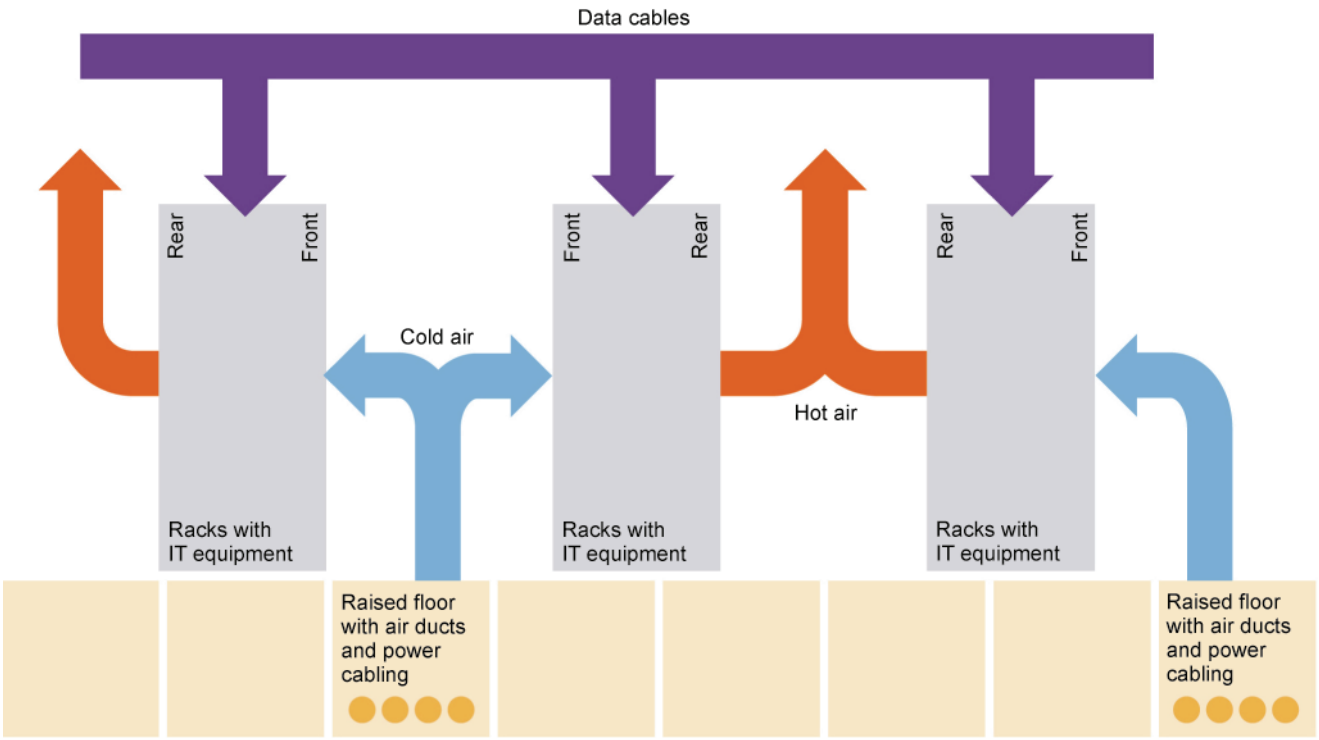
Green Sustainable Data Centres

IT Equipment



This course is produced under the authority of e-Infranet: http://e-infranet.eu/

Course team
prof. dr. Colin Pattinson, Leeds Beckett University (United Kingdom),
course chairman and author of Chapter 1 and 7
prof. dr. Ilmars Slaidins, Riga Technical University (Latvia*),*
assessment material development: Study Guide
dr. Anda Counotte, Open Universiteit (The Netherlands),
distance learning material development, editor- in-chief
dr. Paulo Carreira, IST, Universidade de Lisboa(Portugal),
author of Chapter 8
Damian Dalton, MSc, University College Dublin (Ireland),
author of Chapter 5 and 6
Johan De Gelas, MSc, University College of West Flanders (Belgium),
author of Chapter 3 and 4
dr. César Gómez-Martin, CénitS - Supercomputing Center and
University of Extremadura (Spain),
author of Checklist Data Centre Audit
Joona Tolonen, MSc, Kajaani University of Applied Sciences (Finland),
author of Chapter 2

Program direction
prof. dr. Colin Pattinson, Leeds Beckett University (United Kingdom),
prof. dr. Ilmars Slaidins, Riga Technical University (Latvia)
dr. Anda Counotte, Open Universiteit (The Netherlands)

Hosting and Lay-out
http://portal.ou.nl/web/green-sustainable-data-centres
Arnold van der Leer, MSc
Maria Wienbröker-Kampermann
Open Universiteit in the Netherlands

First edition 2014

IT Equipment

Chapter 3

IT Equipment

Johan De Gelas
University College of West Flanders

INTRODUCTION

In Chapter 2 we have discussed the configuration of the data centre and how the energy efficiency of the auxiliaries like cooling can be improved. In Chapter 3 and 4 we will explain why IT equipment offers the biggest potential for energy savings. In Chapter 3 we will discuss which IT equipment is used in a data centre and how you discover the best strategy to optimize the performance per watt. In Chapter 4 we discuss the techniques for energy reduction by adjusting the operating system and by consolidation and virtualization of servers.

