.

On the back end, multiple ports are needed to connect to drive shelves and storage media. And scale-out FAS configurations running the NetApp clustered Data ONTAP® operating system also require network ports for the cluster interconnect. Essentially, the more scale you want from a storage system, the more I/O expansion is needed. For a storage vendor this means looking at three things:

- The network connectivity you provide for attaching to different client networks
- The back-end storage connectivity you provide for scaling up capacity
- The number of PCIe slots you provide to allow flexibility when it comes to expansion

NetApp storage systems use I/O expansion slots in several ways:

- Front-end network connectivity (Fibre Channel, Ethernet)
- Back-end storage connectivity (SAS, Fibre Channel)
- Performance acceleration (Flash Cache intelligent caching)

Table 2 shows a comparison of NetApp FAS systems with the HP 3PAR StoreServ 7000 and several high-end rack-mount servers from leading suppliers. This isn't meant to disparage the servers in any way, but merely to point out the significant differences between storage systems and even high-end server platforms with respect to ports, maximum drives supported, and the number of available expansion slots.

NetApp provides a complete selection of built-in ports, including Ethernet (Gigabit Ethernet and 10-Gigabit Ethernet), Fibre Channel, and SAS. This provides a base configuration that is already highly capable in terms of connectivity.

4.1 PCIe Slots

The PCIe slots that NetApp builds in to its controllers translate into greater expandability. NetApp FAS8000 models offer up to 24 full-length PCIe Gen3 slots. These slots can be used to add Fibre Channel, SAS, and Ethernet connectivity to supplement the on-board ports that NetApp supplies. They can also be used for NetApp Flash Cache intelligent caching cards to accelerate read operations. Note that full-length PCIe cards can offer greater functionality and I/O advantages. All the non-FAS systems shown in the following table use at least some half-length slots, limiting I/O potential.

Table 2) NetApp expandability and connectivity versus competing storage and servers.

	FAS2500	FAS8080 EX	HP 3PAR StoreServ 7000	HP ProLiant DL380p Gen8	PowerEdge R910
Form Factor	2U or 4U	12U	4U	2U	4U
PCIe Slots	NA	24	4	6 (with optional 3-slot riser)	7
Max Drives	144	1,440	480	12	16
RAID	Integrated	Integrated	Integrated	Optional	Optional
Ethernet Ports	NA	8 (10GbE)	8 (GbE)	1 NIC (2-port or 4-port)	1 NIC (4-port)
FC Ports	0	8–32	8	0	0
Unified Ports*	8	8	NA	NA	NA
SAS Ports	4	8	4–8	0	0

*Unified target adapter 2 ports can be configured either as 10GbE or 16Gb/sec FC.

4.2 Network Connectivity

The data center landscape continues to evolve rapidly, and storage systems have to be able to evolve along with IT needs. The protocols and interfaces you use today might not be the interfaces you need next year, or even next month. FAS systems accommodate this need by providing a wide variety of on-board interfaces and the ability to add substantially more connectivity of any type as needed.

With the exception of our entry FAS2520, all FAS systems provide flexible on-board unified target adapter 2 (UTA2) ports that can be configured either for 10GbE (including FCoE) or 16Gb/sec Fibre Channel. These UTA2 ports provide a more compact design and greater flexibility to support a wide variety of storage needs while leaving expansion slots free for other functions. Large numbers of additional ports can be provisioned. For example, a dual-controller FAS8080 EX supports a maximum of 64 16Gb autoranging FC ports, 64 10GbE ports, or 72 GbE ports.

NetApp has been a leader in the delivery of unified target adapters capable of delivering both Ethernet and Fibre Channel connectivity from the same adapter. Converged networking makes it possible to consolidate Fibre Channel traffic with LAN and IP data traffic on the same network cabling, simplifying

your data center infrastructure. Whatever network connectivity you need can be supported from the same device without hardware changes. If you need more of one network technology and less of another, FAS storage can adapt to your needs.

4.3 Drive Subsystems and Flash Integration

When you deploy a storage system, you want something that can scale to meet your storage needs over an extended period of time. Otherwise, you may end up with many storage systems that are not fully utilized and that make management extremely difficult. For this reason it is important to have an array that is flexible enough to support new storage types as your needs change. This includes capacity-oriented HDDs, performance-oriented HDDs, and SSDs.

As you would expect, the drive subsystem is another point of differentiation for FAS systems. A typical server configuration connects only a few drives at most. The servers shown in Table 2 go up to 16 drives. The maximum number of drives supported in FAS systems ranges from 84 to 1,440. Media options include high-capacity drives (up to 4TB), performance drives (10K RPM, up to 1.2TB), SSDs, and self-encrypting drives.

