Evaluating and modeling the energy impacts of data centers, in terms of hardware / software architecture and associated environment

Yewan Wang

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

THÈSE DE DOCTORAT DE

L'ÉCOLE NATIONALE SUPERIEURE MINES-TELECOME ATLANTIQUE
BRETAGNE PAYS DE LA LOIRE - IMT ATLANTIQUE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

Yewan WANG

Évaluation et modélisation de l'impact énergétique des centres de donnée en fonction de l'architecture matérielle/logicielle et de l'environnement associé

Thèse présentée et soutenue à Nantes, le 9 Mars 2020
Unité de recherche : Laboratoire des Sciences du Numérique de Nantes (LS2N)
Thèse N° : 2020IMTA0175

Rapporteurs avant soutenance :

Pr. Romain ROUVOY Professeur, Université de Lille
Pr. Noël DE PALMA Professeur, Université Grenoble Alpes

Composition du Jury :

Président :	Pr. Romain ROUVOY	Professeur, Université de Lille
Examinateurs :	Pr. Noël DE PALMA	Professeur, Université Grenoble Alpes
	Dr. Anne-Cécile ORGERIE	Chargé de recherche, CNRS/IRISA Rennes
	Dr HDR. Patricia STOLF	Maitre de conférences, Université Toulouse Jean-Jaurès
	Pr. Hamid GUALOUS	Professeur, Université de Caen Normandie
Dir. de thèse :	Pr. Jean-Marc MENAUD	Professeur, IMT Atlantique
Co-dir. de thèse :	Dr HDR. Stéphane LE MASSON	Responsable du département ICE, Orange Labs
Co-dir. de thèse :	Dr. David NÖRTERSHÄUSER	Orange Expert, Orange Labs

ACKNOWLEDGEMENT

First of all, I would like to express sincere gratitude to my thesis supervisor, Pr. Jean-Marc MENAUD, for the continuous support of the whole research process and related research. He taught me what scientific research looks like, and provided me a lot of opportunities in exchanging with the professors, Ph.D and postdoc students working in corresponding academic areas.

I would also like to express my deep appreciation to Dr. David NÖRTERSHÄUSER and Dr. Stéphane LE MASSON of Orange Labs as the co-supervisors of my thesis, for their expertise, enthusiastic encouragement, motivation and useful suggestions during the research work. I'm grateful that they provided me the opportunity to join the GDM/ICE team as Ph.D student. The research will not be realized without the financial support given by both Orange Labs and university IMT Atlantique.

I would like to express my very great appreciation to all the members of the team ICE in Orange Labs, specially: Dominique BODÉRÉ, Bertrand LE LAMER, Alain RINGNET, Olivier FOUCAULT, Jacky GAUTIER and Pascal BODIOU. As a foreign student, they helped me a lot in my integration of work and life in France. In particular, I would like to thank M. Philippe LEVASSEUR, for his kindness and valuable technical support on my experimental work. Moreover, I won't forget my friends for all the good moments of exchange: Kanza SALALIQI, Simon RICORDEAU, Chafika YAHIA CHERIF, Antoine DONALIES, Paul Arnaud and the others. I was inspired as well from the exchanges with my colleagues at the other department: Roland PICARD, Benoit HERARD and Joel PENHOAT.

My sincere thanks also goes to the postdoc fellow at IMT Atlantique, Mr Jonathan PASTOR. I appreciate a lot his technical guidances, useful suggestions and support on the experiments of the cluster, and I really enjoyed the days that we work together. In addition I would like to express my gratitude to Dr. Anne-Cécile ORGERIE and Dr. Patricia STOLF as the CSI members of my thesis.

I would also like to extend my thanks to my jury members: Pr. Romain ROUVOY, Pr. Noël DE PALMA et Pr. Hamid GUALOUS.

Last but not the least, I would like to thank my parents and companion Yiru, their love and understating support me spiritually throughout the whole research time.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOMENCLATURES & ABBREVIATIONS

Physics Constants

C	Capacitance	Farads (F)
C_p	Specific heat of air at constant pressure	$J/kg \cdot^\circ C$
C_x	Thermal capacitance of material x	$J/^\circ C$
f	Frequency	Hertz (Hz)
h_x	Thermal conductance of material x	$^\circ C/W$
P	Power	Watt (W)
R_x	Thermal resistance of material x	$W/^\circ C$
T	Temperature	$^\circ C$
t	Time	second (s), minute (m), hour (h)
V_{cc}	Supply Voltage of processor	Volt (V)

Abbreviations

AC	Alternating Current
ACPI	Advanced Configuration and Power Interface
ACV	Alternating current Voltage
ANN	Artificial Neural Network
API	Application Programming Interface
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning EngineerS
BMC	Baseboard Management Controller
CMOS	Complementary Metal Oxide Semiconductor
COP	Climat Change Conference
CPU	Central Processing Unit
CR	check point and restart
CRAH	Computer Room Air Handler
CV	Cross Validation
DC	Data Center
DC	Direct Current
DCiE	Data Center Infrastructure Efficiency
DDR	Double Data Rate
DIMM	Dual In-Line Memory Module
DMTF	Distributed Management Task Force
DPNM	Dynamic Power Node Manager
DRAM	Dynamic Random-Access Memory
DVFS	Dynamic voltage and frequency scaling
EC	European Commission

ECE	Energy and Carbon-Efficient
EPA	Environmental Protection Agency
EPOC	Energy Proportional and Opportunistic Computing
ESD	Energy Storage Device
FFD	First Fit Decreasing
FPU	Floating-Point Unit
GLB	Geographical Load Balancing
GOC	Green Open Cloud
GPU	Graphics Processing Unit
GUI	Graphical User Interface
HPC	High Performance Computing
HT	Hyper-Threading
HTTPS	Hypertext Transfer Protocol Secure
ICT	Information and communications technology
IEA	International Energy Agency
IHS	Integrated Heat Spreader
IoT	Internet of things
IPC	Instruction Per Cycle
IPMI	Intelligent Platform Management Interface
IT	Information Technology
ITN	Information Technologies & Networks
MAPE	Mean Absolut Percentage Error
MEPS	Minimum Efficiency Performance Standard
NIC	Network Interface Controller
ODE	Ordinary Differential Equation
OS	Operating System
PC	Personal Computer
PCS	Primary Cooling System
PDU	Power Distribution Unit
PM	Physical Machines
PMC	Performance counters
PPA	Power Purchase Agreement
PPW	Performance per Watt
PTDeamon	Power/Temperature Daemon
PTIM	Polymer Thermal Interface Material
PUE	Power Usage Effectiveness
QoS	Quality of Service
RAPL	Running Average Power Limits
RC	Capacitance-Resistance
REC	Renewable Energy Credit
SCS	Secondary Cooling System

SERT	Server Efficiency Rating Tool
SPEC	Standard Performance Evaluation Corporation
STIM	Solder Thermal Interface Material
SUT	Server Under Test
TCO	Total Cost of Ownership
TDP	Thermal Deign Power
TIM	Thermal Interface Material
UPS	Uninterruptible Power Supply
USB	Universal Serial Bus
VM	Vitrual Machine

INTRODUCTION

Context

Internet plays an important role in our daily life, it is accessible at any time once connected to the network. Internet eliminates the physical distance between individuals and makes the whole world accessible. More and more hardware and software resources are connected to the Internet and provide on-demand availability as services, the most important of them are known as cloud services. Nowadays, people are relying on the conveniences brought by cloud services, for both work and entertainment: social networking, real-time video streaming, on-line gaming, mail service, shopping, etc. Emerging services are making the cloud bigger than ever and the tendency is still going on. For example, in 2017, 70 017 hours of videos have been watched per minute by Netflix users, and this number jumped dramatically to 266 000 in 2018 [Des18]. Behind the Internet, all the cloud service are supported and provided by real and large scale physical infrastructure: Data center. Data Centers are large scale facilities composed of comprehensive number of mechanical and electrical infrastructures. They are the cornerstones of the Internet that support the digital life of everyone. Electricity is consumed by IT equipment such as servers and network devices in data centers to provide data processing, transaction, storage, etc. At the same time, due to the activities of the electrical devices, considerable amount of heat will be generated and increases the surrounding air temperature quickly. Cooling system is indispensable as well in a data center environment to keep appropriate operating temperature for these electrical devices, and it requires consuming additional energy. Nowadays, with the increasing demand of cloud services, data centers have became a huge energy consumer in the world. Reducing the energy consumption, especially the part from fossil resources has became a significant issue under discussion around the world. This thesis project focuses on evaluating the energy impacts from different part of a data center, including hardware, software and environment. Then propose a method to estimate the energy required by a physical infrastructure with corresponding configurations. In this section, we introduce the background of this research subject, specify the goals and detail the research plan.

Energy consumption of Data centers: some numbers

With the rapid growth of Internet services in recent years, the dramatic energy consumed by data centers has drawn much attention of data center owners and operators. For both economical and environmental concerns. Recently, more and more discussions care about the side effects on the environment brought by data centers. In terms of energy use, the data centers in the world are estimated to consume 1.4% of the global electricity consumption in 2017 [ABC17]. With the rise of data traffic, the percentage is estimated to reach 13% by 2030 [Sad17]. Power densities of a data center are 538-2153 $W \cdot m^{-2}$ and sometimes can be as high as 10 $kW \cdot m^{-2}$ [Bea13]. Today, data centers are estimated to consume 200 terawatt hours (TWh) each year, which is more than the need of some countries,such as Iran [Jon18]. In terms of CO_2 emission, total ICT (Information and Communication Technology) sector accounts for 2% of the global CO_2 emission, that significantly contributes to the greenhouse effect, and data center is believed to have the fastest growing carbon footprint among the whole ICT sector [WASM14].

Addressing the problem of high energy use, it is essential to have a preview of how data center consume energy. Figure 1 illustrates how energy flows in a typical data center facility. It can be seen from the figure that, for a data center, apart from the servers, power has to be delivered to other parts of the facility so as to keep normal functioning of the entire data center. The indicator Power Usage Effectiveness (PUE) [AAF12] is usually used to evaluate the general energy efficiency of a data center. PUE is defined as the total power entering a data center divided by the power used to run the computing equipment. An ideal PUE equals 1 which means all the power has been used by the computing side (for more details about PUE, refer to "Energy Efficiency metric" at section 1.1.2). Google claims to have achieved a comprehensive trailing twelve-month (TTM) Power Usage Effectiveness (PUE) of 1.12 across all their data centers, in all seasons, including all sources of overhead. Best PUE for individual campus can be as low as 1.09. Which means that over 90% of the energy, including electricity and natural gas, is consumed by computing equipment [Goo19]. While global average PUE of the other largest data centers in the world is around 1.67 [Mem19].

Research background

The evolution of the data center architecture is changing rapidly. New technologies are emerging to enrich IT services, such as 5G, IoT (Internet of things), BigData, etc. The evolution of IT services obligates in the same time the revolution from the aspects of both hardware and

Figure 1 – Power Usage in a Data Center [CK16]

software. Data center designers often encounter the problems of selecting or replacing appropriate equipment for emerging new services. An ideal design should ensure the required Quality of Service (QoS), while minimizing the Total Cost of Ownership (TCO), with respect to the Equipment Life Cycle. The explanations of the aspects are detailed as follows:

— Quality of Service (QoS): Cloud services come in with various forms: voice, video, data and signaling. The single Internet platform should be able to support all type of communication forms on demand, and meet up with desired requirements for running services, such as: expected delay, response time, Bandwidth, loss and error rates, etc [CFY04].Full capacity is expected meet up with required QoS during peak demand period.

— Equipment Life Cycle [GC15]: Electrical equipment has limited lifetime, depend on manufacturing and usage. Taking servers for example, the general server lifetime is about 3-5 years, depending on how it has been used. Sometimes the lifetime can be extended to decades by doing periodic upgrades and component replacements. However, server will likely not remain cost-effective for decades. When the maintaining cost exceed the replace cost, it is usually the time to buy a new one. Table 1 lists the general lifetime for common equipment in a data center.

— Total Cost of Ownership (TCO): TCO for a data center includes total cost of investment and operating costs as well costs related to replacement or upgrades at the end of the life cycle. Major TCO costs consists of five parts: infrastructure, server acquisition, power utilization, networking equipment, and maintenance cost [CIGH17]. TCO varies with physical configurations like the number and type of server, hardware implemented and the software service applied.

Estimating the energy impacts and evaluating the performance contributions of principal parts of the data center allow us understanding the energy and performance trade-off, identifying the low energy-efficiency parts, and finally help us to focus the efforts on specific areas and optimize the global energy efficiency effectively.

As the leader provider of telecommunication in Europe, Orange proposes much kind of telecommunication services across Europe and Africa. Consisting sustainable development and being a responsible operator is always an important concern for Orange. Particularly, in 2007, Orange confirmed its participation in 21^{st} Climate Change Conference (COP21) and made a promise that their emissions of CO_2 will be reduced 20%, and energy consumption will be reduced 15% from 2016 to 2020 [Gro16]. Orange commits also to reduce 50% its CO_2 emissions per customer-usage in 2020 (base line 2006). Therefore, this thesis research project has been

raised as part of the "Green ITN" plan [Gro19], with expectations to help developing an energetic criterion that evaluates the energy efficiency along with contribution of service for data centers of newer generation, so as to minimize energy consumption while meeting up with the required QoS. The goal is planned to be realized by proposing a model that estimates both the energy consumption and service contribution of principal parts of the data centers, including servers and cooling system. Such a model can be very helpful for Orange in selecting appropriate IT equipment and optimizing hardware, software and environmental configurations in terms of designing, re-scaling, updating and renewing their data centers.

Table 1 – General Lifetime of main components in a Data center [Mic17]

Equipment/Services	Lifetime (estimation, years)
Servers	$3 \sim 5$
Firewall	$5 \sim 8$
Switches	$5 \sim 8$
Wireless Access Points	$5 \sim 8$
Uninterruptible Power Supplies(UPS) Devices	$4 \sim 6$

Research problem and goal

The evaluation of energy efficiency of different sub-systems in a data center is a key point in optimizing hardware/software selection and environmental conditions. The subsystems vary from single component like the processors, to a node then to an entire cluster. Over provisioning and inefficient usage of resources result in the waste in Data centers. For example, previous research has highlighted the huge energy waste led by running too many idle servers in data center (idle state: only OS is running, without executing any workload) [MCRS05]. Building energy efficient data centers brings both economic and environmental benefits. Therefore, it could be favorable to understand how energy has been consumed and how external aspects such as environmental conditions and equipment related conditions interfere with the global power consumption. Developing a power consumption predictive model is one of the solutions. Ideally, such a model could be able to predict the power consumption of a system with acceptable error rate, by taking various system parameters into consideration, such as system load, environment conditions (i.e. ambient temperature and space area), system configurations, etc.

Some parameters are known to be correlated with global power consumption (system load), and some are not (equipment models, equipment arrangement, thermal effects, manufacturing related features, etc.). There is a lack of study addressing identifying and characterizing the underlying aspects that may have influence on global power consumption. Therefore, in the first place, an overall evaluation will be performed to locate and quantify these potential aspects. The large diversity and complex hardware/software configurations in a data center environment make a great challenge for this research. Later on, based on the previous results, a global power model is expected to be established to estimate the energy consumption of a physical infrastructure, for both server and cooling parts. The main goal of this thesis is to propose a model to estimate the energy consumption of a computing system. Such model is expected to estimate global energy consumption by providing with the important system related information, such as IT load, hardware and software architectures, environmental conditions, etc. This model could help evaluating the energy efficiency of critical parts in the whole infrastructure, optimizing the configurations for current infrastructure and orienting the decisions on equipment selection and renewal for data centers in the next generation.

In order to achieve the objective, the research in this thesis consists of several consecutive steps, they are summarized as follows:

— Exploring a dedicate state of art study to obtain a systematic survey on the up to date approaches and technologies.

— Performing an overall evaluation on the potential effects that may affect power consumption in computing systems. Evaluating as well the accuracy, reliability for current power metering and modeling approaches.

— Proposing a modeling approach based on previous studies to estimate energy consumption of computing systems. Estimation is expected to include both the consumption from servers and cooling system.

— Validating this modeling approach with real computing systems (in the progress).

Contributions

The principal purpose of this research is to realize a global power consumption model of a comprehensive computing system, which allows us to understand the contributions of different parts, such as system variables and environmental conditions to the global power consumption of the system. So as to facilitate a global optimization. The research begins from a series of experimental evaluations, in order to identify and quantify the potential influential aspects, that

should be taken into account in a power model for a physical computing system. The evaluations have been especially concentrated on some defects and uncertainties appeared in devices, servers and environment several. In the end, we propose a global power consumption modeling approach for a physical cluster, with expectations to take into account important variables in a computing system, such as IT load, environmental temperature, cooling system configurations, etc. The major contributions presented in this thesis are detailed as follows:

1. Before implementing a mathematical power model to a physical computing system, it is substantial to recognize every possible consumer and underlying sources of influence appeared in the system. Therefore, we have firstly performed an overall identification on potential aspects that may affect the global power consumption of a computing system, especially for the parts other than IT load. The computing systems scale from entire cluster to single processor. Several underlying influential candidates that may brought by infrastructure deployment or environment have been explored experimentally, such as construction difference from manufacturing, arrangement of servers in the racks, fluctuating ambient temperatures and voltage variation from power supply. The experiment results turn out that, besides IT load, the power consumption of servers can be obviously varied by two aspects: construction difference between identical servers and temperature variation of environment. Later on, we have focused firstly on the thermal effect and the construction difference will be discussed later in the next part. Especially, we design tests to evaluate and characterize the influence of temperature variation of CPU and of the other components to the power.

2. Secondly, we complete the evaluation on the differences between identical processors caused by imperfection fabrication, as respond to the question left in the first study. We concentrate the attention on processor, which is supposed to be the most consuming component in a server. This manufacturing variability has been observed in previous studies, however, the amount of samples is not adequate to determine the source of the variability, especially from the perspective of thermal characteristics. In this study, we compare the power consumption of two types of Intel processors from different generations, by running the samples at full load and under the same environment conditions. Processors from modern generation appear more power variation between identical samples than the older one. Two hypothesis have been proposed and studied experimentally: the application of TIM (thermal interface material), and parameters of leak current. The observations and results drawn from this research highlight the variability between identical processor samples, they provide solid evidences to help understanding the imper-

fections caused by fabrication processing. The findings remind us of re-evaluating the accuracy of current power models.

3. In addition, we investigate experimentally as well two indispensable parts in terms of building power model for physical computing systems: power characterization approaches and model type. In the beginning, we present a deep evaluation about the power models based on CPU utilization. It is a single indicator power model, which is widely adopted thanks to its convenience. However, according to the analysis, simple linear regression is not sufficient to build reliable power models. The influence of inlet temperature on accuracy of model has been especially explored. Based on the experimental data, we propose a new model to compensate the additional power caused by the rise of ambient temperature. Besides, we evaluate other two possible ways to improve the accuracy of this kind of model: making use of operating frequency information, and applying polynomial regression. Later on, we have investigated also the accuracy and reliability of measurement values obtained from the following power metering approaches: IPMI, Redfish and Intelligent PDU. These approaches provide an alternative and economic ways to get power data in data center environment. Comparing to traditional power meters, they take advantages of the integrated sensors in servers and PDU to get directly the power data. However, few work addressing their accuracy and reliability in real utilization. In order to compensate this missing part, we compared the values obtained from them with a high-accuracy power analyzer, the evaluation has been done for different infrastructures. We concludes the validation results with expectations that the findings could provide more details and guidelines for data center operators, when adopting these kinds of tool as power metering approaches.

4. In the end, we propose an approach to estimate the global power consumption of a physical cluster. The cluster consists of 48 identical servers in four racks and a cooling system. Especially, two models have been realized and validated with real measurements: one model simulates the thermal system of the cluster and an other model allows estimating the cooling power consumption based on inlet temperature. In the first model, the thermal system has been simplified to an equivalent capacitance-resistance (RC) electrical circuit. And the real-time model has been established by using Ordinary Differential Equations (ODE), with respect to the law of heat balance during heat transfer processing. In order to identify the parameters of model, which are combined with complicated physical and thermal properties, we conducted a global optimization, by minimizing the results of the model with the real measurements. The model is able to predict inlet tem-

perature according to total power of servers, cooling power and environment temperature. Further more, we have proposed another model for estimating the real time cooling power consumption based on inlet temperature. Four critical parameters determined the functioning of cooling system have been identified, based on the behaviors and settings of cooling system. A final model is expected to realize a global power consumption estimation of the cluster, which includes both power consumed by cooling system and servers. The model should be able to predict the global power consumption based on IT load of servers, settings of cooling systems and environmental variable. Due to limited time, we are still working on this model, the further work has been well detailed in this study.

Publications

Most contributions mentioned before have been redacted in papers and the work have been published and presented at different national and international conferences and workshops. The last contribution is in the process of redaction and will be submitted later at a corresponding conference or journal.

National conferences and workshops

— Yewan Wang, David Nörtershäuser, Stéphane Le Masson, Jean-Marc Menaud. Etude de l'influence des aspects thermiques sur la consommation et l'efficacité énergétique des serveurs. SFT 2018 - 26ème Congrès Français de Thermique, May 2018, Pau, France. pp.1-8.

— Yewan Wang, David Nörtershäuser, Stéphane Le Masson, Jean-Marc Menaud. Etude de l'influence de la température du processeur sur la consommation des serveurs. Compas2018, Conférence d'informatique en Parallélisme, Architecture et Système, Toulouse, France.

International conferences

— Yewan Wang, David Nörtershäuser, Stéphane Le Masson, Jean-Marc Menaud. Potential effects on server power metering and modeling. CloudComp2018, 8th EAI International Conference on Cloud Computing, Sep 2018, Newcastle, Great Britain.

— Yewan Wang, David Nörtershäuser, Stéphane Le Masson, Jean-Marc Menaud. An empirical study of power characterization approaches for servers, ENERGY2019, 9th IARIA

International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies, June 2-6, 2019, Athens, Greece.

— Yewan Wang, David Nörtershäuser, Stéphane Le Masson, Jean-Marc Menaud. Experimental Characterization of Variation in Power Consumption for Processors of Different generations, July 14-17, GreenCom2019, 15th IEEE International Conference on Green Computing and Communications Atlanta, USA

International journal

— Yewan Wang, David Nörtershäuser, Stéphane Le Masson, Jean-Marc Menaud. Potential effects on server power metering and modeling, Wireless Network (2018), page 1-8, Springer, https://doi.org/10.1007/s11276-018-1882-1

Organization of the Manuscript

The rest of this thesis is structured as follows:

— In Chapter 1, a survey of the state-of-the-art is presented. The survey present the previous studies related to this research in several domains: metrics concerning energy efficiency evaluations for servers and data centers; current examples in using carbon free energy in data center; hardware and software designs for achieving energy-efficient designs for small and large computing design; instrumental and modeling approaches in getting power measurement of servers, their advantages, accuracy and reliability, etc.

— In Chapter 2, we identify and quantify several underlying external aspects that may have influence to the power consumption of a physical cluster. The aspects evaluated include: construction difference, position and arrangement of servers, voltage variation and temperature of components on motherboard. Power consumption variations brought by thermal effects have been specially evaluated on servers of different brands and generations.

— In Chapter 3, we compare the power consumption among identical processors for two Intel processors from different generations. The observed power variation of the processors in newer generation is much greater than the older one. Then, we propose and evaluate hypotheses on the underlying causes in dedicated experiments by precisely controlling environmental conditions.

— In Chapter 4, we compare and discuss the reliability, advantages and limitations in terms of power characterization solutions for servers. We evaluate firstly CPU-utilization based

power models. The findings highlight the challenges in realizing accurate and reliable power models. Later on, we extend the evaluation to the power metering tools: IPMI, RedFish and Intelligent PDU, for the purpose of providing guidelines in adopting these tools in real data centers.

— In Chapter 5, we present our work about realizing a global power consumption model for a physical cluster. We have firstly proposed an approach to simulate the real-time thermal system in using equivalent RC electrical circuit. Then a second model is realized to estimate the cooling power consumption according to inlet temperature.

— In Conclusion, we conclude the work presented at the manuscript and present the perspectives for the further work.

CHAPTER 1

STATE OF THE ART

The state of the art study contains related studies to our research interests. We review firstly different metrics that present the energy efficiency of servers and data centers. Then we decribe some classical and recent studies addressing building a "green" data center, such as replacing fossil sources with carbon-free and clean energy sources, adopting energy-efficient designs on hardware and software solutions of computing systems. The designs presented scale from single component to large scale data centers. For modeling, we need firstly measuring. We evaluate three kinds of power meters: internal power meters, external power meters and embedded power meters, by discussing and comparing the accuracy, availability and reliability. Moreover, we summarize the studies concerning building power models for servers in two ways: using resource usage or counters. We highlight further on the challenges in terms of building accurate power models. Finally, we remind the influence of user behavior in this domain.

1.1 Energy Efficiency evaluation

In this section, we introduce the history and evolution of the theory of Moore, which highlights the importance of evaluating the energy efficiency of modern computing systems. There are different metrics and methods in interpreting the energy efficiency for servers and data centers, several popular ones are presented in this survey.

1.1.1 Power Wall

Moore's Law is not a physical or natural law but an observation that the number of transistors in a dense integrated circuit doubles approximately every two years. The observation is named after Gordon E. Moore, the co-founder of Intel and Fairchild Semiconductor. His prediction proved accurate for several decades, as shown by figure 1.1 and the law was used in the semiconductor industry to guide long-term planning and to set targets for research and development.

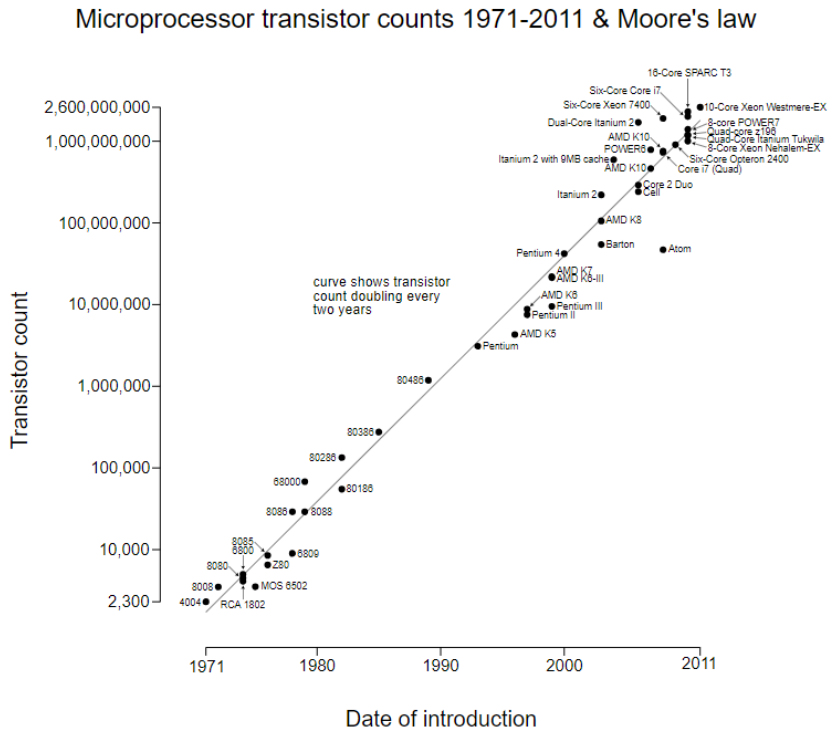


Figure 1.1 – Microprocessor Transistor counts & Moore's Law [Com18]

However, in recent years, some researches show that the growth in Moore's Law may slow even go to an end. For example, the international Technology Roadmap for Semiconductors of 2010, predicted that the growth would slow around 2013, and Gordon Moore in 2015 foresaw that the rate of progress would reach saturation. Most semiconductor industry forecaster, including Gordon Moore, expects Moore's Law will end by around 2025. The increasing number of transistor in one chip also causes some negative impacts. For example, a roughly 45% increase in processor transistors has translated to roughly 10-20% increase in processing power. Dennard scaling [DGR+74] proposed by Robert H. Dennard and his team in 1974 is a scaling law. They defined scaling formula for scaling circuit parameters and dimensions based on transistor size. With the increase number of transistor, the scaling factor reduces the area of circuit parameters and dimensions which contribute to compensate the power rise caused by transistor density [BC11]. By following the scaling formula, transistor density can continue Moore's Law while keeping the power dissipation unchanged. Dennard scaling suggests that performance per watt would grow at roughly the same rate as transistor density, doubling every 1-2 years. This law has been then widely adopted by the semiconductor industry as the roadmap for setting

targets and expectations for coming generations of process technology. However, since around 2005-2007, Dennard scaling appears to have a broken down, while Moore's Law continued on [Boh07]. The primary reason behind the breakdown is that at small sizes, current leakage becomes greater and heats up the chip, which eventually increase further the static power. Another problem is that the voltage can on longer be scaled down as required by Dennard scaling, the voltage scaling has reached the lower limits imposed by threshold voltage (VT) scaling limits for security and reliability requirements [Tau02].

Therefore, density and number of transistors in chip can no longer be an effective indicator to represent the energy efficiency for processors, as also for modern computing systems. Instead, corresponding metrics have been proposed to evaluate the energy efficiency for recent computing systems, they will be discussed for the rest of this section

1.1.2 Energy Efficiency metrics

As discussed in the previous section, transistor density can no longer determine the efficiency of a system. Several metrics have been proposed and applied in the literature to evaluate energy efficiency of data centers. In this section, we detail some energy efficiency metrics for evaluating green computing systems, the computing systems under evaluation including single server and data centers.

Power Usage Effectiveness (PUE) The most adopted one is Power Usage Effectiveness (PUE), introduced by the Green Grid [AAF12]. The metric represents for the radio of total energy consumed by data center to the part delivered to IT equipment. PUE is defined as equation 1.1 below:

$$PUE = \frac{Total\ Facility\ Energy}{IT\ Equipment\ Energy} \tag{1.1}$$

An ideal PUE equals 1, which means every watt of power is contributed to operate IT equipment to do "useful computing work". However, value of 1 is unrealistic to achieve as energy is required at support side to provide the normal operational conditions of data centers, such as power supply equipment, cooling system, lighting, etc. Recently, thanks to the continues improvements achieved for cooling systems (better air flow design, efficient cooling equipment, advanced cooling technologies, etc) and greater adoption of renewable energy, large-scale data centers are beginning to have PUE value of 1.1 or less, while small scale data centers still have PUE values greater than 2.0 [SSS+16]. Determining PUE requires a period of time as it measures energy. Some studies emphasize that relevant data should be collected at least one year

to get a reasonable PUE value [YM13] [BKST13]. Because PUE can be influenced by outside temperature and IT load variation. Even though the PUE value is widely used in the industry, there still remains some limitations. For instance, the method of quantifying the PUE hasn't be unified. PUE is normally calculated and reported by industry operators with their own method, that leaves room to "improve" the final result [Goo19].

Data Center Infrastructure Efficiency (DCiE) Same as PUE, DCiE was also proposed by GreenGrid as a measurement to determine energy efficiency of a data center. It is expressed as the percentage of power consumed by IT equipment power in total facility power. Therefore, DCIE is actually the reciprocal of PUE, which has a value of $1/PUE$. DCiE is defined as equation 1.2 below:

$$DCiE = \frac{IT\ Equipment\ Energy}{Total\ Facility\ Energy} \times 100\% \tag{1.2}$$

Performance per Watt (PPW) Green500 project [FC07] proposes metric PPW to rank supercomputers efficiency based on power requirement and performance achievement. The performance is determined as Giga Floating-point Operations Per Second, by running Linpack benchmark on the system. And the power is determined as the average power during the entire execution of running Linpack [DBMS0s]. The Performance (in operations/second) per watt adopted by Green500 to rank the supercomputer can therefore be expressed as $gigaflops/watt$. For example, as we write right now (November 2019), the lastest first record has achieved 16.9 gigaflops/watt, reported by A64FX prototype supercomputer [Top19]. Green500 aims to encourage long-term sustainability design for high-end supercomputers.

Server Efficiency Rating Tool (SERT) SERT is an industrial standard rating tool for evaluating energy-efficiency for server systems, developed by Standard Performance Evaluation Corporation (SPEC) committee. It measures and analyze the energy efficiency of servers by executing the SERT Test Suit to the Server Under Test (SUT), while measuring the power consumption data in the same time. SERT 2 is the new version of SERT, it proposes the SERT 2 metric, also called the SERT 2 Efficiency Score to represent the energy efficiency of a whole server system with one single number. The SERT 2 metric number can be used to compare the energy efficiency between servers across different brands and scales. SERT 2 has been now accepted officially by U.S.Environmental Protection Agency (EPA) for server energy efficiency labeling with Energy Star certification [(EP19]. European Commission (EC) also adopts SERT

test results as a software requirement for evaluating a potential environment friendly server system design [COM19]. EC proposes Minimum Efficiency Performance Standard (MEPS) based on SERT metric results to set minimum limits for servers in order to filter out the worst performing servers from the market.

SERT 2 metric score can be obtained by executing the complete SERT 2 test suit on the server system by following strict the environment conditions and power measurement device requirements [Com13], the test environment and key components are detailed as follows:

— `Test enviroment`: The test environment is composed of multiple hardware and software components, the simplified system diagram is shown by figure 1.2.

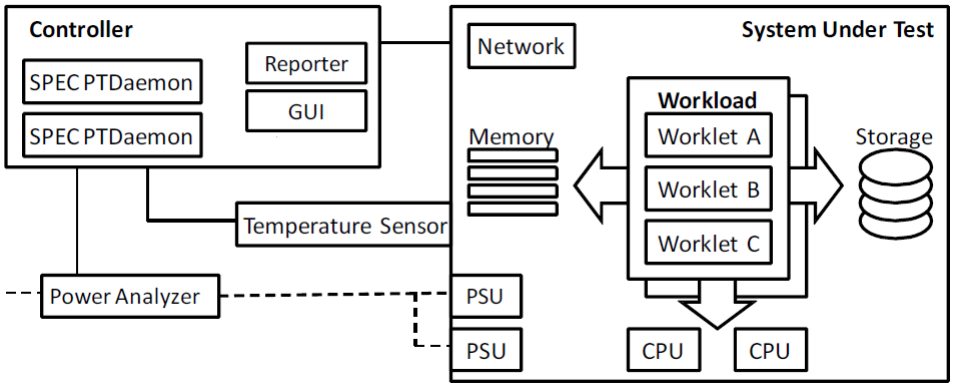


Figure 1.2 – SERT system diagram ©1995 - 2020 Standard Performance Evaluation Corporation (SPEC). All rights reserved.

SERT test environment consists of two independent systems: the system under test, the computing system we want to evaluate the energy efficiency; And a controller system. Controller is operated by the user and composed of several software components:

GUI: configuring the SUT and executing the SERT test suit through user interface (GUI);

SPECPTDeam: communicating with the measurement devices;

Reporters: analyzing performance and power consumption data and generating human-readable test reports for further analysis.

— `SERT Test Suite`: Complete SERT 2 test suit is composed by several workloads targeting at stressing main components of a server: CPU, memory and storage. Each workload includes several mini-workloads called `worklet` with a sequence of different load levels, in order to perform an overall evaluation from different aspects. The details of each workload is shown in the figure 1.3. At the beginning of the execution

of a worklet, program will perform a "calibration" process to determine the maximum throughput of the server as the highest load level, the other load levels are determined based on calibration result as well.

Workload	Load Level	Worklet Name
CPU	100%, 75%, 50%, 25%	Compress
		CryptoAES
		LU
		SHA256
		SOR
		SORT
		XMLValidate
Memory	Flood: Full, Half	Flood
	Capacity: 4GB, 8GB, 16GB, 128GB, 256GB, 512GB, 1024GB	Capacity
Storage	100%, 50%	Random
		Sequential
Hybrid	100%, 87.5%, 75%, 62.5%, 50%, 37.5%,25%, 12.5%	SSJ
Idle	idle	Idle

Figure 1.3 – SERT 2 Test Suite ©1995 - 2020 Standard Performance Evaluation Corporation (SPEC). All rights reserved.

— `SPEC's Power/Temperature Daemon (PTDeamon)` In terms of getting real-time energy consumption and ambiant temperature data during the test, SERT 2 is able to work along with several high-accurate power analyzers and temperature sensors in the market. While executing the test suit, instant power and temperature data will be recorded in the meanwhile by PTDeamon. PTDeamon is a software component installed on the controller system, provided by SERT. It can communicate with the power analyzers and temperature sensors with multiple functions, such as set up configuration requirements directly to the devices, get back instant power/temperature data. For attention, only the devices in the "accepted devices list" [Cor] are able to work together with SERT.

— `SERT metric score` The calculation of final SERT metric score consists of the following steps, as also illustrated in figure 1.4:

1. Calculating per load level energy efficiency Eff_{load} for each load level during a worklet execution as follows:

$$Eff_{load} = \frac{NormalizedPerformance}{PowerConsumption} \tag{1.3}$$

All worklets (except idle) run at multiple load levels. In this equation, normalized

performance refers to normalized throughput, power consumption refers to in this context the average measured power consumption for each load level.

2. Calculating worklet efficiency cores $Eff_{worklet}$ in using the geometric mean of each load level scores, as follows:

$$Eff_{worklet} = exp(\frac{1}{n} \times \sum_{i=1}^{n} ln(Eff_{load_i})) \times 1000 \tag{1.4}$$

Where n represents the number of load level available for each worklet.

3. Calculating Workload Efficiency $Eff_{workload}$ in using the geometric mean of all worklets within the workload, as follows:

$$Eff_{workload} = exp(\frac{1}{n} \times \sum_{i=1}^{n} ln(Eff_{worklet_i})) \tag{1.5}$$

Where n represents the number of worklets available in a workload.

4. Calculating SERT 2 Metric score in using a weighted geometric mean, particular workload weight is determined by expert groups of SPEC: 65% for CPU workload, 30% for Memory workload and 5% for storage workload. The final score is calculated as follows:

$$Eff_{Score} = exp(0.65ln(Eff_{CPU}) + 0.3ln(Eff_{Memory}) + 0.05ln(Eff_{Storage})) \tag{1.6}$$

However, the cost of running a compliance SERT test is not free, except labouring fee, investment is also required for purchasing both equipment (accepted power analyzer and temperature sensors [(SP19]) and software license (2450 € for SERT software license). EC estimates in the report [COM19] that: the total compliance cost would be approximate 21000 € for companies with 15 server models and 30000 € with 25 server models.

With the Dennard Scaling coming to the end, the energy efficiency of a computing system can no longer be achieved easily by increasing the number of transistor in a chip, there are still lots of other ways to realize a "green" data center, for example by using as much as possible the green energy, adopting energy efficiency hardware and software designs on infrastructures, etc. We are going to detail these solutions explored in previous studies in the next section. Nevertheless, SERT metric puts forward a great step to the development of a worldwide agreed standard of energy efficiency for computing systems, like servers and clusters. In this thesis research, for most of the experiments, we utilise the workloads in test suite SERT to benchmark

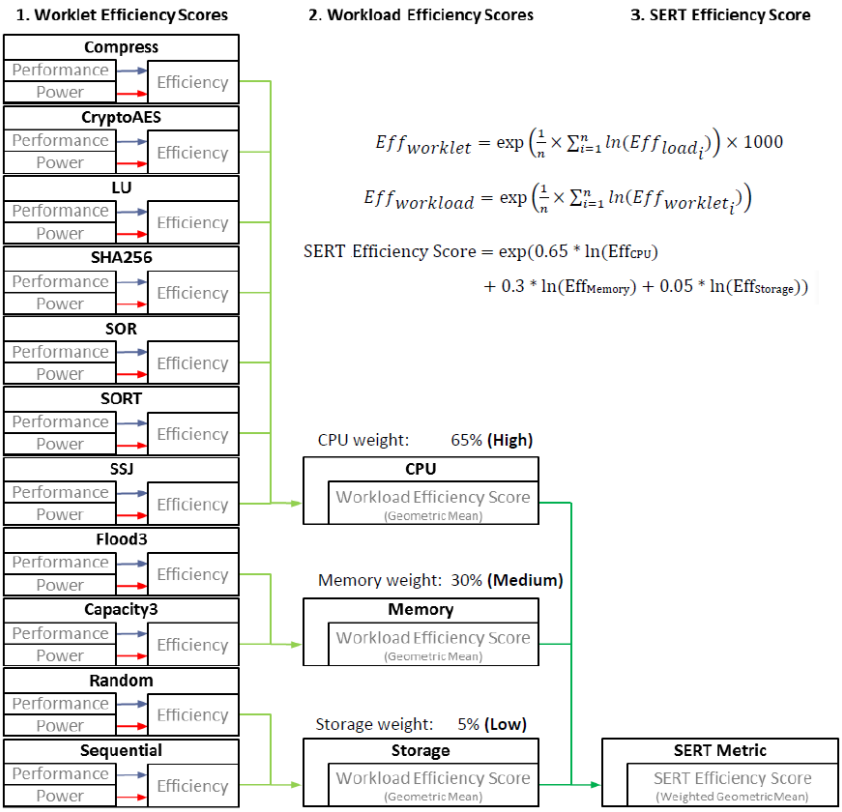


Figure 1.4 – SERT 2 metric calculation [vKLA+17] ©1995 - 2020 Standard Performance Evaluation Corporation (SPEC). All rights reserved.

the performance and energy efficiency of our SUTs.

