NFS Evolution Changes the Landscape of HPC Data Management

Addison Snell

White paper sponsored by: BlueArc

OVERVIEW

In the high performance computing (HPC) market, price/performance reigns. Achieving optimal productivity within the constraints of facilities and budget is the constant struggle of the HPC user. Unfortunately, adequately assessing price/performance is not as straightforward as it once was. Processing elements have continued to get faster and cheaper with metronomic regularity, and as a result, the bottlenecks in application performance are increasingly found in data management.

Furthermore, the cost equation within productivity necessarily includes more than the cost of hardware and software. Staffing and administration are significant expenses, and to lower these costs many organizations prefer to follow standards in their IT solutions, allowing their administrators' expertise to span the entire infrastructure.

Herein lies the problem. In HPC data management, traditional standards-based solutions have been limited in performance and scalability, but proprietary, high-performance solutions have required specific expertise to set up, manage, or scale. On the one hand, a storage architect could implement a standards-based solution, such as one based on the NFS file system protocol, but such solutions may fail to provide performance at scale without significant redundancy and expense. On the other hand there are custom, parallel file systems that are designed for scalability and performance, but these sacrifice standards and interoperability.

BlueArc seeks to address this tradeoff with a hybrid file system combining the benefits of standard NFS and the performance and scale of parallel file systems. Rather than settling for the status quo ("NFS does not scale"), BlueArc questioned it ("Why does NFS not scale?") and challenged it ("What can we do to enable NFS scalability?"). The result is a storage solution that bridges the gap between scalability and interoperability.

Critical enabling features of the BlueArc storage solution include:

- FPGA acceleration: The data that passes through the BlueArc I/O server is parallelized and accelerated with field-programmable gate arrays (FPGAs), which speed up performance regardless of protocol.

- Virtualized storage: BlueArc makes use of SAN hardware and protocols, and the BlueArc file system offers block-level features to the storage administrator. The BlueArc Virtualization Framework gives a single administrative view of disparate pools of storage, enabling SAN functionality in a NAS environment.

- Parallelization: BlueArc's object-oriented filesystem not only separates metadata and data management, but also parallelizes data movement. This parallelization is at the heart of BlueArc's approach to performance and scalability.

Upcoming evolutions in HPC data management, including the emergence of a parallel NFS standard and the potential for consolidation in data management, may raise the value of interoperability still further and put increased pressure on shared disk implementations. BlueArc is well-positioned with a solution that addresses problems today and is compatible with new implementations and protocols on the horizon.

MARKET DYNAMICS: FILE SYSTEMS IN HPC

Ask an average person to name the parts of a personal computer. They will typically start with the obvious external components (keyboard, mouse, monitor or screen) and move to some basic internal components they are familiar with, such as the CPU, motherboard, and hard drive. At a basic level, we consider the hard drive to be part of the computer.

Somewhere between PCs and HPC we tend to lose this distinction. HPC lab directors will proudly show off their data centers, directing you to observe their computers over here (racks and racks of CPUs) and – if they think of it as part of their HPC infrastructure at all – their storage over there. The storage is not part of the computer, but rather a separate appendage, something that the computers are forced to contend with on occasion, and hopefully minimally, because getting data from storage can be painfully slow, wasting those valuable CPU cycles. Thus we have one of the classic, sarcastic definitions of a supercomputer: a device for turning a compute-bound problem into an I/O-bound problem.

The shift in computation away from tightly integrated SMPs toward loosely coupled clusters served to reinforce this tenet. The separate computational domains do not share data access or memory, leading to potential inefficiencies in connecting all of the nodes to the storage network.

The key technology component for optimizing data management across clusters is the file system. Depending on the implementation, file systems can either enable or hinder data access and sharing across a cluster. However, looking at file systems is not a simple price vs. performance tradeoff. There are myriad file systems available in the HPC market. A 2007 Tabor Research study found 17 different file systems being used among only 40 different sites. The file systems also fall into different categories that follow different basic schemata for how data is accessed. In order to compare file systems, it is first necessary to understand these categories and their relative strengths and weaknesses.

