Compliments of IBM

IBM Platform Computing Edition

Software Defined Storage FOR DUMMIES®

A Wiley Brand

Learn to:

- Control storage costs
- Eliminate storage bottlenecks
- Use IBM GPFS to solve storage management challenges

Lawrence C. Miller
CISSP

Scott Fadden

Software Defined Storage For Dummies®

A Wiley Brand

IBM Platform Computing Edition

by Lawrence C. Miller, CISSP
and
Scott Fadden

For Dummies®
A Wiley Brand

Software Defined Storage For Dummies®, IBM Platform Computing Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
`www.wiley.com`

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom For Dummies book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact `info@dummies.biz`, or visit `www.wiley.com/go/custompub`. For information about licensing the For Dummies brand for products or services, contact `BrandedRights&Licenses@Wiley.com`.

ISBN: 978-1-118-84178-5 (pbk); ISBN: 978-1-118-84314-7 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Table of Contents

Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. For details on how to create a custom *For Dummies* book for your business or organization, contact `info@dummies.biz` or visit `www.wiley.com/go/custompub`. For details on licensing the *For Dummies* brand for products or services, contact `BrandedRights&Licenses@Wiley.com`.

Some of the people who helped bring this book to market include the following:

Project Editor: Carrie A. Johnson

Acquisitions Editor: Connie Santisteban

Editorial Manager: Rev Mengle

Business Development Representative: Christiane Cormier

Custom Publishing Project Specialist: Michael Sullivan

Introduction

The rapid, accelerating growth of data, transactions, and digitally aware devices is straining today's IT infrastructure and operations. At the same time, storage costs are increasing and user expectations and cost pressures are rising. This staggering growth of data has led to the need for high-performance streaming, data access, and collaborative data sharing.

If you work with big data in the cloud or deal with structured and unstructured data for analytics, you need software defined storage. Software defined storage uses standard compute, network, and storage hardware; the storage functions are all done in software, such as IBM GPFS, that provides automated, policy driven, application aware storage services, through orchestration of the underlining storage infrastructure in support of an overall software defined environment.

IBM General Parallel File System (GPFS) provides online storage management, scalable access, and integrated information lifecycle management tools that are capable of managing petabytes of data and billions of files. This high-performance, shared-disk file management solution offers fast, reliable access to a common set of file-based data.

About This Book

Software Defined Storage For Dummies, IBM Platform Computing Edition, examines data storage and management challenges and explains software defined storage, an innovative solution for high-performance, cost-effective storage using IBM's GPFS.

Foolish Assumptions

It's been said that most assumptions have outlived their uselessness, but I'll assume a few things nonetheless! Basically, I assume that you know a little something about storage technologies and storage management challenges. As such, this book is written primarily for technical readers and decision makers, such as storage administrators and IT managers.

Icons Used in This Book

Throughout this book, you'll occasionally see special icons to call attention to important information. No smiley faces winking at you or any other cute little emoticons, but you'll definitely want to take note! Here's what you can expect.

REMEMBER

This icon points out information that may well be worth committing to your nonvolatile memory, your gray matter, or your noggin — along with anniversaries and birthdays.

TECHNICAL STUFF

You won't find a map of the human genome or the blueprints for IBM's Watson here (or maybe you will, hmm), but if you seek to attain the seventh level of NERD-vana, perk up. This icon explains the jargon beneath the jargon and is the stuff legends — well, nerds — are made of.

TIP

Thank you for reading, hope you enjoy the book, please take care of your writers. Seriously, this icon points out helpful suggestions and useful nuggets of information.

WARNING!

Proceed at your own risk . . . well, okay — it's actually nothing *that* hazardous. These helpful alerts offer practical advice to help you avoid making potentially costly mistakes.

Chapter 1

Storage 101

In This Chapter

- Recognizing data access and management challenges
- Knowing the basics of what storage does
- Understanding different types of storage
- Distinguishing between different storage technologies
- Looking at cluster file systems

Enterprise users and data consumers are churning out vast amounts of documents and rich media (such as images, audio, and video). Managing the volume and complexity of this information is a significant challenge in organizations of all types and sizes as more and more applications feed on and increase this deluge of file-based content, and as individuals and businesses collaborate for intelligence, analytics, and information sharing.

This chapter provides some background on digital storage technologies and the storage industry. We start by quickly summing up storage industry challenges, and then you take a look at what function storage serves in your IT infrastructure. Next, you look at the hardware nuts and bolts of storing digital data (don't worry, it's not too geeky) and common software approaches to organizing data. You finish up with a look at the beginnings of software defined storage — scalable file system storage.

Data Access and Management Challenges

Digital data — both structured and unstructured — is everywhere and continues to grow at a stunning pace. Every day, approximately 15 petabytes of new information is generated worldwide, and the total amount of digital data doubles approximately every two years. At the same time, storage budgets are increasing only one to five percent annually, thus the gap between data growth and storage spending is widening (see Figure 1-1). The data growth explosion, as well as the nature and increasing uses of data, are creating tremendous data storage challenges for enterprises and IT departments everywhere. Simply put, storage needs to be less expensive in order to keep up with demand.

REMEMBER

Structured data refers to data that's organized, for example, in a database. *Unstructured data* refers to data that doesn't have a defined model or framework — for example, multimedia files.

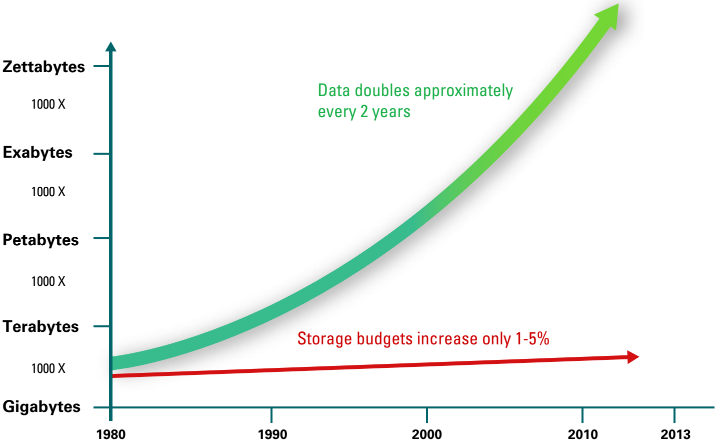

Figure 1-1: Storage requirements are devouring CAPEX and OPEX resources.

Some of this growth can be addressed with larger hard disk drives and networking components getting faster and faster. But as these technologies advance, making the data useful becomes more difficult. Larger hard disk drives enable you to store more data, but in many cases the hardware appliances that utilize these drives aren't able to keep up.

WARNING!

To address this problem, many IT managers today are adding more and more network-attached storage (NAS) devices. NAS devices can be relatively cost effective to purchase, but growing your file storage in this manner causes data administration and management (such as migration, backups, archiving) costs to skyrocket! And while this approach may solve your storage capacity problem, it doesn't necessarily improve application I/O performance and it does nothing to reduce management costs or improve application workflow.

Many data centers have become victims of "filer-sprawl" — IT departments deploying numerous, relatively inexpensive NAS appliances or "filers" in a futile effort to keep up with out-of-control storage capacity demands. Beyond simply trying to keep up with capacity demands, "stop-gap" or temporary fixes often lead to other data access and management challenges, including

- Rising administrative costs to manage file-based data
- Data accessibility that's limited in remote locations
- Continuous data availability and protection that becomes increasingly difficult to maintain
- Backup and archival operations that can't keep pace with growing data

REMEMBER

Data access and management is critical to an efficient computing infrastructure. An efficient infrastructure must be balanced properly between three key components: compute, network, and data. The network and the data are normally the most difficult challenges for enterprise IT departments.

Three Important Functions of Storage

At its most basic level, enterprise storage performs three important functions, which are

- Store data (intermediate and final)
- Protect data from being lost
- Feed data to the computer's processors (so they can keep doing work)

Storage administrators have always recognized the need for storage capacity and data protection, and most storage vendors do a good job of providing solutions that satisfy these first two functions.

However, storage administrators are now increasingly focused on the third function of storage because getting data from the storage to the processor has become a performance bottleneck, and most storage vendors have done a poor job of addressing this performance issue. Consider that over the past decade,

- **CPU** speed performance has increased 8 to 10 times.
- **DRAM** speed performance has increased 7 to 9 times.
- **Network** speed performance has increased 100 times.
- **Bus** speed performance has increased 20 times.
- **Hard disk drive (HDD)** speed performance has increased only 1.2 times.

One result of this storage bottleneck is that your applications may be running slow, which negatively impacts productivity and wastes the capacity of other expensive infrastructure in your datacenter.

Defining Types of Storage

Not all storage is created equal. Storage systems are commonly divided into block-, file-, and object-based storage systems and include direct-attached storage (DAS), network-attached storage (NAS), and storage area networks (SAN).

Block storage

Block-based storage stores data on a hard disk as a sequence of bits or bytes of a fixed size or length (a block). In a block-based storage system, a server's operating system (OS) connects to the hard drives. Block storage is accessible via a number of client interfaces including

- Fibre Channel (or Fibre Channel Protocol, FCP)
- SCSI (Small Computer System Interface) and iSCSI (Internet Protocol SCSI)
- SAS (Serial Attached SCSI)
- ATA (Advanced Technology Attachment) and SATA (Serial ATA)

Fibre Channel and iSCSI interfaces, as well as SAS, are commonly used for storage housed outside of the computer, such as a storage area network (SAN). SAS and ATA are typically used in direct-attached storage (DAS).

Block-based storage systems are commonly implemented as either direct-attached storage (DAS) or storage area networks (SAN).

Direct-attached storage

DAS is the simplest and cheapest type of storage for computer systems. As the name implies, DAS is directly attached to the computer or server via a bus interface (see Figure 1-2).

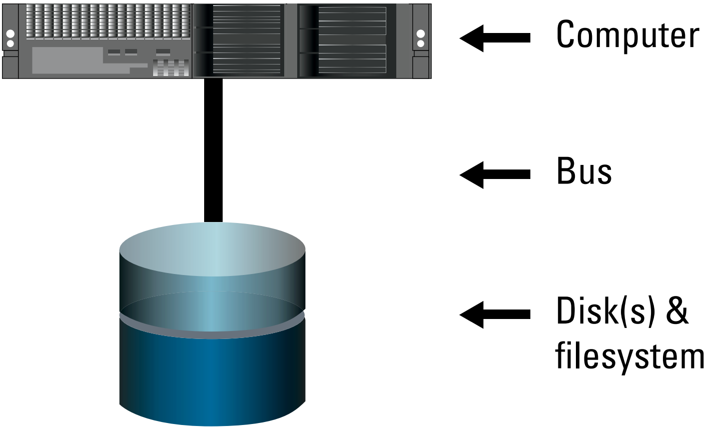

Figure 1-2: DAS connects hard disks directly to the computer or server via a bus interface.

DAS can provide basic data protection functions (for example, RAID and backup) and has very limited capacity because you can only install as many drives as the number of physical slots available on the server.

Storage area networks

A storage area network (SAN) is a separate storage system with its own protection functions connected to a server or servers through a dedicated storage network (see Figure 1-3).

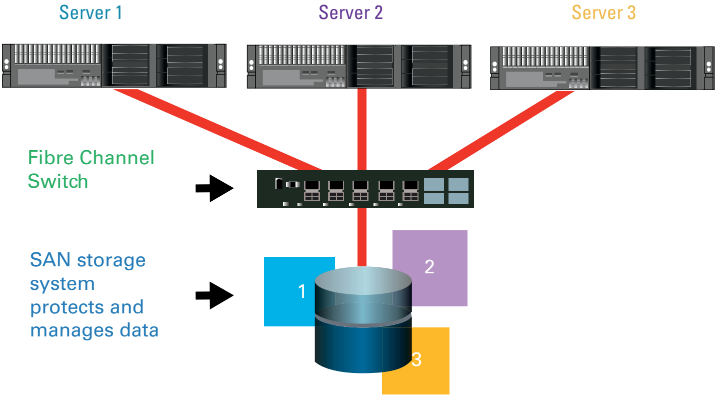

Figure 1-3: A SAN connects a storage array to servers via a dedicated storage network.

A SAN can be used by multiple servers. Each server has one or more fast, dedicated storage connections to one or more storage arrays. A SAN allows multiple computers to share access to a set of storage controllers. This provides great flexibility for maintaining enterprise IT infrastructures. In large organizations, SANs enable a division of labor where the system administrators manage the computers and the storage administrators manage the SAN. On its own, data can't be shared between separate LUNs or volumes, even within the SAN, except when cluster file systems are used in the SAN. Having multiple computers sharing access to the same data is important to many applications and workflows — making a cluster file system a necessary addition to SANs used for these purposes.

A LUN (Logical Unit Number) is an identifier assigned to a collection of disks within the storage system, defined in a storage controller and partitioned so that host servers can access them. A computer can then use these LUNS to store data. For example, you can create a file system on a LUN as a place to store files. A volume is part of a LUN created within volume management software.

The IBM XIV Storage System is an example of block-based SAN attached storage. Learn more at `www.03.ibm.com/systems/storage/disk/xiv`.

