



Data Center Storage in Education

Storage technologies help your organization streamline its information architecture, prevent data loss and deliver access anytime, anywhere.

TABLE OF CONTENTS

- 2** Storage Consolidation
- 2** Storage Area Networks
- 3** Storage Virtualization
- 4** Data Deduplication
- 4** Tiered Storage
- 6** Backup, Recovery and Archiving
- 7** High Availability
- 7** Manufacturer Options

Executive Summary

Data is proliferating at an unprecedented rate in the education sector and beyond. In 2007, the global volume of new information exceeded the world's available space for storing it. By 2011, annual production of digital data will approach 1,800 exabytes (1,800 billion gigabytes), twice the expected available storage capacity.

Although the unit cost of data storage is declining, increased demand is outstripping potential savings that otherwise might accrue. That exponential growth has implications for K-12 schools and higher educational institutions. At a time of tight IT budgets, data storage is consuming ever-larger portions of limited financial resources allocated to information technology.

In addition to budget cuts and added hardware and software costs, IT budgets are sagging beneath the weight of energy bills — as well as the direct and indirect expenses of complying with more stringent data controls mandated by HIPAA and the Patriot Act.

As organizations rely more on IT, an important question for data center managers is how to manage the increase in data without blowing a hole in already thin IT budgets. This leads data centers to look for technologies that will reduce costs and simplify operations associated with hardware, energy and personnel.

As the volume of data grows, so does the imperative of users to quickly and securely access information. Expansion of online education and other remote services requires data to be reliably accessible whenever students, teachers or staff demand it.

The IT landscape is changing, and the old panacea for data-storage challenges — simply add capacity — no longer works. To better manage information, educational institutions are striving to meet three important goals: consolidate storage and streamline information management architecture; mitigate risk of data loss and corruption; and guarantee timely access of data by authorized users.



Storage Consolidation

Gartner reports that many organizations are moving to physically consolidate storage in data centers, despite costs ranging from \$10 million to \$35 million to build 10,000 square feet of new storage space.

For example, the Indiana University Data Center, dedicated in 2009, cost \$32.7 million. In the K-12 realm, schools in Michigan stand to benefit from a recently undertaken consolidation effort. Since January, state officials have sought industry feedback about plans to build a massive 100,000-square-foot data center.

Envisioned as a public-private partnership, the facility would provide storage and IT support to cities, counties, schools and government agencies statewide. The new center will merge three data hubs that were formed by an earlier consolidation initiative. This previous effort combined 36 centers, saved an estimated \$19 million and reduced office space by 30,000 square feet. Reducing duplicate data systems will create additional savings, officials predict.

The challenge for K-12 schools and colleges is that, like most data-intensive organizations, they are generating data and connecting resources faster than data centers can process them. At times, simply moving massive data sets between physical locations becomes problematic.

Serious About Energy Costs

Data center managers began to get serious about energy costs in 2009, when more than half of the respondents to a survey of data center professionals said that consumption had become a major concern.

Data centers use up to 100 times the energy of other buildings, much of it to run and cool IT components that produce vast amounts of heat. As far back as 2005, data centers consumed 1.2 percent of all electricity produced in the United States, a figure that is undoubtedly higher today.

Among data centers that tracked energy use in 2009, 44 percent showed increases, and one in five reported double-digit jumps. Despite increased awareness, an alarming 28 percent of respondents said they didn't know if energy costs were up or down, according to the SearchDataCenter.com 2009 *Data Center Purchasing Survey Report*.

In the post-mainframe era of networked personal computers and standalone servers, an IT sprawl ensued that made managing data increasingly difficult. Islands of data centers emerged and multiplied to handle storage capacity for discrete operations functions.

When servers ran low on storage, IT administrators simply added more capacity through direct-attached storage (DAS). The practice, essentially throwing disks at the problem, proved temporary and unsustainable.

Although it worked in the short term, adding capacity did nothing to improve storage efficiency or bring order to increasingly fragmented data storage silos. If anything, adding storage in the absence of other advances exacerbated the increasingly thorny problem of how to deal with massive amounts of data.

