The State of pNFS: The Parallel File System Market in 2011

Addison Snell

March 2011

White paper

EXECUTIVE SUMMARY

Intersect360 Research has tracked the evolution of hardware and software technologies for data management over the past four years, throughout the emergence of parallel file systems for scalable high performance computing. Historically file systems have been an area with a single, broad standard – NFS – and myriad non-standard options for greater performance. In 2011 the inchoate market for parallel file systems has led to some early consolidation onto GPFS and Lustre. We predict that storage consolidation will accelerate, and additionally that parallel file systems will be adopted by more end users, due to the maturation and availability of Parallel NFS (pNFS).

The appeal of pNFS is twofold. For the IT administrator, pNFS will be a familiar environment, even when it is new. NFS is ubiquitous in Linux environments, and pNFS is a straightforward evolution of NFS. pNFS will therefore have a broad end-user appeal to those who want greater data management scalability. Because it works in conjunction with proprietary backend file systems to project an NFS environment, pNFS can be integrated into most product offerings, although differentiated offerings will go beyond simply using it as a unified protocol to offer optimized performance. pNFS therefore does not short-circuit or diminish existing intellectual property or competitive advantage. It is a tool for bringing proprietary advantages to more end users without forcing them to sacrifice a standard.

Although the development of pNFS seemed to stumble in the second half of 2009, by the end of 2010 the momentum had returned, with several contributing factors:

- BlueArc demonstrated a working pNFS implementation in its booth at SC10, the leading annual supercomputing conference, in November 2010.
- The roadmap for Lustre was muddied by Oracle's acquisition of Sun Microsystems and the subsequent fallout from Oracle's apparent lack of focus on HPC technologies.
- Other key vendors, such as Microsoft, have joined the pNFS effort.

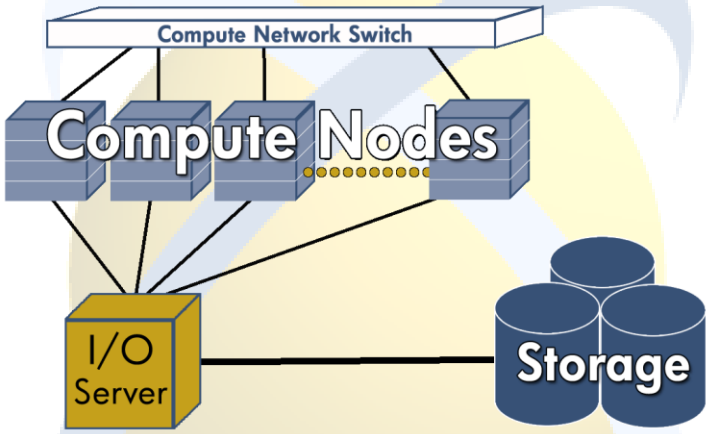
With pNFS becoming generally available, we expect to see some expansion of scalable HPC technologies into new use cases. pNFS will have significant appeal in introducing parallel file systems into environments where none existed before, and many experienced users of parallel file systems will also find value in simplified and standardized implementation and management versus competing technologies. pNFS could additionally become a key technology in addressing the advanced technology gap that has been described as hindering innovation for small and medium-size U.S. manufacturers. pNFS is therefore one of the most important new technologies for 2011.

MARKET DYNAMICS

Parallel File Systems

The central I/O technology supporting any computational infrastructure is the file system, which is responsible for keeping track of where data resides in storage, logged by simple monikers like file names. In clustered environments, the most basic implementation is a network file system, often called a "distributed file system" for certain implementations. NFS (for Network File System) is the aptly named network (or "distributed-access") file system that is standard for UNIX and Linux environments. (CIFS is the predominant network file system for Windows.) In a network file system implementation, I/O requests are essentially routed from the cluster through a designated I/O server (sometimes called a "central file server" or "filer head"). [See Figure 1.]