SANs are commonly used in mission-critical or high-transaction (IOPS, or I/O Operations Per Second) environments, for example, online transaction processing (OLTP) databases, ERP (Enterprise Resource Planning), and virtualized systems. Advantages of SANs include

- **Fast speeds:** SAN speeds are increasing dramatically due to
    - **Fabric interconnects:** Speeds of 40 Gbps and 80 Gbps are common, and InfiniBand EDR (Enhanced Data Rate) in a 12X cluster is capable of 300 Gbps data rates.
    - **More spindles, more speed:** As you add more drives to a SAN, you can increase the read/write access speeds available to the computers using the SAN.

- **Management:** Processing and storage are managed separately.
- **Data protection:** Data protection functions, such as backup and off-site replication, can be done outside the computer running the application and don't choke the performance of the attached servers.

Compared to other storage systems, SANs can be relatively expensive because they're engineered for maximum reliability and performance.

File storage

File-based storage systems, such as NAS appliances, are often referred to as "filers" and store data on a hard disk as files in a directory structure. These devices have their own processors and OS and are accessed by using a standard protocol over a TCP/IP network. Common protocols include

- **SMB (Server Message Block) or CIFS (Common Internet File System):** SMB (or CIFS) is commonly used in Windows-based networks.
- **NFS (Network File System):** NFS is common in Unix- and Linux-based networks.
- **HTTP (Hypertext Transfer Protocol):** HTTP is the protocol you most commonly use when using a web browser.

NAS appliances are relatively easy to deploy, and client access is straightforward using the common protocols. Computers and the NAS appliances are all connected over a shared TCP/IP network, and the data stored on NAS appliances can be accessed by virtually any computer, regardless of the computer's OS.

NAS appliances are fairly common in datacenters today. However, NAS appliances have several significant disadvantages. They're typically slower than DAS or SAN and can be storage performance bottlenecks because all data has to go through the NAS's own processors. NAS appliances also have limited scalability. When a NAS appliance fills up, you add another, and another, and so on. This creates "islands of storage" that are very inefficient to manage (see Figure 1-4).

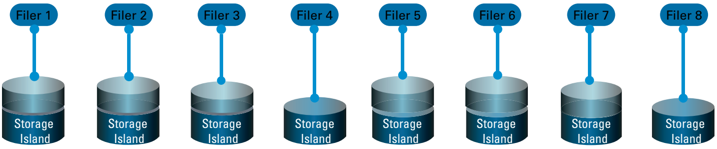

Figure 1-4: Filers often lead to "islands of storage" in the datacenter.

Object storage

Object-based storage systems use containers to store data known as *objects* in a flat address space instead of the hierarchical, directory-based file systems that are common in block- and file-based storage systems (see Figure 1-5).

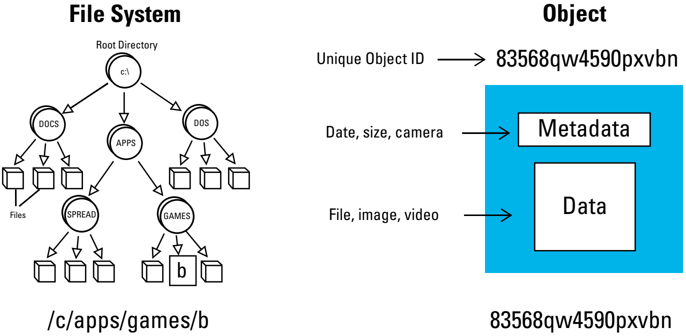

Figure 1-5: Comparing the file system to the object-based storage system.

A container stores the actual data (for example, an image or video), the metadata (for example, date, size, camera type), and a unique Object ID. The Object ID is stored in a database or application and is used to reference objects in one or more containers. The data in an object-based storage system is typically accessed using HTTP using a web browser or directly through an API like REST (representational state transfer). The flat address space in an object-based storage system enables simplicity and massive scalability, but the data in these systems typically can't be modified (other than being completely deleted and an entirely new version written in its place — an important distinction to keep in mind).

TIP

Object-based storage is commonly used for cloud services by providers such as IBM SoftLayer, Amazon S3, Google, and Facebook.

Time to explain RAID

RAID (Redundant Array of Independent Disks, originally Redundant Array of Inexpensive Disks) is a data storage technology that distributes data across multiple drives in one of several ways (called RAID levels), depending on the level of performance and protection required. Eight standard RAID levels are defined by the Storage Networking Industry Association (SNIA) as follows:

- **RAID 0** (block-level striping without parity or mirroring). Requires a minimum of two hard drives; provides maximum performance and usable storage capacity, but no redundancy.
- **RAID 1** (mirroring without parity or striping). Requires a minimum of two hard drives; read performance is not impacted. Write performance is slower than RAID 0 because data must be simultaneously written to both drives in a mirrored set and usable storage capacity is reduced by 50 percent; one drive in a mirrored set can fail without loss of data.
- **RAID 2** (bit-level striping with dedicated Hamming-code parity). Requires a minimum of three hard drives with each sequential bit of data striped across a different drive; this is a theoretical RAID level that has not been implemented.
- **RAID 3** (byte-level striping with dedicated parity). Requires a minimum of three hard drives with each sequential byte of data striped across a different drive; not commonly used.

- **RAID 4** (block-level striping with dedicated parity). Requires a minimum of three hard drives; similar to RAID 5 but with parity data stored on a single drive.
- **RAID 5** (block-level striping with distributed parity). Requires a minimum of three hard drives; data and parity are striped across all drives; a single drive failure causes all subsequent reads to be calculated from the parity information distributed across the remaining functional drives, until the faulty drive is replaced and the data rebuilt from the distributed parity information.
- **RAID 6** (block-level striping with double distributed parity). Requires a minimum of four hard drives; similar to RAID 5 but allows for two failed drives. This is the most common RAID level in use today, but with growing drive capacities (in excess of 3 TB) rebuilds can take days with the entire system being very unresponsive during that time.
- **RAID 10** (mirroring and striping, also known as RAID 1+0). Requires a minimum of four hard drives; data is striped across primary disks and mirrored to secondary disks in an array. Read performance is not impacted, but write performance is degraded similar to RAID 1 and usable storage capacity is reduced by 50 percent; one drive in each span (primary and secondary) can fail without loss of data.

In Chapter 3, you find out how new RAID implementations are providing innovative approaches to these standard RAID definitions. One of these new technologies is GPFS Native RAID.

Hard Disk and SSD Technologies

The most common hard drives in use today include SATA (Serial ATA), SAS (Serial Attached SCSI), and SSD (Solid State Drives). Each drive technology provides different combinations of capacity, performance, and cost (see Figure 1-6).

SATA drives are typically used in desktop and laptop computers, as well as DAS, NAS, and SAN. SATA drives provide the highest capacity (for example, 2 to 4TB) and lowest cost per gigabyte. However, SATA drives are slower (typically 7,200 RPM) and less reliable than other drive technologies. SATA is commonly implemented in SANs for secondary storage and for application data with relatively low IOPS requirements.

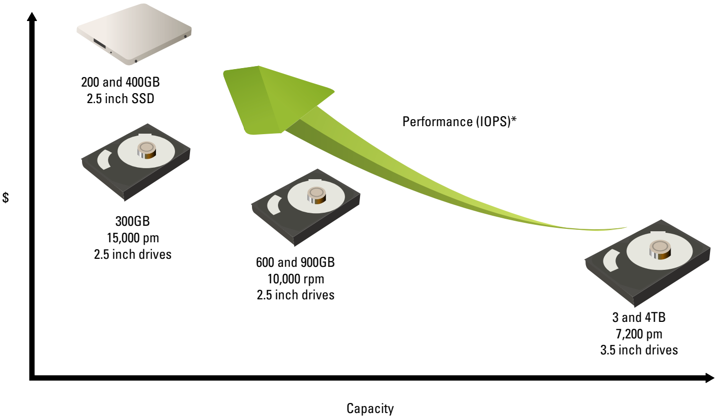

Figure 1-6: Different hard drive technologies require a tradeoff between capacity, performance, and cost.

SAS drives are commonly used in servers (DAS) or SANs. SAS drives provide a tradeoff between performance (typically 10,000 and 15,000 RPM) and capacity (for example, 300, 600, and 900GB). SAS drives are more reliable than SATA drives and their individual components are designed to handle frequent read/writes and high IOPS.

SSDs use flash technology to provide reliable and high-speed data storage. Flash technology uses floating gate transistors to store data as 1s and 0s in individual cells. SSD capacities are increasing rapidly, with current capacities ranging up to 500GB, and are extremely fast: Read/write operations on Flash storage are measured in microseconds compared to milliseconds for hard disk drives, and IOPS are measured in tens of thousands to millions, compared to hundreds for hard disk drives. Although the cost of SSDs is dropping and capacity increasing, this technology still comes at a premium and is most commonly used today in situations where high performance is needed over capacity.

Cluster File Systems

A cluster file system can be accessed from many computers at the same time over a network or a SAN. Yes, this sounds similar to network protocols such as CIFS or NFS, but there are a few key differences:

- Some cluster file systems provide access from multiple nodes over a SAN; this isn't possible with CIFS or NFS.
- Direct access from the computer using a SAN provides a cluster file system several performance advantages over standard network protocols.
- Cluster file systems are tightly coupled and communicate at a more sophisticated level to enable an application, for example, to have multiple nodes reading and writing to a single file.

Some cluster file systems extend the same functionality over TCP/IP or Infiniband. This is similar to NFS and CIFS, because both approaches use the network to access data, but in the case of a cluster file system, the network protocol used to transfer data is part of the cluster file system software. This tight integration allows the cluster file system to provide high performance and advanced access patterns over the network. These protocols can leverage technologies including Remote Direct Memory Access (RDMA) for faster processing of data.

These integrations mean that cluster file system can be fast and provide advanced functionality, but they aren't particularly well suited for workstation access. You wouldn't run a cluster file system on your tablet to access your music, for example. Cluster file systems are designed to provide enhanced file data access for the IT infrastructure, like the systems from which you download your music.

A cluster file system achieves high I/O performance by spreading data across multiple storage servers, all sharing the same global namespace, to increase scalability.

A parallel file system is a type of cluster file system that reads and writes data in parallel across multiple storage nodes providing extremely high performance, scalability, and data protection. You find out more about parallel file systems in Chapter 2.

Chapter 2

What Is Software Defined Storage?

In This Chapter

- Putting the all the pieces together in software defined storage
- Recognizing the benefits of software defined storage
- Introducing GPFS software defined storage

Software defined storage is a relatively new concept in the computing and storage industry and can refer to many different technologies and implementations. Software defined storage is part of a larger industry trend that includes software defined networking (SDN) and software defined data centers (SDDC). This chapter explains exactly what software defined storage is.

Defining Software Defined Storage

At its most basic level, *software defined storage* is enterprise class storage that uses standard hardware with all the important storage and management functions performed in intelligent software. Software defined storage delivers automated, policy-driven, application-aware storage services through orchestration of the underlining storage infrastructure in support of an overall software defined environment. Standard hardware includes

- **Disk storage** such as SAN, NAS, and disk arrays or JBODs (just a bunch of disks)
- **Network devices** such as switches and network interfaces
- **Servers** for storage processing, management, and administration

Additional characteristics of software defined storage can include

- Automated policy-driven administration for storage management functions, such as information life cycle management (ILM) and provisioning
- Storage virtualization
- Separate control and data planes to manage the storage infrastructure and data in the storage infrastructure, respectively
- Massive scale-out architecture

These characteristics are in contrast to traditional storage systems that depend heavily on custom hardware-based controllers to perform storage functions. NAS, DAS, and SAN (discussed in Chapter 1) are examples of typical hardware-based storage systems that rely on special RAID controllers and non-portable custom firmware for their storage functions.

Key Benefits of Software Defined Storage

Enterprises today are recognizing many significant benefits of software defined storage in their datacenters. These include increased flexibility, automated management, cost efficiency, and limitless scalability.

Increased flexibility and agility

Traditional enterprise storage platforms such as SAN and NAS (discussed in Chapter 1) are typically based on proprietary systems and come with a high total cost of ownership (TCO). SAN solutions typically require the use of costly and complex SAN switches, storage arrays, and other proprietary components.

NAS devices are relatively inexpensive but have limited scalability. When you run out of space on a NAS, you simply add more NAS devices. However, this isn't a true scale-out capability because each individual NAS is presented as separate, standalone storage that's separately managed.

A software defined storage solution increases flexibility by enabling organizations to use non-proprietary standard hardware and, in many cases, leverage existing storage infrastructure as part of their enterprise storage solution. Additionally, organizations can achieve massive scale with a software defined storage solution by adding individual, heterogeneous hardware components as needed to increase capacity, and improve performance in the solution.

Intelligent resource utilization and automated management

Automated, policy-driven management of software defined storage solutions helps drive cost and operational efficiencies. As an example, software defined storage manages important storage functions including ILM, disk caching, snapshots, replication, striping, and clustering. In a nutshell, these software defined storage capabilities enable you to put the right data in the right place, at the right time, with the right performance, and at the right cost — automatically.