As data accumulates in poorly organized piles, the challenge of knowing what data is where becomes increasingly difficult, as does the accurate and timely access of information. Data management becomes a piecemeal process that is vulnerable to errors, including missed backups and misidentification of corrupted data. Servers frequently have too much capacity for storing data, or they don't have enough.

In the first step toward consolidation, organizations take an inventory of available storage assets and their use by an institution's various departments and groups. Such audits often reveal storage utilization rates of 30 to 40 percent — and as low as 10 percent in some cases.

Consolidating assets into fewer and larger storage pools can shrink hardware and software needs, trim energy costs and make data management more efficient. Streamlining storage in one or more central locations (as opposed to server-attached storage), for example, makes it possible to switch out servers and perform other such tasks without shutting down the network.

Consolidation can further reduce costs by making it easier for organizations to charge departments for the storage used, a model that fosters accountability and heightens cost sensitivity.

Storage Area Networks

In 2009, an Association for Computer Operations Management (AFCOM) survey of more than 430 data center professionals found that operational growth — not governmental regulation — was the primary factor driving demand for data storage, with more than three of every four respondents citing operational needs as the impetus for expanding storage. Organizations of all types continue to become more dependent on IT tools and electronically stored data to serve their end users.

Leading Causes of Data Loss

A recent survey by Protect Data, an IT consultancy specializing in data backup, found that hardware systems are the leading cause of data loss. Moreover, the exponential increase in storage volume means that data loss continues to rise, despite technological advances in the reliability of magnetic storage media.

Here's a short list of why organizations lose data:

Hardware or system malfunctions: **44%**

Human error: **32%**

Software corruption: **14%**

Computer viruses: **7%**

Natural disasters: **3%**

Source: Protect Data

Worldwide, almost two of every three data centers represented in the survey reported storage needs that had increased dramatically in the past five years. More than 60 percent said they would require more data center space within five years. The AFCOM survey also found that 77 percent of respondents had implemented storage area networks (SANs), whereas 42 percent acknowledged using network-attached storage (NAS).

As the AFCOM survey shows, SANs have emerged as a reliable solution to the challenge of storing and managing data. A SAN offers remote storage capacity to servers by way of an architecture that mimics storage attached to the operating system. A SAN takes the disparate silos of disk arrays that previously had been attached and dedicated to a single application and links them into a high-speed network.

SANs can employ a variety of storage devices, including disk arrays, tapes and optical jukeboxes. They typically are used in combination with a system for storing unstructured file-based data.

SANs have improved efficiency and accuracy by delivering storage as a single unit. Unlike the fractured environment those networks seek to replace, SANs can deliver storage to a large number of servers and manage that storage in a consolidated environment.

Very often, SANs are confused with NAS. Whereas both are connected to a network, NAS is a file-level system that is less robust than SANs that rely on block-based storage. NAS uses Network File System (NFS) or other file-based protocols.

Whereas NAS systems are external to the operating systems they serve, SANs use storage devices such as disk arrays that are attached to operating systems in a way that makes them appear to be attached directly to servers. The architecture of SANs makes it possible to increase storage or otherwise reconfigure storage capacity independent of server cycles.

Storage area networks use storage arrays that are tied together with a "network fabric," similar to how a LAN ties together personal computers. SANs come in a variety of iterations with various costs, performance profiles and technical protocols.

Here's a brief rundown of the main SAN protocols:

FIBRE CHANNEL: FC is a gigabyte technology that migrated from the supercomputer field to become the standard protocol for storage area networks in the enterprise storage environment. FC is relatively expensive. It can use fiber-optic cables, as its name implies, as well as twisted-pair copper wires.

INTERNET SMALL COMPUTER SYSTEM INTERFACE: iSCSI is a networking storage standard based on an Internet protocol. As such, iSCSI can run on existing network cabling, providing cost savings not available with expensive Fibre Channel. IP compatibility also allows storage of data to and retrieval from remote locations.

FIBRE CHANNEL OVER ETHERNET: FCoE is a relatively new standard that adapts high-performance Fibre technology to more ubiquitous and cost-effective Ethernet lines.

INFINIBAND: This technology, commonly used in supercomputer applications, is another SAN option that seeks to compete with FCoE and iSCSI. InfiniBand provides higher availability to stored data, but at a price.

