



The Avere Architecture for Tiered NAS

Avere Systems, Inc.
5000 McKnight Road, Suite 404
Pittsburgh, PA 15237 USA
1-412-635-7170
www.averesystems.com
info@averesystems.com

Part number 0254-001-0171

Executive Summary

The storage industry is at an inflection point. High-performance solid-state drives (SSDs) are threatening the dominance of hard disk drives (HDDs), particularly expensive 15k-RPM Fibre Channel (FC) drives, in storage systems. High-capacity, inexpensive serial ATA (SATA) drives have replaced tape as the archive medium of choice. Incumbent storage vendors are struggling to include new storage media in their products, but each candidate replacement medium has particular strengths and weaknesses that existing storage operating systems do not accommodate. Storage vendors have focused on capacity over performance even as application servers and clients require greater and greater throughput. To keep pace with the demands of your computing infrastructure, you must constantly upgrade to the latest, most expensive generation of NAS servers.

The Avere Systems FXT Series enables you to accelerate the performance of your current NAS infrastructure and lower the costs of NAS acquisition, operation, and expansion. The Avere solution also allows you to use less expensive, lower-performance NAS servers and lower-cost, high-density media such as SATA HDDs to expand your NAS infrastructure, increasing performance and extending the useful lifespan of legacy NAS systems, all while consuming less power, cooling, and rack space than traditional NAS solutions.

The Storage Medium Dilemma

Current NAS architectures are reaching their limits in terms of performance. As HDD capacity continues to rise, NAS servers struggle to keep pace with the volumes of requests sent to ever-denser disk drives. Disk I/O rates have remained relatively constant as capacity has risen; as a result, the number of operations per stored byte continues to decrease. For best performance, storage administrators must overprovision FC drives, wasting a significant percentage of their raw capacity. Even high-end NAS servers cannot overcome the fundamental limitations of increased access times on denser SATA drives while attempting to deliver high-bandwidth file access to clients. Other storage technologies such as de-duplication consume processing power on your NAS servers, also slowing overall performance. Adding new capacity to existing storage systems is relatively simple, but increasing performance requires that you upgrade to your vendor's latest high-end platform, add disk shelves to compensate for the relatively constant number of operations per second per drive (thereby consuming more rack space, power, and cooling), migrate data

from your current platform, and hope that the newest, most expensive generation of hardware meets your organization's performance requirements.

New storage media such as SSDs, whether in the form of large DRAM modules or flash memory modules, hold the promise of overcoming the performance limits of HDDs. However, each medium has its own limitations; for example, DRAM is volatile and flash has a limited number of write cycles. In addition, successfully deploying different types of media in your data center requires specialized knowledge of both the applications being run, which can change over time, and the characteristics of the storage media. Most storage administrators do not have time to learn, let alone implement, this information in today's "do-more-with-less" business environment.

The Avere Solution

To solve the problems of HDD performance and new storage media, Avere Systems introduces Demand-Driven Storage™ with *tiered NAS*; that is, a network-attached storage system that intelligently places your data on the storage medium best suited for its current access patterns. The Avere system, consisting of FXT Series hardware and Avere OS software, is deployed between your current NAS server (in Avere terms, the *mass storage server*) and your clients and application servers, thus enabling you to continue to use your existing infrastructure without disruption to data access. The FXT Series server contains and uses different types of storage media, including solid-state storage and serial-attached SCSI (SAS) HDDs. Avere OS software analyzes how files are being accessed and places the files internally on the most appropriate storage medium for the fastest possible access. For example, a small file that multiple clients are actively reading and writing to is stored in DRAM, while a very large file that clients are predominantly accessing sequentially is stored on SAS drives. A file that has not been accessed is retained only on your mass storage server. Unlike some solutions, the Avere system benefits write loads as well as read-only data. For optimal performance, changes made to data by clients and application servers are stored locally on the Avere system, which writes all changed data back to the mass storage server at an interval specified by the administrator.

