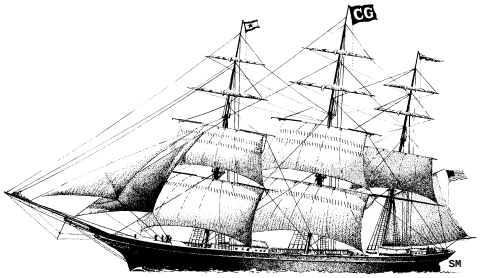
THE CLIPPER GROUP
Explorer™



Navigating Information Technology Horizons℠

Published Since 1993 Report #TCG2003003 January 30, 2003

Consolidated File Systems — Relieving the Pains of Scale in Data Storage

Analyst: Michael Fisch

Management Summary

When an enterprise grows, the world celebrates. Wall Street is pleased because revenues and profits are up. Management is satisfied because their plans have born fruit. Workers are enthusiastic because jobs and opportunity for advancement are plentiful. Customers like doing business with a "winner". **Growth is a good thing.**

But accompanying the chorus of celebration are moans and groans – **because growth is also painful.** Workers and managers alike are overwhelmed with additional activity. Informal procedures and human networks for getting things done become strained and burdened. Facilities are crowded. The old information systems now seem inadequate. **These are the pains of scale, and enterprises must adapt their methods and means to thrive at a larger size.**

These pains also apply to a small but critical component of an IT infrastructure – the file system. In essence, a file system stores and organizes information for easier access, sharing, and management. **Most enterprises use local file systems and/or network-attached storage (NAS) appliances for this purpose, but both can have limitations in large-scale environments.** The resulting "pains" include limited data sharing, unused capacity, disruptive expansion, and management complexities, including backup and restore.

An alternative, up-and-coming technology called *consolidated file systems* (or global/distributed) can solve these problems. It is designed for high scalability, concurrent access by multiple, heterogeneous servers, pooling multiple storage devices, and centralized management. The result is several worthwhile business benefits:

- **Lower storage operating costs** due to simpler management,
- **Lower storage acquisition costs** due to more optimal use of storage, and
- **Greater value from information** due to broader, easier access.

If your enterprise is experiencing pains of scale in data storage, a consolidated file system may provide relief. **Read on for a closer look at this promising technology, what its future holds, and how to find out if it is right for your enterprise.**

IN THIS ISSUE

- A Tale of Scale 2
- Purpose of a File System 2
- File System Technologies 2
- Business Benefits 5
- Gauging the Need: The Four S's of Scale .. 6
- A Promising Future 6
- Conclusion 7

A Tale of Scale

Growth is good, but it also brings challenges. An enterprise that is good (or lucky!) enough to grow may suddenly find itself operating in an ungainly and awkward manner. What worked in the early days must give way methods and means that are better suited to scale:

- Many "hats" or responsibilities held by a few give way to specialized departments and job descriptions,
- Ad hoc procedures for getting things done become structured, even documented, processes,
- Garages and small offices are replaced by larger facilities, perhaps with better amenities, and
- Basic information systems give way to more sophisticated, enterprise-class systems that can more effectively serve a large organization.

Simply put, requirements change with scale. Whether it is a building, a business process, or an information system, managers must adapt and restructure it to maintain effectiveness. **This is true even for a file system – a humble and sometimes overlooked technology component that is essential for storing information.**

Purpose of a File System

A file system is one layer in a storage infrastructure – the part of an information system that handles data storage. To understand the purpose of a file system, consider what the following items have in common:

- Address book (paper),
- Filing cabinet,
- Internet phone directory, and
- File system.

These are all tools for organizing information. **It is important to *access* information in a timely manner**, and thumbing through an alphabetized address book is much faster than rummaging through cluttered drawers and boxes (*or was it the pocket in my jacket?)* to find a business card or slip of paper. And when more than one person is involved, **the ability to *share* becomes important**. For instance, a filing cabinet can keep files on patients so anyone in a doctor's office may access them. Of course, ***security* becomes an issue when files are open for common access**. You wouldn't want private health or financial information falling into the wrong hands!