1.2 Making the data center Greener

Lawrence Berkeley National Laboratory draft the report "United States Data Center Energy Usage Report" in 2016 [SSS+16]. The report estimates the total electricity consumed by data centers in the US from 2000 to 2014, and forecasts the consumption up to 2020, based on the previous studies and historical data provided by several institutions. Current results show that, electricity consumed is estimated to be increased by about 4% from 2010-2014. The same estimated values from 2010-2014 and 2000-2010 are 24% and 90% respectively. Data centers in the US are estimated to consume about 70 billion kWh in 2014. Based on current trend estimates, energy use is expected to rise slightly and reach approximately 73 billion kWh in 2020. Moreover, the reporters also point out that, the improvements at energy efficiency didn't have negative impacts on the normal progress in terms of performances. As conclusion, since 2010, the increase trend of electric demand for data centers in US has been dramatically slowed down since 2010, thanks to the improvements at energy efficiency, regardless of the dramatic increase in data center traffic and demand of services. Another report from International Energy Agency (IEA) [Kam19] found the same tendency, huge strides have been achieved in improving energy efficiency of data centers. The global energy demand of data centers in the world, is expected to slightly decrease from 198 Twh in 2018 to 191 Twh in 2021 . These previous studies confirm the effectiveness of the continuous efforts contributed to improve the energy efficiency of data centers.

In this section, we provide an overview about the recent efforts and studies proposed in several domains for the purpose saving energy and reducing carbon emission of data centers. The state of the art study will be presented by the following perspectives: we introduce firstly the methods concerning the exploitation of renewable energy to power on the data centers in the section 1.2.1. Including both the real use cases which are already deployed by giant companies and several scientific projects in the progress. After that, we summarize relevant researches aim at improving energy efficiency of computing systems through power proportional designs, including the hardware solutions in section 1.2.2 and the software solutions in section 1.2.3. In the last section1.2.4, we present some advanced power management solutions for large-scale data center, including some new propositions which provide possible directions in the perspective of reducing energy consumption for data centers in the near future.

1.2.1 Exploiting carbon-free energy sources

Recently, facing and dealing with climate change become an urgent global priority. In order to minimize environmental footprint, some data center operators are seeking for powering the facilities with carbon-free "green" energy instead of traditional fossil fuel "brown" resources such as coal and natural gas. The attempts in exploring green energy are not for the purpose of reducing the energy consumption of data centers but looking for opportunities to maximize the portion of green energy usage in such big power consumer facilities, so as to reduce the environmental impacts.

One big barrier nowadays in using renewable energy is due to its higher price comparing to fossil fuel energy, which is not commercial effective for many small and medium size industries. In earlier stage, government policies, social responsibility of industry are indispensable for the development and commercialization of renewable energy. Facebook, Microsoft and Google are a few of the technology companies that have committed to go '100% renewable' through the RE100 [Org14], with commitment to source 100% of their global electricity consumption from renewable sources by a specified year. Thanks to the continuous efforts, development and financial investment over decades, renewable energy remains its healthy growth and have become a mainstream source of affordable electricity for millions of people [REN18]. From the perspective of economics, the cost of renewable power has dropped dramatically. Over the past eight years, levelized costs for wind and solar energy have decreased by 67% and 86% [Laz17], respectively. According to a recent study from Bloomberg New Energy Finance [SH18], by 2050 half of the world's electricity will be generated from wind and solar. From the perspective of marketing, according to the estimation from International Energy Agency (IEA) [FB18], in the next five years, renewable energy will have the fastest growth in electricity section. It is estimated to meet up with over 70% if the global energy consumption growth and provide almost 30% of power demand in 2030. [FB18]

In 2018, Google claims to have achieved a great goal in 2017: they have purchased 100% wind and solar energy to match consumption for both data centers and offices operations [RP18]. Except powering some parts of the Data center by in-site renewable energy [GKL$^+$13], this achievement is completed by buying renewable electricity directly from a particular renewable energy providers in the form of a power purchase agreement (PPA). PPA is long-term contract to purchase power at a fixed and negotiated price from a renewable energy producer, the contract lasts usually for decades. PPA helps renewable energy providers getting solid and stable financial commitment that they need to develop new clean energy facilities. In exchange, PPA buyers will be rewarded with Renewable Energy Credits (RECs), and one REC represents for

one MWh of renewable power [Goo13]. However, data center operates 24/7 but most renewable energy sources don't yet. In addition, for some complicated reasons such as environment and politics, renewable energy is not available to customers in some areas around the world. In this case, part or whole of Google's data centers still need to connect to traditional grid to get constant and stable power, which is from fossil fuel resources. Therefore, Google purchase extra RECs to compensate "no-green" part and meet their 100% goal [Goo16].

EpoCloud project [BFG+17] EPOC refers to Energy Proportional and Opportunistic Computing systems. This project focuses on an optimal power management solution for a mono-site and small scale data center, which has all the resources in one location and powered on by both regular electricity from grid and local renewable energy sources (windmill or solar cells). In EPOC, each job is executed in a dedicated Virtual Machine (VM) in order to facilitate the energy-aware resource allocation. Usual tasks which need continuous computing resource like web servers are powered on by electricity from regular grid. For the less urgent tasks that can be delayed and interrupted with a deadline constraint, EPOC takes advantage of the local renewable resources to perform opportunistic tasks. They adopt the concept "green energy virtualization" [HKLP17] in order to avoid using energy storage in the small scale data center. Renewable energy will be used once available. In terms of virtualization, when there is excessive green energy than demand, they use the surplus part to make up for the degraded interval with less green energy. Therefore, the supply of "green energy" appears to ideally meet the demand. In this way, not a watt of green energy will be wasted and there is no need for energy storage. Their solution is shown to be able to reduce "grown energy" consumption by 45% and double the "green energy" utilization compared to the baseline algorithm. As part of EpoCloud project, `Seduce` platform [PM18] is developed to provide the scientific testbed for studying power and thermal management aspects in a data center. `Seduce` enables real-time access to more than 200 measurements around a physical cluster, such as power consumption of servers and cooling system, temperatures at different positions of the cluster (at the front and back of each server and the cooling system, of the room), configurations and working states of cooling system, etc. The data is available through a user friendly API to the public. Now, the developers are working on integrating the renewable energy data to the platform, such as real-time information from solar panels and batteries.

DATAZERO projet [CRGRS18], short for "Data centers with zero emissions and robust management using renewable energy", is an innovation project aims for optimizing management of electricity and service flows for a data center powered with multiple energy resources. The project leads by Toulouse Institute of Computer Science Research and units interdisci-

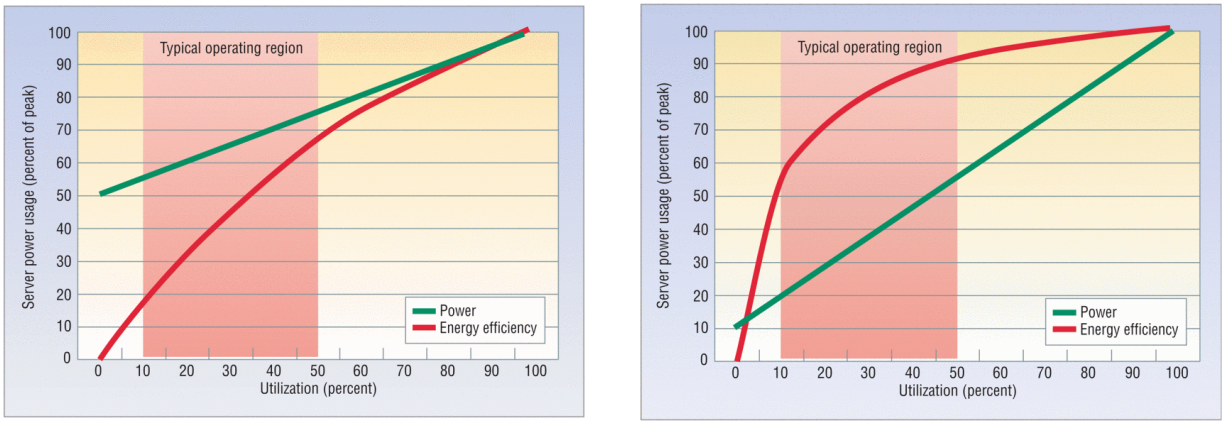
plinary team from both industry and academia. The consortium also includes Eaton Cooperation, LAPLACE laboratory and the FEMTO-ST Institute. For now this project is still in progress. DATAZERO project targets at mid-sized data centers (up to 1000 m2 and 1 MW), where IT loads can be managed through either virtualization or cloud orchestration. The main purpose of the project is to enable using effectively multiple renewable energy sources in a data center by proposing a simulation toolkit. Users are able to test by tuning and comparing several mixes of renewable sources, trying scheduling policies in order to reach a given level of performance in a robust and efficient manner. The simulation can be done by providing necessary descriptions to the simulation toolkit, such as application, IT load and energy source information in forms of XML profiles.

Powering data center with renewable energy is beneficial to reduce the environmental impacts of the facilities caused from using traditional fossil energy. However, these solutions are not aim at reducing the global energy consumption. At the rest of the section, we are going to discuss the hardware and software solutions proposed to reduce the energy consumption, scale from processors to data centers.

1.2.2 Energy proportional design: Hardware solutions

In an energy proportional system, overall power consumption is proportional to its utilization. This concept is firstly introduced by Google engineers Luiz André Barroso and Urs Hölzle [BH07]. In the paper, they argue that servers could be much more energy efficient with an energy proportional design. An ideal energy proportional system is expected to consume nearly no power when idle and gradually more power with the increase of activity level. At that time (2007), they investigate the CPU utilization of thousands of servers at a google data center, for a period of six months. They find that servers operate most of the time at a utilization percentage between 10 and 50. They calculate and compare the energy efficiency at different usage levels for two servers with different power proportional features. The results are shown in Figure 1.5a and Figure 1.5b. Utilization has been defined in this case as a measure of the application performance: normalized value to the performance at peak load. The green line represents the normalized server power to its peak, as a function of utilization. Energy efficiency is derived here as the division result of utilization and power. Red line represents the energy efficiency of the server at different utilization. Figure 1.5a indicates that the server still consume half of the peak when doing nothing useful work (utilization equals 0). Energy efficiency drops quickly while lowering utilization. Most of the time, servers are running with efficiency lower than 50%. When taking a look at an energy proportional server in 1.5b, the energy efficiency within the

typical operating range has been improved greatly, even for the utilization at 10%, efficiency can reach near up to 60%, which almost triples the value of the previous server.



(a) Less energy-proportional server

(b) More energy-proportional server

Figure 1.5 – Server power usage and energy efficiency at varying utilization levels [BH07] © 2015 IEEE

This section details the hardware solutions dedicated to energy proportional designs, including configuring voltage and frequencies settings for the processor, switching to different sleep states to turn off unnecessary hardware resources, improving energy efficiency through multi-threading and hybrid hardware combination, etc. The designs vary from component level to entire and large-scale systems.

Dynamic voltage and frequency scaling (DVFS) (ACPI P-states) For CMOS circuits, dynamic power dissipated per unit of time by a chip depends on voltage and frequency, and can be defined by equation 1.7 [HIG94]:

$$P_T = CfV_{cc}^2 \tag{1.7}$$

Where:

C : dynamic power-dissipation capacitance, in farads (F)

f : frequency, in hertz (Hz)

V_{cc} : supply voltage, in volt (V)

For a given processor, C is normally a constant value according to physical features. Therefore, reducing V_{cc} can lead to power saving according to the formula 1.7. However, constrained by stability requirement in CMOS design, f must be scaled along with V_{cc} [LSH10], and that will result in performance degradation. The advanced configuration and power interface

(ACPI) [Hog04] specification defines the CPU performance states (P-states) and idle states (C-states). Through P-states, ACPI selects different DVFS operating levels in order to adjust dynamically voltage and frequency according to current load, in order to saving active power consumption of a CPU. While ACPI also defines different idle state to adjust power consumption of a CPU during period without activities. We will detail idle state management in next part. Under DVFS, P-states defines the pair of voltage and frequency at which the processor operates on. The experimental research of Cesarini et al. [CBB17] illustrates how different P-states manage DVFS levels. The Levels of Intel P-state range from P_0 to P_n and higher n means slower processor speeds (takes more time to complete a task) and less power consumption. It can be configured dynamically according to system workload [int15a]. P_0 represents for the highest frequency with the possibility to run with at the peak load. DVFS has proven to be a highly effective method of matching system power consumption with required performance. Many advanced usage of DVFS approaches are also proposed to reduce forward energy consumption under specific situations [CSP05] [KDLA07].

C-States C-states refer to idle state, which means the operating system actually executes nothing, except C0 state. Basic idea of C-state is to cut down activities for some subsystems inside the CPU, such as cutting the clock, reduce the voltage or even completely power off. Therefore, C-state x, or Cx, represents for one or some of subsystem of the CPU is powered off. The higher the x is the more parts of subsystem are powered off, and less energy is consumed. At C0 state, CPU is fully turned on and is able to reactive immediately. P-state can be exploited only at C0 state. CPU sleeps at deeper in higher C-state and requires more time to "wake up" and back to 100% operational (C0 state). [Del18]

Intel: Hyper-Threading(HT) Number of transistor implemented in a processor continue to increase so as to keep up with the growth of performance requirement. However, as a matter of fact, transistor count and power consumption goes faster than processor performance. Simultaneous multi-threading is one of the thread-level parallelism approach, where multiple thread can run within one processor without switching. This approach makes better use of the resource available in processor, therefore performance can be improved at a greater rate than power dissipation, which improves significantly the energy efficiency. Intel's HT technology adopts simultaneous multi-threading approach to its architecture and is firstly introduced in Intel Xeon processor family [MBH$^+$02]. For each physical processor that present, HT makes it appears as multiple logical processors to the operating system and shares the workload between

them. Thus, multiple tasks can be executed simultaneously within one processor. Later on, Intel compare the performance of two Xeon processors with and without HT functionality, the results show that, HT is effective to a range of applications of type data-parallel and compute-intensive [MPS02]. On the other side, as quoted by OpenBSD maintainer Mark Kettenis, the implementations of Simultaneous Multi Threading typically share TLBs and L1 caches between threads, and that is suspected to lead to security problems. Therefore, in 2018, OpenBSD operating system decided to disable hyper-threading "in order to avoid data potentially leaking from applications to other software" [Sha18]. Then in 2019, following by critical chip flaws revealed: Meltdown [LSG+18] and Spectre [KHF+19], millions of computers are exposed to security risks [Gre18]. Experts in security recommend disabling HT to prevent potential attacks [Gre19], even through that could lead to a huge performance loss.

Heterogeneous computing Hybrid architectures are proposed on optimize energy proportionality. Such that, workloads can be migrated dynamically between high performance hardware and low performance hardware. Existing hybrid designs vary from one component to the whole data center [WA12].

In terms of the processor, ARM's `big.LITTLE` technology is one of the power management technologies to save power in mobile SoCs [Hol16]. It works with Dynamic Voltage and Frequency Scaling (DVFS), clock gating, power gating, retention modes, and thermal management to deliver a full set of power control for the SoC. It combines with the high-performance CPUs and smaller CPUs in one CPU subsystem to allow software to dynamically move to the right size processor for the required performance. The latest big.LITTLE software and platforms can save 75% of CPU energy and can increase performance by 40% in highly threaded workloads to the appropriate CPU core based on performance needs [AQ14].

Villebonnet et al. inspires the concept of `big.LITTLE` and extend it to a data center scale [VDCL+15], by proposing the infrastructure BML ("Big, Medium, Little"). BML is composed of heterogeneous computing resources, as well as a corresponding BML framework acts as the global scheduler to deal with the applications and resources management. They validate this proposition by running simulations for a stateless web server use case. The performance and energy profiles show that, running distributing requests with BML heterogeneous nodes saves more energy than homogeneous nodes.

The hardware solutions here informed us of how to reduce the power consumption by taking advantages of P-states, C-state and heterogeneous architectures. It is difficult to increase forward processor performance and efficiency through hardware solutions, but there is still much

room left for software solutions. The following section introduces several software solutions to optimize the energy usage efficiency: increasing the load level of the physical machine through resource arrangement.

1.2.3 Energy proportional design: Software solutions

Recently, the break outs at hardware level have been slowed down due to fabrication process. Comparing to hardware design, software consolidation provides more possibilities for researches to improve the energy proportionality of an existing infrastructure, as it has less requirements on hardware architecture and is easier to implement [BCH09]. Likewise, as shown by figure 1.5, more IT load, better the energy efficiency of servers. Conversely, less the servers are utilised, worse the energy efficiency. Therefore, software consolidation can take effect on two things to improve the overall energy efficiency: a) for servers: increasing IT load; b) for clusters/data centers: turning off unused servers. This section details some commonly adopted software consolidation solutions on energy proportional designs, including virtualization technology, server consolidation, VM/container migration and right sizing policies.

Virtualization Virtualization technology is a popular and widely applied method in improving energy proportionality for computing systems. Virtualization can reduce the power consumption at data center level by assembling job into less Physical Machines(PM). In this manner, the rest of the PM could be turned off or turned into deeper sleep mode. In 2011, Koomey [Koo11] published a report for The New York Times. The study found the growth of electricity has slowed down because of energy efficiency improvements brought by virtualization: from 2005 to 2010, worldwide electricity consumption increased only 56% instead of a doubling as observed from 2000 to 2005. Here, we discuss two forms of virtualization which have currently many mature use cases: `Virtual Machine` (VM) and `Container`. `VM` technology provides a complete and isolated environment between multiple systems. Relying on partitioning [BHR89], virtualization enables distributing hardware resources of a PM to multiple VMs. For a VM environment, `guest` refers to the process or system that runs on a VM, and `host` refers to the underlying hardware platform that supports the VMs. Thanks to the isolation, VMs can coexist and run simultaneously on the same hardware platform. The security or system failure that may occur on one guest system will not affect the others [JR05].

Server consolidation and VM migration Server consolidation entails replacing multiples servers running at low utilization with a single server running at a higher utilization [HLM+09]

[SSS+16]. And the consolidation process involves VM migrations between different PMs [FNCR11]. Since VM acts as a fully functional system, the entire VM can be migrated to the other nodes on the network at run time [JR05]. There are two challenges in performing effectively VM migrations among nodes:

— Optimizing VM placement. Hardware sources of VM is allocated and fixed once created (i.e. the amount of capacities of CPU/RAM/Storage and etc.). The size of VM varies according to the job requirement. However, the sum of resources required by VMs should not exceed the fixed capacity of PM. The real capacity of a physical machine is fixed (for example, 64Gb of RAM). Therefore, the capacity volume demanded by any VM should consider the available capacity left in the physical machine, in order to avoid "overcommit". Common objective to save energy in a data center is to reduce the number of active PMs, thus an ideal VM placement algorithm could provide better use of PM resources, so as to minimize the number of active PMs [SSS+16].

— Minimizing delay and energy cost. Even though the VM migration can bring in energy saving benefits, migration operations among different PMs still requires a mount time, this delay could be significant and lead to additional energy cost, if the migration frequency is not suitable [OAL14].

Lefèvre and Orgerie [LO10] propose Green Open Cloud (GOC) architecture as an energy-efficiency framework dedicated to Cloud architecture. They develop a prediction algorithm to anticipate the upcoming recourse requirements in order to switch on the nodes in advance. Frequent On and Off cycles have been avoided by aggregating the reservations. Finally, their solution is validated on a physical cloud nodes and 25% electric consumption is proved to be saved in using GOC.

Khosravi et al. [KGB13] introduce Energy and Carbon-efficient (ECE) VM placement algorithm. Their proposal considers both Power Usage Effectiveness (PUE) and carbon footprint rates. The simulation results of `CloudSim` show that, the ECE can reduce the power consumption by an average of 8% and 20% and the CO_2 emission up to 10% and 45% comparing to other solutions. These reductions have no degradation effect on Quality of Service (QoS) to the platform.

Later on, renewable energy resources implementations have been considered on VM migration. In the work of Li et al. [YAJ17], solutions have been proposed to maximize the use of on-site solar panels for small/medium-sized data center (with 20 to 150 servers). In terms of scheduling VMs, they combine the algorithm of First Fit Decreasing (FFD) and resource over-commit, aims to use as less as PMs. Then they propose solution to find optimal solar panel

dimension and battery size so as to balance the energy losses brought by battery inefficiency and VM migrations cost. Comparing to other approaches based on Energy Storage Devices (ESDs), the proposed solution can save up to 33% more energy consumption.

Container Migration Containers are also referred as OS level virtualization [Sch14]. A container is an isolated software package out of the rest of the system. The package provides all the dependencies, codes and libraries are assembled to a distinct image, so as to support an application running quickly and reliably across different computing environments. Unlike VM, containers do not replicate the entire operating system, they just keep the necessary components needed to operate on. This strategy makes them more lightweight and easier to develop. In terms of migration, containers on a PM are not only sharing the underlying OS kernel but also some libraries. Therefore, destination PM should prepare these libraries before taking in migrated containers. Common container migration technologies including: check point and restart (CR) [MKK08] and another project CRIU [con] (based on CR). Other than containers, VMs have complete and isolated execution environments (an OS) once created, they can be received and managed by other PM [FGXR18]. Recently, container migrations become more popular in the internet of things (IoT) domain. The lightweight containers are very suitable for supporting IoT equipment, which have usually limited computing capacities [PVM+19].

Slow/Shut down policies Considering the inefficiency brought by high energy usage at idle state, shutting down unused servers that are not expected to be used in a long period, is expected to achieve reasonable energy savings. Moreover, when applying consolidation workloads for fewer node, overall energy can not be reduced effectively while switching on unused nodes. Meisner et al. [MGW09] propose `PowerNap` in their work. PowerNap is an energy-conservation approach which is able to switch the entire system rapidly between a high-performance active state and a near-zero-power idle state. Rather than powering off the unused idle servers, authors intend to put them into deep sleep mode in exchanging for shorter wake-up time. This method suits especially for a unpredictable dynamic system, where user demand varies rapidly and randomly.

Benoit et al. [BLOR17] point that, shutting down policies can not be applied at large-scale if no constraint is respected on the target system. The model proposed addresses various constraints, such as time and energy cost of shutting down and waking up, power capping constraints imposed by provider, etc. The author present formal definitions for three shutting down models with different features: Basic Models, Sequence-Aware Models and Power-Capping-

Aware Models. They also provide possible applications of these models through simulations on real workload traces. The simulations show that the models can be used only or combined according to different application scenarios and architectures.

Shutting/slow down policies can combine with VM migrations to move forward the energy proportionality at data center scale. However, how to determine the number of servers to turn into slow mode and how to deal with upcoming requests remain still a challenge. Lin et al. [LWAT11] propose the idea of "right sizing" the data center, in order to avoid overhead brought by infrastructure and make the data centers more energy proportional. "Right sizing" refers to dynamically adapt the number of active servers to match the current workload. Unused servers are allowed to enter a power-saving mode (e.g., go to sleep or shut down). In their work, they propose online algorithm Lazy Capacity Provisioning (LCP) to minimize the total cost. Total cost including two cost models: a) Operational cost related to active server's activities; b) Switching cost, modeling the costs of changing states between active and power-saving modes, such as delay caused by migrating connections/data/etc. (e.g., via VM techniques). Later on, many studies have been propose to optimize and extend the work of LCP [AQ18] [LLWA12] [ZZS18].

1.2.4 Advanced power management

Nowadays, data centers have multi roles and become more and more complex: heterogeneous, geographically-located, mixed energy resources, etc. Apart from power proportional designs targeting at optimize single system energy efficiency, power management is essential for such large scale data centers in regulating hardware and energy resources. Nowadays, many sites have implemented renewable resources on site. In these cases, advanced power management can help for instance: avoiding safety issues, minimizing total cost and promoting the green energy across different sites. In this section, we concentrate on several power management approaches dedicated to large-scale data centers, from traditional "power-capping" approach to newer opportunities brought by Geographical Load Balancing and block chain.

Power Capping Power capping is a hardware mechanism to cap peak power of servers. Most servers have been shipped with this mechanism to limit the peak consumption of servers to stay within a set threshold [LWW08]. For example, Intel apply Running Average Power Limits (RAPL) technology to enforce in hardware a given power limit. RAPL scales up and down core's frequency by tunning the P-states, in order to limit the power under constraint [CBB17]. One of the typical usage is provisioning power consumption of servers lower than observed

peak, so as to ensure safety. In practice, real word application can barely exercise every sub component of server at the peak consumption denoted by nameplate ratings of server. Data center designer measure the peak power of each individual server by running the hosted application at the highest request rate supported by servers. However, peak power may increase after software changes or reboots. Power capping technology can fix the peak power once determined, thus avoid damaging circuits and power distribution units. Capping the peak power of servers can not save too much energy. Instead, changing cap levels dynamically according to workloads will be more efficient. Coordinated power capping systems have been proposed [WCLK12] [WDG+16] to make control decisions efficiently and safely.

Intel has applied the power capping technology by using "Intel Dynamic Power Node Manager (DPNM)", which adopts IPMI interfaces and an add-on software in order to balance and trade off power consumption against performance for a group of servers [Int15b]. One major challenging in applying power capping is reducing maximum energy without losing too much performance. Intel's power capping technology has been successfully applied into a data center of Baidu: for a three-server rack, a 750W power capping policy has been implemented (250W per server), rather than capping the power for each individual server. Before power capping policy, the peak power at rack level can achieve 900W. The implementation policy has been tested on other levels as well, such as 200W and 260W in place, in order to evaluate the final impact on performance. Finally, the optimal policy level set to 250W is confirmed to reduce platform power consumption while maintaining an acceptable performance level [Sam09].

Sun et al. [SLH+16] discover that applications under different hardware resources allocations, such as different CPU capacity, memory size and I/O bandwidth configurations, can have similar performances but distinct power consumption. They defined this phenomenon as "Performance-Equivalent Resource Configurations (PERC)". This observation exhibit the possibility of reducing energy usage of running an application by re-allocating hardware configurations without performance degradation. Generally, power capping methods based on resource-constrained focus on cutting down CPU-related activities when current power usage exceeds power budget, which will more or less lead to performance degradation and be less effective when dealing with non CPU-intensive applications. Authors introduce a new framework of power capping and propose an heuristic algorithm called PowerCap in using PERC replacement. Basic idea of this approach is to select the optimal one among several PERC candidates, which has the highest power reduction and least or no performance degradation.

Recently, some innovative methods combined with up-to-date technologies have been proposed for large scale data centers power managements. These propositions, such as Geographi-

cal Load Balancing (GLB) and blockchain give new ideas and directions for decreasing forward the energy consumption of future data centers.

Geographical Load Balancing (GLB) Recently, concept of GLB and "Right sizing" (as mentioned in previous subsection) provides new opportunities to make the large scale data center (with thousands of servers) more energy proportional [QWB+09] [ZZS18]. GLB distributes dynamically the workloads to various data centers at different locations. In the first place, GLB is proposed for safety and gaining economical benefits for industries. For example, preventing shutting down services or losing permanently data from unpredictable accidents like nature disaster at one location [Avi19]. In terms of saving money, energy prices varies dynamically across different regions according to nature resources conditions (wind, solar variations), local time difference, etc [IA09] [RLXL10]. GLB can be applied to reduce total cost by moving services from one location to the others with lower electric cost [GP13]. However, the action of reducing cost can paradoxically result in the rise of total energy use [LLW+11b] [GP13]. Nowadays, researches are trying to "follow the renewables" in using GLB, which means to use as much as "green" renewable energy instead of "brown" fossil fuel energy. Liu et al. [LLW+11a] investigate the possibility of powering up internet-scale systems entirely or nearly with renewable energy. Their result highlight the effectiveness of using GLB in reducing the brown energy use. GLB could significantly increase the renewable energy capacity. After that, emerging solutions have followed up to optimize the concept [LLW+11b] [GP13] [TQdAB17].

Block-chain technology Blockchain has also been applied to help improving the general energy efficiency of data center. In the work of Xu et al. [XWG17], the authors consider using block-chain technology to minimize the total energy consumption of cloud DCs connected to both power grid and fluctuating green energy (wind, solar and tide) resources, without prior knowledge about future green energy generation. Traditional resource management models adopt usually a scheduler to handle request and VMs migrations across different clould DCs. And the migration process will cost a lot during network congestion. In order to reduce the extra energy cost spent at request scheduling and request migration among DCs, the authors propose a block-chain-based decentralized resource management framework to replace the scheduler. Moreover, block-chain features bring other benefit like robustness to the framework, as failure from one data center will have no impact to the continued resource management, which brings significant robustness to the data centers. In the end, they implement the reinforcement learning (RL)–based method in the framework to minimize the total energy cost from request migrations

among DCs. RL generates the model from learning historical data, therefore, prior knowledge of the upcoming renewable energy generation is not required. In the end, they valid the proposition by running the simulation in Google cluster workload traces. The results achieve 50% and 20% more energy savings than Round-robin(RR) and MinBrown(MB) approaches respectively.

1.2.5 Advanced cooling technologies

High temperature remains the main cause of component failure [AR08]. Data centers have to be adequately cooled for guaranteeing sustainable safety and reliability to IT equipment. As we mentioned in section "Introduction", traditional cooling infrastructures usually take up about 50% of the total power consumption in a data center, and could be even worse in some cases [EJF14]. Therefore, the improvements contributed to optimize the energy efficiency for cooling system represents an effective manner to reduce the global power consumption of data centers. Recently, the evolution of servers increases continuously power density. In order to avoid hot spots, temperature is usually set below the IT requirement for the cooling system [NBA11], which results in considerable waste and inefficiency. Therefore, advanced cooling solutions have emerged to improve the cooling ability while reducing the energy use, for example, increasing heat transfer efficiency by placing cooling equipment closer to the heat generator components (rack, servers, even CPU); taking advantages of the outside environment conditions to reduce the cooling cost; designing buildings with special structures to facilitate the air circulation, etc. In this part, we introduce some advanced cooling system technologies, which helps the energy consumption of cooling system, especially the solutions for the DCs with high power densities.

Air-cooled systems Air conditioning system in a data center has two fundamental functions: cool down IT equipment and manage air distribution. According to the ways of distributing air to the IT equipment, there are three mainstream cooling designs: based on room, rack or row. The basic design concept of the three manners is compared in figure 1.6, distinguished by the position of the computer room air handler (CRAH) units.

In the figure, blue arrows indicate the paths of the primary cooling supply in the room. For Room based cooling system, CRAH units are associated with the room. Room-based cooling has simpler and economical mechanical design, suitable for small data centers with lower power densities (below 5kW per rack [ILA18]). For rack based cooling, each rack is equipped with independent CRAH unit, provides powerful cooling performance, and dedicates to extreme power density situations (up to 50 kW per rack) [DR10], however, additional equipment brings extra

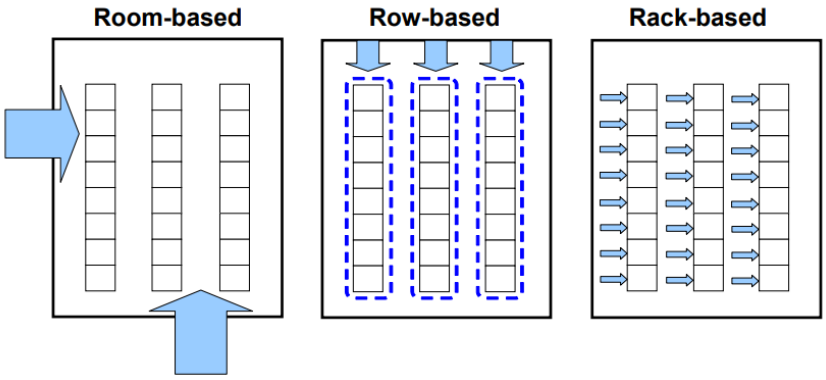
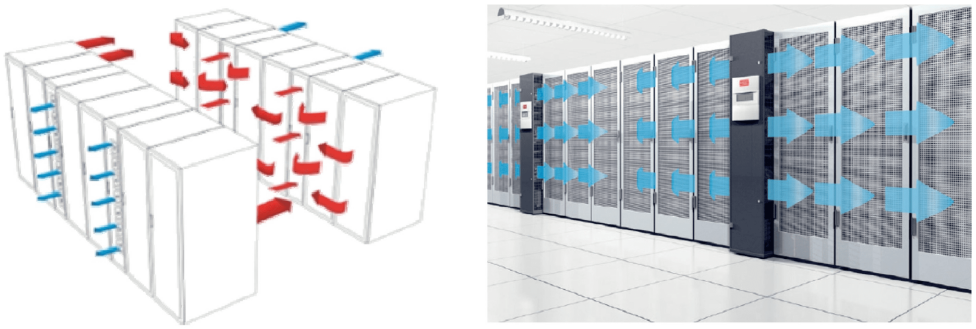


Figure 1.6 – Basic concepts of room, row, and rack based cooling [DR10].



(a) (b)

Figure 1.7 – In-Row cooling system in a DC [PK14] [LZY18]

cost on equipment, which makes rack-based cooling too expensive to afford. Row-based cooling places the CRAH unit between racks, and provide relatively powerful cooling performance to deal with high power density needs (above 6-7 kW per rack [ILA18]), and with less cost. Rack-base and row-based cooling are designed to constrain the cooling power to target and limited area (within rack or row), the airflow paths have been reduced, which economize a lot the CRAH fan power and increases the efficiency. For data centers with light load, fan power is usually wasted into wide space. Figure 1.7 shows the basic concept of In-Row design applied in a DC environment. Nowadays, as the cost of row-based cooling has been reduced and becomes acceptable, this cooling method has been commercialized and is considered to be a good choice for data centers of new generations, for the reasons of high efficiency and redundancy [LZY18].

Liquid-cooled systems Direct air cooled system is still attractive as a conventional cooling technique wherever possible, thanks to its mechanical simplicity [Ete07]. However, air-cooled system is not supposed to be the best solution for DCs with high power density as the cool-

ing capacity and air management are no longer adequate. The poor air management may result in rising local temperatures and harm the IT equipment [CYP14]. Liquid-cooled systems have been proposed in dealing with high power density situations. Among all the liquid cooling techniques, liquid cold plates are the most advanced commercial devices [Ete07], which are equipped by certain server models and provide server-level cooling solutions: cold plates are heat exchangers in direct contact with heat generating sources like CPU, memories. For example, in 2010, IBM built a hot-water-cooled supercomputer prototype called `Aquasar` in Zurich as part of IBM's First-Of-A-Kind (FOAK) program [ZMT+12], the built-in microchannel copper coolers replaced the heat spreader and contact directly to CPUs and DIMMs (Dual In-Line Memory Module, a type of computer memory). The chip level cooling ensures high efficiency heat transfer between components and the water, as shown in figure 1.8. The prototype also disposes of an air cooled part to help comparing the performance between air and water cooled systems. The comparisons show that, thanks to the higher thermal conductivity and specific heat capacity of water, heat can be transferred from chip to water more quickly and easily. For water cooled side, the temperature difference of 15 C° between water and chip is sufficient to meet cooling requirement, while for air cooled side, however, the difference mush reach 35 C°. That allows water being heated up to 60C° without causing overheat of chip (the temperature of chip must be kept below 85C°). Moreover, such high temperature water at outside provides opportunities for heat reuse, such as heating building spaces [IBM10].

Figure 1.8 – Water cooled IBM Blade Center QS22 [ZMT+12]

Free cooling Free cooling, also commonly known as the cooling system in economizer mode, is believed to be the most effective way to save cooling energy consumption. As recognized by the name, free cooling is meant to cool the facility with natural free resources, usually by taking advantage of outside conditions [ZSX+14]. Therefore, it can help even replace the role taken by traditional mechanical cooling. For a long period of time, the adoption of free cooling is suspecting, because the outdoor condition was not considered appropriate for a data center environment, as IT devices requires strict working conditions, especially the operating temperature and humidity [ZWZ17]. In 2008, ASHRAE TC9.9 evolve the requirements to expanded the environmental range for data centers to cover more locations in the world and enable longer operating hours in economizer mode. Later on, in 2011 ASHRAE update the TC9.9 guideline by including expansion of the environmental classes, to provide different operating envelopes that matches different business values and climate conditions [Com11].

The 2011 updated version brings new opportunities in saving energy at cooling side. Since then, free cooling technologies experience emerging developments. As shown in the recent ASHRAE thermal guidelines [Com11], the recommended envelope (Rec) defines the long term operating limits that can ensure the greater reliability of IT equipment. The allowable classes (A1-A3) specify that, the environmental conditions for data centers operated at economize mode are accepted to loose to allowable envelopes for short periods of time, without affecting the overall reliability and operation of the IT equipment [Com11]. Different classes define envelopes for verifying functionality of IT equipment according to different business purposes [Dem15]. According to a white paper of Green Grid [HPB12], 99 % of the locations in Europe is able to use air-side free cooling all year, if data center apply the A2 allowable envelopes. Nowadays, free cooling is becoming an essential component of the modern DC [GE19]. There are different designs for realizing free cooling, such as air-side, water-side and heat pipe system [ZSX+14]. This part concentrates at free cooling through air-side and water-side, their benefits and applications will be discussed as follows.

The most applied free cooling technology is the air-side free cooling. Air-side cooling makes use of outside cold air to take the whole or part of the role of refrigeration component. The outside air can be used in a directly or indirectly way. Direct air-side free cooling draw the outside cold air directly into the data center room, when temperature difference between inside and outside is suitable [Sha12]. In terms of implementation, direct air-side free cooling is the simplest way to apply free cooling technology. Despite for the simplicity and benefits in cold climate zones, direct air-side free cooling can bring in indoor environment humidity, particulates and gaseous contaminants [CHP+11] affecting the IT reliability. Therefore, ventilation system

is required to work along with fresh air handling unit, includes dehumidification device, filters and air cleaners to guarantee the indoor air quality. Lee et al. [LC13] conduct a simulation to examine the potential energy savings of data centers located at 17 climate zones, their study show that direct air-side cooling is not suitable for regions with too dry and humid climate zones, additional cost demanded by fresh air handling unit will overcome the benefits obtained from free cooling. Moreover, for the regions located in hot areas such as Turkey, the energy savings varied with the months in the year [BA11], free cooling has less benefits from June to August as the outdoor air temperatures is too high.

In order to solve the problems led by direct air-side free cooling, other free cooling technologies have been proposed. Firstly, the outdoor air can be applied in an indirect way to cool the indoor air, which we called indirect air-side cooling system. Generally, this design uses an air-to-air heat exchanger, in order to avoid the indoor air mixing with the outside air [NBA11]. Kyoto Wheel is a classical example applying this design [Pot11]. In addition, when difference of temperature between indoor and outdoor is small, this type of cooling system can be used along with the evaporator to extent the operating hours [LNL15].

Water-side free cooling is another solution, due to the significant thermal mass of water, natural cold water is a good chose to transfer heat. Water-side free cooling aims at cooling down the return water in a chilled water directly, or indirectly with a cooling tower [DW17]. Clidaras et al. proposed a water-based data center placed on ship(s), specially for the zones near sea. The cooling system consists of a closed water loop with seawater-to-freshwater heat exchanger, natural cold water from sea is conducted to decrease the temperature of the freshwater flow used for cooling IT devices. The data centers on the sea could also take advantage of the sea waves when applicable by including further a wave-powered electrical generator into the grid. More commonly, the water-side free cooling is introduced to work alone or together with mechanical refrigeration mode, the system can switch to or work partially at economizer mode when environmental conditions meet the requirement [Lui10]. The returned chilled water can be cooled by cold air or cold water. For example, depends on seasons and weather, mechanical refrigeration can stop completely or partially producing chiller water when the outside air is cold enough [ZWZ17]. More advanced system designs propose the combination of both free cooling resources and renewable resources to make the system more efficiency and adaptable to more zones with different conditions. Hamann et al. [HIVK11] propose an energy efficiency data center cooling system combined the usage of free and/or solar cooling when possible. The cooling system includes a free cooling unit and/or a solar cooling unit, one or more modular refrigeration chiller, and a water loop. The direction of the water loop is selected by a control

unit, through one of the above cooling units or in a combination way. The decision is determined by several aspects: data center thermal load, ambient temperature (condition for free cooling), available sunlight energy (condition for solar cooling). As highlighted by Lee et al. [LC13], water based free cooling system worth more attention for too dry as well as humid climate zones.

In this section, we present the solutions of building a "green" data center with low environmental impacts. A "Green" data center can be achieved by increasing the usage of renewable energy resources and reducing the total energy consumption. Servers and cooling system consume most of the energy required by a data center. We briefly introduce the hardware and software solutions to increase the energy efficiency of server, by means of energy proportional designs. The evaluations of the solutions mentioned above, in terms of performance, efficiency, effectiveness and reliability, rely tightly on precise power consumption measurements or simulation results. In the next section, we are going to present the previous studies concerning the power characterization methods for servers, from instrument and from power models.

