

PETER BAER GALVIN

Pete's all things Sun: the "problem" with NAS



Peter Baer Galvin is the chief technologist for Corporate Technologies, a premier systems integrator and VAR (www.cptech.com). Before that, Peter was the systems manager for Brown University's Computer Science Department. He has written articles and columns for many publications and is co-author of the Operating Systems Concepts and Applied Operating Systems Concepts textbooks. As a consultant and trainer, Peter teaches tutorials and gives talks on security and system administration worldwide. Peter blogs at <http://www.galvin.info> and twitters as "PeterGalvin."

pbg@cptech.com

COMPUTER STORAGE HAS EVOLVED

from Directly Attached (DAS) to Storage Area Networks (SAN). Along the way, Sun in 1984 invented NFS, and Network Area Storage (NAS) was born. Since then other NAS protocols have been added, most notably the Windows-based Server Message Block (SMB), aka CIFS. But throughout the history of storage, NAS has been regarded as poorly performing and unreliable compared to SAN and DAS. Certainly NetApp's creation of a NAS "appliance" helped move NAS from being a science project to a mainstream production solution, but in my opinion NAS is still underappreciated and underdeployed. Perhaps in light of the new generation of NAS appliances, that should change.

At a more philosophical level, it's worth asking "what is SAN" and "what is NAS." Fundamentally, they are storage arrays that make disk space available via varying protocols over varying interconnect media. For the most part, both technologies are available with Fibre Channel (FC), SATA, and SAS disks. Both have disks of varying speeds, capacities, and performance. Traditionally, SANs have been FC connected and NAS appliances connected via Ethernet, but many current products provide both interconnects—block transactions occur via FC or iSCSI and file transactions over Ethernet. A proof point of this merger of NAS and SAN is the FCOE protocol which places Fibre Channel frames over Ethernet networks. Perhaps the most straightforward definition is that "SAN" is block-based storage and "NAS" is file storage, and that a given datacenter should choose which to use for any given application or function. After those decisions are made, it is easier to determine the best products to implement the resulting storage architecture. Now let's consider the problem with NAS as well as the solutions it can provide.

The "Problem"

Over the years I've seen many, many computing infrastructures. Back in the "old days" (say, the 1980s), we had servers and SANs for production, and NAS was pushed to the side. It was typically used for home directories and the storage of utility programs, if at all. In those cases, NAS storage was mounted to all servers as well as all workstations.

That helped NAS gain a reputation for unreliability—probably because any failure caused everyone to notice it, and failures were difficult to recover from (with hard mounts never timing out, for example, taking down all computing until the NAS server could be fixed). Also, many situations called for “cross mounts,” where servers would mount each other’s directories via NFS. If one server then failed, all servers would eventually end up hanging until the failed one recovered. NFS also had quirks like “stale file handles” that left a bad taste in the mouth.

So failures of NFS servers were quite painful to the computing infrastructure. Why did NAS servers fail as often as they did? Well, they were non-clustered, while their SAN brethren typically had more redundant components and automatic recovery from problems. Originally, a “NAS server” was just a general-purpose Sun server running NFS. SAN originally and usually still is a purpose-built storage array. Also, they were and still are network-connected. Back in the day, there was typically one network connection to each workstation (and frequently between servers as well). That one link was used for NAS and non-NAS network traffic. Even if there was a separate network carved out for storage communication between the servers and NAS, it was rarely redundant. Multiple use and single points of failure meant NAS was more prone to failure than SAN. Thus the lingering impression that SAN is more reliable than NAS.