Categories of Cluster File Systems for HPC

File systems for HPC clusters can be categorized into three broad groups: distributed file systems, parallel file systems, and shared disk file systems. **Distributed file systems** generally provide data access to a cluster of servers through one or more nodes in the cluster that are designated as I/O servers. A typical example implementation uses one or more nodes acting as NFS exporters of shared mount points for the remainder of the cluster. This category also includes the use of "filers" acting as NFS exporters. In either case, the filesystem is distributed in nature, meaning the NFS exporters offer access in a distributed fashion to the cluster. In such environments, each of these I/O nodes has access to a portion of the data, and I/O traffic for the cluster is routed through the appropriate I/O server or filer.

The most common distributed file system for HPC is NFS, although Microsoft's CIFS/SMB and Novell NetWare (NCP) also fit the category. NFS protocols have been around for many years and are the de facto standard for data access in UNIX and Linux implementations. While NFS has the plug-and-play advantages of standards conformity, most NFS implementations are perceived to have limitations in scalability or performance for large-scale HPC data sets.

Another traditional limitation of distributed file systems is that only one node can access a given file at a time. Files are "locked," usually with a standards-based locking mechanism or locking manager to assure single access. **Shared disk file systems** were designed to get around this hurdle, allowing multiple nodes to read and access files simultaneously.

Shared disk file systems are generally associated with storage area networks (SANs), although in most cases they are capable of running in network-attached storage (NAS) modes as well. Shared disk file systems also offer separation of metadata and data operations in an attempt to virtualize large SAN environments. These file systems frequently advertise specialized features or performance advantages over traditional NFS implementations, but because they are proprietary they frequently carry greater costs, not only in the price of the file system itself, but also in specialized administration.

Shared disk file systems use metadata servers in order to separate metadata from data management, but parallelization remains an issue. Even at large scale many SAN implementations will employ a single metadata server, effectively creating a bottleneck and a potential single point of failure.

Another path for HPC data management, **parallel file systems** attempt to combine the advantages of distributed and shared disk implementations. Like distributed file systems, parallel file systems also use designated I/O servers to manage data traffic; however, parallel file systems attempt to bolster scalability by striping data into locations that can be accessed by multiple computational nodes simultaneously. Like shared disk file systems, the main benefit is enhanced performance and scalability over common distributed file systems, at the potential cost of standardization.

> **distributed file system:** Provides data access to clusters through one or more selected I/O servers. NFS is the most common distributed file system. *Advantages:* Generally cost-effective, standards-based, with low administration costs. *Disadvantages:* Most implementations are limited in scalability and performance.
>
> **shared disk file system:** Provides universal shared access to all data from all nodes in a cluster, frequently used with a SAN. *Advantages:* Concurrent access to data. *Disadvantages:* Proprietary offerings that may carry higher costs. May be difficult to scale metadata traffic horizontally across cluster. Standards moving away from shared disk.
>
> **parallel file system:** Stripes data into locations that can be accessed by multiple nodes in order to enable greater scalability and performance. *Advantages:* Greater scalability and performance than traditional NFS implementations. *Disadvantages:* Greater expense in setting up multiple I/O servers for data redundancy. Proprietary offerings with potentially higher administration costs.
>
> **pNFS:** Parallel NFS, an emerging implementation of NFS that is a multi-vendor standard. pNFS has the potential to become a transition platform for distributed and shared disk file systems.

These three categories – distributed, shared disk, and parallel – fairly cover the cluster file system landscape at present. One other important product under development could change (and perhaps simplify) the HPC landscape. A consortium of storage vendors, led primarily by NetApp, Panasas, and Sun, with additional contributions from companies such as BlueArc, EMC, and IBM, is currently working on a parallel version of NFS, known simply as **pNFS**. The goal of pNFS is to build parallelism into the NFS protocol standard, providing a common, high-performance file system for clusters. Because it incorporates scalable design aspects, pNFS is a possible transition platform and consolidation point for many HPC data centers.

Distributed File Systems

Distributed file systems provide access to data through fixed access points. All distributed file systems follow a basic prescribed architecture for the data network. Among a cluster of servers, one or more designated I/O servers are selected to provide data to the compute nodes.

