



Energy-Efficient Computing and Data Centers

**Luigi Brochard, Vinod Kamath
Julita Corbalán, Scott Holland
Walter Mittelbach and Michael Ott**



ISTE

WILEY

Energy-Efficient Computing and Data Centers

Series Editor
Serge Petiton

Energy-Efficient Computing and Data Centers

Luigi Brochard
Vinod Kamath
Julita Corbalán
Scott Holland
Walter Mittelbach
Michael Ott

ISTE

WILEY

First published 2019 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

© ISTE Ltd 2019

The rights of Luigi Brochard, Vinod Kamath, Julita Corbalán, Scott Holland, Walter Mittelbach and Michael Ott to be identified as the authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

Library of Congress Control Number: 2019940668

British Library Cataloguing-in-Publication Data

A CIP record for this book is available from the British Library

ISBN 978-1-78630-185-7

Contents

Introduction	ix
Chapter 1. Systems in Data Centers	1
1.1. Servers.	1
1.2. Storage arrays	3
1.3. Data center networking	4
1.4. Components	5
1.4.1. Central processing unit.	5
1.4.2. Graphics processing unit.	7
1.4.3. Volatile memory	8
1.4.4. Non-volatile memory	10
1.4.5. Non-volatile storage	10
1.4.6. Spinning disks and tape storage	13
1.4.7. Motherboard.	15
1.4.8. PCIe I/O cards.	16
1.4.9. Power supplies	17
1.4.10. Fans	18
Chapter 2. Cooling Servers	19
2.1. Evolution of cooling for mainframe, midrange and distributed computers from the 1960s to 1990s.	19
2.2. Emergence of cooling for scale out computers from 1990s to 2010s	20
2.3. Chassis and rack cooling methods	23
2.4. Metrics considered for cooling	27
2.4.1. Efficiency	27
2.4.2. Reliability cost	28
2.4.3. Thermal performance	29

2.5. Material used for cooling	31
2.6. System layout and cooling air flow optimization	32

Chapter 3. Cooling the Data Center 37

3.1. System cooling technologies used	37
3.2. Air-cooled data center	38
3.2.1. Conventional air-cooled data center	38
3.3. ASHRAE data center cooling standards	40
3.3.1. Operation and temperature classes	40
3.3.2. Liquid cooling classes	41
3.3.3. Server and rack power trend	42
3.4. Liquid-cooled racks	43
3.5. Liquid-cooled servers	46
3.5.1. Water heat capacity	46
3.5.2. Thermal conduction module	47
3.5.3. Full node heat removal with cold plates	48
3.5.4. Modular heat removal with cold plates	50
3.5.5. Immersion cooling	50
3.5.6. Recent DWC servers	51
3.6. Free cooling	54
3.7. Waste heat reuse	54
3.7.1. Reusing heat as heat	54
3.7.2. Transforming heat with adsorption chillers	55

Chapter 4. Power Consumption of Servers and Workloads 65

4.1. Trends in power consumption for processors	65
4.1.1. Moore's and Dennard's laws	68
4.1.2. Floating point instructions on Xeon processors	72
4.1.3. CPU frequency of instructions on Intel Xeon processors	73
4.2. Trends in power consumption for GPUs	74
4.2.1. Moore's and Dennard's laws	77
4.3. ACPI states	78
4.4. The power equation	83

Chapter 5. Power and Performance of Workloads 87

5.1. Power and performance of workloads	87
5.1.1. SKU power and performance variations	87
5.1.2. System parameters	89
5.1.3. Workloads used	92
5.1.4. CPU-bound and memory-bound workloads	92
5.1.5. DC node power versus components power	93

5.2. Power, thermal and performance on air-cooled servers with Intel Xeon.	94
5.2.1. Frequency, power and performance of simple SIMD instructions	95
5.2.2. Power, thermal and performance behavior of HPL	98
5.2.3. Power, thermal and performance behavior of STREAM.	103
5.2.4. Power, thermal and performance behavior of real workloads	107
5.2.5. Power, thermal and frequency differences between CPUs.	115
5.3. Power, thermal and performance on water-cooled servers with Intel Xeon.	124
5.3.1. Impact on CPU temperature.	124
5.3.2. Impact on voltage and frequency.	125
5.3.3. Impact on power consumption and performance	127
5.4. Conclusions on the impact of cooling on power and performance	131

