Storage Considerations in Data Center Design

November 2011

Introduction

Storage is an increasingly important and complex component in modern data centers. Requirements for fast data access combine with other often conflicting requirements for retention, legal discovery, longevity, integrity, accessibility, security, disaster recovery and so on. All these requirements must be carefully balanced.

The recent move toward cloud infrastructures signals an additional requirement--to convert traditional capital expenses (CAPEX) into operational expenses (OPEX). This has the effect of shining a spotlight on the existing operational costs of storage, which brings the power usage of the equipment into much sharper focus than has been traditionally seen. Increasing power and rack densities in modern gear have accentuated this trend.

But the single most important consideration related to storage when planning a data center is the intended use of the data center. Large monolithic data centers being built by giant internet companies such as Microsoft and Google are fundamentally different in model and operation from data centers designed to support existing "systems of record," i.e. database-driven business applications. These in turn tend to be different from data centers intended to support engineering operations. Data centers intended to support cloud-oriented operations tend toward the monolithic model, but have added requirements for switching and virtual storage infrastructure that pull them somewhat away from it. Finally, data centers supporting SMBs and SMEs typically have relaxed data requirements because of financial constraints.

This paper will briefly discuss the various considerations that come into play when designing the storage component of a data center. It is not intended to replace consultations with storage professionals and IT architects, who bring a lifetime of accumulated wisdom and experience to bear on the problems that their customers face. It is intended, rather, to help data center designers ask the right questions.

Storage Considerations

There are many considerations that must be taken into account when designing the storage layer of a data center. We briefly sketch these here. While it is up to experienced storage consultants and IT architects to make the decisions around these considerations, data center designers need to understand the complexity of a data center's storage infrastructure in order to defend against simplistic assumptions that may limit flexibility going forward. In general, the operational power used by storage is only one of many decision vectors in a storage deployment decision, each of them critical in its own right.

Reliability

Reliability is the first leg of the famous RAS (reliability, availability and serviceability) triad. Fundamentally, it means that data, once stored, will not go away or change. Typically, this is reflected in

requirements for RAID protection and other types of redundant data copies—often at distance—to protect against disk and other types of media failures.

Availability

It is of little use to a real-time application to be assured that its data is safe, if it cannot be accessed because of a protocol, equipment or communications failure. Availability is typically measured in the percentage of time that data is available to be served to an application. A five nines storage installation maintains data availability 99.999% of the time, on average. That translates to downtime of less than 5 minutes per year.
Historically a difference has been drawn between planned and unplanned downtime. In today's increasingly 24x7x365 world, however, opportunities for planned downtime are correspondingly rarer, which leads to an increased focus on...

Serviceability

Highly serviceable systems are able to continue to function and serve data even while they are being serviced. Typically this means that, in addition to not having any single points of failure (SPOFs), all components are hot pluggable or otherwise able to be taken out of service and replaced without interruption to processing. It is usually taken for granted that non-disruptive upgrade (NDU) actually may involve a somewhat reduced level of service or performance for some time. As service events are generally more frequent than unplanned outages at traditional data centers, designers must be careful to determine what level of service impairment is acceptable during these events and for what interval.

Retention and compliance

Many industries now carry requirements for data retention in compliance with various legal directives such as Sarbanes/Oxley, HIPAA and so on. In addition, a company's legal department must be consulted to determine what policies must be placed on email and other core electronic services having a data component.

Any kind of data retention requirement usually involves a dedicated records management system and storage that is congenial to it. This may limit the storage system choices for that component of the data center.

SNIA

Legal discovery

It is prudent during a data center design cycle to liaise with the Corporate Legal department to determine its requirements for legal discovery. Depending on the quantity of discoverable data, and the timeframes imposed by the discovery process, certain types of media, such as tape, may actually be found to be unsuitable for long-term storage of said data.

Switching infrastructure

Traditional data centers usually have two different fabrics and switching infrastructures, one for storage and one for host and user connectivity. The increasing use of virtual machines which may be

provisioned and booted many hundreds of times a day places increased load on whichever of these fabrics is used for the task, resulting in a requirement for very high bandwidth links between the
storage hosting the VM images and the hosts that are consuming them. Alternatively, VMs may be hosted locally and invoked from local storage, but this introduces other management and distribution challenges.

Disaster recovery

It is well known that businesses that lose their core systems for more than a certain period of time—often quoted as two weeks—may actually fail. It is less well understood what the costs are of an extended data center outage, but for any business in need of an actual data center they may be assumed to be significant. The most common method of dealing with this risk is to maintain compute and storage capacity at another site that is kept synced to the main data center within some delta of time (often seconds). The ability of the compute and storage infrastructure to fail over to the remote site and continue operations, within the parameters set by the needs of the business, is key to business assurance. Depending on the business critical nature of the data and operations, and type of data center, data center designers may find themselves actually designing two or more data centers.

Storage Technologies

There are numerous storage technologies that impact data center design, some by affecting fabric requirements, and others by impacting power levels or overall energy requirements.

SANs

Storage Area Networks (SANs) evolved in a reasonably linear path from their ancestors, the direct attached disk. Enterprise level open systems computers use a protocol called SCSI

(pronounced "scuzzy") to communicate with their hard drives. The earliest SAN systems aggregated several drives together and gave admins the capability of carving out partitions, called Logical Units (usually called "LUNs" for short), that were portions of the total available space rather than of just one drive. These LUNs were then presented to the host computing systems as SCSI partitions, exactly as if they were ordinary SCSI hard drives. Later, systems began mirroring each drive, using two drives to do the work of one. This allowed both greater performance and the ability to tolerate drive failures without interruption of service. Today, systems use several types of virtualization to allow sophisticated data management techniques such as growing and shrinking of LUNs, mirroring of data to remote sites, and so on. While there is now no perceptible correspondence between a logical block number (LBN) and a physical block number on the target hard drive, the host still sees what looks like a physical SCSI drive and uses it exactly as before.

