HAL
archives-ouvertes.fr

Energy Efficient Storage Technologies for Data Centers

Alan G. Yoder

▶ **To cite this version:**

Alan G. Yoder. Energy Efficient Storage Technologies for Data Centers. John Carter and Karthick Rajamani. WEED 2010 - Workshop on Energy-Efficient Design, Jun 2010, Saint Malo, France. 2010. <inria-00492884>

HAL Id: inria-00492884
https://hal.inria.fr/inria-00492884

Submitted on 17 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Energy Efficient Storage Technologies for Data Centers

Alan G. Yoder, Ph.D.
NetApp, Inc.
Storage Networking Industry Association (SNIA)

Abstract

An impressive amount of work has been done to date on improving the electrical efficiency of various data center components. The data storage industry has begun to see the fruits of this effort, with increased power supply and fan efficiencies. However, storage presents other significant opportunities for energy conservation, through various types of capacity optimization, that are not captured in electrical efficiency discussions. As data storage uses on the order of 25% of IT power in an average data center, these other opportunities bear examination. This article presents a survey of emerging storage technologies which positively impact energy usage and presents current thinking in the storage industry regarding their relative effectiveness. It also attempts to set a baseline for configurations against which improvements in capacity and energy use can be made.

Introduction

The idea of optimizing storage is as old as data processing. Compression is one of the oldest and most venerable, as well as one of the best understood computational disciplines. Newer techniques such as data deduplication and advanced RAID, however, have only recently become well enough understood to admit of industry standard definitions [1]. Many of these technologies offer far greater gains in efficiency than can be gotten via electrical optimizations (e.g. power supply efficiency, component idle modes).

However, the gains offered by various technologies are confusing. Some, such as thin provisioning, are much better than it might seem at first blush. Others are much worse. Disk spindown, for example, has seen much interest, but few products have successfully shipped, and existing research leaves open the question of what to do about background data housekeeping tasks.

Data protection requirements add further complexity to the modern data center storage picture. Many academic lab environments are content with relatively simple backup schemes. But advanced data centers often have recovery point objectives (RPOs) of zero, meaning that no transaction loss is acceptable. And their recovery time objectives (RTOs), the time allowed for recovery from a fault, may be minutes or even seconds. Even at less stringent sites, planned outages longer than a standard network delay window are often unacceptable.

In addition, recent studies of power consumption in storage systems have outlined the complexity of behavior in the field, including the fact that maximum performance and maximum power may not match [19] and that interactions between storage and servers must often be considered together [20].

These issues have caused a disconnect between industrial and academic conceptualizations of exactly what "storage" means.

Section 1 outlines a set of "standard" requirements for data center storage, with the reasoning behind them. Section 2 presents the storage optimizing technologies found noteworthy by the Storage Networking Industry Association (SNIA). Section 3 discusses the efficiency gains achievable by these technologies, as measured against the base configurations outlined in Section 1. Section 4 concludes.

It should be noted that this paper deals almost exclusively with technologies and configurations suitable for primary storage. Space restrictions prevent a full discussion—which could easily equal this one in size—of secondary storage.

Section 1. Base storage configurations in industrial data centers

In the home and small or home office business sectors, almost all the data from almost all disk crashes can be recovered by a disk forensics expert for a fee of about $1000. This makes expensive RAID arrays and backup systems seem unnecessary. Indeed, a system like Apple's Time Machine[1] with a decent power conditioner and UPS, and an offsite copy of the Time Machine disk

[1] All trademarks mentioned herein are the property of their respective owners.

to guard against site wide catastrophe, can seem like all the data protection an average small office needs.

But this level of protection is inadequate for systems that must support 24/7 operation and thousands to hundreds of thousands of users. For Data Center class storage, the acronym RAS (*reliability, availability* and *serviceability*) is used to describe the requirements. In plain English, we can characterize these three qualities as follows:

- Reliability means that the system rarely if ever goes down or loses data.
- Availability means that when it does go down, it doesn't stay down very long. A system that meets the "five nines" standard has an average downtime of less than 5.26 minutes per year.
- Serviceability means the system can be serviced—replacing broken parts, upgrading firmware—without impacting availability.