Cost efficiency

Rather than using expensive proprietary hardware, software defined storage uses standard hardware to dramatically lower both acquisition costs and TCO for an enterprise-class storage solution. The software in a software defined storage solution is standards based and manages the storage infrastructure as well as the data in the storage system.

In many cases, organizations can leverage their existing investments in storage, networking, and server infrastructure to implement an extremely cost-effective software defined storage solution.

In a July 2012 report, Gartner, Inc., found that the average acquisition cost per gigabyte of traditional multi-tiered storage systems ranged from $0.9/GB to $5/GB. By comparison, software defined storage solutions average $0.4/GB.

Limitless elastic data scaling

Unlike traditional storage systems, such as SAN and NAS, software defined storage enables you to scale out with relatively inexpensive standard hardware, while continuing to manage storage as a single enterprise-class storage system. As you scale out your storage infrastructure, performance and reliability continue to improve. As an example, IBM General Parallel File System (GPFS) delivers orders of magnitude more in I/O performance improvement as hardware is added, compared to conventional NAS (see Figure 2-1). GPFS is covered in more detail in Chapter 3.

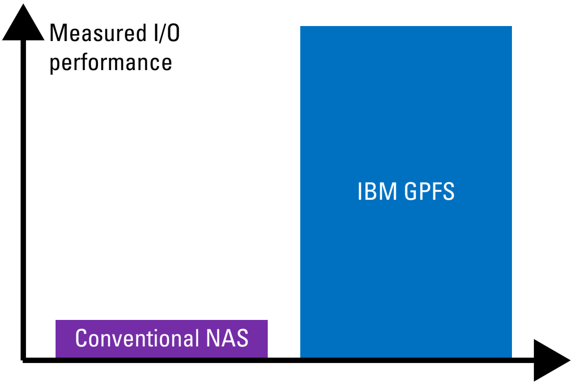

Figure 2-1: IBM GPFS delivers extreme I/O performance.

Software defined storage provides massive, virtually limitless scalability. For example, IBM GPFS supports

- A maximum file system size of one million yottabytes
- 2^{63} (or approximately 9 quintillion) files per file system
- IPv6
- 1 to 16,384 nodes in a cluster

REMEMBER

A yottabyte is equal to one trillion terabytes.

Supporting Files, Blocks, and Objects

Software defined storage typically refers to software that manages the creation, placement, protection, and retrieval of data. IBM GPFS is an enterprise-class software defined storage solution that runs on IBM and third-party platforms and is available as software that can be deployed on a variety of commodity hardware systems and as part of an integrated "appliance" — the GPFS Storage Server (GSS), described in Chapter 4.

Developers of large Cloud-scale applications have been particularly interested in software defined storage. For them, existing heavy hardware controlled storage solutions simply don't scale, are prohibitively expensive, and are too inflexible to dynamically expand capacity for application data they envision driving their business needs moving forward. Many of them have focused their development on OpenStack — the open source cloud computing platform for public and private clouds. See the nearby sidebar "Focusing on OpenStack" for more information.

For OpenStack developers, GPFS can unify your storage with a common way to store VM images, block devices, objects, and files. Software defined storage allows this type of integration. GPFS functions like GPFS Native RAID (GNR) and policy-based data placement give you the flexibility to put the data in the best location on the best tier (performance and cost) at the right time. Software defined storage means that you can deploy this on heterogeneous industry standard hardware. This is shown in Figure 2-2.

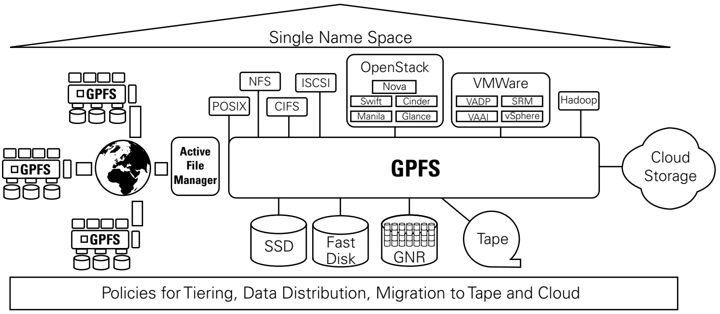

Figure 2-2: GPFS provides a common storage plane.

Focusing on OpenStack

OpenStack has a modular architecture with various components including

- **OpenStack Compute (Nova):** A cloud computing fabric controller
- **Block Storage (Cinder):** Provides persistent block-level storage devices
- **Object Storage (Swift):** Scalable redundant storage system

With OpenStack, you can control pools of processing, storage, and networking resources throughout a datacenter. And while OpenStack provides open source versions of block and object storage, many OpenStack developers have identified a need for more robust storage to support Cloud-scale applications. While many OpenStack developers feel they can architect around limitations in OpenStack compute capabilities and robustness, storage has a much "higher bar" as far as resiliency and reliability go.

Responding for the need for robust software defined storage, the OpenStack "Havana" release includes an OpenStack Block Storage Cinder driver for IBM GPFS, giving architects who build public, private, and hybrid clouds access to the features and capabilities of the industry's leading enterprise software-defined storage system. And the Cinder is just the beginning. IBM's vision for GPFS and OpenStack is to create a single scale-out data plane for the entire data center or multiple connected data centers worldwide.

GPFS unifies OpenStack VM images, block devices, objects, and files with support for Nova, Cinder, Swift, and Glance, along with POSIX interfaces like NFS and CIFS for integrating legacy applications. The ability to use a single GPFS file system to manage volumes (Cinder), images (Glance), shared file systems (Manila), and use file clones to efficiently/quickly shared data within and between components will be a big advantage for Cloud-scale application developers using OpenStack.

The robustness and features of GPFS combined with OpenStack Swift object extensions could provide an enterprise-grade object store with high storage efficiency, tape integration, wide-area replication, transparent tiering, checksums, snapshots, and ACLs — capabilities most object-based storage offerings can't match today. OpenStack on GPFS delivers compelling efficiencies in a single unified storage solution that can support object and file access to the same data with robust and efficient GPFS Native RAID data protection. OpenStack Swift object storage on GPFS can reduce the amount of raw storage you need to use compared to object storage systems that rely strictly on replication.

Chapter 3

Digging Deeper into IBM GPFS

In This Chapter

- Understanding IBM GPFS
- Removing storage performance bottlenecks
- Eliminating storage management complexity
- Exploring different clustering options

To really understand IBM General Parallel File System (GPFS), you can take four servers and use a storage-area network (SAN) to attach the storage to all four servers so all servers can see all the disks. You install evaluation GPFS software on the four servers and from one server create a file system, mount that file system on all the servers, install your application, and you're off and running. In this chapter, you discover the basics of the GPFS, including the core concepts, key features, and cluster configuration options.

Understanding GPFS Concepts

GPFS started out as a clustered file system and has evolved into so much more. Today it's a full-featured set of file management tools, including advanced storage virtualization, integrated high availability, automated tiered storage management, and performance to effectively manage very large quantities of file data. GPFS is designed to support a wide variety of application workloads and has been proven extremely effective in very large, demanding environments.

GPFS allows a group of servers concurrent access to a common set of file data over a common SAN infrastructure, a network, or a mix of connection types. The servers can run any mix of AIX, Linux, or Windows Server operating systems. GPFS provides storage management, information life cycle management tools, centralized administration, and it allows for shared access to file systems from remote GPFS clusters providing a global namespace.

A GPFS cluster can be a single server, two servers providing a high-availability platform supporting a database application, for example, or thousands of servers used for applications such as the modeling of weather patterns. GPFS was designed to support high-performance workloads and has since been proven very effective for a variety of applications. Today, GPFS is installed in clusters supporting big data analytics, gene sequencing, digital media, and scalable file serving. These applications are used across many industries including financial, retail, digital media, biotechnology, science, and government.

GPFS provides a unique set of extended interfaces that can be used to provide advanced application functionality. Using these extended interfaces, an application can determine the storage pool placement of a file, create a file clone, and manage quotas. These extended interfaces provide features in addition to the standard POSIX (Portable Operating System Interface) interfaces.

POSIX is an IEEE (Institute of Electrical and Electronics Engineers) family of standards for maintaining compatibility between different variations of UNIX and other operating systems.

Removing Data Related Bottlenecks

Over the past decade, processors, memory, network, and bus performance have all increased exponentially, but disk speed performance has only increased 1.2 times. This performance gap slows data heavy applications, delays schedules, and wastes expensive infrastructure. GPFS accelerates time to results and maximizes utilization by providing parallel access to data. GPFS provides extreme performance and eliminates storage bottlenecks, by providing parallel access to data. (see Figure 3-1).

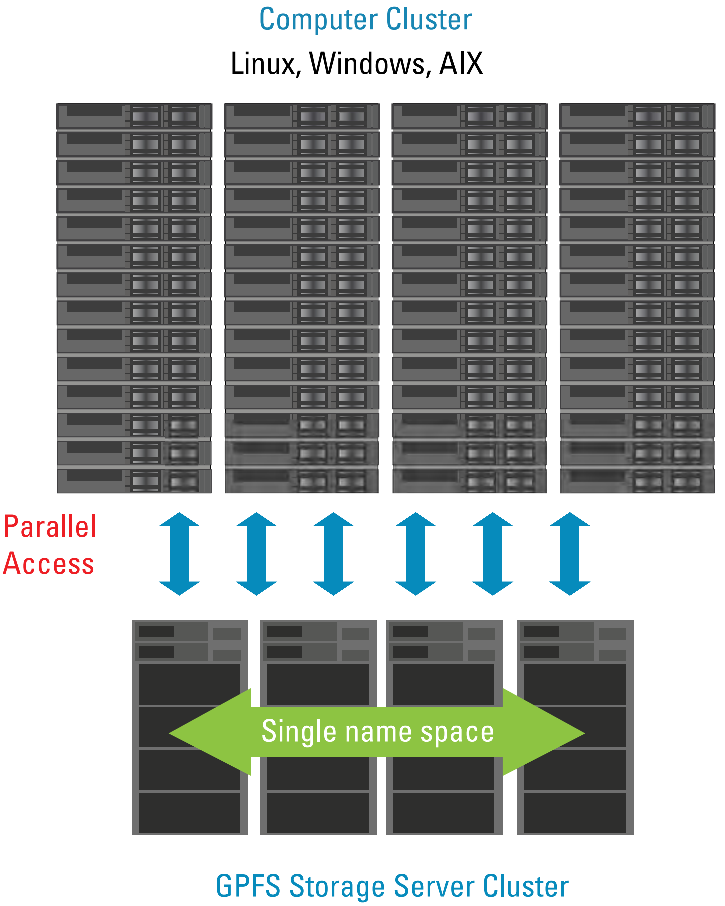

Figure 3-1: GPFS eliminates storage bottlenecks.

GPFS achieves high-performance I/O by

- Striping data across multiple disks attached to multiple servers
- Providing efficient client side caching
- Executing high-performance metadata (inode) scans
- Supporting a wide range of file system block sizes to match I/O requirements
- Utilizing advanced algorithms that improve I/O operations

- Using block-level locking based on a very sophisticated token management system to provide data consistency, while allowing multiple application servers concurrent access to the files

When many servers need to use the same set of files at the same time, the file system needs to ensure that all the files are protected, so one server can't change a file without the other servers knowing about the change. Keeping thousands of servers "in the loop" on file status is difficult and scaling up is even harder.

GPFS provides file integrity protection through a token process that keeps file data consistent by always ensuring there is only one owner for any given file. This method scales well in GPFS because any server in the cluster can be assigned file protection duty. There are two parts to managing tokens and file consistency: handing out the tokens and keeping file metadata up to date.

The server(s) that initially have the token for all files that are not in use is called the *token manager*. You can assign one or more servers to be a token manager. Multiple token managers help each other out by sharing the workload and by taking over when a fellow token manager fails. When a file is opened, the token manager hands off the token for that file to the server that's opening the file. The server using the file is now responsible for all metadata changes to that file. If a server wants to open a file that is already open on another server, the token manager redirects the request to the server that already has the file open and lets the two servers work out the details among themselves. This sharing of metadata maintenance across the entire cluster is what makes GPFS scale very effectively. Many other file storage technologies rely on a single metadata server or centralized database. Such a single/centralized approach quickly limits how much data can be stored. GPFS addresses this limitation by sharing the workload.

Simplifying Data Management

Creating a billion files isn't that difficult, but maintaining a storage solution containing a billion files or more takes industrial class tools. GPFS has the tools needed to manage petabytes of data and billions of files. The global namespace is easy to administer and can be scaled quickly, as desired, by simply adding more scale-out resources — eliminating "filer sprawl" and its associated issues.

GPFS has an administration model that's easy to use and consistent with standard file system administration practices. These functions support cluster management and other standard file system administration functions such as user quotas and snapshots.

A single GPFS command can perform a file system function across the entire cluster, and most can be issued from any server in the cluster. Optionally, you can designate a group of administration servers that can be used to perform all cluster administration tasks, or only authorize a single login session to perform admin commands cluster-wide. This allows for higher security by reducing the scope of server-to-server administrative access.