The second step in the evolution of SAN infrastructures came with the development of Fibre Channel, initially both a fabric (a physical medium, e.g. fiber-optic cable) and a protocol encapsulating SCSI that allows switching and deployment at much larger distances than SCSI itself allows. With this, it became possible for very large arrays of disks to service hundreds or thousands of hosts. Separating the storage from the hosts and aggregating it like this restored at least some sanity to the former morass

of backup and archival activities. Very large tape jukeboxes evolved, able to back up the data from hundreds of servers in a single location.

Modern variations of SAN technology include iSCSI, which encapsulates SCSI over a TCP/IP network on an Ethernet fabric, and FCoE, which encapsulates SCSI using Fibre Channel rather than TCP/IP over an Ethernet fabric. While SAN technologies have their detractors, they remain the dominant storage networking paradigm in the storage market today, particularly for systems of record. Vendors with significant SAN market share in 2010 include EMC, IBM, HP, NetApp, Dell and HDS.

NAS

Network Attached Storage (NAS) has many of the same goals as SAN--consolidation of storage into a central location, removal of the storage burden from host OSes, and so forth. Owing probably more to circumstances of birth than to any architectural imperative, NAS systems evolved in an environment dominated by unstructured data. The SNIA (Storage Networking Industry Association) defines unstructured data as the opposite of structured data, perhaps because structured data is easier to define. In general, data generated by database

applications--systems of record--has a known format specified by a data definition language or schema of some kind. Unstructured data, by contrast, is almost always file based, and can be thought of as all the data generated by office productivity and other user-level applications. Novell and NetApp are usually thought of as the early pioneers in the NAS market. Vendors with significant NAS market share in the enterprise in 2010 include EMC, NetApp, IBM and Dell.

The distinguishing characteristic between NAS and SAN is the location of the file system. In a SAN based system, the host is responsible for mounting and maintaining any filesystems it might need on the raw block storage offered by the SAN. NAS systems, on the other hand, use a file sharing protocol such as NFS or CIFS to create, read, write and delete files in a POSIX-like way; the filesystem resides on the storage server rather than on the host.

In recent years, some NAS systems have tapped a basic insight—that both LUNs and files are essentially contiguous arrays of bytes—and begun offering block storage built on top of the filesystem. In other words, they maintain LUN images as files on their local storage and present them to hosts as LUNs via FibreChannel, iSCSI or FCoE. This allows them to employ sophisticated file-based data management strategies on the "block storage", at the cost of some filesystem overhead. Evaluating this tradeoff is part of a storage consultant's job.

Both SAN and NAS installations require a significant switching infrastructure, and lead to a common three-way characterization of data center equipment into computing, networking and storage components. Estimates of the relative power consumption of these components vary, and shift over time as technologies advance, but a convenient rule of thumb is 60% for computation, 10% for networking and 30% for storage.

Storage hardware components

At a high level, storage for traditional data centers can be thought of as storage arrays comprised of two fundamental parts: the controller(s) and a quantity of disk shelves (or drawers). This allows for a convenient componentized approach to rack space and power sizing calculations.

The power used by a controller is usually dominated by a large memory cache composed of DRAM (Dynamic Random Access Memory). In 2010, a 4GB memory module uses a bit under 2W of power, with caches in excess of 256GB not being uncommon. For this reason, it is important to include cache sizing when determining power loads of controllers.
Controllers are often combined into HA (high availability) pairs for RAS purposes. These may be in active-active pairs or active-passive pairs. N-way configurations are also possible. The details of these configurations are not important from a power standpoint; what a data center

designer needs to know is how many controllers there will be and how much power each uses under normal and exceptional use conditions.

Disk shelves, from a power standpoint, are composed of disks and fans. To save power, fans must be specified as variable speed. This is more significant than it may seem at first glance, as fan power increases quadratically with rotational speed, and a large array with over 1000 disks will likely have a couple hundred fans running to cool them.

Because of the rapidly moving nature of the technology, the disks in an array present the greatest challenges. First, disk capacities double every couple years or so. This trend has been predicted to hit a wall of fundamental physical limits, but the wall keeps receding as new discoveries in physics are made at the research labs. So a "forklift upgrade" of five year old equipment to new equipment may yield densities near 10x the old densities. That means that instead of 10 racks of high-performance disks to store 144TB, one rack may now suffice. The power requirement per rack may be somewhat higher, but even counting in a density tax, a power savings of 80% or more on the disk shelves can be expected. There are other factors too, discussed below.

It is also routine for a rack of drives to weigh more than it did in the past, as customers pressure vendors for ever increasing densities. Designers must be careful to design floor systems sufficient to the loads of modern equipment.

Disk drives

Disk drives come in several flavors, each with different power, capacity and performance characteristics.

SSDs (Solid State Drives) have improved rapidly since their introduction to the market. They generally offer the lowest cost per I/O owing to their silicon-based extremely low latency. They are also the most expensive per terabyte stored, at about 10x the price of lower cost disk drives.

High performance FC (Fibre Channel) drives have been the mainstay of the enterprise storage industry for 20 years. In general, they spin at rates two to three times as fast as slower SATA drives, and have

1/2 to 1/3 the capacity of SATA drives. Since 2008, they have largely been replaced in new equipment by SAS (Serial Attached SCSI) drives. SAS is slightly more performant than FC, because it does away with the Fibre Channel protocol layer. Because of this, however, it is not

suitable for distances, and is therefore mostly used for the backend storage networks that move data between array controllers and their disk drives. In either case, though, a rule of thumb that works well is to consider that high-performance drives use about 6 times the power per Terabyte of raw capacity as SATA. In exchange, they offer greater reliability and greatly increased performance because of their increased rotational speed. They remain critical in high-volume database environments.

