FlashNAS ZFS

Enterprise NAS Functionality at a Fraction of the Price

May 7, 2014

Executive Summary

Today's dominant Network-Attached Storage (NAS) vendors are well known for their products' scalability, availability, and 24x7 support. However, they also tend to continually add features and functions—even as the number of customers using them declines. Meanwhile, most other vendors have chosen not to participate in this arms race, staying instead at the opposite end of the storage-product functionality spectrum.

As a result, too many customers are forced to choose between two extremes: "high-end capability for a high-end price" *vs.* "low-end capability for a low-end price." That might be acceptable if it weren't for the 60-70% of storage capacity in most data centers consumed by redundant copies of data.

Winchester Systems offers FlashNAS ZFS to the underserved "middle ground" between these two extremes: organizations that depend upon their storage infrastructures to survive, but don't want to pay premium prices for capabilities they don't need. That means providing the right features for less—without requiring additional (and expensive) licenses to enable valuable functions—as well as providing US-based 24x7 call center support.

The rest of this white paper describes "must have" NAS storage requirements of such organizations, and how Winchester Systems FlashNAS ZFS addresses them—while at the same time keeping customers' capital and operational costs to a minimum.[1]

[1] *For a discussion of Enterprise requirements for block-based storage, see the Winchester Systems white paper "FlashDisk FX: Enterprise DAS and SAN Storage Functionality at a Fraction of the Price."*

Keeping “Enterprise” Real

Many IT vendors use terms like “Enterprise” to describe their NAS products, listing attributes and functions they say no Enterprise organization can do without. Too many times, a better translation would be “stuff our products offer that others don’t.” Or, for the more cynical among us: “high priced.”

The problem is, this game has been played for too long. In order to continue reaping premium prices, top-tier storage-equipment makers have continued adding more and more features and functions. Beyond a certain point, however, the value delivered by those premium functions starts to decline. Then there’s the low-end extreme of the storage-product spectrum, which can effectively be described as “Slightly better than the stuff you’d find in servers, desktops and laptops.”

Here’s a radical notion: what if someone were to define “Enterprise Storage” from the perspective of those whose businesses’ daily survival depend on their IT infrastructure? What would be on their “must-have” requirements list? The must-haves we’ve consistently heard from customers and partners are:

Scalability. Regardless whether it’s “scale up” or “scale out,” an Enterprise Storage platform must be able to adroitly handle very large amounts of data. Yes, “very large” is a moving target. At this writing, it’s in the Petabyte range. Equally important, performance must scale as well, handling workloads ranging from real-time OLTP databases to write-once-read-many archives.

Availability. An assurance that data is accessible whenever it’s needed. This requires reliable storage that operates continuously despite hardware, software, and sometimes even human-created failures. The level of data protection required depends on the value of the data, which can vary over time. The more a storage system can take corrective actions on its own (“self-healing”), the better it can guarantee availability of the data. Mirroring or replication of data across storage systems—both locally and over geographic distance—is also a must.

Confidentiality. A major aspect of information security: ensuring only authorized personnel and applications can access specific data. This usually involves the “Three A’s” of security: Authentication, Authorization and Auditing and, in most cases, must integrate seamlessly with incumbent identity-management services such as Active Directory (used mainly by Windows systems) and NIS+ (used mainly by UNIX systems).

Integrity. An assurance that the data retrieved is identical to what was originally stored. Threats to data integrity such as tampering and vandalism, typically considered aspects of information security, readily come to mind for most. And, of course, Enterprise storage must protect against any detected data errors. But there are more insidious threats, such as undetected hardware data errors—which RAID technology alone *cannot* address—leading to data loss or corruption.[2]

Snapshots. Preserving a self-consistent copy of an entire file system or disk volume at a specific point in time. This capability is used for multiple purposes: eliminating the “backup window” for traditional backup software; recovering from failed software installations, upgrades or patches; recovering accidentally-overwritten files, and much more. The more space-efficient the snapshots, without sacrificing performance, the better.