LEARNING OBJECTIVES
After you studied this chapter we expect that you are able to
- describe the main building blocks of IT equipment inside the rack
- measure power consumption of the IT equipment inside the rack, or make extrapolations based on industry standard benchmarks.

Study hints
The purpose of this chapter is to give an overview of iT equipment: servers, networking and storage equipment.
The workload is 6 hours.

CORE OF STUDY

1 IT Equipment: the Source of Energy Consumption

Performance

The key process in a data centre is computing. In green IT we aim for the highest performance per watt, without loss of other quality aspects such as availability and response time. *Performance* is characterized by the amount of useful work accomplished by a computer system or computer network compared to the time and resources used.

1.1 THERE IS MORE THAN PUE

Just improving PUE (see chapter 2) should not be the ultimate goal. PUE can be useful as an indicator of energy efficiency of data centre facilities, but it is a very bad metric when it comes to judging how green your ICT infrastructure is.

REFLECTION 1
Why is PUE a very bad metric for energy efficiency?

The most important reason why your organization has a datacenter is that it wants to process data to support the organization in its business activities. The purpose of all facilities such as cooling, ventilation and power conversion is to make that processing possible. Therefore, getting as much productive processing work done with the lowest energy bill and carbon footprint is the ultimate goal. The optimization of IT equipment should be the starting point for any energy optimization or 'greening' strategy.

1.2 IT EQUIPMENT: OVERVIEW

IT equipment can be categorized as:

Servers compute units
1 *Servers* or '*compute units*' which perform most of the processing
Storage
2 *Storage* which store and archive the data
Communication equipment
3 *Communication equipment* which transfer the data

Depending on the category completely different strategies are necessary to optimize energy consumption. We will discuss servers in section 2, storage in section 3 and communication equipment in section 4.

2 IT Equipment: Servers

The energy consumption of a server or compute unit depends on:
1 The components
2 The form factor
3 The software part: settings of the power management system, cluster based operating system and the software workload
4 The workload

We will discuss the latter two in the next chapter and focus on components and form factor in this one.

2.1 COMPONENT CHOICES

Depending on the workload, the power consumption of a certain component will vary greatly, see Figure 1.

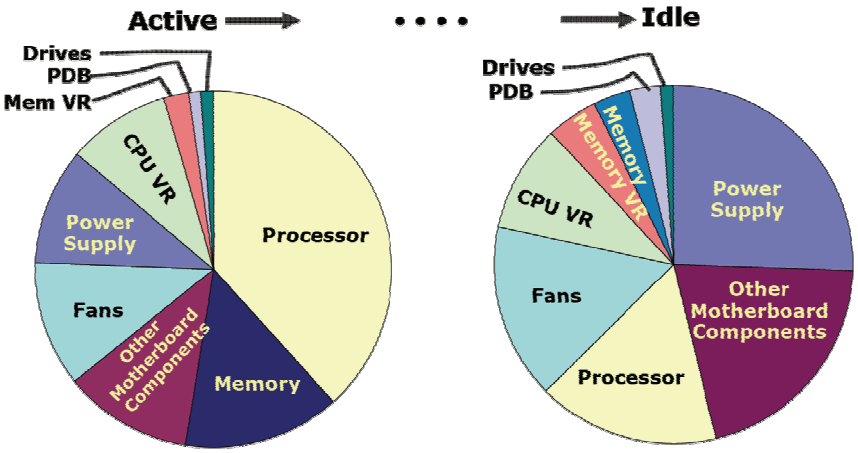


FIGURE 1 Energy consumption under load and idle (source: Intel)

Consumption Losses

When a server workload gets lots of requests, the processor and memory energy *consumption* greatly increases. At idle load, the energy *losses* at the power supply increase.

So your first task should be to determine whether or not your servers run most of the time close to idle or close to maximum load. You can use the operating system's monitoring tools. For Windows, this is typically 'Perfmon'. For Linux and Unix operating systems, tools like 'vmstat', 'dstat' and 'mpstat' are used. VMware vSphere and other virtualization solutions have their built-in monitors. Or you can use the Datacenter Information System Management (DCIM) software, described in chapter 6.

If you notice that most of your servers are running at low load, it is important to understand that this means that your processing per Watt ratio is low. The relatively high losses in the PSU, voltage regulation and the fact that only a small amount of power is spent on processing are the cause of this.

Increase server usage

So it is good to consider a strategy to *increase server usage*. For general server workloads such as web-, mail- , file- and database servers virtualizing your workloads might be the path to more efficiently working servers. If you have already virtualized most of your workloads, detecting and resolving storage bottlenecks is probably the best strategy.

Once you have exhausted these possibilities, you can focus on choosing the best components.

Low power CPUs

Low power CPUs: low power CPUs do not necessarily save energy. Modern CPUs place their cores in deep sleep states (c-states) as soon as they are not required. As a result, high performance CPUs can put cores in a longer sleep state more quickly than low performance CPUs (De Gelas, 2010). Most of the time, low power CPUs run at lower frequency and/or have fewer cores.

Lower power CPUs are handy to 'cap' power consumption. If your maximum available power is limited by your data centre, it is worth considering them.

If you are not limited by the memory capacity, choose one high performance CPU over two lower performance ones. The interconnections between the two CPUs and the synchronization communication between the two CPUs can easily add up to 15% of the total energy consumption of the CPU. (Rakesh Kumar, 2005)

Take into consideration though, that the popular NUMA based dual socket system typically offers twice as much memory capacity. In other words, memory capacity scales with the number of sockets.