Chapter 6. Monitoring and Controlling Power and Performance of Servers and Data Centers

6.1. Monitoring power and performance of servers	133
6.1.1. Sensors and APIs for power and thermal monitoring on servers . .	134
6.1.2. Monitoring performance on servers	140
6.2. Modeling power and performance of servers	142
6.2.1. Cycle-accurate performance models.	142
6.2.2. Descriptive models	142
6.2.3. Predictive models.	144
6.3. Software to optimize power and energy of servers.	149
6.3.1. LoadLeveler job scheduler with energy aware feature	150
6.3.2. Energy Aware Runtime (EAR).	151
6.3.3. Other run time systems to manage power.	153
6.4. Monitoring, controlling and optimizing the data center	154
6.4.1. Monitoring the data center	154
6.4.2. Integration of the data center infrastructure with the IT devices . .	156

Chapter 7. PUE, ERE and TCO of Various Cooling Solutions

7.1. Power usage effectiveness, energy reuse effectiveness and total cost of ownership	159
7.1.1. Power usage effectiveness and energy reuse effectiveness	159
7.1.2. PUE and free cooling	162
7.1.3. ERE and waste heat reuse	163
7.2. Examples of data centers PUE and EREs	164
7.2.1. NREL Research Support Facility, CO.	164
7.2.2. Leibnitz Supercomputing data center in Germany	166
7.3. Impact of cooling on TCO with no waste heat reuse	173

7.3.1. Impact of electricity price on TCO	177
7.3.2. Impact of node power on TCO	178
7.3.3. Impact of free cooling on TCO	181
7.4. Emerging technologies and their impact on TCO	183
7.4.1. Waste heat reuse	184
7.4.2. Renewable electricity generation.	189
7.4.3. Storing excess energy for later reuse	192
7.4.4. Toward a net-zero energy data center	193
Conclusion	195
References	199
Index	209

Introduction

As shown by a recent news article in *Nature* (Jones 2018), data centers consume about 1% of the total electricity demand while information and communication technology (ICT) as a whole, including personal devices, mobile networks and TV consumes about 10%. It shows also that the demand will grow exponentially in the near future leading in 2030 to a ratio of electricity demand varying depending on the estimations between 3% and 8% for data centers and between 8% and 21% for ICT. An article in the *Guardian* (2017) shows a similar accelerating trend.

The energy consumed by a data center during a given period is the sum of the energy consumed by all the workloads that have been executed, plus the energy consumed when devices are idle, plus the energy loss to convert the electricity from the power lines down to the IT devices, plus the energy to cool the IT devices. This book will tackle all four aspects of energy.

The energy consumed by a workload when running on a system is the integral of its power consumption over its execution time from beginning to end. If the power consumed by the workload is constant, it would simply be:

$$\text{Energy} = \text{Power} \times \text{Time}$$

One trivial way to minimize the power of a workload while running on a system is to reduce the frequency of the processor. But this can be counterproductive since the elapsed time will very often increase to a point where the total energy is constant or has increased. From an investment perspective, it is also obvious that a system and a data center have to be used as much as possible and run as many workloads as possible. That is why energy efficiency, and not only power efficiency, is critical.

Existing data centers have two physical limits they cannot exceed: floor space and power supply. They also have an economical limit, which is their budget for operational costs.

To keep up as much as possible with Moore’s law, power consumption of IT devices is also on the rise, so that existing data centers are facing the dilemma of either keeping the same density of servers per rack (leading to a tremendous power and cooling challenge) or giving up on the density. That is one reason why numerous new data centers are being built around the world by hyperscale, cloud and high-performance computing data centers.

This increasing energy demand impacts not only the planet’s ecosystem but also the ICT business as increasing electricity demand leads to even more increased energy cost since the price of electricity itself is on the rise.