Storage Virtualization

In recent years, storage virtualization has emerged as a significant advancement in data management. In its simplest form, virtualization adds a level of abstraction between storage assets and end users or system administrators. As such, virtualization allows users (and applications, for that matter) to view a system's storage as if all available capacity existed on a single master disk.

The upside is access to storage whenever it is needed, regardless of where the storage physically resides. Virtualization simplifies the creation of storage pools and makes it easier to add new capacity when the need arises. Streamlining those processes also reduces administrative oversight.

Storage virtualization, by providing a unified view of data resources, hides details from users and obviates the need for users to match the size of requested data resources and the bandwidth for conveying them. Software working in the background automatically manages these issues.

Virtualization also enables the practice of tiered storage and thin provisioning, which lets a system administrator allocate to individual users total storage space in excess of the system's actual capacity. In such an environment, sophisticated software operating on the back end of a system manages user storage needs.

With thin provisioning, also known as flex volumes or dynamic provisioning, storage space that has been overpromised to an application is made available to other applications as needed.

Storage virtualization has evolved in recent years, from consolidation projects involving noncritical storage to enterprise production systems, according to the Burton Group. The research group says advancements in virtualization technologies have now made them enticing for critical operations services by enabling high availability, disaster recovery and added flexibility in IT systems and processes.

Despite the upside, selecting the optimal system among many choices isn't easy. Symantec's 2009 *State of the Data Center* report, which surveyed 1,600 organizations worldwide, found that even as storage capacity continues to grow, inefficiencies in its management result in underutilization of costly capacity. Median raw storage reported by organizations in 2008 was 128 terabytes, up from 100TB a year earlier, according to the survey.

Driven to better manage and more efficiently use storage, four out of five respondents to the Symantec survey were exploring storage virtualization.

According to the SearchDataCenter.com 2009 *Data Center Purchasing Survey Report*, 54 percent of IT administrators planned to increase spending on virtualization. Despite tight budgets, that figure is virtually unchanged from last year.

The rationale for investing in virtualization has changed dramatically, however. Whereas administrators previously had characterized virtualization as something that would be nice to have, in 2009 the view shifted to virtualization as a necessity for reducing costs associated with hardware, power and cooling.

Data Deduplication

With predictions for modest economic growth throughout 2010 and continuing sluggishness in many sectors, organizations eager to trim costs are looking to consolidate data. Some 62 percent of

Reshaping the Data Center

In the past, new data centers were designed to handle an organization's needs for 15 to 20 years. A more efficient model is to design for future needs and build only what's needed for five to seven years. The incremental approach reduces operating expenses that would otherwise be spent on unused capacity.

Source: Gartner

respondents to the previously mentioned AFCOM survey reported that they are in the process of consolidating one or more data centers or are seriously considering it.

More than half indicated they would relocate their newly consolidated data center to another existing facility, or build an entirely new one to accommodate the additional requirements. Others are turning to another approach to consolidating data: data deduplication.

The ongoing quest to manage ever-larger data stores with greater efficiency has led to the emerging popularity of this technology. Data deduplication's ability to reduce storage space by 50 to 80 percent and shorten the time needed to back up data has made it a go-to consolidation technology.

Deduplication applications work by seeking out and identifying data that appears in more than one place, saving it once, recording all the locations where that data should appear and making it available on demand.

Deduplication works for program and data files. For example, using deduplication software, backing up 100 notebooks loaded with 100 copies of Microsoft Office would require saving one version of the operating software rather than 10,000 separate versions. The same principle applies to a document distributed to and saved by each of an organization's staff members, whether they total 10 or 10,000.

Seventy-seven percent of organizations in Symantec's data center survey reported that they were working on data deduplication solutions, and seven of every 10 were pursuing replication. Data replication is a way of sharing information among redundant storage devices. The goal is to improve reliability and accessibility and to decrease the likelihood of system failure.

Tiered Storage

In the world of data centers, top-shelf storage comes with a commensurate price tag. Using a Rolls Royce-quality storage device to keep low-value data simply doesn't make sense. The rapid growth

of information has focused attention on storing data in less costly, lower-performance drives at a time when organizations are expected to manage larger data stores with higher degrees of accountability.