Lower power memory

Lower power memory runs at a lower voltage and as a result the maximum frequency is limited. Few applications benefit from high memory speeds, so low power memory is a good trade-off in many cases.

Power supplies

Power supplies are rated by the voluntary 80+ certification program[1], discussed in Chapter 1. Several efficiency levels can be reached: Bronze, Silver, Gold, Platinum and Titanium.
At the minimum the PSU must reach 80% energy efficiency at 20%, 50% and 100% of the rated load. In other words, the PSU converts 80% of AC energy into useable DC energy, and converts 20% into heat. Most modern server power supplies get the gold level, which means they achieve 88% efficiency at 20% load (230V). An efficient PSU is extremely important as there is a 'cascade' effect as described later.

Local Storage

Local Storage: many servers use the local storage only to start up and connect to some form of network attached storage. Therefore using magnetic disks with spinning platters is not a good practice. Some vendors offer the possibility to boot from a small amount of internal flash memory. Since flash memory uses a negligible amount of power at idle, this is a good choice for many servers.

Voltage regulation (VRs)

The quality of *the voltage regulation (VRs)* is another important part.
The voltage regulation of the CPU and memory is an important part and some vendors reduce power at low CPU load by reducing the amount of 'steps' that are used to convert the higher DC voltage to the very low CPU voltages. However, there is no simple way to gauge the quality of the VRs.

The rest of the motherboard is less important, in most cases the power consumption of USB controllers, video chip and other components are negligible. Network and storage controllers are discussed in the storage and networking part.

2.2 FORM FACTOR CHOICES AND RACK UNITS

Chassis

In Chapter 2 we saw that a modern data centre consists of server racks, through which cool air is blown, and we saw that the air flows this creates gives rise to hot and cold aisles. The server rack is also called the *chassis* and the form factor of the chassis determines the airflow, heatsink and fan size and other thermal characteristics. As a result the form factor has an important impact on the energy efficiency of IT equipment.

We can discern five different form factors:
1 Tower servers
2 1U rack servers
3 2U, 3U and higher rack servers
4 Blade servers
5 Micro servers

We will discuss the form factors. But to understand form factors, we must first discuss rack units.

Rack units

A *rack unit*, U or RU is a unit of measure that describes the height of equipment designed to mount in a 19-inch rack. The rack unit size is based on a standard rack specification as defined in the industry standard EIA-310[2].

[1] http://www.plugloadsolutions.com/80PlusPowerSupplies.aspx
[2] http://engineers.ihs.com/document/abstract/SBSSIBAAAAAAAAAA

Multiple of Units

Equipment mounting frame

One rack unit is 1.75 inches or 44.45 mm high. One rack unit is typically abbreviated to 1U. IT equipment that follows this industry standard will use *multiples of Units*. Two rack units are described as '2U', three as '3U' and so on. A typical full size rack is 42U, although half-height (21U) and higher racks exists (47U).
The 19-inch (482.6 mm) or 23-inch (584.2 mm) dimension refers to the width of the *equipment mounting frame* in the rack.
For a 1U Rack server, see Figure 2.

FIGURE 2 1U server: notice that the airflow is always obstructed

2.2.1 *Tower Servers*

Tower servers are upright, free-standing servers. Typically 3 to 4 U wide, these servers come with large fans and thus relatively good airflow. They were very popular in the 1990s but do not have a place in the modern data centre based upon the standardized 19-inch mounting rack. Using tower servers that are not mountable in rack cabinets makes many energy saving techniques hard to implement. For example, making sure that the hot and cold aisles are working – paramount for an efficient datacenter – is much easier to achieve with rack servers.

Some tower servers mounted as 4U rack servers, but these servers should be categorized as 4U rack server, not as tower servers.

Individual units are fitted horizontally into a standard 19 inch mounting rack, with units located vertically above and below each other. These rack servers are optimized for maximum density as they use (floor) space and rack space very efficiently. Unfortunately, the height of 1U allows for only the smallest fans (40 mm) and the dense packing of components obstructs good airflow. These 40 mm fans have to turn much faster to keep the components cool.

As the energy consumption of a fan is proportional to the third power of the turning speed, power consumption can quickly grow exponential when the load on the server is high.

To add insult to injury, the expansion slots, known as mezzanine slots, for adding network interface cards (NICs) and other expansion cards, are vertical and further obstruct the airflow.

2U, 3U and higher Rack servers

2U, 3U and higher Rack servers

2U and *3U servers* are also designed for vertical arrangement within a standardized 19-inch mounting rack, but as the height of the rack server is increased, so is the fan size: 80 mm and higher, see Figure 3.
Heat sinks can of course be higher too.

FIGURE 3 HP DL380 2U server: notice the large grill allowing air to enter unobstructed

The airflow is in general a lot better than a 1U server as it is possible to design the server in such a way that the airflow can enter the server without obstruction and reach the hottest components (CPU, chipset) very quickly.

In addition, 2U and higher rack server allow the use of half-height horizontal expansion cards.
It is important to note that relative to server blades and enclosures, rack servers are more limited in the number of new drives and memory you can install.

Rack servers are generally designed to work as a logical and cohesive whole but without the tight integration found with server blades, which makes rack servers more flexible in some situations. In addition, you can run servers from different manufacturers in the same rack unit because the servers do not share proprietary components.