The I/O server becomes the gateway for all data distribution to the cluster. When a particular process running on the cluster needs data, the node that needs it passes the request to the I/O server. The I/O server accesses the data from storage and passes it back. Storing data works the same way; the data is sent over the network to the I/O server with instructions to store it. Typically the I/O server is connected to the compute nodes through an independent network – usually over the I/O port on the node – in order to keep the data traffic separate from the computational message passing.

Figure 1: Diagram of Network Attached Storage (NAS) with Distributed File System
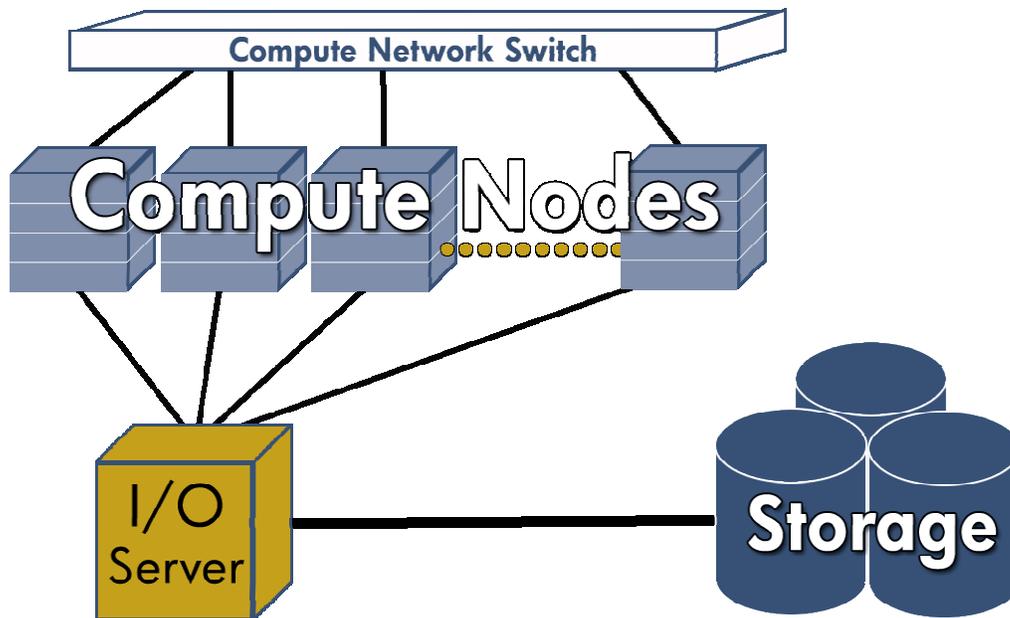
NFS is the most common distributed file system, and it follows this scheme. [See Figure 1.] The main advantage of this architecture is simplicity. NFS solutions are plug-and-play, and the administration of NFS servers is broadly understood. NFS implementations are therefore common for low-end clusters, for which performance and scalability are not at issue, and for larger scale implementations that nevertheless do not require a large degree of data movement.

The limiting factor with NFS is scalability, especially for applications that are data-intensive. As the system scales, more and more data traffic must pass through the I/O server. Eventually the amount of data exceeds the I/O server's bandwidth capabilities, creating a bottleneck.

The most straightforward way to attempt to circumvent the scalability problem is to add additional I/O servers, providing additional paths for data access. The disadvantage of this approach is that it is expensive to carry the overhead of additional I/O servers beyond the number necessary for path failover and redundancy. The implementation of multiple I/O servers therefore can be an inefficient way to deal with large amounts of data, especially if the data load is inconsistent or "bursty," requiring the user to maintain an infrastructure capable of handling peak activity levels.

Shared Disk File Systems

The scalability and performance limitations of NFS led some vendors to develop their own shared disk file systems. In general shared disk file systems allow multiple nodes to access data structures simultaneously, so there is no central I/O server to be a potential bottleneck. Because each node can access the data without going through a central I/O server, the data appears to be directly attached to the node. For this reason, shared disk file system implementations are commonly used with storage area networks.

Shared disk file systems allow multiple nodes to read the same file simultaneously, but only one process has the ability to write to a specific file at a given time. Other processes can continue to read the file at that time. The file locks prevent "interceding updates," in which a file is edited simultaneously by two different processes. [See Figure 2.]