SATA (Serial ATA) drives were originally targeted at the consumer market. They have moved into the enterprise market however, as manufacturers were able to increase reliability and extend the lifetimes of these drives. As stated previously, they offer by far the best operational power efficiency per Terabyte of any random access media.

It is the job of the storage consultant and IT architect to weigh the data access needs of the application mix in the data center and arrive at a sane balance between performance, capacity and power utilization.

Architectural implications for facilities

The data center designer's principal interest is in the physical infrastructure required by the storage. In a traditional mixed-use data center this nearly always means two data fabrics: a switched Ethernet infrastructure and a switched Fibre Channel infrastructure. Most designers find that a "star" or tree-like arrangement works best, with small switches at the rack level and director-class switches routing traffic between the smaller switches. This minimizes the amount of physical cabling that must be placed and maintained, at the cost of more switching levels and possible decreased response times. In very high performance requirements, physical collocation of servers and storage and minimizing switching in the fabric is a major consideration, but most loads do not require this.

With respect to the power used by the storage itself, there are two considerations. One is the peak power used by the array, and the other is operational power.

Peak power

Peak power is typically determined by first finding out how many controllers and shelves will be needed in a given system, and then adding up the peak power from the manufacturer's spec sheets to get a whole-system number. This number is used to size the power feed to the equipment. While safe, it is usually too high. First, manufacturers are forced to assume the worst when publishing these specs. They assume fully loaded equipment, and voltages at the bottom of the listed range, which drives up amperage pull.

Therefore, when possible, equipment power during boot should be measured on-site to get an accurate number.

It is possible to significantly lower the inrush current required by a large array through the use of staged booting. If the array itself does not have this capability, the SNIA and TGG recommend the use of PDUs that are able to do staged power-on cycles, providing the additional time required to boot is not prohibitive and the array manufacturer is able to support such a practice.

Care should also be taken to distribute the peak load over the three legs of a three-phase installation equally.

Operational power

The SNIA has published an operational power measurement spec and established the Emerald™ program to maintain a repository of information on idle and active power usage by various products. It is not always clear, however, how to extrapolate from a published number for one configuration to a trustworthy number for another. Designers should check all published operational numbers and calculations against the power calculators offered by all reputable vendors to their customers.

Storage Technologies that reduce operational power

The SNIA has identified a number of technologies, mostly software-based, that can dramatically reduce the overall amount of operational power required to store a given amount of data in a data center.

Parity RAID

RAID 1—simple drive by drive mirroring—has been the gold standard for data protection for two decades. Its obvious shortcoming is that it takes twice the amount of raw capacity and power to store a given amount of data as would seem to be required by merely looking at disk capacities.

Parity RAID improves on this by dividing data drives into stripes and calculating a parity stripe that is stored on another drive. The details of parity organization are unimportant in this context. What is important is that one or more drive failures can be tolerated in a given "RAID group" without data loss: the data on the failed drive is recalculated using the parity information plus the information on the remaining drives. RAID 4 and RAID 5 can survive a

single loss per RAID group (usually 5 to 8 drives). RAID 6 can survive two failures per group (usually about 16 drives).

There are numerous tradeoffs in using parity RAID that storage consultants and IT architects must consider. For data center designers, the savings in number of drives required provide a benefit both in terms of footprint and overall operational power required per TB of required storage.

Recommendations to use RAID 1 (simple mirroring) or RAID 1/10 and cousins (not-so-simple mirroring) should be vetted by IT staff to be sure that the increased operational power is really worth it. In particular, it can be asked whether arrays that are mirrored at distance for disaster recovery purposes really need RAID 1 protection at either endpoint, since a mirror already exists at the other endpoint.

Thin Provisioning

Recall that arrays have traditionally served up contiguous slices of capacity in the form of LUNs. This has the effect of requiring storage to be allocated "up front". Furthermore, because applications believe that LUNs are of the advertised size, they usually are not written to tolerate write failures or running out of space. This leads to paranoia-induced allocations that are larger than necessary. A large fraction of usable space—half, according to industry sources—is made sealed off and unavailable by this dynamic.

As storage has become virtualized, however, systems have acquired the capability to defer the provisioning of each block of storage in a LUN until it is actually written to. These thin provisioning systems can save potentially half of the energy required per TB of required storage, as systems can be sized smaller at original purchase time and scaled up as data needs grow.

Data Deduplication

Especially in backup scenarios, it is common for the same data to be written to a device multiple times. Data deduplication replaces these multiple copies with a source copy and multiple pointers to it, resulting in space savings quoted to be as high as 99%. This to date can only be done on systems with random access media--usually spinning disk drives. Deduplication of online data in primary storage is also possible but the percentages are not as high--a 2010 vendor announcement reports a deduplication percentage of 27% over an exabyte+ of customer data.

Deduplication to near-online disk-based systems has many attractions: especially in companies with legal discovery requirements the ability to keep backup data online and searchable is

invaluable. Smaller companies, which usually do not have the organization and discipline to manage complex tape-based backup systems very well, often also find the simplicity of disk-based backup to be very compelling.
However, larger data centers with sophisticated IT departments are still buying tape backup systems in volume, as of this writing. Tape uses less operational power than any kind of disk-based media, and is preferred from a power standpoint when it meets the business needs of the organization. But data center designers need to understand the other factors that may enter into backup decisions.

Compression

Tape systems have been doing compression for decades, usually achieving a compression ratio of around 50%.

Compression is more difficult on disk-based storage systems, as data is stored on them in block format (typically 4KiB at a time) instead of in a streaming format as is done on tape systems. But systems featuring it have begun to ship. There is little available data on how much power may be saved by this methodology, but in general any technology that requires less raw storage capacity to store a given quantity of data is a potential energy-saver.