Blade servers

Blade servers

Blade servers are small form factor servers housed in a blade enclosure or chassis. These blade chassis offer shared components like the PSUs, fans, networking and other interconnects. As 10 or more blade servers share these components, there are some efficiency gains.

For example, fewer and larger PSUs (2+2) equates to less heat losses per PSU. Also the reduction of network interface and storage cards can lead to reduced power consumption.

The typically 7 to 10U height blade chassis can also use very large and thus efficient fans.

REFLECTION 2
Why are large fans almost always more efficient than small ones?

However, the energy efficiency of such large fans can quickly be negated by the fact that extremely dense blade servers (sometimes up to 14 in a 7U chassis) need a much higher airflow to stay cool.

Micro servers

Micro servers are similar to blade servers, but the chassis vary wildly from 2U to 10U and larger. A micro server is not a standardized form factor. There are some common characteristics though. An example is found in figure 4.

Micro servers typically consist of many small form-factor system-on-a-chip (SoC) boards. The SoC consists of the CPU, memory, I/O interfaces and other integrated in one die. Micro server SoCs are typically based upon the low power (<20 W), relatively simple ARM or Intel Atom processors. However, low power, highly integrated versions of the 'brawny' AMD Opteron and Intel Xeon are also used.

These small SoC boards interface with a fabric, and get access to an Ethernet and/or storage network this way.

Micro server chassis are thus by definition server clusters that contain many small server nodes. In fact, the small size of the boards allows tightly packed clusters of micro servers to be built, making micro server clusters among the most dense server chassis in the datacenter.

Micro server nodes are stripped down to the essence: the idea is to save power by tailoring the performance of an individual server node to a small task. As there little room for extensive cooling, the server node is limited to one low power SoC and much less RAM than in traditional servers.

Micro servers are typically suited for lightweight workloads that scale easily over many nodes. Serving up static web content is one example, as this workload is dominated by loading items and the actual processing part is very small.

FIGURE 4 SeaMicro micro server cluster chassis

High Performance Computing (HPC)

Micro-servers are thus only efficient for some tasks. As the RAM and CPU power per node is limited, they are not suitable as a platform for running *High Performance Computing (HPC)*, virtualization or database applications. Examples of micro servers are SeaMicro's SM clusters and HP's Moonshot servers.

Using micro servers can be an effective strategy to reduce power consumption providing that the right workload is run upon them.

3 Networking Equipment

3.1 NETWORKING EQUIPMENT POWER SAVING STRATEGIES

Power saving strategies can be categorized according to the level at which they happen:

Networking device level Technology
– Optical node
– Clock gating
– Sleep control
– ALR
– Circuit switch
– Thermal design

Network server techniques
– Cache server
– Filtering
– Sorry server
– CDN

Network techniques
– Lightweight protocol
– Energy-aware network planning

We will discuss these techniques in the next subsections.

3.2 DEVICE LEVEL TECHNOLOGY CHOICES

The following power saving techniques occur inside the network devices or components.

Optical node
It is interesting to note that optical transmission technology – given a certain power budget – can transfer at much higher rates compared to copper based networking technology. Or in other words, at the same speed, optical networks consume less energy.

The savings are especially large at the physical level.

An Ethernet PHYceiver is a chip that implements the hardware send and receive function of Ethernet frames; it interfaces to the line modulation at one end and binary packet signaling at the other.

The PHY chip that transcodes the binary packet signal (for example: Ethernet Frames) into a voltage signal (line modulation or vice versa) on the copper medium typically needs up to 3 times more than an optical PHY. [3]

Therefore, although optical signalling is more expensive than transmitting over a copper medium, it is important to remember that the former is more energy efficient.

Clock gating

Clock gating logic avoids that the clock signal spreads to parts of the chip that are not used. The clock gated parts of the circuitry do not have to switch states and save power this way. We will discuss it in more detail in the next chapter.

Sleep control

Simply put, *sleep control* is a technology to put equipment and interfaces to 'sleep' when they are not used, and save the energy consumption.

This is however difficult to implement as the current protocols results in continually probing and polling devices (for example ARP, RARP).

To realize sleep control, the network nodes and attached nodes must support sleep control as described in the energy efficient Ethernet protocol developed by IEEE P802.3az[4] Energy Efficient Ethernet Task Force.

ALR

ALR or Adaptive Link Rate is the technology that allows switching between link speeds depending on the network traffic demand. Typically, an interface will switch between 10/100/1000/10k Mbit/s. As the higher speeds come with more complex encoding, they require more energy. A 10 G Ethernet interface can use up to 15W, a 100 Mbit/s less than 0.5 W.

Circuit switch

Circuit switches consumes less energy than packet switches, because of the simplicity and the fact that these switches does not need extra energy-consuming devices such as SRAM and CAM. Circuit switching cannot realize statistical multiplexing, so it can degrade network performance.

Thermal design

*The thermal desig*n of a network node can be significant. Clever and large heat sink designs can avoid the use of fast turning fans, saving a considerable amount of energy.

3.3 NETWORK SERVER TECHNIQUES

Power Savings can not only come from reducing the power of the individual network devices, but also from making that the core server and networking infrastructure is not used unnecessary.

[3] http://www.intel.com/content/dam/doc/brochure/ethernet-controllers-phys-brochure.pdf
[4] http://www.ieee802.org/3/az/

Cache/proxy server

Cache/proxy server

Caching the most used web content can seriously reduce the amount of bandwidth and the load on servers and storage. As a result less servers and storage devices are necessary to support the same amount of traffic and users. Typically, these servers use either a distributed memory cache (like Memcached) or some form of proxy.