The good news is that you have options. Information needed immediately for a critical application is much more valuable, for example, than a redundant backup of data that will be retrieved far into the future.

Acknowledgement of that fundamental inequality has compelled IT managers to embrace tiered storage architecture that uses high-performance assets, such as solid-state disk drives of Fibre Channel solutions, to store high-value data at relatively high cost. In turn, less valuable data is assigned to less robust, yet more cost-effective, means of storage such as Serial Attached SCSI (SAS), Serial Advanced Technology Attachment (SATA) and tape drives.

A quality hierarchical storage system has the flexibility to move data among tiers as the value of data changes. On top of the cost savings, moving less valuable data to lower tiers improves performance of higher tiers that typically are critical to organizations' core functions.

In addition to creating a hardware system to keep data of different values, tiered storage also relies on software to segregate data and optimize its storage. The success of tiering hinges, to some degree,

on how well an organization determines the value of its data. Poor identification or inadequate classification undermines storage tiers.

Another factor to consider when building a storage tier is the operational costs of various storage assets. Technologies deemed low-cost (in terms of initial capital layout for a finite capacity) may be more costly to manage, over time, than higher-end technologies. At times, the true cost of less expensive low-end storage can be higher than that of high-end expensive data storage.

A tiered system can have as few as two and as many as six tiers of differentiation. Categories assigned to various tiers can reflect varying requirements for data protection, performance needs and demand for the data.

The versatility of tiered storage can benefit storage area networks' block-based storage, as well as file-based network-attached storage. Tiers also encompass different data classes, including unstructured files, structure databases and semistructured data, such as e-mail.

Here's a rundown of a common tiered storage configuration provided by GreenPages Technology Solutions, an IT consultancy.

TIER 1 STORAGE: For operations-critical 24x7 databases, file servers, e-mail applications and data warehouses, a redundant, cache-based tiered storage model (called Tier 1 storage) is the best option. The Tier 1 storage model offers quick response times and fast data transfer rates. As such, Tier 1 storage is a great solution for organizations that need to effectively store high-performance data that demands high availability.

TIER 2 STORAGE: For seldom-used, noncritical databases (historical data, for instance), a Tier 2 storage model is a great option, because Tier 2 data can generally be stored on less expensive media in a SAN. Tier 2 storage is a good option for organizations that have a large amount of data that does not require 24x7 availability or extensive backup. Tier 2 storage can also help reduce hardware costs and management overhead.

TIER 3 STORAGE: For rarely accessed data, a Tier 3 model offers further economies of scale, because data can be stored on even less expensive media, such as recordable compact discs. Tier 3 storage is a convenient and simple way for IT administrators to protect large amounts of noncritical data from fire, theft and computer malfunctions.

Delivering High Availability

More than ever, organizations are demanding that data centers have high availability, which is the ability to continue operations when components fail. Here are four elements to consider:

- Disaster recovery requires multiple data center facilities to meet high availability objectives, balancing operations risk and cost.
- Resilience is achieved by bolstering infrastructure via load balancing. Clustering and backup are firmed up via data replication and mirroring.
- Management must be sufficient to ensure system capacity and performance as well as automated service provisioning to sustain availability levels.
- Operations maintain high availability through oversight of processes and continuity of operations planning.

Source: *ThruPoint*

Backup, Recovery and Archiving

Like most aspects of data storage, backup and recovery don't come in one-size-fits-all packages. Putting in place an effective system requires answering some fundamental questions: Why are you backing up? What are your backups doing? What are you protecting against?

Just remember that backing up data is easy; recovery can be hard.

The best way to execute disaster recovery is to prevent disaster loss in the first place, yet many data centers are not prepared to prevent or recover from data lost in a natural or human-caused disaster.

Symantec's most recent *State of the Data Center* report found that just 35 percent of data centers have an "above average" disaster recovery plan. More than one in four respondents concede that their plan needs work, and almost one in 10 report having a plan that is informal or undocumented.

Disaster recovery lags even as organizations are becoming ever more reliant on IT and systems' data. Organizations that run more than 1,000 applications constituted 41 percent of respondents to the Symantec survey. Half of those applications were deemed mission-critical by data center managers.