Delta Snapshots

There are many uses for point-in-time (PIT) copies in modern data centers, from static copies that can be used as the basis for backups, to copies used for testing and what-if scenarios, to copies of golden images used to run VMs in highly virtualized environments.

Traditional copies, generally called snapshots or clones, have been full copies of their target data sets. But whether writeable or read-only, it is rare for them to have more than 10% or 20% of their data changed in production use scenarios.

Delta snapshots are made using various forms of copy-on-write technology--blocks are only written when new data is written to them. Before this, they are shared with the PIT copy's target, taking only as much space as pointers and metadata that tracks the pointers.

As a result, delta snapshots generally save 80% to north of 98% of the size of the raw capacity required to hold the target data set, per snapshot. Depending on how heavily they are used, the potential energy savings are significant.

Storage "tiers"

Storage tiering is a technology that attempts to place data on the most economical media that is justified by its use pattern. Vendors claim that they can store approximately twice as much

data on a given array, at equal or even lesser capital cost, using this technology. Power usage is generally less than on a pure high-performance array, because both SSDs and SATA drives use less power than SAS or FC drives.

Another practice, similar to tiering in effect, involves very large (Terabyte-scale) flash caches in front of pure SATA storage. This configuration is not suitable for workloads that are heavy in streaming writes, but for the more common random access workloads performance may equal or even exceed that of a traditional array with high-performance SAS drives. Operational power is impacted in two ways. First, the large amount of flash cache uses a significant amount of extra power at full I/O load. Second, the use of SATA drives rather than SAS or FC drives results in a significantly larger storage capacity per unit of energy. Designers need to consult with vendors and their power calculators, and with storage consultants and IT architects to determine what the actual power loads are likely to be.

Storage Power Efficiency

There are three types of power efficiency of interest to users of storage. The first is the electrical power efficiency of the equipment. The second is the I/O power efficiency, i.e. the number of I/Os delivered per watt of power used by the equipment. The third is the capacity power efficiency, which is the amount of data that can be stored per watt of power used by the equipment.

Electrical power efficiency

This is the efficiency that the US Environmental Protection Agency and Power companies are the most concerned with. Designers and architects should pressure their vendors to provide silver, gold or

platinum level power supplies. Which to choose is a function of increased price and potential power saving. The inequality that should be satisfied is

$$newprice - baseprice < \left(1 - \frac{base\_efficiency}{new\_efficiency}\right) \times hours\_used \times \$/KW \times KW\_draw$$

Using the right side of this formula, it can be seen that a server pulling 225W at 60% efficiency will cost about $50 a year more at $0.10/KWh than the same server with an 80% efficient power supply, pulling 168W. Depending on the ROI expected for the investment in increased power supply efficiency, it might be difficult to justify price increases of over $25 to $50 for the more efficient unit, even though $150 would be saved over a typical 3-year depreciation pay down.

SNIA

In general, electrical efficiency improvements yield the smallest returns of the three efficiency types, and as efficiencies get closer to 100%, the returns available from increasing them will become vanishingly small.

I/O power efficiency

I/O efficiency is fairly easy to characterize, but more difficult to specify. This is because to a large extent the full I/O efficiency of a storage system in a typical data center is never used. Large-scale streaming applications and highly loaded database applications are an exception, but in general, a storage server running at about 60--80% of its rated I/O capacity is considered to be well used by its admins. "Scale-out" systems attempt to help with this by making it possible to simply add more servers when an I/O or capacity limit is reached, thus making it more feasible to run "closer to the wall," but system granularities and other effects make it difficult to "thin provision" I/O in the uncomplicated way that one thin provisions capacity.

I/O power efficiency is usually characterized in maximum IOPs per dollar, i.e. IOPs/$. Solid state disks (SSDs), as of this writing (2011) have a significant advantage over rotating disks in this department, though the advantage may be only marginal or even negative for sustained write workloads. Systems with large caches typically see similar benefit due to the electronic storage used in the cache.

Capacity power efficiency

Capacity power efficiency is defined as the number of stored TB (terabytes) per watt of power used by the storage system. Through the use of parity RAID, thin provisioning, delta snapshots, deduplication, SSDs and various forms of tiering, capacity power efficiency may fairly easily be approximately doubled.

Capacity power efficiency is related to pure capacity efficiency, which is defined as the number of stored TB per unit of raw capacity used in the storage system. A mirrored array will have a capacity efficiency of less than 0.5, as half of the raw capacity is used just for mirroring, and there are other system overheads as well. Modern systems employing all or most of the capacity optimization technologies mentioned herein are able to actually store more data than they have raw capacity for, had the data been stored naïvely, thus yielding an efficiency number of over 1.0.

Capacity power efficiency is the most important of the three power efficiency types at present in the storage industry. This is because the predominate form of online storage in the modern

data center is spinning disk drives, and these use 85% or more of their peak usage just to spin the drives in an "idle" state, according to extensive testing at SNIA. In fact, some arrays may use more power at idle than when delivering data, as an idle state may be an indication to the array of a good time to kick off seek-intensive background processing and housekeeping tasks. Given that large arrays may have hundreds of drives per controller, the disk + fan power load generally overwhelms the controller power load in the calculations. So capacity power optimization, in addition to offering the greatest gains mathematically, operates on the largest component in the storage power equation, and hence yields the greatest returns. And since it works by decreasing the number of drives that must be bought and powered, it also favorably impacts the capital cost of the equipment.

SSDs offer more or less ideal power profiles for data at rest, as flash-based storage electronics require little to no power to maintain their state. So in terms of capacity per watt, they excel. In terms of pure capacity efficiency, they have the same profile as rotating storage, assuming the storage system treats them the same as any other disk drive on the system. In terms of capacity per dollar they are still at about 10x what SATA media costs, in 2010. Obviously there are significant tradeoffs to be made, depending on which factor is the most important to the architects and designers of a data center.