According to the U.S. Bureau of Labor Statistics, prices for electricity were 66% higher in 2018 versus 2000 which is 21% higher than inflation during the same period (Official Data Foundation 2018). A similar trend is seen in Europe (Eurostat n.d.).

Figure I.1 shows the wide range of electricity prices in 2018 around the world.

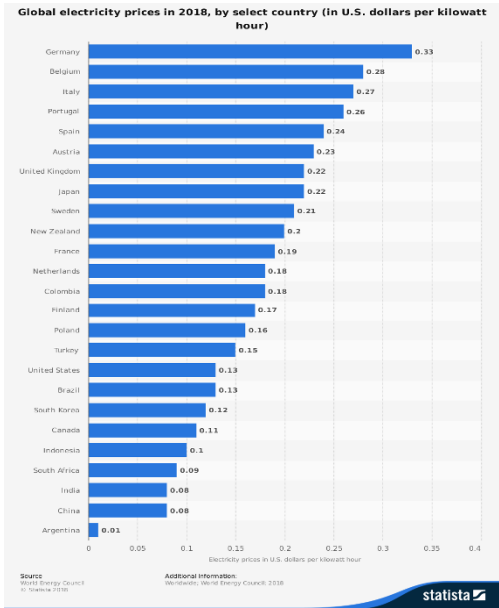


Figure I.1. Price of electricity around the world in 2018, from Statista.com

These prices are average price across different locations in a country. For example, the average price of electricity in the United States is \$0.13 per kW/h, but it differs widely across the different states with locations where it can be as low as \$0.06 per kW/h in Colorado. This variability of electricity price associated with the environmental conditions (temperature, humidity, etc.) of the location does play a very important role in the power and cooling design of a data center.

Power usage effectiveness (PUE) is a measure of how efficiently a computer data center uses its power. PUE is the ratio of total energy used by a data center facility to the energy used by the IT computing equipment. The ideal value is 1.0.

$$PUE = \frac{\text{Total energy}}{\text{IT equipment energy}} \quad [I.1]$$

In the following, we will use PUE as the cooling efficiency not taking into account the power conversion losses. Legacy data centers usually have a PUE of about 2 and with a careful design it is possible to get around 1.1. Table I.1 shows the numbers of years needed for the energy cost to power and cool a server to equal the server acquisition cost with various PUE and electricity prices.

PUE =	2	1.1
electricity price	number of years	number of years
\$0.06	12.7	23.1
\$0.21	3.6	6.6
\$0.33	2.3	4.2

Table I.1. *Impact of PUE on price of electricity versus price of server*

It shows how the energy costs can quickly equal the acquisition cost with a high PUE and medium to high electricity price. With a low electricity price, even with a high PUE, this will take more than 10 years and even close to 20 with a low PUE. This illustrates the importance of the data center location that impacts both PUE (through the possible use of free cooling) and electricity cost on the cooling infrastructure design.

Although PUE is an important metric, it does not take into account how much power the IT device is consuming and its power efficiency.

IT energy effectiveness (ITUE) measures how the IT equipment energy is optimized where voltage regulator (VR), PSU and fan are the energy consumed by the voltage regulator, power supply and fan. The ideal value is 1.0.

$$ITUE = \frac{IT\ equipment\ energy + VR + PSU + fan\ power}{IT\ equipment\ energy} \quad [I.2]$$

We will study all ITUE components not in isolation, but in relation to the type of workload the server is running since the server power consumption depends on the workload running on it.

But even with a low PUE and ITUE, the heat produced by the systems in the data center is still lost without any waste heat reuse.

Energy reuse effectiveness (ERE) measures how efficient a data center reuses the energy dissipated by the computer. ERE is the ratio of total energy used by the facility minus the waste heat energy reuse divided by the energy used by the IT computing equipment.

$$ERE = \frac{Total\ facility\ energy - energy_reuse}{IT\ equipment\ energy} \quad [I.3]$$

An ideal ERE is 0.0. With no heat reuse, $ERE = PUE$ and we have always:

$$ERE \leq PUE \quad [I.4]$$

Many articles have been focusing on the PUE of data centers but very few have been looking at the global equation $ITUE + PUE + ERE$.

This is what we will address in this book.