There is also an impression that SAN has better performance than NAS. First, consider the communications protocols. For SAN, the Fibre Channel medium carries SCSI protocols between servers and storage arrays. SCSI is (by definition) optimized for storage operations. TCP/IP is a general protocol used for everything from sending one character at a time (telnet, for example) to bulk file transfers (ftp and NAS). In addition, TCP/IP runs over a shared medium, so it has to deal with collision detection and recovery. The TCP/IP communications are therefore more chatty and less efficient than the equivalent SCSI commands (where there are equivalents). Also, the caching of NAS I/O is less effective than SAN, due to NAS storage being shareable. As one example, consider metadata caching. On a SAN, once a LUN is mounted, the mounting server “owns” that LUN. Over the course of I/Os it can cache all the data and metadata it needs, infinitely. With NAS, because other systems might be accessing the same directories and files, NAS clients must recheck with the NAS server periodically to see if any metadata has changed. Those timings can be modified via mount options but are typically measured in seconds not minutes. If the NAS client detects that its cached data is invalid, the clients have to throw out the cached data and metadata and reload it in the worst case (depending on file open modes, for example). Thus the overhead of NAS operations is higher than SAN operations.

All of this adds up to NAS performance challenges. With NAS, a single user can seemingly cause more of a performance hit than on SAN. For example, again back in the ‘80s, we would debug NAS performance problems by watching the network traffic and finding a user flooding the networking with NAS requests. Frequently, the problem would be a single user running a UNIX “find” command across some directory structures mounted via NFS. A single user running a single command could bring the NAS server to its knees. The equivalent operation across a SAN would be less onerous, most likely due to the large caches included in most SANS.

That Was Then, This Is Now

NAS is not just for sharing anymore, and is past most of its adolescent problems. In fact NAS is now quite mature, fast, and reliable. But NAS, in

many datacenters and in many instances, is still relegated to tasks of lesser importance. Tier-1 use of NAS (for non-stop production) seems to be rare, but shouldn't be. Consider the latest generations of two great NAS products: NetApp's FAS series of "filers," or NAS appliances, and the Oracle/Sun Storage 7000 line. Both product lines scale from small to very, very large capacities. And both scale up to very high performance, although that is harder to prove. The SPECsfs2008 benchmark (<http://www.spec.org/sfs2008/>) is one source of performance information about NAS servers, and there are many posted results, but Sun is not one of the contributors. Sun (rightly in my opinion) considers it to be a severely flawed benchmark, but it's about the only thing we've got that shows comparative NAS performance. On-site testing of real environments is always the best indicator of performance but usually difficult to do and not commonly done. In testing in my company's lab my colleague Sean Daly drove VMware to push 990.48 MB/s of throughput and 149,227 IOPS (I/O operations per second) from a NetApp filer. That is certainly a lot of performance. And both NetApp and Sun NAS servers can be configured as high-availability clusters, with fast failover in the case of component failures (with the NetApp failing over faster than the Sun 7000 in our testing).

Why then is NAS not taking the world by storm? In some ways it is, as indicated by the rapid growth rates of NetApp and Sun's storage group. But there are certainly many cases when NAS could and should be used but where DAS or SAN is used instead. The reasons for that are as varied as computing infrastructures and the managers that run them. In many cases it's a simple case of familiarity. Storage managers have more experience with SAN than NAS, and go with what they know. In other cases it is for simplicity. Running one kind of storage, from one vendor, is simpler than running two kinds of storage solutions from one or two vendors (the existing SAN vendor or a new NAS vendor). And in some cases the lack of NAS use is based on previous, painful experiences, or a lack of understanding of the state of NAS servers and their features. It is this last group that I'm hoping to address with this column.

The Case for NAS

If SAN storage arrays also have high reliability and high performance (for the most part), then why not just run SAN instead of NAS? Consider some of the more potent features of good NAS storage. Also consider that even though some of these features are available with SAN storage, they are frequently more expensive, require extra devices, or are much more limited than their NAS brethren.