Types of Data Centers

There are several fundamental types of data centers. Traditional "systems of record" data centers support often hundreds of traditional database-driven business applications. Large monolithic installations serve immense, globalized, highly parallel internet applications such as those embodied by the major search engines and social media. Cloud-oriented data centers can be thought of as a variation of these, with additional switching and virtual storage requirements. Engineering data centers tend to incorporate a mix of cloud and traditional models. Finally, data centers for smaller businesses (SMBs) typically use a less robust grade of storage because of economic considerations, and therefore deserve special consideration.

Data centers for systems of record

Traditionally, systems of record have used Direct-Attached Storage (DAS). This has the advantage of simplicity in initial setup, as well as very high performance, as the number of protocol and switching layers between the computing engine and the storage is kept to a minimum.

However, as businesses and systems grow, DAS tends to become more and more operationally impaired. Without homogeneity—very difficult to maintain when systems are being rolled in and out continuously—management of the storage becomes a logistical mess. Further,

effective backup and recovery of a myriad device types and patch levels becomes more and more difficult and expensive, in terms of both equipment and administrative overhead. Both NAS and SAN storage systems have evolved to bring some amount of sanity to the situation. Both allow consolidated backup and recovery strategies. From a designer's point of view, the principal difference is the location of the various fabrics. In a pure FC SAN shop the LAN fabric—always Ethernet—doesn't penetrate much into the storage

area, while in a pure NAS shop Ethernet is the only fabric installed. In a DAS deployment, there is little in the way of a storage fabric, except as possibly used to consolidate tape backup.

Remote duplication of data

Systems from the major storage vendors are able to automatically mirror data to remote sites. and perform almost magical recovery actions. While completely flawless disaster recovery actions are still the exception rather than the norm, they are increasingly in demand as data centers grow and equipment failures themselves become the norm rather than the exception.

> *Enterprise Data Center: A data center so large that something in it is always broken.*
> *-- Alan Yoder, NetApp*

No modern data center designer can afford to overlook this fact and the mitigation strategies that accrue around it.

Database storage

Systems of record are mostly database driven. Until a few years ago, databases almost always used raw storage, which meant that only DAS or FC arrays were suitable candidates as a storage type. More recently, however, vendors have improved their NFS implementations to the point that file-based storage now works as well as raw storage for the well-known enterprise-level databases. This in theory frees the database system from doing storage management. However, database vendors have also made forays into storage management on DAS or JBOD, intending to use their knowledge of data layout to improve performance and reliability. All three systems—DAS, FC arrays, and NFS servers—enjoy significant market presence today, and there is little indication that any one of them has or will overcome the others at this point. Designers should therefore avoid involvement in database-to-storage communication format wars and concentrate on manageability, database function and other significant issues that impact the business more directly. The final decision on format will likely affect fabric deployment, however, and so must be taken into account at that level.

Email

Email is one of the more visible and vexing database-driven applications today. Several considerations apply.

First, a robust spam filter may save easily 90% of the data storage requirements of an email system. This is because well over 90% of all email today is spam, according to several studies, and the best place for spam is the big bit bucket in the sky, where it uses no storage and no power. Most users seem able to tolerate the occasional "false positives"--leading to dropped emails that may or may not be important--that all spam filters will occasionally generate. Businesses that cannot tolerate false positives have retention and review requirements that will cut into the savings, but spam filtering remains even then an aggressive green technology.

A second major consideration for email systems is compliance and retention policy, meaning, how long should email be kept before it is automatically deleted? And how quickly must it respond to discovery

requests? There is no standard policy for these questions; emerging case law seems to indicate that the important things are to have a policy and implement it rigorously. The company's legal department is the best originator of the actual policy. In some businesses, it is not uncommon for Legal to insist that all mail be kept forever on searchable media (which may impact the use of tape). In others, 1- 3- and 5-year automatic deletion intervals are all reported.

For sizing, records on existing email volumes must be examined and extrapolations made to determine the likely amount of storage that will be needed over the lifetime of the data center. Obviously, a thin provisioning storage system will be of benefit, allowing capacity to ramp up with storage needs. A storage or email management system that deduplicates attachments and mail messages can also save a significant amount of space and operational power.

Unstructured data

Traditional data centers have also seen increasing volumes of unstructured data generated by office productivity applications. All of the capacity optimization technologies mentioned in the foregoing section—parity RAID, delta snapshots, thin provisioning and so on—are applicable to storage of unstructured data. Designers should encourage storage consultants and IT architects to make maximum use of these technologies to shrink the amount of storage that must be initially deployed and scale it as capacity requirements expand. In making the required calculations, records going back several years of the actual data capacities stored will be of more benefit than any records of raw capacities deployed.

Backup and Archive

All data centers must deal with backup and archive at some level. To a large degree, data is retained in tape archives. From a designer's point of view, the requirement is a medium that can be taken off site and stored at a secure remote facility. Archives are rarely read, and must be managed so that tasks such as destruction of aged-out email may be carried out expeditiously. Woe to the business whose data center relies on archives for the restoration of business in the event of a disaster; restore times run to days or weeks.

There are many issues related to long-term retention, and the field is still in its youth. At present, it appears that a robust virtualization capacity—both at the compute and storage layers—will be necessary to solve the problems, so designers presented with requirements in the long-term area should keep this in mind. In particular, storage virtualization appliances (SVAs) which can be configured to "look like" storage equipment from a decade or more ago, will likely be required. It is possible, however, that these capabilities may be met by outsourcing.

Backups, which are intended to be useful in the case of a restore event, are another matter. Some businesses still use tape for this, but online media, in the form of deduplicating backup appliances or delta snapshots on primary or secondary storage, are now considered best practice for many data centers. Thanks to their highly optimized use of raw capacity, both technologies provide operationally cheap and yet high-speed access to data required to restore broken systems back to health. Both also

are useful in reducing the size and expense of the tape arrays that are still required for archive purposes: the windows required to write data to tape can be expanded without significant impact to business operations.