1.3 Power Characterization for Servers: Hardware Solutions

Cloud data centers can be formed by a great number of servers. Reliable power characterization is an essential part of power management for data centers. Accurate power measurement data is also indispensable for building reliable power models. We are going to give more details in the next section 1.4. In this section, we list and compare some alternative economical instrumental solutions by discussing principally the following properties: accuracy, availability, we classify the power measure instrument methods into three categories and detail them in the following subsections: internal power meters in 1.3.1, external power meters in 1.3.2 and embedded power meter in 1.3.3.

1.3.1 Internal power meters

Internal power meters allows studying the contribution of each component to the total power of server. Traditional technology uses a digital meter to measure the voltage drop across a shunt resistor across the server motherboard, and to compute the power dissipated on the wire. Several previous studies use this solution to analyze the power consumption behaviors at component granularity while running different workloads: [BJ08] [BEK+02] [DS12]. With the increasing demand of analyzing power consumption for individual components in the server, several solu-

tions have been proposed to target this problem.

PowerPack `PowerPack` is a framework composed of hardware components (e.g., sensors and digital meters), and software tool (e.g., drivers, user-level Application Programming Interfaces (APIs)) [GFS+10]. Together, PowerPack allows power consumption profiling at component level in a high performance cluster environment. Total power consumption can be isolated to each subsystem, such as processors, disks, memory, Network interface controllers (NICs). Later on, These measurements can be correlated to code segments of an application. Power consumption of individual component is captured by a precise sensing resistor tapped into each DC power line, then use a digital meter to measure the voltage difference between two ends of the resistor. Total AC power is also measured by an inline sensor device between system power cable and the wall. The combination of AC and DC measurements allows evaluating the loss of AC to DC conversion as well. Software has been developed to collect, synchronize and analyze data recorded by all the sensors in a manner of "out-of-band", which means using a separate computer system to trait the data. Software compatible with many widely used power instruments available in the marketing, including NI Device, Watt'sUp Pro, Yokogawa, RadioShack and Baytech PDU.

PowerMon2 `Powermon` and `Powermon2` are low cost power monitoring devices designed for analyzing both entire and sub systems' power consumption inside commodity computer systems. Powermon devices are integrated with power meter that connects between an ATX power supply and the other internal components, such as motherboard and hard disk. Powermon can make separate voltage and current measurements on each of 6 DC power tails, at the frequency of 50Hz through a USB interface. PowerMon2 disposes of 2 more measurement channels for additional peripherals such as disks and graphic processing unit (GPU), it can reach up to 1024Hz for single channel and 3072Hz divided for multiple channels measurement sampling. It features also a smaller form that allows fitting in a standard 3.5" hard drive bay of a 1U server chassis [BLFP10]. Figure 1.9 shows the Powermon2 device.

PowerInsight: `PowerInsight` follows the same principle as PowerMon2 and is built on top of BeagleBone board that uses an ARM Cortex A8 processor with 256 MB of DDR2 memory. BeagleBone has small size but powerful enough to satisfy capability and connectivity requirements. Scaling is able to provide raw value for current in mA, voltage in mV and power in mW. Sapming frequency can be greater than 1KHz, which is limited by user-space over-

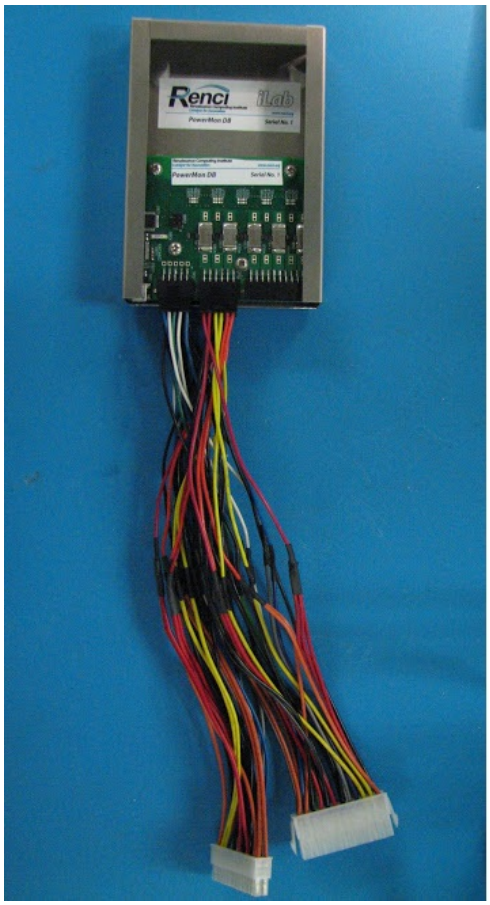


Figure 1.9 – Powermon2 [BLFP10] ©2010 IEEE

head.The board can be connected via the USB device and the onboard 100/100 Ethernet. Other expansion headers includes JTAG interface, 48 pin connectors, SPI links, UARTs, analog inputs, GPIO. PowerInsight can be connected up to 15 components and is used to acquire power measurements from custom power sensing boards connected to it. Each board is then connected through Ethernet and can send the acquired data to a master node [LPD13]. Figure 1.10 shows the connections diagrams of PowerInsignt within a motherboard and among several computing nodes.

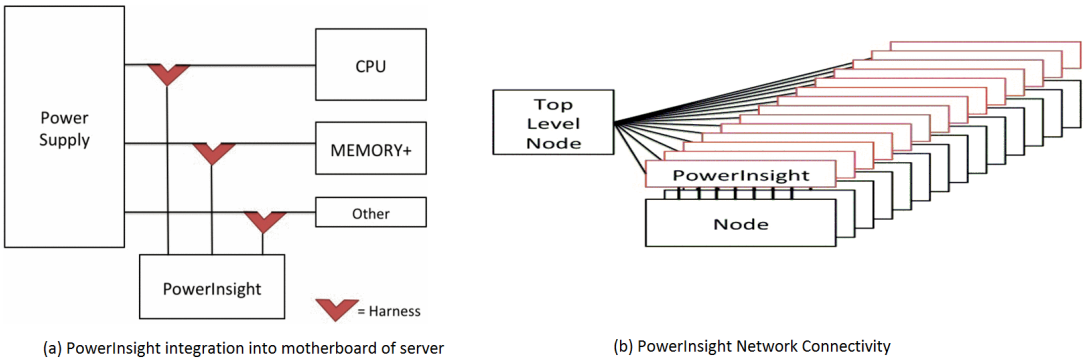


Figure 1.10 – PowerInsight internal and external connections [LPD13] ©2013 IEEE

Internal power meters are capable of conducting component level power measurements, which enables the researches to specific architectures within computing systems. However, when dealing with a large-scale data centers with a great number of servers, each node need independent device, the installation, management and maintenance will be complexe. Moreover, most internal power meters are made for laboratory research purpose instead of being commercialized, the products are not usually available to get in the market. Therefore, internal power meters are more suitable for specific laboratory researches, such as understanding the energy flow within independent or small scale servers. In real situations, the commonly used power meters for getting power data of servers are external 1.3.2 and embedded 1.3.3 power meters, we are going to give some examples in the follow parts.

1.3.2 External power meter

External power meter usually stands between wall plug and the power supply of device. High accuracy power analyzers such as devices accepted by SPEC PTDaemon Tool in this list [Cor], are capable of providing highly accurate power consumption data (refer to "Server Efficiency Rating Tool (SERT)" at section 1.1.2 for more details). However, their limitations

like high cost, weak support of integration with server systems make it unrealistic in widely using high accuracy power analyzers in a data center environment. Energy measurement errors can be accumulated according to instrumental imperfections and data processing inaccuracies, some principal reasons are summarized as follows:

— Measurement errors from Voltage and current sensors. Power cannot be measured directly, the value is calculated by multiplying instantaneous voltage and current data from sensors. Each sensor exhibit a certain measurement error which is inevitable due to manufacturing tolerances.

— Analog to digital convection (ADC) error. Signal come from electrical sensors need to be filtered to reduce aliasing and noise then be converted from continuous time domain to discrete steps.

— Low pass filters. Data sampling rate may not be fast enough to track the highly dynamic behaviors of computing system (in particular for CPU).

— Data conversion. Signals from sensors needed to be converted to machine processing data, multiplied to be power value then to human readable data format. Each conversion/manipulation may introduce additional errors due to limited resolution.

— Time stamp latency. Each measurement is tapped with a time stamp. However, it is hard to determine the latency between the time stamp and the exact time when measurement sample is taken.

In this section, we investigate several low cost external power meters in the marketing to provide some references in choosing power meters for servers.

WattsUp Pro: `Wattsup Pro` was widely used to monitor real-time electricity usage and cost. Currently, the provider of `Wattsup` series has gone out of business and new product is no longer available in the market. Once installed, the device collects and calculates a wide variety of data, such as power (in Watts), Root Mean Squared (RMS) potential present on the power line, current (in ameres), estimated electricity cost (daily, monthly) [Tec15]. The device stands between the power supply and the wall plug, can be connected via USB cable. Official logging software tool can no longer be found but there are several open source libraries for python to redirect the measurements from `WattsUp pro` to PC [Lin16] [yyo15].

PowerSpy: `PowerSpy` is an advanced energy analyzer, it can work in both autonomous data logging mode and real time mode. In real time mode, `PowerSpy` can serve as an oscilloscope, data such as voltage, current, power (RMS and peak), line frequency, etc. can be transferred in real time via Bluetooth to user's PC. In data logging mode, energy information can be logging into internal memory (4 GB) in a CSV format. Internal storage is able to store

data for one month with a resolution of 20ms and up to 20 years with a resolution of 5 years. Reliable measurement is guarantee of 1% precision for a large range, specially for very low power: current from 1mA to 6A, voltage from 90 to 240V (AC), power from 10mW to 1300W and frequency from 45 to 65Hz. A windows based software `PowerLog` is also developed to facilitate the collection, virtualization and analysis of the data. [Alc]

Power distribution Unit (PDU): PDU in data center is used to distribute AC power to multiple servers and related IT equipment. PDUs vary from simple 120 volts power strips to units that bread out 240 volts into 120 volts and three phases. [Dav]. Advanced high edge intelligent PDUs are equipped with multiple functions to satisfy advanced power management requirements in data centers, such as power filtering to improve power quality, monitoring remotely via the SNMP/LAN protocol from a web browser. Moreover, intelligent PDU allows power metering at rack level helping operators to determine real-time power usage and rack capacities in order to identify underutilized servers, avoid downtime brought by overloaded circuits and realize load balancing intelligently to utilize power resources efficiently [Inc] [Tec]. The PDU is designed for power management purposes in data center, therefore device like PDU may have limited temporal resolution. [HIS+13].

External power meters require additional investment, besides external power meters, nowadays, more and more servers propose embedded power meters in their product, gives an other option. The power data can be requested directly from the operating system through specific interfaces such as IPMI and Redfish.

1.3.3 Embedded power meter

IPMI and Redfish are usually available in modern high performance servers. They can be used to monitor system state information such as power consumption, inlet and exhaust temperatures through specific interfaces. IPMI represents for Intelligent Platform Management Interface, created by Intel, Dell, HP and NEC in 1998. It is a standardized hardware management interface and has been widely implemented on more 200 server vendors nowadays [Int]. IPMI is designed to realize system-management independently without passing through OS. Administrators are allowed using IPMI to manage the machine locally or remotely regardless of its state (on or off). Monitoring system status is one of the functionality of IPMI. IPMI can communicate with Baseboard Management Controller (BMC) to retrieve data of certain hardware components (temperature probe, Fans, power supplies, etc.). BMC is a specialized micro controller embedded on the motherboard by the vendors. There are several open source tools supporting IPMI protocol, such as ipmitool, freeipmi, OpenIPMI, etc. Then, with the massive growth in size and

complex of Data centers, traditional IPMI is not sufficient to manage the modern scalable data centers anymore. Hence, In 2010, Distributed Management Task Force (DMTF) proposed Redfish to overcome the limitations of IPMI in terms of scalability, performance, simplicity and interoperability [KSSC17]. In comparison with IPMI, Redfish is a standard API adopts HTTPS protocol, which is considered more secure than UDP protocol (adopted by IPMI). In addition, Redfish use human readable technologies like JSON and OData, which makes the operations such as request and response more user friendly.

Table 1.1 compares the different hardware power characterization solutions.

Table 1.1 – Comparisons between different hardware power characterization solutions

Hardware	*Features*	*Precision*	*Price*
Raritan's PX intelligent rack PDU series	400V three-phase power distribution	Power(kWH): $\pm 1\%$	404 €$\sim$ 1087 €
WattsUp Pro	USB interface communication Energy cost estimation Low cost	Power(W): $\pm 1.5\%$	117 €
Powermon Powermon2	Subsystems monitoring USB interface communication Low cost	Voltage:$\pm 0.9\%$ Current:$-6.6\%/+6.8\%$	125 €
Power Insight	Component level measurement In band and out-of band collection Large scale instrument	Voltage:$\pm 0.3\%$ Current:$\pm 1.8\%$	N/A
PowerSpy	Reliable measurement for low power Bluetooth link between device and software Large internal storage of 4GB Competitive price comparing to high level analyzer	Power(W): $\pm 1\%$	239 €
Powerpack	Component and total power are both available Compatible with many widely used meters Compete power collection and analysis solution	N/A	N/A
IPMI/Redfish			N/A

In this section, we discussed the instrumental solutions to get power data of servers through

multiple types of power meters. Depending on the types, power meters can provide measurement on hardware levels, for nodes, servers (external and embedded power meters) or components (internal power meters). Today, for optimization purposes, the power consumption on software levels are quite demanded by researches, such as the energy consumed by an application, a VM, a process, etc. These requirements are completely beyond the capabilities of power meters. In order to meet the requirements, power models are proposed to provide power characterisation solutions on software level. In the next section, we are going to present previous work concerning different methods of building power models, then discuss the problems and challenging encountered in this research domain.

1.4 Power characterization for servers: Power Models

Modeling power consumption of servers is an active area of research. Power models are built by correlating system activity data, with the power measurement through mathematical analysis. Comparing to physical power analyzers, power models have several advantages. First of all, it provides an economical way to get power data, no more investment on power meters is required. More interesting, power models are built from activity data of system, they have potential abilities to link the energy data with system activities. That gives opportunities to derive the energy consumed on software level, such as single process or Virtual Machine (VM). Next, in using power models, it will be possible to identify the performance bottlenecks, inefficiency of algorithm and optimize the software design in a comprehensive way. Even more, power models are easier to integrated with server system, they can be used to orient some power management and optimization propositions, such as VM Migration, shut down technologies, etc. In this section, we present several examples of building power models, in two classical ways: based on resource usages (1.4.1) and performance counters (1.4.2). The problems and challenges of achieving high precision are discussed as well (1.5).

1.4.1 Power models based on resource usages

In early stage, power models adopt the utilization of CPU as the only input. One of the most notable study is conducted by Fan et al. [FWB07], whose study has shown that the power consumption of servers can be accurately represented by CPU utilization by using simple linear relationship. The error is validated less than 5% for dynamic system activities. Economou et al. [ERKR06] introduced a method called Mantis to model full-system power consumption.

The model is built with linear regression based on component utilization metrics: CPU Utilization, off-chip memory accesses, disk and network I/O rates. The model achieves an overall error range from 0% to 15% for two different server systems. Especially, the blade model has errors less than 5% for all cases. After that, with the evolution of manufacturing, server architecture becomes more complex, the accuracy of models based on CPU utilization has been questioned in many ways. Orgerie et al. [OLG10] highlight that CPU consumption is not linear to its load. The results of their experiments showed that even applying the same CPU load, they observed three different power consumption values. Hence, they concluded that it is indeed not possible to get a linear function between CPU utilization and power consumption. Zhang et al. [ZLQZ13] validated the linear model for 392 published results, which composed of different kind of servers. They use R-squared values to evaluate their model. The authors show that, among 395 published results, 6.5% (25 kinds of servers) have the R-squared values less than 0.95, which means the CPU utilization is not always correlated significantly with server power usage.

1.4.2 Power models based on counters

Furthermore, researchers try to build power estimation models with performance monitoring counters (PMC). PMCs record and store the counts of system-related activities. The principal of models based on PMCs is the selection of several PMCs, which have good correlation with power consumption. The models can be then illustrated through mathematical methods such as linear, non-linear regression formula, or even by neural network. PMCs based power models usually have better accuracy in comparing with single indicators based model. Some previous studies deem the model-based power consumption to be reasonably accurate [GMG+10] [CM05].

Model evaluation metric The model quality can be evaluated with different metrics, they represent for the statistical measure of how well the values of prediction approximate the real data. Here, we present two widely used metric: Mean Absolut Percentage Error (MAPE) and coefficient of determination (R squared).

— Mean Absolut Percentage Error (MAPE), serves as an metric to evaluate the expected value of the absolute error loss. MAPE is calculated as the average absolute percentage error between the actual values and the forest values at each time point [Mak93]. MAPE

is defined as equation 1.8, where A_t is the actual value and F_t is the forecast value.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}|\frac{A_t - F_t}{A_t}| \tag{1.8}$$

— R-squared or R^2 is also known as coefficient of determination. The value of R-squared ranges from 0 to 1, describes the goodness of prediction, the higher the better. R^2 equals 1 means the prediction data fits perfectly the real data. In `Numpy` - a fundamental python package for scientific computing [vCV11], the R^2 is calculated as equation 1.9 :

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.9}$$

Da Costa et al. [DCH10] evaluate the power consumption of a PC by using performance counters, then extend the conception to predict the power consumption of single applications. Training data is collected by running several applications and synthetic benchmarks. A small number of optimal variables combinations within 165 different counters are selected for each synthetic benchmark, which has the best regression result with the real power consumption measurements. Then, a global model for the entire PC is derived by including and analyzing the possible candidate variables. The global model has R-squared values greater than 0.94 for all the cases when applied respectively to each benchmark. Even though they did not evaluate the global model with some real applications, their results confirm the feasibility by using PMC to predict the power consumption of the IT systems.

Running Average Power Limit (RAPL) proposed by Intel is one of the nobel work in power modeling area [STIH11]. Dongarra et al. [DLLW12] compare the energy measurements from both RAPL model and PowerPack framework with a sampling frequency of 100ms. The experiments are performed on two multicore architectures: one dual-socket quad-core Intel Xeon and one quad-socket quad-core Intel Sandy Bridge. They consider RAPL to be a trustworthy alternative to physical measurements.

Bircher et al. [BJ12] propose a method to create power models for six subsystems (CPU, memory, chipset, I/O, disk and GPU) by using performance counters within processors. They chose several performance events which are highly correlated to power consumption in subsystem including memory, chipset, IO, disk and processor. Resistors are connected in series with power source to capture the power consumption for each subsystem. Their models are validated by a wide range of workloads and achieved an average error less than 9% for each subsystem. However, they did not mention the model accuracy for the entire system.

Witkowski et al. [WOPW13] present a practical approach to estimate power consumption of

applications in High Performance Computing (HPC) environment. Their models are represented as regression functions by using only a few variables related to CPU, motherboard and memory. Variables will be included in the model once the coefficient of determination increases. Some of the original variables are transformed to increase model accuracy. When validated with the same synthetic workloads during training phase, their model reports an average error between 1% and 4% comparing to real measurements. However, the average error is increased to a range of 3% -7% when tested with a real HPC application.

Some state-of-the-art power models provide platform-specific solutions [LGT08] [BdM12], which makes the model more accurate and adaptive in current situation. However, the methods used to build the model lack always portability, the usage is limited to specific platforms or conditions. Some other researchers suggest advanced machine learning techniques to improve accuracy of PMC-based models for general use. Some of them point that, the accuracy of model can be greatly increased by removing some irregular outliers of measurements [WOPW13] [DCH10].

Cupertino et al. [CDCP15] propose to use Artificial Neural Network (ANN), one of the computational intelligence technologies to improve the model accuracy. They compare the MAPE between an ANN model and a traditional capacitive model, and show that the ANN can decrease the MAPE from 5.45% to 1.86%. MAPE is a widely adopted indicator to evaluate the model quality, less error means better quality.

Wang et al. [WCS11] point out that for a given processor, the usage of PMCs is limited by the available event counters and the maximum number of PMCs that can be read simultaneously. Even more, power models with less PMC can be more flexible and applicable. Based on this fact, the authors have then proposed a power model with only Instruction Per Cycle (IPC) and frequency as inputs. In order to improve the accuracy, running benchmarks will be divided into different categories based on IPC values, then they build the models separately for each category. The authors also develop a tool “SPAN” to realize run-time power profiling and correlate power dissipation to source code functions. Their power model is validated by using two benchmarks from SPEC2008Cjvm, and achieve absolute error rate of 5.17% and 4.46% respectively. Tool "SPAN" achieve accuracy as high as 97% on average by running FT benchmark from NAS Parallel benchmark suite and synthetic workloads.

Mair et al. [MHEZ14] present their power estimation model called “W-Classifier”. The model classifies different workloads into 5 categories by using some power-dominant PMCs: INT, FPU, FPU/cache mixed, INT/cache mixed and memory/idle. They validate W-Classifier with OpenMP multi-threaded benchmarks from NAS Parallel Benchmark suite on all 16 cores.

They find that W-Classifier has an average MAE of 6.95% for all benchmarks, while traditional multi-variable model achieves an average MAPE of 40.74%. However, authors admit that W-classifier has difficulty to estimate the power consumption of benchmarks with large range of power variation. They have then proposed to improve the model by adding more kinds of classification categories as further work.

Power models have wider interesting applications, they are essential for conducting energy efficiency optimization studies. However, according to some evaluation studies, the precision of power models are not so promising as stated by the propositions. Actually the high precision is not easy to achieve and the accuracy could be doubtful under different situations. We will go through the potential challenges and problems stated by previous studies in the next section.

1.5 Challenging in building accurate power models for modern processors/servers

Building high accuracy power models is challenging, not only due to the advanced techniques of mathematical analysis. With the evolution of processors and servers, advanced technologies have been applied into the product and rise the challenges of building high-accuracy power models. Many evaluation reports and studies show that, the precision of power models can be potentially influenced by defects and uncertainties appeared in devices, servers and environment. In this section, we list these underlying influences raised by the observations mentioned in these researches.

1.5.1 Power characterization instruments

Internal watt meter mentioned in 1.3.1 seems to be a perfect power characterization approach to take a deeper look at how power is distributed among different components. However, the precision provided by internal watt meter is not always satisfying. [DDG+13] performed a test and compared the readings between external power meters and internal power meters when running the same benchmarks. They pointed out that internal power meters like `Powermon2` may not have the same reading compared to external power meters. The values collected by internal power meter are more dispersed and there are more outlier samples. Internal device can cause a difference of more than 50% for low consuming benchmarks like `hdparm` and `iperf`.

Hackenberg et al. [HIS+13] evaluate the RAPL accuracy for a Sandy Bridge system in using the physical measurements taken by 12V P8 LMG450 as the reference. They find that

the accuracy of RAPL depends on the type of workloads. Computationally intensive worload has little deviation comparing with the reference measurement. However, RAPL in the tested system underestimated the energy consumption for a particular memory workload. The authors blame the causes to the Sandy Bridge architecture, as the DRAM domain is not included for RAPL measurements. In contrast, RAPL overestimate the energy consumption for idle state. In additional, RAPL provide energy rather than power consumption data without time stamp information updated to each sampling.

In terms of the power characterization approaches under data center environment: intelligent PDU, IPMI and Redfish are commonly used solutions to realize large-scale power management for a group of servers. However, there is a lack of study addressing the precision of these tools. Moreover, there isn't a specification uniform the precision for IPMI or Redfish adopted for servers. Especially, at the written time, the work for Redfish hasn't finished and is still a "Work in Progress" on the website of DMTF [DMT].

1.5.2 PMC related problems

As mentioned in section 1.4.2, power models built on PMC could have better correlation result than models with single indicator. Obtaining accurate results of consumption behavior at the whole system level or individual component level is not straightforward. According to previous studies, the difficulties include but are not limited to the following reasons:

1) Diversity. Physical architecture of server differs very much between manufactures and becomes more complicated from generation to generation, with the emerging of new features. The availability of PMCs differs among different machines [WOPW13]. The problem of the diversity makes the power models less portable between heterogeneous servers in Data centers.

2) Evolution. Evolution of system is somehow rapid and random. Some indicators used to build the original model would no longer exist with the evolution of computing system. For example, four years after the introduce of Mantis [ERKR06] (mentioned in 2.1), John C et al [MAC+11] have noticed that, some of the original indicators used by Mantis no longer exist in current systems.

3) Hidden system behaviors. Some component provider make optimizations without exposing to any of the existing counters, which makes some device behaviors invisible to OS [MAC+11]. High precision will be difficult to achieve without being aware of these changes that affect power draw.

1.5.3 Environmental influences

Some recent studies concluded that except IT load applied to the components (CPU, memory, network and storage), the power of servers can be affected by external factors, such as original fabrication process [CQP14] [vKBB+16a], ambient temperature [OLG10] [WKV11] [Sam12], way of placement in a rack. Among all of the components, processors are responsible for most of the power consumption and the variations [DGLM13] [vKBB+16b]. Normally, the consumption of CPU depends mainly on IT load. As the power increases with the load, CPU works harder and dissipates more heat, if the heat is not evacuated in time by the cooling system, the temperature of the CPU becomes higher and leads to the rise of leakage current, which will in reverse increase the power of CPU [KC09] [MB09]. Patterson et al. [Pat08], the ambient temperature affects server power in two ways: through temperature sensitive components (i.e. CPU) and through server internal cooling fans. They draw the conclusion theoretically by analyzing a typical data center configuration. CPU temperature draws much attention, as a lot of work has confirmed the strong correlation between CPU temperature and server power [HSP+15] [CQP14] [GAAS+16].

Mair et al. [MHEZ13]observed the power latency when running unchanged system load in the server. Moreover, the duration of power latency follows tightly the CPU warm-up period, and the system's fan speed remained steady at 3600 rpm during the test. Therefore, the power latency in this case was not due to the consumption of fan but to the rise of CPU temperature. The conclusion was that the CPU temperature can result in notable variation of power consumption before and after the stable state. They suggested to prolong the execution time to eliminate this thermal impact and increase the model precision. However, the correlation between the CPU temperature and the server power hasn't been discussed and only the AMD architecture servers are concerned in their studies.

El Mehdi Diouri et al. [DGLM13] find that different nodes from a homogeneous cluster have different power consumption at idle state. The power consumption of two nodes stay unchanged even after exchanging the positions. They blame the causes to the age of the processors, as the server equipped with older processors shows more variation.

1.5.4 Variability between identical systems

Recent scientific observations altered that, the fabrication discrepancy between the printed transistors can result in visible difference on performance and power consumption among high-performance microprocessors. Moreover, the variation is becoming worse in modern proces-

sors [AMK16] [MZB+17]. This observation rises doubt about the real precision of the existing power models. For example, if a power model is built upon a server and the validation is also done with the same one, will the precision remain the same to other servers in the homogeneous cluster or even to the whole data centers? According to the experiment done by John C. McCullough et al. [MAC+11], they found that when applying a power model trained on Intel Core i5 labeled 540M-1 to an identical processor labeled 540M-2, mean prediction errors could be increased from 10% to 23%. They conclude that, power instrumentation is the only way to perform accurate power characterization for servers. There are other evaluation studies hold the similar doubts.

Henry C Cole et al. [CQP14] conducted several tests among three server manufacturers (three from Intel, one from Dell and one from Supermicro) with similar mechanical and electronic specifications, in order to determine whether the energy use and efficiency of server had the relationship with their brands. 5% difference of power consumption was observed among three identical Intel servers. They switched their main components in the motherboard in order to identify the source of the difference. The results showed that the difference was mainly brought by CPU. However, these tests were performed within only three servers and the conclusion was suggestive rather than definitive.

Marathe et al [MZB+17] performed several tests to compare both performance and energy efficiency variation among identical nodes on Sandy Bridge, Ivy Bridge and Broadwell clusters. The variations are compared separately with and without hardware-enforced power limit. They found that processor performance and energy efficiency variation is becoming worse with the evolution of computation capacity on modern Intel processors.

Balaji et al. [BMGA12] compare the power consumption variation for modern mobile processors. Their data shows power consumption variation among processors ranging from 5% to 17% when processors operate at the lowest and highest frequency respectively. Different power management settings such as Turbo Boost and C-state can also affect the value of variation.

Acun et al. [AMK16] investigate the processors under Turbo Boost in HPC systems. They point out that dynamic overclocking feature of processor is responsible for substantial frequency difference among the processors, which explains the up to 16% of core-to-core performance variation. The faster processors usually consume more than the slower ones.

Jóakim et al. [vKBB+16b] characterize the variation on CPU power consumption. Experiments are performed on three different platforms and different processors are picked for each platform. Identical processor samples are exchanged after each run to guarantee the identical conditions. The power consumption can differ as much as 29.6% in idle and 19.5% at full load

for identical samples. Their observations also show that CPU power directly influence system power. Additionally, the authors use worklets in SERT (Server Efficiency Rating Tool) [LT11] as workloads to stress the SUT (System Under Test). To our knowledge, during the phase of calibration, SERT will identify the maximum rate at which transaction can be executed for each worklet. This value is highly reproducible for one server in run-to-run test, but may vary from server to server calibration [WNLMM18a]. However, as the performance variation among processors has been observed in previous studies, and the authors has not mentioned the core-to-core performance variation reported by SERT in this paper, we cannot tell if the workloads (worklets after the phase of calibration) used to stress each sample of processor are exactly the same.

Apart from frequency variability, S. R. Sarangi et al [SGT+08] emphasize that, within-die parameter variation can result in process variation including both random and systematic effects, can also negatively impact a processor's frequency and leakage power.

However, most of the studies focus on the difference at the server level, since processors are placed in different servers or different sockets, the results cannot eliminate the influences caused by system noise, such as the influence of the nearby processors, the platform bugs [MZB+17] [TAZ+17] [LKM16]. Moreover, thermal control strategy is rarely mentioned. As processors are temperature sensitive components, different operational temperature can affect the results.

Even though high-accuracy power model is hard to realize, efforts should not be given up on this area. The requirements of accuracy depend on concrete objectives. In terms of consolidating workloads, the base power of total servers is relatively high, modeling within 5 to 10% error is acceptable to support power aware decisions. While for scheduling applications/VMs on a heterogeneous processor platform, the accuracy should be precise enough to capture the difference of power consumption between the cores. Otherwise the decision policy will not be able to make correct guide [MAC+11].

1.6 Conclusions

Environmental impacts brought by data centers in the world has been a great concern, in terms of the huge energy consumption and emission CO_2. However, the scale of worldwide data centers continues to expand dramatically due to emerging demand in cloud computing services, efforts and actions much be taken in to reduce the side environmental effects from running data centers. The subject of the thesis has been part of the "Green IT" project of Orange. The objective of the research aims to optimiser the energy efficiency of a computing system by

proposing a comprehensive power model, which takes into account the hardware and software resources, activities and environmental variables.

In the beginning this chapter, we introduce different metrics representing energy efficiency of computing systems, from data centers to a server. The indicators of these metrics have different evaluation purposes and standards.

Then, in the next section of the chapter, we present the recent technologies trying to build "green" data centers with less electrical power consumption and emission CO_2. Firstly, we introduce the current advancement of powering data centers completely or partially with renewable energy resources. These solutions cannot reduce the total power consumption but are effective of reducing the environmental impacts in long term perspectives. Then we focus on the studies regarding energy proportional designs, including both hardware and software solutions. Since the end of Dennard Scaling, energy efficiency has no more increase possibility by rising transistor densities on a chip, and energy proportional designs provide another promising way to increase the energy efficiency.

Later on, we investigate previous studies addressing obtaining power data of servers, through different instrumental power meters. We explore as well, two classical methods of building power models: based on resource usages and performance counters.

Moreover, during the state of the art studying, we find that the precision of power models is not determined only by mathematical techniques, they can be influenced by various of potential uncertainties around: such as human related operations, measurement tool, complexity and diversity of modern servers, environmental conditions, etc. There lacks obviously comprehensive evaluations with sufficient experimental evidences, to identify and characterize these uncertain impacts. Therefore, the scientific research of this thesis has been conducted by covering two parts: (1) Chapter 4-6: Experimental evaluations on the potential effects from different aspects that may influence the power consumption of servers and the precision of power models. (2) Chapter 7: Estimation the total power consumption of a physical cluster through power models, including the electrical power consumption from cooling and servers.

CHAPTER 2

IDENTIFICATION AND CHARACTERIZATION OF THE MYSTERIOUS AMONG IDENTICAL SERVERS

Suitable configuration of hardware resources in data center is indispensable to make good use of the energy resources [Com]. Much work has been done on building accurate power predictive models for servers in data centers. Some of them proposed high-accuracy software-level solutions as replacements to physical analyzers in order to allocate efficiently physical resources and make the system more energy-aware, as the approaches we mentioned in 1.4.2. The question is, if we build and validate the power model for one server, can we rely on the precision obtained and apply directly the model to the other servers in the homogeneous cluster or even in the whole data centers?

During our experimental measurements, we observe that there exist some external factors that may cause unexpected power variation among identical servers and result in extra errors to the original precision stated by power model designers. Similar observations appear in previous studies also highlight that, other than IT load, the server power could be varied by external factors, several assumptions have been detailed in section 1.5.3. Moreover, among all the assumptions, power variation brought by CPU temperature has drawn much attention as a lot of work has confirmed the strong correlation between CPU temperature and server power. Correspondent findings can be found in 1.5.3. However, there lacks sufficient experimental evidences supporting and characterizing these assumptions, which make some of the statement less persuasive. Therefore, in this chapter, we try to fill the gap by providing more evidences to clarify the mysterious. Series of experiments have been designed and performed to answer the following questions:

— Do identical servers stressed at the same load have the same power consumption?

— Do the factors list here have influence on the power consumption of servers: position and arrangement in the cluster, fluctuating neighboring temperature and variation of source voltage ?
— How thermal effect influence the power consumption of servers?

2.1 Context and objective

In this study, We investigate the mysterious of power consumption among 15 identical servers, several workloads in SERT test suit has been used to perform stress test to the servers. The potential factors evaluated include: different positions and arrangement of servers in the cluster, fluctuating neighboring temperature, source voltage variations. Especially, we design and perform experiments to evaluate and characterize the influence of temperature variation from CPU and from the other components to the power.

The objective of this study is to identify and characterize the influences introduced by these potential factors, that may contribute to the power variation of servers. The findings are expected to remind the further power model designers to consider the factors identified into the model building, and correcting the precision of previous theoretical power models built from mathematical analyses. This chapter is organized as follows: Firstly, in section 2.2, we compare the power consumption of 15 nominally identical servers under the same load. An industrial-standard benchmark has been used to stress different components of the servers at different target levels. This experiment aims at clarifying how power varies among identical servers in a real data centers. Then in section 2.3, we explore several potential factors that may contribute to the power variation of servers in real data centers, including: server arrangements in racks, fluctuating neighboring temperature and variations of source voltage from power supply. Especially, in section 2.4 we study the influence of temperature variation on power consumption and performance of servers. Temperature is varied separately on CPU and on the other components in the motherboard. Conclusions are given in section 2.5.

2.2 Variations of power consumption among identical servers in a cluster

In this part, we execute the same test suite to 12 identical servers in a rack, in order to investigate whether the servers consume the same under the same load. The experiments are

performed to compare both the power and performance variations of the servers, by executing different type of workload of multiple load levels.

2.2.1 "Ecotype" cluster overview

The experiments are performed to the servers of the cluster named "ecotype". Ecotype is one of the clusters belonging to Grid5000 [BCAC+13], and geographically located at the university IMT Atlantique Pays de la Loire, in the city of Nantes in France. Grid5000 is a large-scale testbed for all areas of computer science research, such as distributed computing, Cloud, HPC and Big Data. Until Avril in 2019, as shown by figure 2.1, Grid5000 testbed contains 31 clusters located at 8 sites across France, users have access to 828 compute-nodes grouped in homogeneous clusters.

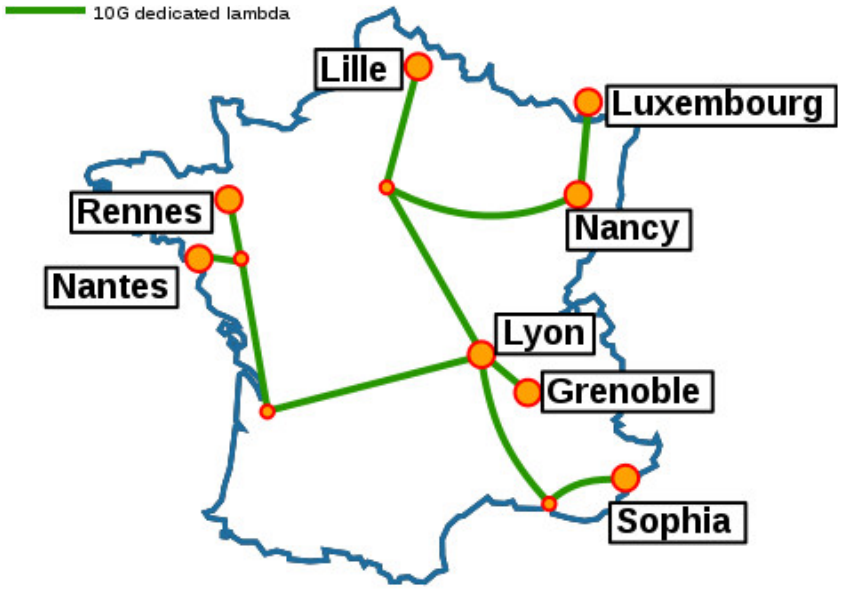


Figure 2.1 – Grid'5000 sites across France

The "ecotype" cluster is placed in an independent room of the university. Figure 2.2 show the front and back views of the ecotype cluster in the room.

Ecotype contains 48 identical servers labeled ecotype 1 to 48, they are installed in 4 server racks of the cluster, each server rack has 12 servers of model Dell PowerEdge R630. Placement and arrangement of the servers in the racks are illustrated in 2.3. Ecotype applies two kinds of server arrangements in the racks. As can be seen from the figure, in rack 1 and 2, the 12 servers are placed loosely in the rack, there are spaces without obstacle between servers. While for servers in rack 3 and 4, they are placed right next to each other without space between them.

Figure 2.2 – Ecotype cluster overview

The later arrangement allows putting more servers in one rack, and the power density will be increased as well. Characteristics of servers can be found in table. 2.1.

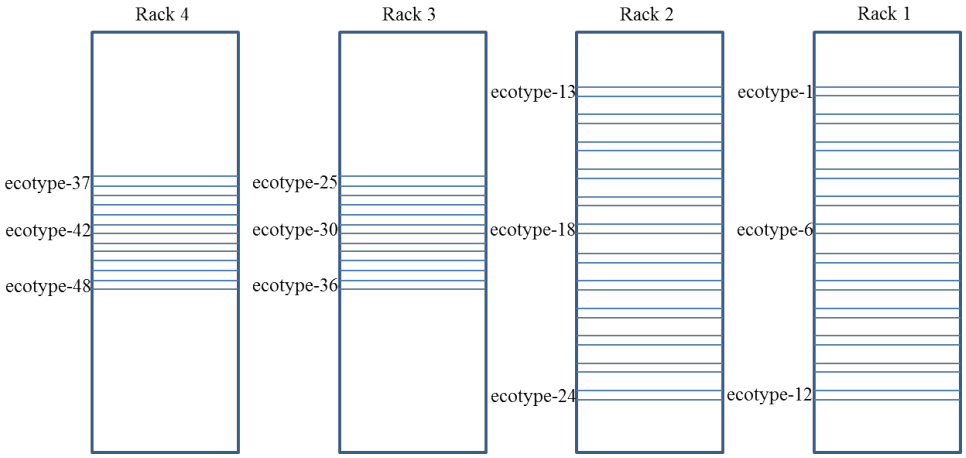


Figure 2.3 – Position and arrangement of servers in ecotype cluster

Table 2.1 – Characteristics of servers in ecotype cluster

Items	Dell Power Edge R630
Processor	Intel Xeon E5-2630L v4
Processor Release date	Q1'2016
Base frequency	1.7 GHz
Cores	10 cores, 20 threads per CPU
RAM	128GB
Solid State Disk (SSD)	400 GB

Table 2.2 – Test suite information

Worklet	Components	Description	Load Levels
LU	CPU	Dense Matrix operations	100%, 75%, 50%, 25%
SHA256	CPU	SHA256 hashing transformation	100%, 75%, 50%, 25%
Sequential	Storage	Reads and writes data to/from file	100%, 50%
Capacity3	Memory	XML Validation	Base, Max
Idle	System	No load on SUT	None

2.2.2 Experiments setup

In terms of benchmarking, five typical micro-workloads called worklets from test suite SERT are chosen to stress the key components of the SUT: CPU, memory and storage system (refer to "Server Efficiency Rating Tool (SERT)" in section 1.1.2 for more details about the SERT Test Suite). Features of these worklets chosen are shown in Table 2.2. Before testing, we have evaluated whether the SERT result is reproducible for same server: all the worklets are executed twice on S1-S5. The results show that the power and performance (throughput) variation is within 1%, the result of SERT is highly reproducible.