Distributed memory cache the most used objects in the unused RAM of several servers, reducing the amount of complex server and storage requests.

A proxy intercepts all clients requests and checks its local cache for a copy of the resource. If none is found, it makes the request on the client's behalf and then relays this back to the client, caching a copy itself in the process. The next time the same resource is requested, a cached copy is then already available. As some resources tend to be very popular, the load on the rest of the infrastructure is reduced and hence also the energy consumption.

Filtering

Filtering

Filters/firewalls blocks unnecessary/unwanted accesses and thus reduce the amount of traffic and load on the infrastructure. A decently configured firewall can thus save quite a bit of energy, besides the security benefits.

Sorry server

Sorry server

A *sorry server* is a server which warns clients that some services cannot be reached, hence lowering the chance that they repeatedly try to reload the content. As a result it can reduce the amount of peak traffic. Of course, the sorry server must be able to run at very low power when its services are not needed, otherwise the energy balance might be negative as organization size for high traffic and try to avoid congestion at all costs.

CDN

Content Delivery Network (CDN)

A *Content Delivery Network (CDN)* is a network of servers deployed in multiple data centers across the Internet. When a user requests a webpage, the closest CDN server will respond. CDN mostly serve up static content (media files, static text, scripts and documents) of the requested website. CDNs lower the response latency and save energy as fewer network devices are handling the packets (Neilson, 2010). Of course, the CDN must be able to serve up the requested content most of the time. Just like a cache server, the 'hit rate' must be high to save energy.

REFLECTION 3

What similarity exists between all network server power saving techniques?

3.4 NETWORK TECHNOLOGY TECHNIQUES

Lightweight network protocols

Lightweight network protocols

Simpler network protocols might save energy. One example is using the UDP transport protocol when TCP connections are not necessary. Other examples are the LWAPP and CAPWAP protocol that reduces the signalling overhead for WLAN developed by IETF CAPWAP workgroup[5].

The choice of routing protocol can also have an energy impact. (Joseph Chabarek, 2009)

Energy-aware network design

Energy-aware network design

Current networks are usually designed with the goal of optimal performance and reliability. Energy consumption is rarely considered. Avoiding overprovisioning of network bandwidth in network sections can save some energy. Another example is avoiding multicast traffic being spread over a larger part of the network than necessary.

The network can also be designed such that power-hungry packet processing operations are limited to a subset of the routers. (Joseph Chabarek, 2009)

4 Storage Equipment

4.1 OVERVIEW OF STORAGE EQUIPMENT POWER SAVING

Power saving strategies can be categorized according to the level at which they happen:

At disk level

At disk level

Replacing magnetic disks with flash
Consolidate drives
Spinning disks down
Variable rotations per minute (RPM)
Hybrid disks with Flash cache

At the storage system level

At the storage system level

Tiering
Deduplication
Hot and Cold Zones

4.2 AT THE DISK LEVEL: INTRODUCTION

The disk level is the physical magnetic or flash media (e.g., hard disks, SSDs…) used for persistent storage.

Hard disk

The *hard disk* is the most common storage medium in use today for laptops, desktops, and data centres. A hard disk consists of one or more rotating magnetic platters. Each platter is divided into concentric tracks and sliced into arc-shaped sectors, containing fixed-size chunks of bytes (e.g., 512 bytes, 4KB, etc.). A sector is accessed when the rotating motor spins and the head – mounted on an actuator arm – is positioned over it.

[5] http://www.ietf.org/html.charters/capwap-charter.html

The time it takes to access data on a hard disk is referred to as the disk access time and consists of the seek time (the time it takes for the arm to move to the correct track), the rotational delay (the time it takes for the platter to rotate so that the target sector is under the disk head) and the transfer time (the time it takes to transfer the sectors).

A modern disk has multiple power modes, each of which represents a level of tradeoff between power and access time:
– Active
– Platters spinning, head working
– Platters spinning, head idle
– Spun Down
– Disk spun down, head idle

Spinning disks down is thus interesting for servers that only use storage to boot up. However, as hard disks are the most important bottleneck in most storage systems, this is not option in all other situations.

4.2.1 *Flash as the 'Performance Provider'*

Flash storage is becoming popular as an alternative to hard disk drives. The largest advantage is that NAND flash can achieve much faster read and write access times than magnetic disks. Samsung specifies[6] 0.29µs for reads and 100µs for writes for the Samsung SM825 Enterprise drive for example.

Flash stores data bits in memory cells, each cell is made from floating-gate transistors, see Figure 5.

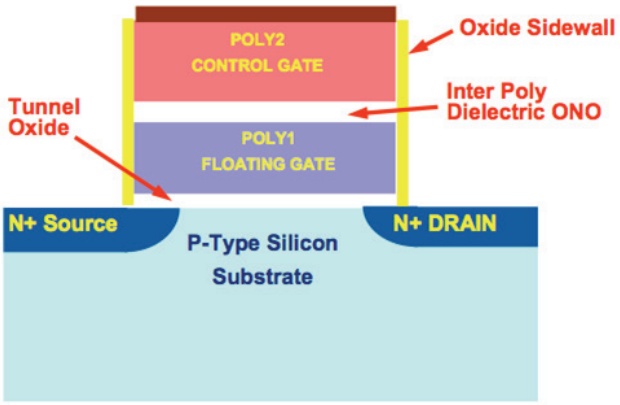


FIGURE 5 Flash cell

The number of electrons trapped in the floating gate can control the resulting charge level of a memory cell which is then used to represent discrete bits of data.