You can set quotas by user, group, or at the directory level to monitor and control file system usage across the cluster. When using quotas, you can easily see usage reports that include user, group, directory, inode, and data block usage. You can use snapshots to protect data from human error. You can create a snapshot of an entire file system or a subdirectory (called a *fileset*). A snapshot is used to preserve the file system's contents at a single point in time. It contains a copy of only the file system data that has changed since the last snapshot was created and keeps that data in the same pool as the original file, which keeps space usage at a minimum. Using the same pool simplifies storage administration because you don't have to set aside additional space for snapshot data. Snapshots provide an online backup capability that allows you (or an end user) to easily recover from an accidental file deletion, or the ability to compare a file to an older version.

Most application clusters need a method to get data into and out of the cluster. Because GPFS runs directly on a standard server, you can use a variety of tools to get file data into and out of a GPFS file system. To better enable end user access to a GPFS file system, the file system can be exported to clients outside the cluster through NFS (Network File System), including the capability of exporting the same data from multiple servers. This GPFS feature is called *Clustered NFS* (cNFS). Clustered NFS allows you to provide scalable file service with simultaneous access to a common set of data from multiple servers. The cNFS feature includes failover capability, so if a NFS server fails, the clients connected to that server automatically connect to another server in the cluster.

NFS is a Network file system protocol that enables access to storage by using a standard protocol over a TCP/IP network. NFS protocol access is commonly provided by a network-attached storage (NAS) appliance or similar device. Samba enables file and print services for Microsoft Windows clients from UNIX and Linux based servers.

Cluster Configurations

When it comes to cluster configuration options, GPFS is a multi-function tool. The same GPFS software is installed on all the servers in a cluster. What a server does, and how it participates in the cluster, is based on the hardware it has available and what you need it to do. Cluster configuration is independent of which file system features you require. Cluster configuration options can be characterized into the three categories covered in this section.

Shared disk

A shared disk cluster is the most basic environment. In this configuration, the storage is directly attached to all servers in the cluster, as shown in Figure 3-2. Application data flows over the SAN, and control information flows among the GPFS servers in the cluster over a TCP/IP network.

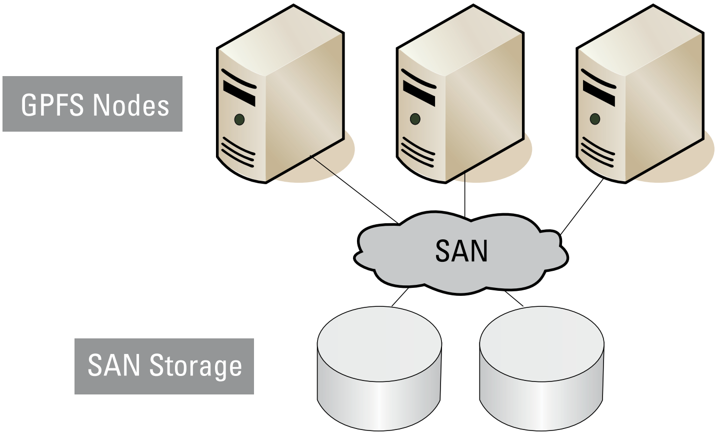

Figure 3-2: SAN attached storage.

This configuration is best for small clusters (1 to 50 servers) when all servers in the cluster need the highest performance access to the data. For example, this configuration is good for high-speed data access for digital media applications or a storage infrastructure for data analytics.

Don't have a SAN? No problem!

When every server in the cluster can't be attached directly to a SAN, GPFS uses a block-level interface over the network called the Network Shared Disk (NSD) protocol. GPFS transparently handles I/O requests, whether using NSD or a direct attachment to the SAN, so the mounted file system looks the same to the application.

In general, SAN-attached storage provides the highest performance, but the cost and management complexity of SANs for large clusters is often prohibitive. In these cases, network block I/O provides a viable option.

GPFS clusters use the NSD protocol to provide high-speed data access to applications running on LAN-attached servers. Data is served to these client servers from an NSD server. In this configuration, disks are SAN-attached only to the NSD servers. Each NSD server is attached to all or a portion of the disks. Multiple NSD servers should be deployed to serve each disk in order to avoid a single point of failure.

GPFS uses a network to transfer control information and data to NSD clients. The network doesn't need to be dedicated to GPFS, but it should provide sufficient bandwidth to meet your GPFS and other applications sharing the bandwidth.

In a NSD server configuration, a subset of the total server population is defined as NSD servers. The NSD servers are responsible for the abstraction of disk data blocks across an IP-based network. The fact that I/O is remote is transparent to the application. Figure 3-3 shows an example of a configuration where a set of compute servers are connected to a set of NSD servers via a high-speed interconnect or an IP-based network (such as Ethernet). In this example, data to the NSD servers flows over the SAN, and data and control information flows to the clients across the LAN.

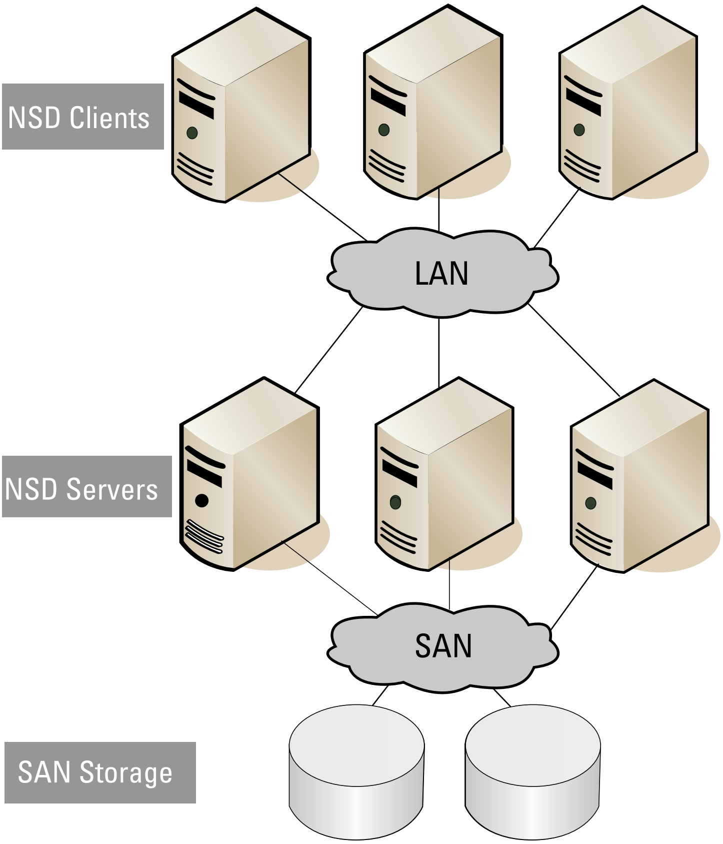

Figure 3-3: Network block I/O.

The choice of how many servers to configure as NSD servers is based on individual performance requirements and the capabilities of the storage subsystem. High bandwidth LAN connections should be used for clusters requiring significant data transfer. To enable high-speed communication GPFS supports TCP/IP using whatever hardware you have available (1Gbit and 10Gbit Ethernet, link aggregation, IPoIB), and RDMA on InfiniBand for control and data communications. InfiniBand is a switched fabric communications link commonly used in High Performance Computing (HPC) installations. With speeds up to 56Gbit/sec and very low latency it is well suited to workloads with high I/O demands.

GPFS provides the ability to designate separate IP interfaces for intra-cluster communication and the public network. This provides a more clearly defined separation of communication traffic. An NSD server architecture is well suited to clusters with sufficient network bandwidth between the I/O servers and the clients. For example, statistical applications like financial fraud detection, supply chain management, or data mining.

Empowering global collaboration

GPFS provides low latency access to data from anywhere in the world with Active File Management (AFM) distributed disk caching technology. AFM expands the GPFS global namespace across geographical distances, providing fast read and write performance with automated namespace management from anywhere in the world. As data is written or modified at one location, all other locations get the same data with minimal delays. These game-changing capabilities accelerate project schedules and improve productivity for globally distributed teams.

Sharing Data Across GPFS Clusters

You can share data across GPFS clusters via two methods: GPFS multi-cluster and Active File Management (AFM).

GPFS multi-cluster

GPFS multi-cluster allows you to utilize the GPFS NSD protocol to share data across clusters. With this feature, you let other clusters to access one or more of your file systems, and you mount file systems that belong to other GPFS clusters for which you've been authorized. A multi-cluster environment permits the administrator access to specific file systems from another GPFS cluster. This feature permits clusters to share data at higher performance levels than file sharing technologies like NFS or CIFS.

GPFS multi-cluster isn't intended to replace file sharing technologies that are optimized for desktop access or for access across unreliable network links. Multi-cluster capability is useful for sharing across multiple clusters within a single physical location or across multiple locations.

In Figure 3-4, Cluster A owns the storage and manages the file system. It may grant access to file systems that it manages to remote clusters, such as Cluster B. In this example, Cluster B doesn't have any storage but that isn't a requirement. Commonly in the case where a cluster doesn't own storage, the servers are grouped into clusters for ease of management. When the remote clusters need access to the data, they mount the file system by contacting the owning cluster and passing required security checks. In Figure 3-4, Cluster B accesses the data through the NSD protocol.

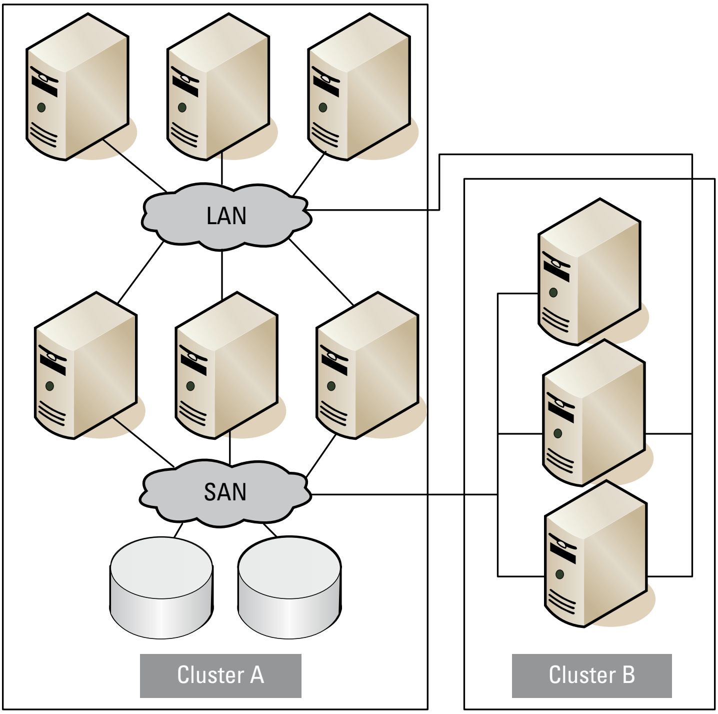

Figure 3-4: Multi-cluster configuration.

Multi-cluster environments are well suited to sharing data across clusters belonging to different organizations for collaborative computing, grouping of clients for administrative purposes or implementing a global namespace across separate locations. A multi-cluster configuration allows you to connect GPFS clusters within a data center, across campus, or across reliable WAN links.

World-wide data sharing

So what can you do if your network link is really long or not so reliable? You can use a feature in GPFS called Active File Management (AFM). AFM allows you to create instead of direct access to the data in the other cluster, as in a multi-cluster situation, a copy of the data when and where you need to access it.

At first glance, AFM may seem like any other cache. But when you start looking at what you can do with these basic behaviors, the options start multiplying. To understand some of the possible options, take a look at how AFM can be used.

A cache relationship is defined when you create a fileset. You can have up to 1,000 independent cache relationships in each GPFS file system. Each cache fileset can have its own mode and configuration parameters, and still share a common set of storage. A cache file system can be an entire file system in size, or it can be limited by quotas. A read-only cache contains data that can't be modified — it's read-only. Data is copied into the cache fileset by GPFS, on demand, by opening a file or running a command to prefetch (or pre-cache) a list of files. A single target can have as many read-only cache relationships as it has bandwidth available, because the cache does all the work. A target doesn't know there's a cache, and a cache only knows about its own target. Two filesets in the same file system can be caching data from the same target, and they wouldn't know about each other.

The isolation between the cache and target is what makes this model scale so well. You can have as many as 1,000 read-only caches because each cache only has to track one relationship. And then it gets interesting! Beyond the many-to-one cache relationships, you can also cascade caches. A target can have a dual personality; it can be a cache and a target at the same time. For example, say a target exists in New York. The data originates in London and is cached in New York. The office in Tokyo needs a copy of the same data, so the cache in Tokyo uses the copy in New York as the target.

In read-only mode, data consistency is simple: The target is the only place where the data can be created and modified. But there are other caching modes that provide additional functionality. For example, in an *independent writer* cache, data can be created and modified in the cache and at the target. When a file is created or modified in the cache, the write operation completes locally and the changes are automatically copied to the target. You can also have multiple independent writers updating data for the same target.