The experiment is performed on individual server one after one in the rack 1, with respect to the run rules defined by SERT commit [Com13]. The system's connection diagram is shown in Figure 2.4, which includes principally three parts: Measurement system, Controller and SUT.

— The measurement system is composed by two devices:
Yokogawo WT330: a power analyzer measuring AC (alternative current) power provided to PSU (power supply unit) of server, with maximum measurement error less than 1%.
Testo176 + Thermocouple: a thermometer with two thermocouples connected (type K,

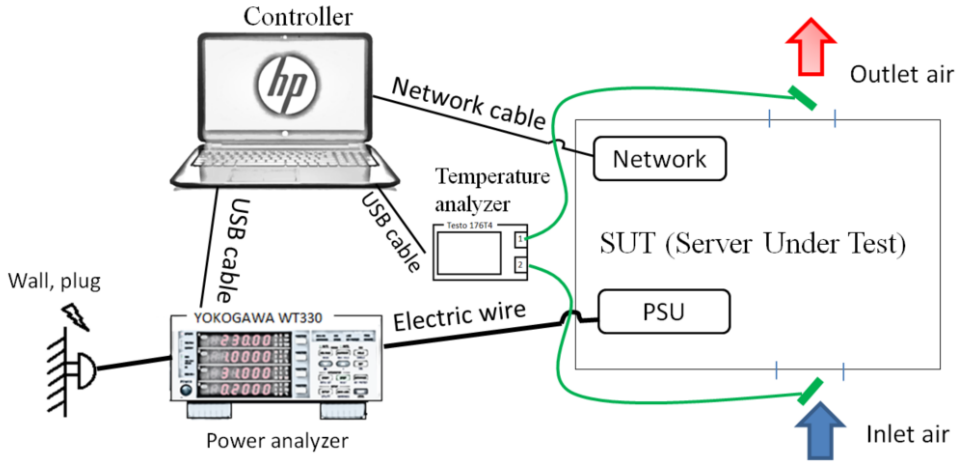


Figure 2.4 – System diagram

with 0.1mm diameter)

— Controller: controls SUT by sending commands via a network cable. Controller also gathers measurement data recorded by power and temperature analyzers with 1Hz sampling frequency.

— SUT: installed with Linux OS compatible to SERT, systems are configured according to the guidelines described in the SPEC Methodology [Com14]

2.2.3 Power variation for 12 identical servers in the same rack

The test suite is repeated on S1-S12 in sequence in rack 1. Figure 2.5 shows the average power of S1-S12 while running different worklets and figure2.6 precises the maximum-minimum power and performance variation on percentage.

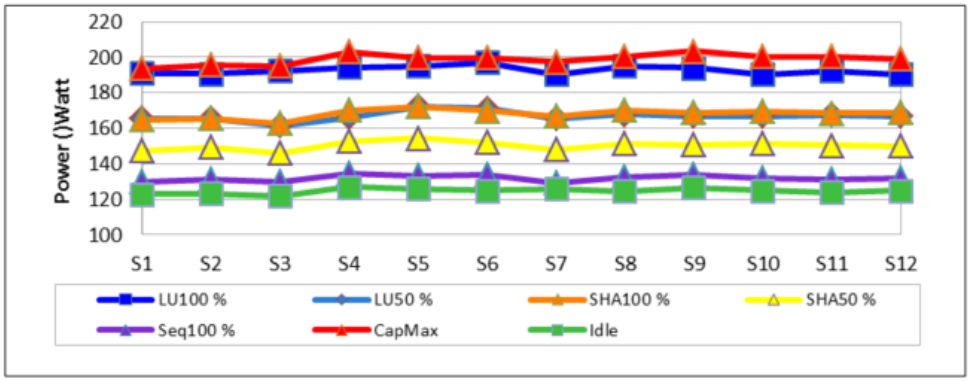


Figure 2.5 – Average power of 12 identical servers while running different worklets

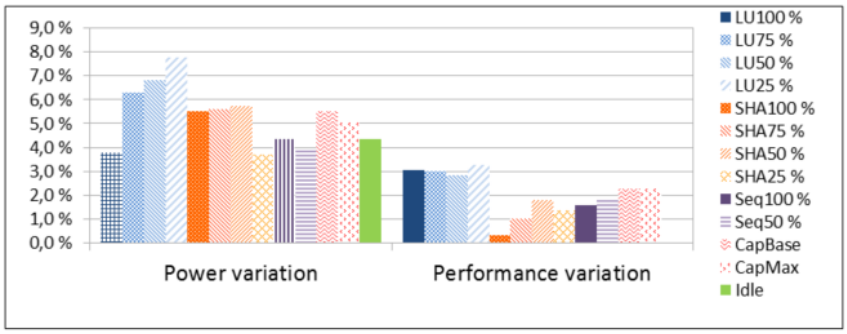


Figure 2.6 – Power and performance variation in percentage among 12 identical servers

It can be observed that under the same load, power variation among identical system in the same rack can reach up to 7.8%. According to our observations and previous state-of-the-art study, this variation may be caused by several potential factors, such as fabrication process variability, fluctuating neighboring temperature, way of placement in the rack and source voltage variation. One of the effective approaches to determine the influence of fabrication variability between SUTs is to run the test suite one more time after switching their positions in the rack [DGLM13] or switching their key components if necessary [CQP14]. However, we are not allowed to open or to move the servers belonging to Grid5000 because of security and assurance issues, the impact of manufacturing variability will be further discussed and studied in the chapter 3.

2.3 Evaluations of potential impacts on power variation among servers

In this part, we identify and characterize experimentally the potential impacts that may result in power consumption variation among identical server. The following aspects have been investigated: different arrangement of placing servers in the rack; fluctuating ambient temperature and variation of source voltage.

2.3.1 The impact of arrangement densities of servers and neighboring temperature

As we mentioned in 2.2.1, servers are placed in the four racks in two kinds of arrangements. For a working cluster with active servers, rack 1,2 and rack 3,4 have different situations in terms

of air circulation and heat distribution between hot and cold sides of the servers. All the racks have same amount of servers, therefore, the power density remains the same. This configuration allows us to investigate whether the way of arrangement has impact on the power consumption of servers. In this case, four SUTs at the same height in the racks are chosen: S6, S18, S30 and S42 (refer to Figure 2.3).

Test suite is repeated twice in two cases. In case one, only the SUT is turned on in the rack. In case two, we increase the neighboring temperature by turning on the other servers. Figure 2.7 shows the percentage of power variation in two cases for four SUTs. The results turned out that, for the same server under different neighboring temperatures, the power varies from 0 (idle, S42) to 5.6% (LU25%, S30). Comparing the variation on percentage between S6, S18 and S30, S42, different way of arrangements have no obvious impact on power consumption of servers.

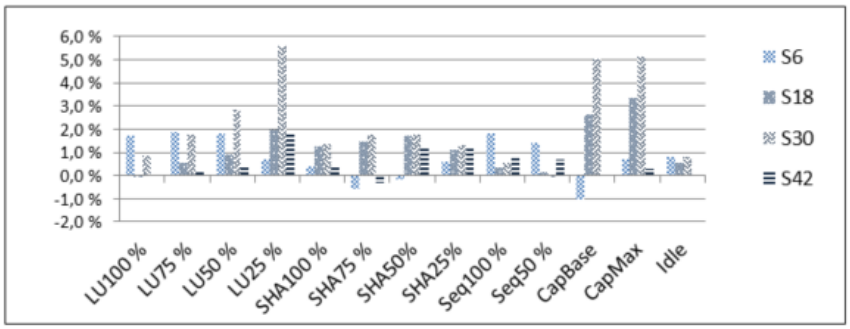


Figure 2.7 – Power variation on percentage for servers under different placement densities

2.3.2 The impact of source voltage variation

While performing the tests in section 2.2.3 , we have noted that source voltage provided by grid varies randomly from 230ACV to 240ACV. In order to eliminate these variables, we perform additional test in a thermal laboratory to determine whether the variations of voltage from grid has impact on power consumption of devices. In the laboratory, the ambient temperature is controlled and remained at 23°C. The power supply of SUTs is replaced by an AC power generator, which can provide stable AC voltage. The tests are performed on two servers with very different characteristics: Gigabyte mw50-sv0 and Dell PowerEdgeR630, servers' details can be found in Tab. 2.3. Test suite SERT is performed three times on each server under the following voltages: 207V, 230V and 253V. The results show that, for Gigabyte, the power variation is less than 2% for all the worklets. For Dell, the variation is less than 1.5% for worklets except the "Sequential" (storage, 2.8 %). These observations demonstrate that, the voltage variation from

Table 2.3 – Characteristics of the SUTs

SUT	Gigabyte mw50-sv0	SuperMicro x10sdv-tln4f	PowerEdge R630
Processor	Intel Xeon E5-2609v3,6 cores, 1.9GHz	Xeon D-1540, 2.0 GHz, 20 cores	2 x Intel Xeon E5-2650L v4 56 cores, 1.7GHz
Release date	Q3 2014	Q1 2015	Q1 2016
Memory	4 x 16Go DDR4 2133MHz	4 x 16Go DDR4, 2400MHz	4 x 32 Go DDR4 2400MHz
Storage	480Go SSD	400Go SSD	400Go SSD

source power will not bring additional impact on the server power. Therefore, voltage variation is not the reason for the power variation between identical servers.

2.4 Evaluations of thermal effects on power consumption of servers

In this part, we focus on the influence of thermal effects on the power consumption of servers. Results of the experiments in previous section 2.2 show that, apart from the system load, ambient temperature is one of the major contributors to the power variation between identical servers. However, ambient temperature variations can act on different components of the servers. According to previous studies, CPU is supposed to be the most temperature sensitive component. However, besides CPU, there lacks evaluations studying the temperature sensitivities of the other components. In addition, dedicate and precise temperature control techniques at component levels is a must to conduct this study. Therefore, the experiments presented in this section aim at filling the study blanks in this area. The impacts of temperature variations on power consumption of servers, will be characterized separately on two parts of the server: CPU and the other components, through variables controlled experiments.

The impact of CPU temperature (leakage current) and of the other components are studied separately in section 2.4.2 and 2.4.3. SUTs include servers from different manufactures of different sizes. Fans, DC power generator and climatic chamber are provided to help controlling precisely the thermal conditions at component levels.

Table 2.4 – Characteristics of the SUTs

SUT	Gigabyte mw50-sv0	SuperMicro x10sdv-tln4f	PowerEdge R630
Processor	Intel Xeon E5-2609v3,6 cores, 1.9GHz	Xeon D-1540, 2.0 GHz, 20 cores	2 x Intel Xeon E5-2650L v4 56 cores, 1.7GHz
Release date	Q3 2014	Q1 2015	Q1 2016
Memory	4 x 16Go DDR4 2133MHz	4 x 16Go DDR4, 2400MHz	4 x 32 Go DDR4 2400MHz
Storage	480Go SSD	400Go SSD	400Go SSD

2.4.1 Experiments setup

Cluster environment is not able to realize the component level temperature control, therefore, the experiments of this part are performed in a thermal laboratory. Moreover, different type of servers have participated in the evaluation, which makes our final conclusion applicable to most cases. Four servers from different providers are chosen as SUT in this evaluation. They are equipped with different Intel CPUs from different generations, their characteristics are shown in table 2.4.

With the rise of ambient temperature, integrated fan may consume more power to reject the heat of CPU, that could influence the final result. Therefore, in order to eliminate the electricity consumed by fans, integrated fans in the motherboard are removed and replaced by external fans powered by a separated DC power supply. The system diagram of the testbed to evaluate thermal effects on CPU and on the other components are shown in figure 2.8.

The testbeds include mainly four parts: measurement system, controller, SUT and temperature control system.

— The measurement system: Fluke 430T, a high-accuracy power analyzer, allows measuring AC (alternating current) power provided to PSU (power supply unit) of server, with maximum measurement error less than 1%.
— Controller: controls SUT by sending commands via a network cable. Power analyzer are connected to controller via USB cables, controller gathers power readings on server power with a sampling frequency at 1Hz.
— SUT: installed with Linux OS.
— Temperature control system: External fan powered by a DC generator and climatic chamber.

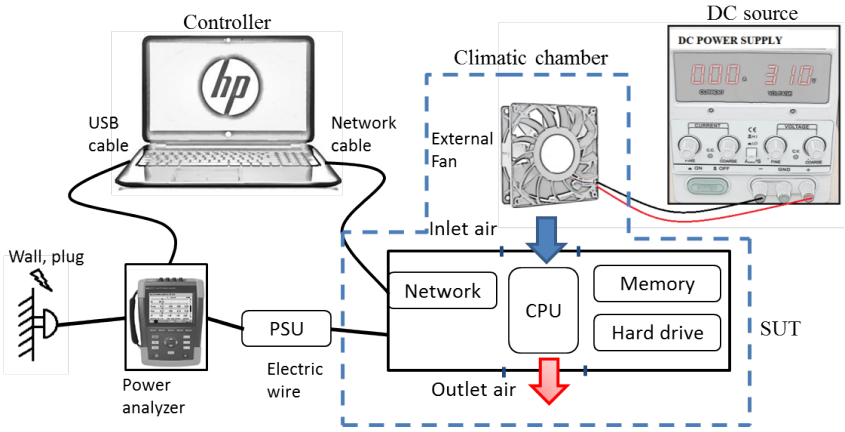


Figure 2.8 – Test system diagram: evaluate influence of temperature variations on CPU and on the other components

2.4.2 Impact of CPU temperature (leakage current)

Benchmark cpuburn [Rob11] is used to stress the SUT in this test. It is a CPU intensive benchmark that keeps the load level at 100% and maximizes the heat production of CPU. Cpuburn is executed on each SUT for more than 30 minutes. While running the cpuburn, we vary manually the surface temperature of CPU via the external fan by adjusting the air flow between fan and heat sink. CPU temperature is varied by following a same pattern on SUTs: firstly, CPU temperature is maintained high for a period of time. Then the fan is put immediately close to CPU to cool down its temperature as low as possible. At that point, we remove the fan, CPU temperature will increase quickly as heat generated by transistors can not be evacuated efficiently. The fan should be replaced before temperature exceed the threshold allowed by manufacture, otherwise system will trigger the protection policy by lowering the CPU frequency. The CPU temperature is varied manually as demonstrated in 2.9. CPU temperature is recorded and marked by red points in the figure.

Results of four different servers are shown from figure 2.9, to figure 2.12. Blue points are the records of server power. It can be observed that under a stable load, the instant data of CPU temperature and server power consumption are highly correlated. The spearman coefficients are larger than 0.93 for all the SUTs. The power varied more than 10 Watt for servers Gigabyte and Dell PowerEdgeR630, R740.

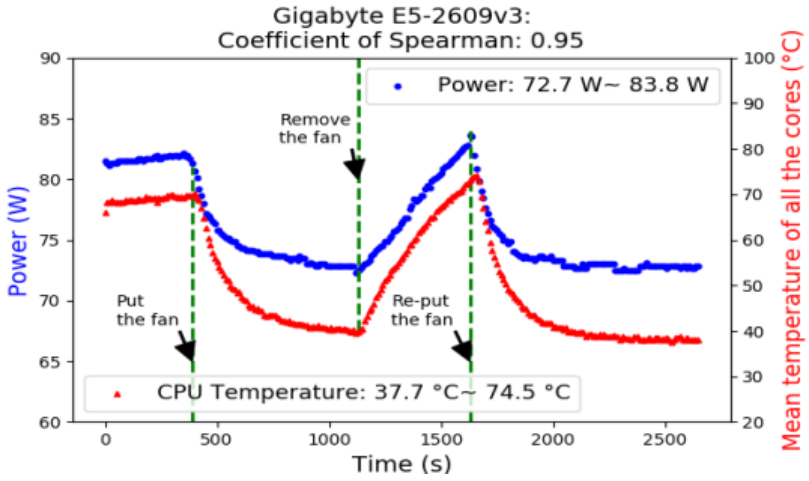


Figure 2.9 – Relationship between CPU Temperature and server power-Gigabyte

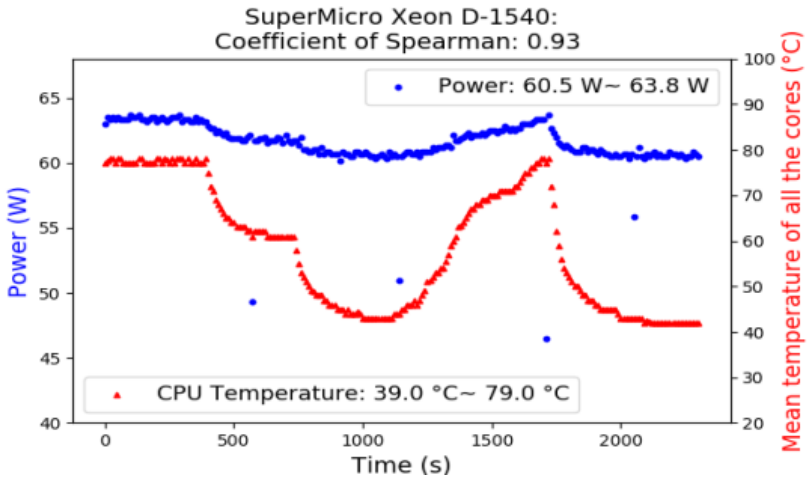


Figure 2.10 – Relationship between CPU Temperature and server power-SuperMicro

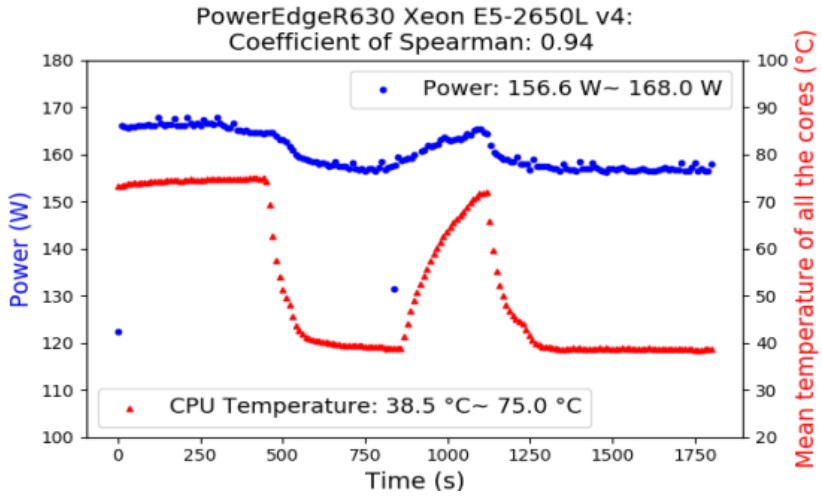


Figure 2.11 – Relationship between CPU Temperature and server power-PowerEdgeR630

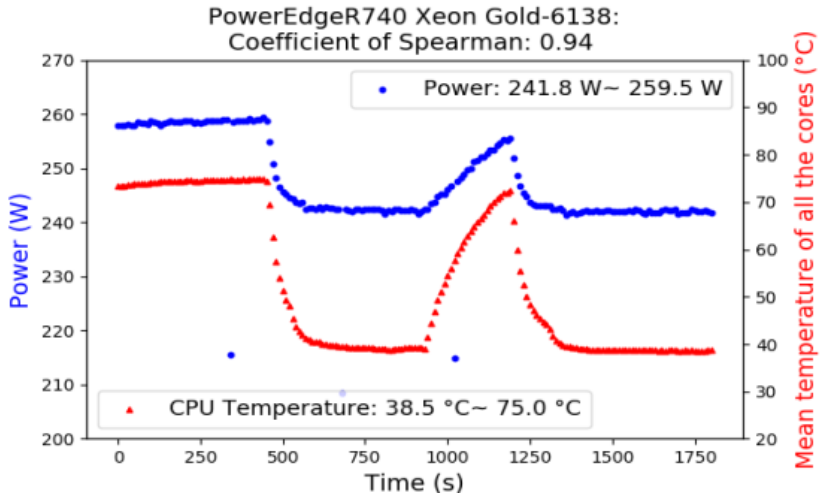


Figure 2.12 – Relationship between CPU Temperature and server power-PowerEdgeR740

2.4.3 Impact of the temperature of the components other than CPU in the motherboard

In this part, we try to characterize the influence of the temperature variation on the other components except CPU in the motherboard. CPU temperature should maintained unchanged while varying the temperature of the other components. In order to do so, we put the whole server into a climatic chamber, where ambient temperature can be well configured and maintained. The temperature of the components will be varied with the ambient temperature in the climatic chamber. Meanwhile, CPU temperature is remained the same (in average) by adjusting the air flow speed of the external fan. As shown in figure 2.8. CPU and memory intensive benchmark stream [Joh] is used to stress the SUTs. The test is repeated twice with climatic chamber's temperature configured respectively at $25°C$ and $35°C$. Results of two different servers are shown in figure 2.13 and figure 2.14. Only the server Gigabyte and SuperMicro are concerned in this test because of the limited size of the climatic chamber. CPU temperatures are remained at $47°C$ for Gigabyte and $59°C$ for SuperMicro during the execution when the ambient (climatic chamber) temperature increases from $25°C$ to $35°C$. **The results prove that the power remains nearly the same when just varying the temperature of the other component than CPU.** Execution time of the benchmark stream is recorded for the two servers under different CPU temperatures, as shown in figure 2.15. **CPU temperature has no obvious impact on server performance.**

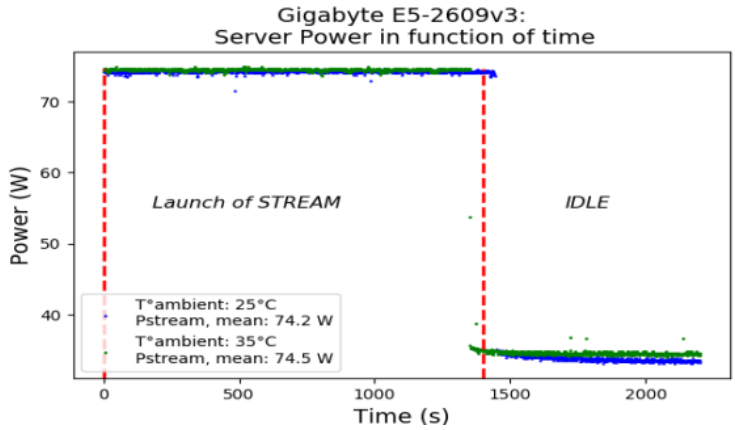


Figure 2.13 – Relationship between temperature of other components and server power - Gigabyte

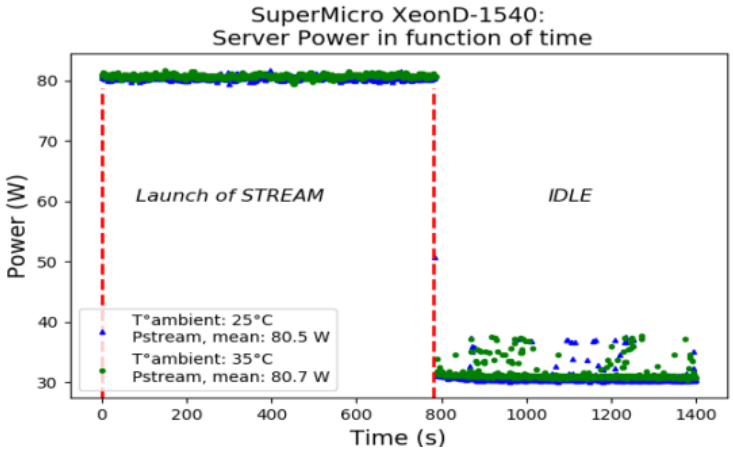


Figure 2.14 – Relationship between temperature of other components and server power – SuperMicro

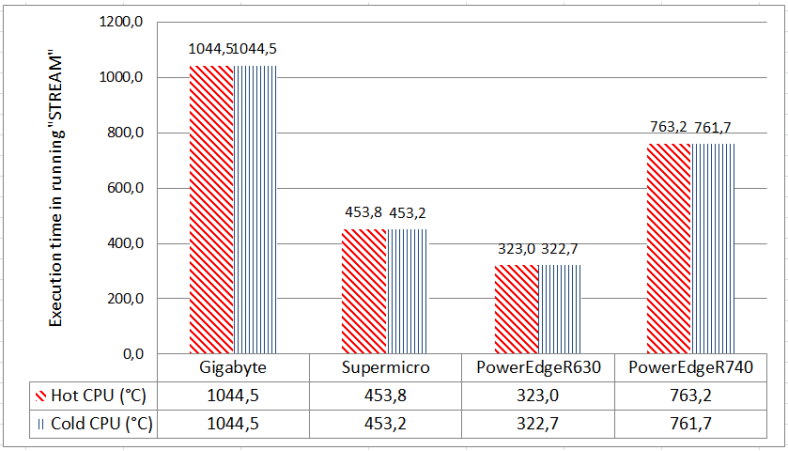


Figure 2.15 – Impact of CPU temperature on server performance

2.5 Conclusions

This chapter addresses how power varies by other external aspects except IT loads. In order to identify and quantify the influences, physical experiments have been performed to investigate several assumptions made by previous literature. In the first place, we investigate the power variation between 12 identical servers in a physical cluster in section 2.2.3, 7.8% power variation is observed while executing a same test suites. After that, we perform several tests to verify the following aspects that may result in power variation: different server arrangements in the racks, fluctuating neighboring temperature and voltage variation from power supply. The results turn out that, influences from different arrangement and voltage variation are obvious. However, we find that, fluctuating neighboring temperature can contribute up to 5.6% power variation for a same server. Thermal effect, turned out to be one of the major contributors to the power variation. Therefore, later in section 2.4, we have further investigated how the thermal effects vary the power consumption of servers in a thermal laboratory. Thanks to the climatic chamber, external fans and DC power supply. We are able to control precisely the surface temperature at component levels. The influence of temperature variation on CPU and on the other components to the power consumption of server have been studied separately. The results prove that CPU temperature can introduce important power variation: for the server Gigabyte equipped with an Intel Xeon E5-2609v3 CPU, the server power (fan is not included) is increased by 16% while the CPU temperature varying from $37.7°$C to $74.5°$C. However, the server power remains nearly the same when just varying the temperature of the other components while maintaining the temperature of CPU. Therefore, CPU is supposed to be the most temperature sensitive component. The ambient temperature affects the power consumption of servers via two ways: the consumption of fans and CPU (leakage current).

The results presented here are significant for the development of a predictive model to estimate the power consumption of servers. Thermal effect from CPU has great impact on server power. The results also emphasize the effectiveness of liquid-cooled solutions at component level (refer to 1.2.5). According to the observations of the first experiment, another potential impact that way cause the power variation between identical servers is the fabrication difference. Servers in a homogeneous cluster seem not to be created "equally". However, with respect to security and assurance policy of Grid5000, we are not able to characterize the influence brought by fabrication variability by opening the servers. Therefore, in the next chapter 3, we are going to discuss and explore deeply the fabrication variability problems by comparing the power variation between the most consuming component in a server: processors.

CHAPTER 3

VARIATIONS BETWEEN IDENTICAL PROCESSORS

In previous chapter, we observe the power variation between identical servers in a rack under the same load. Several assumptions have been investigated by experiments. Among all the assumptions, the influence brought by fabrication variability between servers still remains unsolved. In such situation, we require more conditions to guarantee the same ambient temperature for each server in a rack, while the security and assurance issues not allow us to do further investigations in using these servers. Among all of the components, processors are believed to be the biggest consumer comparing to other components, according to previous studies such as [DGLM13] [vKBB+16b]. Moreover, recent studies altered that, the tiny fabrication discrepancy between the printed transistors can result in visible difference in terms of both performance and power consumption among high-performance microprocessors. Therefore, in this chapter, we try to address the fabrication variability problem by identifying the differences between identical processors.

3.1 Context and objectives

Physical experiments have shown that even under the same conditions, identical processors consume different amount of energy to complete the same task. The variation is becoming worse in modern processors [AMK16] [MZB+17]. In the domain of modeling, variability could rise problem in accuracy. For example, John C. McCullough et al. [MAC+11] found that when applying a power model trained on Intel Core i5 labeled 540M-1 to an identical processor labeled 540M-2, mean prediction errors is increased from 10% to 23%. They suggest using power instrumentation instead for accurate power characterization.

Recent experimental studies have identified several sources of variation such as: frequency variation introduced by advanced performance enhancement technologies like Turbo Boost and Multi-Threading [MZB+17] [AMK16], within die parameter variation [SGT+08], aging

[DGLM13], etc. Even though this manufacturing variability has been observed and studied before, according to our acknowledgement, there lacks of experimental evidence supporting the hypotheses due to limited amount of samples, especially from perspective of thermal characteristics.

The objective of this study is to investigate the power variation between identical processors and identify the variation sources brought by fabrication processing. Addressing the lack of study on physical experiments and limited samples, we investigate the power variation between samples of two processors: Intel Xeon E5345 and Intel Xeon E5-2603v2. They have the same number of cores and Thermal Deign Power (TDP), but come from different generations. Thirty identical processor samples participated in this evaluation for each generation. Evaluation of the variability is done by switching processors in the same platform in order to eliminate the influences introduced by platform design. Thermal parameters such as ambient and CPU temperatures are controlled and varied with the help of a climatic chamber and an external powered fan.

The chapter is organized as follows: Firstly, in Section 3.2, we compare the power consumption between processor samples came from 2 generations. Test environment is strictly controlled to provide equal environmental conditions. Then, in 3.3 we propose and evaluate two hypotheses that may eventually led to difference between processor samples, from the perspectives of thermal characteristics: different application of Thermal Interface Material (TIM) and variation of leakage current. Conclusions are given in 3.4. The approaches proposed give new directions in term of identifying and characterizing the influence of manufacturing variability between processors. Results of our experiments highlight that more attention should be paid to the difference between identical information systems.

3.2 Exploring power consumption variation between identical processor samples

Newer technologies bring diverse features in improving performance for processors, and in the same time, introduce more variation regarding performance and energy usage to processors with the same design. In this section, we compare with physical experiments, the power variation between identical samples for Intel Xeon E5345 and Intel Xeon E5-2603 v2. Details about the processors can be found in Table 3.1. The two processors both have 4 physical cores, the same TDP. TDP cannot represent the actual power usage of CPU [And13]. Indeed, TDP

Table 3.1 – Characteristics of the processors

Platform name	G41(socket 775)	Just Game LGA2011
Processor ID	Xeon E5345	Xeon E5-2603v2
Processor Release date	Q1'2007	Q3'2013
Architecture	Cloverdown	Ivy Bridge
Base frequency	2.33 GHz	1.8 GHz
Cores per processor	4	4
Lithography	65nm	22nm
TIM	Solder TIM	Polymer TIM
TDP	80 W	80 W
Turbo Boost	No support	No support
Hyper Threading	No support	No support
OS	ubuntu-18.04-live-server-amd64	ubuntu-18.04-live-server-amd64

refers to cooling system requirement for supporting long-term sustainable workload [HP11]. Usually, it is determined below the peak power but higher than the average power that could be sustained for long-term use (defined by CPU provider, generally determined as the average power during the execution of a given computation). For example, some modern Intel processors equipped with Turbo-Boost technology, which allows CPU running at the highest possible frequency achievable for short duration [Cor13]. The power level at this point is higher than the TDP configurations. Objective of Turbo-Boost is to complete an intensive task at maximize performance for short-term use. In this case, this task is not a long-term sustainable workload for the system. Even more, if execution time at Turbo-Boost is asked to be extended, temperature limit could be exceeded, conditions will not meet the Turbo-Boost running requirements, and heat management system will slow down the clock speed until meet the sustainable average power specified by TDP. Therefore, TDP is a reference for cooling system designer, the cooling system should be designed to match or exceed the TDP [HP11]. In terms of the software test environment, same operating systems Ubuntu 18.04 OS are installed on the platforms. Turbo Boost and Hyper Threading technologies are identified as important impacts result in variability between processors, corresponding researches have been presented in section 1.5.4. In our case, none of the two technologies is available for the two processors. That provides an opportunity for us to concentrate on exploring deeper sources of variability behind the design, especially the thermal characteristics. Otherwise, the two processors dispose different designs on architecture, TIM application and operating frequency.

3.2.1 Experiments setup

The testbed platform is designed to create strict identical operating environment for the processor samples. The whole platform (excluding the power supply) is placed in a climatic chamber, where the ambient temperature is configured at $35°C$. The ambient set point is set little beyond the ASHRAE typical requirement envelope (15- $32°C$, for products require a stable and more restrictive environment) [Com16], so as to exposure as much as possible the thermal characteristics difference between samples. Fan is placed on top of the processor's heat sink and powered by an external DC (Direct Current) source so as to avoid power variation brought by motherboard's fan. Homemade CPU intensive workload called `bzip2` (compress folder with different type of files) and `pi_calculator` (calculate bits of pi) are chosen as stress tools to maximize the usage and heat dissipation of CPU. Homemade workloads are designed with the following features: a) The workloads can active all the threads available in the system and maintain them at the full load for a configured period of time; b) During the test time, instant power of server has stable value without much variation, it is important to compare with different samples with stable value; c) The homemade workloads simulate the execution of actual application in the real world, performance data (execution time) can also be obtained for further analysis; d) Moreover, we found that workload `bzip2` can make processors generate more heat than `cpuburn` (with higher CPU temperature). That meets better our expectations to maximize the thermal characteristics of samples. Samples of processors are tested and exchanged one by one in the same platforms. When exchanging the samples after each run, we try our best to uniform the heat paste applied between processor and fan. The platform test diagram shown in figure 3.1 includes three parts: SUT, control and measurement system.

— SUT: Server Under Test, installed with Ubuntu Server OS.

— The control system:

Controller: normal PC (personal computer) with Linux system installed, in the same local network with SUT and controls SUT remotely by SSH. Controller also gathers and synchronizes measurement data recorded by power and temperature analyzers.

Climatic chamber: Servathin RC01, controls ambient temperature, provide equal test environment.

External Fan: controls CPU surface temperature, powered by external DC source. External Fan is used to replace the integrated fan in the motherboard, in order to eliminate the power consumption variation from the integrated fan.

— The measurement system:

Power: measured by Wattsup-Pro [Tec15]. Measures contain the power consumption of

the whole server except the fan's.

Ambient temperature: measured by a thermometer with a thermocouple connected (type K, with 0.1mm diameter).

CPU Temperature: collected by Linux command line tool `lm-sensors`.

Sampling frequency for all the measures are configured at 1Hz.

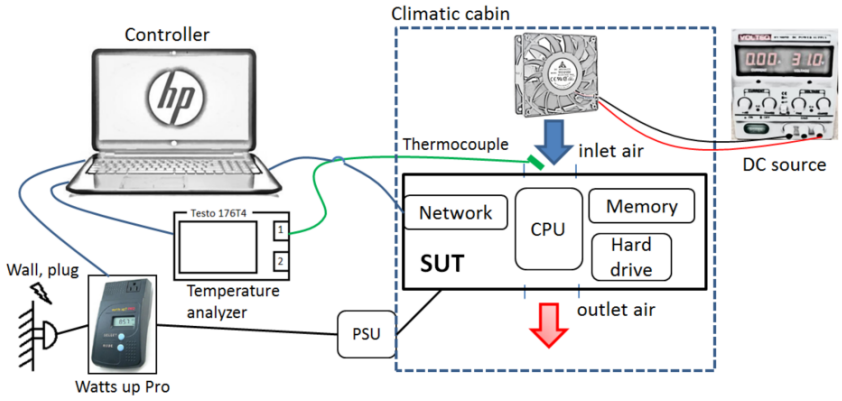


Figure 3.1 – Platform test diagram for exploring power variations between identical processors

3.2.2 Power variation between identical processor samples

In this part of study, we investigate the power consumption variations between two Intel processors, by executing the same workload under equal test environment. Environmental conditions and test processing have been detailed in previous section 3.2.1. The SUT equipped with different processor samples has been stressed at full load during about an hour. The power taken for analyzing the variation between processor samples is the average power during the last minute, when thermal condition tends towards stable.

For each processor, the power consumption data are collected for all the 30 samples. While in order to have a better view and understanding, for each type of processor, we choose 18 representative samples for presenting the results: 9 samples that consume the most and another 9 samples that consume the least among the 30 samples. Figure 3.2 and Figure 3.4 show the rank of power consumption by running application `pi_calculator`. Moreover, CPU temperatures are also collected during the test. The two processors both have four physical cores, and temperatures are collected per second for all the cores by the integrated sensors. During the stress test, all cores in the CPU are activated to run at full load. We take then, the average temperature of four cores per second as the CPU temperature value of each sample. The distri-

butions of CPU temperature for each sample is presented in Figure 3.3 and Figure 3.5. Right below the power consumption figures, and take the same order. Box plot is chosen to illustrate distributions along the test (variation of CPU temperature around about 1 hour).

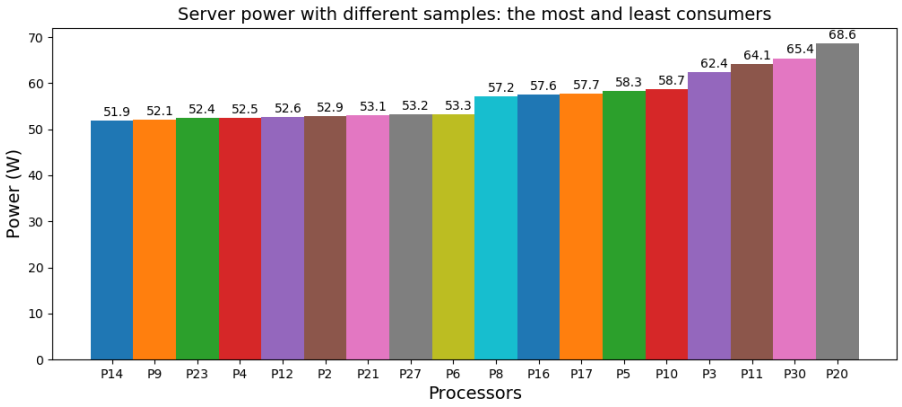


Figure 3.2 – Server power comparison for samples of Xeon E5-2603v2

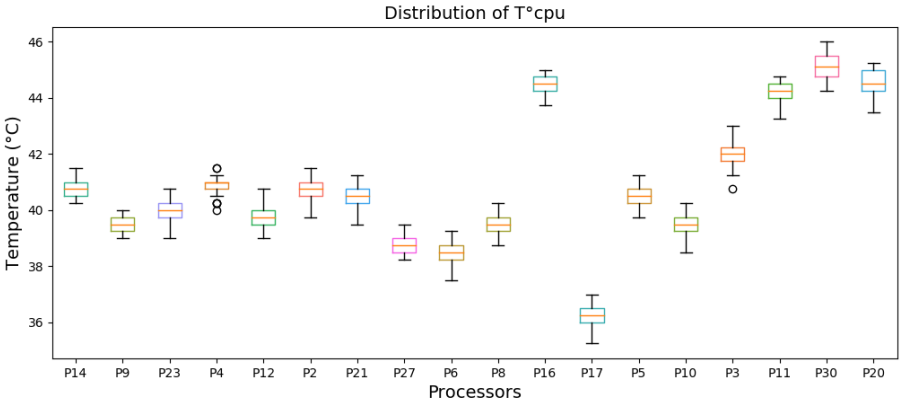


Figure 3.3 – CPU distribution for samples of Xeon E5-2603v2

Power variation between processor samples is calculated as the maximum percentage difference between the best and worst cases. It can be obtained form figure 3.2 and 3.4 that, processor samples from Xeon E5-2603v2 (newer generation) have the power variation of 30% (16.1W / 51.9W), which is much greater than the power variation obtained from samples of Xeon E5345 (older generation): 2.8% (2.8W / 98.4W). In order to make sure that the power consumption variation is not caused by execution of workload, we re-verify the power variation of Xeon E5-2603v2 by performing the same experiment with another CPU intensive workload `bzip2`. The rank of power consumption of the samples remains the same. Therefore, the power variation

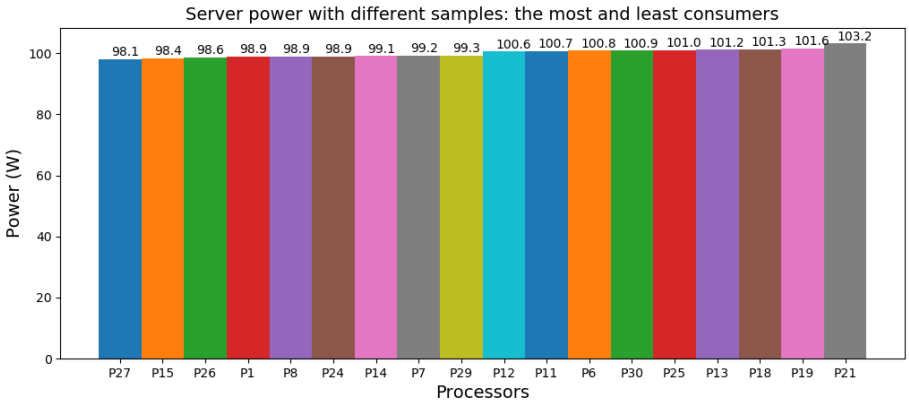


Figure 3.4 – Server power comparison for samples of Xeon E5345

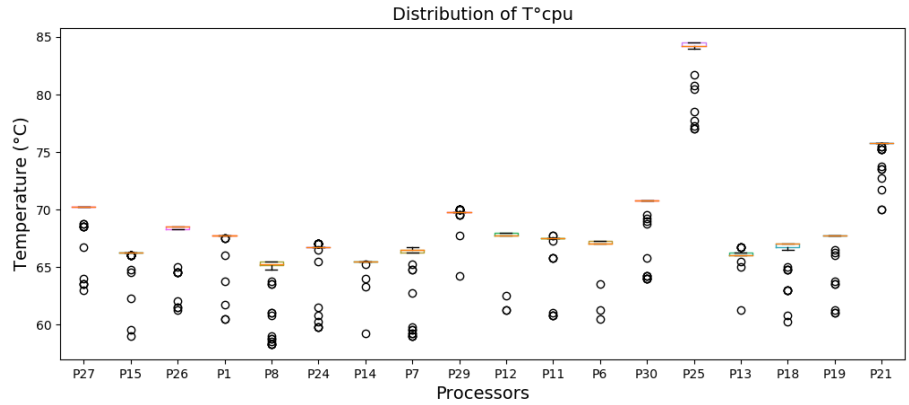


Figure 3.5 – CPU distribution for samples of Xeon E5345

observed in Xeon E5-2603v2 does exist and has nothing to do with the type of workload. In addition, we notice that the server power at full load of all the samples of Xeon E5-2603v2 are below the TDP (80W, refer to Table 3.1). As we already mentioned in 3.2, TDP is actually a cooling system design requirement. It has no direct reflect to the actual working power of CPU, it can be lower or higher than actual CPU power [And13].