The above is not intended to dismiss tape as a storage strategy—it offers the best operational power profile and compactness of any storage technology to date. The capabilities of the medium and the requirements for the data must be carefully balanced, however.

DR sites sometimes offer the best path back to health after an outage. But when a business has a robust DR strategy in place, both dedup'ed and snapshotted backups are also useful on an ongoing basis for archive staging and lost file recovery, in addition to their primary business continuance mission.
Fabric and power requirements for backup and archive are too various to usefully describe in a guide such as this one. Designers should verify that the business requirements are fully

considered by their storage consultants and IT architects, and proceed based on their recommendations and the chosen equipment manufacturers' best practices.

Monolithic data centers

Some modern internet companies—Google and Microsoft especially—have been aggressively building a new breed of data center. The design of these data centers is shaped by several considerations that are quite different from traditional ones

- Applications are easily and highly parallelizable
- All applications can be hosted on the same architecture
- The applications are at incredible scale (global)
- The applications are purpose built
- Storage and computation are advantageously collocated
- Small amounts of data loss are excusable

At present a "brick" architecture is considered the most advantageous for these data centers. The concept of a brick ranges from a 1-U server with several disk drives in it to a large shipping container filled with CPUs, storage and switching, but tends toward the former.

The brick architecture typically replicates data at least twice throughout the network. When anything in a brick fails beyond its internal redundancy limits (if any), the entire brick is targeted for replacement, and any data and computation formerly hosted on it are moved to a new brick.

It is difficult to endorse the resulting designs from a green storage standpoint, in part because the companies building them have declined to disclose the details of their architecture on account of the incredible competitive pressures in today's internet marketplace. Several things seem clear, however. First, the use of storage in these data centers is probably not well optimized. It's possible that some of the capacity optimization technologies mentioned in the first part of this chapter are used, but unless they are provided by the host OS (Linux or Windows), they must be home-rolled. The heavy reliance

on replication to allow easy and worry-free replacement of bricks seems inevitably to lead to greater raw storage capacity use than in traditional data centers.

Second, there are far more CPUs in such an architecture than in a traditional centralized storage architecture. However, if the designers are able to balance CPU density against the combined storage and computational requirements of the applications, this may actually

become a strong point of the architecture, as there are undeniable benefits to having storage and computation as close to each other physically as possible.

Third, it is difficult to see how disaster recovery (DR) is done with such a broadly distributed data pool, short of replicating entire data centers, which would only decrease storage efficiency. This guide is unable to offer much guidance in this area, as so little practice has been published and reviewed in industry fora. Before going with a monolithic architecture, designers should ask their IT architects some questions.

- How will backup be done?
- Are there any retention requirements? If so, how are these handled?
- How is remote replication handled? Does it allow for follow-the-sun computation in addition to DR?

One thing that can be predicted with some confidence regarding monolithic data centers, however, is change. Change in computational units, change in storage units, and change in application mixes and optimization strategies. For this reason, best practice for data center designers from an operational perspective seems to be to modularize power and network distribution networks as much as possible. In any architecture that leverages commodity hardware, total reconfigurations are likely on 3- to 5-year boundaries. Large power buses with moveable and interchangeable power drops, and networks of large upgradeable blade-style director switches with easily relocatable terminal switches give maximum flexibility with respect to these future reconfigurations, thereby mitigating future fabric and electrical change costs.

Cloud data centers

Data centers intended to host cloud infrastructures are in some ways similar to monolithic data centers. However, there are some very important differences from the monolithic model that the designer must be aware of. These include a central VM repository and infrastructure, the backup and recovery strategies for same, and the storage infrastructure.

Clouds can be built with built-in DAS in the bricks, exactly as in monolithic data centers. However, this may limit the storage available for each node or force reconfiguration of some compute nodes into pure storage nodes--usually not an optimal use of the node's resources. More flexibility may be obtained by the use of heavily virtualized storage delivered over the network to pure compute nodes having no storage at all except perhaps a cache. These nodes boot over the network from a storage array as well. This arrangement allows the compute farm to be pure computation--possibly using blade servers for compactness and serviceability. Using blade servers allows a more economic use of higher-

bandwidth fabric such as 10G Ethernet, as the fabric does not need to be run to every node. The intermediate switching infrastructure must then be 10G, at a corresponding increase in cost, but this is offset somewhat by the built-in switching in the blade enclosures.

Storage in such a facility must be capable of delivering virtualized storage containers pre-configured with OS images and applications suitable to a given task, often many hundreds in a short time, as a given large application is loaded and prepared for a run. An ability to clone many virtual machine images from a single golden image is a differentiator. Each storage array must be able to handle the bandwidth requirements from the entire collection of blade servers that it is backing--an array for every couple hundred blade servers may be necessary depending on the storage bandwidth requirements of the applications. Secure multi-tenancy is another requirement, to assure end users that their data cannot be accessed or manipulated by third parties.

As in all data center models that feature standalone storage, storage that is amenable to hot aisle/cold aisle racking is a must.

Engineering data centers

Many engineering data centers have had grid-based processing farms for many years, performing work such as compiles of large applications, VLSI routing of chip designs, large-scale pharmaceutical and biochemical computations, structural analysis, wind tunnel simulations and other industrial tasks. These grids closely resemble compute clouds at a high level. The main difference between grids and clouds in practice seems to be that grids have been largely purpose-built, and function largely without benefit from much virtualization, whereas clouds, thanks to heavy use of virtualization, can be built in a way that makes them a general-purpose resource.

Data center designers must ascertain whether the engineering team is planning a transition from grid to cloud approaches in its new engineering data center. The increased virtualization requirements in a cloud approach drive numerous switching and fabric infrastructure choices, as detailed in the previous section.

