Panasas White Paper - Scientific Computing World (UK)
November 1, 2015

Avoiding Compromises in Scale-out NAS For Technical Computing

As large volumes of unstructured data (sometimes referred to as "big data") are increasingly used to drive innovation and discovery, the data sets to be processed and analyzed continue to grow rapidly, presenting significant challenges to IT staff. Enterprises have a range of choices when it comes to storage architectures, but more and more are turning to network-attached storage (NAS) instead of direct-attached storage (DAS). NAS is attractive because it scales better than DAS and enables more flexible architectures.

However, the challenges for network attached storage increase with scale. While most NAS products scale to some degree, simultaneously achieving linear scalability, high performance, and high reliability with ease of management is a challenge unmet by legacy architectures.

While performance is critical, performance that comes at the expense of manageability can hamper workflows and impact productivity. For example, legacy NAS systems create islands of storage making them unmanageable and costly at scale. Some scalable NAS products use clustering to improve scalability and manageability, but cluster overhead and serial file processing in the data path often bottleneck performance. Finally, reliability and availability at scale is critical to workflows as hard drive failures occur directly in proportion to the number of hard drives in the system. As a result, the benefits of greater capacity and performance at scale can easily be negated by lower data availability.

There is a new approach to address a breadth of technical computing requirements at scale. It starts with a hybrid architecture and parallel data paths that enable performance while increasing data reliability at scale. All this while maintaining an environment that is easy to manage, whether a single storage node or hundreds of nodes.

Parallel Data Flow

Storage becomes the weak link in the overall performance chain when compute clients wait on storage systems for data input and output. Exceptional storage performance is necessary to accelerate workflows for greater discovery, innovation, productivity, and profitability. The storage architecture must get past the limitations of standard network protocols and serialized data paths for greater performance – all while not compromising performance for any specific file workload (large file, small file, and mixed file workloads).

Similar to a multi-lane freeway that enables greater traffic flow, parallel data paths substantially increase data throughput to and from storage. In addition to data parallelism, avoiding performance robbing metadata accesses in the data path can greatly accelerate performance, but this is an approach that is not seen in legacy architectures. Enabling direct client to file data in object storage devices (OSDs) and processing metadata outside the data path can eliminate the vast majority of metadata-related performance penalties. Unfortunately, parallel data flow with direct client to storage device data access is not found in legacy scale-out storage architectures.

Linear Scalability

While it is relatively easy for a single legacy NAS-head product to double performance with an additional clustered head and associated storage, the real scale-out NAS challenge is to achieve one hundred times the performance from one hundred NAS heads.

There are better ways to achieve efficient linear performance at scale. We have already mentioned out-of-band metadata processing – in which a distributed metadata approach enhances linear scalability. Second, the RAID parity compute engine can be distributed across client machines so that RAID performance scales with the number of clients. This assumes data protection using software erasure codes. By eliminating the performance scalability limitations of hardware RAID, each client node calculates RAID at a per-file level, so RAID performance scales with the number of clients accessing storage. RAID performance is driven across an entire compute cluster — avoiding bottlenecks by NAS heads, NAS clusters, and hardware RAID controllers.

Thirdly, a storage architecture with a “shared nothing” design has no performance-robbing communication between storage blades. Done right, the end result is seamless scaling of concurrent users, client nodes, and linear performance scaling across the entire storage solution. By eliminating scaling inefficiencies that commonly occur at scale, users experience the same cost and performance efficiencies with 100 storage nodes as would be realized with a few storage nodes. Performance degradation at scale is eliminated. Storage nodes of shelves can be architected to non-disruptively and linearly scale out performance. Lastly, linear performance at scale cannot be accomplished without automatic load balancing to optimize performance and eliminate performance hotspots.

All of this assumes that storage shelves can be configured to accelerate small file I/O, large file streaming, or a mix of both file workloads. If a storage shelf is “tuned” for large files, for example, a small file or mixed file workload will impact performance – and ultimately linear performance scalability. It is therefore imperative that the base storage shelf be “tunable” for the expected file workload for maximum scalability.

Reliability at Scale With Triple-Parity Erasure Code Data Protection

As scale increases, reliability decreases in legacy NAS architectures. At scale, drive failures become more frequent simply because there are more drives to fail. Traditional hardware RAID solutions are architected for protection, but disk rebuild performance does not increase with scale. A complete rebuild can take or days – or even longer – for large systems. The bottom line is that long RAID rebuild times lower overall data availability. IT groups have learned to accept lower reliability and availability as the price for scaling, but it’s time to put that notion to rest.