Figure 2: Storage Area Network (SAN) with Shared Disk File System
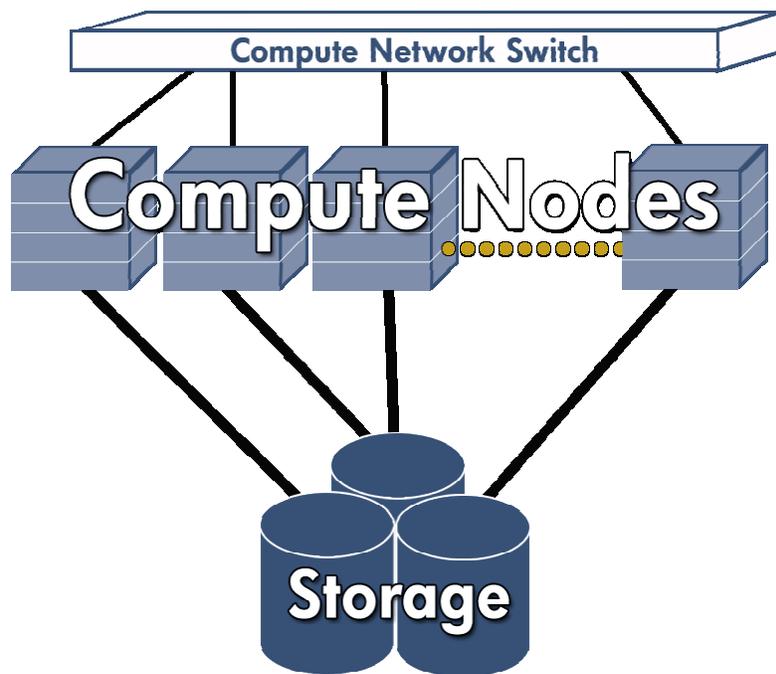
The other challenge for data management scalability is metadata. All data generates metadata. Metadata contains information about a file; it is the data that represents the data. For a given file, the metadata might communicate the size of a file, the location in which the bits are stored, a record of processes accessing or editing the file, etc. While parallel file systems share metadata information among multiple cluster nodes by spreading the management task equally across the I/O nodes, shared disk filesystems typically make use of a single metadata manager.

Because shared disk file systems are largely proprietary, the details of specific implementations vary. Some of the most common shared disk file systems are QFS (Sun Microsystems[1]), CXFS (SGI), PolyServe (now owned by HP), and GFS (Red Hat), as well as custom offerings from

---

[1] As of this white paper's printing, Oracle had agreed to acquire Sun Microsystems, but the transaction had not yet closed. Oracle has not commented on future development of QFS or Lustre.