[6] http://www.samsung.com/us/business/oem-solutions/pdfs/SSD-FAQs.pdf

NAND flash is the most popular form of flash because it offers the lowest cost-per-gigabyte ratio. NAND flash is addressable in pages and must be read in this larger 'page' granularity (Most of the time 4 KB). Additionally, NAND flash is erased in erase blocks, which consist of many contiguous flash pages (typically 512 KB). Therefore, a single erase of any size on NAND flash will erase many pages.

One major disadvantage of flash is that it can only write to empty pages. This means that in order to over-write a flash, the entire erase block which contains the page must first be erased. Erasing in flash can thus be orders of magnitude slower than simply reading.

Additionally, each memory cell may only be erased a finite number of times, typically in the range of a few thousand times. In order to increase the lifetime of a flash device, wear leveling is often employed, which attempts to spread erase requests evenly over the erase blocks on a flash device. This allows vendors to guarantee that flash devices can survive a very high amount of writes.

Although Flash drives do not consume much less per GB than magnetic they can deliver an average access time that is easily 100 times lower. The performance/watt is thus a magnitude better. As the performance of one flash device can easily replace a few tens to a hundred magnetic disks, it is clear that huge power savings are possible by replacing magnetic spindles by flash drives.

When you make sure that your performance requirements are being taken care of by using flash devices, it is possible to reduce the power usage of magnetic disks in different ways, as we will describe in the next sections.

4.2.2 *Consolidate Drives*

One simple way to decrease power consumption of the storage devices at the hardware level is to consolidate data from many small drives onto fewer, bigger hard disks. This can be illustrated easily by comparing the specs of two drives of different capacity:

Name	Rotational speed	Capacity	Average Operating Power
ST4000NM0053	7200 rpm	4 TB	11.27W
ST1000NM0021	7200 rpm	1 TB	7.84W

[7] Source: Seagate

One 4 TB disk needs a bit more than one third of four 1 TB disks. Of course, having fewer disks also means that you have fewer spindles and thus performance.

[7] http://www.seagate.com/internal-hard-drives/enterprise-hard-drives/hdd/enterprise-capacity-3-5-hdd/

Flash disks for performance Magnetic disks for capacity

As discussed above, as hard disks come with a poor performance per Watt ratio, an energy-savvy strategy would always use *flash disks for performance* and *magnetic disks for capacity*.

In other words, all performance requirements are tackled by using flash devices while the magnetic disks are only used for applications where capacity is far more important than performance (archiving, back-up etc.)

4.2.3 *Other Techniques at the Disk Level of Storage*

Spinning down

Spinning down

One approach for power savings in magnetic disks is to predict periods of disk idleness. The basic strategy is to spin the disk down when the predicted idle period is long enough to save more energy than that required to spin the disk back. Until recently, this was a hard to implement strategy as spin-up time of a disk is several seconds, which results in unacceptable long response times in most applications. However, when flash disks take care of the latency sensitive accesses, this can be viable.

Variable RPM

Variable RPM

In some cases, spinning down is not an option as several seconds of waiting time – even if this happens on rare occasions – is not acceptable.

Reducing the RPM might be an option as the total access time will be a few microseconds higher, but not seconds. Spinning the platters faster than necessary wastes power, and a slower RPM may be sufficient to meet some workload demands.

The offering of several levels of rotational speed, is often controlled by the value of a register on the disk. Several 'Green disk drives' support this, but the power savings are very modest. As a result, disks with variable RPM have been phased out for models that use different power saving features but deliver the same performance as other hard disks.

Hybrid Drives

Hybrid Drives

Hybrid disks are magnetic hard disks that contain a small amount of flash memory for caching and storing the most frequently read data.

If data is being requested out of the drive's NAND cache, the rest of the disk can spin down to save power. As a result, hybrid disks have been proven to be faster and more power efficient than traditional drives in read intensive environment, especially if random reads happen a lot. [8]

Hybrid drives are priced slightly higher than other drives of the same capacity. However, hybrid disks are still not common in the enterprise market. Most storage vendors prefer a tiered storage instead (see next subsection).

[8] http://www.anandtech.com/show/5160/seagate-2nd-generation-momentus-xt-750gb-hybrid-hdd-review

4.3 AT THE STORAGE SYSTEM LEVEL

While hyperscale datacenters such as those of Google and Facebook typically use off the shelf disks, most datacenters are equipped with network storage: Storage Area Networks (SAN) or Network Attached Storage (NAS). As a result, you have

4.3.1 *Tiered Storage*

Given that
1 most performance bottlenecks in the datacenter are located in the storage systems
2 the key to saving power for disks is to create opportunities for them to spin down
3 flash disks can handle the most random I/Os,

it is possible to improve the energy efficiency of a storage system by combining flash drives and magnetic disks in a tiered system. Tiered storage or *Hierarchical storage management (HSM)* is a data storage software technology, which automatically moves data between low-cost storage magnetic disks and fast, low latency media. In most cases, the first tier is based upon flash storage, and the higher tiers are based upon magnetic disks.