New approaches to data protection at scale, including per-file triple-parity distributed RAID (such as PanFS RAID 6+) can result in a 150-fold increase in reliability over dual-parity data protection approaches by lowering the percentage of files affected by too many drive failures as the system scales. Per-file RAID with triple parity protection minimizes the amount of data that must be rebuilt after one or more failures, providing availability and reliability at scale that meets the requirements of business critical workflows and establishes a new standard for enterprise-grade data reliability.

	Data Reliability	**Volume of Rebuild Data**	**Time to Rebuild Data**	**Extended File System Availability (EFSA)**
Legacy NAS w/RAID 6	• Decreases with scale	• Substantial amount of data to be rebuilt after disk failure • Rebuilds damaged and undamaged files • Higher risk of prolonged exposure in degraded mode	• Rebuild in days • Rebuilds performance not scalable • Serial data channels slow rebuilds	• File system not available after triple disk failure • Complete file system restore often required
PanFS w/RAID 6+	• Increases with scale	• Minimal data to be rebuilt after disk failure • Only repairable files rebuilt in most cases • Lower risk of prolonged exposure in degraded mode	• Rebuild in hours • Rebuilds performance scales with the system • Parallel data channels for faster rebuilds	• File system is available after triple disk failure • Restores limited to a small list of specific files in most cases

Ease of Management

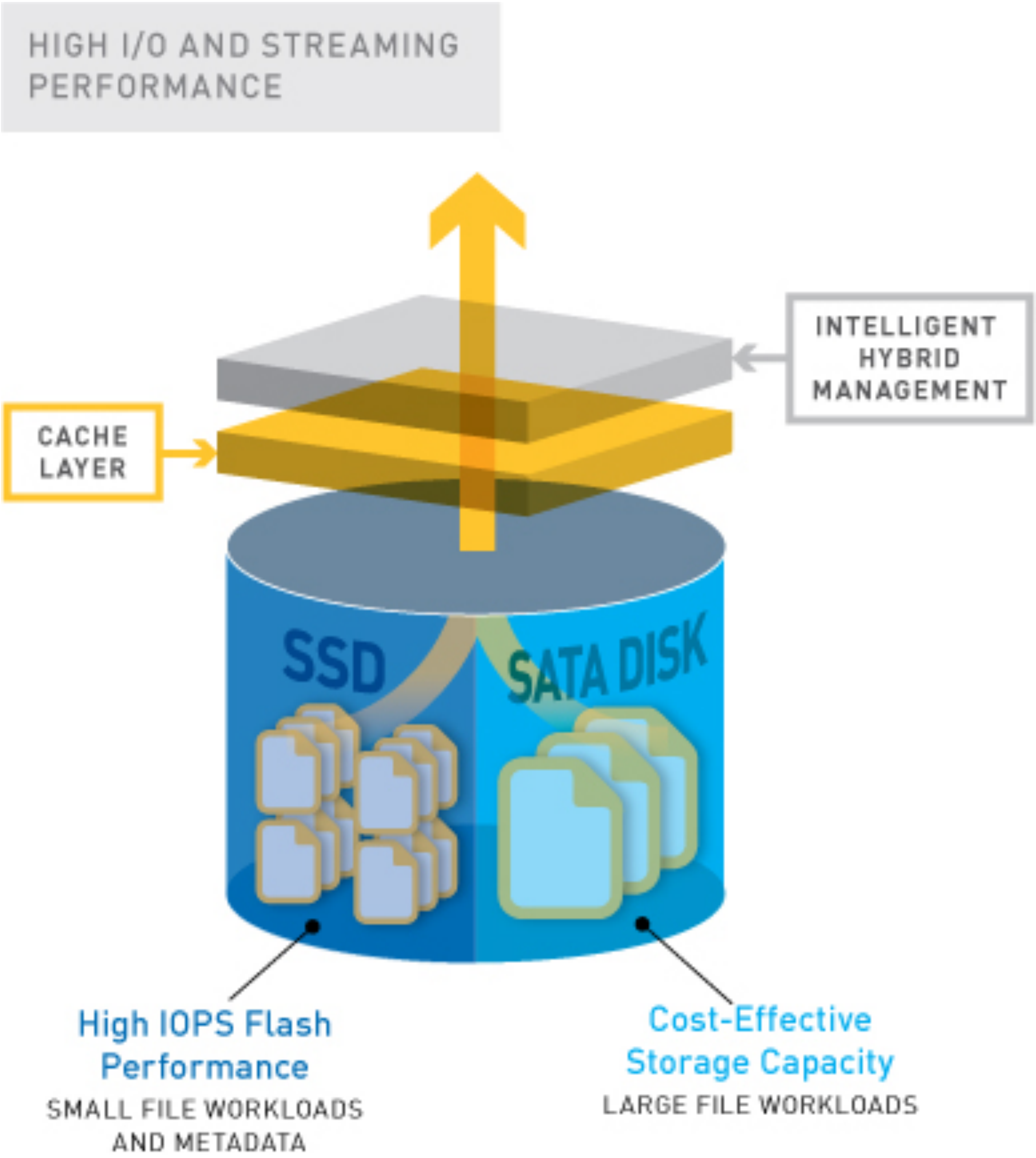
Increased scalability more often than not means increased complexity, resulting in greater management effort and cost. Legacy scale-out NAS is designed for scaling capacity and performance by adding storage nodes. Scaling to high capacity is easily supported, but substantial deployment and management effort may result if scalability is not joined with a global namespace, adaptive automation, linear scalability, and simple centralized management. A solution in which the level of effort required to manage a single node is the same as for managing hundreds of storage nodes is essential to lowering management effort and costs.