Figure 1: Diagram of a Network File System Implementation
Source: Intersect360 Research



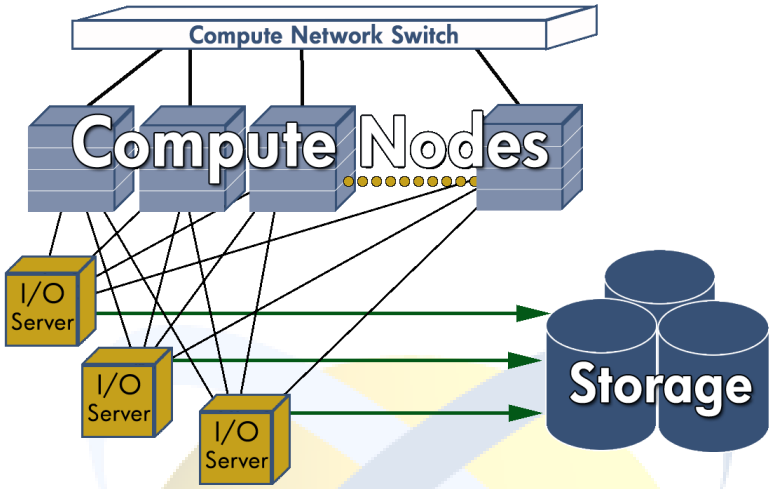
Network file systems are straightforward and relatively easy to manage. Furthermore, NFS is a widely used standard, and administrators' knowledge of it is easily portable from one system to another. On the other hand, the scalability of NFS or other distributed-access file systems is necessarily limited. For larger systems with greater I/O needs, the I/O servers can quickly become bottlenecks.

Today this scalability problem can be addressed by introducing a "parallel file system," which uses multiple I/O servers, each equally serving all nodes in a cluster and jointly accessing all of the data in storage. By providing multiple, parallel access routes to data in a single name space with distribution of the data across multiple nodes, parallel file systems can deliver dramatic increases in scalability and throughput over network file systems. [See Figure 2.]

Until very recently, the unfortunate drawback to parallel systems has been a lack of standards. Each high-end storage vendor would have its own proprietary solution – some more successful than others – and these non-NFS paradigms would add to the administrators' burden. The result in many cases was poorly optimized or even abandoned data management implementations.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

Figure 2: Diagram of a Parallel File System Implementation
Source: Intersect360 Research



GPFS and Lustre

By the end of 2009, two parallel file systems had begun to emerge as relative leaders ("standards" would be too strong a term), GPFS from IBM, and Lustre from Sun Microsystems. The roadmap for GPFS is stable. It is a supported product that is sold by IBM and other storage vendors.

The path forward for Lustre is not so clear. Sun originally acquired Lustre from its original designer, Cluster File Systems Inc., in 2007. After Oracle subsequently acquired Sun, many HPC users were concerned about future development of Lustre, which Oracle stated it would support only on Oracle hardware. Responding to that concern, a consortium of both users and vendors have rallied together as OpenSFS, an HPC-focused organization focused on developing ongoing features of the open-source Lustre distribution. In addition, Whamcloud has emerged as one commercial organization offering multi-platform Lustre support.

Heading into 2011, Lustre currently leads the parallel file system category in installations, followed by GPFS.[1] Intersect360 Research believes that over the next few years, Lustre will continue to be popular in open-source-oriented user environment, especially academic and government research labs with existing in-house Lustre expertise, but potential fragmentation of the roadmap and the support community will continue to be a concern for many users. GPFS will continue to have some success in commercial environments, but its breadth of appeal will necessarily be limited by the proprietary nature of the offering.

[1] Intersect360 Research HPC Market Advisory Service: "HPC User Site Census: Storage," December 2010.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

LOOKING FORWARD: THE POTENTIAL OF PARALLEL NFS

The State of pNFS in 2011

Since the mid-2000s there has been a multi-vendor industry initiative to develop pNFS (for "Parallel NFS"), a set of NFS extensions, built into the NFS standard, that will allow NFS to act as a parallel file system. The first pNFS marketing campaign was launched by Panasas at the International Supercomputing 2007 (ISC'07) Conference and Expo in June 2007, and pNFS development has included Panasas, BlueArc, EMC, NetApp, Oracle, Microsoft, and other vendors, as well as open-source efforts to integrate pNFS clients into Linux distributions.