Hierarchical storage management (HSM)

Tiered storage can improve energy consumption: by diverting most of the I/O operations to the first tier, the second disks stay idle for much longer periods.

However, the implementation of the HSM can make a huge difference. Most vendors use automatic tiering, meaning that hot data is moved from the slow magnetic disks to the flash disk. This is a rather simple algorithm, which can waste a lot of energy. As the process is not real-time, the disks can handle a lot of requests before the data is finally moved towards the flash tier. Every migration of data, up or down the hierarchy, also wastes quite a bit of processing, storage bandwidth and thus energy.

Using the flash tier as a cache layer can be a lot more efficient in read intensive environments: data read is kept immediately inside the cache and an eviction algorithm decides which data is evicted and replaced.

Newer storage devices also use 'flash pools' for read *and* writes. These flash caches lower the energy consumption by making sure that random writes are 'serialized' before they are committed to the magnetic disks. As random writes cause much more activity on magnetic disks, this is quite effective. In many cases serial writes are directed immediately to the disks. The simplest caches use the FIFO algorithm, the more sophisticated and also energy efficient use an LRU (Least Recently Used) eviction algorithm.

4.3.2 *Deduplication*

Data deduplication is a data compression technique for eliminating duplicate copies of repeating data. Decrease storage utilization by avoiding duplicate copies may also yield energy savings as a side effect.

Deduplication can result in a storage system with fewer disks, or a storage system of which the upgrade to more disks can be postponed.

However, the tradeoff is the computational cost of hashing and bookkeeping to identify identical blocks. The processors inside the storage controllers and the deduplication algorithm have been as efficient as possible. The administrator may influence this by lowering the 'aggressiveness' of the deduplication algorithm.

Consolidating blocks with identical content may also lower the spatial locality of original blocks belonging to the same files, and increase the probability of disks performing more random seeks.

So again, it is important that a pool of flash disks makes sure that the most random accesses are not handled by the magnetic disks.

An empirical study (Lauro Beltrao Costa, 2011) assessed Data deduplication tradeoffs from an Energy perspective. They measured the power of the entire system and reported that energy savings can be achieved once the level of content similarity is above 18%. To break even with the deduplication computational overhead of the test system, the similarity level needs to achieve 40%. However, as deduplication software gets better and storage processors get better performance/watt, these percentages may decrease.

4.3.3 *Hot and Cold Zones*

In a study based upon real world servers, (Bhandarkar, 2010), the researchers observed that 89% of files are actively referenced for about 10 days after their creation, and 60% of files become inactive after 20 days.

These observations led to the design of GreenHDFS, which classifies files into hot (frequently accessed) files, and cold (low access frequency) files. Hot files should be stored on high performance storage,. Cold files are compressed and stored on high capacity storage with as many power management features enabled as possible. When hot files are no longer 'hot', they are moved from hot-zone machines to cold-zone machines. As cold files become popular, they are moved from the cold zone to the hot zone. This is in fact similar to tiered storage, but with a 'file focus' instead of a 'block' focus.

5 IT Equipment Energy Models

Processing is done by the CPU, and every other component around it is built to support this processing. That is why some IT equipment energy models consider the energy consumption of IT infrastructure as a sequential chain of energy losses starting with the processor.

One is called 'the cascade effect', and this model has been championed by the software division of Emerson[9].

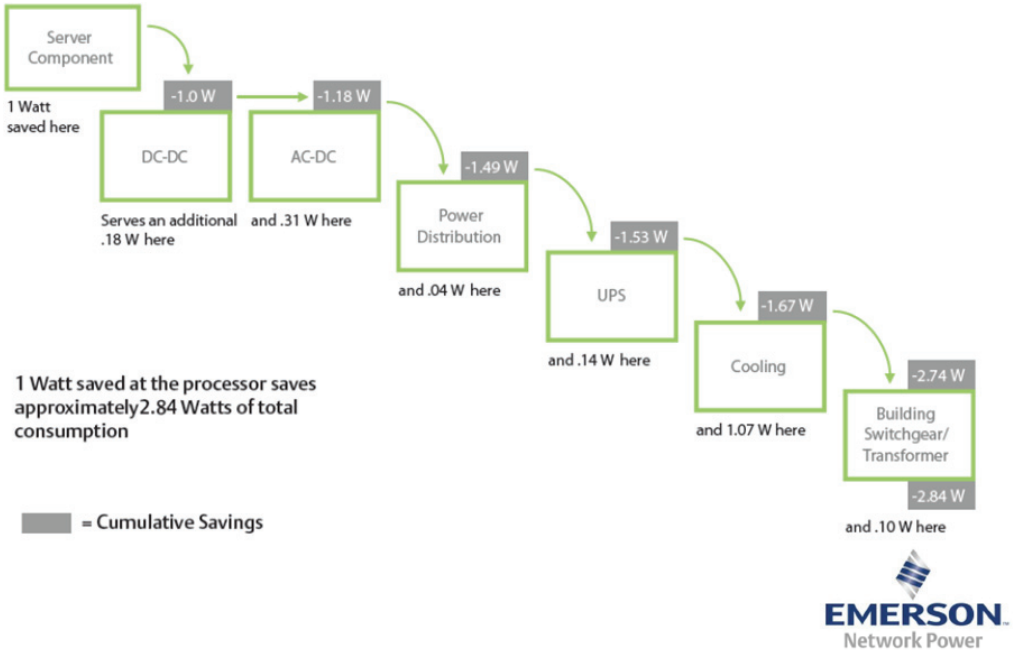


FIGURE 6 The cascade datacenter energy model (source: Emerson)

Although it is debatable that every 1Watt saved at the processor will result in total saving of 2.84W, the model clearly shows how important it is to understand the energy management of IT equipment. Saving power inside the rack can result in immediate and free savings in the power distribution, UPS, Cooling and transformer facilities of the datacenter. Most datacenters today have taken the opposite approach: as IT and facilities are two different worlds and people, both have been focusing on their own priorities. Only by making both worlds work together, the largest gains in energy efficiency can be made. The message of Emerson is quite unique as it is a company that is focused on datacenter facilities. That is why we mention this model in this course: it integrates several chapters.