Using a storage system that offers a single global namespace, adaptive automation, and simple centralized management not only takes the headache out of scale-out NAS, but also reduces cost and administrative effort. An entire storage subsystem can be managed as a single appliance at any scale by automating key functions such as new storage discovery and load balancing, reporting, snapshots, user quota enforcement. Add to that a Simple Network Management Protocol (SNMP) for integration with other data center management tools and worries about storage administration become a thing of the past.

Intelligent Hybrid Architecture

Incorporating hybrid storage hardware (hard drives and flash devices), a file system and protocols in a fully integrated solution creates an ideal environment for mixed-workload performance, with rapid access to small and large files alike.

An intelligent hybrid architecture uses flash not simply as a cache in front of hard drives or to accelerate metadata performance. It will also ensure that files and metadata are located on the best type of storage technology for optimum performance and cost-effectiveness. Intelligent hybrid architecture places file data onto flash or enterprise SATA drives based on which device will deliver optimum performance and cost for a given file. This assures optimum performance for small file, large file, and mixed file workloads.


HIGH I/O AND STREAMING PERFORMANCE
INTELLIGENT HYBRID MANAGEMENT
CACHE LAYER
SSD
SATA DISK
High IOPS Flash Performance
SMALL FILE WORKLOADS AND METADATA
Cost-Effective Storage Capacity
LARGE FILE WORKLOADS

An intelligent hybrid architecture can be built on three storage tiers that are managed for maximum performance in relation to cost:

- Tier 1: Fast RAM
  - Caching for all files (read-ahead and write-behind for data and metadata)

- Tier 2: Flash storage
  - Solid State Disk (SSD) drives accelerate small files and file metadata performance

- Tier 3: Enterprise SATA drives
  - Large files reside on cost-effective enterprise SATA drives

- Per-file RAID striping for optimum performance

The intelligent hybrid approach enables the highest performance possible for cached files while delivering flash performance for small files that are not cached. Metadata also resides on flash, and not hard drives, to improve file system responsiveness. Large files are striped for performance across multiple cost-effective SATA drives. Hybrid management should be completely automated so that no management effort is required.

Storage Operating System Functionality

Most of the functionality we have already mentioned will reside in a storage operating system as seen in next-generation scale-out architectures. As such, expect next generation solutions to include some or all of the following as functionality delivered by an advanced storage operating system:

- Scale-out design with linear scalability for low TCO and for avoiding scaling inefficiencies
- Parallel data paths (vs. serial paths) for maximum performance
- A direct data access approach (vs. access delays with in-band metadata)
- Integrated hybrid approach – with intelligent tiering between hard drives and flash to keep small and large file workloads in the most optimum place for cost and performance
- Easy management (single global namespace, GUI, CLI, auto load balancing, reports)
- Erasure code software RAID (scales with clients and offers low overhead triple parity protection)
- Extended file system availability (file system availability even after 3 simultaneous failures)

Conclusion

Conventional NAS has been a workhorse for enterprises, but now that exponential growth of big data is exposing the limitations of scale-up NAS, there is a demand for next-generation scale-out solutions. By moving to a hybrid scale-out NAS solution with linear scalability, higher performance, increased reliability, and ease of management, business innovation can realize its full potential.

Author: Andre Franklin, Senior Product Marketing Manager - Panasas, Inc.