There are reasons for anticipation. pNFS has appeal for both the user and vendor communities. For users, pNFS represents the ability to deploy a scalable, parallel file system without having to give up NFS standards, saving on both cost and administrative effort. For vendors, pNFS provides a way to deliver a standard interface to their own proprietary enhancements, because pNFS can be layered on top of backend solutions that then become transparent to the end user. The promise of pNFS is to bring more scalable I/O throughput solutions – and in some cases HPC in general – to new categories of users.

network file system: Provides data access to clusters through selected I/O servers. NFS is the most common network file system. *Advantages:* Generally cost-effective, standards-based, with low administration costs. *Disadvantages:* Most implementations are limited in scalability and performance.

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.

GPFS: A proprietary parallel file system developed by IBM and sold by IBM and some other storage vendors. *Advantage:* Developed and supported by IBM. *Disadvantage:* Proprietary environment unfamiliar to many administrators.

Lustre: A parallel file system previously maintained by Sun Microsystems and then acquired by Oracle, which going forward will support Lustre on Oracle hardware only. Grass-roots consortia have formed to shepherd ongoing development and support for Lustre for HPC. *Advantage:* Open-source efforts. *Disadvantage:* Uncertainty in ongoing roadmap.

pNFS: Parallel NFS, an emerging implementation of NFS that is a multi-vendor standard. pNFS has the potential to become a transition platform for both distributed and parallel file systems. *Advantages:* Scalable, parallel file system that enables greater scalability and throughput over NFS. Extension of standard NFS environment that will be familiar to most storage administrators and therefore easier to adopt and maintain than other parallel file systems.

From mid-2009 to mid-2010 development of pNFS seemed to slow, but Intersect360 Research attributes this perception to the normal delays associated with merging a new standard into an existing open source tree. These garden-variety complications were exacerbated by each vendor's specific requirements. (For example, one vendor might be seeking a file-based implement, while others are block-based or object-oriented.) By late

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

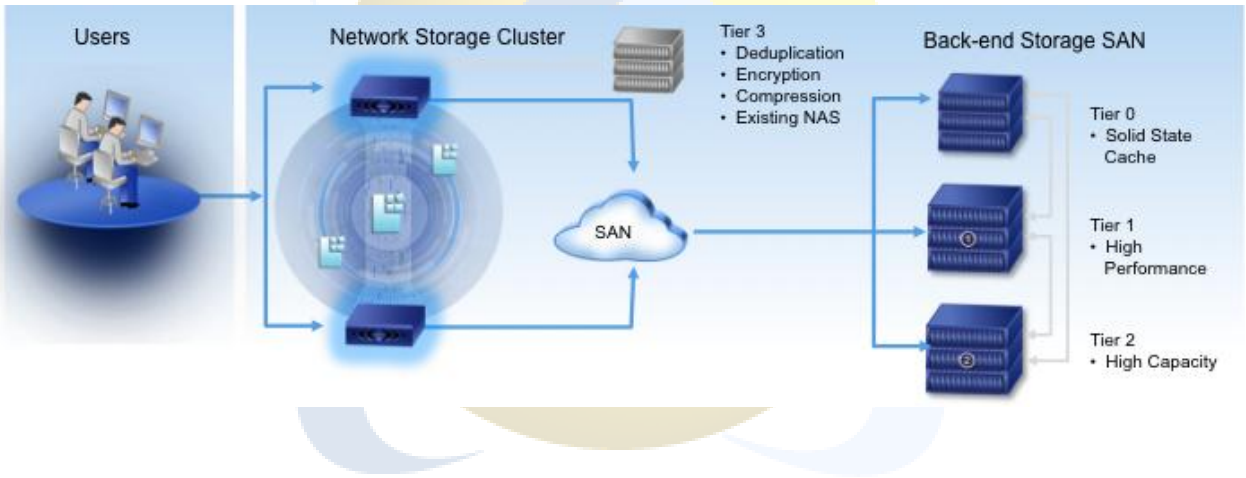
2010, pNFS had been committed to be merged into the Linux kernel as part of the NFS 4.1 standard. The table is set for pNFS implementations to begin in 2011.