Therefore the goal must be to reduce the PUE and the energy consumption of the IT equipment first before starting to look how the facilities can improve.

9 https://www.cisco.com/web/partners/downloads/765/other/Energy_Logic_Reducing_Data_Center_Energy_Consumption.pdf

SUMMARY

IT equipment, being the core of the datacenter should be the starting point of any power saving strategy. Every watt saved inside the rack multiplies to several more in the rest of the datacenter and ultimately is very visible in the carbon footprint of the datacenter.

IT equipment can be categorized in three categories: processing/servers, networking/communications and storage.

You can influence the power efficiency of your servers by choosing the right form factor, adapting your servers to your applications and carefully configure your servers with the right (low power) components.

You can influence the power efficiency of networking equipment by choosing the right network devices, by using some form of caching/filter servers and by designing your network not only for availability and performance but also for energy efficiency. The design of a power efficient network is a complex matter, way beyond the scope of this course.

The power efficiency of storage is more straightforward. It is clear that the starting point should be the use of flash, as the performance/watt is an order of magnitude better. The random read/write performance of one flash device can easily replace (10,000 IOPs vs 100 IOPs) a few ten to hundred magnetic disks. In a modern power efficient storage solution, flash devices should take care of the performance intensive applications, while magnetic disks should store the 'colder' data.

Once that goal is achieved, you can start looking at reducing the power usage of magnetic disks in different ways, such as spinning them down when they are not needed.

Literature

Bhandarkar, R. K. (2010). GreenHDFS: 2010. Towards An Energy-conserving, Storage-efficient, Hybrid Hadoop Compute Cluster. *Proceedings of the 2010 International Conference on Power-aware Computing and Systems* (pp. 1-9). Berkeley: USENIX Association.

De Gelas, J. (2010). *Saving server power in the real world.* Kortrijk: Sizing Servers.

Joseph Chabarek, J. S. (2009). *Power Awareness in Network Design and Routing.* Wisconsin: University of Wisconsin-Madison.

Lauro Beltrao Costa, S. A.-K. (2011). Assessing Data Deduplication Tradeoffs from an Energy and Performance Perspective. *Proceedings of Green Computing Conference and Workshops* (pp. 1-6). IEEE Computing Society.

Rakesh Kumar, V. Z. (2005). *Power consumption in point-to-point inter-connect architectures.* San Diego: IEEE.

SELF–ASSESSMENT[10]

1 What is the first priority of your 'green' datacenter?

2 Why should saving power in the IT equipment be the starting point? Give one example.

3 Why are blade servers not always an energy efficient solution despite the large fans?

4 Which kind of server form factor has no place in a 'green' datacenter?

5 What relation has ALR with network speed, and why should high network bandwidth and power savings be compatible with this technique?

6 What is the starting point in an energy efficient storage infrastructure?

[10] In a self assessement question the student can integrate the content of the chapter

MODEL ANSWERS

1 Answers to Reflection Questions

1 A little thought experiment. Consider the situation where you let the room temperature of your datacenter increase. If your datacenter is cooled by conventional CRAH units (see chapter 3), those CRAH units will consume less energy. At the same time, it is likely that the IT equipment will consume more as the fans have to turn faster to keep the components at the desired temperature. So the PUE is *higher whereas the total energy consumption is lower*!

As

PUE = (Energy equipment + Energy Facilities/ Energy Facilities)

An *increase* in IT equipment energy consumption results in fact in a better PUE. Therefore PUE is a bad metric.

2 Simple physics: smaller fans need higher RPM to offer the same amount of airflow. As power is linear with the third power of RPM, it is clear that high RPMs are very power inefficient.

3 They all try to lower the load of the infrastructure that is 'behind' them.

2 Answers to Self-Assesment Questions

1 Processing the data at a speed that satisfies your customers at the lowest carbon footprint possible.

2 Because every watt saved in IT equipment can result in further savings in the rest of the datacenter. The opposite is not always true. A more efficient UPS will not lower the energy bill of your servers for example. But low power servers will need less UPS infrastructure and thus save a quite a bit of energy.

3 Because some of them are built for density. The very narrow blades leave little space for the airflow requiring a high velocity airflow which requires low temperatures or high RPM fans.

4 1U-servers or 'pizza box' servers. The low height results in tiny fans with very high RPM, which will consume high amounts of energy just to cool the server. If density is the priority, micro or blade servers are a better alternative.

5 Adaptive Link Rate allows the network device to use a lower power link speed when traffic is low, but can change quickly to higher link speeds if necessary.

6 Making use of the incredibly low random latency of flash for the performance requirements. Performance is to be handled by flash, capacity by magnetic devices.