The first penitential cause came to our mind is the differences of CPU temperatures between samples. However, the variation of CPU temperature can not explain the power variation: there is no obvious correlation between CPU temperature and power consumption. Otherwise, processor samples from Xeon E5-2603v2 have wider CPU temperature distribution than Xeon E5345 during the whole test. As can be seen in the box plots 3.3 and 3.5, CPU temperatures of processor samples from Xeon E5-2603v2 varies within an interval. While for processor samples from Xeon E5345, CPU temperatures are almost stable at a constant for each sample.

Moreover, we evaluate as well the performance variation among processor samples. The performances of samples are represented as the execution time for completing the same task (defined by workload). The results show that, little performance variations are observed for both Xeon E5-2603v2 and Xeon E5345, between the shortest and longest execution time, the performance differences are 0.3% and 0.6% respectively. We find the results normal, since all the samples operate at the same frequency, there is neither influence coming from Turbo Boost nor Multi-Threading technologies. According to previous studies, operating frequency is responsible for most of the performance variation. The results indicate also that, the power variation observed before has no relation with the performance variation neither.

In the following research, we try to find the reasons behind the power variation between samples, especially from perspective of thermal characteristics.

3.3 The differences of thermal characteristics behind identical processors

Results of the experiments in previous part tell that, in this case, processors from newer generation have much more power variation than the ones from older generation. CPU temperature distributions along the test cannot perfectly explain the power variation. We confirm as well that the power variation is not introduced by the performance variation. Power and CPU temperature do not show correlation, although the different range of CPU temperature distribution between the two processors still got our attention. In this situation, we try to identify

the source of variability by investigating the differences of thermal features between samples. The power required by processor in a server system can be divided into two parts: IT load and associated cooling system (sink + fan). With the increase of IT load, processor consumes more energy to meet the service requirement and in the same time dissipates more heat into the air. If the heat is not evacuated in time by the cooling system, the temperature surrounding processor becomes higher and leads to the rise of leakage current, which will in reverse increase the power of CPU [KC09] [MB09]. However, there is doubt about whether the processor samples have the same behavior in generating and rejecting heat. We are wondering if the differences of thermal feature can also lead to the power variation among processors under the same load.

In this section, we propose two hypotheses regarding different thermal features between different samples of Xeon E5-2603v2. In subsection 3.3.1, we study the influence of TIM (Thermal Interface Material), in order to verify if the new PTIM (Polymer Thermal Interface Material) can lead to the different thermal characteristics between identical samples. In subsection 3.3.2, we investigate the parameters of leakage current among processors (presented by static power). We start from the mathematical formula of static power and analyses the relationship between the leakage current and the temperature. Fans, DC power generator and climatic chambers are provided to help controlling precisely the thermal conditions.

3.3.1 Hypothesis 1 : Thermal Interface Material (TIM) applications

Thermal conductivity of a material describes its ability to transfer heat, measured in watts per metre kelvin (W/mK). Higher the thermal conductivity, higher the ability to transfer heat. CPU is an electrical component with high density of transistor and circuits, lots amount of heat can be generated quickly in a concentrated small area, that's why we need specific structure to help dissipating heat of a working CPU. Figure 3.6 demonstrates a simplified structure of the thermal package design of a CPU in a motherboard. In such a structure, heat is generated within the silicon die area. Integrated Heat Spreader (IHS) is a mental made component served as transferring the heat from silicon die to CPU cooler (fan or liquid cube, refer to "liquid-cooled system" in section 1.2.5), so as to dissipate heat quickly. The problem is, silicon die and IHS are made by different material, sometimes the micro air gap between them is inevitable. However, air is a poor conductor of heat with very low thermal conductivity value of 0.026 W/mK (at $25°$C). While for metal such as copper, the value can reach 384.1 W/mK (at $18°$C). Therefore, air gap reduce effective contact area between silicon die and IHS, which affects dramatically the heat transfer efficiency. TIM is therefore applied to fill the air gap between IHS and silicon die, so as to facilitate heat exchange between them [Rac17] [Pra06].

A thermal model proposed by Huang W et al. [HHSS05] shows that, the thickness variation of TIM can affect a lot silicon die temperature distributions across processors. Since Ivy Bridge processor generation, Intel decides to apply Polymer TIM (PTIM) to the processors as a replacement of Solder TIM (STIM), for reasons of economics, environment and better cooling performance [RMPV06].

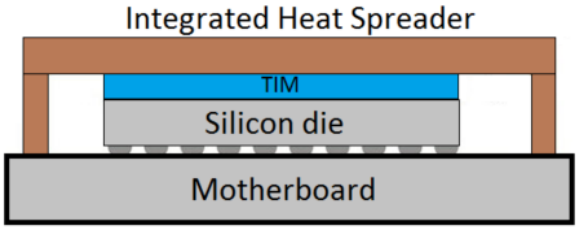


Figure 3.6 – Processor thermal package structure

In the previous experiment, Xeon E2603v2 applied with PTIM is observed to have more power variation than Xeon E5345 applied with STIM. Type of TIM could be one of the suspects leading to the power variation. If the hypothesis is correct, then the removal of TIM from the original package of processor will effectively reduce the power variation extent between samples. Therefore, we try to recognize the possible influence brought by TIM by removing PTIM from the Xeon E2603v2 and re-perform the same test. TIM is hidden behind IHS from exposure directly to outside, that challenges a lot the manipulation: we have to remove TIM along with the entire IHS. Moreover, this manipulation is irreversible and can result in permanent damage to the processor. We start the verification with a small quantity of samples, in order to reserve enough samples for further studies. Three samples consuming less: P14, P23, P9 and two samples consuming more P30, P20 are chosen in this study (refer to figure 3.2 for consumption information). Figure 3.7 shows the example of the processor sample after the removal of TIM, the right part shown in the figure has been removed. After the manipulation, the processor's silicon die (lower half in the figure 3.7) is cooled directly with the heat spreader of fan as illustrated by Figure 3.8). We applied equal amount of heat paste between silicon die and heat spreader to facilitate the heat exchange between them.

Then we re-perform the test as described in section 3.2 and re-analyze the power variation. Table 3.2 shows the results of power consumption and average CPU temperature in two circumstances: with and without PTIM. Unfortunately, removing the PTIM does not help reducing the power variation extent between samples. After removing PTIM, most of the samples (except P9) just have little power variation as well as with PTIM. Therefore, we think that PTIM is not the major source leading to the power variation between identical processor samples.

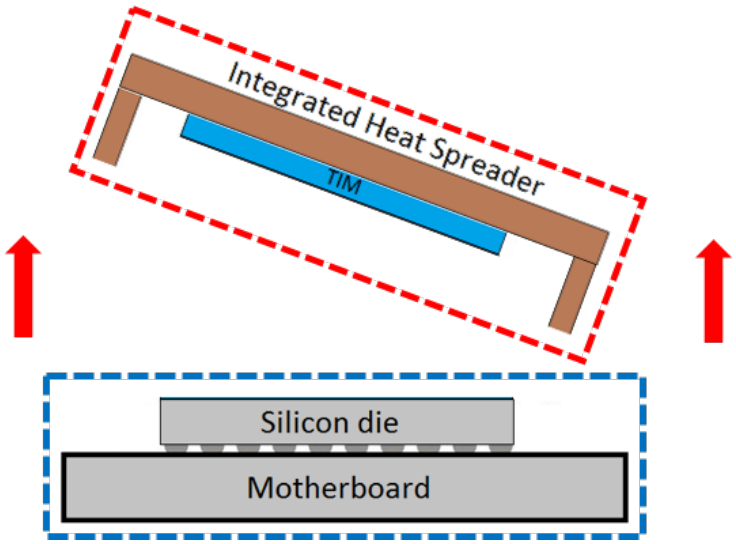


Figure 3.7 – Remove the TIM from the processor [Har19]

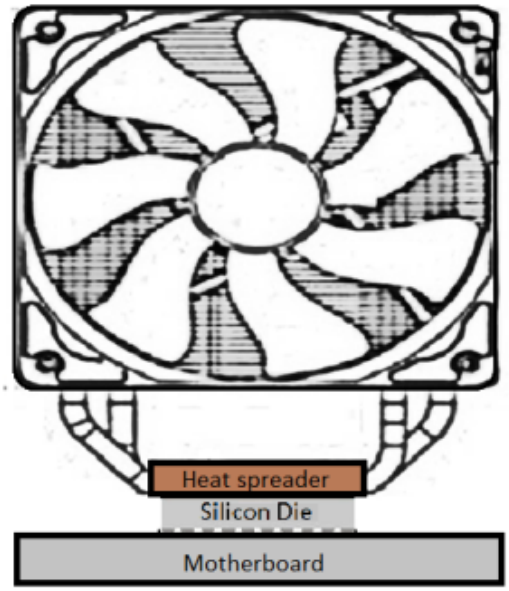


Figure 3.8 – TIM remove schematic diagram

Table 3.2 – Power and CPU temperature for Xeon E5-2609V2: with and without PTIM (workload: pi_calculator)

Processors	With PTIM		Without PTIM		Δ (Watt & °C)	
	P(W)	T(°C)	P(W)	T(°C)	Δ P(W)	Δ T(°C)
P14	50.0	38.2	50.3	40.0	+0.3	+1.8
P23	50.6	40.3	51.0	36.0	+0.4	-4.3
P9	50.6	39.5	56.4	36.7	+5.8	-2.8
P30	63.9	42.9	63.3	38.9	-0.6	-4
P20	66.8	42.7	68.3	47.7	+1.5	+4.7

3.3.2 Hypothesis 2 : The parameter variation of static power

The overall power dissipation in today's microprocessors is mainly composed by two sources: dynamic power and static power. Other power loss such as short circuit occurs at whenever gate switch is relative small, it can be absorbed by dynamic power [KAB+03]. In this sub-section, we review firstly the CMOS technology as theoretical inspiration for our hypothesis. Then we present the details of the investigation and analysis of the test.

CMOS Technology review

Dynamic power comes from charging and discharging the processor's capacity loads. It can be described by the equation (3.1) [STD94] [Eti18]:

$$P_{dynamic} = \alpha \sum C_i f V_{dd}^2 \tag{3.1}$$

In this equation, f is the switching activity (operating frequency), $\sum C_i$ is the sum of gate and interconnection capacitance, V_{dd} is the supply voltage and α is the activity factor of the overall circuit. Static power is the product of voltage supply and leakage current. There are different kinds of leakage modes in MOS transistor, and the most dominant leakage mechanism is sub-threshold leakage I_{Dsub} [KRH97]. I_{Dsub} is the current flow between source and drain at off-state. Off-state current becomes now a limitation factor for down-scaling the threshold voltage, since it determines the power consumption of a chip in its idle state. Therefore, the static power dissipation representation can be simplified and described by equation (3.2). I_{Dsub} can be described by equation (3.3), according to previous studies [HIG94] [BS00].

$$P_{static} = V_{dd} I_{Dsub} \tag{3.2}$$

$$I_{Dsub} = e^{\frac{-qV_{th}}{akT}} \tag{3.3}$$

where q, k, a and k are physical related constants, T is the absolute temperature and V_{th} is the threshold voltage of the transistor, which sits between ground and the supply voltage.

In earlier years, traditional low-power microprocessor design focus mainly on reducing dynamic power consumption. At that time, static power consumption is not a limitation and is negligible compared to dynamic power [STD94] [HIG94]. In pursuing higher performance and lower consumption, CMOS technologies scaled the chips from generation to generation by following Moore's law "The number of transistors and resistors on a chip doubles every 24 months" [Moo98]. The chips then have more transistors, denser CMOS circuitry and smaller dimension. In 2007, in applying 45nm process technology, there were 3.3 million transistors per square millimeter (MTr/ mm2) for chips of Intel. Ten years after, in 2017, Intel announced its latest chip generation: 10nm technology, with density as high as 100.8 MTr/ mm2, which is 30 times denser than in 2007 [Rac17]. Note that, the designation like "45nm", "10nm" refers to commercial name for certain lithography process technology. The number doesn't indicate the size of any particular feature of the chip and can vary significantly between manufactures [Nod16]. As a general rule, the smaller the number is, the denser the circuitry becomes.

As the chip dimension scaled, for the purpose of reliability required by constant field scaling, supply voltage V_{dd} has to be decreased by the same factor for chips to keep the electric fields the same across different generations [DGR+74]. That brings additional benefit of dynamic power saving as suggested by formula (3.1). The threshold voltage (V_{th}), has also to be scaled down along with the V_{dd} in order to avoid performance degradation as shown by formula (3.4) [SN90].

$$Delay = \frac{1}{f} = \frac{C_{gate}V_{dd}}{I_{Dsat}} \propto \frac{V_{dd}}{(V_{dd} - V_{th})^{1.3}} \tag{3.4}$$

Yet, as a result, the I_{Dsub} increases exponentially with the V_{th} decreases [KAB+03] [ZBS+04] according to formula (3.4). Therefore, in modern processors, static power takes greater and greater part in consumption and becomes increasingly dominant. Moreover, advanced lithography process with thinner xnm may have more fabrication deviations to impact the parameter values of leakage current among identical processor samples, then leads to eventually power consumption variation. Comparing the two processors, the newer generation E5-2603v2 with 22nm technology has 10 times more power variation than E5345 with 65nm technology. This provides a great opportunity to empirically evaluate the impact of parameter variation brought by CMOS technology evolution on the processors variability.

However, there is no way to measure directly the leak current of the processor by physical device, without "opening" some key components hidden and protected by Intel packing technology. Such manipulation is too risky and may cause permanent damage to the motherboard. The only measure of consumption accessible is the whole consumption of server composed of consumption of CPU (P_{cpu}) and of the other components including motherboard (P_{chip}). The processor samples are switched one by one in the same motherboard, P_{chip} remains the same, the only item can vary the consumption of server is P_{cpu}. As discussed before, P_{cpu} is composed by dynamic and static power. It can be seen from the equation (3.1) (3.2) and (3.3) that, dynamic power is frequency dependent value but independent from temperature variation. On the contrary to dynamic power, static power is temperature sensitive value, but cannot be affected by frequency scaling. The consumption of the whole server can be then simply represented by the equation (3.5) and (3.6):

$$P_{server}(V_{dd}, T_{cpu}, f) = P_{chip} + P_{cpu}(V_{dd}, T_{cpu}, f) \tag{3.5}$$

$$P_{cpu}(V_{dd}, T_{cpu}, f) = P_{dynamic}(V_{dd}, f) + P_{static}(V_{dd}, T_{cpu}) \tag{3.6}$$

Our test platform has no support in BIOS to regulate manually V_{dd}. f is governed by frequency governor "performance" and adjusted by frequency driver `intel_pstate`. f varies according to system current load. In this case, the data obtained is too limited to separate and quantify the two sources of consumption by statistical analyses. Deriving the models of static and dynamic power of processor is interesting, but it is beyond the scope of the study discussed here. If interested, Goel et al [GM16] present a systematic methodology for modeling static and dynamic power consumption of individual cores and uncore components in their work. In our cases, for a given processor sample, we fix f and V_{dd}, and suppose that the value of $P_{dynamic}$ is constant and independent from temperature variation. The parameters of leakage current as expressed in equation (3.2) and (3.3) can be simply identified by varying the T_{cpu} whiling keeping the f and V_{dd} unchanged.

Experiment setup

Same platform as described in section 3.2.1 is adopted. CPU intensive benchmark `cpuburn` [Rob11] was used to stress the SUT in this test. The benchmark is usually used to maximize the heat dissipation of processor. It can active and stress every thread available in the processor at 100% utilization. During the execution of `cpuburn`, operating frequency (f) is maintained

constantly at maximum (1.8GHz). Besides ambient and CPU temperature, we monitor as well the CPU frequency (provided by python library `psutil` [Psu09]). Climatic chamber is used to vary the ambient temperature. However, as presented in section 3.3.1, five processors: P14, P23, P9, P30 and P20 have been used to validate the first hypotheses, their TIM and HIS have been removed. We worried that, this manipulation may change their original features and introduce unexpected influence to the following experiment. Therefore, they are not included in this test. Other four samples consume less: P12, P21, P6 and P4, as well as the four samples consume more: P3, P10, P5 and P8 are chosen to validate the second hypotheses.

For each sample, we repeat the test procedures as follows. Figure 3.9 illustrates the whole procedure.

1. Keep the server on idle state for 3 minutes at 22°C ambient temperature.
2. Execute `cpuburn` for 30 minutes. Along the test, configure the climatic chamber to increase the ambient temperature from 22°C to 50°C.
3. After the execution of cpuburn, keep the server on state idle for 3 minutes at 50°C ambient temperature.
4. Shutdown the server, switch for the next processor sample and wait for the whole platform cool down to 22°C ambient temperature.

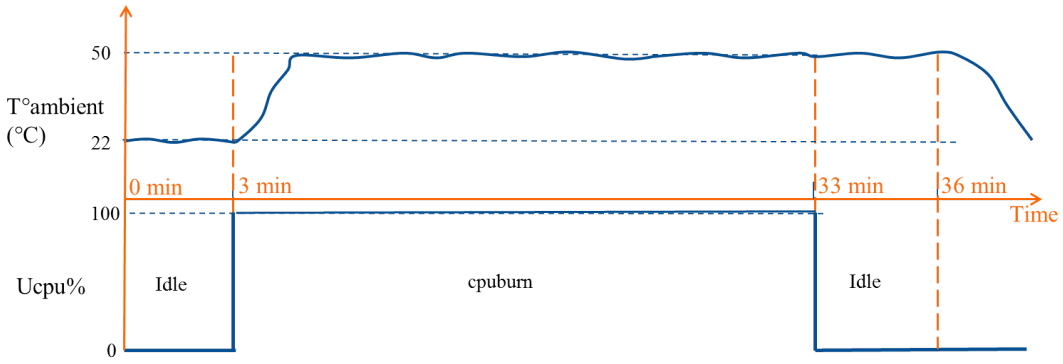


Figure 3.9 – Leakage current variation investigation: experiment illustration

Experiment results and analyzes

On high load state High load state occurs at the second step, where server is stressed by cpuburn. Figure 3.10 shows the evolution of CPU temperatures over time, with the ambient temperature increases from 22°C to 50°C, as illustrated by figure 3.9. It can be observed that,

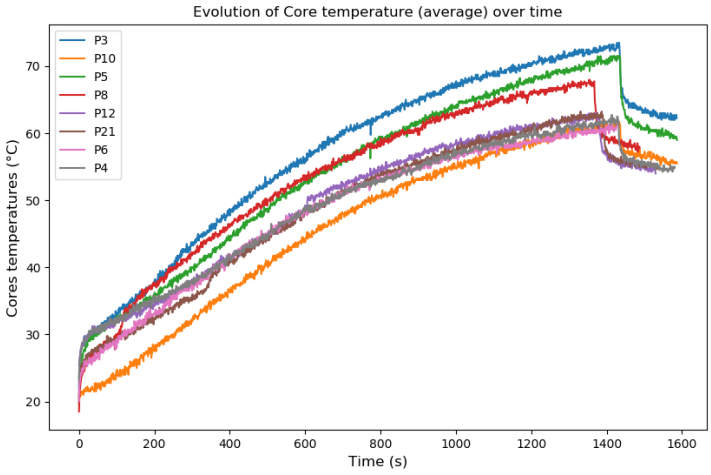


Figure 3.10 – CPU temperature evolutions on function of time

the change of CPU temperature between different samples differs from each other. Under the same load and the same ambient temperature, CPU temperature increases gradually with the rise of ambient temperature but up to different maximum values. For example, CPU temperature of P3 can be increased to more than $70°C$ while P10 just reaches $60°C$. Figure 3.11 shows the value of server power during high load while varying CPU temperatures. Processor samples are marked by different colors, the points represent for the real measurement data. Not two of the samples have the same Thermal behaviors. Moreover, we were able to fit the server power and temperature data with a mathematical function as shown by equation 3.7. The equation has an exponential form of $exp(-k/T)$ which links to the static power dissipation formulas based on temperature as described before in equations 3.2 and 3.3. We calculate the Mean Absolut Percentage Error (MAPE) (refer to "Model evaluation metric" in section 1.4.2) to evaluate the fitting quality, and MAPE is less than 0.4% for all the samples.

$$P_{server}(T) = a + b * e(\frac{-k}{T}) \tag{3.7}$$

In this equation, P represents for total server power in Watt, P varies according to CPU temperature T in Celsius degree. Comparing this equation with equation 3.5 and 3.6, intercept a actually contains P_{chip}, $P_{dynamic}$ at high load and P_{static} at temperature $20°C$. Parameters b and k represent leakage current parameters. It can be seen from the fitting equations shown in figure 3.11 that, the variation of leakage current parameters between samples, result in different power change rates of server power while varying CPU temperature. Samples consume the most (P3,

P10, P5 and P8) in the first experiment 3.2 also have higher power change rates than samples consume the least (P12, P21, P6, P4). In another word, processor samples P3, P10 P5 and P8 dissipate heat at higher rate under the equal test conditions (same load and ambient temperature is varied in the same way). Power change rates among samples have been illustrated separately in Figure 3.12, where we present only the exponential part of the power function, and for a wider range.

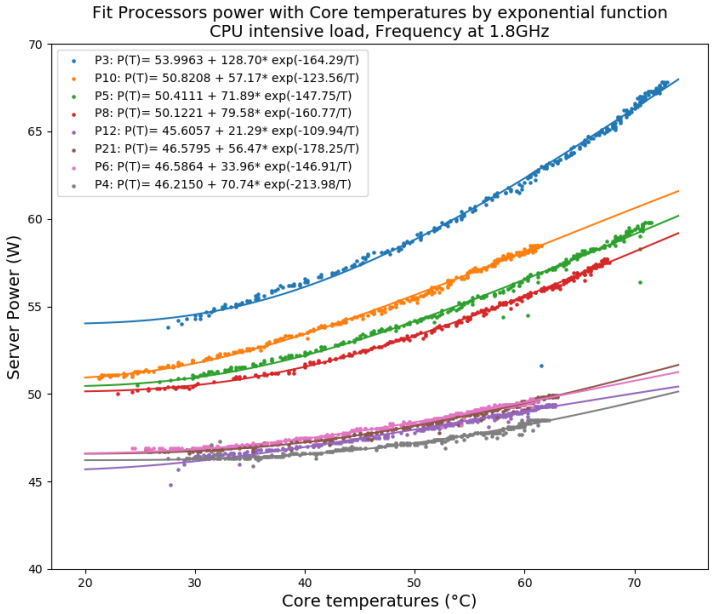


Figure 3.11 – Relationship of CPU temperature and Server power for different processor samples with execution of cpuburn

On idle state On idle state, there is no workload, only OS keeps running in the server. Server power on idle state installed with different processor samples are compared at $22°C$ and $50°C$ ambient temperature respectively. Results are shown in Figure 3.13. We can observe that, power variation between samples on idle state is greater at $50°C$. Moreover, samples belong to higher consumption group (P3, P10,P5 and P8) have bigger power increment when ambient temperature passes from $22°C$ to $50°C$. Server power installed with P3 has a rise of 6.9 Watt. While with P4, total power increases only by 0.7 Watt.

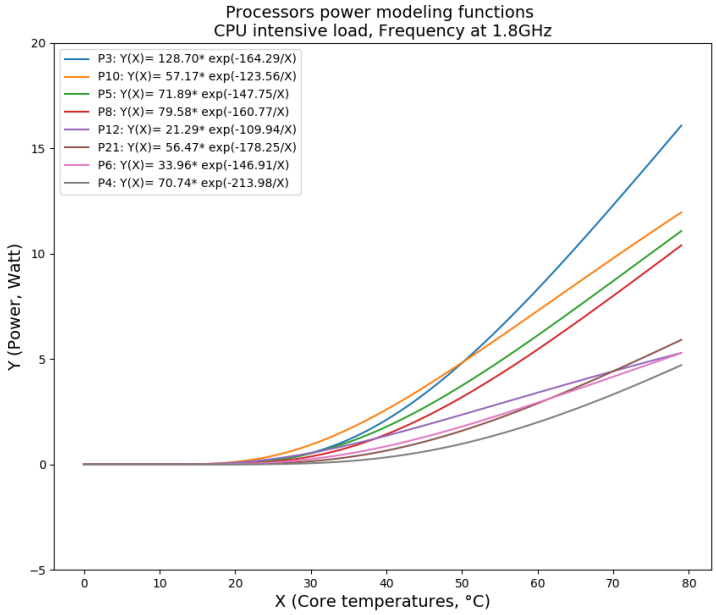


Figure 3.12 – Illustrations of Server Powers functions for different processor samples

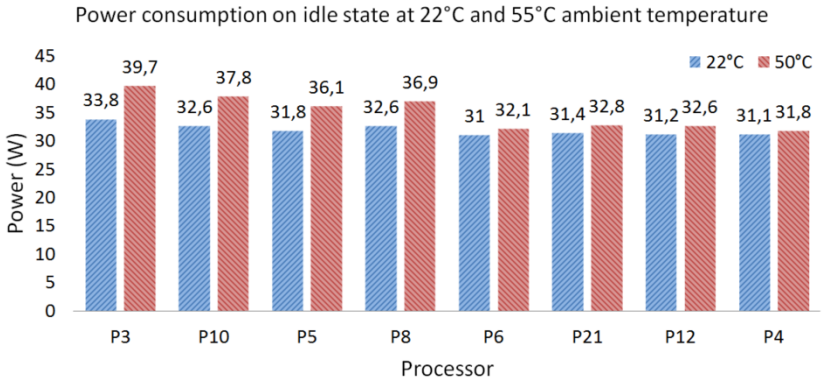


Figure 3.13 – Server power on idle state at 22°C and 50°C ambient temperatures

3.4 Conclusions

Processors are becoming smaller, more powerful and less consuming from generation to generation. On the other side, processors become more complex than ever. Leakage current variations arise from imperfections in the fabrication process among modern processors, such as lithographic length aberration. According to our observations, leakage current variations have impact on cooling ability, samples dispose of different increase rate with the rise of temperature.

With the decrease of lithography size for modern processors, technology today can hardly control precisely the variability between processors in fabrication process. Apart from the performance variability brought by the operating frequency variation, our experimental results show surprisingly high variability of power consumption within modern processor. 30% power consumption variation is observed among 30 identical processor samples, which has no correlation with the performance (frequency variation). In the first time, we confirm and characterize by means of physical experiments, the power consumption variation introduced by leakage current parameter variation. Samples can have different heat dissipation under the supposed same conditions (ambient temperature and load level), which results in different leakage current distributions then finally affect the static power consumption among samples. Moreover, fabrication variability has random effects on processor samples and cannot be detected by OS. Our demonstration highlights the challenges of modeling techniques posed by the processors variability. The finding highlight that, same servers in the production line are not created equal. The variability between identical servers in homogeneous clusters do exist, and the difference should not be ignore in terms of modeling their power consumption under equal test conditions. As a result, the precision of power models presented in previous studies could be questionable when applying to the other identical SUTs. The variability issue from imperfection of fabrication processing can not be avoided so far. Therefore, power modeling techniques based on mathematical analysis should respect the potential errors that may introduce from the variability between identical samples, especially for modern ones.

In addition, the findings present in this chapter can be also applied to optimize the energy management strategies in data center. By developing a model allows identifying static and dynamic power during an execution of an application, we could improve the performance of current optimization algorithms. Processor samples with worse thermal features will not only have more static power, but also require more energy for cooling, same for the servers. It worth the efforts to identify the servers with better and worse thermal features in a data center, so as to better orient energy management strategies.

CHAPTER 4

POWER CHARACTERIZATION APPROACHES FOR SERVERS

Reliable power characterization approaches are essential for supporting energy-aware solutions. High accuracy power analyzers are capable of providing accurate power consumption data. However, limitations like high cost, weak support of integration with computing systems make it unrealistic to use power analyzers in a data center environment. Recently, some alternative economical solutions are adopted widely. Such as embedded power meters (refer to 1.3.3) and power models (refer to 1.4). The reliability of these solutions need to be evaluated thoughtfully before applying to different real situations. In this study, we investigate experimentally the following approaches: IPMI, Redfish, Intelligent PDU and power models.

4.1 Context and objectives

Power consumption characterization of servers is an essential part to achieve energy-aware adaption strategies in a data center environment. Physical instruments such as wattmeter and power analyzer can get accurate measurements for electrical devices. However, in a data center environment, it costs too much. Moreover, for optimization requirements, estimating the power consumption for further use cases is demanded as well, and power meters cannot help. Recently, economical alternative power characterization approaches, include both hardware and software solutions are applied widely in data center environments. Popular hardware approaches include: a) Intelligent power distribution units (PDU); b) Standard specifications that provide interface with integrated sensors, such as Intelligent Platform Management Interface (IPMI) and Redfish. Besides, software solution like power models, provide a device less solution to get power data based on system activities. Moreover, they can be part of the optimization strategies of data centers, by estimating the power consumption in further use cases. Despite of the wide employments of the above solutions mentioned, as far as we know, few works addressing the data precision obtained through these approaches.

In this study, we try to fill up this missing part by evaluating the data precision of these economical power characterization approaches.

The chapter is organized as follows: in Section 4.2 we evaluate the precision of Redfish and IPMI applied in a new series of IBM servers. The measures of power consumption recorded by Redfish and IPMI are compared with a high-accuracy power analyzer. Similar evaluations has been done to IPMI and Intelligent PDU for Dell Poweredge servers in a cluster. Then, in Section 4.3, we conduct a deep research to the power models based on CPU utilization. The source of inaccuracy has been discussed and the solutions are proposed to improve the accuracy. Conclusion is given in section 4.4.

4.2 IPMI, Redfish and Intelligent PDU: power data precision evaluations

The employment of PDU and embedded power meters are discussed in section 1.3.2 and 1.3.3. Apart from getting power values, IPMI, Redfish and Intelligent PDU also usually take participation to realize the power management for the entire data center, there are several use cases in the real world: [Inc] [Tec] [Int15b] [Len19]. As discussed in state of the art, their concepts make us curious about their actual abilities of power characterizations.

In this section, we present our comprehensive evaluation of these power characterization solutions.

4.2.1 Evaluation experimental setup

Basic idea of the evaluation is to compare their readings in real time with a high accuracy power analyzer.

The evaluations have been conducted separately on two different servers: a prototype server from Lenovo Skylake series, installed in a cluster at Orange Labs in Rennes; and a Dell PowerEdge R630 server, installed in a cluster at univsersity IMT Atlantique in Nantes. The precision of IPMI, Redfish and Intelligent PUD have been obtained and evaluated as described below:

— IPMI precision is evaluated on both the two servers, as they all have equipped with IPMI technology. We use an open source API tool freeipmi [Tea] to get the power readings from IPMI. In addition, Dell propose Original Equipment Manufacturer (OEM) specific IPMI commands. For certain Dell servers supported this function, users are able to read instantaneous power consumption data through " get-instantaneous-power-

consumption-data" Dell OEM IPMI command [Sys12]. This option is proposed by server manufacture to provide additional functionality in their product. It is realized by adding hardware extensions (like specific sensors) on the motherboard. This function is available for our Dell server, gives us the opportunity to evaluate the precision of IPMI-oem function as well.

— Redfish precision is evaluated on Lenovo server. The prototype server integrated the newest `XClarity Controller` - a software server management solution proposed by Lenovo. The software controller includes a Redfish REST API to get server power in real time by the latest Redfish technology.

— Intelligent PDU is evaluated on a Dell server belonging to "ecotype" cluster (refer to 2.2.1 for more details).

Server is stressed with the test suite SERT as mentioned in "Server Efficiency Rating Tool" in section 1.1.2, with a total execution time of about 2 hours. The SUTs have both double power supplies design, the measurement diagram is shown in figure 4.1. High accuracy power analyzer Yokogawa WT330 is placed between server power supply unit and wall plug to measure and record power consumption data as the evaluation reference (with maximum measurement error less than 1%). A program has been developed as software collector. The software collector is running in the server while executing the test, serves as synchronizing readings from several channels to local database based on Network Time Protocol (NTP). Sampling frequency is set to 2 Hz.

4.2.2 Experimental results and analysis

We evaluate the power measurement precision of IPMI, Redfish and PDU by calculating their MAPEs (refer to "Model evaluation metric" in section 1.4.2) with the power measurement taken by Yokogawa WT330 - a high accurate power analyzer. The evaluations of IPMI and Redfish performed on Lenono server is presented in figure 4.2. The evaluation of IPMI and Intelligent PDU on Dell server is presented in figure 4.3.

The measurements taken by power analyzer is marked with green points in the two figures. MAPE of IPMI and Redfish in Lenovo server are 3.7% and 2.9%. As demonstrated by figure 4.2, IPMI and Redfish are less capable to capture server power when the power varies at a high frequency. We find that, the sampling frequency of power data collected from Redfish and IPMI is not as frequent as power analyser. However we didn't find any document specifying their sampling frequencies. We have then configured the sampling frequency of power analyser to 50 ms, and compare the data with the power measurements taken per second from IPMI and

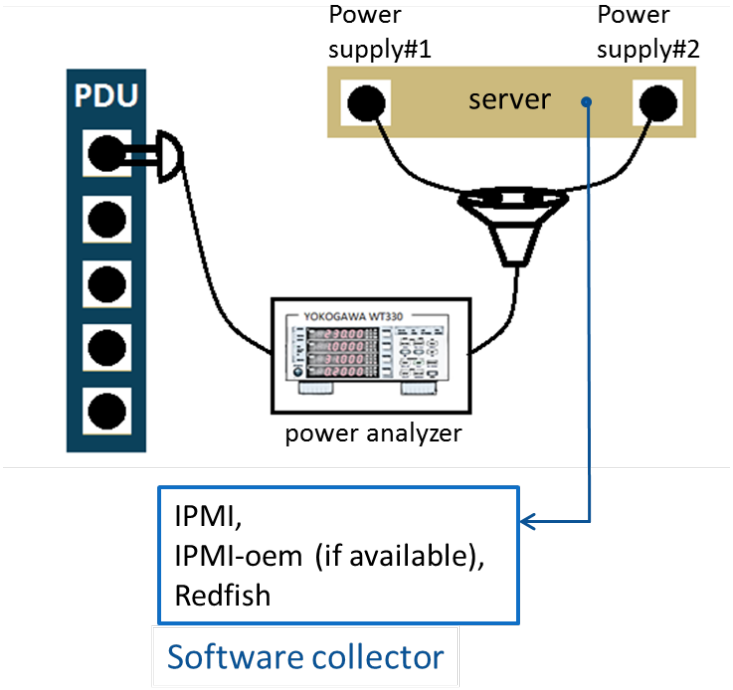


Figure 4.1 – Measurement diagram: evaluate power measurement precision for IPMI, Redfish and Intelligent PDU

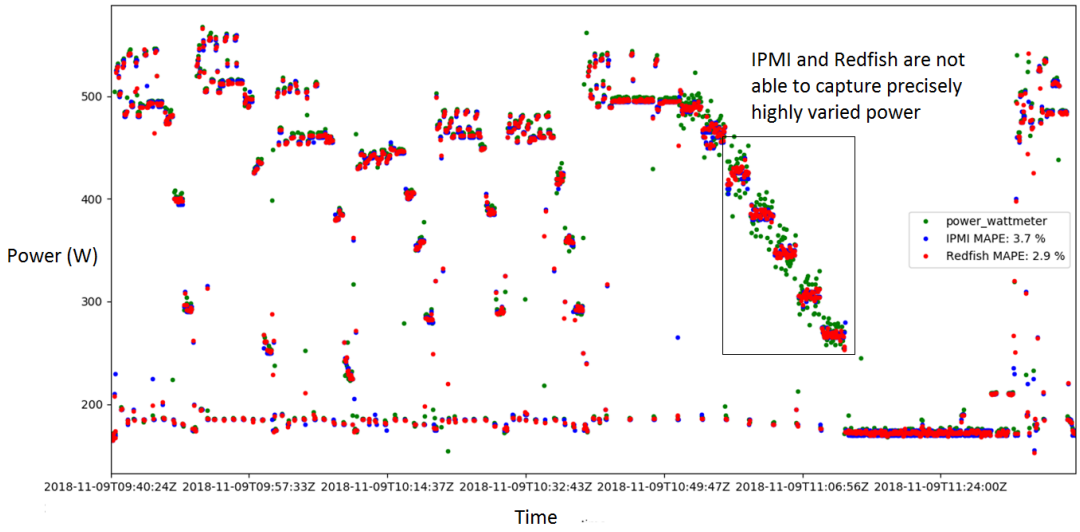


Figure 4.2 – Evaluating precision of IPMI and Redfish on Lenovo server

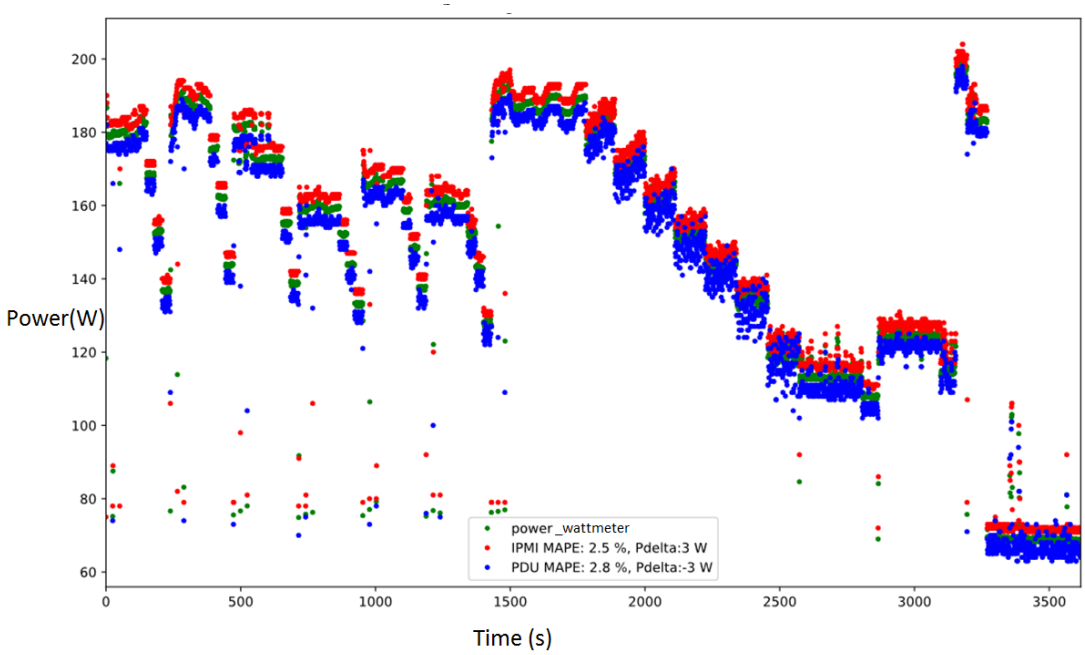


Figure 4.3 – Evaluating precision of IPMI and Intelligent PDU on Dell server

Redfish. A roughly sampling frequency of 200 ms is identified. The power data obtained from IPMI and Redfish could be less accurate when power varies in a high frequency (higher than 5Hz).

We calculated later the MAPEs at three server power ranges with the same data: below 200 Watt, from 200 to 400 Watt and above 400 Watt. Results in table 4.1 show that, the precision has worse results at lower power range. In fact, power consumption is observed to have more variations at lower system load (lower power). Unfortunately, we didn't find concrete explanations. According to our knowledge, the frequent power variation at low load may due to the execution of some low proprieties system processes. Since during unoccupied periods, servers have more available resources, some system-related background processes are likely to be awake and executed.