Serviceability and availability requirements combine to mean that data center class equipment must have *no single point of failure* (no SPOF). The power supplies in such a piece of equipment are hot swappable and redundant, and therefore don't run at more than 50% capacity under normal circumstances. This means efficiency work should focus on 10-50% utilization scenarios. Variable speed fans are also critical to good efficiencies, as they too are redundant.

The reliability and availability requirements lead to a downstream requirement for redundancy in the storage media. This has led to the use of RAID 1 (mirrored disks) in data centers as a policy mandate. Not surprisingly, data center class disk and system vendors have historically not struggled much against this seeming waste of rotating media.

The availability requirement also means that some means of restoring quickly to a good copy of a dataset must be available. This is usually done through the use of point-in-time (PIT) copies, termed *snapshots* or *clones*. These are often augmented by write journals—also stored on redundant media—that can be replayed against the PIT copy to restore the dataset to its exact state at the time of the failure.

Finally, a significant percentage of data center class gear is used for *structured data*, which is, at a first order, data generated and maintained by databases. These databases frequently are configured to use raw storage. *Storage area networks* (SANs) have provided data center operators with configurations that can be run and backed up in a more centralized and efficient manner than direct-attached disks. However, given that the SAN-attached array must emulate direct-attached disk, this class of equipment has traditionally required the use of high-performance fibre-channel (FC) or serial-attached SCSI (SAS) disk drives. Comparison of costs between these drives and the SATA drives available at Fry's Electronics neglects key differences and technology design choices [17], [18].

The point of this exposition is that when comparing configurations in a data center context, the RAS requirements must be used to establish a baseline configuration against which improvements can be evaluated. The SNIA baseline configuration can be summarized as "no SPOF, FC drives, RAID 1".

Section 2. Storage optimizing technologies

While "moving the needle" (i.e. using less power to accomplish tasks) is important, there are other ways of saving not only energy, but equipment assets as well. If an array can be made to store twice as much data, the hardware savings are obvious. For this reason, updating arrays to new drive technologies every several years remains a tried and true way of saving on both equipment and power (the gains gotten from drive generations are so obvious and longstanding that they don't appear on the following list).

SNIA's analysis indicates that there are five major capacity optimizing software technologies

1. Delta snapshots
2. Thin provisioning
3. Advanced RAID
4. Data deduplication
5. Compression

and a capacity optimizing hardware technology

6. Slow high-capacity drives + flash

that are driving significant storage efficiencies. We will treat each of these, as well as the "move-the-needle" technologies

7. Disk spindown
8. Power supply and fan efficiency, and
9. SSDs

We'll also mention

10. Facilities optimization

First, however, a classification scheme[2] will help to organize the various technologies according to effect. There are four basic strategies for reducing power consumption in data center storage equipment

1. Make the equipment more power-efficient
2. Use less redundancy
3. Commit less space
4. Squeeze more data into available space

Section 2a. Making equipment more power efficient

This class of technology includes facilities optimization, power supply and fan efficiency, disk spindown and high-capacity disks.

Facilities optimization

The PUE metric [5] developed by The Green Grid [6] nicely demonstrates the gains possible when optimizing facilities. Briefly, the PUE is the ratio between the amount of power entering a facility and the amount of power reaching the IT equipment. Traditional data centers often had PUEs of 2.5 and higher, meaning that 40% or less of the power actually reached the equipment it was intended to drive. Modern innovations include flywheel UPSs, air economizers, variable speed fans, hot aisle/cold aisle technologies, and improved monitoring and feedback control systems. Collectively, these have allowed the construction of facilities with PUEs of 1.25 on a routine basis, with better than that achievable. This is a doubling of overall power efficiency, more than can be gotten from any IT-side adjustment.

Power supply and fan efficiency

Traditional power supplies have had efficiencies in the 60-70% range. Efforts such as the 80 Plus Program [7] and the US EPA's ENERGY STAR program [4] have resulted in efficiency targets in the 80-95% range, depending on load.

Variable speed fans would seem to actually offer greater gains than this, because of the near-quadratic reduction in power required as they spin slower. Unfortunately, although the ENERGY STAR program for home PCs and small servers requires use of these, data on the amount of power that has been actually saved is scarce. More quantified study in this area needs to be done before even educated guesses are appropriate.

[2] The word "ontology" is eschewed by this author.