Human error, the biggest cause of unplanned downtime, is the culprit in 25 percent of failures. Along with staffing shortages, this statistic also explains the zeal with which data centers are embracing automation. Some 37 percent of data centers were understaffed in 2008, and almost half reported that finding qualified applicants is a serious problem.

Distributed data centers exacerbate the lack of staffing, providing further impetus to consolidate data centers and storage capacity.

As high availability of data becomes a requirement for organizations, data centers are embracing innovative ways of securing their data that improve traditional methods of directly backing up large volumes of data. Backing up data while systems continued to run didn't always capture changes to data that occurred during the backup, resulting in aberrations.

To overcome that problem, some backup protocols prevent the writing of new data during the duration of a backup, a technique that amounts to a planned service interruption. For many organizations and their data centers, service interruptions, planned or otherwise, are anathema.

High-availability systems have turned to snapshot backups that use a read-only approach to capturing and backing up data at a point in time, while allowing users and applications to continue writing data to the system's memory.

Another data backup technique, mirroring, lets a system simultaneously write information to multiple hard disks. Mirroring, also referred to as disk mirroring and database shadowing, is used to ensure high availability of data sets and the applications they support.

Archiving is a subtype of specialized backup that uses a different set of IT tools. Storage characteristics of archiving are different as well. Archives typically are less immediately accessible, but tend to be better organized and more secure. It's like the difference between the guaranteed return on a nonliquid static asset such as a municipal bond — safe, secure, guaranteed — and the more fluid yet potentially higher yield of a stock.

Many organizations require keeping data archives for five to 10 years.

4 Performance Considerations

Most storage managers grapple with four major issues when it comes to storage performance:

- **CAPACITY SCALING:** The ability to easily upgrade storage without disrupting the system.
- **PERFORMANCE SCALING:** The necessity of maintaining acceptable service levels as storage capacity grows and the number of supported hosts increases.
- **AVAILABILITY:** The concept of redundancy that ensures failover; the ability to immediately switch to a redundant system when a failure occurs. Failover ensures availability of data despite events that would otherwise make information inaccessible, possibly resulting in systemwide service interruptions.
- **MANAGEABILITY:** The idea that systems should, as much as possible, automate scaling, manage capacity and ensure failover with minimal direct human intervention.

Source: *SearchDataCenter.com*

High Availability

Data storage is only as good as its ability to deliver what you need when you need it. Taken to its logical extreme, that axiom leads to the gold standard of high availability.

To satisfy the demand of end users, some applications require constant and instantaneous availability of data. Service failures because of inaccessible data would be disastrous at the beginning of a college's semester when students are finalizing their schedules, or at the end of the quarter for a K-12 school right before grades are issued.

As organizations rely increasingly on web-based transactions, service disruptions because of inaccessible data or other causes become more costly. This is certainly true for educational institutions that increasingly depend on online instruction. They cannot afford data center errors that could result in service downtime. Schools and universities must ensure that faculty and students have access to online classrooms, course materials and services around the clock.

Critical networks that cannot afford to go down rely on failover systems that continue to run when a data storage node crashes. Systems attain high availability by simultaneously writing data to multiple hard disks that exist within a data center or, for greater security, at a remote site. In such a data mirroring environment, a disk failure would trip an automated switch that would move system access to the next disk, preventing data loss or application failure.

Almost four of every five respondents to the Symantec survey were pursuing strategies of continuous data protection at the time the poll was taken.

Manufacturer Options

EMC is the world's largest provider of data storage platforms. In 2009, EMC released Symmetrix V-Max, Version 8 of its primary product. The core of the product is EMC's proprietary operating software, Enginuity, which manages all components in the EMC Symmetrix storage array. EMC's products are frequently found in large data centers.

HEWLETT PACKARD unveiled its P4000 series of IP storage area networks last year with the iSCSI protocol. The company is marketing the P4000 SAN line as a cost-effective solution to high-end data center challenges. The storage device is scalable and uses autobalance technology to smooth distribution of data across resources. Its storage clustering architecture boasts thin provisioning and built-in integrated snapshots.