BlueArc and pNFS

BlueArc has been a supporter of pNFS and a participant in its development since its early stages, and for the past two years has taken a leadership role in promoting it. BlueArc coordinated "Birds of a Feather" sessions at the past two Supercomputing conferences, and at SC10 in November, BlueArc ran demonstrations of pNFS in its booth.

While many vendors are planning to incorporate pNFS into their offerings, BlueArc has traditionally differentiated itself with NFS optimizations, and the company is therefore justifiably optimistic about the benefits it can deliver to users with pNFS. The current BlueArc storage server families are architected around making NFS and other common protocols more scalable. The internal BlueArc file system is object-oriented and multi-protocol, and the tiered-storage architecture provides built-in hardware acceleration of NFS and other common file system and network protocols, including CIFS, iSCSI, and TCP/IP.

Figure 3: Diagram of a BlueArc Storage Implementation
Source: BlueArc



While pNFS levels the playing field for BlueArc and other storage vendors by providing a standard administrative interface, it thereby accentuates the areas in which BlueArc has been providing differentiation. BlueArc's strategy of providing enterprise storage functionality – global name spaces, virtualization, high availability, read caching, data migration and replication, etc. – with an accelerated NFS stack, is enhanced by the addition of pNFS accessibility for administrators.

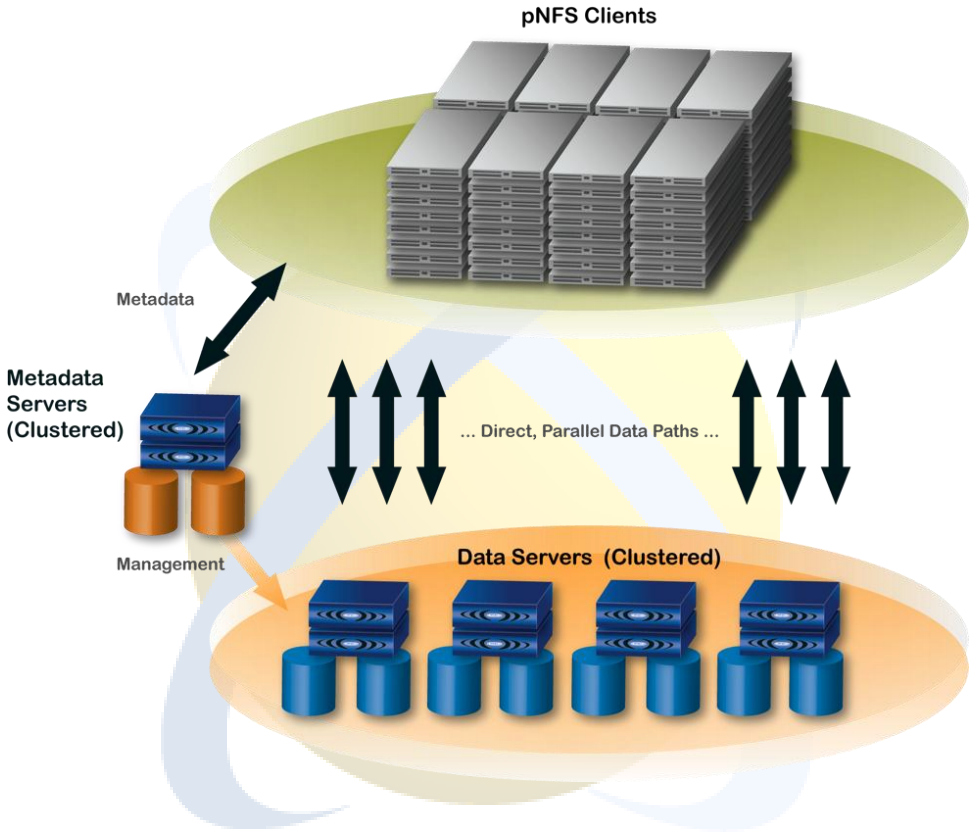
Furthermore, whereas with many implementations pNFS acts merely as a front-end extension to a custom parallel file system, BlueArc has taken it a step further. BlueArc has built a next-generation clustering architecture that is specifically optimized for pNFS. This strategy is consistent with BlueArc's history of embracing and enhancing the performance of a standard.