Disk spindown

Spinning down unused disk drives has received a lot of attention, relative to other technologies. However, all analyses in the literature that the author has seen, such as PARAID [12] and SRCMap [13], make simplifying assumptions that render the research unusable for primary storage. First, what to do about background housekeeping activities such as parity and RAID scrubbing, absolutely essential in an enterprise-class array, is left as an open question. Second, accesses to spun down drives can cause latencies of 10 to 15 seconds. This relegates these systems to secondary storage (the SNIA defines "online", i.e. primary storage as having a max time to first data of 80ms).

One possible use for spindown in primary storage systems is for hot spares. A data center class array may have 1-2% or even more of its disks in hot spare mode, so that amount of the disk (and possibly the disk drawer fan) power can be reclaimed by this method.

SSDs

Solid state disks (SSDs) appear to have nearly ideal energy characteristics: energy used is proportional to IOPs, and data at rest is nominally free. This is a huge improvement over rotating media, which uses about 85% as much power at idle as it does when active [3].

But SSDs are not presently a viable drop-in replacement for traditional storage media, as the price points are simply not there. At present, they are useful for caching and storage tiering.

High capacity drives with flash

In modern data centers, storage systems are normally separated into two classes—high performance (and expensive) fibre-channel (FC) or serial-attached-scsi (SAS) systems, and high-capacity serial ATA (SATA) systems. SATA drives offer about 4x the capacity per spindle that FC drives do, spin at lower speeds (5400 or 7200 vs. 15000 RPM) and cost less. Based on the spec sheets, current technology from Seagate [15] has statistics as follows

	Barracuda XT SATA	**Cheetah 15K.7 FC**
Raw capacity	2 TB	600 GB
Idle power	6.39W	11.61W
Response time (avg)	4.16 ms	2.0 ms
Sustained data rate	138 Mb/s	122 MB/s +

This yields a power difference per unit data of about 6.05. In other words, it takes one sixth as much power to store data at idle on SATA as it does on FC. This is a significant savings in the power required for data at rest. But it has traditionally come at a significant cost—a reduction in response time and bandwidth, as shown in the table. This has historically made arrays using SATA drives useful mainly for secondary storage and specialized applications such as video capture.

However, recent innovations in flash and solid state storage have enabled the construction of disk arrays consisting mainly of SATA drives, with large tertiary caches of flash or SSDs, and even quaternary caches of FC or SAS drives. These hybrid systems approach or even in some circumstances exceed the performance of FC and SAS arrays, at price points below those of FC or SAS storage. Given the present cost differential between traditional drives and SSDs, which shows few signs of narrowing significantly in the near future, it is possible that this class of system will become predominant until the value proposition of pure SSD storage becomes more compelling.

The topic of optimal sizing of the various elements in a DRAM/NVRAM/flash/SSD/rotating media storage hierarchy is of great present interest to the industry. The SNIA's Solid State Storage Initiative (SSSI) is attempting to address the space, and has projects underway in several areas. [23]

Section 2b. Use of less redundancy

As explained in the first section, redundancy is a requirement for data center class data protection. This is reflected even in the Uptime Institute's data center tier framework; tier 3 data centers have redundant UPSs, while tier 4 data centers have redundant power coming from entirely separate power grids and suppliers [8].

RAID

The data protection requirements in a data center result in a requirement to survive the loss of any one storage media device in a RAID array[3] without loss of data or access to data (a small performance degradation during "RAID rebuild" is considered acceptable). RAID 1—mirrored storage—achieves this admirably. It can survive the loss of many drives in an array, in fact, so long as no two of them belong to the same mirrored pair.

RAID 5 (and the less common 4 and 3) also offer protection from a single disk failure per RAID group. However, the RAID groups are larger (typically 5 to 8 for RAID 5 vs. 2 for RAID 1),[4] and reconstruction of a RAID group may affect performance more than in a mirrored array. These RAID group formats cannot tolerate as many failures overall as RAID 1, because of the larger RAID group sizes.

Nonetheless, there are many scenarios in which data center operators decide that RAID 5 is worth the additional risk over RAID 1 to save on the number of drives that must be bought. Reductions of 37% or more are easily possible (5 disks in a RAID 5 group vs. 8 RAID 1 disks), with a corresponding reduction in the required fan and cooling power needed.