- Snapshots—read-only point-in-time, fast, low-space-use file system copies—and clones, read-write versions of the same, are “magical” in their function and utility. When I first tested ZFS, for example, I was taking snapshots every minute of every day of every month for a year. I had thousands of snapshots, each representing the state of the file system at that minute. The power to undo and redo any file system changes is extreme and not used enough.
- Diskless booting allows servers to run as “field replaceable units,” running interchangeably except for their knowledge of which remote boot disk image they are associated with. For this model to work, the servers must be configured similarly, and must all have access to all external storage units. That is certainly easier with NAS storage than with SAN storage. But consider combining diskless booting with cloning. A datacenter manager could create a “golden image” of a server operating system, configured

exactly as needed, and then clone it hundreds of times to make hundreds of identical boot disks (for hundreds of servers). When a change is needed, a new golden image can be created, cloned, and the servers rebooted to use the new versions. Many versions of golden images (and boot disks) can be kept for revision control, testing, disaster recovery, and so on. This functionality is similar to that touted by virtualization vendors, but done at the disk level rather than the virtual disk level. Both have their place in the datacenter, but with diskless booting no virtualization (or virtualization license) is needed. To improve performance, you could consider using an internal disk for swap space, keeping swap traffic off of the network and the NAS array.

- Replication of snapshots allows disk-to-disk backups as well as easy disaster recovery site synchronization. When combined with diskless booting, a single NAS server replicating to a similar server in a remote datacenter “solves” the data part of disaster recovery. Set up a farm of servers at the remote site, and the compute portion is solved as well. That remote site can also be your disk-to-disk backup site, with production replicating to disaster recovery. Some environments are using such a scenario instead of backing up the data to disk. While some others only put tape drives in the remote site, and perform disk-to-disk-to-tape backup in that manner.
- Sharing is probably the most compelling feature of NAS over SAN. Home directories of users can be shared to all servers that the users log in to, giving them their environment across all servers. Less common but equally useful is the storage of applications on NAS. Those applications can be installed once, and maintained in one location (with snapshots or other methods for revision control), and all servers can have access to the same versions of all applications. Also becoming more popular is the storage of application data on NAS. For example, Oracle happily recommends using NAS to store Oracle Database data. Even Oracle’s RAC clustering can use NAS storage for the data, and is actually much easier to set up that way than using SAN storage. As always, when in doubt check with your application vendors to see what they support. You might be surprised to find out that NFS is on the list.
- Ease of management is something rarely said about traditional storage arrays, although some newer arrays (such as 3PAR and IBM’s XIV) are great improvements over their older counterparts. Tasks that take many steps and lots of time on a traditional SAN can take minutes or even seconds on NAS. Consider the pain of expanding the amount of storage available to a host on both a SAN and a NAS. Also consider standard, complicated tasks performed by your storage administrators. Compare the effort and risk (the more commands, the more likely a mistake) to performing the same task on NAS. If you don’t have NAS on-site, consider a demonstration by a NAS vendor to show you the differences in administration.
- Deduplication is all the rage, and for valid reasons. It can reduce the amount of storage used by a given set of data, and, depending on the implementation method, it can maximize the use of caches by only storing the deduplicated block once in the cache. Likewise, it can decrease the amount of data replicated between datacenters by only sending original blocks, not duplicates. And it is especially useful in environments such as virtualization where many copies of the same blocks are stored (operating systems and binaries). SANs have a difficult time including deduplication, and in many cases an external device is needed. Both NetApp and Sun NAS devices include free deduplication.
- Flexibility is the watchword with NAS, as most major NAS solutions (including the two being discussed here) can be used as SAN as well as NAS. These products provide both iSCSI and Fibre Channel connectivity. iSCSI

is useful for connections where NAS is not supported (e.g., the Microsoft Exchange datastore), and where the complexity and expense of FC cables, switches, and HBAs is not wanted. But where maximum performance and reliability via SAN storage from a NAS appliance is desired, the ultimate step of adding an FC SAN attachment between your hosts and your NAS appliance is available.