Using independent writer with multiple caches accessing the same target is similar to having multiple people share an NFS project space. If you copy your version of a project plan after your colleague has saved their version of the same project plan, the next time it's opened, the reader sees your version. Independent writer works in the same way, except it's the last cache to update that wins, so you need to think about the file updates that could occur in a remote data center when designing your independent writer cache relationships.

By using these basic caching techniques, you can allow all your sites to see the same data all the time. Figure 3-5 shows one example of using AFM to create a single view of all the data from multiple sites.

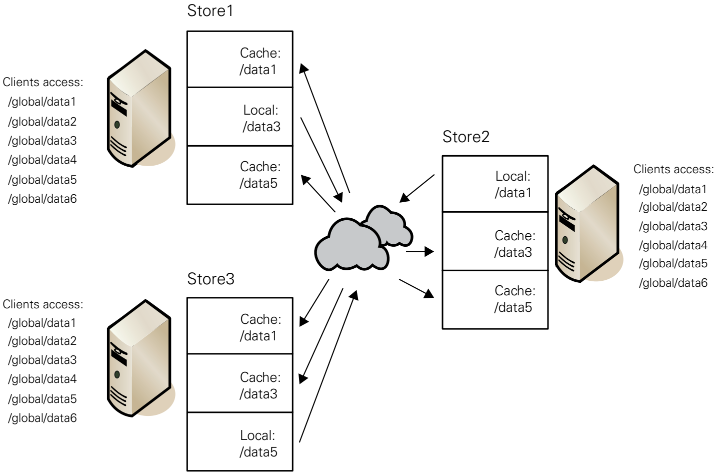

Figure 3-5: Global namespace using AFM.

In this example, each site is the target for one third of the data. The other two sites have cache relationships with the other sites. This means that no matter which site you log into, you see exactly the same file system structure and you have access to all the data — it may just take a little longer to read a file if it has not yet been copied to your site.

Managing the Full Data Life Cycle at Lower Costs

GPFS enhances information life cycle management and lowers your data management costs significantly by using multiple tiers of storage including tape. With powerful policy-driven automation and tiered storage management, you can create optimized tiered storage pools by grouping disks based on performance, locality, or cost characteristics. Data migrated to tape remains visible in the file system and is directly accessible by end-users. Migration policies transparently move data from one storage pool to another without changing the file's location in the directory structure.

For example, you can create a rule for thresholds that moves files out of the high performance pool if it's more than 80 percent full, thereby mitigating potential bottlenecks in the high performance pool. GPFS information life cycle management capabilities and benefits include

- Policy-driven automation and tiered storage management
- Flexibility to match the cost of storage to the value of data
- Storage pools create tiers of storage that include high-performance SSD, high-speed SAS Drives, and high-capacity NL SAS Drives
- Full integration with IBM Tivoli Storage Manager (TSM) and Linear Tape File System (LTFS) that provides the following functionality
    - GPFS handles all metadata processing, then hands the data to TSM for storage on tape
    - Data is retrieved from the external storage pool on demand when an application opens a file, for example
    - Policies move data from one pool to another without changing the file's location in the directory structure
    - Thresholds can move files out of the high-performance pool if more than 80 percent full, for example

Tape migration provides tiered data storage for approximately one-fifth the cost per terabyte of disk.

GPFS-powered Enterprise Solutions

GPFS is under the covers of some of the most game-changing enterprise products today, including SAP HANA in-memory database appliance, IBM DB2 PureScale, IBM SONAS (Scale-out NAS).

SAP HANA in-memory database appliance

SAP's High-performance Analytic Appliance (HANA) enables organizations to instantly explore and analyze all of their transactional and analytical data from virtually any data source in near real-time. Delivered on optimized hardware, HANA realizes the efficient processing and analysis of massive amounts of data by packaging SAP's intelligent use of in-memory technology, columnar database design, data compression, and massive parallel processing together with essential tools and functionality (for example, data replication and analytic modeling), business content, and the SAP BusinessObjects Business Intelligence (BI) solutions.

The three major components of this solution include IBM Blade Server and Chassis, IBM GPFS File System, and IBM DS4700 Storage equipment. The flexibility provided by this solution enables customers to grow their information warehouse as the data grows through smooth extension and upgrade paths in a pay-as-you-grow model.

IBM DB2 pureScale

DB2 pureScale is a cluster-based, shared-disk architecture that reduces costs through efficient use of system resources and allows users to scale out a database on a set of servers in an "active-active" configuration delivering high levels of availability, scalability, and application transparency. A DB2 server that belongs to a PureScale cluster is called a member. Each member can simultaneously access the same database for both read and write operations. Designed mainly for OLTP (online transaction processing) workloads with many concurrent transactions, DB2 PureScale offers almost unlimited scale-out in terms of number of members that it can support. Tests with up to 128 members show near-linear scalability.

GPFS is the underlying file system of IBM's DB2 PureScale offering for businesses with applications running on DB2 that require on-demand storage growth and continuous availability without any changes to the existing applications. GPFS allows the DB2 engine to run on several host servers that cooperate to provide coherent access to database clients from any of the member servers. Automatic load balancing takes care of client re-routing without the client having to reconnect to a different server during load redistribution when a server is added, removed, or fails.

This kind of data sharing and availability is provided by the GPFS storage infrastructure. GPFS enables DB2 PureScale to significantly reduce the risk and cost of business growth and provides unlimited capacity. Pairing GPFS file systems and DB2 for Linux, UNIX, and Windows databases creates an unbeatable platform for high-performance database systems.

IBM SONAS

Excessive proliferation of NAS (discussed in Chapter 1) and standalone file servers to cater to rapidly growing enterprise storage demands has resulted in an increase in storage system disparity and administrative and energy costs of IT assets. Typically, more than 50 percent of user data in an enterprise is inactive. This is because projects that once used such data actively are no longer current, or because technologies and the businesses have evolved. Manual data life cycle management in such cases is neither a scalable proposition nor does it address demanding I/O-intensive application workloads and performance.

IBM's Scale-Out NAS (SONAS) can manage multiple petabytes of storage and billions of files using a single file system providing much needed operational efficiency with automated, policy-driven tiered storage. SONAS can satisfy bandwidth-hungry applications with its scale-out performance and provide business continuity and disaster recovery using asynchronous replication. SONAS leverages GPFS to provide scalable storage and efficient data retrieval through content metadata policies. SONAS supports storage at scale through GPFS global namespace and hierarchical storage-management features.

GPFS can easily scale from a two server high-availability cluster to a 1,000 server compute farm. It is very cost effective at any scale and provides the same set of enterprise class functionality to a local health care clinic as it does to the world's largest banks. Although you may not need a file system that supports 18 petabytes and billions of files, you will. Inevitably, our data storage needs grow. Today, you are very likely carrying a smartphone or USB thumb drive that has more capacity than enterprise hard drives did just 15 years ago — yet it still seems to never be enough!

Chapter 4

Getting to Know the IBM GPFS Storage Server

In This Chapter

- Taking a look at GSS components
- Understanding GSS fault tolerant features

Disk drive performance hasn't changed significantly in many years, but in the last ten years disk drive capacity has increased from a maximum of 100GB per disk to 4,000GB — a 40x increase! To put this in context, assume that you can read a single disk drive at a speed of 60MB per second. In 2003, backing up that drive took 28 minutes. Today, because disk drives haven't gotten much faster, backing up a single disk full of data takes almost 19 hours. This same performance issue also impacts disk maintenance operations, such as RAID rebuilds. When a 4TB disk fails, how long does it take to read and write enough data to put a new disk in its place? These types of performance issues are one of the key driving forces behind the invention of the IBM GPFS Storage Server (GSS).

This chapter talks about the IBM System x GSS. GSS combines the performance of System x servers with GPFS software to offer a high-performance, scalable building-block approach to modern storage needs, known as software defined storage. GSS allows you to start with a configuration that meets your organization's current needs and expand capacity and bandwidth with each additional GSS to meet your future needs.

Delivering an End-to-End Solution

A GPFS Storage Server is made up of IBM System x servers, storage enclosures, and hard drives (JBODs, or just a bunch of disks), and includes all of the software and networking components in a fully integrated solution. IBM thoroughly tests and optimizes each GSS for reliability, interoperability, and performance so organizations can quickly deploy the system and get to work achieving their business goals.

GSS is built and tested by IBM and includes solution-level support that includes both IBM and third-party components.

Features of GSS include

- Scalable building-block approach to storage
- Two configurations: Model 24 (232 Near Line Serial Attached SCSI, or NL/SAS Drives) and Model 26 (348 NL/SAS Drives)
- Up to 4TB drives
- Up to1392TB of capacity
- FDR InfiniBand, 10 and 40Gbit Ethernet interconnects, or both
- Declustered RAID for sustained, predictable performance and rapid rebuilds
- No hardware controllers; disk management and RAID are performed by the GPFS Native RAID (GNR) feature

GSS packaging

GPFS Native RAID (GNR) software is capable of running on a variety of hardware platforms, but to be a reliable solution it needs to be tested with each server and disk tray. The GPFS native RAID software was first delivered on the IBM P775 Supercomputing platform. This hardware has a specially designed disk tray that holds 384 drives. This hardware is excellent but not practical for many customers, so IBM released the GPFS Storage Server as a means to provide this advanced storage technology on standard hardware to more customers.

How does GSS defy physics?

Well, it doesn't. As physics professors will be quick to point out, you can't defy physics. But you can take a completely new approach to solving the problem. GSS incorporates a new approach to solve the issue of long RAID rebuild times with really large disks called declustered RAID.

In a traditional RAID 6 array, for example, a set of disks are grouped together, and data is stored on all of the disks in the set along with chunks of information that can be used to "rebuild" a failed drive (see Chapter 1 for an explanation of RAID 6). These chunks of information are called parity bits. When a drive fails in a RAID 6 array, parity bits on the other disks are read to re-construct the data from the failed drive, which is then written to a new disk. The entire contents of the failed disk must be written to the new disk before the array is again completely protected from further disk failures.

In a declustered RAID array, the number of disks in a group, called a declustered array, is dynamic and the data and parity are across the available disks as the data arrives. In a declustered array each block of file data is cut into what is called a strip. So if you use an 8+2 configuration you get 8 data strips and 2 parity strips the difference is that this striping is independent of the number of available disks. This means that if, for example, you have 58 drives and you are using 8+2 parity, the first block of data is spread over 10 of the drives, the second block of data is spread over 10 other drives, and so on.

So, how does this improve rebuild speeds? Instead of having to rebuild a failed disk as in RAID 6, a declustered RAID array just needs to ensure that all the data that was on the failed drive is copied or computed (parity is computed) on other drives. So GSS doesn't defy physics, it just spreads out the work. Instead of writing 4TB to a single spare drive, the surviving 57 drives just need to swap some data and move on.

A rebuild on a RAID 6 array with 4TB drives can negatively impact the performance of your entire array for 17 to 24 hours. In a declustered RAID array, the impact is less than 15 minutes.

Figure 4-1 shows the hardware that's used in the GSS and the different shipping configurations.

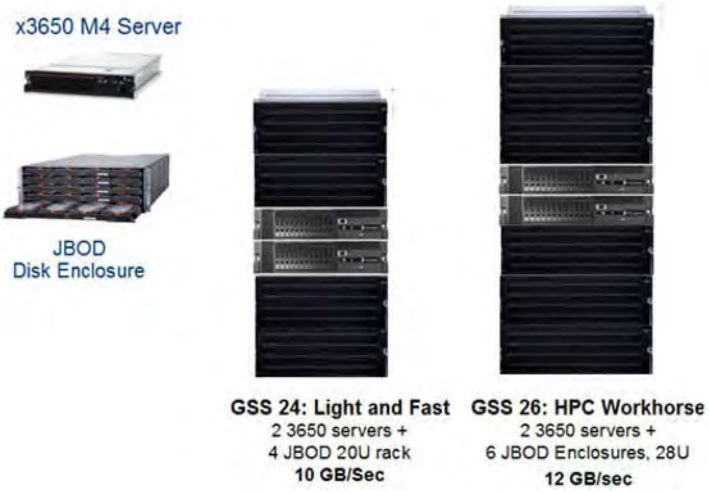

Figure 4-1: The IBM System x GSS.

Bringing Parallel Performance to Fault Tolerance

The speed of recovery along with other advanced data protection features makes GSS much more reliable, even as the number of drives increases. GSS declustered RAID technology is not tied to a set number of drives. In GSS, there are standard configurations, but the RAID technology can work with as few as 11 drives to several hundred drives. This keeps GSS relevant, even with new and different drive packages and types.

Advanced data protection

For a long time, GPFS has been able to rely on the RAID controller to take care of things like making sure what you write to the disk last week is correct when you read it this week. Managing the data all the way down to the raw device comes with additional responsibility and some great functional benefits. One of these responsibilities is to make sure the data hasn't changed since it was written to disk. In GSS, this is accomplished using a data checksum. A *checksum* is a number that's computed by looking at a chunk of raw data. This number is stored in a separate area and used when the data is read to ensure it hasn't changed since it was written. Because GPFS now has that checksum information, it can be used to ensure data integrity all the way from the disk to the GPFS client, thereby protecting data from storage, system, or network errors.