Michigan's K-12 Data Consolidation Project

Mirroring a national trend, Michigan's K-12 schools and the state's Department of Education are moving to consolidate educational data. The state DOE merged three data centers into one several years ago, part of a larger statewide consolidation that eliminated 35 data centers.

Driving the movement is a desire to create and leverage longitudinal data systems that promise new insights into school performance. The consolidation has led to more reliable storage of educational data at a lower cost, state officials say. Uptime for servers is now 99.999 percent. Prior to the consolidation of disparate storage assets, determining performance metrics was a challenge.

Consolidation has improved the performance and maintenance of redundant systems, backup capabilities and systems monitoring. Combining assets has also eliminated single points of failure and made it easier to share data among agencies.

Next up is a plan to combine all of the state's data storage into a super-data center. Traditionally, the state's 57 independent school districts have had a great deal of autonomy in the collection and management of data.

However, some districts have created consortia that store student data in mini-data warehouses. The state's superintendent of public education has encouraged the creation of such consortia as a way to improve operational and financial efficiencies.

As a result, the consortia have become more competitive for grants disbursed by the federal government and private foundations. The cash flow stems from a belief in reform circles that measuring educational outcomes is a necessary component of improving educational systems; reliable measurement requires reliable data.

The federal Department of Education's *Race to the Top* fund, for example, disburses \$4.35 billion to improve education quality and results across states. Another grant called "Data for Success" uses information stored in Michigan's Macomb County data warehouse to improve teachers' professional development.

The public-private partnership envisioned by the state could cost upwards of \$90 million. Independent schools could store data at the site, which would further reduce redundant data collection efforts and reduce costs.

NETAPP is the data storage and management company that introduced the world's first networked storage appliance in 1992. The breakthrough made shared storage an affordable reality and popularized network-attached storage. Continuing to innovate its product line, the company produces sophisticated storage devices with terabyte capacity and a reputation for reliability.

QUANTUM markets magnetic tape data storage media. The company's storage applications also include tape automation, data deduplication and scalable file storage software. In the 1990s, it acquired Digital Equipment Corporation, the developer of digital linear tape. Quantum became the largest independent supplier of backup, recovery and archive solutions when it acquired Advanced Digital Information Corporation in 2006.

The Hoosiers' New Data Center

Indiana University's new \$32.7 million data center sprung from a desire to mitigate risk.

At the physical level, the center is designed to withstand catastrophic natural events. A concrete-reinforced bunker protects the university's computers, servers and 2.8 petabytes of data against flooding, power outages and tornados — no small concern in a state frequently beset by twisters. The old data center, housed in a 50-year-old former elementary school, was one strong gust of wind away from disaster.

Weather events are hardly the only foe. The university also seeks to forestall damage of another sort. Recognizing that it's virtually impossible to have a great university without first-rate IT facilities and infrastructure, Indiana wanted to bolster and fortify the electronic backbone that supports the statewide teaching, research and administrative activities of 115,000 faculty, staff and students.

Consolidating storage of data held by campuses across the state lessens the risk of errors (human and otherwise) that threaten data security and operations. Implementing common layers of security and trimming the number of people involved with managing security has the effect of diminishing some risk of these threats.

The Indiana University Data Center, dedicated in November 2009, sprawls over almost 83,000 square

feet, including 11,000 square feet of computer equipment rooms.

One of the largest education data centers in the region, the center houses critical computers, networks and data storage equipment for the entire university. The hub-and-spoke infrastructure transmits information throughout the state by way of Indiana's high-speed fiber optic network, called I-Light.

The center is home to the supercomputers known as Big Red and Quarry, and it keeps records of students' coursework and degrees as well as the institution's financial data. Toward that end, it uses storage area networks, storage virtualization and storage tiering. It has not yet invested in data deduplication.

The center offers backup data space and network connectivity to the state, a benefit likely to be noted by lawmakers charged with appropriating funds to the university, and which may help keep in place state and federal funding sources.

University officials are also hoping that the data center will prove enticing to foundations and federal agencies that fund IT projects. The National Science Foundation recently made a \$10.1 million grant to Indiana University at Bloomington to create a national network of universities' supercomputers. Researchers will use the experimental network to work on large-scale scientific problems.