RAID 6, when acceptable performance is offered, is another alternative. Due to the mathematics, larger RAID groups are used, and the economies work out roughly the same as for RAID 5. However, availability is better than for RAID 5 or even RAID 1. First, RAID 6 protects against the chance—becoming more significant with every disk drive generation—that a second drive will fail during a RAID reconstruct of today's large SATA drives. Second, in a group of 16 drives, any two can fail under RAID 6, while under RAID 1 7%, i.e. 1/(N - 1) of the second failures would take out a mirrored pair, bringing down the array.

Delta Snapshots

Standard snapshots are point-in-time (PIT) copies. High-end storage arrays offer the ability to restore access quickly after a fault by remounting a broken volume on a recent snapshot, and possibly replaying a log.

Delta snapshots work on a principle very similar to the *vfork* method used in Unix process management: pages (blocks) that are shared between the live system and the snapshot are not copied; blocks and the inodes or other metadata pointing to them are only copied when new data is written.

Again, only anecdotal evidence is available, but for a snapshot kept one day, it appears to be quite rare that more than 10% of the data in it has been

[3] Space does not allow a discussion of RAID concepts and types. Please see any of numerous sources on the Web, especially [1] and [9].

[4] See manufacturer's data sheets and best practice guides for recommended RAID group sizes.

written to. A space savings estimate of at least 90% is therefore appropriate for a delta snapshot.

Section 2c. Committing less space

Storage containers are often over allocated to protect against future unplanned outages. In addition, much storage is wasted by never being allocated in the first place—in a sense the storage in the array itself has been over allocated.

While there are no conclusive research results that we know of, anecdotal and widely accepted evidence gives a figure of roughly 30% for storage utilization across the industry [10]. That is 30% of the *usable capacity* of that storage. Working backwards, storing 30 TB of data requires 100TB of formatted storage, which under RAID 1 would require 200TB of raw storage plus up to 10TB for the system. That means users of traditional storage may be paying for about 7 times as much capacity as they are using.

Thin provisioning

The technique of thin provisioning avoids much of this cost. The system allocates storage on demand, and the storage admin tracks the total amount of storage used and adds more as it is required. As with quotas on filesystems, the total amount of quota or allocation given out may greatly exceed the amount of real storage backing it.

Estimates of utilization when thin provisioning is in place tend to run to 80%. The SNIA has begun using a conservative figure of 100% as the amount of gain to be expected in an average scenario (going from 35% to 70% utilization) [11].

Section 2d. Squeezing more data into available space

Finally, the data itself often offers opportunities for further reduction in space requirements. Compression ratios of 2:1 are commonly accepted in tape backup, but tape is a streaming format well suited to compression. Compressing to block-oriented storage is more difficult and less efficient on account of quantization roundoff, but is now being shipped in the marketplace.

Data deduplication is in some ways more suited to block storage. The greatest gains are seen in backup scenarios, where similar datasets are repeatedly written to online or near-online media. But as these scenarios compete with tape, which uses no energy on the shelf, they are generally left out of "green" discussions.

However, in-place dedup of live data is also an increasingly available alternative. Savings estimates vary considerably depending on data set and on whether any financial guarantees are attached to them. A recent vendor announcement gives a figure of 27% for an exabyte of field data. [22]

Section 3. Savings

The following table summarizes the *approximate* savings available via the technologies discussed herein, as determined by the Green Storage Technical Working Group of the SNIA. The numbers given are therefore industry consensus numbers. They are relative to the component, so facilities optimization applies to the whole facility, for example, while power supply savings apply only to that percent of energy used by the power supplies and the things they power in the facility.

Technology	Savings
Facilities optimization	50%
Power supply improvements	20%
Variable speed fans	unknown
Large capacity drives	80%
Advanced RAID	40%
Delta snapshots	90%+
Thin provisioning	50%
In-place data deduplication	27%
Compression	20% +

There is no good way to tell how these savings will combine when multiple technologies are implemented at once. A marketplace-based metric, however, suggests that 50% is a conservative figure for the software technologies alone. Multiple vendors are now guaranteeing that their systems will use 50% less storage than a competing system without these technologies.

While rigorous analyses and data comparisons are desirable, these have proven problematic to obtain. First, data sets vary; there is no known predictor for the energy and space consumption characteristics of a specific application on an arbitrary system. Second, large storage arrays are expensive and hard to procure. Third, storage vendors are notoriously shy about publication of performance statistics; non-disclosure agreements commonly prohibit equipment users from publishing the results of their internal research. For these and other reasons, the SNIA has so far used a consensus-based approach to obtain numbers when publishable empirical studies have been unavailable.