Engineering data centers are also typically required to store a significant amount of unstructured data, in the form of product design and documentation discussions, documents, presentations and so on. Sharepoint, Exchange and other collaboration tools may also be hosted separately from corporate IT for various reasons. Designers must ascertain how much unstructured data storage will be required for a new data center; at this point (2010) cloud

computing resources are an inefficient method of doing this. Best practice for this storage resembles that of a traditional data center for this reason. Storage vendors should be queried for products that meet the needs of both cloud and file storage, when possible, with an eye toward minimizing the number of dedicated storage personnel needed for the new data center.

The above is complicated by "follow the sun" computing practices that have emerged in many leading edge engineering companies in the past decade. These designs usually incorporate at least three globally distributed data centers that are kept synchronized with each other, at least for core data. By

using equipment that is close to the engineering team working in daylight hours, WAN latencies are minimized and computational efficiency increased. The storage systems in these centers must be able to perform data synchronization efficiently and reliably to avoid data write conflicts and slowdowns. WAN optimization products are a must-have in such an architecture unless the storage incorporates WAN optimization natively in its synchronization protocol.

SMB data centers

Smaller businesses generally have reduced needs and reduced requirements, in part driven by their smaller financial profiles. The data center for a small business may consist of a small room or even a closet with several servers in it. Many of the facilities-side approaches to a green data center are probably unavailable to the designer or consultant because of the small scale.

On the storage side, best practices include small highly integrated storage arrays that can be easily expanded when the company experiences growth. Very easy management or inexpensive outside maintenance contracts are a must—this often implies a Windows-based or very well designed web-based management interface on the storage. Backup is best done to a trusted cloud backup provider or an internal disk-based backup with dedup capability that is mirrored to another site; while tape is optimal, very few SMBs have the self-discipline to securely and effectively manage a tape backup rotation including offsite storage. Better small storage arrays are also able to offer PIT snapshots at frequent intervals, giving the business a much better RPO and RTO profile. They also generally offer at least some of the capacity optimization technologies mentioned earlier in this chapter.

Care should be taken to avoid seemingly easy green shortcuts with hidden pitfalls. One example is direct use of outside air for cooling—easy to accomplish when a data closet is on an outside wall. The pitfall here is that humidity levels above 80% can cause condensation on disk

SNIA

drive electronics leading to early and non-warranteed failure of the drives [Seagate], especially in smoggy areas.

Dehumidification is normally taken care of automatically by air-conditioning equipment, and is therefore easy to overlook as a requirement. A consultation with an experienced HVAC contractor should be undertaken to determine whether economical use of outdoor air can be made.

Operational power monitoring and management

There are at this point at least as many methods of monitoring operational power as there are vendors of equipment. We suggest several here that are susceptible of some level of standardization.

SNMP

SNMP reporting is almost universally available -- even small home JBOD boxes support it. However, there are no standard MIBs at the time of this writing. This means writers of monitoring software must detect the make, model and version of a piece of equipment and then use a vendor-supplied MIB for querying its management information. This isn't generally as hard as it sounds, and SNMP remains the undisputed king of status reporting for networked objects.

SNMP supports a rudimentary form of active management via the SET operation, but earlier versions of the protocol had such weak security that the protocol became known as being useful only for read-only management. Additionally, the lack of a transaction mechanism made complex management operations awkward or impractical. So while version 3 of SNMP has largely addressed the security concerns seen with earlier versions of the protocol, it has not seen wide acceptance. To date, active management via SNMP is not available on most devices.

In spite of SNMP's ubiquity, SNMP reporting of power usage is in its infancy. Component manufacturers must include the capability of reporting up to the enclosing storage system, which must then aggregate and report the results via SNMP. Widespread adoption of this technique for reporting power is probably several years away (say 2015), and no industry roadmap for it exists.

PDU-level monitoring

It is normally straightforward to configure storage equipment with rack-level PDUs that support power monitoring. Vendors should be queried for their support of this practice, as

proper vendor support will include the PDUs in the cabinet at manufacturing time. The alternative is retrofitting new PDUs after purchase of the equipment.

The reasoning behind rack-level monitoring is that it introduces reasonably fine-grained power reporting without significant fuss or expense, and it does so now. Several vendors currently offer solutions based on rack-level PDUs, and their number is growing.

CIM

SNIA's SMI-S specification, version 1.5, contains an Operational Power profile. It specifies a means of collecting statistical data from a ComputerSystem. This statistical data can include--depending on the manufacturer's implementation and the elements of the device--the power used by a system's fans, disk drives, disk shelves, HBAs, power supplies and the whole system itself. The CIM classes involved are included in the DMTF CIM Schema v2.28 and later. Quantities are reported in milliwatts.

To date, there has been little implementation of this profile, and manufacturers of management consoles have not joined in efforts to promote it. Widespread adoption is hard to predict, and in any case will probably be several years away from the date of this writing (2011). No industry roadmap exists.

Operational Temperature monitoring

Temperature monitoring is at roughly the same level as power monitoring in terms of availability and robustness. Almost all storage system components have temperature sensors built in so that the devices may shut down to prevent media failure. These readings are typically not to the precision that

modern data center designers would like, and are also not usually reported out to the enclosing system.

So work needs to be done before fine-grained temperature monitoring is available via either SNMP or CIM. Timeframes will most likely roughly parallel those for operational power monitoring.
Power monitoring PDUs, discussed in the previous section, generally offer inputs for several temperature sensors. The availability of this needs to be taken into consideration when selecting a rack-level PDU and vendor. It is the recommendation of this guide that three to five sensors be used in storage equipment racks, with the majority of them placed in back (on the hot side) of the equipment.