Table 4.1 – MAPE of Redfish & IPMI between different power ranges

Power Ranges (Watt)	*Redfish MAPEs (%)*	*IPMI MAPEs (%)*
From 0 to 199	4.1	5.5
From 200 to 399	4.0	4.4
From 400 to 600	1.8	2.3
From 0 to 600	2.9	3.7

Table 4.2 – Summary of evaluations for the hardware power characterization solutions

Solution(server)	Max MAPE (%)	Limitations
IPMI (Lenovo) IPMI (Dell)	5.5% 2.8%	Less accurate for highly varied and weak values; The offset could exist, calibration is needed
IPMI-oem (Dell)	1.1%	Available only for certain models
RedFish (Lenovo)	4.1%	Less accurate for highly varied and weak values;
Intelligent PDU (Dell)	2.8%	Less accurate for highly varied values The offset could exist, calibration is needed

Figure 4.3 show the precision evaluation results of IPMI and Intelligent in a Dell server. In this evaluation, we find that, other than inaccuracy brought by sampling frequency. Power measurements obtained from IPMI and Intelligent PDU can both have the problem of calibration. As shown by the figure, instant power values have a stable difference of +3Watt and -3Watt for IPMI and PDU respectively. MAPE can be improved to less than 1.5% after correcting the calibration. The evaluation of precision for Dell OEM IPMI command has been completed with an other test by following the same procedure as described before. The specific OEM IPMI command "get-instantaneous-power-consumption-data" in the tested server can provide high accuracy power consumption data with MAPE value of "1.1%". This command is available for certain Dell server models.

Finally, in table 4.2, we conclude the precision and possible problems could be encountered during the usage of these tools.

The work presented in this part expected to give references in choosing hardware power consumption characterization solutions in a data center environment. We suggest that, before applying IPMI, Redfish and Intelligent PDU in a data center environment, for the purposes like power management and optimization, it worth checking firstly the precision and calibration of these tools and make sure that measurement is accurate enough for supporting decision makings.

4.3 Power models based on CPU-Utilization

Power meters are able to provide real-time power consumption data by measuring. Hardware investment is indispensable. IPMI, Redfish are available only on modern servers. Equip a data center with intelligent PDU requires higher budget. Basic idea of power model is to correlate power consumption with some system related data through mathematical analysis. Power

models is a software solution providing hardware free power characterization approaches. Besides financial benefits, sometimes, we are will to know the power consumption situation at a future time point for optimization requirements. Power models can be used also to estimate the power consumption according to system activities occurred in the future. Power models can be built with different system related data. In this study, we evaluate power model of servers based on CPU utilization, since it is one of the most classical power models. This type of model has particular advantages comparing to the others. It is easy to deploy to all kinds of servers regardless of the server architecture, processor types, or providers. However, beyond these advantages, the accuracy of model remains as a big problem, since only one indicator is used to build the model.

In this section, we discuss the power model based on CPU utilization by exploring experimentally two questions:

— Firstly, in reality, how power spreads out for a fixed usage value?

— Secondly, is power linear to usage for a given CPU frequency range?

At the end of the evaluations, we propose two ways to improve the model accuracy: by applying polynomial regression function and by adding ambient temperature data into the model.

4.3.1 Question 1: How power spreads out for a fixed CPU utilization value?

Power model is nothing but a math function. Single utilization value corresponding to single estimated power value. In reality, power varies for a given utilization value, since resources in server can be stressed in different ways according to service requirements. The distribution of power values for a given utilization value affects directly the model accuracy. In this section, we stress the SUT by executing different kinds of workload and see how largely can power spread for the same CPU utilization.

Experiment setup and methodology The experiment is performed on a Gigabyte mw50-sv0 server, equipped with one Xeon E5-2609v3 processor. Different types of workload from SERT [LT11] (refer to "Server Efficiency Rating Tool (SERT)" in 1.1.2 for more information) are chosen, in order to stress SUT in different ways. The test suit includes six CPU-intensive workloads (`Compress`, `CryptoAES`, `LU`, `SOR`, `Sort` and `SHA256`); one CPU and memory hybrid workloads (`SSJ`); two memory-intensive workloads (`Flood3` and `Capacity3`). Consumption at idle state is also measured. Details about the worklets used can be found in Table 4.3 [Com13].

Table 4.3 – Test Suite Information

Worklet	Components	Description	Load Levels
Compress	CPU	Compress and decompress data	100%, 75%, 50%, 25%
CryptoAES	CPU	Encrypts and decrypts data	100%, 75%, 50%, 25%
LU	CPU	LU factorization of dense matrix operations	100%, 75%, 50%, 25%
SOR	CPU	Jacobi Successive Over-relaxation workload	100%, 75%, 50%, 25%
Sort	CPU	Sorts randomized 64-bit integer array	100%, 75%, 50%, 25%
SHA256	CPU	SHA256 hashing transformation and encryption/decryption	100%, 75%, 50%, 25%
SSJ	CPU/Cache/Memory	simulates Online Transaction Processing (OLTP) operations	100%, 87.5% ... 12.5%
Flood	Memory	Measures memory bandwidth across arrays	Full, Half
Capacity	Memory	Exercises Java's XML Validation	Base, Max
Idle	System	No load on SUT	None

The power consumption data is collected by Yokogawa WT330, a high-accuracy power analyzer, with maximum measurement error less than 1%. CPU utilization information is collected by redirecting the information from linux system directory "/proc/cpuinfo". Data sampling frequency is set at 1Hz. At the end of test, box plot from python library `matplotlib` [Hun07] is used to interpreted the dispersion of power for a given CPU utilization. Box plot is widely used for displaying statistic distribution, a simplified manner in comparison to a histogram or density plot. For a normal distribution, 50% of the data is within the box. Two short lines beyond the box represent for the minimum and maximum values within 99.3% of the data. Outliers represent for the remaining 0.7% data.

Evaluation results The whole test suite mentioned above is executed on SUT for more than one hour, the dispersion of instant power at different CPU utilization is shown by box plots in figure 4.4. In general, server power is supposed to increase with the rise of load levels (CPU utilization). However, we are surprised to see that, the power distributes in a considerable range at a certain CPU utilization. For instance, when CPU utilization of server is at 20%, the power can vary between 47 Watt and 57 Watt (99.3% of the data). Sometimes, the power is even higher than the CPU utilization at 30% and 40%. In fact, even through the workloads occupy same CPU time, they make use of the other hardware resources of the server in a different way: Cache and memory access, I/O operations, network resources, etc. The situation is worse when server run at full load. Power ranges in a large distribution from about 62 Watt to 78 Watt. There are also lots of power outliers (represent by the small circles) at full load. Actually, system spent time in activating electrical components when current load change dramatically from a lower level to highest level, the period is very short but still account for about 0.7% of the total power data. The power ranges are relatively lower at 10%, 50%, 70% and 90%, due to a lack of experimental data at these load levels (refer to "load levels" in table 2.2).

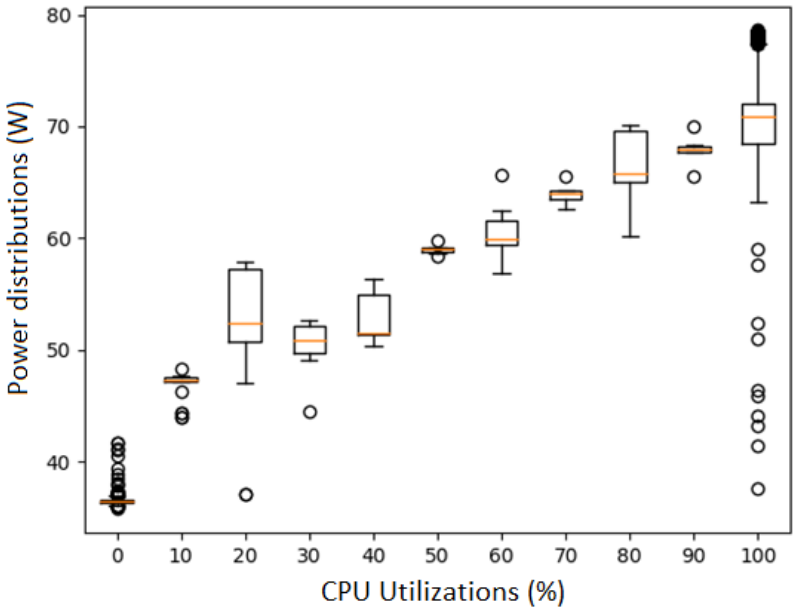


Figure 4.4 – Distribution of power under different CPU-utilization.

Different ways in using hardware resources in a server is one of the major explanations for the power distribution at certain CPU utilization. In order to understand the differences of resource usage between workloads, we show the instant power measurement data based on CPU utilization in figure 4.5, the measurement points have been classified by workloads with different colors. The result illustrate the phenomenon: even for workloads with the same type "CPU intensive", the relation between power and CPU utilization can be represented with different linear functions. For instance, comparing with the other CPU intensive workloads, while executing worklet `LU` (green points in the figure), server power increases at higher rate with the rise of CPU utilization. Still, we can also see that: even though server power spreads a lot at one certain utilization level while executing different workloads, the power distribution of one workload at one CPU utilization is relatively much smaller, the power model precision could be increased by workload classification technology.

Therefore, the power model based on CPU utilization could be optimized through workload classification. The classification could help identifying the way of using the other hardware resources in a server. As further work, we think it could be an effective way to improve the accuracy of power models based on CPU utilization, if workloads could be identified into different classifications with PMCs and attribute them with different models.

The accuracy of power model is limited by number of parameter (only one), and workload

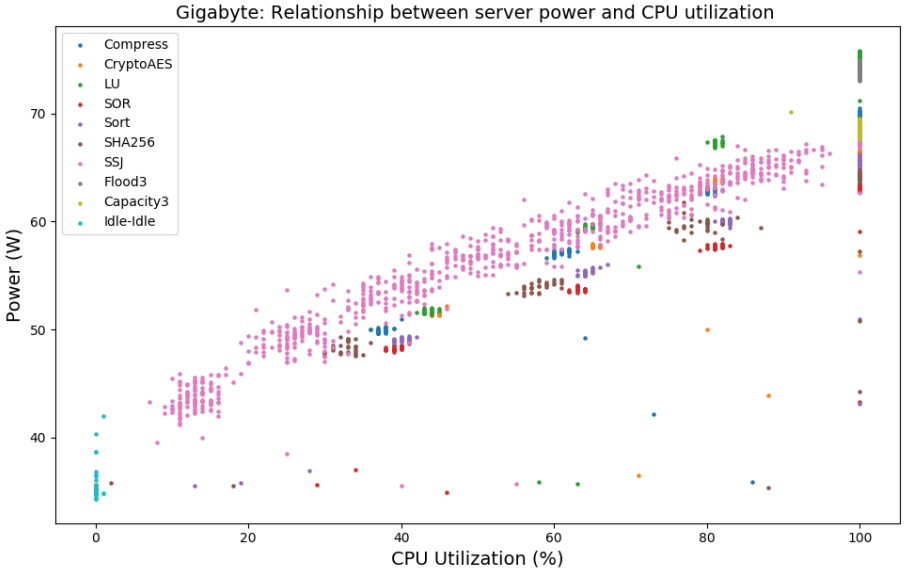


Figure 4.5 – Relationships between CPU-utilization and server power of different worklets.

seems to be a possible way to optimize the model accuracy. In the following section 4.3.2, we are going to explore if operating frequency of CPU can be also used to improve the precision effectively. In addition, we are wondering if the polynomial regression instead of linear regression could help increasing the model precision. The application of polynomial regression will be discussed in the section 4.3.3. Moreover, inspired by the study of section 2.4, we've also evaluated the influence of ambient temperature variation on the model. A delta power based on ambient temperature has been proposed to correct the thermal influence.

4.3.2 Question 2: For a given frequency, is power consumption linear to the CPU utilization?

In the study of [HCD+17a], server power is observed to be somehow linear to the number of cores running at full speed. However, the linear function needs to be updated under different p-state available in the system. `SimGrid` [HCD+17b] adopts this idea and develop the energy plugin in their simulator. SimGrid is designed to do simulations for analyze distributed application running at distributed computer systems. In the framework of Simgrid, an energetic profile is requested to configure in advance to the energy plugin [v3.]. For a multi core system with DVFS, the energy profile need to be configured for each p-state, with the following four power

values in Watts:

— **idle**: when OS is up and running but does nothing;
— **OneCore**: when one core is running at 100%;
— **AllCores**: when all cores are running at 100%;
— **Off**: when host is turned off. For the other number of cores running at 100% the power is estimated by linear extrapolation between **OneCore** and **AllCores**

However, it is impossible for SimGrid to estimate the power for a given core with a utilization between idle and full. That makes us wondering whether power could be linear to CPU usage for a narrow frequency range?

In order to perform the experiment, we determine to run the test in a modern server: Dell PowerEdge R630, the server has wider operating frequency: from 1200Hz to 2000Hz. A python library `psutil` (python system and process utilities) [Psu09] has been used to retrieve current CPU frequency in real time, with the function "psutil.cpu_freq". We divided the operating frequency into 10 small ranges as illustrated by figure 4.6. Measurements data points are marked by different colors according to the frequency ranges, as shown by figure 4.7.

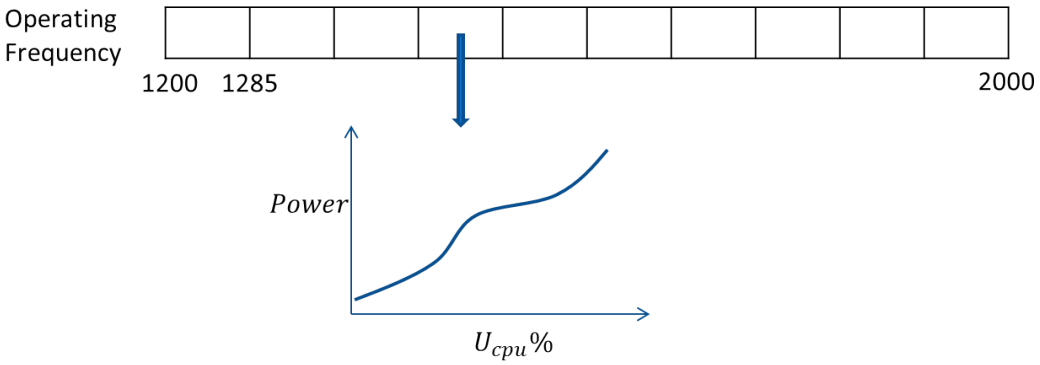


Figure 4.6 – Will server power be linear to utilization within a narrow frequency range?

We can see form the results that, **within a small frequency range, server power is not linear to CPU utilization, the power varies occasionally.** For example, data belonging to the highest frequency range is presented in blue sky, the data distributes randomly across different CPU utilization and server power. Even more, as illustrated by the comment in the figure, server power can be different for same CPU frequency and CPU utilization. We blame as well this phenomenon to the type of workload: the other hardware resources are being used differently. Therefore, making use of CPU frequency data is not an effective way to improve the accuracy of power models based on CPU-utilization.

In the following parts, we propose our own ideas to improve the model accuracy. The ideas

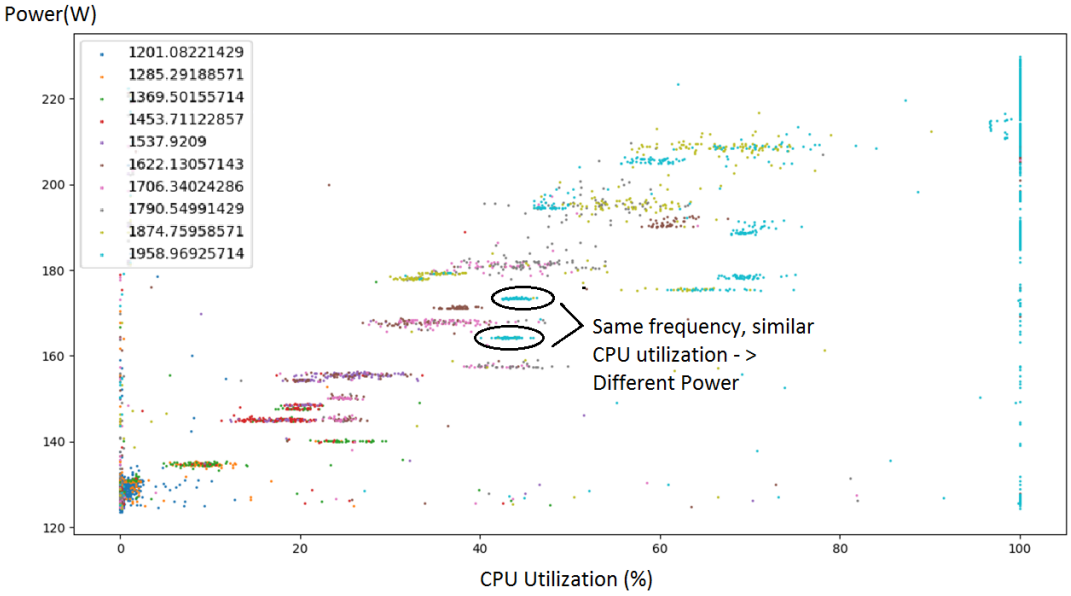


Figure 4.7 – Relationships between CPU-utilization and server power under different frequencies

are based on real observations on a physical machine. The approaches include applying polynomial function and adding additional temperature data into the model.

4.3.3 Proposition 1: Applying polynomial function

In question 1, we explain how accuracy is lost while building the power model with single CPU utilization value. Moreover, traditional method adopts linear regression to fit data with a straight line, which seems not the perfect relationship between CPU utilization and server power. According to the curves shown in figure 4.5, the data is supposed to be fitted better with a non-linear model. In this part, we try to improve the model accuracy by using polynomial regression. Polynomial regression is used to capture the non-linear features between variables with the method of least square [STI74]. Theoretically, polynomial regression belongs to linear model, as the model is expressed as linear in the form of coefficients associated with the variables. In practical, each x^n is treated as an independent variable, and the fitting can be done by least squares analysis to a multiple regression model [Smi18].

For a single independent variable, polynomial regression of degree n is defined as equation

4.1:

$$y = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n \tag{4.1}$$

Higher degree is able to capture more non-linear features for the data and improve the model accuracy. However, model could also learn noise if degree n is set too high. Model has perfect fitting result in trainning data set, but poor precision when dealing with new data that hasn't participated in the trainning process. This problem is called over-fitting. For a model over-fitted , even through the model gets closer to more data in the training set, it fails to capture precisely the unseen data and becomes less generalized [Aga18]. The overfitting can be prevented by providing more data to the training data set. However, in most of the cases in real situation, only one dataset is available for training. In this case, overfitting is possible to be avoided by algorithm. In our example, we evaluate quantitatively the overfitting issue by using Cross-Validation (CV). In CV mode, the whole data set is divided into two parts, one part is used to train the model, and the other part won't participate the trainning process but be reserved for model validation. CV accuracy is a qualification value between 0 and 1, higher accuracy in validation set represents for a more generalized model. We use the `PolynomialFeatures` class provided by scikit-learn [PVG+11] and train the model with polynomial regression under different degree. Then, we compare the CV accuracy of the validation data set, with the MAPE value (refer to "Model evaluation metric" in section 1.4.2) obtained from whole data set. The values of CV accuracy and MAPE with the increase of polynomial degree (from 1 to 9) are shown seperately in figure 4.8. MAPE values at different degrees are marked in red, it quantifies the overall error loss of the model. Accuracy of CV at different degrees are marked in blue, it quantifies the model quantity on validation data set.

As can be seen from the figure 4.8: in the beginning, with the rise of polynomial degree, both the accuracy of CV and the MAPE have been improved dramatically, especially when degree increases from 1 (equal to linear regression) to 2. However, starting from degree 6, MAPE continues to decrease but accuracy of CV becomes worse and worse. In fact, in this example, after degree 5, model starts to learn noise from the training data. Therefore, in this specific example, degree 5 is believed to be the optimal polynomial regression degree of the model to fit the data. At this point, validation shows the best accurate result.

After determining the polynomial regression as 5, we apply the polynomial regression model of degree 5 to the whole data set, the fitting result in shown in the figure 4.9. Blue points represent the real measurement data, and the power estimation result of the fitting model is showed in orange. It can be seen from the plot that, **polynomial regression has better abilities**

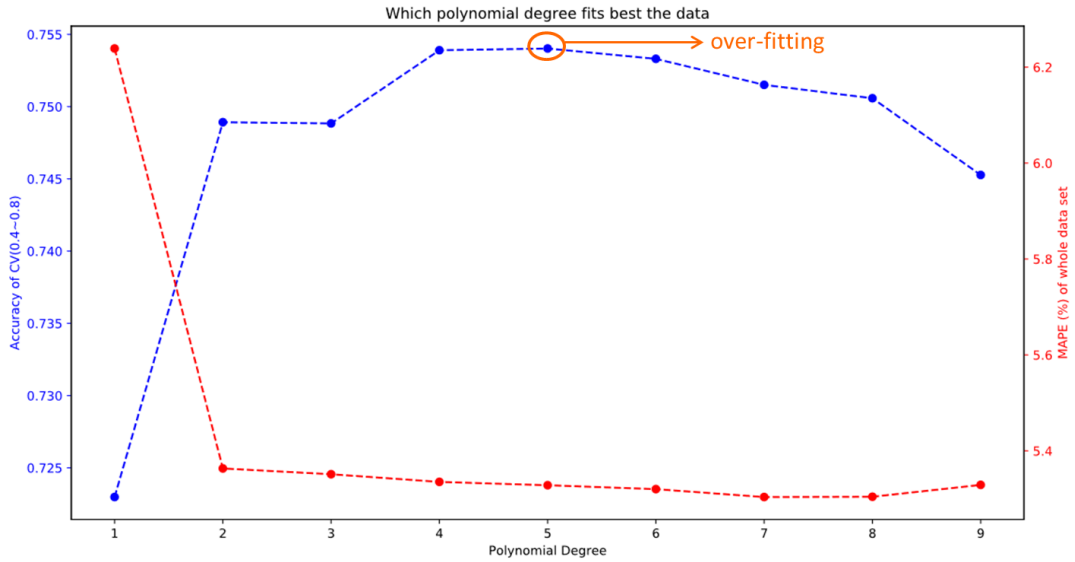


Figure 4.8 – Which polynomial degree fits best the data

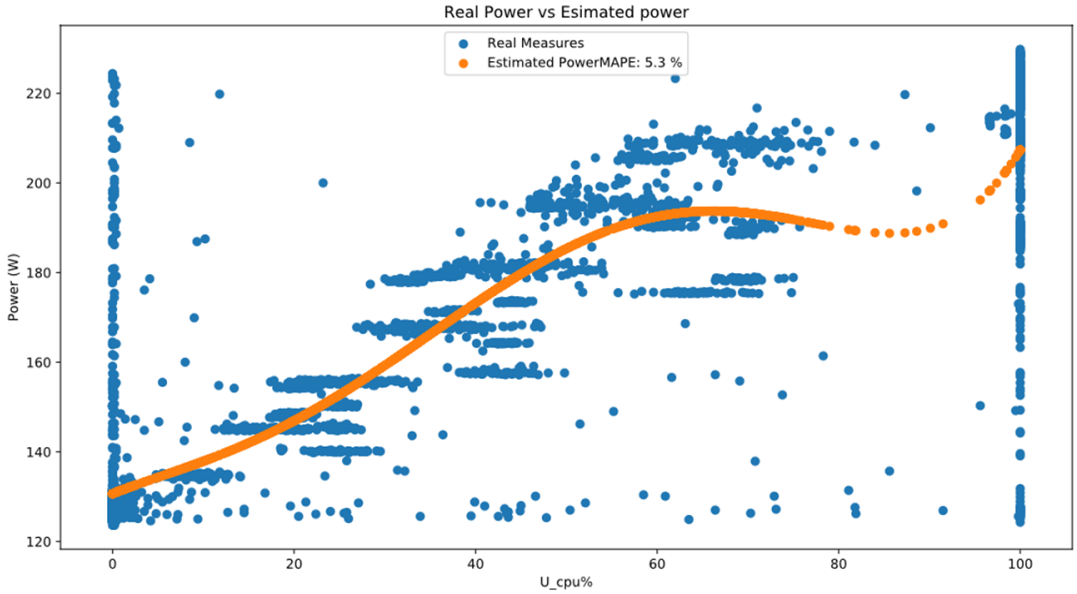


Figure 4.9 – Fitting power data with a polynomial function of degree 5

than linear regression in terms of capturing the non-linear features between independent variables: model built by linear regression can have error more than 6.2% (refer to the MAPE value at degree 1 in figure 4.8, polynomial regression with degree 1 equals linear regression). In using polynomial regression, the MAPE can be improved to 5.3%.

For attention, the most advanced math technology can not resolve the errors result from the lack of data. for example, actually, the model shown in figure 4.9 doesn't have ideal estimation of power for CPU utilization between 80-100%. The major problem is due to the lack of available data in this area. The precision of the power model is expected to be better by providing training data with full-scale coverage.

4.3.4 Proposition 2: Considering the influence of ambient temperature

Furthermore, we evaluate the influence of inlet temperature to the power consumption of server. The server is placed in a climatic chamber, where we can control the ambient temperature precisely. Inlet temperature is measured by a thermocouple of type K. Test suite SERT is executed three times on the server at 22°C, 35°C and 45°C ambient temperatures respectively. The results of the server power under different ambient temperatures can be seen in Figure 4.10. Server power increases with the rise of inlet temperatures. As we studies before, the increment of power is contributed mainly by fans and leakage current of CPU (refer to 2.4 for more information) [WNLMM18b]. Unlike power, CPU utilization remains all the same under different ambient temperatures, the results are shown in Figure 4.11. Therefore, as demonstrated by the experiments, there is a risk of losing accuracy without considering variation of ambient temperature in the models.

Taking the data sets of worklet "SSL" as a use case. Function (4.2) describes a baseline model proposed by [FWB07]. Data set is collected at 22°C ambient temperature. Estimated power is simply represented by a linear function by using the power values at idle and full load. Beyond the baseline model, we propose the power increment function $P_{Delta}(T)$. $P_{Delta}(T)$ represents the increment of server power due to the rise of ambient temperature $Delta_T$ (temperature base line is 22°C), with the unit of $W/°C$. $P_{Delta}(T)$ can be interpreted by a quadratic equation as shown in function (4.3). There are parameters are determined: a_0 (W), $a_1(W/°C)$ and $a_2(W/°C^2)$. The final power model is built by adding $P_{Delta}(T)$ to the baseline model, as shown by (4.4).

$$P_{estimated} = P_{idle} + U_{cpu\%}(P_{100\%} - P_{idle}) \tag{4.2}$$

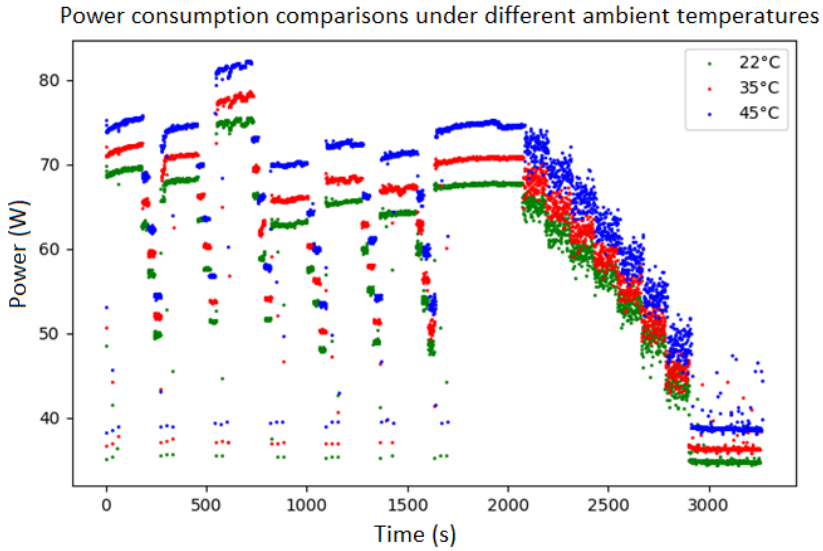


Figure 4.10 – Server power under three different ambient temperatures.

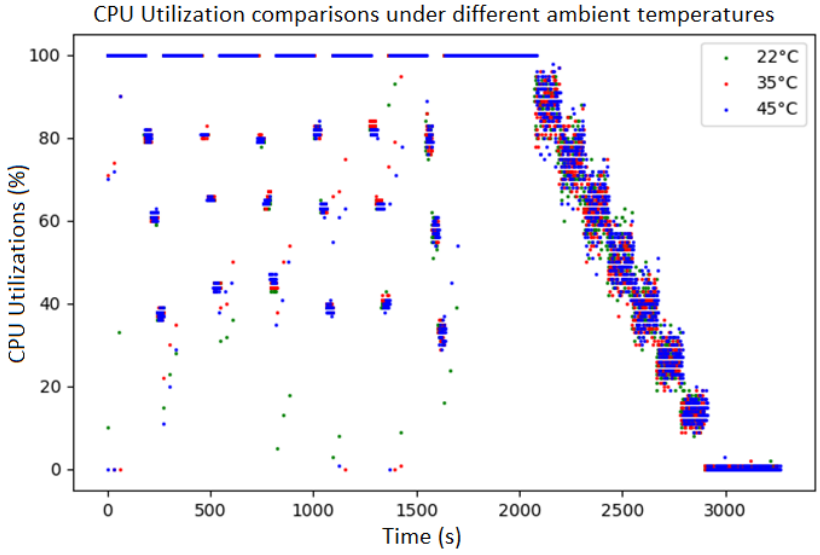


Figure 4.11 – CPU utilization under different inlet temperatures.

$$P_{Delta}(T) = a_0 + a_1 T + a_2 T^2 \tag{4.3}$$

$$P_{estimated} = P_{idle} + U_{cpu\%}(P_{100\%} - P_{idle}) + P_{Delta}(T) \tag{4.4}$$

Within the formulas, $U_{cpu\%}$ represents CPU utilization in percentage and T is ambient temperature. P_{idle} and $P_{100\%}$ are the average power (Watt) when server running at idle ($U_{cpu\%} = 0$) and full load ($U_{cpu\%} = 100$). The models are trained and validated with the same data set by using cross validation from function `cross_val_score` of scikit-learn [PVG+11], cross validation (cv) generator is set to 4 to realize a 4-fold cross validation. The average MAPEs after CV for model (4.2) and (4.4) shown in Table 4.4 demonstrate **the effectiveness of including the power increment function $P_{Delta}(T)$ (4.3): the MAPE of updated model (4.4) increases dramatically with the ambient temperature rises from $22°$C to $45°$C.**

TABLE 4.4. MAPE OF MODELS AT DIFFERENT AMBIENT TEMPERATURES

$T_{ambient}$ (°C)	22	30	45
Model without Delta (T°)	5.6 %	9.2 %	16.8 %
Model with Delta (T°)	4.6 %	4 %	3.2 %

4.4 Conclusion

The influence of thermal effects have been discussed and evaluated on servers and processors. In this study, we have further studied the thermal effects on accuracy of power models. We present a deep evaluation about the power models based on CPU utilization. The influence of inlet temperature on models has been especially discussed. According to the analysis, one regression formula by using CPU utilization as the only indicator is not adequate for building reliable power models. First of all, workloads have different behaviors by using CPU and other hardware resources in server platforms. Therefore, power is observed to have high dispersion for a fixed CPU utilization, especially at full workload (CPU utilization = 100%). At the same time, we also find that, power is well proportional to CPU utilization within the execution of one single workload. Hence, applying workload classifications could be an effective way to improve model accuracy. Moreover, inlet temperature can cause surprising influence on model accuracy. The model reliability can be questioned without including inlet temperature data. In a use case, after including inlet temperature data, we have greatly improved the precision of model outputs

while stressing server under three different ambient temperatures.

Using industrial specifications, such as IPMI and Redfish is another popular way to get power consumption data for some modern HPC servers. The experiment results show that, the precision of both IPMI and Redfish differs from different power ranges, the higher the better. We blame the loss of precision due to the latency during requests. Comparing to IPMI, Redfish is observed to have less latency in our experiments.

CHAPTER 5

ESTIMATING POWER CONSUMPTION OF CLUSTERS

5.1 Context and objectives

In this chapter, we present our research concerning the modeling of the global energy consumption of a computer cluster - the ecotype cluster of Grid5000 (refer to section 2.2.1 for more information about ecotype cluster). The cluster is located in an independent room with 48 identical servers and equipped with two separated air cooling systems. One is at the top the room and the other is close to the server racks. These physical facilities allows an overview about how the global power consumption varies according to operating variables in a data center, such as the computing resource usage, external environmental temperature, cooling system configurations, etc. In general, the energy consumption of a typical data center contains [IT07]:

— Servers & Storage: Doing actual computing, data processing and storage work.
— Cooling system: Maintaining favorable temperature conditions for hardware to operate, according to the requirements suggested by American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). [Com16]
— Network Hardware: Network devices which are responsible for transferring data packets across connecting devices on a network, including routers, switches, firewalls, modems, etc.
— Power conversion: Accessories such as cables, Uninterruptible Power Supply (UPS) and Power Distribution Units (PDU), etc.
— Lighting: Human operator related facilities.

The first two parts: servers and cooling systems consume the most of the energy required, as introduced in . Apparently, the energy consumed by servers and cooling is positively related to each other: when intensive work coming, servers increase the load level and occupy more computing resources to provide processing capacities. Server power increases with the increase of load, and in the same time generate more heat. As a result, cooling system works harder to

Table 5.1. Nomenclature

Symbol	Description	Units
T°_{inlet} or T°_{i}	Temperature of the air entering into the servers (measure)	C°
T°_{outlet} or T°_{o}	Temperature of the air left the servers (measure)	C°
T°_{room} or T°_{r}	Temperature of the room (measure)	C°
T°_{server} or T°_{s}	Temperature of the servers (virtual) [1]	C°
$T^{\circ}_{container}$ or T°_{c}	Temperature of the container (virtual)	C°
$R_{1,2,3,4}$	Server Racks, #1 to #4	
PCS	Primary Cooling System, In-Row	
SCS	Secondary Cooling System, air conditioning in the room	
P_{server}	Instant power consumed by servers	W
$P_{cooling}$	Instant power consumed by In-Row	W
h_r	Thermal conductance between room and the container	C°/W
h_c	Thermal conductance between container and cold air entering into servers	C°/W
h_s	Thermal conductance between servers and container	C°/W
h_o	Thermal conductance between container and hot air leaving servers	C°/W
$C_{a,1}$	Thermal capacitance of air at the container's cold side	J/C°
$C_{a,2}$	Thermal capacitance of air at the container's hot side	J/C°
C_c	Thermal capacitance of container	J/C°
C_s	Thermal capacitance of servers	J/C°
M_s	Flow rate of air passing through servers	kg/s
C_p	Specific heat of air at constant pressure	J/kgC°
t	time	s
T	Temperature	C°

meet the heat transfer requirement. Within a closed and stable system, according to heat balance policy, it is possible to build a model to estimate the energy required by cooling system based on the power consumption of servers. Another key point is the temperature outside of the cooling system, the heat exchange could be important if the thermal isolation is weak.

A dedicated power model of data center could be useful for predicting the global energy consumption by feeding common operational data such as server usage information and thermal conditions. Once the model is realized, by varying configurations, we could be able to optimize the global energy efficiency at run time level. For example, finding the optimal cooling system set points according to the current and/or the upcoming loads.

This study is composed of both physical experiments to the clusters and numeric modeling of the global energy consumption, including the power consumption from servers and cooling

systems. In order to facilitate the reading, notations that have been used in this chapter are detailed in table 5.1. The contents for the rest of this chapter are organized as follows:

In 5.2, a global description of the cluster is given, including the position and placement of the servers and cooling system in cluster; the role and function of cooling system and the sensors used for getting data. Our basic idea of building the model based on observations of the real time system is presented at the end of this section. In section 5.3, we present our study about the digital simulation of the cooling behavior for the cluster. The model is able to estimate the real time temperature of the cold air entering into servers and of the hot air leaving the servers. The model takes three operating variables as inputs: power of servers, power of cooling system and the temperature outside the cooling system. In section 5.4, we propose a cooling consumption model based on the real time temperature of the cold air entering into servers. The conclusion is given in section 5.6.

5.2 Cluster overview

Basic information about ecotype cluster has been presented in previous study at section 2.2.1. In this part, we present the following information: power management platform Seduce in 5.2.1; architecture and function of the cooling system in section 5.2.2 and thermal behavior in section 5.2.3. Finally, we present our idea of building the model based on the knowledge of the system in section 5.3.

5.2.1 Power and thermal management platform: Seduce

Seduce platform is used to collect data in this study. *SeDuCe* is a scientific testbed [PM18] designed for power and thermal management in data centers. Seduce has been developped for cluster *ecotype* of the Grid5000 infrastructure. It enables researchers studying both power and thermal aspects of servers while conducting experiments. Uses from Grid5000 can have real-time access to operating information about the cluster. In terms of individual server, users can get power, the temperatures at the front and back of each server. Server power data is retrieved from intelligent PDU installed in the cluster, and temperatures are getting from the measurements taken by thermal couples of type k. Besides servers, the room temperature is also monitored with a temperature sensor. In addition, *Seduce* also integrates the data provided by cooling system of cluster. Such as the inlet and outlet temperatures of the cooling system, temperature thresholds, fan speed, cooling consumption, etc. Users can visualize the real time data of the

system through the dashboard tool. If needed, the raw data can be obtained via the web portal and user-friendly *Seduce API* by specifying the start and end time in the scripts. All the measurements are collected at the frequency of 1Hz.

5.2.2 Cooling systems of the cluster

Figure 5.1 provides a top view of the "ecotype" cluster. It disposes of five air-tight racks, design is based on *Schneider Electric IN-ROW* model ACRD600 (200-240V, 50/60Hz) [ele19]. We have introduced this row-based air side cooled system the state of the art study at 1.2.5. In addition, the model applies indirect air-side free cooling technology to take advantage of the outside cold air. The system is equipped with an evaporator to help increasing the temperature difference between the outdoor air and the liquid in the heat changer (refer to "Free cooling" at section 1.2.5 for more information about free cooling of this kind). The functions of the In-Row cooling system will be explained later.

The cluster disposes of two cooling systems, the primary cooling system (PCS) also called In-Row is installed in the middle of the racks for servers. In-row gets back the hot air at the back side of the servers (T°_{outlet}), cools down the hot air then blows the cold air to the front side of the servers (T°_{inlet}). This In-Row model consists in a direct expansion cooling system with fans and controllers. Four server racks and one In-Row rack are placed inside an closed cooling container equipped with Plexiglas doors, which limits the air and heat exchange between inside and outside of container. Equipped with advanced air management strategy, In-Row is capable to work efficiently, as the cooling area is limited within the container, the influence from outside of the container (room temperature) is restrained as well thanks to the design. Besides In-Row, on the top of the room, there is an air-conditioning served as the Secondary Cooling System (SCS) to keep a stable room temperature. This study concentrates on the cooling power consumption of the PCS (In-Row). The power consumed by SCS is wished to be completed in future study.

In order to better understand the functions of In-Row, a 3-D architecture of the cluster is presented in figure 5.2. Arrow marks in the figure indicate the circulation of the airflow within the container, between servers and the In-Row. Red and blue represent for hot and cold air respectively. Cold air passes through the working servers from the cold side of the container, be warmed up and gets out to the hot side. Afterwards, hot air will be taken in by the In-Row, cooled down quickly while passing through the cooling module of the In-Row then return back to the cold side. Cooling module of In-Row is actually composed by two parts. One part is placed inside the rack and another part is installed outdoor. Figure 5.3 shows the details of its architecture and illustrates the actions taken by the cooling module to lower the temperature

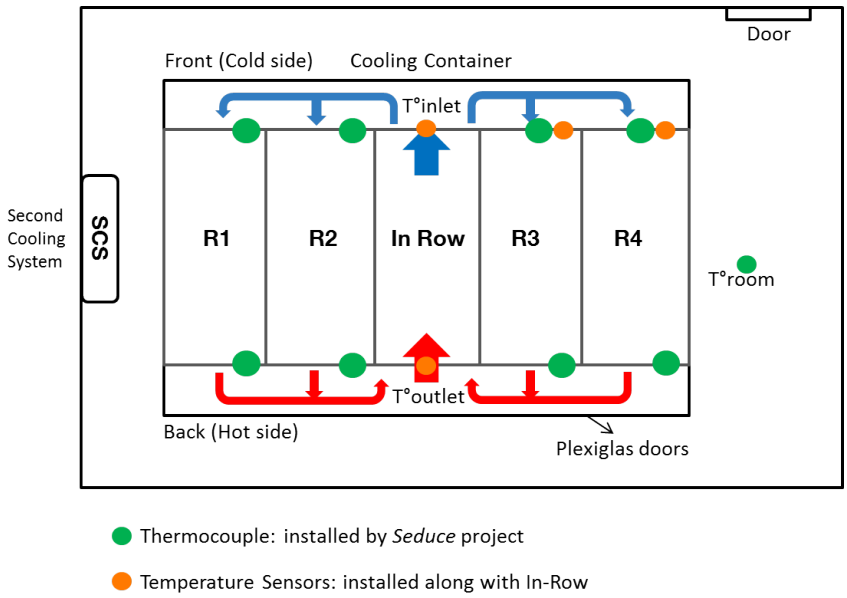


Figure 5.1. "Ecotype" cluster topview

of hot air. The compressor constricts the refrigerant vapor, raising its pressure, and pushes it into the condenser. When the hot gas in the condenser meets the cooler air temperature of the outside, it becomes a liquid. The refrigerant absorbs the heat from the server, cooling down the air. Condenser has large surface area to evacuate heat from liquid to outside air quickly and efficiently. On the way out of the condenser, expansion device helps lowering down the pressure within the metal tube to facilitate cooled liquid return back to the evaporator.