Pooled Storage. The ability to virtualize disk drives, RAID sets, and so on into storage pools that can be dynamically expanded and carved up into NAS shares or SAN volumes as needed. This capability separates storage services provided over the network from the physical devices that comprise them, enabling IT admins to add or change storage hardware without disrupting those storage services—and avoid disrupting applications that depend on the storage.

Immutability. Also referred to as “Write Once, Read Many” or its acronym, “WORM,” immutable storage provides an unalterable data archive. Demand for WORM is largely driven by regulatory requirements to maintain records of various

[2] *This problem has been discussed at length in numerous studies over the past decade or so.*

types for months, years, and in some cases lifetimes. Once dominated by optical storage media, the need for immutable storage that's accessed as easily—and as quickly—as regular read/write storage has been driven by regulatory, legal discovery, and other requirements for timely archive search and access.

Support. No-hassle, no-finger-pointing, 24x7 product support. Organizations that depend upon their storage infrastructure also depend on their vendors to solve any problem that arises, and provide reliable guidance on storage-architecture best practices.

The Under-Served Middle

A chasm has opened up in the storage industry between "Enterprise Vendors" (such as EMC, HP, IBM and NetApp), and "Everyone Else." And it's growing. The largest storage vendors have answers for most, if not all, of the Enterprise "must-have" needs just outlined. Most are well known for their scalability and availability. Several offer immutable storage, though usually as an add-on product at significant cost. 24x7 support can be taken for granted (but don't expect personalized service unless you're a *very* large customer). And even consumer products today can integrate with Active Directory and NIS+ for NAS-client authentication and access control.

However, the "Enterprise Vendor" group has continued adding features that are highly valuable to some customers—but a declining percentage actually use them. Most of the "Everyone Else" group decided long ago not to attempt competing directly with the Big Guys in this feature/function arms race. As a result, too many customers end up forced to choose between two extremes: "high-end capability for a high-end price" *vs.* "low-end capability for a low-end price."

For most companies that need Enterprise capabilities, that might be acceptable—if it weren't for the fact that some 60-70% of their storage capacity is being consumed by redundant copies of data. Many have started asking, "Do I really need to pay the same price for storing my copy-data?"

The easy answer, given by many, is "use SATA disks." After all, they're cheaper than SAS or Fibre Channel drives.

The problem is, those inexpensive SATA drives are still housed in "Enterprise Vendor" storage enclosures. And they need more protection (using RAID-6, for example) because of their lower reliability. Customers using SATA do indeed pay a lower price. But let's not kid ourselves. They're still paying a significant premium for that name on the front bezel.

What's the alternative? Consumer-grade storage? Good luck getting the scalable performance, availability, integrity-assurance, or support needed by most organizations dependent on their storage.

A fair number of IT shops and systems integrators have resorted to "rolling their own," cobbling together storage platforms using industry-standard servers, installing software products themselves—and taking on support for those platforms themselves. Many such projects start out appearing less expensive, but end up costing much more because of continued (and sometimes rising) labor expenditures.

At Winchester Systems, we consider this growing gulf between the Big Guys and everyone else to be unacceptable. So we created FlashNAS ZFS with a simple goal in mind: to serve the middle ground in between these two extremes. FlashNAS ZFS provides Enterprise features real customers and partners have told us matter most, at prices far below those of "Enterprise" vendors.

Balancing Capability, Simplicity, and Price

To provide Enterprise data protection and reliability at down-to-earth prices, Winchester Systems engineers began with a hardware platform purpose-built for continuous availability and long-term durability. That meant using controllers, power supplies and fans that are modular, redundant and hot swappable—technology long used successfully in our FlashDisk block-storage arrays. For the software platform, Winchester System engineers chose the revolutionary ZFS file system for its end-to-end error detection and correction, self-healing, and SSD optimization features for unsurpassed protection of data integrity throughout the storage system.