EMC, Quantum, Symantec, and Oracle. None of these is dominant in the market, with more recent innovations taking place in the parallel file system category.
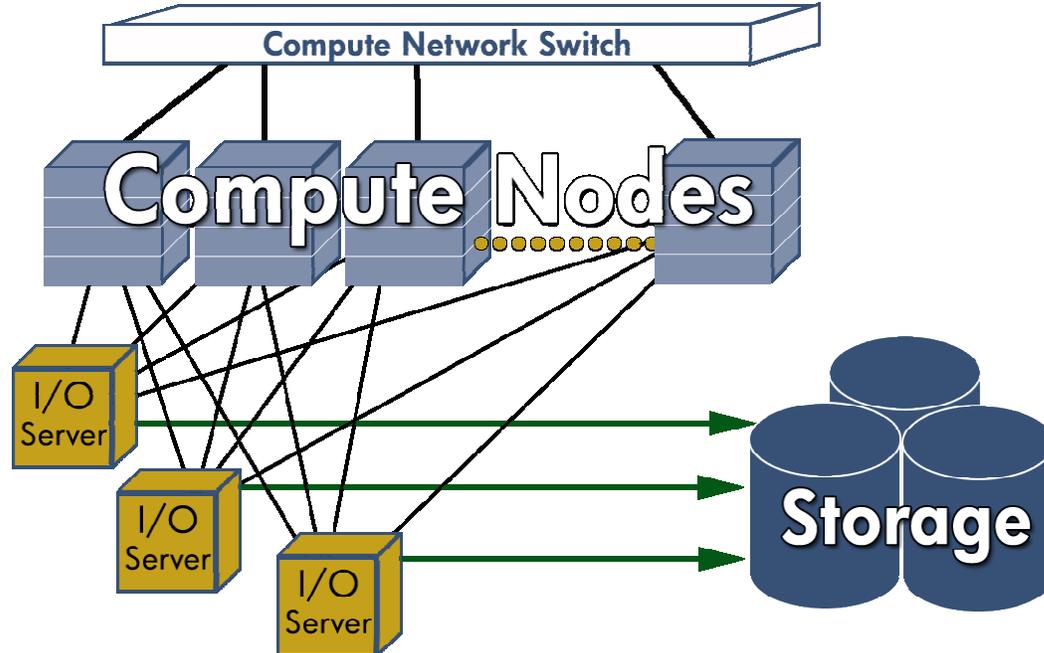
The primary advantage of shared disk file systems is their scalability, and therefore their performance, for large-scale, data-intensive applications. Most SANs can be readily expanded, and some are heterogeneous, meaning they will support multiple operating systems. However, no single shared disk file system is widely adopted to the extent NFS is. In addition, Tabor Research studies confirm that organizations tend to prefer NAS solutions until they are "forced" to implement a SAN to achieve greater scalability. This transition from NAS to SAN can be difficult, because of the change in file system and the resultant need to re-optimize the data workflow.

The most significant challenge for shared disk file systems is effective horizontal scaling of metadata. Metadata management is critical for allowing concurrent data access at scale, and while SANs do tend to perform well for large amounts of data, shared disk file systems can deteriorate significantly in performance as the number of files increases, because of the high degree of metadata-related communication required between I/O nodes.

Parallel File Systems

Most of the recent development in HPC data management has gone into the growing category of parallel file systems. Parallel file systems offer several potential advantages to distributed and shared disk file systems by combining features of both. Parallel file systems employ multiple I/O servers and stripe data into multiple locations, so that a given file could be accessed through more than one I/O server (although not necessarily simultaneously). [See Figure 3.]

Figure 3: Data Management with a Parallel File System

In this schema, the primary advantage over shared disk file systems is the ability to scale with multiple I/O servers as opposed to through a single metadata server. The extra paths avoid bottlenecking and thus boost performance while also allowing options for additional features like fault tolerance.

Development in parallel file systems has led to many competing offerings. Some of the more common file systems in this category are PVFS2 (open-source), Lustre (designed by Cluster File Systems, Inc., now owned by Sun Microsystems[2]), GPFS (IBM), PanFS (Panasas), and OneFS (Isilon).

The newest entry in this category, pNFS, has the potential to change the landscape of HPC data management. pNFS is based on NFS and therefore looks and feels like a standard distributed file system; however, it is properly classified as a parallel file system due to its ability to scale with multiple I/O servers. For some users pNFS will therefore appeal as a sort of Holy Grail for HPC data management: it promises to provide scalability and performance without sacrificing standards or administration costs. Tabor Research has found that the majority of HPC users are aware of pNFS's development and planning to evaluate it for their own needs.

Because it aims to address problems on both sides of the standards vs. performance issue, pNFS provides a potential evolutionary path for both. As a result we may expect to see consolidation in HPC file system usage in the years ahead, but until pNFS is released and battle-tested, HPC users will continue to search for the solutions that best fit their individual criteria for productivity.

---

[2]   See previous footnote on Oracle and Sun.

BLUEARC SOLUTIONS: ACHIEVING HPC PERFORMANCE WITH NFS

Amidst the myriad file system implementations competing in the three categories outlined above, BlueArc is following a unique strategy that seeks to combine the advantages of each. The essence of BlueArc's hybrid approach is to achieve scalability with differentiated, high-performance hardware, while still using standard data access protocols like NFS as a front-end. BlueArc's intent is to combine the simplicity of NFS with the scalability and performance of parallel file systems, while still leveraging SAN hardware to enable robustness, performance, and scalability.