Another responsibility GPFS now holds is to make sure the disks are healthy. Now all storage components are monitored and maintained through the disk hospital. If there are media errors, path issues, or simply disks behaving badly, GPFS responds to these events with a series of actions, including rerouting requests and power-cycling disks.

Hard disks don't report some read faults and occasionally fail to write data, while actually claiming to have written the data. These errors are referred to as silent errors, phantom-writes, dropped-writes, and off-track writes. GPFS Native RAID implements an end-to-end checksum calculated and appended to the data by the client to detect silent data corruption.

If the checksum or version numbers are invalid on read, GPFS Native RAID reconstructs the data using parity or replication, and returns the reconstructed data and a newly generated checksum to the client. Thus, both silent disk read errors and lost or missing disk writes are detected and corrected.

More efficient use of disk performance

Compared to conventional RAID, GPFS Native RAID implements a sophisticated data layout scheme that uniformly spreads (or "declusters") user data, redundancy information across all the 58 disks of a declustered array.

With declustered RAID, you don't have idle spare disks sitting and waiting to be called into service. Spreading data over all available drives including spares, a hot spare becomes hot spare space. So, instead of assigning a drive to be a spare, the RAID software just keeps space free for failure events.

A declustered array can significantly shorten the time to recover from a disk failure, which reduces the rebuild overhead for client applications. When a disk fails, data is rebuilt using all 57 operational disks in the array, thus providing bandwidth that is 6 times greater than the 10 disks in a conventional RAID 6 group.

The disk hospital is a key feature of GPFS Native RAID that asynchronously diagnoses errors and faults in the storage subsystem. GPFS Native RAID times out an individual disk I/O operation after about ten seconds, thereby limiting the impact from a faulty disk on a client I/O operation. The suspect disk is immediately admitted into the disk hospital where it is determined whether the error was caused by the disk itself or by the paths to it. While the hospital diagnoses the error, GPFS Native RAID uses the redundancy codes to reconstruct lost or erased stripes for I/O operations that would otherwise have used the suspect disk.

Bringing It All Together

GPFS Storage Server is a storage appliance that follows in the GPFS tradition of being adaptable. All of the GNR software exists separate from the existing GPFS functionality. This means that in addition to the new RAID capabilities of GNR you can continue to use all of the GPFS features including the ability to mix any storage into the cluster. You can use NSD's on a GSS in an existing cluster and even in an existing file system.

Already proven in the field, GPFS Native RAID (GNR) has demonstrated the ability to succeed under very demanding I/O workloads. GSS is a key next step in a series of next generation software-based storage technologies.

Chapter 5

Software Defined Storage in the Real World

In This Chapter

- Reviewing high-performance computing in higher education
- Discovering storage solutions for energy and telecommunications
- Building better systems with engineering analysis
- Examining solutions in the life sciences industry
- Taking a look at media and entertainment uses
- Studying research and development applications

IBM's General Parallel File System (GPFS) and server and storage systems are ideal for many different workloads in various industries. This chapter takes a peek into several organizations around the world that are using GPFS to solve real-world challenges.

Higher Education

Higher education institutions must support research in many fields over years of study. This research requires high-performance computing (HPC) systems and storage arrays. The United Kingdom's Durham University Institute for Computational Cosmology is one example of a research institute with extreme computing and storage needs.

Established in 2002, Durham University's Institute for Computational Cosmology (ICC) has become a leading international center for research into the origin and evolution of the universe, using HPC clusters to simulate cosmological events and answer some of the most fundamental questions in science: What were the first objects in the universe? How do galaxies form? What is the nature of dark matter and dark energy? Where does the large-scale structure of the universe come from? What is the fate of the universe?

Following the success of the original DiRAC (Distributed Research utilizing Advanced Computing) initiative, the UK government allocated a second round of funding for a new generation of HPC investment, known as DiRAC 2. The ICC made a successful bid to create a new cluster, which would be known as COSMA5.

“The strategy behind DiRAC 2 is that the clusters are designed for different types of research problems,” says Dr. Lydia Heck, Senior Computer Manager at the ICC. “COSMA5 is optimized for ‘big data’ projects, while, for example, the Leicester cluster and the Cambridge HPCS cluster are designed for optimal high-performance inter-process communication, and the Cambridge Cosmos system for shared memory. Running the right code on the right platforms makes a huge difference to efficiency, as well as enabling the UK’s universities to tackle a wider range of projects. To make sure COSMA5 would deliver the petascale storage and data-processing power required for our role in DiRAC 2, we needed to put the best possible infrastructure in place. We performed a full EU procurement exercise to select the right hardware and implementation partners to help us build the cluster, and a joint proposal from IBM and OCF achieved the highest score of all the bids we received. Their success was not just based on price and technical capability, but also on support and service levels — and they had good references to support their proposal, which made us confident that they could deliver what we needed.”

“COSMA5 is four times as fast as the previous-generation COSMA4 cluster — achieving sustained performance of 126 TeraFLOPS at 91 percent efficiency on the LINPACK benchmark,” says Dr. Heck. “It is also 28 times as fast as the old air-cooled COSMA3 cluster, which we have now retired and removed from the machine room. “Replacing COSMA3 with a more powerful water-cooled cluster means that our total HPC landscape is much more energy efficient: The machine itself not only uses considerably less electricity per FLOP, it also requires no air conditioning.”

The ICC now needs no air chillers; all of the cooling requirements are met by the iDataPlex water cooling system. As a result, the machine room's power utilization effectiveness (PUE) score has been reduced to around 1.2 — which means that more than 80 percent of the electricity used by the room goes directly to the clusters themselves, rather than powering ancillary systems.

COSMA5 has also seen the ICC adopt new software for cluster management: IBM Platform HPC, which includes IBM Platform MPI for communication between processes and IBM Platform LSF for workload management.

"The IBM Platform Computing software is more stable and user friendly than our old open source software," comments Dr. Heck. "One of our objectives is to sell HPC capacity to business customers, and we feel that these tools will put us in a better position to deliver professional levels of service."

Meanwhile, the scientific community is already working with COSMA5. A cosmological research project known as EAGLE is using the cluster to model galaxy formation. The UK magnetohydrodynamics community is also harnessing the cluster to investigate star formation and interstellar clouds.

Dr. Heck concludes: "COSMA5 is playing an active role in the advancement of cosmology and astronomy, helping the UK maintain its position as a leader in these areas. We're excited by its potential to support academic and commercial science projects across the country and around the world."

Energy

The energy industry has massive processing and storage challenges. For example, the Oil and Gas (O&G) companies explore vast, remote areas around the world in search of new oil reserves. Seismic imaging and ultrasound data enables these companies to map what is beneath the surface — on land and sea.

Sui Southern Gas Company (SSGC) Limited implemented IBM business analytics and data warehousing solutions to help tackle its data processing and management challenges. SSGC is Pakistan's leading natural gas provider, with a transmission pipeline network more than 3,220 km long and a distribution network that spans more than 1,200 towns. The company has more than 2.2 million industrial, commercial, and domestic customers and supplies approximately 390 million cubic feet of natural gas each year. SSGC also owns and operates Pakistan's only gas meter manufacturing facility, with annual production capacity of more than 750,000 meters.

To manage its transmission and distribution networks, SSGC had more than 5,000 paper maps and was steadily adding new maps as the scope of its operations grew. The day-to-day handling, storage, and maintenance of these maps were a significant drain on resources. More important, the information they contained was outdated almost as soon as they were created, and there was no consistent scale or format for the maps. Reconciling operational data with mapping data to get a clear picture of supply-and-demand issues was practically impossible: Not only were the maps inconsistent and difficult to use, but also there was no single source of reliable data.

SSGC wanted to be able to visualize data by geographic location, to improve its efficiency in matching supply and demand, and to help plan its ongoing corporate expansion. The company also wanted better visibility of gas leakages and unaccounted-for gas (UFG) so that it could close the gap between the value of gas delivered to the network and the value of the revenues generated from its sale.

SSGC implemented a business analytics warehouse solution based on the IBM Smart Analytics System 7700 (ISAS 7700). The first stage of the project was the creation of a consolidated data warehouse. In the second stage, Infosphere DataStage was implemented to enable seamless integration with Cognos Business Intelligence Reports and Dashboards for routine analysis and reporting.

The IBM solution aggregates geographic data from more than 5,000 maps with data on assets (pipelines, fittings, depots), road and rail networks, and data from SSGC's ERP and billing systems. By combining real-time operational and current geographic information, the solution enables timely, sophisticated analytics that reveal previously hidden insights. Uniting operational data with geographic data brings reporting and analysis to life, enabling SSGC to visualize supply and demand issues as never before. What-if analyses enable the business to test new ideas faster and understand the results with greater clarity, supporting key business transformation initiatives.

The IBM Smart Analytics System 7700 solution is a pre-integrated and optimized stack of data warehouse management software, analytics tools, storage, and IBM Power Systems servers. Designed to deliver optimal performance and flexibility for business analytics, the IBM solution was also fast and easy to deploy. As a complete business-ready analytics solution, the IBM Smart Analytics System 7700 offers simple deployment and operation, yet provides a rich and complete stack of technologies with the resiliency required for SSGC to analyze data with confidence and focus on business issues rather than on platform integration.

The key components of the data warehouse are IBM InfoSphere Warehouse, IBM InfoSphere Business Glossary, and IBM InfoSphere DataStage. The IBM Smart Analytics System 7700 deployed by SSGC is a pre-integrated stack of data warehouse management software and analytics tools, with an IBM System Storage DS5300 array and an IBM Power 740 server with IBM POWER7 processors. Tuned to deliver optimal performance and flexibility for business analytics, the IBM solution was also fast and easy to deploy, enabling SSGC to focus on business issues rather than on integration.

The combined GIS and analytics solution provides a web portal that delivers detailed, multi-layered digital maps overlaid with business and customer information for improved decision-making. Users can generate up-to-date versions of pre-built reports tailored to meet the needs of different job functions and management levels. Executive dashboards provide at-a-glance views of performance, and managers can use the intuitive drag-and-drop interface of IBM Cognos Report Studio to create their own tailored reports.

The IBM solution gives SSGC faster, more accurate, and more comprehensive reporting and analytical capabilities, helping the company to identify hot spots for leaks and pilferage, to visualize supply-and-demand issues, and to undertake what-if analyses to plan more efficient processes. The creation of a data warehouse to store operational data for analysis and reporting has given SSGC a single version of the truth, improving the consistency and reliability of reporting. Thanks to the data warehouse, SSGC can also now easily leverage information from a wide variety of sources, providing a richer view of operational performance across multiple different dimensions and a better understanding of customer behavior.

In SSGC's old environment, resources were assigned manually to workloads — all that happens automatically in their new IBM software defined environment. User-driven reporting with real-time data updates has replaced the laborious, time-consuming creation of static reports. Different views and dashboards are provided to each different type of information consumer within SSGC. With a single trusted version of information in a centrally managed environment and the ability to geo-reference data, SSGC now has a scalable integration platform to handle the variety of data, and dramatically improve the speed and clarity of its insight into supply and demand issues.

Engineering Analysis

Engineering analysis enables design engineers to analyze the individual components of a complete system, applying scientific analytic principles and processes, to understand its operation.

Infiniti Red Bull Racing uses a HPC cluster, featuring IBM GPFS, to power its high-performance computing infrastructure for both design applications and near real-time race analytics, giving the racing team the edge it needs to design and run the best cars on the track.

While races may be won by drivers at the wheel, building a championship-winning Formula One car requires the ongoing efforts of designers, scientists, and engineers, backed by crucial support from a vast array of partners.

Al Peasland, Head of Technical Partnership at Infiniti Red Bull Racing, elaborates, "Formula One regulations are becoming more stringent every year, with governing bodies limiting the amount of wind-tunnel and on-track test time we can use. We rely heavily on virtual analysis and simulations, such as computational fluid dynamics (CFD), to form the backbone of our testing and development work."

With CFD, Infiniti Red Bull Racing can perform virtual wind-tunnel testing on new car designs as a first step to determining the impact of design changes on a vehicle's aerodynamics. Simulation is a critical factor in analyzing design improvements and requires huge amounts of processing power.

To support top performance for design and analytics applications, Infiniti Red Bull Racing uses an HPC cluster, featuring IBM GPFS, which enables high-speed file access to applications running on multiple nodes of the cluster.

The Infiniti Red Bull Racing team knew that as demand for the data generated by the HPC cluster took off, managing the workload would be essential. After initially examining open source alternatives, the team turned to IBM Platform Computing to make better use of its enormous compute resources using intelligent scheduling and resource allocation.

Nathan Sykes, CFD and FEA Tools Group Team Leader at Infiniti Red Bull Racing, remarks "With the IBM Platform LSF family of products, we have seen a 20 to 30 percent reduction in the time it takes to complete simulations simply by being able to design complex, interdependent workflows and schedule individual jobs automatically. This means that we can continually run more and more analyses significantly faster than before. As a result, we can redesign, model, and build a new Formula One car in about half the time that it used to take in the past."