Building on this foundation, FlashNAS ZFS designers added Enterprise functions such as storage-pool mirroring, remote replication, and data archiving and retention compliance. All of these capabilities are combined into a single integrated system with an easy-to-use, web based management interface—with no hidden or “extra” costs to enable specific features. All functionality is available for the base system price.

Why did Winchester Systems use this approach? How is it better than those used by competitors? We’ll start by explaining why we chose ZFS, and then how we added Enterprise-required functions while still keeping things simple—and at significantly lower prices.

Why ZFS?

Data Integrity Protection and Self-Healing

The majority of modern file systems and volume managers, including those embedded within most commercial NAS products, assume that no component in a storage system is safe from failure. The better ones are designed to handle the loss of one or more power supplies, fans, disks, I/O paths, and even controllers without disrupting access or losing data.

But what happens when a device doesn’t fail gracefully? What if instead it *misbehaves*? Things can go wrong within a device even though it appears to be working perfectly. That includes *silent data corruption*, an IT-staffer’s—indeed any businessperson’s—worst nightmare.

A common technique for handling this within storage arrays is to format disk drives using larger block sizes, and append checksum data to the user data stored within each block. The problem is, most arrays rely on the *drive* to perform that check and report the results. This exposes the system to two vulnerabilities: (1) a drive that reports success but returns bad data, and (2) data that’s corrupted during the transfer from disk to memory.

NetApp’s approach in its Write Anywhere File Layout (WAFL) file system is stronger, but similar. WAFL groups long-formatted disk sectors into multi-sector blocks with a combined 64-bit checksum, and verifies that checksum *after* the data is read from disk. That’s better than relying solely on disk firmware, but it still only validates that a delivered data block is consistent. It doesn’t guarantee it’s the *correct* data block.

The ZFS file system, created by Sun Microsystems (now Oracle), is designed with a focus on end-to-end data integrity. That requires a fundamental distrust of all components beneath it in the stack: drives, interconnects, busses, and so on. So it verifies each block against an *independent* checksum *after* it’s been transferred to memory. The checksum for each block of data is stored in the pointer to that block, not with the block itself. That pointer is, in turn, checksummed and the resulting value stored in its pointer, and so on. This includes not just file data, but also the entire hierarchy of volume and file metadata throughout the storage being managed.[3] This end-to-end data verification is unique to ZFS. NAS products based on older file system technology cannot match this level of data-integrity protection.

Merely detecting corruption isn’t enough, however. ZFS is also designed to recover the original data whenever possible—in other words, provide self-healing data. When a bad block is detected (its checksum verification failed), ZFS automatically fetches the correct data from a redundant copy and repairs the bad block by replacing its contents with the correct data. Instead of waiting for application I/O to uncover a problem, FlashNAS ZFS also performs scheduled media scanning (also called “disk scrubbing”) to proactively uncover and repair any silently corrupted blocks.

Online-Expandable Storage Pools

Instead of disk devices, ZFS always allocates space for its file systems and iSCSI targets from “Storage Pools.” Each Pool is composed of virtual devices, each being a set of disks with ZFS RAID 0, 1, 5, or 6 protection. Expanding a pool involves simply adding one or

[3] *A detailed discussion of ZFS integrity protection mechanisms can be found in former Sun Fellow Jeff Bonwick’s blog at https://blogs.oracle.com/bonwick/entry/zfs_end_to_end_data.*

more RAID sets; from then on ZFS stripes data across all RAID sets within the pool (the resulting combinations sometimes called RAID 10, 50 and 60).

Unlimited Snapshots

ZFS snapshots are extremely efficient in both disk space and storage-processor usage because of the way the file system handles write operations. All ZFS disk transactions use an “allocate-on-write” process. When an application updates a data block, the file system allocates a new block from the storage pool and writes the new contents in the newly allocated block, updating related metadata pointers and checksums in a similar manner along the way.