As with other customized approaches BlueArc relies on its own engineering enhancements to achieve scalability and performance at levels relevant to HPC sites, and BlueArc has indeed built its internal file system from the ground up. However, one critical difference is that BlueArc began with straightforward NAS as its basic model and still uses NFS as a front-end; enhancements are added to this basis. BlueArc's view is that NFS is capable of scaling, given the proper approach and technological investment.

BlueArc Titan Product Family

The fundamental question driving BlueArc's technology roadmap is, "How can we make NFS more scalable?" The basic challenge is to alleviate the everyday I/O bottlenecks that come as standard-issue with scalable file system deployments. BlueArc takes a two-step approach with its Titan product family: separate metadata management from the data pathway, and then accelerate the data protocols for the traffic that remains. [See Figure 4.]

Figure 4: BlueArc Titan Implementation
Source: BlueArc
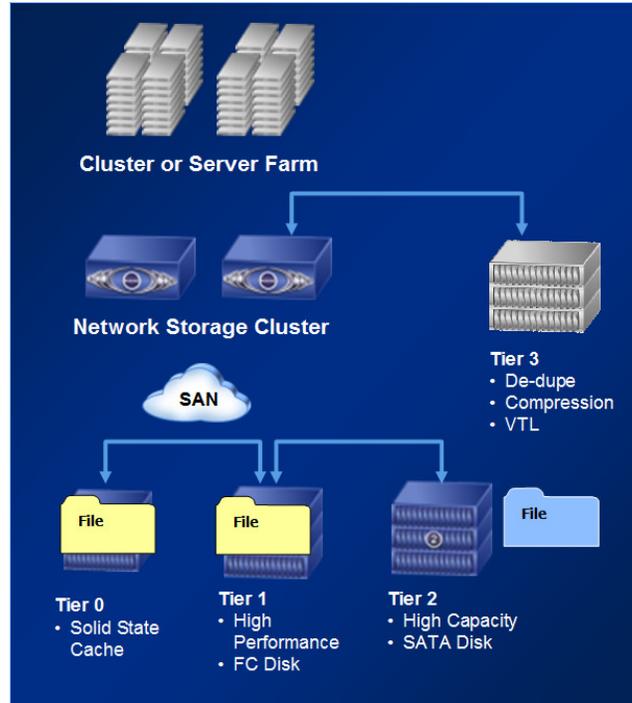
In any object-oriented file system, the responsibility of metadata management is passed off to intelligent storage devices. The host operating system sends requests for data to storage but does not know, for example, upon which spindles the data is stored. The effect is similar to giving

your car over to valet parking: you know approximately where the car is, but the valet knows precisely. The storage system manages the metadata for the cluster.

Object-oriented storage is used in other parallel file systems, but BlueArc is noteworthy in deploying it with NFS. The use of object-oriented storage is a common-sense approach to redefining NFS scalability. In order to address the main problem with I/O server scalability, metadata management is moved off to intelligent devices: in this case, BlueArc's Titan server. BlueArc has replaced the NFS server with a device that the cluster sees as a traditional NAS I/O server, but which in reality is a much more evolved design.

Object-orientation notwithstanding, what makes the Titan product family unique is that BlueArc accelerates the traffic on the NFS server with FPGAs, a set of accelerator chips that are specifically programmed to speed up both data and metadata processing. FPGAs are starting to see greater adoption across HPC, especially for applications that are text- or integer-based and highly repeated at scale: a perfect description of an I/O server's workload. BlueArc has programmed NFS protocol acceleration into FPGA hardware, providing a dramatic speed-up of a task that was already being moved off of the compute cluster CPUs and into storage I/O servers anyway, so there is no latency penalty to computation.

BlueArc reasoned that the vast majority network traffic that surrounds NFS implementations could benefit from FPGA acceleration. Examples include network protocols, metadata communication among clustered Titans, and common I/O calls such as *getattr* (get attribute) and *rw* (read/write). Only a few features such as quotas (a file system option that allows administrators to budget disk allocations) would not benefit materially from acceleration and were therefore left alone.

Not exclusively an NFS server, Titan accelerates CIFS traffic as well. BlueArc's FPGA implementation simultaneously preserves both CIFS and NFS presentations, under a single namespace to many petabytes if desired, so that hosts see what appears to be a scalable yet standard Windows or NFS server.