With IBM Platform Process Manager, the team can manage and control the end-to-end process flow and submission of jobs to Platform LSF. The solution is used in conjunction with IBM Platform Application Center, which supports flexible, web-based submission of simulation workloads.

This sequence of jobs forms an overall design workflow, providing a consistent and defined path for the simulations submitted by the designers. During the process, valuable application licenses are managed with Platform LSF License Scheduler, which ensures optimum utilization of expensive software resources.

"The performance improvements that we have gained with the IBM Platform LSF family allow us to do much more in shorter timeframes," comments Nathan Sykes. "We can design and test more components, and do so very quickly, which gives us a greater chance of making the design breakthroughs we need to build the fastest and most aerodynamically efficient cars on the track."

Building on its investment in the IBM Platform Computing family of products, Infiniti Red Bull Racing is currently working to deploy IBM Platform Analytics. Once fully operational, the advanced analysis and visualization tool will help turn massive amounts of workload data into valuable insight and provide powerful reporting capabilities.

In addition to using virtual testing to optimize car design, Infiniti Red Bull Racing takes advantage of analytical applications to power faster decisions on the racing track. For this, IBM Platform Symphony supports the processing of in-race sensor information from the car and track to return recommendations for the team on tires, designs, and strategy.

“We have seen an increase in performance with the analytics work we are doing, both in terms of volume and speed,” notes Christian Horner. “Instead of it taking days or hours to run a simulation, we are now completing them in minutes. This means that we are able to get results to the track side much faster, which allows the team to make intelligent adjustments and decisions during the course of a race weekend.”

Infiniti Red Bull’s software defined environment optimizes their compute, storage, and network resources so they can adapt to the type of work required — from structural analysis and fluid dynamics in the race car design process — to real-time analytics driving tactical decisions during a race.

The partnership between Infiniti Red Bull Racing and IBM Platform Computing has helped to fuel impressive results for the racing team and will serve as a solid foundation for future success.

Christian Horner concludes, “IBM Platform Computing has played a huge role in speeding up design, simulation, and analysis, and now forms an integral part of how Infiniti Red Bull Racing develops championship-winning Formula One cars. For an energy drink company like Red Bull to take on the might of some huge engineering companies and beat them is something that we are all fiercely proud of. We rely on partnerships such as the one with IBM Platform Computing to power engineering and design excellence, and to sustain the kind of success that we have enjoyed in recent seasons.”

Life Sciences

The life sciences industry encompasses numerous fields including agriculture, biochemistry, food science, genetics, health and medicine, medical devices and imaging, and pharmaceuticals, among others.

The University of Bonn uses IBM technologies, including clustered IBM BladeCenter servers, IBM System Storage arrays, and IBM GPFS, to support critical research in the field of genetics.

The University of Bonn's Institute of Medical Biometry, Informatics and Epidemiology is a world-leading medical research center in the field of medical genetics research. The Institute makes significant use of computational models and carries out complex statistical and analytical calculations on large data sets. To help gain new insights into important research topics, the institute had set up a small cluster solution with 14 nodes. As the number and scope of projects and the amount of data grew by orders of magnitude, this solution became increasingly outdated. The Institute looked for a cost-efficient yet powerful computational platform that would enable more complex operations at higher speed.

With limited space available in its server room, the Institute required a high-performance computing system with a small physical footprint. The second key requirement was for high-speed connectivity between the computational cluster and the data storage environment. Finally, the Institute needed a highly scalable file system that could handle the enormous — and growing — amounts of data that genetics researchers expect to be able to manipulate.

"The only file system we knew on the market that would be capable of scaling as required was IBM GPFS (General Parallel File System)," says Waldemar Spitz, System Administrator at the Institute of Medical Biometry, Informatics, and Epidemiology.

The Institute deployed a total of 26 IBM BladeCenter HS22 servers and 8 IBM BladeCenter LS42 servers, providing 34 cluster nodes with 504 processor cores in total, supported by 1.8TB of main memory. Two IBM System Storage DS3400 arrays provide a total of 40TB of storage capacity for the IBM BladeCenter cluster — and the Institute is actively adding more capacity.

To satisfy the institute's networking and storage performance needs, IBM recommended IBM System Storage DS3400 devices, connected via Fibre Channel to IBM System x3650 M2 servers, set up as IBM General Parallel File System I/O servers, which are linked via InfiniBand connections to the high-performance computing cluster.

Today, researchers in two teams use the computing resources provided by the IBM solution and the Institute also makes the resources available to guest researchers from around the world. New research enabled by the IBM cluster covers all aspects of genetically complex diseases and population genetics.

The researchers not only benefit from the massively increased performance, but also experience fewer interruptions. "The new IBM cluster solution runs in a more stable manner, especially under high load," says Spitz.

The cluster solution from IBM has also made management much easier for system administrators. In the past, the team used customized scripts to manage its cluster. Today, the Extreme Cloud Administration Toolkit provides standardized tools and methodologies for cluster administration, increasing the systems management efficiency at the Institute. This is a particularly important benefit, because it frees up researchers from routine administration and enables them to focus on science.

The IBM BladeCenter cluster processes significant volumes of useful medical research data, evidenced by the fact that the Institute is now expanding its storage environment with an IBM System Storage DCS3700 array and an IBM System Storage DCS3700 Expansion Unit to 400TB of net capacity. Designed for applications with high-performance streaming data requirements, the IBM DCS3700 offers optimal space utilization, low power consumption and high performance. With up to 60 SAS drives in just 4U of rack space, it can reduce operational costs for capacity-intensive applications. And with up to 4000 MBps in sustained drive reads, the IBM DCS3700 storage system is equally adept at delivering throughput to bandwidth-intensive applications.

The use of GPFS will make it easier for the Institute to continue expanding the storage environment, as the file system has practically unlimited scalability.

Media and Entertainment

Media and Entertainment (M&E) companies create and store tremendous amounts of data through various stages that include capture, post-production, editing, and archiving.

Speicher M1 GmbH, based in Bremen, Germany, offers cloud-based media asset management services. The company provides a collaboration platform to host and manage video data for customers in the movie and video industry. Secure archiving on durable media is crucial to protect media assets and enable long-term monetization of movie footage. The industry faces enormous challenges around archiving and indexing large volumes of digital video data over long periods, prompting many companies to outsource data hosting.

Peter Flory, Managing Director of Speicher M1 GmbH, comments, “Our customers typically work in large, distributed teams. Providing fast access to video data and supporting effective collaboration among team members are key priorities, yet most managed cloud solutions offer data storage only.”

Speicher M1 saw a gap in the market for an end-to-end platform combining digital media storage, management and distribution. There was an opportunity for the company to distinguish itself from competitors by offering a cloud-based service for full-lifecycle management of media assets. Speicher M1 recognized that capturing as many new customers as possible would also depend on competitive pricing. “Our research showed that many of the top service providers were charging in excess of one hundred euros a month per terabyte of data,” notes Flory. “We aimed to offer our full solution at a much more competitive price.” Speicher M1 decided that IBM offered the optimal platform for rapid, secure, and cost-effective management and deployment of cloud-based applications and services.

The cloud platform supporting the new solution is based on four IBM System x 3650 class servers and an IBM System Storage DS3512 Express with an IBM System Storage EXP3512 Express Storage Expansion Unit.

Speicher M1 implemented an IBM System Storage TS3500 Tape Library with IBM TS1140 Tape Drives to provide long-term archiving capabilities. The company uses IBM Tivoli Storage Manager to provide hierarchical storage management for its tape storage, fully automating video archiving on tape drives. The IBM servers and storage systems are connected using IBM System Storage SAN24B-4 Express SAN switches featuring fast 8 Gbps Fibre Channel links. For improved performance and high availability, Speicher M1 operates its systems in a cluster configuration, running Red Hat Enterprise Linux and using IBM GPFS.

The cloud-based platform provides a host of industry-specific features which support efficient processing and collaboration among distributed teams. These include advanced search and filter options that enable customers to easily navigate through extensive footage. Customers can access the service via an online portal, making it quick and easy for teams to manage, edit and market projects regardless of location.

Speicher M1 has gained a cost-efficient archiving and media asset management platform, which will help it to better support customers in the movie and video industry. Combining different storage technologies allows the company to offer its customers an integrated, high-performance solution, at a significantly lower rate than competitors.

Speicher M1 now has a fast and secure end-to-end media asset management platform running on a private cloud, which enables a full-service solution at a fraction of the price of competing storage-only solutions. Powerful software defined, policy-driven automation dynamically moves large media asset data through tiered storage pools, based on performance or cost requirements. Data migrated to tape remains visible in the file system and is directly accessible to end users while lowering storage costs by 80 percent and giving Speicher M1 a significant competitive advantage in their service offerings.

"By using a mix of IBM disk and tape storage, we can balance price and performance, improving the cost-efficiency of the solution significantly," says Flory. "This approach allows us to offer a high-performance solution at a competitive price, which we believe puts us ahead of our competitors."

The new solution provides support throughout the entire media asset lifecycle, and features easy integration with web shops and other third-party systems. This enables media companies to process, distribute and market their videos with minimal effort.

"Our customers can preview footage online without downloading large video files," says Flory. "They can categorize and discuss clips while working on the movie. Finally, they can share the finished movie with partners, host online screenings or even connect their archive directly to web shops. This provides companies with one central platform for all video production and distribution needs — without large upfront investment costs."

Research and Development

Research and development (R&D) activities in practically every industry generate large amount of data and have intensive computational requirements.

Jülich Supercomputing Centre (JSC) is part of Forschungszentrum Jülich GmbH, the largest research center in Germany. JSC supports research into a wide range of fields including fundamental physics, life sciences, climate change, and energy by providing the powerful infrastructure and technical expertise needed to run large and complex simulations quickly and effectively.

JSC had to put major effort into managing its previous storage infrastructure, which relied on expensive "hardware defined" data protection in the form of dedicated RAID controllers. Replacing that with a software defined storage environment (IBM GPFS Storage Server) allowed them to utilize the abundant computing power available across the GPFS cluster to be redirected on demand, to dynamically support RAID rebuilds at much higher speeds and at half the cost, compared to conventional hardware defined storage systems.

To remain at the forefront of global research, JSC deployed an IBM Blue Gene/Q supercomputer with 28 racks, at the time the fastest and most energy-efficient supercomputer in Europe. The center uses the IBM supercomputer for research into complex systems such as the human brain.

Modeling complex systems requires huge amounts of data — especially the so-called "scratch data" that is used during simulations, but not stored for the longer term. JSC must provide the IBM supercomputer with reliable, rapid access to scratch data.

JSC also needed a new storage infrastructure to match the capabilities of the IBM supercomputer. The most important requirements for this new storage infrastructure were reliability together with high bandwidth and capacity.

JSC decided to implement an IBM System x GPFS Storage Server (GSS) solution, eliminating the need for separate physical storage controllers. GPFS uses GPFS Native RAID software to deliver not only outstanding throughput, but also extreme data integrity, faster rebuild times, and enhanced data protection.

Chapter 6

Ten Ways to Use Software Defined Storage

In This Chapter

- Looking at innovative uses for software defined storage in your organization

There are many different computing and data challenges, each with a unique set of requirements related to scale, performance, retention, access, availability, and capacity. This chapter give you ten (okay, seven) innovative uses for software defined storage using IBM GPFS (General Parallel File System) that can help you solve your storage challenges.

Improve Application Performance

Data-intensive applications are defined by the fact that they need to read or write a large amount of data to get the job done. Speeding up data-intensive applications is easy, use a faster storage infrastructure. In reality, it may not be so easy because many storage solutions don't scale efficiently. When the storage device does not scale the only way to improve I/O performance is to use many small data containers and modify your applications to utilize these separate containers concurrently. If you don't spread the data for the project across multiple containers adding containers can create data hot spots. A data hot spot occurs when you have a high level of I/O required by a set of data on a single NAS appliance. GPFS eliminates hot spots by spreading data across all of the available storage hardware.

GPFS is designed to support I/O workloads run by a large number of processes and servers concurrently accessing a common set of file data. GPFS stores and accesses file-based data in blocks that are striped across many disks. Striping data provides the best performance for whatever set of files you are using because you can get the full performance of the underlying storage.

For many storage products you have to use separate containers to provide better overall throughput. The problem with this approach is that you partition up your data, greatly complicating management tasks, and it is difficult to keep all of your hardware busy. For GPFS, the opposite is true. GPFS allows you to fully leverage the performance of all of the underlying storage hardware. It does this by spreading the file data over all of the available storage all the time. This means that you don't have idle disks and more importantly not wasting money.

For example, in the entertainment industry, any time a celebrity makes the news their fans want to download pictures and movie clips of that celebrity, causing a set of data associated with that celebrity to become "hot." To support these types of bursty workloads, some storage products use techniques such as file replication or RAM caching.

The problem is that replication-based products require extra space for files and extra bandwidth to make the copies. Also, in the case of streaming media, the load against each file varies with time, and replication has to be done continuously. And RAM caching is cost-effective only for very large servers and thousands of parallel media streams.

Making all of this work takes advanced data integrity and sharing methods. GPFS is very "fair," providing equal access to the data for all servers and processes. The efficient use of storage space along with fault tolerance ensures data integrity in GPFS-based solutions.