Because they’re not overwritten, old block contents and their metadata pointers can be easily retained. As a result, snapshots of a file share or volume can be created and maintained with very little processing overhead. As data within the active share/volume changes, each snapshot preserves only replaced data and metadata blocks, while also preventing them from being reused until that snap is deleted.

That means a snapshot takes up no additional space until blocks within the file share or iSCSI volume are changed. Better still, the number of snapshots possible is limited only by available storage-pool space. FlashNAS ZFS admins can also define automatic snapshot creation schedules and snapshot-deletion rules for each share or iSCSI volume based on age or desired maximum number of snaps.

Several competing snapshot implementations require specifying “reserve” capacity somewhere to hold changed-block contents. Their theory: using a separate reserve ensures space is available in active file shares. An example many use to illustrate this point is deleting files, which does create more available space in the accessed file system—while consuming it in the reserve. However, this approach imposes an arbitrary limit on both the size and number of snapshots for each volume or file share. And they create potentially unavailable—and thus wasted—storage capacity when underutilized. Worse, when enough data changes to exhaust that reserve, the storage system must either destroy the snapshot or start using space in the active file system.

ZFS eliminates all of this snapshot-reserve sizing guesswork.

RAM and SSD Based Performance Acceleration

To increase read performance, ZFS caches recently read data blocks in RAM using an Adaptive Replacement Cache (ARC) algorithm, and uses SSDs as a second-level ARC cache (also called L2ARC). To accelerate write operations, SSDs are used to hold a copy of each write transaction in the ZFS Intent Log (ZIL) until the transaction contents are safely committed to disk media. The ZIL can also be mirrored to further protect data integrity.

In-line Compression

For space savings, and better storage efficiency, ZFS offers data compression on designated file shares and volumes—with faster performance for files updated in place than is possible with most file systems. If an updated data block after compression is smaller than the original compressed block, FlashNAS takes advantage of the ZFS file-system variable block size and “allocate-on-write” algorithm by simply allocating and writing a smaller-sized block. For files that are updated in place, this provides superior performance because updated blocks can be compressed without rereading and rewriting the rest of the file’s content.

Although ZFS has gained significant respect and adoption on server platforms, its use within commercial storage has been largely limited to high-end products from Oracle and software-only products based on ZFS bundled with an open-software NAS stack (or industry-standard servers with such bundles pre-installed).

Keeping ZFS Storage Simple

The benefits of ZFS are compelling, but a good file system alone is not enough to comprise a reliable, easy to use, Enterprise Storage product. A focus on simplicity also required. Simplicity from a *customer* perspective. Winchester Systems engineers sought to combine hardware and software technologies that address "must have" needs of Enterprise Storage, while at the same time keeping both hardware and software—as experienced by the customer—as simple as possible.

It's been clear to Winchester Systems designers that ZFS provides top tier data integrity. They're not alone, of course. A number of storage vendors offer ZFS-based products, using approaches that also focus on simplicity—from an *engineering* perspective.

Most ZFS-based storage products are built using general-purpose Intel servers running ZFS on Linux or OpenSolaris. To offer High Availability most add a second server, external storage, and a general-purpose cluster software product such as RSF-1 that's designed to protect multiple applications in a wide variety of cluster configurations. Because of its general-purpose nature, such cluster software offers numerous options and settings, requiring cluster admins to make numerous configuration decisions. Making management of this software and hardware bundle simpler for customers is no easy task. Especially when customers are also cabling and configuring the hardware being managed. So it's no surprise most end up passing a lot of that complexity on to their customers.

Winchester Systems engineers took the opposite approach. Instead of bundling commodity hardware and off-the-shelf software, they chose to *tightly integrate* a product using plug-in hardware modules, passive backplanes and purpose-built software. And by purpose-built software, we mean software designed specifically for our high-availability NAS product—and only that product.