NetApp has worked closely with silicon vendors such as PMC-Sierra to go far beyond the typical 50–100 drives seen with other storage vendors' products. NetApp has moved aggressively to SAS as a storage interconnect from Fibre Channel for three main reasons:

- **Point-to-point isolation.** SAS provides complete isolation for each attached device, thereby avoiding many of the problems that can affect resiliency.
- **Bandwidth.** NetApp primarily uses 6Gb/sec SAS links. Each SAS "wide port" supports four SAS lanes for an aggregate bandwidth of 24Gb/sec per port or 96Gb/sec per four-port card.
- **Connectivity.** The number of drives that can be connected to a single SAS port is limited primarily by performance considerations, not disks, enabling a higher degree of expansion.

However, simply being able to cable up a large number of drives is not the whole story. You have to be able to utilize the drives effectively and protect the data on those drives. The optimized I/O paths of FAS controllers and our closely integrated RAID design enable NetApp to take advantage of the large numbers of drives we support.

Flash integration is another area in which FAS storage excels. All-flash FAS systems (configured with SSDs only) provide maximum performance and minimum latency for use cases such as virtual desktop infrastructure and database. In hybrid configurations, FAS storage controllers accommodate up to 24TB of flash in PCIe slots. NetApp FAS storage systems can also use SSDs in NetApp Flash Pool™ aggregates that combine HDDs with a relatively small amount of SSD capacity to accelerate random reads and writes.

SSDs provide the most benefit for random, transactional workloads. HDDs are quite good at sequential performance, especially on a cost basis. Combining the two types of media lets you benefit from flash for its transactional performance and HDDs for sequential performance without having to know the exact I/O behavior of your storage workloads; many workloads do both transactional and sequential I/O.

4.4 What It Means to You

NetApp FAS systems match a tremendous level of I/O optimization with the connectivity and expandability that storage systems demand. At the enterprise level, our current flagship, the FAS8080 EX, provides 24 expansion slots and is capable of supporting dozens of Fibre Channel, Ethernet, and SAS ports. A single FAS8080 EX HA pair supports up to 1,440 drives for a maximum raw capacity of 5760TB. A cluster of such systems running clustered Data ONTAP scales to a maximum of 17,280 drives and over 100PB of raw capacity. No other storage offers this level of scalability.

5 Reliability, Availability, and Serviceability

Delivering the best possible user experience with the highest level of reliability and availability isn't about any single thing. It's about everything that goes into a storage system, from the initial design through the product lifecycle, including the eventual retirement of that product from your data center. This includes hardware and software design, component selection and sourcing, testing, manufacturing processes, integrated monitoring and alerting, and both automated and hands-on support. It also means measuring how storage systems perform in the field and committing to make continuous improvements.

5.1 FAS Storage Design

When NetApp designs a storage system, resiliency isn't an afterthought. Every FAS system is designed to deliver full component redundancy and HA, and every design is carefully tested.

We do exhaustive quality assurance testing to make sure we are doing everything we can to deliver the best product possible. In many cases, a vendor will run only a few thousand system hours of testing; however, some issues might not appear until you go well beyond 10,000 hours.

The FAS8000 went through half a million hours of test time before we shipped a single unit. The annual component replacement rate for the FAS8000 series is less than half the rate of our earlier systems, which were already highly reliable. Testing is one of the crucial elements that differentiate enterprise storage from non-enterprise gear.

Component Selection and Redundancy

All FAS designs use enterprise-class components (processors, chipsets, drives) rather than less-expensive components with lower duty cycles and higher failure rates. Component redundancy is designed in to controllers, drive shelves, and interconnects to further protect against component failure. NetApp RAID-DP® technology protects against disk failures, and dual-controller, active-active high-availability configurations protect against full controller failures.

Error Detection and Correction

FAS systems are built to detect and stop errors as data moves through the system. Errors are a possibility in any electronic device due to alpha particles, component issues, and so on. NetApp FAS storage includes safeguards in both hardware and software to detect and stop the propagation of corrupt data.

For example, all disk blocks have a checksum to verify that when a block is read it is the same as when it was written. The NetApp Flash Cache design uses a cyclic redundancy check on each cached block so that the correct data is returned. This is in addition to the typical error correction used for flash data.

Out-of-Band Management

All FAS controllers include a service processor that remains operational even when a controller is down. This provides remote power cycle, call-home notification, FRU reporting, current/voltage and temperature sensing, and more. This capability is typically not found in entry-level storage. In many cases, the service processor can inform NetApp of issues before a failure occurs, so that service can be performed proactively.