Chapter 1 presents the different IT devices in a data center and the different components in a server from a power consumption perspective.

Chapter 2 presents the different cooling technologies in a server, their evolution and the trade-off between density and cooling.

Chapter 3 presents the different cooling technologies in the data center based on air cooling and liquid cooling, the ASHRAE standards and how waste heat can be reused in particular with adsorption chillers.

Chapter 4 presents the Xeon processor and NVIDIA accelerator evolution over the last 15 years in terms of performance and power consumption with reference to Moore's and Dennard's laws. It also presents the impact of microarchitecture evolutions and how the evolution in microprocessor design is power driven.

Chapter 5 analyzes the power, thermal and performance characteristics of a server when running different types of workloads in relation with the type of instruction executed. It also compares these metrics depending on the server cooling technologies.

Chapter 6 presents hardware and software to measure, model, predict and control power and performance in a server and in a data center in order to optimize their energy consumption.

Chapter 7 analyzes the PUE, ERE and total cost of ownership (TCO) of existing and new data centers with different cooling designs and the impact of power consumption, electricity price, free cooling and waste heat reuse. Through two data center examples it highlights also how a careful data center design, free cooling and waste heat reuse can reduce PUE, ERE and TCO and save up to 50% of energy. It concludes showing how renewable energy can be used to provide the remaining energy needed by the data center.

Acknowledgments

This book is the result of a 12-year experience with IBM and Lenovo designing energy-efficient servers and data centers in collaborations with high-performance customers around the world. I would like to mention and thank in particular the Leibniz Supercomputing Center in Munich (LRZ) for the work we did since 2011 to design and deliver the most energy-efficient solutions. This book would not have been possible without the help of colleagues at Lenovo Data Center Group, with the application team (Eric Michel, Peter Mayes, Christoph Pospiech), with the power management team (Robert Wolford) and with the data center team (Jerrold Buterbaugh), at Intel (Andrey Semin), NVIDIA (François Courteille), IBM (Ingmar Meijer, Vadim Elisseev) and LRZ (Herbert Huber). Special thanks also to Peter Mayes for his rereading.

Systems in Data Centers

There are different types of IT equipment that serve different functions depending on customer application. This chapter provides an overview of servers, storage arrays, switches and their components.

1.1. Servers

A server is a broad term describing a specific piece of IT equipment that provides computing capability and runs software applications in an environment networked with other IT equipment, including other servers. Most servers contain the following major hardware building blocks: processors, memory, chipset, input/output (I/O) devices, storage, peripherals, voltage regulators (VRs), power supplies and cooling systems. Additional application-specific integrated circuits (ASICs) may be necessary, such as an onboard redundant array of independent disks (RAID) controller and a server management controller.

Volume rack-mounted servers are designed to fit within commonly available rack sizes, such as the 19 in. (0.5 m) rack form factor defined by EIA/ECA Standard 310-E specification¹. The vertical dimension is expressed in terms of rack units or just units (U). One U or 1U represents 1.75 in. (44.45 mm) of vertical height within a rack. Servers used for computing are available in standard rack-mount and custom configurations. Typical dimensions and sizes for standard rack-mount compute servers are full-width 1U, 2U or 4U. A single-server chassis may contain multiple server

¹ Available at: <https://www.server-racks.com/eia-310.html> [Accessed May 7, 2019].

nodes. Each node is defined as containing all key components, except power supplies, needed to make up a complete server. These nodes simply share the larger chassis infrastructure to conserve data center space. For more dense servers, there are 1U and 2U server enclosures that house several 1U $\frac{1}{2}$ -width servers.

Microservers are an emerging technology. They are based on system on a chip (SOC) design where all the functions, which are located on a motherboard for a classic server, are integrated on a single chip with the exception of memory, boot flash and power circuits. SOC are usually less power hungry than usual microprocessors leading to microservers that are more dense than classic servers. Although microservers and SOC are not analyzed in the following chapters, they are worth mentioning. These servers generally provide sufficient, targeted performance with optimized performance-per-watt capability, while being easily scalable with shared power and cooling infrastructure for individual servers.