The last two items on the list, the Internet phone directory and file system, store data in digital form rather on paper. They are part of an electronic information system. An Internet directory is a singular, universal source for phone numbers, which similar to what a consolidated file system attempts to be (*more on this later*). The main point is that all of the items above serve the same conceptual purpose – organizing, accessing, sharing, and securing information.

The purpose of a file system is to organize digital information so that it is more easily and safely accessed, shared, and managed. Specifically, it:

- Presents data to applications and users,
- Distinguishes data by file name and type,
- Organizes files into a directory structure (logical location),
- Describes files by size, date created, last modified, and so forth, and
- Allows file sharing, locking (during an update), and rights management among multiple applications and users.

In short, a file system creates structure around raw data and makes it simpler and more meaningful to use (*see sidebar on the following page*). **Most applications and all users view data through a file system.**

File System Technologies

Local File System

A local file system is the first rung on the "ladder of scale". It provides file services principally for the server on which it resides. If an organization has only one

server, then a local file system presents no apparent scalability issues. It works well. However, if there are many servers, perhaps running different operating systems, then limitations become apparent:

- **Limitations of data sharing** – Each server has its own captive pool of files. Though it is possible for other servers to access them, it can be complex to administer[1], reduce application performance on the host server, and create update problems in heterogeneous environments.
- **Too much unused capacity** – Captive file pools also lead to too much unused capacity since each file system carries its own "overhead" of room to grow. The aggregate overhead of many file systems is much greater than one consolidated file system.
- **Difficult to manage** – Finally, each file system must be managed individually, which consumes time and administrative resources. The management tasks include backup and restore, capacity monitoring, planning, and expansion, replication for purposes of disaster recovery or data sharing, and other namespace administrative tasks, such as managing rights for file access. Multiply these by the number of file systems, compounded by different versions and operating systems, and it becomes clear why it would be much less labor-intensive to manage just one consolidated file system.

Network-Attached Storage Appliances

Moving up the ladder, network-attached storage (NAS) is the next step. **NAS is a *dedicated* appliance or server that makes files broadly-accessible over a local-area network (LAN).** Multiple, even heterogeneous servers and PCs can access files using standard protocols like CIFS for Windows and NFS for Unix. Some NAS appliances also offer advanced features that enhance availability and performance.[2] NAS improves over local file systems through more robust file sharing, scalability to potentially multiple TBs in a single appliance, and simpler management.

But NAS can run into scalability problems too. As the number of servers and users increase and as the size and number of files grow, multiple NAS appliances and file system instances are necessary. And eventually they cause pains of scale:

[1] Such as managing a web of mount points (which direct a file system to access files on a remote system) as well as access rights.

[2] Such as redundant components, failover, point-in-time copy, and remote mirroring.

File and Block Data – Storage Bedfellows

File and block data are both integral parts of a storage infrastructure. They are storage bedfellows, serving different but important roles.

Block-level data is comparable to the shipping containers in a cargo ship. The containers are standard-sized so they can be easily loaded, stacked, and unloaded. They are full of cargo – anything that someone might want to ship. However, one cannot tell at a glance what is inside because the containers are sealed and all appear the same. Likewise, block data comes in discrete chunks and appears nondescript without higher-order intelligence to interpret and assemble it. **Block data represents the "raw" building blocks of storage – the physical foundation upon which a storage infrastructure is built.**

To become useful, the shipping containers must be unpacked and their contents organized, such as on the shelves of a store or warehouse. Then people can then see and use the goods. **A file system serves a similar purpose by organizing block data and making it distinct, defined, shareable, and ultimately more meaningful.** (Databases can perform a similar function with block data, though sometimes they also run on top of file systems.)