Alternate control path (ACP) provides out-of-band management for NetApp drive shelves. This allows a storage controller to reset a misbehaving storage channel without communicating over that channel and provides other capabilities to recover from shelf errors. Our next-generation design will feature ACP functionality over the SAS cables, providing the same value but simplifying cabling.

Media Reliability

At NetApp, we have perfected our hardware designs and hardened Data ONTAP over the last 20 years to provide a high level of resilience to all modes of drive failure—both alone and in combinations. For instance, our dual-parity RAID implementation (RAID-DP) provides protection against double drive failures (and other failure modes) without sacrificing write performance. (For a full explanation of write performance, see the Appendix, "Intelligent Caching of Write Requests," at the end of this paper. For more information about RAID-DP, see the Tech OnTap® article Back to Basics: RAID-DP.)

For HDDs, Maintenance Center performs proactive health monitoring and distinguishes between transient events and real underlying issues using drive diagnostics. HDDs can often be replaced proactively before they fail. Maintenance Center automatically manages disk failures through a systematic failure-verification process. When a disk is identified as a potential failure, Maintenance Center takes over. Data is migrated from the disk onto a spare through Rapid RAID recovery (which copies data directly from the disk before failure occurs) or reconstruction. The process occurs without user intervention. If transient errors can be repaired, the disk is returned to the spares pool.

The functions of Maintenance Center do not apply to SSDs, because SSD failure modes are different. NetApp SSDs are rated for high levels of daily write activity. Because SSDs are rated for a certain number of write cycles, we track important SSD metrics such as:

- Percentage of device life used
- Percentage of spare data blocks consumed
- Threshold limit based on the percentage of spare blocks consumed

If an SSD wear threshold is exceeded, the Data ONTAP event management system sends a message and logs an event to initiate replacement.

Serviceability

The simple servicing design makes it easier to access and replace components. For example, all drives are accessible from the front of the rack and designed for easy access, even when shelves are installed at the top of a rack.

5.2 Protecting Availability of Deployed Systems

FAS systems include a wide variety of capabilities to further protect the availability of the storage systems you deploy.

The NetApp AutoSupport™ tool sends system alerts and weekly logs to your administrators and to NetApp. The tool uses continuously updated risk signatures to help identify and proactively address issues that could affect the availability or efficiency of your storage. Our component-level monitoring and alerting benefit from an understanding of the specific components deployed and work to detect problems before they lead to failures. With advanced service analytics you can extend risk detection beyond the boundaries of storage to include networks, servers, and virtual machines.

Nondisruptive operations with clustered Data ONTAP eliminate the need for planned downtime for maintenance and lifecycle operations. You can perform firmware and software updates so that your storage runs the latest and most reliable code. Because these processes are nondisruptive, storage can be kept more up to date, lessening risk.

Data-in-place upgrades, which are practically unique to NetApp, allow you to replace controllers while reusing shelves and media.

Finally, if a problem arises, our Support Service teams are available 24/7 to troubleshoot and help resolve issues.

5.3 What It Means to You

As a result of all the measures that NetApp takes to protect reliability and availability, you can expect NetApp storage to deliver an exceptionally high level of resilience and uptime. But don't just take our word for it. A white paper from IDC explores enterprise availability requirements in detail, and explains how NetApp delivers greater than 99.999% availability as measured across the NetApp installed base. By following best practices with clustered Data ONTAP, it's possible to achieve availability of 99.9999% or greater.

6 Conclusion

Storage systems face unique requirements for I/O performance, network connectivity, expandability, scalability, and reliability. NetApp continues to make significant efforts to differentiate its FAS storage line according to these attributes. Through tireless attention to design details, our FAS systems are able to deliver tremendous performance, scalability, and availability, with superior manageability, ease of service, and ease of use in a package that combines significant hardware innovation with proven software value and broad industry integration.

Appendix: Intelligent Caching of Write Requests

Caching writes has been used as a means of accelerating write performance since the earliest days of storage. NetApp uses a highly optimized approach to write caching that integrates closely with the NetApp Data ONTAP operating environment to eliminate the need for the huge and expensive write caches seen on some storage arrays while enabling NetApp to achieve exceptional write performance, even with double-parity RAID.

Journaling Write Requests

When any storage system receives a write request, it must commit the data to permanent storage before the request can be confirmed to the host or client. Otherwise, if the storage system experiences a failure with data only in volatile memory, that data will be lost, and underlying file structures could become corrupted.

Storage system vendors commonly use battery-backed, nonvolatile RAM (NVRAM) to cache writes and accelerate write performance while providing permanence, because writing to memory is much faster than writing to disk. NetApp provides nonvolatile memory (NVRAM or NVMEM) in all of its current storage systems, but the NetApp Data ONTAP operating environment uses nonvolatile memory in a much different manner.