Once configured, the In-row works automatically along with sensors installed around the container. The positions of the temperature sensors are marked in the figure with orange points in figure 5.1 and 5.2. One sensor is placed at the cold side in order to get the value of the inlet air temperature (T°_{inlet} or T°i). It's positioned right after the fan, where cold air just gets out from the In-Row. Another one is placed at the hot side to get the value of outlet air temperature (T°_{outlet} or T°o). It's positioned on the top the In-Row before the fan, as the hot air is usually lighter than the cold air. There are three other temperature sensors placed in front of rack 2,3,4 to provide additional information when necessary. The In-row has a central processing system to manage the measurements from sensors and guide the operations of In-Row. Operators can access the processing system remotely to view or modify some cooling configurations and retrieve the measurements. On the top of the room, there is a temperature sensor (marked by a green point) installed by ourselves to monitor the room temperature (T°room or T°r).

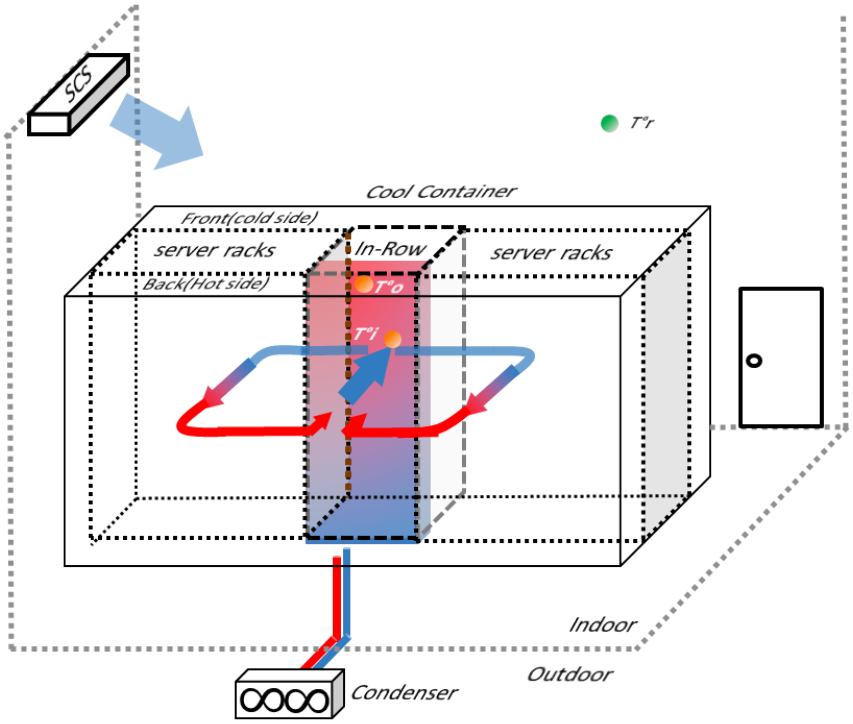


Figure 5.2. "Ecotype" cluster 3D architecture

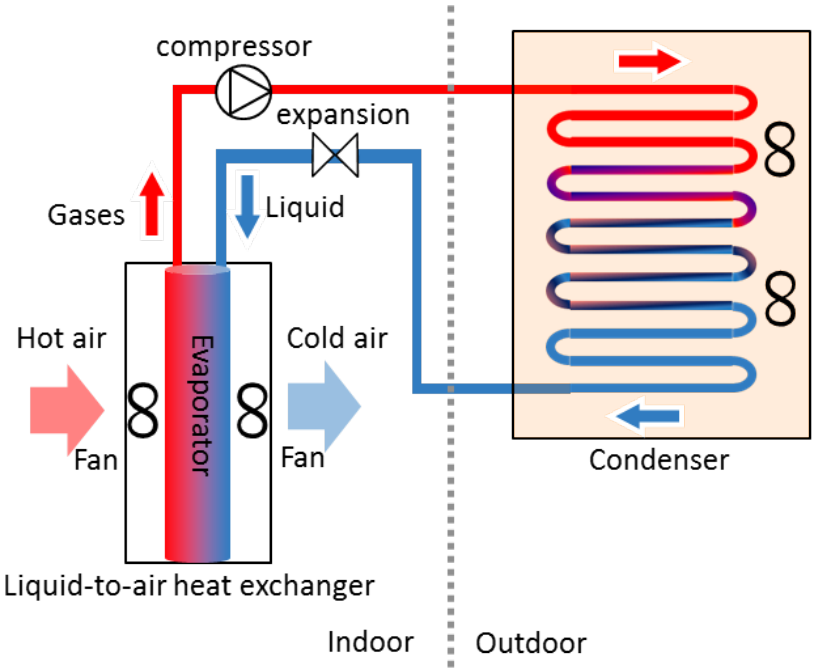


Figure 5.3. Architecture of the cooling module of In-Row

5.2.3 Observations of the thermal management of the cluster

We start building the model by observing the real time data in the *Seduce* platform (refer to 5.2.1). In order to have a clear view, we zoom in the system to a short period of time. Figure 5.4 illustrates a typical thermal behaviors of the In-Row for a about two hours.

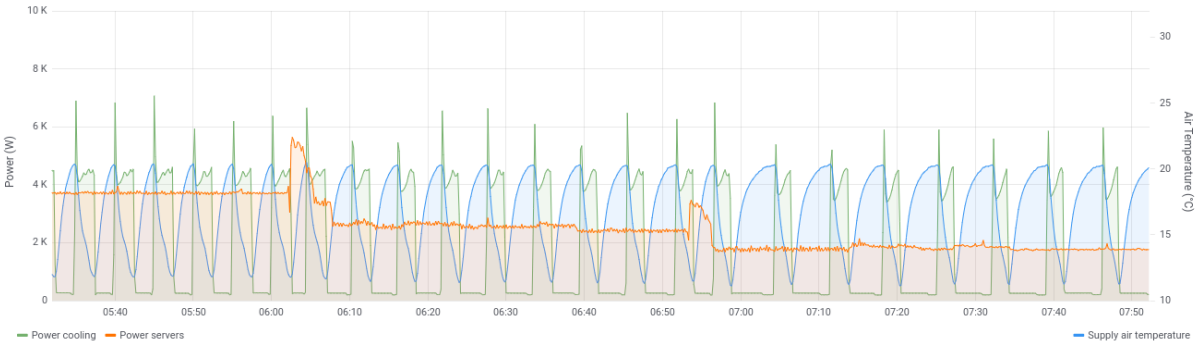


Figure 5.4. Thermal behaviors of In-Row

In the figure 5.4, orange line represents the $P_{server}(t)$, green line represents $P_{cooling}(t)$. $P_{server}(t)$ represents the power of all the servers in the cluster at time t. $P_{cooling}(t)$ represents the total power of In-Row (including all the parts) at time t. $P_{server}(t)$ and $P_{cooling}(t)$ own the left label with the unit of kw. Blue line represents T°_{inlet}, owns the right label with the unit of Celsius degree. As shown in figure 5.4, $P_{cooling}$ within the container is not a constant value. Actually, In-Row activates compressor regularly according to T°_{inlet}, T°_{inlet} varies in an infinite loop and acts as the trigger for cooling module. As shown by figure 5.5, there are two phases in a cycle, in the first phase, compressor is turned off, hot air continues to increase the liquid temperature within the evaporator and T°_{inlet} becomes more and more higher. In this phase, In-Row works at a constant and lower power about 460 Watt, energy is consumed mainly by fans, condenser, regulator and control units. Once the T°_{inlet} exceed the high temperature threshold about 20°C, compressor will be activated to lower down the liquid temperature and the second phase begins. T°_{inlet} starts to decrease. During this phase, with the activation of compressor, power of In-Row will be increased more than 10 times. Once T°_{inlet} is lowered down to the low temperature threshold about 12°C, compressor will be then deactivated and the system returns back to the first phase. The temperature of the high and low thresholds of T°_{inlet} are initialized at the installation by technicians from the manufacture. The configuration of the thresholds are not allowed to be modified by users because of warranty policy.

Heat is generated mainly by servers, therefore, power of servers can influence the increasing rate of T°_{inlet} at the first phase. We show an example in figure 5.4, where P_{server} varies from

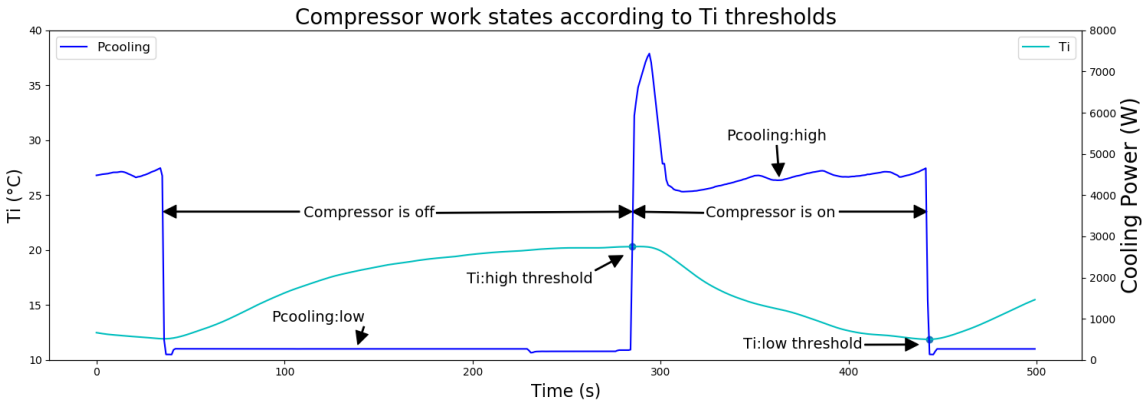


Figure 5.5. Demonstration: two phases in a cycle

about 1.8kw to about 3.8kw. When P_{server} becomes higher, servers generate more heat to the space in the same unit of time, T°_{inlet} increases quicker and activates compressor in a shorter time. Otherwise, when P_{server} is lower, less heat will be generated, T°_{inlet} will take more time to reach the high temperature threshold. Except for the P_{server}, we think that T_r could also be an important key indicator, as cooling container is not built by perfect thermal isolation material. Even though the container is closed and it is capable to prevent massive air flow between the cooling container and the room, the heat exchange between them could still be important to affect the temperature of air inside the container (T°_{inlet} and T°_{outlet}).

Figure 5.6 shows a stable system state: both P_{server} and T_r have stable and constant values over time. In this case, T°_{inlet} has highly similar form at each cycle time.

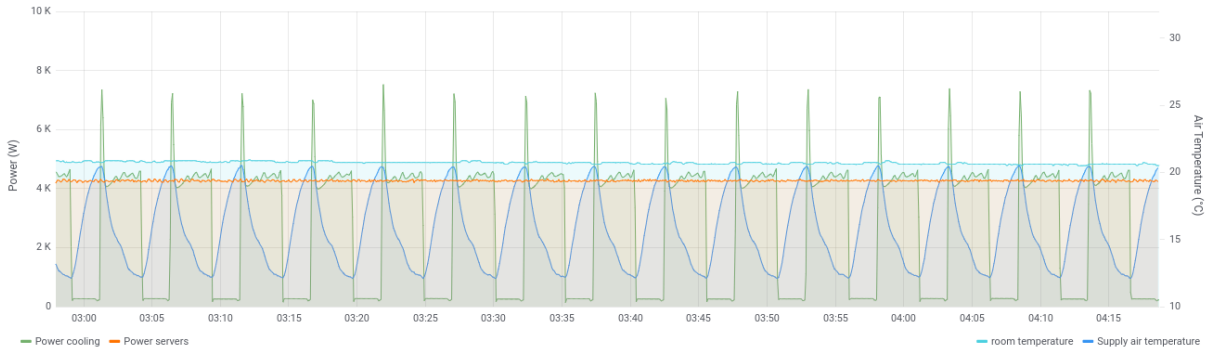


Figure 5.6. One case: In-Row works at a stable state

5.2.4 Idea of building cooling model, based on the conclusions from observations

As conclusion, for a normal working cluster, T°_{inlet} ranges between the lower and higher temperature thresholds in an infinite loop. It continues to increase when compressor stop working and turns to decrease with the activation of compressor. P_{server} and T_r can change the thermal conditions within the container then lead to variation of loop cycle time. According to the observations, we draw several rules to determine the operations of the cooling system: In-Row relies on T°_{inlet} to activates and deactivates compressor. T°_{inlet} varies according to the P_{server}, $P_{cooling}$ and the T_r.

In reality, In-Row follows regular working patterns and these patterns can be simulated through thermal analyses. However, classical thermal analyses for such a system need to solve Navier-Stokes equations which requires a lot on design for grid and huge computational resources. Taking all the situations into consideration, we decided to build a simplified one dimensional thermal model of the cluster by using only four measurements into the model: P_{server}, $P_{cooling}$, T_r, T°_{inlet} and T°_{outlet}. Once the parameters of the thermal model are determined, $P_{cooling}$ could be estimated by providing the other variables. The idea of building the cooling consumption model could be realized by following the steps below:

— Modeling T_i based on P_{server}, $P_{cooling}$ and T_r:

$$T_i(t) = f(P_{server}(t), P_{cooling}(t), T_r(t)) \tag{5.1}$$

— Modeling $P_{cooling}$ based on T_i:

$$P_{cooling}(t) = f(T_i(t)) \tag{5.2}$$

— Modeling the $P_{cooling}$ based on P_{server} and T_r:

$$P_{cooling(t)} = f(T_r(t), P_{server}(t)) \tag{5.3}$$

5.3 Modeling thermal behavior of In-Row by digital simulation

In this section, we present our method of building the models to simulate the thermal behavior of the cluster. The idea comes from observing the real time data of the cluster as described

Table 5.2. Thermal-Electrical Analogy; symbols and units [Dav04] [FVLA02] [PW08]

Thermal			Electrical		
Temperature	T	C°	Voltage	V	V
Thermal Resistance	R	C°/W	Electrical Resistance	R	Ω
Thermal capacitance	C	J/C°	Electrical Capacitance	C	F
Heat Flux	$\dot{Q} = \frac{\Delta T}{R}(W \cdot m^{-2})$		Current through Capacitor	$I = C\frac{dV}{dt}(A)$	
Heat balance			Kirchhoff's Current Law		

in section 5.2.3. So far, we have realized two modelings, one for T_i and another for $P_{cooling}$ respectively as described by equation 5.1 and 5.2. In this section, we present our work concerning the model for T_i. The model for $P_{cooling}$ will be presented later on in section 5.4. The method of building model for T_i is introduced in subsection 5.3.1. Value of T_i in real time is actually related to several system variables. In order to estimate the T_i, we build an entire thermal system for cluster. The whole system has been simplified and represented as an equivalent RC circuit to facilitate the data processing and calculation. Then in subsection 5.3.2 we describe how to identify the parameters of the model. Finally, we show the model validation results in 5.3.2.

5.3.1 Method for the In-Row thermal system simulation: equivalent RC electric circuit

During the observation, we notice that the variation of T_i over time seems similar to the form of current in a resistor-capacitor(RC) electrical circuit. Actually, in 1942, Paschkis [Pas42] has already proposed a description about how to simulate thermal behavior in buildings by using electrical analogy analyses. In this study, we present our approach to model the thermal system of a cluster by analysing the corresponding equivalent electrical circuit. Similar methods have been proposed in previous studies like [TWR02]. Table 5.2 list the analogy between electrical components and thermal quantities that have been used in this study.

Thermal resistance (R) and thermal capacitance (C) are fundamental elements in a thermal system [Che]. They are both physical properties of material. Figure 5.7 shows a typical heat flux through an object. Q represents for heat, according to the second law of thermodynamics, the heat flows (shown by the red arrow) from the hot side of an object to the cold side to equalize the temperature difference [Dug18]. Thermal resistance (R) and thermal capacitance (C) determine

the behaviors of heat flow when temperature difference presents.

Three Heat transfer modes can be distinguished: by thermal conduction, thermal convection or thermal radiation. In terms of thermal conduction, for an object with thermal conductivity $k(W/(mK))$, area $A(m^2)$ and thickness $L(m)$, thermal conductance (h) of an object can be defined by equation 5.4, measured in W/K or W/C°. Thermal resistance is the inverse of thermal conductance (note as h and $R = \frac{1}{h}$) [BK03].

$$h = \frac{kA}{L} \tag{5.4}$$

In terms of thermal convection, with heat transfer coefficient h_v, we have:

$$h = h_v \times A \tag{5.5}$$

In our model, thermal convection is taken into account through conduction coefficient. And heat transfer by radiation is integrated in convection coefficient. For the rest of this study, the three heat transfer modes are all included and simplified by thermal conductivity variable h.

Thermal conductivity h measures the ability of a material to conduct heat [AA66]. Heat will be transferred faster in a material with higher thermal conductivity. For example, material with high thermal conductivity are widely adopted in heat sink to dissipate heat, while material with low conductivity material are commonly applied to provide thermal insulation. The heat flux through an object can be obtained by dividing the temperature difference by thermal resistance as shown by equation 5.6.

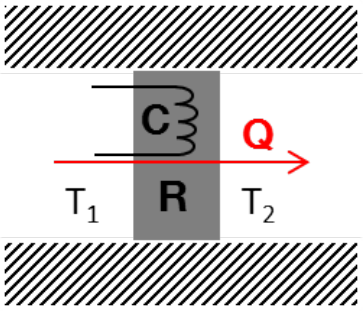


Figure 5.7. Typical heat flow in a thermal system

$$\dot{Q} = \frac{T_2 - T_1}{R} = h \times (T_2 - T_1) \tag{5.6}$$

Besides thermal conductance, when heat diffuse through an object, the temperature of object could be raised because of thermal capacitance (also known as thermal mass), and T_2 will

be lower as heat flows out. Thermal capacitance describes the capability of an object to store heat, providing "inertia" against temperature variation [KPG+01]. For an object of uniform composition, the thermal capacitance of an object can be approximately determined by equation 5.7, where m is the mass in kg and C_p is the specific heat capacity of the material in $J/(kgC^\circ)$. The heat flux depends on the temperature difference between two sides and can be defined by equation 5.8.

$$C = mC_p \tag{5.7}$$

$$\dot{Q}(t) = C\frac{dT}{dt} \tag{5.8}$$

Figure 5.9 presents the equivalent electrical circuit of the cluster. In order to simplify the modeling, the whole system has been considered as single layers, materials like container and server are considered to be homogeneous. It is therefore an one-dimensional thermal system. All the servers present in the cluster are simplified to one point. $P_{server}(t)$ is the heat resource of the system. We suppose that, the power consumed by servers is all turned into heat. In this case, $P_{server}(t)$ represents for the total instant server power at time t, the measurement value is provided by a high accuracy power analyser. $P_{cooling}(t)$ provides the cold source and lead to the drop of $T_i(t)$. Unlike the electrical power, in thermal system, the cold production power of In-Row is called "cooling demand". Cooling systems have usually an energy efficient design, with one unit of electrical energy consumption, they are able to produce several units of cold (usually from 2 to 5). Compressor is the key component to produce cold in a cooling system, most of the electrical energy is consumed by compressor. The efficiency of cooling system can be obtained by dividing the "cooling demand" with the electrical power. For example, figure 5.8 shows the relationship between cool demand and electrical power for the In-Row. Cool demand is zero when compressor stops working, while during the activation of compressor, the In-Row works around electrical power of 4.4kW has a cold production power at about 23.3kW, in order to produce 23.3kW of cold air. The In-Row has therefore an efficiency of about 5.3.

In our modeling, in order to facilitate the calculation of electrical energy consumption, we applied $P_{cooling}(t)$ as the instant electrical power consumption of In-Row at time t, the measurement is provided by control panel of In-Row. The points noted as T_x represent temperatures at different positions of the cluster. T_r, $T_i(t)$ and $T_o(t)$ are real measurements taken by temperature sensors at time t. Their positions are precised in 5.2 and 5.3. $T_c(t)$ and $T_s(t)$ are "imagined" virtual temperatures of container and servers. They are necessary to be included in the model to

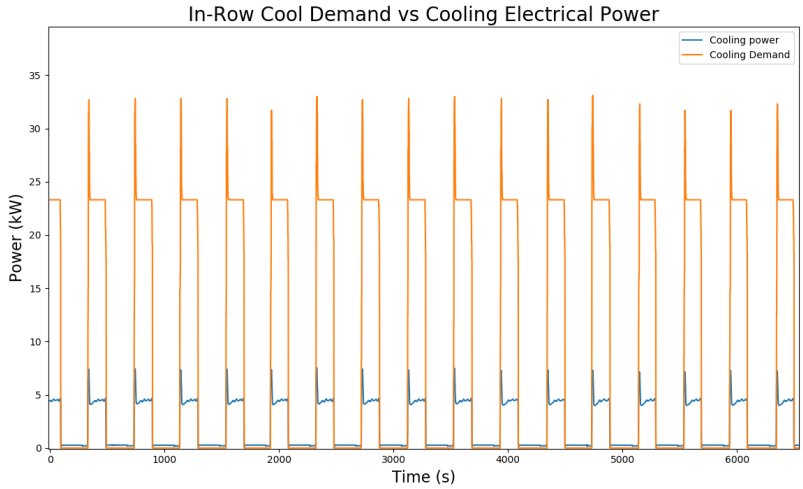


Figure 5.8. In-Row Cool Demand vs Cooling Electrical Power

present the existence of properties for container and servers. All the physical measurements are provided by Seduced platform with a sampling rate of 1 second.

In terms of parameters of the thermal system, C_x represents the thermal capacities of different material with certain volumes. And h_x represents heat transfer by conduction, convection or radiation. For example, h_r models the heat exchange between the air of the room and the surface of the cooing container. Details of the notations are explained in Table 5.1.

Each point in the model represents a volume (with corresponding physical properties, i.e, mass, specific heat, thermal resistance). At each time step, energy balance at the volume level is computed, as demonstrated by equation 5.6 and 5.8, the thermal system model of the cluster can be formed by a group of ODEs (Ordinary Differential Equation) in 5.9:

$$\begin{cases} C_{a,1}\dfrac{dT_c}{dt} = h_r(T_r - T_c) + h_s(T_s - T_c) + h_o(T_o - T_c) \\ C_c\dfrac{dT_i}{dt} = -P_{cooling} + h_c(T_c - T_i) \\ C_s\dfrac{dT_s}{dt} = P_{server} + m_s C_p(T_i - T_s) + h_s(T_c - T_s) \\ C_{a,2}\dfrac{dT_o}{dt} = m_s C_p(T_s - T_o) + h_o(T_c - T_o) \end{cases} \tag{5.9}$$

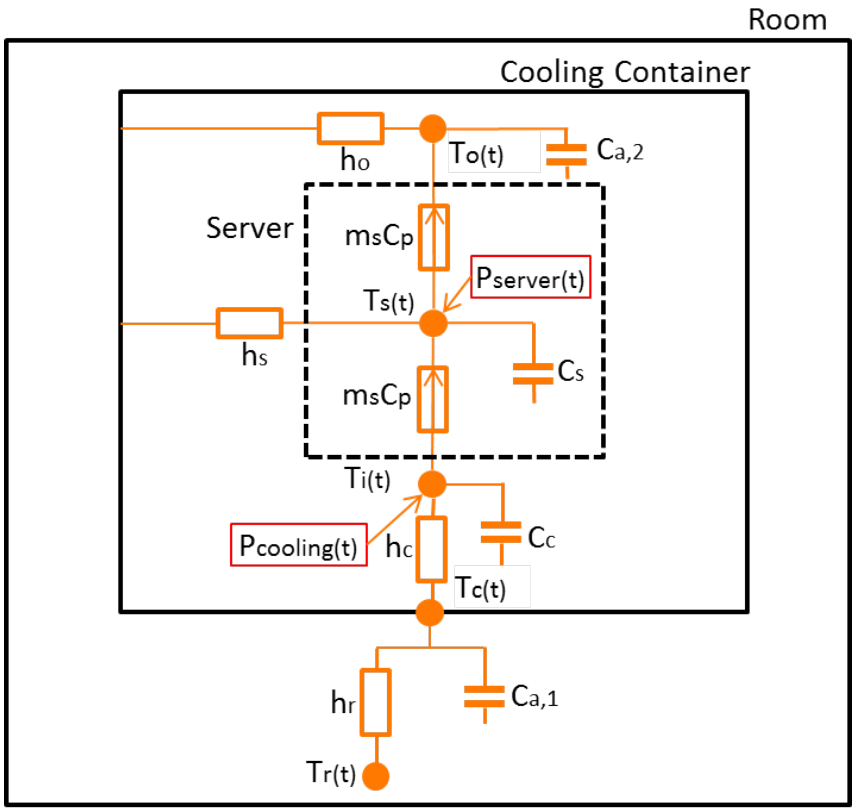


Figure 5.9. Equivalent circuit of the cluster's thermal system

5.3.2 Parameter identification

As explained in previous section, the thermal model proposed for the cluster has been much simplified. The parameters presented in the model are mixed with several objects or combined with several forms of heat transfer. It becomes impossible to recognize their values through physical or thermal properties. Therefore, we need to identify these parameters by taking use of available physical measurements of the system: T_r, $T_i(t)$, $T_o(t)$, $P_{cooling}(t)$ and $P_{server}(t)$. Searching the optimal parameters of a model to fit existing data is an optimization problem. We choose the python package `LMFIT` [MS+18] for perform the optimization. `LMFIT` is designed for fitting complex models to real data by performing non-linear least-squares minimization and curve-fitting methods. The problem can be expressed as minimizing the difference between the model estimation results and the real measurement values. Inputs of the model are operating variables as $P_{cooling}(t)$, $P_{server}(t)$ and $T_r(t)$, and the outputs are the temperatures of four points: $T_c(t)$, $T_i(t)$, $T_s(t)$ and $T_o(t)$. Among them, T_c and T_s are virtual variables, we have only the measurements of $T_i(t)$ and $T_o(t)$. Therefore, the identification of the parameters in our cooling model is done by progressively comparing the estimation results $T_i(t)$ and $T_o(t)$ from the model with the real physical measurements. The principal procedure of identifying the parameter can be described by figure 5.10. In the figure, T_i_e and T_o_e are estimations from model simulation. Optimization algorithm calculates the value of the sum of the last square value between the estimations and real measurements with the parameters provided, then tries new parameter combinations in order to obtain a lower value.

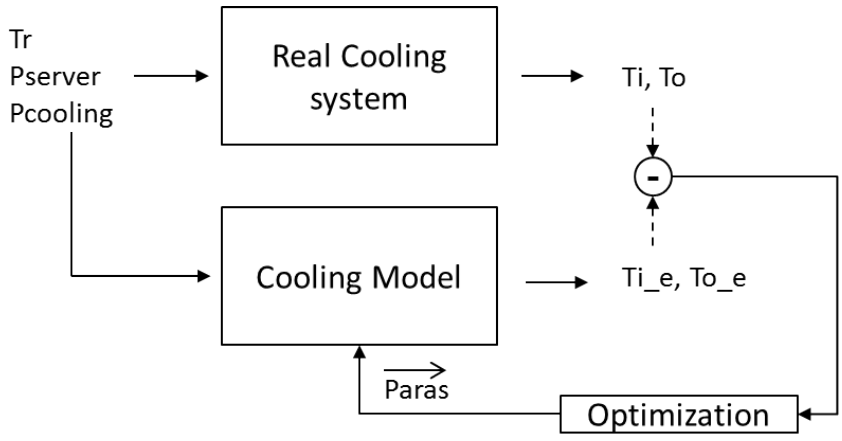


Figure 5.10. Diagram of the identification procedure

There are two key functions in the program of identification:

1. Cooling system model functions. It is a time-series model composed by four ODEs. The function takes in operating variables as inputs: T_r, P_{server} and $P_{cooling}$, thermal system parameters $\overrightarrow{C_x}$, $\overrightarrow{h_x}$ and M_sC_p. Outputs of the model are the temperatures from different positions of the cluster as function of time. Pseudo code of the model is exhibited below:

```
function cooling_model(T_initial, Paras, Data)

    # Initial data
    Tc, Ti, Ts, To = T_initial
    # Operating variables, from physical measurements
    Tr, Pserver, Pcooling = Data['Tr', 'Pserver', 'Pcooling']
    # Cooling system parameters
    Hr, Hc, MsCp, Hs, Ho = Paras['Hr', 'Hc', 'MsCp', 'Hs', 'Ho']
    Ca1, Ca2, Cc, Cs = Paras['Ca1', 'Ca2', 'Cc', 'Cs']
    # Cooling model expressed by differential equations
    Tc_t = (-Hr - Hs - Ho) / Ca1 * Tc + Hr / Ca1 * Tr + Hs / Ca1 * Ts + Ho / Ca1 * To
    Ti_t = -Pcooling / Cc + Hc / Cc * Tc + (-Hc)  / Cc * Ti
    Ts_t = Pserver / Cs + MsCp / Cs * Ti + (- MsCp - Hs) / Cs * Ts + Hs / Cs * Tc
    To_t = MsCp / Ca2 * Ts + (-MsCp - Ho) / Ca2 * To + Ho / Ca2 * Tc

return Tc_t, Ti_t, Ts_t, To_t
```

2. Objective minimize function. The function returns the differences between real T_i, T_o from measurement data and the their estimations from cooling model, by calculating the sum of the differences. Pseudo code of objective function is exhibited below:

```
function residu(T_initial, Paras, data):

    % Get the estimating results from cool model
    T_estimate = cooling model(T_initial, Paras, data)
    Ti_estimate = T_estimate['Ti']
    To_estimate = T_estimate['To']
    % Calculate the sum of difference for Ti and To
    residu = (data['Ti'] - T_estimate['Ti']) + (data['To'] - T_estimate['To'])
return residu
```

We applied the function `minimize` from `LMFIT` to perform the optimization. According to the documentation, function `minimize` will do a least-squares optimization of the return array: calculate the sum of squares of the array, then send the result to the optimization method to be minimized. Optimization method searches the optimal parameter values within the constraints defined by users. Our cooling model contains nine parameters. The procedure of searching the parameters are realized by two steps. Firstly, determining the range of the parameters according to physical thermal characteristics of object, and secondly determining the value of parameters by running optimization algorithm with several training sequences.

Identification the thermal capacitance parameters

Cool model shown in 5.9 contains nine unknown parameters, including four thermal capacitance parameters, four thermal conductance parameters and a thermal fluid conductance $M_s C_p$, which represents air flow passing through servers. Thermal conductance values may vary according to some system variable such as flow. Thermal capacitance represents materiel property, they are supposed to be constant in our system. Therefore, in the first place, we start by identifying the values for the thermal capacitance parameters. The training set chosen is shown in figure5.11, during this period, system has a stable state: P_{server} varies around 1.7kW and within a limited range, T_r is stable and between 23C° to 24C°. All the parameters during the training set are approximately constant.

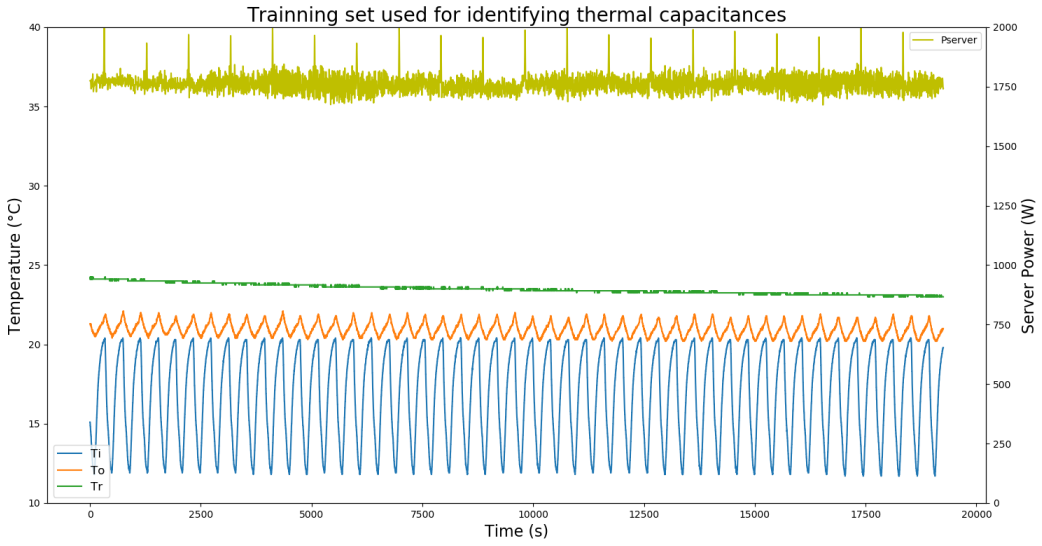


Figure 5.11. Training dataset for identifying thermal capacitance

Even though all the parameters are supposed to be constant in this training dataset, possible parameter combinations would be too large to select without specifying a reasonable researching gird. Fortunately, all the parameters represent certain thermal and physical properties of objects. We start by assigning ranges for each of the parameters according to the prior knowledge about the system. For example, $C_{a,1}$ represents for the capacitance of the air in the room. The value ranges can be obtained according to equation 5.7, where the specific heat capacity of air ($C_{p,air}$) is about 1004 $(J{\cdot}kg^{-1}{\cdot}K^{-1})$. The mass of air in the room can be obtained thought the production of density of air ($\rho_{air} = 1.225kg/m^3$ at $15C°$, sea level) and the volume. The

Table 5.3. Initial value and researching range for system parameters

Parameters	Searching range
$C_{a,1}$	[3800, 26600]
$C_{a,2}$	[175, 1550]
C_c	[23300, 186400]
C_s	[200000, 2000000]
$M_s C_p$	[301, 903]
h_r	[0.04, 3]
h_c	[80, 4860]
h_s	[0, 2]
h_o	[0, 2]

volume of the room except the cooling container is estimated between 15 m^3 to 103 m^3. Therefore the range of parameter ($C_{p,air}$) suppose to between 19000 and 133000. Considering that we apply electrical power instead of cool demand for $P_{cooling}$, the values of parameters should be divided by a coefficient of five in the model. Therefore, the range for $C_{p,air}$ is set to [3800 26600]. We define the range for each parameter by following the same method. Table 5.3 lists for each parameter the researching range for optimization algorithm.

The cooling model to be optimized contains multi variables and has a large searching space, it is hard to generate initial guess values due to limited prior knowledge. Considering the model features, we determine to choose "Differential Evolution (DE)" as the optimization method. DE is a global optimization method, firstly introduced by Storn and Price [SP97]. DE aims to find the global minimum of a multivariate function. Specially, DE can be applied to optimization problems with large searching scale, to find solutions for multiple and constrained objective functions, under dynamic and uncertain environments [DS11]. Users just need to specify the constrains for each parameters (minimum and maximum value), initial guess is not required. DE starts solving the optimization problem by randomly proposing initiated candidates within the large scale searching scale, then focus on searching around several interesting candidates if exist. Comparing to DE, the estimations searched by classic gradient descent local optimization techniques such as Levenberg-Marquardt or Gauss-Newton. The final solution can be greatly influenced by initial values, as the final estimation may converge to local solutions [Dat15]. Local optimization approaches usually requires less time than global optimization approaches to find the solution if initial guess is well provided. However, they are not suitable for finding the possible better solutions which are far from the initial guess. Table 5.4 shows the obtained

Table 5.4. Identification of the thermal capacitance parameters

Thermal capacitance	$C_{a,1}$	$C_{a,2}$	C_c	C_s
Obtained values (J/K)	25120	1200	36263	210700

capacitance values from running the algorithm based on DE.

Identification the thermal conductance parameters

Five training data sets have been used to identify the thermal conductance. For each dataset chosen, P_{server} runs steadily at different average power, varying at about 1.7kW, 1.8kW, 3.8kW, 4.2kW and 6.3kW respectively. Thermal conductance h_c and fluid conductance M_cCp are observed to have different values according to P_{server}. In order to determine the expression for h_c and M_sC_p based on P_{server}. We search the optimal h_c, M_sC_p combinations for each data sets, by fixing the other parameters to a constant that we obtained while searching the capacitance in previous optimization. After that, we apply the logarithmic curve fitting to find the expressions. Figure 5.12 shows the log equations fitting the P_{server} with h_c and M_sC_p that we find under each data set.

In fact, M_cC_p represents the thermal fluid conductance, this parameter relies on the quantity of air flowing during a unit of time. When server power increases, more heat is generated, therefore integrated fans work at higher speed, more air will pass through the server during the same period of time. Value of M_sC_p increases with the rise of P_{server}, the expression that we find match exactly the real situation. In terms of h_c, it increases with P_{server} as well, because heat convection is boosted with higher air temperature.

So far, we have identified all the parameters of the model, table 5.5 presents the value or expressions that we found for each ones.

5.3.3 Validation on the identification process

Once the parameters are determined, we can already verify the correctness by comparing the model estimations results with the real measurements on the data sets that we use during the identification process. Figures from 5.13 to 5.17 shows the estimation results obtained on the trainning data set. The MAPEs (refer to "Model evaluation metric" in section 1.4.2) are less than 4% for estimations of T_i and 3% for T_o.