BlueArc Data Management Features

BlueArc Titan can manage disparate physical storage devices as a single, virtualized storage pool. The BlueArc file system (the back end that sits underneath the CIFS and NFS protocols) is designed to provide distributed access over a range of virtualized devices regardless of storage access protocol. The Titans and the data volumes exported to the cluster can also be further virtualized themselves, creating what appear to be independent domain spaces or distinct file systems among shared storage resources. This is particularly useful when data or computational spaces need to be separated between departments or teams for accounting or security purposes.

Because of the protocol independence, BlueArc storage solutions can be integrated into a wide range of IT infrastructures. In addition, the storage has a "future-proof" aspect, because it can be adapted to new protocols that are introduced over time. Common protocols that are accelerated by the FPGAs in the hardware include NFS, CIFS, iSCSI, NDMP, and TCP/IP, among others.

BlueArc offers several other enterprise features with its storage solutions to complete the data management environment:

- *Virtualized storage:* Although BlueArc uses a parallel file system, Titan employs SAN hardware, controllers, and protocols to virtualize access and offer block-level features to the storage administrator. The BlueArc virtualization engine gives a single administrative view of disparate pools of storage, enabling SAN functionality in a NAS environment.

- *High availability:* Multiple Titan servers can access common data in a virtualized pool with true N-way failover capabilities for clustered implementations.

- *Global namespace:* Disparate storage volumes and physical storage devices can be presented with a common global namespace to provide a single administrative view. BlueArc's global namespace implementation scales to several usable petabytes and integrates both NFS and CIFS environments.

- *Replication:* For replicating data either remotely or locally, BlueArc provides both file-level and block-level solutions. The file-level implementations are policy-based; administrators can do either full or incremental asynchronous replication. The block-level implementations are designed to allow integration with other storage vendors' solutions, including synchronous mirroring and long-distance block replication.

- *Compliance adherence:* In some cases legislation (such as HIPAA in the medical community) requires unaltered data to be kept for long periods of time. For these situations BlueArc offers its Write-Once Read-Many (WORM) file system option, which prevents data from being deleted or altered until after a set date.

- *Data migration:* BlueArc has product history in multi-tiered storage and brings this heritage forward in the Titan line. BlueArc's Data Migrator product provides data lifecycle management capabilities for hierarchical storage management (HSM) environments, balancing the data as required between higher-performing and lower-cost devices while maintaining a consistent data presentation to the hosts.

- *Cross-volume links:* Cross-volume linking is BlueArc's method for extending the reach of the Data Migrator product to third-party storage. While Titan maintains a consistent file system presentation to the hosts, and Data Migrator provides policy-based movement of data, cross-volume links allow the BlueArc filesystem to interface with de-duplication, encryption, compression, and indexing appliances. Cross-volume links also integrate with third-party storage products such as virtual tape libraries, HSM-to-tape file systems, and even competing storage solutions from other vendors.

- *Read caching:* Not to be confused with caching accelerator products that sit in front of or alongside a storage solution, Titan's read caching is actually a separate file system unto itself which maintains a copy of recently accessed data. Each Titan has its own read caching file system. In this way each Titan in a clustered implementation – usually with a global namespace – has its own copy of recently accessed files, and it can respond quickly to multiple hosts' requests to read data. (For single Titan implementations, the read caching filesystem still caches one Titan's files.) In HSM environments, read caching solves the problem of reverse data migration; there is no need to find and restore a migrated file that becomes needed again. Any file that is recently accessed is by definition copied to the read caching filesystem for subsequent access.

BlueArc's goal is to design a scalable, high-performance storage solution that can still be easily managed. This feature set is indicative of BlueArc's efforts to combine benefits of distributed, shared disk, and parallel file system implementations with interfaces that will be familiar to IT administrators. Done successfully, this approach can solve the scalability vs. manageability tradeoff.

TABOR RESEARCH ANALYSIS

BlueArc storage solutions are significant for the HPC industry, in that BlueArc is taking a unique approach to combine scalability with simplicity. Rather than casting off NFS as inherently limited, BlueArc has taken on the challenge of simply making NFS more scalable.