Multiple FXT appliances, or *nodes*, form a scalable *cluster* that increases performance and working-set size linearly as new nodes are added to the cluster. Just as you can add new storage to your mass storage server to increase capacity, you can add new nodes to the Avere cluster to increase performance. A cluster can contain up to 25 FXT nodes. The resources of each node become part of the cluster's combined resources. For example, an FXT 2300 node has 1.2 TB of SAS storage, so a four-node cluster has a total SAS capacity of 4.8 TB.

By offloading the “heavy lifting” of processing file requests from the NAS infrastructure, the Avere cluster separates data retention on the mass storage server from high-performance data delivery and frees processor cycles on your mass storage server for tasks such as data mirroring, de-duplication, and backup operations.

The advantages of the Avere solution, which are discussed in more detail throughout this document, include the following:

- Accelerate the performance of your current NAS servers to increase the performance of your most demanding applications
- Preserve your investment in your existing NAS infrastructure by dramatically improving its performance and extending its useful lifespan
- Enable the use of less expensive NAS servers and lower-cost, higher-capacity SATA drives as primary storage to expand the capacity of your NAS infrastructure without sacrificing performance
- While maintaining or improving performance, save money in the following areas by decreasing the number of expensive NAS servers and disk shelves in your data center:
 - Cost per terabyte
 - Power
 - Cooling
 - Rack space
- Pay only for the performance you need, with the option of scaling performance in the future by adding more FXT nodes to your Avere cluster

Inside the Avere Cluster

The Avere system consists of two primary components, the FXT Series platform and the Avere OS software. An FXT appliance is an Intel[®]-based server that uses a combination of DRAM SSD, flash SSD and/or SAS HDD, and battery-backed NVRAM as internal storage; it provides high-bandwidth network access via 1GbE or 10GbE network ports. Avere OS includes the following major features:

- Tiered File System (TFS)—Avere’s network file system that intelligently distributes data across different tiers of storage media

- Avere Control Panel—An intuitive browser-based GUI that provides a single system image of the cluster; you can administer any entity in the cluster by logging into the Avere Control Panel on any node
- Automatic assignment and failover of network ports and IP addresses
- Support for NFSv3 file servers
- Support for NFSv3 clients
- Support for NFS export policies
- Optional support for CIFS clients
- Optional N+1 high availability to ensure continued service in the event of a node outage

The core of Avere OS is TFS, a journaled file system that analyzes data-access patterns from client requests, places files on the optimal storage tier on the FXT nodes, moves files to different tiers as needed, and commits changed data to the mass storage server within the limits of an administratively specified interval. TFS's allocation algorithms can differentiate, for example, among large sequential write operations, which are most efficiently handled by a combination of HDD and NVRAM; random read operations, which is most efficiently handled by flash memory; and small read and write operations, which are most efficiently handled by DRAM. See "How the Cluster Handles File Operations" for more details.

The data that TFS stores on the Avere cluster is called the *working set*. The size of the working set can be multiple terabytes, depending on the amount of active data and the capacity of the cluster. As clients and application servers request new files, the Avere cluster retrieves them from the mass storage server and adds them to the working set. As files become less active, TFS moves them to slower storage tiers and eventually removes them from the working set, at which point they are located only on the mass storage server.

Avere OS and TFS support three *write modes*, or ways that the Avere cluster handles data written from clients:

- In *write-back mode*, the typical operating mode, TFS handles read and write operations from the working set and writes updated data to the mass storage server within an administratively specified interval (the *maximum write-back delay*). This mode accelerates read, write, and metadata operations.
- In *write-through mode*, TFS updates the mass storage server after each write operation, but otherwise performs as in write-back mode; it can be used as an alternative to the Avere cluster's high-availability solution. This mode accelerates read operations and provides complete data

reliability. This mode accelerates both data and attribute read operations, but the performance of data and attribute write operations is limited to that of the mass storage server.

- In *write-around mode*, TFS updates the mass storage server after each write operation and checks the attributes on each file before using cached data, to enable the use of an FXT cluster while other systems both read and write the same data directly to the mass storage server. This is the expected mode for initial installation and setup of the Avere cluster, as clients are migrated from connecting directly to the mass storage server to connecting through the Avere cluster. This mode accelerates data read operations, but the performance of attribute reads, attribute writes, and data write operations is limited to the performance of the mass storage server.

