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.

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.

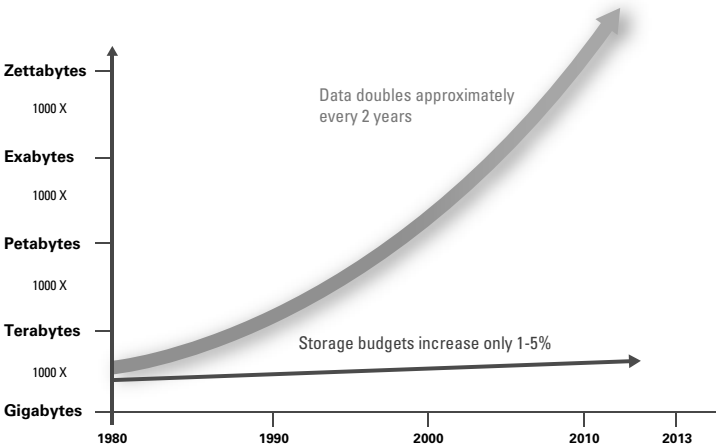

Figure 1-1: Storage requirements are devouring CAPEX and OPEX resources.

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.

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

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).

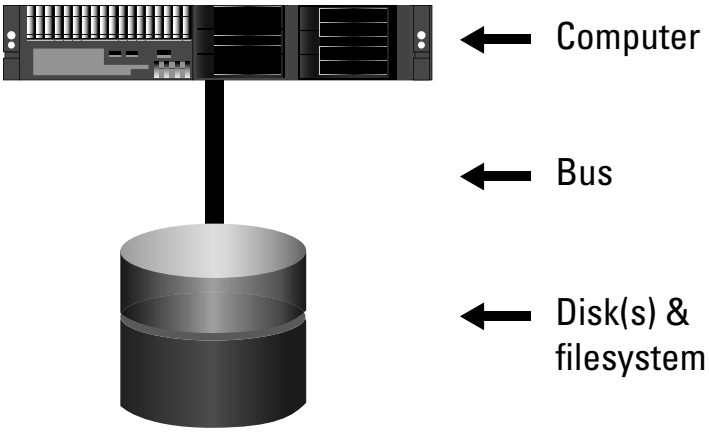

Figure 1-2: DAS connects hard disks directly to the computer or server via a bus interface.

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).

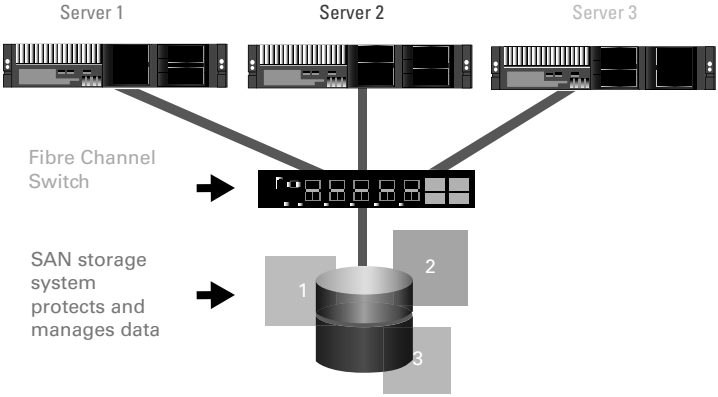

Figure 1-3: A SAN connects a storage array to servers via a dedicated storage network.

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).

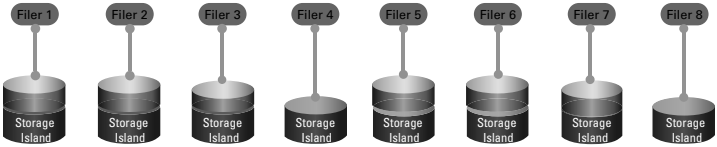

Figure 1-4: Filers often lead to "islands of storage" in the datacenter.

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).

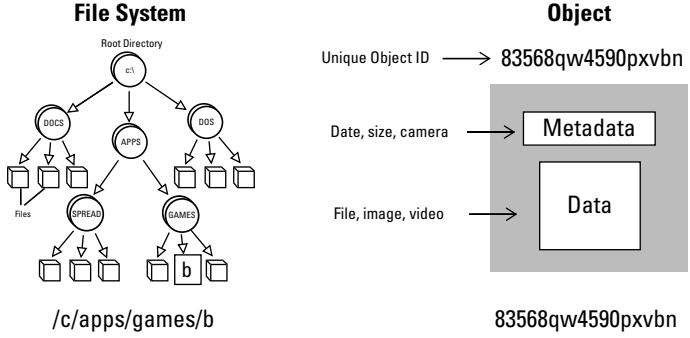

Figure 1-5: Comparing the file system to the object-based storage system.

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).

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.

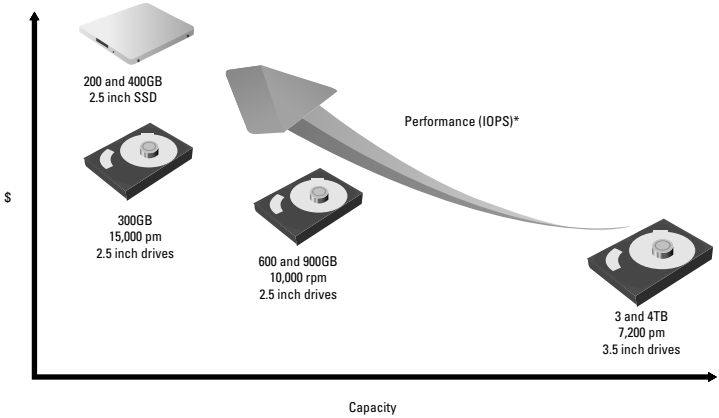

Figure 1-6: Different hard drive technologies require a tradeoff between capacity, performance, and cost.

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.