Section 4. Conclusion

Before concluding, a caveat. This article has not much discussed the performance implications of various technology choices, except peripherally. In general, thinking in terms of energy efficiency for *data in flight* is in its infancy. It is too early to discuss such efforts definitively. Even the SSD space is moving so rapidly that rational prediction is difficult.

As a result, the various technologies discussed herein mostly pertain to the *data at rest* axis of the storage performance space. In other words, we have focused on quantities with units like energy$$/GB, where data in flight metrics would use units like energy$$/IOP [21].

In reality, energy use is just one vector in a many-dimensional storage procurement decision space. The capacity optimizing technologies treated herein, however—delta snapshots, thin provisioning, advanced RAID, deduplication, compression, and the use of slower high-capacity drives with large caches—do not only decrease energy cost. They do so by allowing the purchase and powering of less storage media than would have otherwise been needed. This means that capital expenditure is favorably impacted along with ongoing operational expenditure. As a result, these technologies are powerful forces in the move toward green data centers.

Systems shipping in the near future will have most or all of these technologies in place; future studies of data center class storage should evaluate their work and proposals in the overall context of systems loaded with a full suite of RAS and capacity optimization technologies.

Acknowledgements

Erik Reidel and the workshop reviewers made valuable comments and suggestions. Thanks go also to many collaborators at SNIA, especially Jody Glider and Patrick Stanko.

References

[1] *The SNIA Dictionary*, a lexicon of terms related to the storage industry. www.snia.org/dictionary.
[2] The Storage Networking Industry Association. A 501 (c) 6 trade association of storage industry vendors and concerned individuals. www.snia.org.
[3] The 85% figure was arrived at after testing over a dozen commercial systems at the SNIA Technology Center in Colorado Springs. Due to NDAs, detailed test results are not available except to members of the SNIA
[4] *ENERGY STAR Data Center Storage Specification*. Under development. See www.energystar.gov/index.cfm?c=new_specs.enterprise_storage
[5] PUE is discussed in the Metrics and Measurements whitepaper available at www.thegreengrid.org/en/Global/Content/white-papers/The-Green-Grid-Data-Center-Power-Efficiency-Metrics-PUE-and-DCiE
[6] The Green Grid, an industry consortium. www.thegreengrid.org/.
[7] The 80 Plus program. www.80plus.org/.
[8] TierStandard.pdf. uptimeinstitute.org/content/view/302/281/
[9] RAID is a complex topic. See en.wikipedia.org/wiki/RAID as an introduction.
[10] John Tyrrell (storage architect and veteran of over 600 data center visits), private conversation.
[11] SNIA Tutorials, located at www.snia.org/education/tutorials/2009/fall#green
[12] Weddle et al. *PARAID: A Gear-Shifting Power-Aware RAID.* Proceedings of FAST 2005.
[13] Verma et al. SRCMap: Energy Proportional Storage using Dynamic Consolidation. FAST 2010.
[14] See googleenterprise.blogspot.com/2009/10/q309-spam-virus-trends-from-postini.html
[15] www.seagate.com/www/en-us/products/desktops/barracuda_xt/#tTabContentSpecifications and www.h-online.com/priceinsight/a396893.html
[16] Reinsel, IDC #212714, "The Real Costs to Power and Cool All the World's External Storage", June 2008
[17] Anderson, Dykes, Riedel, "More Than an Interface – SCSI vs. ATA" FAST conference, March 2003
[18] Whittington "Desktop, Nearline & Enterprise HDDs What's the difference?" www.snia.org/education/tutorials/2008/spring/
[19] Allalouf , Arbitman, Factor, Kat, Meth, Naor "Storage Modeling for Power Estimation" IBM Haifa Research Labs; SYSTOR 2009; May 2009
[20] Lange "The Next Frontier for Power/Performance Benchmarking: Energy Efficiency of Storage Subsystems" SPEC Benchmark Workshop, January 2009
[21] SPC Benchmark 1/EnergyTM (SPC-1/ETM), Overview Presentation, October 2009, www.storageperformance.org/press
[22] See www.storagenewsletter.com/news/business/one-exabyte-de-duped-with-netapp
[23] See the several recent tutorial-level presentaions at www.snia.org/education/tutorials/2010/spring#solid