Because of the nature of BlueArc's hardware differentiation, the company has been well-positioned to compete in HPC. pNFS gives BlueArc another way to sell its advantages to more users, many of whom will value the

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

notion of a commercially supported, ready-to-run parallel file system solution. As pNFS becomes available, BlueArc will be among the first in line to promote it as an option with its current and future lines of high-performance storage products.

Figure 4: Diagram of Planned BlueArc pNFS Architecture
Source: BlueArc



Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

INTERSECT360 RESEARCH ANALYSIS

The potential impact for pNFS to have on the HPC industry should not be underestimated. While hard-core, dyed-in-the-wool supercomputer users are certainly discussing the ongoing open-source development of Lustre, pNFS should have a broad appeal. Academic users – many of whom have not taken the step toward parallel file systems because of management complexity – may see a reduced administrative burden relative to Lustre. Therefore pNFS is likely to find a home among both research and production environments in the broad market below the top few dozen supercomputers.

pNFS will have its greatest effect for commercial organizations that are striving to achieve greater performance while conforming to IT standards. In many of these cases, companies have continued to use distributed file systems beyond their optimal limits of cost or scalability. pNFS gives these users a pathway to greater scalability without introducing the unwelcome complexity of an unfamiliar file system.

At the boundary of adoption, pNFS will even play a role in helping introduce HPC technologies to new users. In 2011 this topic is gaining increasing attention in the U.S., with an emphasis on addressing the advanced technology gap that hinders innovation for small and medium-size manufacturers, fewer than 8% of which are using HPC technologies today.[2] Standards are a piece of the requirements for bringing HPC to this potential new set of users.

BlueArc has potential advantages here as well. The BlueArc storage server families accelerate CIFS traffic as well as NFS, and BlueArc will provide gateway access to pNFS from CIFS[3] (providing access to data but not a parallel performance boost in such an implementation). Microsoft has become involved in the pNFS effort as well, pursuing a Windows client for pNFS. Therefore we may very soon see Windows-based solutions that not only allow new users to scale their I/O infrastructures without an exotic file system, but also allow them to do it without introducing Linux into IT environments that are currently 100% Windows.

Finally there are ways in which pNFS can introduce new use cases for parallel file systems. The most common implementation of parallel file systems is for scratch space – i.e., high-performance disk that is serving up active data for an application that is currently running. (Within shared work environments it is typically important for scratch disk to offer balanced scalability, delivering either the bandwidth for relatively fewer large files or the throughput for relatively greater numbers of small files.) But because pNFS will be more familiar for many administrators to manage, some users will explore the idea of bringing the features of the parallel environments – such as greater scalability and a shared namespace – to adjacent areas such as home directories or shared repositories.

Intersect360 Research predicts that pNFS implementations will begin in 2011 for HPC environments, and that pNFS will be particularly valuable at first to midrange commercial organizations that are in the process of scaling their I/O systems. In the longer term, pNFS represents a significant potential nexus of consolidation, allowing IT administrators to centralize their now-disparate data management environments. While GPFS and Lustre will play an ongoing role among the users that have already adopted those technologies, pNFS is a positive development whose time has come, expanding the role of parallel file systems to organizations who will value both standards and scalability.

[2] Intersect360 Research, "Modeling and Simulation among U.S. Manufacturers: The Case for Digital Manufacturing," September 2010.
[3] This is described further in a BlueArc architecture white paper available at http://www.bluearc.com/pnfs.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com