Avere OS supports a single write mode per mass storage server. However, the general write mode on individual directories and file systems on the mass storage server can be overridden by using *Avere control files*, which are small text files that can specify a new write mode to be used in the directory containing the control file and its subdirectories. Additionally, for mass storage servers that support snapshots (point-in-time, read-only backups of file systems), the Avere Control Panel enables you to specify the naming convention of snapshot directories so that TFS automatically uses the appropriate write mode within the snapshot directories.

How the Cluster Handles File Operations

When a client or application server makes a file request, the cluster determines whether the file is in its working set; that is, whether it already holds the file's attributes or any of the file's data blocks. If the cluster does not currently hold the section of the file being accessed, it reads that section from the mass storage server. When the cluster has the data, it uses a number of criteria to determine the most suitable storage tier in which to store it. The criteria include the type of file operation being performed (read or write), the file's size, and whether the file is being read sequentially or randomly, along with other considerations. As the cluster receives additional requests for the same file, it dynamically modifies its handling of the file. As an example, if the cluster initially receives only a few random read-only requests for a large file, it places the file in DRAM and eventually writes it to its SAS storage, keeping only the hottest data in DRAM. If the cluster then sees multiple random reads for an increasingly large number of blocks, from many clients, it moves some blocks from DRAM to SSD, retaining the hottest data in the highest-performance storage medium. If the file is modified with write operations, the cluster also writes the changes back to the mass storage server within the time period specified by the maximum write-back delay setting.

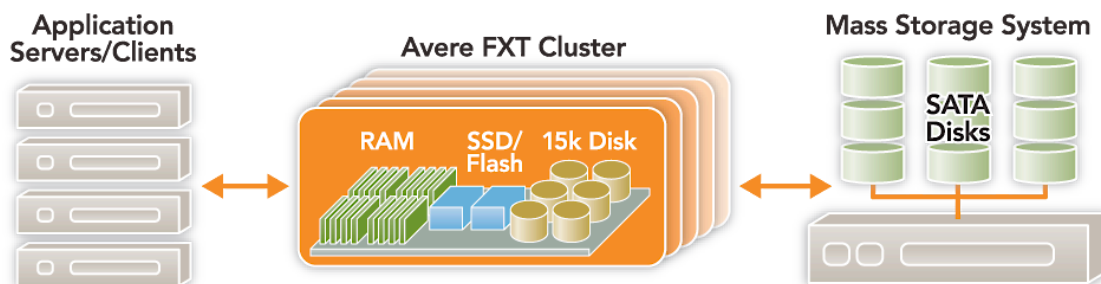
In all cases, the contents of the file are distributed across the pooled resources of all of the nodes in the cluster, preventing data from becoming bottlenecked on a single node. TFS serves the file's data as blocks and permits different clients to access and update different parts of the file. This is particularly useful for very large files that are accessed by multiple clients or threads simultaneously; for example, client A can write one part of a file while client B simultaneously writes a different part of the same file. Additionally, if the access patterns indicate the need, TFS can place read-only copies of the file on multiple FXT nodes in the cluster.

Deploying Avere in Your Data Center

The Avere solution is designed for simple, non-disruptive setup and operation in new or existing NAS installations. During setup and configuration of the cluster, your clients and application servers can remain connected directly to the mass storage server. Setup of the first node in a cluster takes only a few minutes; for additional nodes, you can specify that new FXT appliances on the network automatically join the cluster or manually select new unconfigured nodes from a list and add them to the cluster.

After setup, you have the option to gradually migrate a number of client connections from the mass storage server to the Avere cluster with the cluster in write-around mode, thus ensuring that all client-changed data is written directly to the mass storage server. Write-around mode enables you to migrate the rest of your clients to the Avere cluster on a schedule that meets your business needs; no interruption of data service is required. After all of your clients are connected to the mass storage server through the Avere cluster, you can change the cluster's write mode to write-back mode for maximum performance acceleration.