To achieve even higher compute density than the 1U form factor, blade servers are another option. Each manufacturer designs their blades based on their own packaging and design goals. These blade chassis range from 3U to 10U tall and can house many blades. Blade servers are the result of technology compaction, which allows for a greater processing density in the same equipment volume. The greater processing density also results in greater power and heat density, further complicating data center cooling. Server components that had previously been packaged inside the tower/pedestal or rack-mounted server (e.g. fans and power supplies) are still required, but these components are now located within a chassis (or enclosure) that is designed to house multiple blade servers in a side-by-side or stacked configuration. Most of the time, the blade chassis includes networking, management and even storage functions for the blade servers, while the blade server integrates at least one controller (Ethernet, fiber channel, etc.) on the motherboard. Extra interfaces can be added using mezzanine cards.

Figure 1.1 illustrates 1U, 2U, 4U full width servers, a 2U chassis hosting four $\frac{1}{2}$ -width nodes and a 5U blade chassis hosting eight blades.

Examples of such servers will be given in Chapter 2.

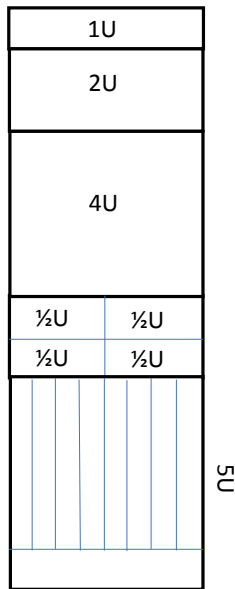


Figure 1.1. 1U, 2U, 4U full width servers, 2U chassis with four 1/2-width nodes and a 5U blade chassis with eight blades including a 1U base for power supplies

1.2. Storage arrays

Disk arrays are enclosures that provide ample non-volatile storage space for use by servers. Like servers, and depending on the scale of deployment required, the storage configuration may be a standard rack-mounted system with varying unit height or possibly a custom stand-alone piece of equipment. Disk storage arrays are typically designed for use in EIA/ECA Standard-310-E-compliant racks. The enclosure contains the storage in either small or large form factor drives in addition to the controllers, midplane and batteries for cache backup. The storage array enclosure typically uses redundant power supplies and cooling in the event of component failures. One of the more challenging aspects of storage arrays is the design of a battery backup system to prevent data loss in case of a power interruption or loss. For an in-depth discussion of storage array thermal guidelines, please consult the ASHRAE storage equipment white paper (ASHRAE TC 9.9 2015).

While disk storage arrays are typically used for online storage, backup and disaster recovery, tape storage is known for its low cost and longevity

for archiving purposes. Tape systems come in a variety of different formats based on the type of tape media.

1.3. Data center networking

A computer network, or simply a network, is a collection of computers and other hardware interconnected by communication channels that allow sharing of resources and information. Networking equipment facilitates the interconnection of devices and the sharing of data both within the data center and beyond. Networks tend to be designed in a hierarchical manner, with many devices (such as servers and storage devices in the case of a data center) connected to a switch that is connected to another switch at the next level of the hierarchy, and so on. Another common topology is a mesh configuration in which many peer network switches are connected to one another to form a single level of the hierarchy. We will consider three different levels of a network hierarchy: core, distribution and top of rack (TOR). More elaborate configurations will not be covered. The core network function can be thought of as the gateway through which all data entering and exiting the data center must pass. As such, the core network equipment is connected either directly or indirectly to every device in the data center. The core switch is also connected to a service provider, which is the “pipe” through which all data passes from the data center to the Internet. The distribution level of the network acts as an intermediate level between the core and edge levels of the network, and as such can offload some of the work the core network equipment needs to do.

Specifically, the distribution level is useful for passing data between machines inside the data center or aggregating ports to reduce the number of physical ports required at the core. The TOR network level consists of switches that connect directly to devices that are generating or consuming the data, and then pass the data up to the distribution or core level. A data center network implementation may have all three of these levels, or it may combine or eliminate some of the levels. In large data centers, the load on the networking equipment can be substantial, both in terms of port count and data throughput. The end-of-row (EOR) distribution equipment offloads the core equipment in both throughput and port count. The TOR edge networking equipment offloads the distribution equipment in both throughput and port count, in the same way as the distribution networking equipment offloads the core networking equipment. Switches enable communication between devices connected on a network. In the case of a

data center, servers and storage arrays are connected to multiple switches in a hierarchical manner.