Every few seconds, Data ONTAP creates a special Snapshot™ copy called a consistency point, which is a completely consistent image of the on-disk file system. A consistency point remains unchanged even as new blocks are being written to disk because Data ONTAP does not overwrite existing disk blocks. The nonvolatile memory is used as a journal of the write requests that Data ONTAP received since creation of the last consistency point. With this approach, if a failure occurs, Data ONTAP simply reverts to the most recent consistency point and replays the journal of write requests from nonvolatile memory to bring the system up to date and verify that the data and metadata on disk are current.

This is a much different use of nonvolatile memory than that of other storage arrays, which cache write requests at the disk driver layer, and it offers several advantages.

- **Requires less nonvolatile memory.** Processing a write request and caching the resulting disk writes generally take much more space in nonvolatile memory than simply journaling the information required to replay the request.

  Consider a simple 8KB NFS write request. Caching the disk blocks that have to be written to satisfy the request requires 8KB for the data, 8KB for the inode, and—for large files—another 8KB for the

indirect block. Data ONTAP merely has to log the 8KB of data along with about 120 bytes of header information, so it uses half or a third as much space for the same operation.

It's common for other vendors to point out that NetApp storage systems often have far less nonvolatile memory than competing models. However, because of their unique use of nonvolatile memory, NetApp storage systems actually need less nonvolatile memory to do the same job.

- **Decreases the criticality of nonvolatile memory.** When nonvolatile memory is used as a cache of unwritten disk blocks, it becomes an integral part of the disk subsystem. A failure can cause significant data corruption.
- **Improves response times.** Both block-oriented SAN protocols (Fibre Channel protocol, iSCSI, FCoE) and file-oriented NAS storage protocols (CIFS, NFS) require an acknowledgment from the storage system that a write has been completed. To reply to a write request, a storage system without any nonvolatile memory must update its in-memory data structures, allocate disk space for new data, and wait for all modified data to reach disk. A storage system with a nonvolatile memory write cache performs the same steps but copies modified data into nonvolatile memory instead of waiting for disk writes. Data ONTAP can reply to a write request much more quickly because it needs only to update its in-memory data structures and log the request. It does not have to allocate disk space for new data or copy modified data and metadata to nonvolatile memory.
- **Optimizes disk writes.** Journaling all write data immediately and acknowledging the client or host not only improve response times, but also give Data ONTAP more time to schedule and optimize disk writes.

Optimizing Write Performance

Storage systems that cache writes in the disk driver layer necessarily have to accelerate processing in all the intervening layers in order to provide a quick response to host or client, giving them less time to optimize.

No matter how big a write cache is or how it is used, eventually data has to be written to disk. Data ONTAP divides its nonvolatile memory into two separate buffers. When one buffer is full, that triggers disk write activity to flush all the cached writes to disk and create a consistency point. Meanwhile, the second buffer continues to collect incoming writes until it is full, and then the process reverts to the first buffer. This approach to caching writes—in combination with the WAFL® system—is closely integrated with NetApp RAID 4 and RAID-DP and allows NetApp to schedule writes such that disk write performance is optimized for the underlying RAID array. The combination of NetApp nonvolatile memory and WAFL in effect turns a set of random writes into sequential writes.

In order to write new data into a RAID stripe that already contains data (and parity), you have to read the parity block and calculate a new parity value for the stripe and then write the data block plus the new parity block. That's a significant amount of extra work required for each block to be written.

NetApp reduces this penalty by buffering protected writes in memory and then writing full RAID stripes plus parity whenever possible. This makes it unnecessary to read parity data before writing and only requires a single parity calculation for a full stripe of data blocks. WAFL does not overwrite existing blocks when they are modified, and it can write data and metadata to any location. In other data layouts, modified data blocks are usually overwritten, and metadata is usually required to be at fixed locations.

This approach offers much better write performance, even for double-parity RAID (RAID 6). Unlike other RAID 6 implementations, NetApp RAID-DP performs so well that it is the default option for NetApp storage systems and has been used regularly for performance benchmark submissions since its release. Tests show that random write performance declines only 2% versus the NetApp RAID 4 implementation. By comparison, another major storage vendor's RAID 6 random write performance decreases by 33% relative to RAID 5 on the same system. (RAID 4 and RAID 5 are both single-parity RAID implementations. RAID 4 uses a designated parity disk. RAID 5 distributes parity information across all disks in a RAID group.)

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NetApp provides no representations or warranties regarding the accuracy, reliability, or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

www.netapp.com