It's an ambitious approach. But Winchester Systems engineers had an ambitious goal: to make a redundant-controller NAS product that's as easy to install, operate and maintain as a single-controller system. In other words, the engineers were fine making things harder for themselves if it meant keeping them simple for customers.

Let's compare the results:

A "simple" software-based ZFS cluster requires two servers running an OS + ZFS file system stack, one or more JBOD disk enclosures directly connected to each server via SAS cables, network and/or serial "heart-beat" interconnects between the two servers, installed clustering software, installed cluster-aware iSCSI-target software, an IP address for each NAS folder or iSCSI LUN to be served, an IP address for each ZFS volume created internally (regardless whether they're exposed to storage clients), and DNS entries for each "service name" in the cluster.

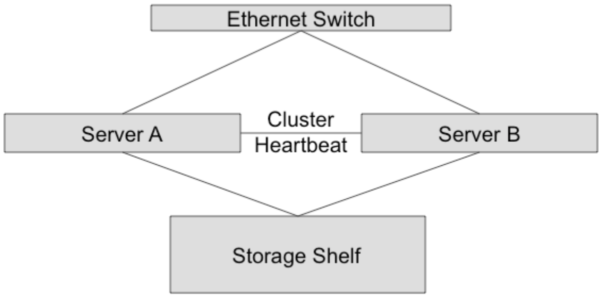

Figure 1: Dual-Controller Software Based ZFS NAS Cluster

Why all of the IP addresses and DNS names? They're required by the clustering software to handle failover. The fun for IT admins doesn't stop there. Each volume created by an admin must also have designated "heartbeat drives" (two are recommended) that are separate from the volume, and a "failover host" (translation: storage controller). Wrapping all of this within a graphical user interface doesn't help matters all that much, either.

Figure 2: Dual-Controller FlashNAS ZFS, with 16 disks

FlashNAS ZFS, by contrast, deeply integrates an entire ZFS cluster within a single 2U or 3U hardware enclosure—including up to 16 disk drives. There are no external cables. No software to install or configure. No cluster management. In fact, FlashNAS ZFS storage admins can't tell there's a cluster running at all.

Winchester Systems engineers went out of their way to make the clustering within FlashNAS ZFS invisible to IT admins. To illustrate this, let's compare how

FlashNAS ZFS handles major cluster events compared to software-based competitors:

Automatic Failover. If a FlashNAS ZFS storage controller should stop working, failover is automatic. No operator involvement is required—except for replacing the failed hardware module, of course. Most software ZFS clusters handle this failover case pretty well.

Automatic Fail-Back. Restoring storage services to their original state, also known as "fail-back" in clustering terminology, is completely automatic in FlashNAS ZFS. When a replacement module is inserted and powers up, FlashNAS ZFS puts the replacement module into service completely automatically. No operator intervention is required. In other words, replacing a FlashNAS ZFS controller module can't be any simpler: Just slide it in. The rest is automatic. Most software-based ZFS clusters cannot do this at all. Instead, they require manually "failing" each individual service back to its original location.

Assured Data Integrity. Last, but not least, the ZFS file system ensures data integrity throughout failover and fail-back, checking the ZFS Intent Log (ZIL) while mounting a failed unit's file systems and properly completing any in-flight I/O for disk writes that had been acknowledged to a client.

This tightly integrated approach has resulted in a ZFS based, redundant-controller NAS system that's quite literally as easy to setup, configure and operate as a single-controller system.

Keeping WORM Simple

The same laser-like focus on simplicity is evident in the way Winchester Systems added WORM (Write Once, Read Many) capability to FlashNAS ZFS. WORM functionality can easily be applied to any shared folder, protecting its contents from being edited, altered, overwritten, deleted or corrupted.

Many companies now face regulations such as SEC Rule 17a-4 for securities firms, HIPPA for healthcare organizations, gaming-commission rules and so on; requiring them to retain data for a period of time while still providing immediate accessibility. Litigation and other e-Discovery events can add significant data-retention needs with little or no notice. Furthermore, some firms have decided, independent of government regulations, they need to protect vital information from tampering by *anyone*—disgruntled employees, hackers, etc.