- Performance analysis and tuning are inarguably easier with NAS devices. Seeing what is happening at a file level is much more revealing than at the block level. There have been many instances in my debugging efforts where the SAN was a black box that we worked around, rather than a source of information useful in determining the cause of the problem. While both NetApp and Sun provide useful tools in their appliances, Sun has done an astonishing job of integrating DTrace into their device, providing never before available details (e.g., heat maps that depict the time each I/O request took to be satisfied).
- Cost can vary dramatically between SAN and NAS, and between vendors and configurations. Certainly a blanket statement such as “NAS is cheaper than SAN” cannot be made. But pricing out a NAS solution, in cases where NAS is a valid fit, is a worthwhile exercise. If possible consider the total cost of ownership over a period of time that suits your site’s replacement schedule, say three or five years. Add into that the costs of software licenses, including host-side licensing (such as backup software, EMC PowerPath, and Veritas File System and Volume Manager). Frequently, soft costs are not considered, or are considered unimportant, but if possible think about staff time as well.

Making NAS Work

NAS is not a panacea for all things ailing your datacenter. Although NAS performance can be very good, certain workloads can perform worse than very good SANs. Consider the total throughput of your solution, especially as limited by per-spindle IOPS abilities. Fewer large disks in SAN will provide fewer I/Os than a larger number of smaller disks in a NAS, for example. Make sure the I/O being provided by the device is sufficient for your needs.

Also, badly implemented technology will not perform as well or as reliably as well implemented ones. Both SANs and NASes need careful deployment planning, disk layout, and feature utilization. Consider especially mount options and block alignment during implementation. One other likely cause of admins thinking of NAS is less reliable than SAN is the interconnect technology. FC switches and cables can only be used for storage connectivity. Ethernet can be used for host and storage connectivity. But those two tasks should not be shared. If the server to storage connection in NAS is treated as nicely as FC is, then NAS will run very well indeed. Certainly, for maximum reliability (and performance) dedicate a VLAN, and if possible two LANS (for redundancy), to NAS I/O. Do not use those networks for other purposes (even backups should be kept separate). Such segregation can go far toward an optimal NAS experience.

The choice of NAS solutions is of course important. I have already mentioned NetApp FAS filers and Sun’s Storage 7000. Note that there is currently a patent lawsuit between the two companies. I don’t believe that such legal actions should affect your decision-making process, as lawsuits rarely impinge on end-users. Other commercial NAS solutions exist, and many are fine products. However, some are “one protocol ponies.” Why settle for a device that can only provide data across one protocol when many-protocol appliances are available? The trade-offs of simplicity versus utility (and cost)

need to be considered, but rarely have managers been unhappy that they had too many protocols available to them.

Finally, rather than purchasing an appliance, many sites “roll their own” NAS services by using standard servers and SAN storage. I think many of those sites would be better off with an appliance, given the performance, reliability, feature sets, and ease of administration of appliances. Frequently, once an appliance is deployed in a datacenter, the datacenter managers find more and more uses for it and move datasets from the existing SAN to the new NAS. A roll-your-own approach might limit performance, reliability, and utility, artificially limiting the use of NAS in an environment.

Conclusion

Many SAN storage devices have many of the NAS features discussed here, but few have all of them. The combination of all of these features makes NAS a very useful, dare I say “compelling,” component of datacenter strategies. NAS is flexible, efficient, and can perform well and reliably. It can also be much easier than SAN to implement and administer. One of our clients recently replaced an EMC DMX 8000 (a high-end SAN) with a cluster of two NetApp FAS arrays. They are very pleased with the trade, citing improved convenience and good performance and reliability. They also note that their purchase of NAS, including three years of maintenance, cost less than renewing one year of maintenance on the DMX 8000. I suggest you consider the benefits your data center could enjoy with an increased use of NAS in production.

Special thanks to Adam Leventhal, Sean Daly, Jesse St. Laurent, and Paul Deluca for contributing to this column.