Conclusion and best practices

Storage is a complex and demanding piece of the data center infrastructure. Except in purpose-built monolithic data centers with clear and compelling reasons for a mixed architecture (i.e. bricks with embedded storage), best practice involves consolidating storage as much as possible. This allows a maximum use of green technologies, administrative effort and centralized backup and archive resources. Furthermore, planning for increasing levels of virtualization is likely to pay dividends; this means emphasizing flexibility especially in the fabric and switching layout, and investing in storage systems that can be easily expanded in capacity and I/O capabilities, and that support storage virtualization easily and well.

The heavy hitters

Some green storage technologies are particularly rewarding. Parity RAID, thin provisioning, and tiering or caching with SATA drives can easily double a storage system's capacity power efficiency. In primary storage, delta snapshots and deduplication also help. In online backup scenarios, streaming deduplication is king, but the other mentioned technologies also will play important roles if deployed in the backup system.

Designers and architects should press potential vendors to demonstrate their level of commitment to storage optimization of all types.

Other reading

SNIA green tutorials may be found at

http://www.snia.org/education/tutorials/2010/spring#green
http://www.snia.org/education/tutorials/2009/spring#green

A SNIA Green Storage Initiative (GSI) whitepaper on green storage is at

http://www.snia.org/forums/green/knowledge/GSI_Best_Practices_V1.0_FINAL.pdf

Acknowledgements

Author — Alan G. Yoder, Ph.D. (NetApp)
Reviewers at SNIA — Mike Dutch (EMC), Larry Freeman (NetApp), Wayne Adams (EMC)

Annex A. Question list for prospective vendors

The following questions are representative of those that should be addressed by storage vendors during any major storage purchase. Their relative importance depends on the installation and the intended use of the storage.

SNIA

Reliability

1. Is no single point of failure (no-SPOF) a requirement?
2. How many simultaneous failures in a RAID group must be tolerated?
3. DR - is a remote live copy of the data required to mitigate site-wide disasters?

Availability

4. How many "nines"? Five nines (99.999% uptime) means an average total unplanned downtime per system of five minutes per year. Four nines allows an average downtime of 50 minutes total per year. Three nines allows an average total downtime of 8 hours and 20 minutes. A "nines" specification is an SLO, not a hard guarantee.
5. RTO - recovery time objective - how long of an outage at any one time is permissible? An RTO is an SLO. [This question is related to question 3].
6. RPO - recovery point objective - how much data that was generated but not yet written can be lost in the event of an outage? This is usually specified in minutes or number of transactions. This too is an SLO.

Serviceability

7. Is NDU (non-disruptive upgrade) required? For some, or all FRUs?
8. What level of service must be maintained during a service event?

Connectivity

9. What kinds of systems will be connecting to the storage? What protocols do they use? (CIFS, NFS, CKD, FC, SAS, SCSI, iSCSI, FCoE). Does the vendor fully support all the protocols in the requirements mix?

Suitability to purpose

10. What applications will have their data hosted on the storage? Is the storage supported by the application vendors?
11. Does the vendor have a loan program enabling you to test the product on site in your environment? No matter how much data, architectural details and how many references are presented, the final test of any system is how well it performs for you and whether it solves the business problem that you are attempting to address.

Retention and compliance

12. Are there retention requirements (from corporate Legal dept.)?
13. If so, can storage support them directly? There are security advantages to having the protocol enforcement point be at the storage itself.
14. Can storage be made secure against "back-door" or "out-of-band" access to supposedly locked-down data?

SNIA

15. Can storage be "scrubbed" of data in the event that it is accidentally written to a non-secure container? (Another often-used term is "data shredding").
16. Does storage support auditing of all attempts to change or delete locked-down data?
17. Can audit data be kept on a remote machine in a different security domain, to protect against malicious administrative access to it?
18. Is there a records management system? What are the storage requirements for it?
19. What is the timeframe in which legal discovery requests must be satisfied?
    a. Are there third-party solutions that can satisfy the requirements?

Backup and Archive

20. How often must backups - to recover from accidental deletion or data corruption - be made?
21. How often must backups - to recover from equipment damage or failure - be made?
22. How often must backup copies be archived to an offsite location?
23. How quickly must backup data be made online and accessible when requested? (can range from seconds to days)
24. How quickly must archive data be made online and accessible when requested? (can range from seconds to days)
25. How many backups are anticipated between each archival operation?
26. How many backups must be kept locally?

Support

27. Is automated "phone-home" support available? This can allow vendors to perform remote diagnostics and suggest repairs before problems become critical.
28. Are the offered support packages adequate? Options can range from full on-site maintenance to a 48-hour response time.
29. Is the pricing for support spelled out for the suggested lifetime of the equipment? (Usually 5 years).

ROI and storage efficiency

30. Do ROI calculations provide adequate information?
    a. Are $/IOPs over the life of the equipment calculated? If so, attempt to normalize the number using the expected I/O load of the equipment, say 50% of max.
    b. Is an average ratio of useful capacity to raw capacity presented? If so, how is it calculated? Do the calculations inspire confidence?
    c. Are $/TB of capacity over the life of the equipment calculated? If so, attempt to normalize it using the ratio in (b), above.
31. Do ROI calculations include support, installation and architecture fees?

Upgrade path

32. What is the required procedure to upgrade from your present system to the proposed new system?
33. Does the vendor offer a smooth upgrade path from the proposed new system to the next level that would be required if your data needs grow significantly?

Power and temperature monitoring

34. Are options for power and temperature monitoring available pre-configured on the equipment?
35. Do the available options work with any monitoring packages you may have installed already?
36. Does the company publish a SNIA Emerald™ Data Sheet?

Company and product viability

37. How long has the vendor been in business?
38. Is the vendor a "going concern"? Obvious acquisition targets and financially weak vendors raise concerns that they may not be able to support a product for its intended lifetime.
39. Is the product well known? Marginal and experimental products similarly raise concerns about viable lifetime.

.

SNIA