The ASIC chip decides where that data need to go and sends it back out through the correct port. The central processing unit (CPU) controls both the PHY and the ASIC. The CPU can take data from the network, process it and then send it back out onto the network.

1.4. Components

1.4.1. *Central processing unit*

The processor, also referred to as the CPU, is one of the greatest sources of heat generation within a server. Aside from the basic processing of data and instructions to provide an output result, processors may also have many more features for managing data and power throughout a system. The processor die is generally housed in a package that includes a substrate (i.e. a small printed circuit board, or PCB, for bringing out signals) and a lid, as shown in Figure 2.1. The lid, or case, more evenly distributes heat to an attached cooling component such as a heat sink (air-cooled) or cold plate (liquid-cooled), as shown in Figure 2.1. In most cases, a socket is used to enable removal and replacement of the processor on the motherboard. Some lower powered processors are lidless using direct attachment of a heat sink on top of the die. In the volume server segment, limitations of low-cost thermal solutions and a greater focus on performance per watt have slowed the generational increases in thermal design power (TDP). The transition to multicore processors has maintained Moore's law (a doubling of the transistors on a chip every 18 months), an improvement even within this limited power envelope. However, recent trends, which include the integration of functions previously implemented in a chipset or external devices, greater numbers of high-performance I/O interfaces and memory channels, larger internal caches and incorporation of specialty processing units, are pushing power up despite significant improvements in power management features. Although integration of additional features into the processor provides an overall improvement in server efficiency, a greater percentage of the overall server power is concentrated in a smaller volume, creating new cooling challenges. Over the last several years, customer demand and regulatory requirements have driven vendors to increase server efficiency across their portfolios. Power management features within the processor have become prevalent, with increasing capability to optimize

performance within power, acoustical and thermal limits. Processor, subsystem and system level power-capping features are also expected to continue improving, enabling more options for system designers to tailor servers to the customer's power and cooling constraints. Microserver processors have appeared with power and die area is optimized for smaller, simpler and targeted/limited workloads. As this technology has evolved, it has improved its performance, capacity and reliability, availability, serviceability (RAS) features required for enterprise applications. These processors use fine-grain power optimizations developed for the mobile market and as such can be computationally efficient. Although the power of these individual microservers is quite low, system designers typically place many of them on a single board or in a single chassis (tens to hundreds). From a total power-per-unit-volume standpoint, aggregate power may be equivalent or higher than a traditional volume server but with a more even power distribution (versus concentration of power in traditional volume server CPU sockets). CPU power continues to slowly increase from generation to generation. Higher performance processors, typically used in high-performance computing applications, have TDP ranging from 100 to 200 W, with the prospect that over the next few years it will increase. For telecommunications and lower performance volume servers, the TDP has historically been in the 50–100 W range and will likely increase to the 150 W range over the same period of time. Historically, it has been difficult to extrapolate today's processor power into the future, and this remains the case today.

Table 1.1 presents a power summary of different Intel Xeon Scalable Skylake processors available in mid-2018.

CPU	Idle Power (W)	Max Power (W)
Intel Skylake CPU (205W)	16.2	205.0
Intel Skylake CPU (200W)	15.6	200.0
Intel Skylake CPU (165W)	19.0	165.0
Intel Skylake CPU (150W)	18.0	150.0
Intel Skylake CPU (145W)	18.0	145.0
Intel Skylake CPU (140W)	17.6	140.0
Intel Skylake CPU (135W)	17.9	135.0
Intel Skylake CPU (130W)	17.9	130.0
Intel Skylake CPU (125W)	18.1	125.0
Intel Skylake CPU (115W)	22.0	115.0
Intel Skylake CPU (105W)	16.8	105.0
Intel Skylake CPU (85W)	13.6	85.0
Intel Skylake CPU (75W)	13.7	75.0
Intel Skylake CPU (70W)	13.7	70.0

Table 1.1. Power summary of different Intel Xeon Scalable Skylake processors

1.4.2. Graphics processing unit

Graphics processing units (GPUs), also called general purpose GPUs (GPGPUs) in the industry, enable heterogeneous computing (multiple CPU types within one server) when installed into a server. A common example of this system is a server that has both GPUs, which have their own processor, and the server's CPU(s). This application is used for high-throughput computing with a more mainstream CPU for latency-sensitive computing. This model offers significant performance and capability improvements in computing while creating challenges in data center thermal management and planning.