The following figure shows how an Avere cluster fits topologically into a typical NAS installation.



After you create the cluster, you can use the Avere Control Panel to configure additional features such as NFS export policies, CIFS access, and high availability (HA).

Monitoring and Administering the Avere Cluster

As with setup, monitoring and administration of the Avere cluster is designed to be as simple and as hands-off as possible, leaving administrators free to concentrate on other parts of the storage infrastructure.

You perform monitoring and administrative tasks by using the Avere Control Panel, which is divided into tabbed pages by functional area. The upper right-hand corner of each page provides an at-a-glance summary of the cluster's status in the form of a "traffic light": a green indicator means that all is well with the cluster; a yellow indicator means that the cluster is serving data but a condition exists that an administrator needs to investigate to ensure optimal operation; a red indicator means that the cluster cannot serve data. Clicking on the traffic light from any page takes you to the Status tab, where you can view system messages and monitor system performance. Other monitoring mechanisms for the cluster include:

- An operations-per-second display
- A CPU performance gauge
- An advanced performance graphing display
- Email alerts
- Support for remote syslog servers
- Support for SNMP discovery
- The ability to display information about client connections, CIFS clients, hot clients, and hot files; see "Take Control of Your NAS Environment" for more information

After the cluster is configured, it requires little administration. Tasks that administrators might need to perform include adjusting the maximum write-back delay, adding or changing NFS export policies and CIFS shares, and upgrading the Avere OS software.

The Avere Control Panel includes a Support page that is designed to facilitate interactions with technical support. If issues that require technical support occur, the Support page provides administrators with a one-stop interface to get

troubleshooting information to technical support, thus speeding the resolution of problems.

System Resilience

The Avere system includes enterprise-class failover and resiliency features to help ensure that data continues to be served in the event of a component failure and that data is not lost in the event of a catastrophic failure such as a power outage.

Any data that is written to the Avere cluster is retained on battery-backed NVRAM. This protects your data in the event of a node failure or power outage.

When you create an Avere cluster, you specify the network (IP) addresses that the cluster uses to communicate with clients and the mass storage server and among the cluster's constituent nodes. You do not specify the physical ports to which these addresses are bound. Instead, Avere OS automatically binds addresses to physical ports; in the event of a port or NIC failure, it automatically migrates addresses to other functioning ports and rebalances the load on all ports, ensuring the optimal continued flow of data. No manual setup or administrative intervention is needed.

For environments that cannot tolerate downtime, you can enable high availability (HA). When HA is enabled, each cluster node mirrors its write data to another node; if the node fails, its data continues to be served and written back to the mass storage server by the mirror node. Because Avere OS's HA implementation is N+1, not 1+1, it can be enabled in any cluster that has two or more nodes; it does not require an even number of nodes or any manual configuration other than enabling HA from the Avere Control Panel.

Take Control of Your NAS Environment

Because of its placement in the storage environment, the Avere system provides visibility into all activity between your mass storage server and clients and application servers. You can use Avere OS's monitoring tools to locate potential problems such as especially demanding clients or application servers, heavily accessed files, and slow or overloaded NAS servers. You can use this information to balance demand across the storage network by distributing workload across different clients, adding capacity to the mass storage server, or increasing NAS server performance by expanding the Avere cluster.

Bringing Unparalleled Performance and Value to NAS

There are two dimensions to consider when evaluating performance. First, there's the application workload. Avere solutions provide high performance across a wide range of workloads by accelerating read, write, and metadata operations. In addition, these operations are accelerated across the full range of access patterns, including random access to small files, sequential access to large files, and a mix of both. The second dimension is the amount of performance (in operations per second) or throughput (in MB per second) required for a given workload. The Avere FXT Series meet the needs of most applications because they provide high performance on a single appliance and linearly scale performance as appliances are added to a cluster. In small file, random access tests, FXT clusters can achieve millions of operations per second. In large file, sequential tests, FXT clusters can achieve tens of GB per second of throughput.

The best way to size an Avere system for your performance requirements is to work with an Avere systems engineer. To contact an Avere systems engineer, please submit a request [here](#).