So it is not a question of file or block, NAS or SAN. You need both, and one is built on top of the other. **The question is: What is the favored point of data access for applications and administrators?** Do you deal in files or blocks? As this bulletin explores the virtues of consolidated file systems, a strong case emerges for the humble, ubiquitous file.

- **Unused capacity** – The NAS appliances themselves become islands of data, each with its own spare capacity. Some appliances may even contain multiple file systems due to size limitations (e.g., 1 TB). As with local file systems, an unnecessary amount of overhead capacity can result.
- **Management complexities** – The same difficulties of managing multiple file systems and appliances can occur, like backup and namespace administrative tasks. Administrators must also deal with a plethora of mount points for directing remote file access on clients.
- **Performance concerns** – A NAS appliance can become a bottleneck to file access speeds, such as when it has scaled to the high end of its capacity. This can affect application performance.

Both local file systems and NAS work very well to a point, but limitations become apparent and even painful *as data requirements and IT infrastructure scales.* These limitations result in a higher total cost of ownership (TCO) for storage as well as a potential impact on the applications, workers, and business processes that depend on fast, consistent file access.

Consolidated File Systems

Relief for the pains of big scale can be found in a *consolidated file system*. There are many adjectives used to describe this file system category: "global", "distributed", "clustered", "universal", "shared". However, the distinctions and nuances between these terms is not entirely clear, so we chose to coin our own phrase: consolidated file system. The descriptor "consolidated" is widely used and understood in the context of server and storage consolidation. **It describes aggregating resources (storage, processing) to deliver better scalability, simpler management, higher utilization, and ultimately a lower TCO.** The same can be said of a consolidated file system.

The primary characteristics of a consolidated file system are:

- **Highly scalable** – capable of scaling to the range of 10s of TBs to 1000s of PBs,
- **Centralized management** – a single management console with one logical view of the file system,
- **Single pool of data –** spans and allocates files across multiple storage devices, and
- **Broad access –** Concurrent read/write access by multiple, possibly heterogeneous servers.

The main point is that a consolidated file system is capable of handling *big scale*. It can contain a large amount of data and span many computers and storage systems. As it scales, it brings management efficiencies and minimally- or non-disruptive expansion. Another way to think about a consolidated file system is in terms of file virtualization – analogous to the block virtualization technologies that span multiple storage arrays.

Furthermore, depending on the particular product or technology, there are a number of possible secondary characteristics:

- **Single namespace** – possibly with the ability to create logical partitions,
- **Wide-area geographical reach,** possibly with caching to improve remote performance,
- **High availability –** such as redundant servers or controllers,
- **Rapid recovery –** such as through journaling or metadata separation,
- **High performance,**
- **Point-in-time copy,** and
- **Tiered storage classes**[3] – the ability to place files among multiple quality-of-service tiers or media types (e.g., online, nearline, offline).

The secondary characteristics can add to the value of a solution, **but none are necessary**

[3] See *Tiered Storage Classes Save Money – Getting The Most Out Of Your Storage Infrastructure* in **The Clipper Group Explorer** dated August 29, 2002, at www.clipper.com/research/TCG2002030.pdf.

to qualify as a consolidated file system, per se. A variety of products in this category are available now, and several new or enhanced ones are expected to roll out in 2003. The box on the right describes various consolidated file system architectures.

Business Benefits

At the end of the day, the reason for deploying any new technology is to support and benefit the overall business. In this context, consolidated file systems can deliver:

- **Lower operating costs due to simpler management.** The general principle here is straightforward: it is easier to manage one of something than many. In this case, one consolidated file system is much easier than many, discrete, and possibly different file systems. Less training and fewer people are required. It provides a single point to act upon. Data services and administrative tasks can be leveraged over the whole file system, including backup and restore, virus scanning, content indexing, capacity monitoring, file system expansion, data migration and replication, lifecycle data management across storage tiers, storage resource management[4], and automation. The mount points on clients are fewer, simpler to manage, and do not change as the system scale. Lowering operating costs is very significant because they are by far the largest component of storage TCO.
- **Lower acquisition costs due to more optimal use of storage.** A consolidated file system eliminates replication and redundant files through sharing, so less capacity is required. Furthermore, it is possible to attain a higher utilization of capacity with a single file "pool" versus many discrete "islands". Both benefits help lower upfront storage procurement costs, also contributing to lower TCO.
- **Greater value from information due to broader, easier access.** The real purpose for storing information is to use it to make smarter decisions and better manage the business. If a consolidated file system makes it easier and perhaps faster for users and applications (that are duly authorized) to access the right information, then the business itself can be more productive and competitive.

Finally, all of these can contribute to the bottom line, which is what makes them true *business benefits*.

Consolidated File System Architectures

There is more than one way to consolidate a file system, so to speak. Approaches that vendors use today include:

Aggregated NAS Appliances

In this approach, multiple NAS appliances are united to act as a single file system. The "aggregator" can be a front-end platform or a distributed file system that runs on the NAS appliances (or possibly on the clients). The individual appliances may include their own disks or they may be NAS heads or gateways that access block storage on a storage-area network (SAN).

SAN File System

Here, servers access files directly over a SAN. Information about the files such as locking and location (i.e., metadata) can reside in a dedicated server connected to the LAN. In this case, a client would first request metadata via the LAN then access the file contents directly over the SAN. Alternatively, clients may access both metadata and file data directly over the SAN.

Clustered File System

In this approach, multiple servers in a cluster share access to the same data through the framework of a clustered file system. The purpose is to allow multiple servers to act as a single, dynamic, highly-available computing entity, and the clustered file system is an essential technology to enable it. This is similar to a SAN file system but is for the narrower purpose of server clustering. Scalability is typically more limited due to heavy inter-processor communication.

[4] See *Storage Resource Management – Conducting a Symphony of Storage* in **The Clipper Group Explorer** dated June 30, 2002, at www.clipper.com/research/TCG2002024.pdf.

Gauging the Need: The Four S's of Scale

So the technical benefits of a consolidated file system can bring value to an enterprise. The next question is: Do you need one? There are many helpful technologies on the market, so is a consolidated file system what your particular enterprise needs? To help answer this question, **consider how *the four S's of scale* apply to your environment**:

- **Storage capacity** – Is there a large amount of data contained in the file systems – whether due to the size or number of files? Does the data span multiple storage arrays or NAS appliances? **Higher capacity requirements may justify a consolidated file system.**
- **Sharing of data** – Do many, possibly heterogeneous systems require access to the files? Is high performance a requirement? **Broad and/or high-speed sharing requirements can point to the need for a consolidated file system.**
- **Simplicity of administration** – Is there a perceived need to simplify file system administration in order to save on operating costs? Is too much time spent doing it, or is it hard to find skilled people who can? Looking forward, do you expect data growth to surpass the ability of the IT staff to administer it? **Making file system management less labor-intensive is the most important reason to implement a consolidated file system.**
- **Singularity of data** – Finally, is there a requirement to maintain only single instances of files rather than multiple? For instance, an enterprise may want to conserve capacity by avoiding or minimizing data replication. Or it may take too much time or resources to replicate files so different systems can access them in a particular workflow process (e.g., files are large and sequential). **The requirement to minimize file instances, while preserving broad access, is also a reason.**

Each of the four S's contributes to the need for a consolidated file system, and a strong requirement in one area alone may justify it.

Certain industries and applications are more likely than others to have these sorts of requirements, based on the nature of their computing requirements. For instance, film editing deals with very large files and multiple systems operating on the same data. Specific examples include:

- Video / film editing and graphics,
- Seismic analysis (oil & gas),
- Drug modeling and analysis (pharmaceutical),
- Financial modeling (financial services),
- Other scientific applications that require high-performance computing (physics, astronomy),
- Data mining,
- Server clusters,
- Grid computing[5], and
- Web server farms.