Metadata in GPFS is distributed similar to data. There are two aspects to metadata attribute storage and data consistency. As with other data, metadata is spread across all available storage and metadata management is distributed across the entire cluster. Also, many metadata-intensive workloads perform much better with GPFS, leveraging its distributed metadata and load balancing features. Applications that require dynamic load balancing need a file system that has excellent I/O performance and is very reliable. GPFS performs like a local file system, with the added advantage of flexibility, increased scalability, and the reliability of a clustered file system.

GPFS enables all of the servers in a cluster to access all system data equally and metadata in parallel. This improves performance for metadata-intensive applications by speeding up application I/O operations. Since GPFS allows any server in a cluster to read from or write to any of the disks, applications that perform concurrent I/O can achieve very high data access rates.

GPFS is POSIX compliant, so you don't need to modify applications to implement an enterprise class file-based storage solutions layered on top of GPFS. Applications can leverage the concurrent data access capabilities of GPFS and achieve horizontal scaling and load balancing. Because the file system is distributed over all the GPFS servers, you can run the same version of a web server application, for example, which is accessing a database on all GPFS cluster servers. This allows an update to the database to be seen on all servers as soon as one server inserts a record.

Many application workflows, for example, in weather forecasting and financial analysis, operate on a large number of individual parts that run as parallel jobs. If one component fails, it could impact the total time taken by the job and reduce application performance.

Using GPFS, I/O overhead can be significantly reduced — especially those dealing with storing and retrieving the application state data from a checkpoint — in order to restart the failed server (or replace it with a new one) and resume the job. With the GPFS massive single namespace feature, you can easily add a new server to replace a failed server and restart only the failed job from the last saved checkpoint. In a GPFS-based storage infrastructure supporting a multi-part application running on multiple servers, the application data is not tied to any single server.

GPFS provides a highly scalable, low-latency, high-performance, reliable file system for large-scale storage infrastructures, with capabilities for distributed parallel data streaming and no single point of failure. It's not uncommon to have GPFS file systems of a petabyte or more, containing hundreds of millions of files.

Data Sharing and Collaboration

A single GPFS file system can support a mix of workload types using the same disk or even different types of storage for two files in the same directory.

With the continuing rise of data-driven applications, the gap between performance requirements and system capabilities for storage, data management, and I/O access has become much wider, and the need for distributed storage and processing has grown. Some examples of these data-driven applications include

- Applications that synthesize and process results generated by large scale simulations
- Business transactions
- Online trading and data processing
- Long-running compute and analysis jobs
- Applications that analyze large datasets for
    - Medical imagery
    - Seismic data for oil and gas exploration
    - Industrial design
    - Internet collaboration
    - CRM (customer relationship management)
    - Social marketing trends
    - Market analysis

In many cases, datasets are generated at multiple sites. For example, investigation of alternative production strategies for optimizing oil reservoir management requires large numbers of simulations using detailed geologic descriptions. The output from reservoir simulations is usually combined with seismic datasets that are often stored on different systems and in different locations, in order to better predict the reservoir properties and drive new simulations.

In the automotive and manufacturing industries, product design and innovation often happen across the globe among distributed teams. CAD (computer aided design) and other file-based data from multiple teams located in different geographic sites may need synthesis and collaboration to integrate various aspects of a design. Such data is often fragmented into small storage islands that use different technologies and are managed by independent teams, each with their own local use policies. Cross-site projects are often hindered by this inconsistent approach to technology and management. Also, many of the technologies employed are not designed for enterprise scale computing and teams struggle to expand their IT environments to meet the growing needs of cross-site collaboration.

With an enterprise-wide GPFS environment, you can achieve cost-effective and efficient collaboration with features such as a common file system and massive global namespace across computing platforms. Users can seamlessly access data from any storage cluster server without having to first transfer the data from another location. This streamlines the collaboration process and is more cost effective and energy efficient because enterprises don't have to purchase additional disk space to store duplicate files. By pooling storage purchases, you can build a much larger common shared data infrastructure. In addition, the data is available in a highly parallel manner, making access to massive amounts of data much faster.

Information Life Cycle Management

It's been estimated that as much as 90 percent of all file data is never needed again after it is initially created and, unlike fine wine, most data gets less valuable the older it gets. With petabytes of data and billions of files, it isn't practical to just ask each application group to "clean up their stuff." Plus, government regulations often require you to keep certain data around for many years. Solving this problem requires automation, and GPFS provides a set of tools designed to help you.

GPFS helps simplify storage administration through its policy-based automation framework and Information Life Cycle

Management (ILM) feature set. GPFS policy-based ILM tools can be used to manage sets of files and pools of storage, and automate the management of file data. Using these tools, GPFS can automatically determine where to physically store user data, regardless of its placement in the logical directory structure. Storage pools, Fileset, and user-defined policies provide the ability to match the cost of storage resources to the value of your data.

Storage pools allow you to create groups of disks within a file system. You can create tiers of storage by grouping your disks based on performance, locality, or reliability characteristics. For example, one pool could be solid state disks (SSDs) and another could be more economical SATA (Serial ATA) storage.

A fileset is a sub-tree of the file system namespace and provides a way to partition the namespace into smaller, more manageable units. Filesets provide an administrative boundary that can be used to set quotas, take snapshots, and be specified in a policy to control initial data placement or data migration. Data in a single fileset can reside in one or more storage pools. Where the file data resides and how it is migrated is based on a set of rules in a user-defined policy.

GPFS has two types of user-defined policies

- File placement
- File management

When a file is created, GPFS needs to know where to put it. This is done by using file placement policies that direct file data as files are created to the appropriate storage pool. You can create file placement policies based on anything GPFS knows about a file when it is created, including filename and the user who is creating the file.

After a file has been created, GPFS knows much more about it. In addition to the attributes available when a file is created, GPFS now knows additional information including the size of the file, how long it's been since someone accessed the file, and whether or not it's been changed. Policies that operate on existing file are called *file management policies* and allow you to move, replicate, or delete files. You can use file management policies to move data from one pool to another without changing the file location in the directory structure. One popular use for file management policies doesn't involve moving data at all — you can use it for reporting. The policy syntax is very powerful, allowing you to generate custom reports, for example, on the type of files using the most space. On Linux and AIX, you can use similar tools to get this information, but the policy engine is very fast — it can look at the metadata of millions of files per second.

ILM tools need to have rich features, be automated, and be capable of operating on very large data sets to be useful when you are storing petabytes of data. The GPFS ILM toolset is well suited to this environment and is capable of managing billions of files.

System Administrators, Take Back Your Weekend

Downtime — whether planned or unplanned — is costly. Making sure your applications are always available requires a robust, high availability solution, so applications can continue to run in the event hardware fails, capacity is expanded, or hardware and software is upgraded.

The core GPFS software is fault tolerant. If a server or even a storage system fails, the other servers can continue to access the data. It does this by continuously monitoring the health of the cluster and file system components. When a failure is detected, the appropriate recovery action is taken automatically. Extensive logging and recovery capabilities maintain metadata consistency when application servers holding locks or performing cluster services fail.

For additional protection, you can have GPFS create two or three synchronous copies of the file system metadata and/or the file data. When data is replicated, you can use either copy of the data for reads in an active-active mode. You can even tell GPFS which copy to read from, in order to keep read access local if you are replicating across sites. This can help for read-intensive workloads and to keep certain traffic off the WAN. When both copies are located in the same data center, for example, reads come from both copies, thereby doubling your read bandwidth.

Software defined storage introduces new ways to manage data like file-based replication in an active-active configuration. This gives you flexibility to keep data safe and be able to use it at the same time, across various types of hardware.

Hadoop MapReduce Workloads

There is an industry trend toward taking a MapReduce approach to analyzing large amounts of data, sometimes referred to as big data. The expected cost benefit of using MapReduce is realized by achieving high I/O throughput on very inexpensive hardware. The approach is to leverage the internal I/O performance of a server, while being able to task a bunch of these servers to solve big problems. To run a workload of this type requires storage software that can support this hardware architecture, and provide the right interface for MapReduce applications to find the right data. Both the Hadoop Distributed File System (HDFS) and GPFS are designed to provide a storage platform for data supporting MapReduce workloads, on large-scale standard hardware consisting of thousands of servers.

MapReduce is a programming and data organization model for processing large data sets in parallel on a distributed computational cluster.

HDFS and GPFS both provide the basic storage tools needed for MapReduce workloads, but that is where the similarities end. HDFS is a basic storage solution for Map Reduce, whereas GPFS is an enterprise storage software solution that supports MapReduce. Some limitations of HDFS include

- Centralized master-slave architecture
- No file locking
- File data stripped into uniformly sized blocks that are distributed across cluster servers
- Block-level information exposed to applications
- Simple coherency with a write-once, read-many model that restricts what users can do with data

GPFS features include

- High-performance, shared-disk cluster architecture with POSIX semantics
- Distributed metadata, space allocation, and lock management
- File data blocks striped across multiple servers and disks

- Block-level information not exposed to applications
- Ability to open, read, and append to any section of a file

GPFS includes a set of features that support MapReduce workloads called GPFS File Placement Optimizer (FPO). GPFS-FPO is a distributed computing architecture where each server is self-sufficient and utilizes local storage. Compute tasks are divided between these independent systems and no single server waits on another. GPFS-FPO provides higher availability through advanced clustering technologies, dynamic file system management, and advanced data replication techniques.

In addition, GPFS supports a whole range of enterprise data storage features, such as snapshots, backup, archiving, tiered storage, data caching, WAN data replication, and management policies. GPFS can be used by a wide range of applications running Hadoop MapReduce workloads and accessing other unstructured file data.

In laboratory tests, benchmarks demonstrate that a GPFS-FPO-based (modifies system) system scales linearly so that a file system with 40 servers would have a 12GB/s throughput, and a system with 400 servers could achieve 120GB/s throughput.

Cloud Storage

Cloud storage provides a scalable, virtualized infrastructure as a service, hiding the complexity of fine-grained resource management from the end-user. According to IDC (International Data Corporation), the amount of information in the world is set to grow 44-fold in the next decade, with much of that increase coming from the rise in cloud computing.

Software defined storage systems running GPFS are well suited to address cloud storage requirements given their extreme scalability, reliability and cost efficiency. Software defined storage systems can scale to thousands of servers while supporting hundreds of GB/sec of sequential throughput and can be configured using lower-cost standard parts without the need for costly enterprise-class storage solutions, such as storage area networks (SANs).

Unlike many cloud storage solutions available today, GPFS supports traditional applications that rely on POSIX file APIs and provide a rich set of management tools. A cloud storage stack built on a POSIX-based cluster file system makes it easy to support existing applications by providing a scalable infrastructure that doesn't require application modifications.

Cloud storage is shared across different classes of applications so standard file semantics are important. The standard interface needs to support new workloads including MapReduce style applications so you don't have to use separate point solutions for these workloads. For mixed workload environments which require access to large amounts of data in a cloud environment, GPFS can help leverage standard components, along with its POSIX-complaint interfaces and support for the latest technologies, such as InfiniBand (IB) Remote Memory Data Access (RMDA).

Enterprise Analytics

For enterprise analytics, like detecting credit card fraud or generating customer specific marketing offers, scalability, reliability, and ready access to data is critical. Analytics is a data-driven application that delivers measureable business value.

The IBM Smart Analytics System is a pre-integrated analytics system designed to deploy quickly and deliver fast time-to-value. Engineered for the rapid deployment of a business-ready solution, the IBM Smart Analytics System includes the following features

- Powerful data warehouse foundation
- Extensive analytic capabilities
- Fully integrated, scalable environment

GPFS is a core component of IBM's Smart Analytics offering providing a high availability data storage platform. The IBM Smart Analytics System core warehouse servers have an active-passive high-availability configuration whereby data is available even in the case of server failures. The system can be scaled up as data grows to meet new business needs such as intelligence, smart analytics, and data warehousing.

Reduce storage costs using any hardware with storage controlled in software

Traditional storage systems have become costly performance bottlenecks for enterprises struggling with ever-growing data challenges. This book explains how software defined storage enables organizations to significantly reduce their storage costs while improving performance, reliability, and scalability with any hardware and intelligent software that performs essential storage functions.

- ***Increase flexibility* — traditional storage systems limit your options and lock you in to a rigid and unadaptable solution**
- ***Simplify management* — automated policy-driven storage management makes it easy to implement policies for information life cycle management and other storage administration tasks**
- ***Empower global collaboration* — low latency access to data from anywhere in the world to enable innovation and increase productivity**

Open the book and find:

- **Software defined storage systems that meet your business needs**
- **Performance bottlenecks that exist in your storage infrastructure**
- **Features and capabilities of IBM General Parallel File System (GPFS)**
- **Turnkey software defined storage solutions that are ready to deploy**
- **Innovative IBM GPFS use cases for complex storage challenges in different industries**

Making Everything Easier!™

Go to Dummies.com®
for videos, step-by-step examples, how-to articles, or to shop!

FOR DUMMIES®
A Wiley Brand

ISBN: 978-1-118-84178-5
Not for resale