Typical Avere Deployments

The Avere system can enhance the usefulness of almost any NAS installation, as discussed in the following scenarios. You and your Avere Systems representative can discuss the value that Avere can bring to your specific storage environment.

Cost Savings

You have a high-end clustered NAS installation for high-performance computing and need to add 200 new clients to the compute cluster. To provide storage for the added computing capacity, the storage vendor's suggestion is to add two new high-end file servers to the storage cluster. To support the required performance, each NAS server will require 12 FC disk shelves at a low percentage of utilization instead of a third the number of SATA disk shelves. The additional system components will require two new racks, 15.66 kW of power, and 40,652 BTU/hour of cooling. The alternative solution from Avere is a four-node FXT cluster with no additional disks on the mass storage server. It requires 8U of rack space, 2.7 kW of power, and 9248 BTU/hour of cooling. The initial cost of the Avere system is a fraction of the incumbent vendor's proposal. The annual cost of power, cooling, floor space, and administrative overhead for the Avere system is an order of magnitude smaller than that of the incumbent vendor's system.

New Installation

You need to install 50 TB of NAS storage for a new project with 300 compute nodes. Instead of buying 170 300-GB FC drives and installing them on six high-end file servers, you buy 35 1.5-TB SATA drives, install them on two entry-level file servers, and place a six-node FXT cluster in front of the NAS servers. Not including the reduced up-front capital expenditure, the savings include 50U of rack space, 3.5 kW of power, and 50,421 BTU/hour of cooling.

WAN Installation

You need to implement a midrange NAS system for a remote office with a mirror back to corporate headquarters for archiving. Instead of purchasing a midrange NAS server mirroring back to the corporate office's archival infrastructure, you purchase a single FXT node, which stores the working set from the corporate office's archive of the remote office's data. The remote office gets the performance of a midrange NAS server through the Avere node.

Increasing Performance

The EDA group has added 500 TB of capacity to their farm of NAS servers and plans to add another 500 TB over the next 12 months. Because of a corporate cost-savings program, they purchased a midrange server with SATA disks instead of a high-end server with FC disks. Members of the group are now complaining that data access is slow on their workstations, and their productivity is lagging. After determining the performance requirements and consulting with your Avere representative, you purchase a six-node Avere cluster to accelerate the performance of the new midrange NAS server. As you look ahead to the next 500-TB capacity upgrade, you determine that it is more cost-effective to purchase a midrange NAS server with SATA disks and an eight-node Avere cluster than to purchase a midrange or high-end NAS server with FC disks.

Extending Equipment Lifespans

When planning a company-wide storage upgrade, you have the requirement to provide a 640-TB midrange NAS system to the training department. You are already planning to replace a 720-TB archive system consisting of a pair of low-end NAS servers and SATA disk shelves with a 1.2-PB archive system consisting of a single higher-capacity server and denser SATA disks. Instead of purchasing both a new midrange system and the new archive system, you purchase the 1.2-PB archive system and a two-node Avere cluster to increase the performance of the 720-TB SATA-based archive system to that of a new FC-based midrange system. The training department is pleased because they get the performance they requested plus an extra 80 TB of storage.

Conclusions

The Avere system's unique combination of multiple storage tiers in a single appliance and software that dynamically organizes data onto those tiers enables organizations to maximize performance and minimize costs in their NAS infrastructures. Instead of using numerous expensive NAS servers with FC HDDs, Avere allows you to use small numbers of lower-performance servers and lower-cost, higher-capacity SATA HDDs, at a great reduction in total cost of ownership but no reduction in performance. Because the Avere system works with any NFSv3 server¹, you no longer have to buy the latest high-end equipment to achieve the performance your applications and users require. Tiered NAS makes it simple to maximize the density and performance of your NAS infrastructure while minimizing the costs of administration, equipment, power, cooling, and rack space. In an era when critical business decisions must be made at a moment's notice based on the latest information available, Demand-Driven Storage from Avere Systems helps ensure that time-to-market objectives are met and storage expenditures stay down.

¹ Contact Avere Systems for a list of NFSv3 servers that have been tested with the Avere system.