A scalable, NFS-compatible solution will be a welcome relief to many users who lament that scalable storage is necessarily non-standard. BlueArc naturally has done a great deal of customized engineering to make its solutions scale and perform, but this work is generally made transparent to the end user. In a perfect implementation, users could be unaware that there is any FPGA-based protocol acceleration; they only know that data is accessed more quickly than with other solutions.

BlueArc is also notable for its combination of elements of all three filesystem schemata: distributed, shared disk, and parallel. The BlueArc architecture maintains a distributed presentation to hosts using standards-based protocols like NFS, uses SAN components in a shared disk style, and contains a high degree of object-oriented parallelization as well. FPGA acceleration notwithstanding, the BlueArc file system stands out in its ability to cross traditional file system boundaries. With pNFS still on the horizon, BlueArc is ahead of the curve in combining features of previously exclusive file system implementations.

pNFS

pNFS is one of the most significant new technologies currently in development in HPC. pNFS has the promise to enable parallel file systems with "standard" NFS, thereby undercutting the need for custom file systems over time. Although non-NFS parallel file systems certainly will not be eradicated, Tabor Research predicts pNFS will bring about some consolidation in the cluster file system market.

Assuming the pNFS standard is successful, older non-parallel NFS versions may eventually age out. Commonality in parallel file systems would push shared-disk implementations further from the norm. In addition, current parallel file systems will find their main points of difference are beginning to erode. Over time fewer users might choose to abandon the NFS standard in order to achieve greater performance. pNFS could be therefore become a natural evolution path for current distributed, shared-disk, and parallel file systems.

The emergence of pNFS increases the relevance of BlueArc's approach. The BlueArc solution offers a smooth transition while still providing a strong performance path for storage architects who choose to delay or avoid pNFS adoption. In other words users can take advantage of pNFS when it is ready, but they are not committed or forced into it if it is not. Furthermore, the parallelism of pNFS only solves part of the scalability problem. BlueArc's protocol acceleration will continue to provide performance benefits over other NFS servers. Whenever pNFS is ready for the HPC market, BlueArc will be in position to take advantage of it.

Future Outlook

The primary challenge for BlueArc is gaining attention in a convoluted market. The current market dynamics have large system vendors, large storage vendors, and smaller, specialized providers such as BlueArc all hawking their own data management solutions, with improvements to NAS, SAN, HSM, TCP, or any letter you like followed by FS, all coming out as an alphabet soup that can be daunting even to a seasoned IT professional. Confusion in the market benefits larger providers with established relationships. An HPC user buying clusters from IBM or storage from EMC is likely to stay with data management solutions from those incumbents. It is therefore difficult for a company such as BlueArc to gain in awareness, regardless of the merits of its products.

In the near term, BlueArc provides a scalable NFS-based solution that will be attractive to many users. Over the long term, BlueArc has a breakout opportunity in the potential consolidation of the HPC data management market. Some HPC system vendors do not currently have clear or cohesive data management stories. If BlueArc is perceived as a strong, safe option, the potential partnerships with these system vendors could drive huge gains for BlueArc. BlueArc's openness regarding protocols and storage types remove potential hindering factors for these sorts of partnerships.

BlueArc does have an established track record of success in both traditional and edge HPC markets. While not as well established as HPC-specific filesystems at larger HPC sites, BlueArc does have many well-known and well-respected customers in specific vertical markets, most notably ones where large-scale data challenges are part of the business cycle. BlueArc solutions can currently be found in traditional HPC application areas (life sciences, oil & gas exploration, engineering design and simulation, digital media) as well as a few emerging edge HPC environments (most notably Internet services). This experience shows BlueArc's potential to cross over from traditional HPC to edge HPC markets with solutions that appeal to both sets of customers.

Regardless of the potential of pNFS and consolidation to act as catalysts for BlueArc, there is a clear market need today for scalable storage solutions, and BlueArc brings compelling differentiation with accelerated NFS. For many HPC applications scalable data management – getting everything from memory to disk or from disk to memory as quickly as possible – is the main driver of productivity. HPC users who have been frustrated at having to choose between a scalable solution and a standards-based one may be pleased to find they can have both.