FlashNAS ZFS helps organizations meet these requirements by using WORM folders. Authorized users can write files to, and read files from, a WORM folder shared over the network. But they're *not* allowed to modify or overwrite the original data; the only way to store a modified version of a file is by creating a copy with a different name (*e.g.*, using "Save as…" in an application).

WORM can be enabled for any shared folder at any time. An IT administrator simply clicks the "WORM" checkbox and selects a retention-period end date (or "forever") in the folder settings. An admin can later extend retention periods just as easily when regulations or corporate governance requirements change. They cannot, however, shorten the period or turn WORM off—ensuring data protection regardless of any employee's access privileges.

Keeping Availability Simple

Rounding out the capabilities of FlashNAS ZFS are functions designed to keep data available through numerous scenarios:

Full-System Upgrade or Replacement

When replacing or upgrading a FlashNAS system, an administrator can simply remove disk drives from the old system, and insert them into another FlashNAS ZFS system where they will be automatically recognized. Admins don't need to worry about getting the exact placement right, and can even insert drives that comprise RAID sets from different FlashNAS systems. Called "disk roaming," this feature stores configuration metadata on the drives themselves, enabling relevant pools and other settings to be restored and made fully operational automatically.

Localized Failures (Rack, Power, Cooling, etc.)

FlashNAS ZFS offers Pool Mirroring to synchronize copies of a storage pool between two FlashNAS ZFS systems. All file share and iSCSI volumes within a mirrored Storage Pool are copied, ensuring data availability even in the event of a complete system failure. Online business applications can continue to be served without major interruptions. Pool mirroring can be real-time (synchronous) or scheduled (asynchronous).

Site-wide Failures

Individual file folders or iSCSI volumes can be replicated to another FlashNAS ZFS system at a remote site to ensure data availability even in the event of a site-wide disaster. The FlashNAS ZFS system uses block bitmaps to record which data blocks have changed. Only the data blocks that have changed since the last synchronization are transmitted, even after an extended communications-link failure.

Transmitted data can optionally be compressed using LZJB, a lossless compression algorithm, and built-in encryption is available to protect the data in-flight. Replication scheduling can be highly customized to suit the needs of each organization or application.

Backup Software Integration

For seamless integration with incumbent enterprise backup regimes, FlashNAS ZFS supports industry-standard Network Data Management Protocol (NDMP) compliant data backup and migration software. The NDMP protocol enables data transport between devices, such as networked storage and tape backup systems.

Conclusion

By staying focused on those Enterprise features deemed most valuable by real customers, shunning those that add more operational complexity than realized value, and keeping the user experience as simple as possible, Winchester Systems has produced a low cost network-attached storage system—FlashNAS ZFS—with the kind of deeply integrated scalability, availability, data integrity, and performance that's been limited to high-end storage products for far too long.

FlashNAS ZFS combines controllers, power supplies and fans that are modular, redundant and hot-swappable with the ZFS file system and its end-to-end error detection and correction, self-healing, and advanced SSD optimization features. Rounding out its Enterprise capabilities is purpose-built software to enable storage-pool mirroring, remote replication, data archiving and retention compliance (WORM), anti-virus scanning, and NDMP backup and migration compatibility.

All of these capabilities are combined into a single integrated system with an easy-to-use, web based management interface. And there are no hidden, or "extra" costs to enable specific features. All functionality is available for the base system price. The standard warranty includes 24x7 call center support and next day on-site service. Four hour on-site response service is also an option.

Instead of being forced to choose between two extremes, "high-end capability for a high-end price" *vs.* "low-end capability for a low-end price," customers now have a third option: "Must-have capability for an aggressive price."

FlashNAS ZFS offers a cost-effective choice for primary storage, online-backup and archive tiers, or any situation where must-have functionality is exactly what's needed. In other words, the right storage at the right price.