Originally designed for computer graphics, GPUs are increasingly being used for other computational purposes that benefit from the massively parallel architectures of these devices. Today GPUs are designed to address parallel, high-throughput computing problems (e.g. rendering a screen's worth of data). This is similar to single instruction, multiple data (SIMD) vector processors used in old supercomputers. The aggregate compute capability of these small processors exceeds that of general-purpose processors. Memory bandwidth is also higher in GPUs compared to CPUs. Several companies have invested in this new paradigm, with GPU products currently available. Others are pursuing a somewhat different heterogeneous computing strategy, offering many integrated core (MIC) processors using a core architecture more similar to standard x86 CPUs. While both GPUs and MIC processors are similar in that they consist of a much higher number of smaller cores than in a typical CPU, there are also significant differences in their specific architecture and programming models. Each has its particular advantages and disadvantages and computational applicability, but these differences are beyond the scope of this book and will not be covered.

A downside to increased compute capability is a surge in power usage. GPUs and MICs that are relevant to servers typically draw between 75 W and 300 W, depending upon the application, with future solutions likely to draw greater than 400 W. While GPUs and MICs are usually more efficient in terms of performance per watt, the large increase in performance often results in server designs with different and larger power requirements compared to CPU-only nodes. Thus, while the overall server count and power draw of the solution may decrease, the power draw of individual servers and racks may increase, therefore making it more challenging to cool.

These GPUs can typically be deployed in a peripheral component interconnect express (PCIe) form factor; however, many manufacturers are developing products that integrate these processors on the main system board. They can have thermal solutions that are passive (no dedicated fans, instead relying on the base server air movers and heat sinks) or active (dedicated fans onboard the PCIe card itself). An active solution generally supports increased processor power and performance and has a lower platform fan power requirement for cooling than a passive solution. However, an active GPU/MIC thermal solution typically does not support fan redundancy, may increase airflow resistance and may not integrate as tightly with the host server base board thermal management strategy. Passive thermal solutions have been developed to support fan redundancy requirements and provide better integration with server management. Airflow and temperature requirements for passive solutions require custom shrouding/baffling to better manage airflow.

Table 1.2 presents a power summary of different GPUs available in mid-2018.

GPU PCIe Cards	Idle Power (W)	Max Power (W)
ThinkSystem NVIDIA Quadro P2000	20.0	75.0
ThinkSystem NVIDIA Tesla P100 PCIe 16GB	20.0	250.0
ThinkSystem NVIDIA Quadro P6000 PCIe Active GPU	20.0	250.0
ThinkSystem Intel Xeon Phi 7240P	20.0	275.0
NVIDIA Tesla M60 GPU, PCIe (passive)	20.0	300.0
NVIDIA Tesla M10	20.0	225.0
NVIDIA Tesla P40 24GB	20.0	250.0
NVIDIA Quadro P4000 8GB	11.1	105.0
NVIDIA Tesla P4 8GB PCIe Passive	20.0	75.0
NVIDIA Quadro P600 2GB PCIe Passive	4.4	40.0
NVIDIA Tesla V100 16GB PCIe Passive	20.0	250.0
NVIDIA Tesla V100 16GB FHHL PCIe	20.0	150.0
AMD Radeon Instinct MI25 16GB PCIe Passive	35.0	300.0

Table 1.2. *Power summary of different GPUs*

1.4.3. Volatile memory

Memory temporarily stores data that have been processed or are to be processed. Multiple dynamic random-access memory (DRAM) chips are packaged with many devices on one PCB. An example of form factor is the dual in-line memory module (DIMM). These PCBs have card edge connectors