A Promising Future

The day is soon coming when consolidated file systems will be mainstream technology. Though deployments today are typically for applications with special requirements, it will eventually become a core, standard layer in the IT infrastructure, especially among larger enterprises and large-scale data centers. Several long-term trends are driving toward this future.

The first is greater market awareness coupled with steady technological advancements. Many vendors, both established and startup, are making R&D investments in consolidated file system products. It has been pegged as an important area of innovation, and the industry will deliver a steady stream of refinements. At the same time, vendors will invest in market awareness and education to promote the products. **The combination of broader awareness and improved product**

[5] See *Computational Grids – Server Consolidation for a Distributed, On-Demand World* in **The Clipper Group Explorer** dated June 21, 2002, at www.clipper.com/research/TCG2002021.pdf.

sophistication will help the technology gain traction.

The second trend is an ever-rising tide of information that will push enterprises across the threshold of "large-scale". Fast data growth is a byproduct of the Information Age. The Internet, enterprise applications, e-business, rich media, reference information, and so forth all contribute to this digital deluge. Even when business growth slows, growth of data still occurs, like a rising tide that never recedes. As a result, more enterprises will face the pains of scale that consolidated file systems can help relieve.

Finally, storage is following an evolutionary path toward greater cost-effectiveness. **Through its history, this industry has constantly looked for the next lever of efficiency to pull that would lower storage TCO.** Witness the major shift from direct-attach to more cost-effective, networked storage. Today, there is increasing interest in centralized storage management, automation, and intelligent storage networking. These technologies also promise even better TCO. Faced with spiraling data requirements, enterprises need storage to cost less, and vendors must continue to innovate to meet this rising requirement.

This is where the file system will play an important role in the ongoing development of storage. **It is a logical point for consolidating, aggregating, or virtualizing data – however you wish to put it – and squeezing more efficiency out of the storage infrastructure.** Compared to blocks, files present a higher level of abstraction, mask complexity, and are simpler to manage. They provide more contextual information and finer granularity for storage and data management applications to exploit. They offer the ability to share, eliminating redundancy and broadening data access. Then comes the consolidated file system with the scope to apply these characteristics broadly, across an entire storage infrastructure. The result is major leverage on storage TCO.

All of this makes a strong case for the broad adoption of consolidated file systems in the future evolution of storage, as it drives down TCO. It is also the reason why files, not blocks, will increasingly be the favored point of access for applications and users and the preferred point of management for administrators. **It's all about leverage, and files systems have it.**

Picture a sphere that contains all of an enterprise's information. The outer surface is what everything sees. It's the point of access. **For many – especially larger enterprises and large-scale data centers – the surface of that sphere will be a consolidated file system that abstracts, simplifies, and enriches everything inside.**

Conclusion

In short, consolidated file systems offer a more cost-effective and manageable way to access data as requirements reach a certain point of large scale. If your enterprise is experiencing the pains of scale, look closely at your data requirements. The four S's – storage capacity, sharing of data, simplicity of administration, and singularity of data – are helpful indicators to determine if a consolidated file system may be at least par of the solution you need.

Furthermore, this is a technology to watch. You may find that your enterprise grows into it – or it grows into your enterprise. **In any case, expect consolidated file systems to play an increasingly strategic role in information storage because they help take the pain out of scale.**

About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Michael Fisch **is Director of Storage and Networking with The Clipper Group.** He brings over seven years of experience in the computer industry working in sales, market analysis and positioning, and engineering. Mr. Fisch worked at EMC Corporation as a marketing program manager focused on service providers and as a competitive market analyst. Before that, he worked in international channel development, manufacturing, and technical support at Extended Systems, Inc. Mr. Fisch earned an MBA from Babson College and a Bachelor's degree in electrical engineering from the University of Idaho.

- ***Reach Michael Fisch via e-mail at Mike.Fisch@clipper.com or at 781-235-0085 Ext. 25.***

Regarding Trademarks and Service Marks

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.