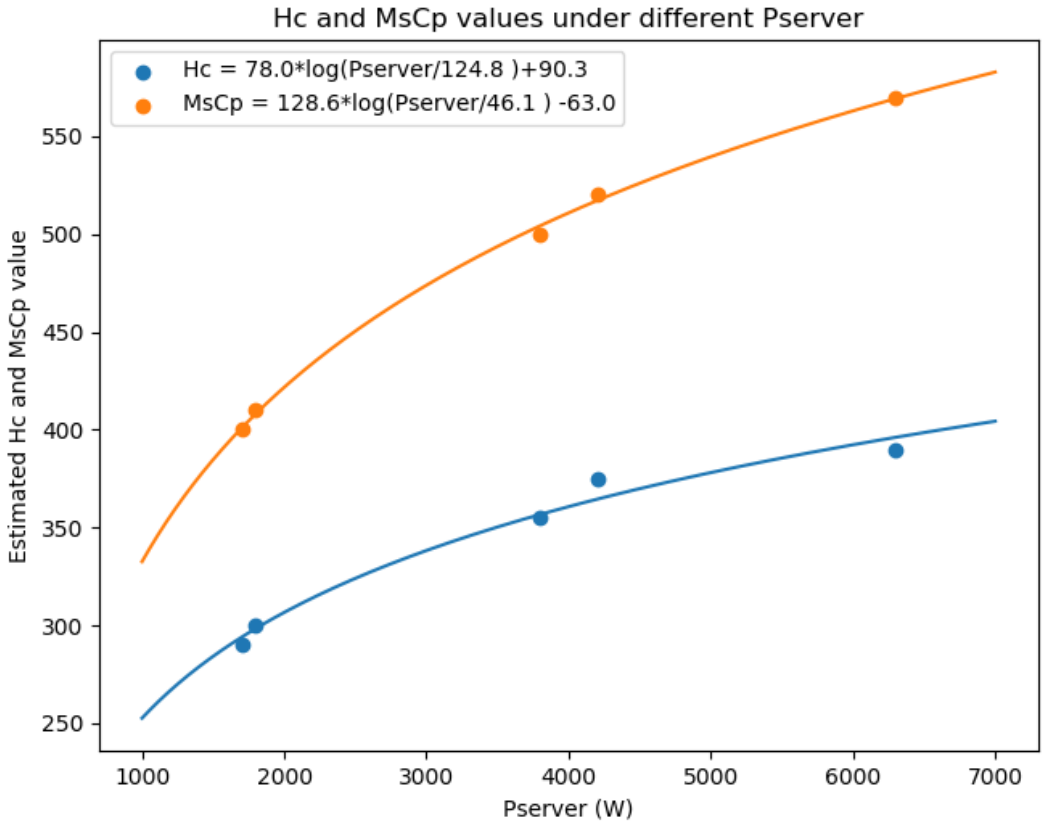


Figure 5.12. Values of Hc and $MsCp$ based on P_{server}

Table 5.5. Obtained values or expressions for system parameters

Parameters	Searching range	Obtained value or expression
$C_{a,1}$	[3800, 26600]	25120
$C_{a,2}$	[175, 1550]	1200
C_c	[23300, 186400]	36263
C_s	[200000, 2000000]	200700
M_sC_p	[301, 903]	$128.6 \times log(P_{server}/46.1) - 63$
h_r	[0.04, 3]	1
h_c	[80, 4860]	$70 \times log(P_{server}/124.8) + 90.3$
h_s	[0, 2]	0.4
h_o	[0, 2]	0.44

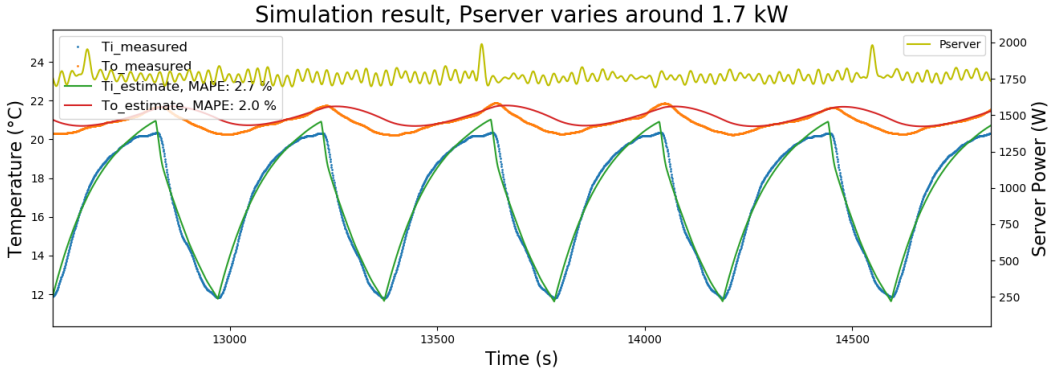


Figure 5.13. Estimation result on trainning set 1: P_{server} runs around 1.7kW

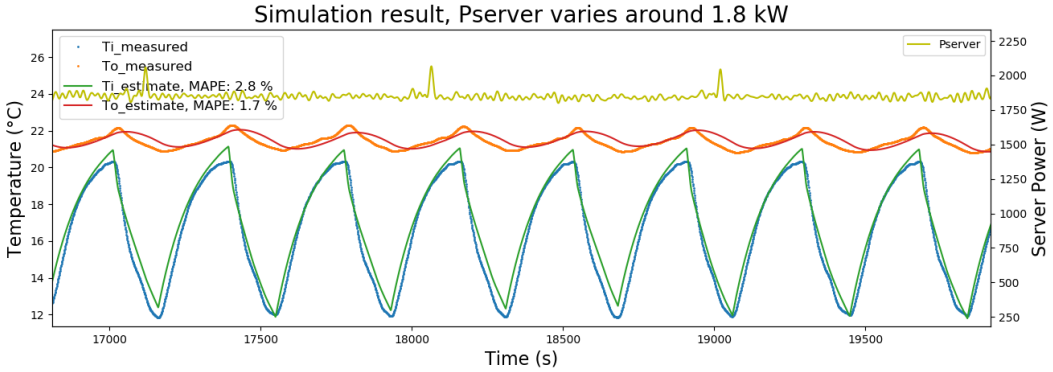


Figure 5.14. Estimation result on trainning set 2: P_{server} runs around 1.8kW

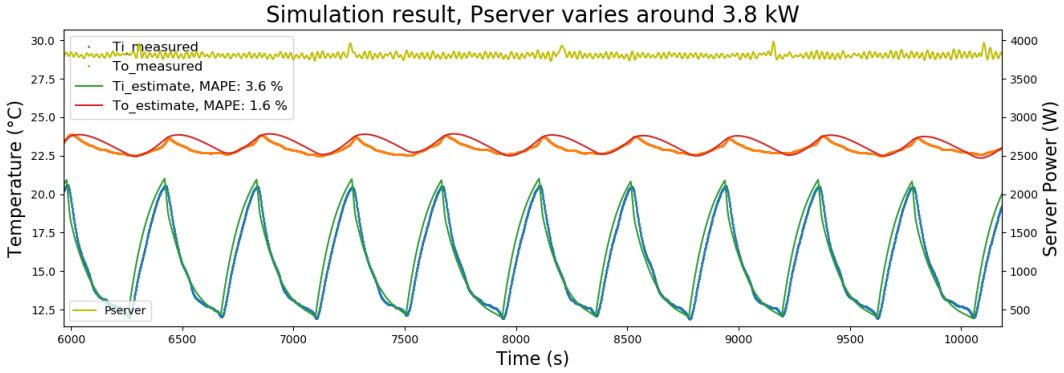


Figure 5.15. Estimation result on trainning set 3: P_{server} runs around 3.8kW

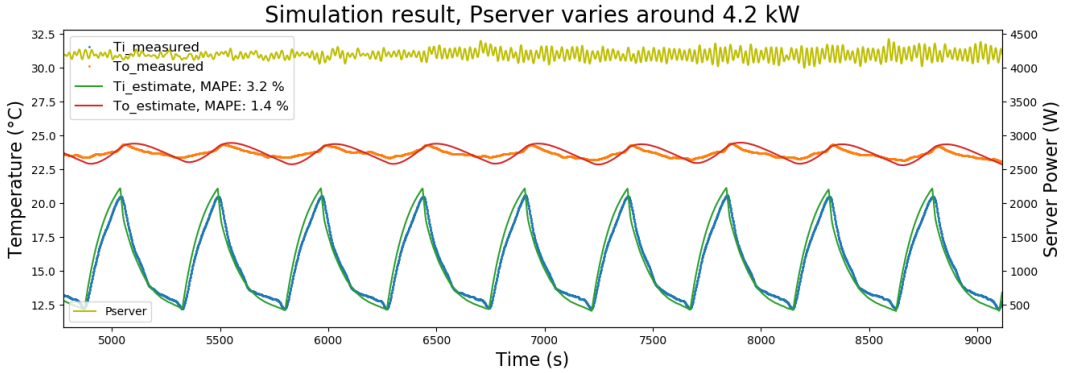


Figure 5.16. Estimation result on trainning set 4: P_{server} runs around 4.2kW

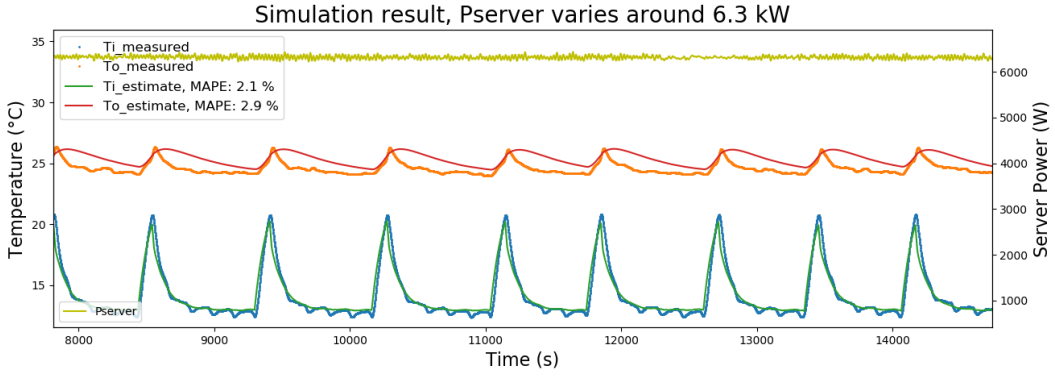


Figure 5.17. Estimation result on trainning set 3: P_{server} runs around 6.3kW

In addition to the training set, we choose another data set that we didn't use during the training processing to complete the model validation. This data set consists in 33 hours of the measurements. In this data set, P_{server} varies dynamically between 4.2kW and 6.8kW. The validation result is shown in figure 5.18. The MAPE for T_i and T_o are 3.4% and 2.1% respectively. The estimation result starts to have more distance from real measurements after about 40000s, where T_r starts to have more fluctuations. It seems that one or more parameters may also vary according to T_r. We will identify these parameters and optimize the model in a future work.

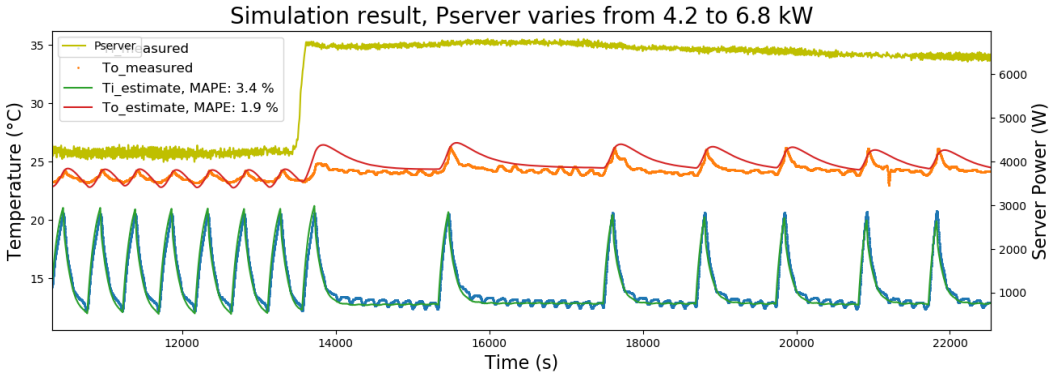


Figure 5.18. Estimation result: P_{server} runs between 4.2 and 6.8 kW

5.4 Modeling cooling power of cluster based on inlet temperature

Previous section introduces modeling the evolution of T_i based on operating variables of the cluster: P_{server}, $P_{cooling}$ and T_r. In order to get the estimation of $P_{cooling}$ directly from P_{server} and T_r, the relationship between T_i and $P_{cooling}$ is required, we have already explained the details in section 5.2.3. In this section, we present our study on modeling the $P_{cooling}$ based on T_i.

Actually, as we explained in section 5.2.3, T_i plays as a trigger for cooling module of In-Row. The cooling power is therefore modeled based on four parameters:

— $T_{i,low}$: Compressor stops working once T_i is cooled down to $T_{i,low}$ (value is about $12.4°C$).

— $T_{i,high}$: Compressor starts working once T_i goes up to $T_{i,high}$ (value is about $20.3°C$).

— $P_{cooling,low}$: Compressor is working, power of In-Row is about 4600W.

— $P_{cooling,high}$: Compressor is not working, power of In-Row is about 260W.

It has been observed as well that, at the beginning of restart of compressor, In-Row runs suddenly at a peak power (between 7kW to 9kW), but during a very short of time, which is a normal

behavior of electrical motors. Moreover, the peak power varies randomly, hard to analyse and estimate. We didn't consider modeling this peak power in our model, energy consumption is the production of power and time, the peak power has little influence on the final result, and we are capable to compensate this peak by increasing a little the value of $P_{cooling,high}$. Figure 5.19 illustrates our basic idea of modeling the $P_{cooling}$ based on T_i in a simple way. In the figure, $P_{cooling}$ represents for real cooling power and $P_{cooling_e}$ represents for the estimated cooling power from model. In this model, $P_{cooling}$ switches between two values: $P_{cooling,low}$ and $P_{cooling,high}$, turning points are determined by the predefined values of $T_{i,low}$ and $T_{i,high}$, which are caused by the working states (on or off) of compressor.

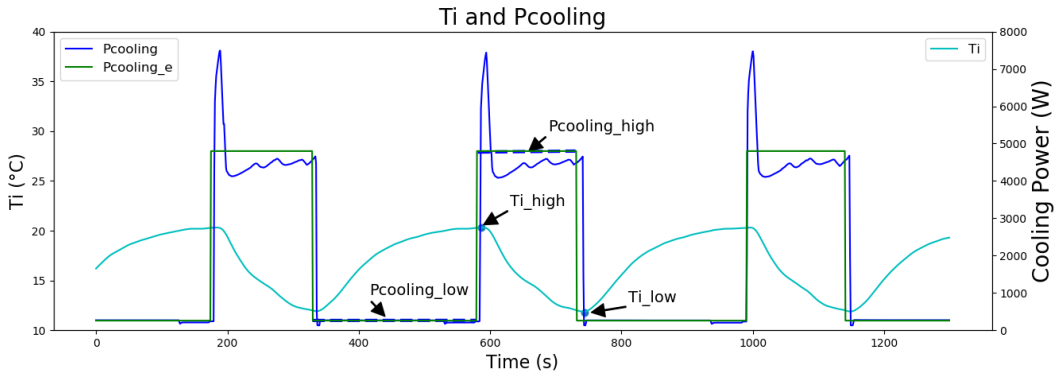


Figure 5.19. Idea of modeling $P_{cooling}$ based on T_i

The cooling can be determined by identifying the four parameters according to operating variables.

5.4.1 Identification

In this section, we present our method concerning identifying $P_{cooling,low}$, $P_{cooling,high}$, $T_{i,low}$ and $T_{i,high}$. We use the same data sets that have been used for building T_i model. Among them, $T_{i,low}$, $T_{i,high}$ and $P_{cooling,high}$ will increase with the rise of P_{server}, even though the increment is very small, the values of turning point have great impact on the precision of cooling model. After several experiments, we decided to assign different values to each parameter based on P_{server} range. For example, in data set 1, servers run steady at around 1.7kW, we identify firstly all the turning points (the sets of $T_{i,low}$ and $T_{i,high}$) in data set 1, part of the result is shown in figure 5.20.

In this data set, P_{server} varies between 1592W and 2236W and has an average power of 1762W. $T_{i,low}$ of each cycle varies between 11.72 °C and 11.95 °C and $T_{i,high}$ varies between

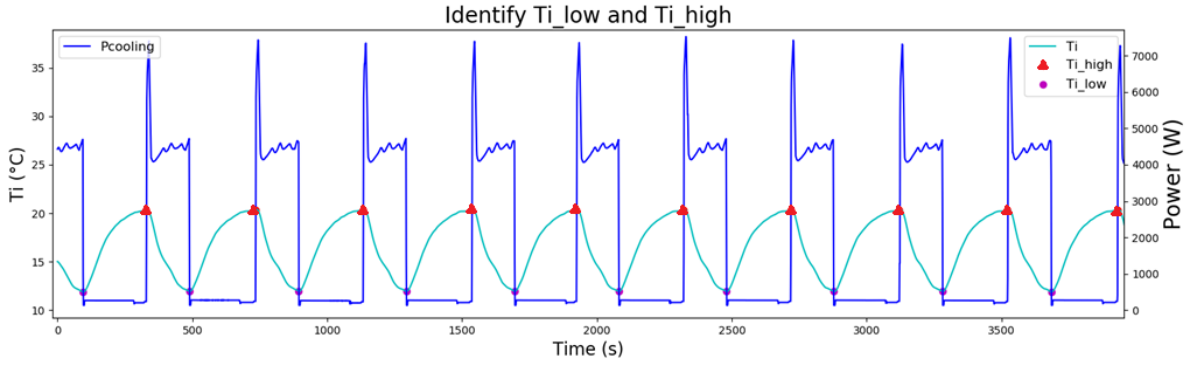


Figure 5.20. The values of $T_{i,low}$ and $T_{i,high}$, P_{server} runs at around 1.7kW

20.30 °C and 20.39 °C. In order to contain all the turning points, the value of $T_{i,low}$ in the model need to be set a little higher than the maximum result of the $T_{i,low}$ in the data set. For the cycles in the data set, if $T_{i,low}$ is set too low, sometimes, the T_i cannot be cooled down to the set temperature, cooling model will consider that the compressor should keep working and maintain the cooling power at $P_{cooling,high}$. Eventually, cooling model will miss several cycles and loss the precision. Similar to $T_{i,low}$, the value of $T_{i,high}$ in the model is set a little lower than the minimum result of the $T_{i,high}$ in the data set. Therefore, we set 12 °C for $T_{i,low}$ and 20.3 °C for $T_{i,high}$. We follow the same method and identify the best $T_{i,low}$ and $T_{i,high}$ for the rest of the data sets. The turning points for cool model are therefore defined based on the range of P_{server}. In terms of $P_{cooling,low}$ and $P_{cooling,high}$, they do not have much variation across different P_{server} range, average value are selected to fit the data.

An exception: servers run at higher power The models built with initial parameters fit well the training sets where P_{server} below 4700W. However, the model didn't fit very well the data of training set 5. In this data set, P_{server} varies between 6077W to 6693W. For previous data sets, $T_{i,low}$ are equal or less than $12.4°C$ and they are always the minimum temperature value at each cycle. However, if we investigate the T_i values of data set 5, it has been found that, for servers run at higher power, heat will be generated by servers at higher rate, in this situation, the cooling system may not be powerful enough to cool down the $T_{i,low}$ below the low temperature set limit $12.4°C$ ($T_{i,low}$ is about $12.4°C$). In this case, the working state of compressor is hard to estimate: compressor may stop at a T_i temperature higher than $T_{i,low}$. We show an example in figure 5.21, where we zoom in to a part of data set 5. Blue line represents the real cooling power and estimation of cooling power from the model is in green. On the first cycle, the minimum temperature of T_i in this cycle is $12.42°C$, but it is still higher than $T_{i,low}$ of $12.4°C$. Instead of

stopping at 12.42°C, compressor continues to work for a while and stops when T_i is at 12.51°C. However, our model is not able to estimate this behavior, for the first cycle, the compressor in our cooling model still stops when T_i reached 12.42°C, that results in an under estimation of the energy consumption due to this exception period. The third cycle shown in the figure 5.21 has the same problem. Actually, the real range of $T_{i,low}$ in data set 5 is between 12.43 and 12.80°C. We set the $T_{i,low}$ value as the minimum temperature achievable during the data set in order to have lowest loss. In addition, we increase the value of $P_{server,high}$ to 5000W so as to compensate the possible energy loss during the exception period.

Table 5.6. Parameters of cooling model based on T_i

P_{server} range(W)	[0 1800]	[1801 2500]	[2501 6000]	> 6000
$T_{i,low}$	12	12.11	12.4	12.42
$T_{i,high}$	20.29	20.29	20.40	20.42
$P_{cooling,low}$	260	260	260	260
$P_{cooling,high}$	4800	4800	4800	5000

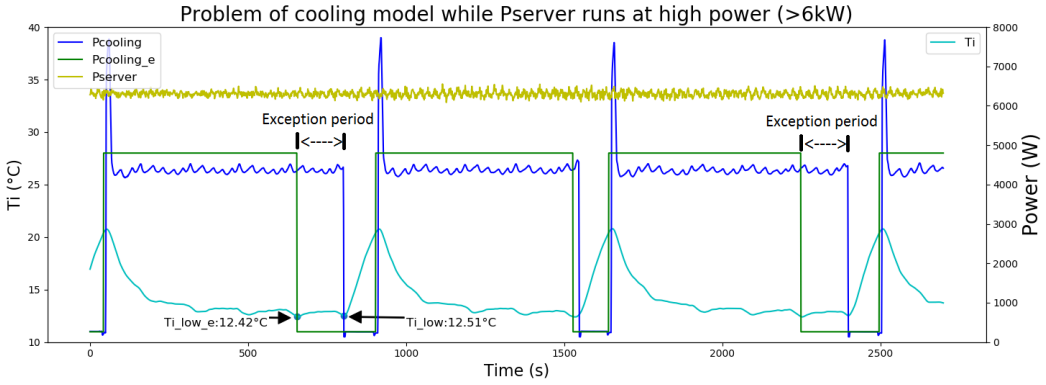


Figure 5.21. Cooling model challenge: servers runs at high power

Table 5.6 shows the values of the parameters chosen for different P_{server} ranges.

5.4.2 Validation

We evaluate the cooling model by using the same data sets in previous section. Electrical energy consumption is determined as the sum of all the instant cooling powers (power data is recorded per second). The error of the cooling model is calculated by computing difference

percentage error between the measured and the estimated energy consumption. The formula used to calculate the percentage error result is shown in equation 5.10:

$$Error_{cooling} = \frac{\sum_{t=0}^{n} P_{cooling}^{estimated}(t) - \sum_{t=0}^{n} P_{cooling}^{measured}(t)}{\sum_{t=0}^{n} P_{cooling}^{measured}(t)} \times 100\% \tag{5.10}$$

Same as the validation process for T_i model, we have firstly evaluate the model correctness directly from the training data sets that have been used to determine the model parameters. Figures 5.22 to 5.26 show the validation results on the training data sets. Model error is calculated by using all the data in the data set, as there are lots of cycles, only part of the data has been shown in the figures. As shown by the figures, model has a percentage error of -3.17% for training set 5. For the rest of the data sets, the percentage errors are all less than 1.1%. Later on, we validate as well the model on the data set excluded from the training process. As shown by the figure 5.27, the model has a percentage error result of -3.21%. **The current model shows better result for servers run at lower power (less than 4.7kW), otherwise, the accuracy can be influenced by unexpected periods as we showed in 5.4.** In general, the percentage error is kept at an acceptable level for all cases (less than 3.3%). Further work can concentrate on these exceptions occurred while servers run at high power, the accuracy of model could be improve further by correcting the unexpected energy loss.

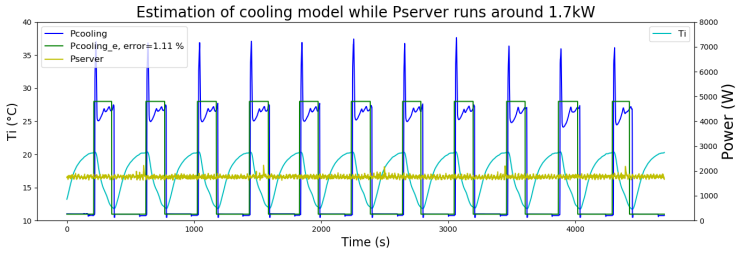


Figure 5.22. Estimation result on trainning set 1: P_{server} runs around 1.7kW

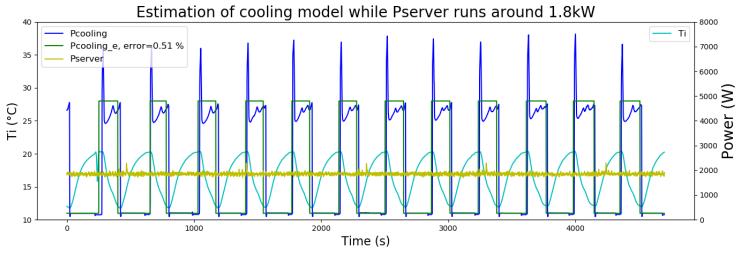


Figure 5.23. Estimation result on trainning set 2: P_{server} runs around 1.8kW

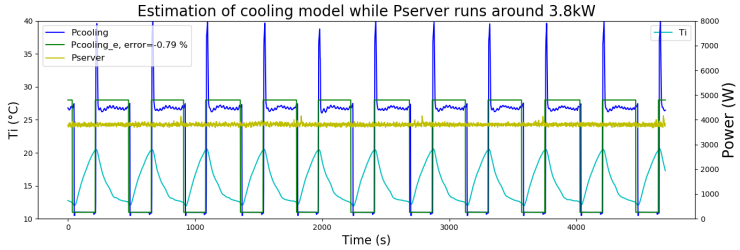


Figure 5.24. Estimation result on trainning set 3: P_{server} runs around 3.8kW

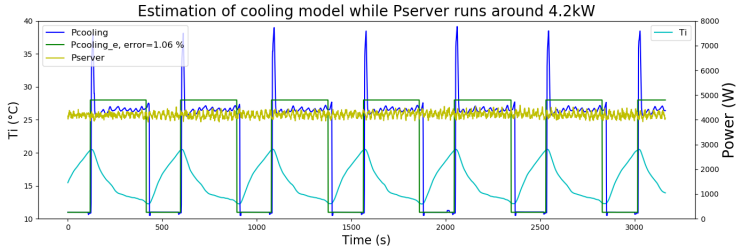


Figure 5.25. Estimation result on trainning set 4: P_{server} runs around 4.2kW

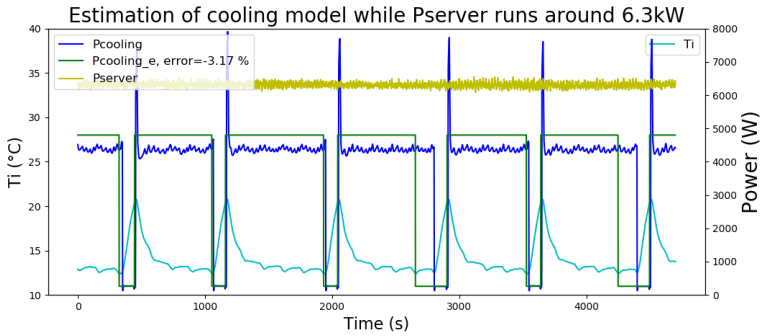


Figure 5.26. Estimation result on trainning set 5: P_{server} runs around 6.3kW

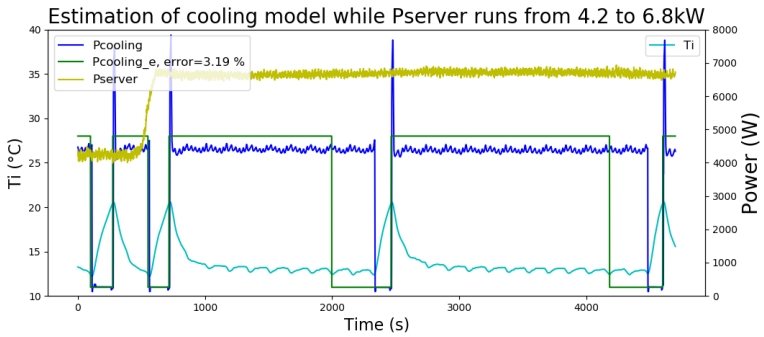


Figure 5.27. Estimation result on data set exclude from training: P_{server} runs dynamically from 4.2 to 6.3kW

5.5 Real time and global power consumption modeling of cluster: Further work

According to the initial idea described in section 5.1, the global cooling model can be realized in three steps. We have completed at the moment the first and second step. In terms of the first step, the model simulated the thermal system behavior has been realized in section 5.3. Taking use of this model, we are able to estimate the value of $T_i(t)$ by providing $P_{server}(t)$, $P_{cooling}(t)$ and $T_r(t)$. After that, in section 5.4, we present our work on modeling the cooling power consumption based on the description of second step. The model estimates the $P_{cooling}(t)$ according to $T_i(t)$ and the corresponding cooling system configuration parameters:$T_{i,high}$, $T_{i,low}$, $P_{cooling,high}$ and $P_{cooling,low}$. The final model is expected in the third step, to estimate the global power consumption $P_{total}(t)$ by feeding directly with operating variables $P_{server}(t)$, $T_r(t)$, and the cooling system configuration parameters: $T_{i,high}$, $T_{i,low}$, $P_{cooling,high}$ and $P_{cooling,low}$. The model is combined by the previous two models for the estimations of T_i and $P_{cooling}$. The idea of building the final model is described as follows: firstly, we provide initial $P_{cooling}(0)$ and $T_i(0)$ to the system, in order to initialize the working environment of the cooling system. In terms of setting the initial values for $P_{cooling}(t)$ and $T_i(t)$ at $t = 0$, in fact, what we expect from the model is estimating the total electrical power consumption, $P_{cooling}(t)$ starts with either $P_{cooling,high}$ or $P_{cooling,low}$ has little importance to the final estimation result. Similarly, the start temperature of T_i could be set randomly between thresholds $T_{i,low}$ and $T_{i,high}$ as well. Once having initial the $P_{cooling}(0)$ and $T_i(0)$, $T_i(t)$ could be generated by using the model realized in the first step. After that, the value of $P_{cooling}(t)$ would be switched between $P_{cooling,high}$ and $P_{cooling,low}$ at the moment $T_i(t)$ reaches the thresholds. Finally, the total cooling energy consumption during the period could be obtained by calculating the sum of $P_{cooling}(t)$.

Further more, the value of $P_{server}(t)$ could be actually replaced by real time server activity related data, such as the current load, number of active VMs, etc. Then, a model estimating the total power of the cluster can be realized by providing operation/activity related data, and the cooling system configuration parameters, as show by the equations 5.11 to 5.13.

$$
\begin{aligned}
config &= [T_{i,high}, T_{i,low}, P_{cooling,high}, P_{cooling,low}] \\
P_{server}(t) &= f(Load_{servers}(t)) && (5.11)\\
P_{cooling}(t) &= f(P_{server}(t), T_r(t), Pc_{config}) && (5.12)\\
P_{total}(t) &= P_{server}(t) + P_{cooling}(t) && (5.13)\\
&= f(Load_{servers}(t), T_r(t), config)
\end{aligned}
$$

A model estimating the real time global power consumption of a physical cluster is very helpful, once realized, much work can be realized in using the model. One can vary certain operating conditions (different T_r, cooling configurations such as T_i thresholds) and compare the simulated power consumption between different conditions. This allows us realizing a full optimization of the global energy power consumption for the cluster. Unfortunately, restrained by deadline, we haven't completed the final model. The final model is expected to be optimized and completed in further work.

5.6 Conclusion

In this section, we describe our method of building a cooling model for a physical cluster. The cooling of the cluster is provided by a specific cooling system called "In-Row", it is a high efficiency cooling equipment which is widely adopted for data centers equipped with large clusters. The cluster we modeled contains five racks, there are four server racks, with a total of 48 servers and an In-Row rack. The whole cluster is placed within a closed container which avoids air and heat exchange between the racks and the outside room. The In-Row rack is installed in the middle of the four server racks. Basic function of In-Row is to take in the hot air generated by servers at the hot aisle, cool down the hot air then release the cold air to the cold aisle. Compressor is the key component of the cooling module of the In-Row, which consumes most of the energy required by In-Row. In order to save energy, compressor does not work all the time, the working state of the compressor is activated and deactivated according to the temperature of air released to the front side (T_i), which is measured and recorded by a sensor installed along with the In-Row. Compressor restarts and stops in an infinite loop when T_i reaches the low and high temperature thresholds. This phenomenon indicates that, if we know how T_i varies according to the operating variables of the system, the cooling power can be estimated based on the value of T_i. Therefore, in our point of view, a global power consumption

model can be realized in three steps.

In the first step, we try to estimate the real time T_i value based on the following operating variables: power of total server $P_{server}(t)$, total cooling electrical power of In-Row ($P_{cooling}(t)$) and the room temperature ($T_r(t)$). The estimation can be achieved by building the thermal system model for the cooling system of the cluster, with respect to the heat balance within the system. After analysing the elements of the system, we propose a simplified equivalent RC circuit to mimic the heat transfer among different objects of the cluster. The simplified one-layer model has been be proposed by a group of ODEs (Ordinary Differential Equation). However, as the thermal model has been much simplified (for example, all the servers are represented with one point, resistances are mixed with thermal conductance, convention and radiation, each point has different mass and volume behind them, etc), the parameters in the model represent complex thermal and physical properties, it is impossible to get their values in a direct way in this case. Therefore, we need to identify the values of the parameters in the model by using the available real measurements of the system. We propose in this situation a global optimization method to solve the parameters in the model. Moreover, two parameters in the model (h_c and M_sC_p) are found to vary with $P_{server}(t)$. We retrieved eventually the expressions of these parameters by curve fitting the best solutions under different $P_{server}(t)$ range. Our model have been validated on all the training data sets and a data set excluded from training process. The MAPE results are less than 4% and 3% for estimations of T_i and T_o comparing to the physical measurements. Especially, this thermal model of the physical cluster has been realized with limited computational resources (one HP laptop with four Intel i5-7300@2.6GHz cores). We believe that, the model can be more accurate with more details, for example: building the thermal model with multiple layers, providing more physical temperatures such as the temperatures at different positions of the cluster and room, which requires obviously, more computational resources. Thus, our approach demonstrate a feasible simpler way to model such a system, with acceptable error range.

In the second step, we build the cooling power consumption model of In-Row based on T_i values. The principle objective is to determine four parameters of the model: the low and high temperature thresholds which switch the on and off working states of compressor ($T_{i,high}$ and $T_{i,low}$), and the power of In-Row during the activation and deactivation of compressor ($P_{cooling,high}$ and $P_{cooling,low}$). Actually, ($T_{i,high}$ and $T_{i,low}$) will vary within different P_{server} ranges. We train the model by using data sets with different range of $P_{server}(t)$, and determine the precise $T_{i,high}$ and $T_{i,low}$ values according to different $P_{server}(t)$ ranges. Model has been then validated by calculating the difference of energy consumption obtained from the model and the

measured one, the result is presented by percentage error. Model shows better fitting with the data while servers run under about 6kW. For higher power, cooling system may not be capable to cool down the T_i to a desired temperature (12.4°C). In this case, instead of stopping working at the minimum temperature in the cycle, compressor will continue to work for a while and stop at a higher T_i temperature. According the observations, this exception happens occasionally, and the cooling model is not capable to detect and react properly at these exceptions yet. In general, we have a absolute percentage error less than 3.3% for all the validation data sets. As further work, the study can concentrate on dealing with these excepted situations, so as to improve further the model accuracy.

In terms of the third and final step, we have not completed the final model estimating the global power consumption of the cluster yet, restrained by thesis deadline. The idea of achieving the model has been detailed in section 5.5. The model is expected to complete and optimized in further work.

CONCLUSION

Summary of the Dissertation

Internet is probably the largest thing built by us, as the physical infrastructure, data centers takes unsurprisingly tremendous amount of energy to our needs in digital life. According to recent predictions, in 2025, share of global electricity used by data centers is predicted to account for no less than 4.5%, and 3.2% of the total carbon emissions [And17]. Building "green" data centers and reducing environmental impact become a great concern, especially for large Internet companies [C$^+$]. Therefore, for an operator telecom like Orange, it is essential to have a preview of the energy consumption before planning the construction, for both economic and environmental benefits. Energy predictive power model is one of the approaches [VDBDCJ$^+$14]. Much work has been done to realize an energy predictive model for servers, we have introduced some of the representative work in section 1.4.1 and 1.4.2. However, according to our experiments in section 2.2.3, for a physical homogeneous cluster, a maximum of 7.8% power variation has been found between 12 identical servers under the same load. This observation indicates that, the accuracy of a power model built upon one server can be questioned if applied to the other identical ones.

Based on this finding, we try to identify the underlying causes for the power variation between identical computing systems. We find through other experiments that, fluctuating neighboring temperature can vary the average power of a same server to 5.6% , with execution of the same workload. Previous studies have also emphasized the power variation brought by temperature changes, as mentioned in 1.5.3. However, few work addressing about how temperature variations affect the power of computing systems. We try to provide more concrete explanations in the study described in section 2.4. The study has been performed on two parts of physical servers: the influence brought by temperature variation from CPU and from the other components. During the experiments, we found ways to only vary the surface temperature for one part while keeping the temperature of the other part constant. The results demonstrate that: the rise of the ambient temperature can increase the power consumption of servers in two ways: through the consumption of cooling system (integrated fans) and static power dissipated by CPU. The evaluation proved as well that, except CPU, the other components in the server are almost not

sensitive to ambient temperature variation. Moreover, we have correlated the power of server with the static power of CPU. The static power dominated by leakage current increases dramatically fast with the rise of ambient temperature. Taking a Gigabyte server for example, during a stable CPU intensive application execution, server power experienced a 16% rise by only raising CPU temperature.

Fabrication process discrepancy can be another cause for the power variation observed among identical servers. Relevant studies are detailed in section 1.5.4. However, due to limited samples, there lacks enough physical evidence and deeper exploration addressing how the fabrication process makes the processor samples differs from each other. Therefore, in section 3 we explore this subject by deep exploring the variability between identical processor samples. Two processors of Intel from different generations have been participated in this evaluation. We test 30 samples for each type. Except the CPU samples, environmental variables have been well controlled: samples are switched one by one in the same motherboard, stressed with the same workload, and the whole testbed is placed in a climatic chamber with the same ambient temperature. In this case, the modern type is turn out to have more power variation between samples. Inspired by the previous studies, we propose and analyze two potential possibilities: TIM applied and the parameter of leakage current. However, removing TIM didn't help reducing the variation. After that, we try to exposure the difference of thermal characteristics between samples and finally find the reasons. In fact, with the decrease of lithography size in modern processors, leakage current becomes more important and parameters within leakage current can differ from one to another. This difference affects finally the static power consumption among samples: static power increases with different rate while the rise of CPU temperature. In this study, we have characterized in an innovative way, the power variation brought by the difference of leakage current parameters between processors.

Thermal effects have been explored deeply to servers in chapter 2 and to processors in chapter 3. Later on in chapter 4, we have discussed influence of thermal effects on the accuracy of power models. We evaluate the classical power model based on CPU utilization. We find that, power model can loss much accuracy for server running under different ambient temperature. According to the studies conducted in previous chapters, we propose to take in ambient temperature as another indicator in the model. Other than thermal effects, single indicator based modeling approach has limitation to build reliable model for modern servers: power could vary within a large range for a fixed CPU utilization level. Even with a fixed CPU frequency, utilization is not perfectly linear to power. Polynomial regression function is capable to fit better the real measurements. However, higher degree could lead to eventually over-fitting. In order

to avoid this problem, we have proposed an algorithm to research and determine the best polynomial degree for building power models. Besides power models, in the same chapter, we have also evaluated the reliability of power measurement data obtained from IPMI, Redfish and Intelligent PDU. Recently, these tools are becoming popular and often play an essential role in realizing the entire power management in a data center environment. However, there is not much work discussing the reliability of these tools. We compare the power measurements from these tools with an high accuracy power analyzer. The experiment results show that, the precision of both IPMI and Redfish can differ from different power ranges, usually, the higher the better. After analysing, the loss of accuracy is believed to be brought by the latency from requests between controller and sensors. Comparing to IPMI, Redfish is observed to have less such latency. Besides latency, sometimes the tools under evaluation have not been well calibrated before using. The accuracy was greatly improved after the calibration. This evaluation is not quite included in the scope of our research, but we still hope that, the results presented can provide some references and guidelines for data center operators, when applying IPMI, Redfish or Intelligent PDU as power characterization approach.

In the end, we present our idea of modeling the global power consumption of a physical cluster. The model is expected to estimate the global power consumption of the cluster, by providing operational configurations of cluster and server activity related data, such as room temperature, configurations of cooling system and load of servers. The objective global power consumption includes the power consumed by both servers and cooling system. The cluster adopts the "In-Row" model as the cooling solution, which consists in a direct expansion system with fans and controllers. In-Row has been nowadays widely used for large scale data centers. The whole cluster is placed within a closed cooling container which includes four servers racks with 48 servers and one In-Row rack served as the primary cooling system of the cluster. The final model has been planned to be realized in three steps, and at the moment we have realized the required models in first and second step. The first model concerns stimulating the thermal system of the cluster. It is a one dimension thermal model. The model is able to estimate real-time inlet and outlet temperatures, by providing three system variable: power of total servers, power of cooling system and temperature outside the cluster (room temperature). The thermal model proposed is simplified with one dimensional, it represents the heat transfer between different elements of the system. The thermal model has been built through equivalent RC circuits, heat flux at four positions of the cluster have been established with ordinary differential equations (ODE), with respect to the heat balance. As the thermal model has been much simplified, parameters presented in the model have mixed with complex physical and thermal properties,

it is hard to retrieve their values according to real world physical or thermal rules. In this case, we propose a global optimization method based on differential evolution (DE) to find solutions of the parameters in the model, in order to approach the outputs of the model as close as possible to the physical measurements. For all the data sets, the validation result shows the MAPE values less than 4% and 3% for the estimations of inlet and outlet temperatures. In terms of the second step, we propose a cooling power consumption model based on the evolution of inlet temperature of the cluster. The model is simply composed by four cooling configuration parameters: the high and low thresholds of inlet temperature, which determine the activation and deactivation of compressor in the cooling system. And the high and low cooling power of cooling system. The power of cooling system is observed to switch between the high and low power depends on the working states on and off of the compressor. However, the thresholds of the inlet temperature vary according to the total power of server. In this case, we propose to provide different thresholds temperatures based on the range of total server power. The model has validation results with percentage error less than 1.2% for servers running at lower power (lower than 4.7kW), and less than 3.3% for servers running at higher power. The errors is increased due to some hard-to-estimate exception of the inlet temperature, occurring specially at higher server power. The model can be improved if these exceptions can be well estimated. We consider finding the solutions in further work. In the end, as the third step, a model targets at estimating the global power consumption of the cluster, includes both server and cooling power is expected. However, we haven't completed the final model due to the thesis deadline. The idea and procedure of realizing such a model has been well detailed in section 5.5. We hope that the work can be carried on and the model is expected to be realized and optimized in the future work. Such model will be very helpful in realizing a global power consumption optimization, by taking environmental conditions, operational configurations and IT loads all into consideration.

Power consumption of a data center is a large subject. During the whole research, we spent a lot of time on experimental evaluations, from single elements like processors, servers, measurement devices to a whole functional cluster. The principal objective and direction of the research have been inspired and cleared a lot after analysing the data obtained from these experimental evaluations. Instead of concentrating on the analysis of the activity related data generated by servers, which has been much and well explored during the last decades, we put more focus on the uncertainties of the power consumption brought by external variables: thermal effects led by temperature variation, differences on identical samples caused by imperfect of fabrication processing, accuracy problems come from power measure approaches. Each factor has been deeply studied and evaluated in our research. Besides experimental evaluations, we put forward as well,

the modeling of the global power consumption of a physical cluster, equipped with advanced and widely used cooling system. Generally, traditional data-oriented methods train the model with "big data", and the model is presented as a "black box". Once the initial conditions change, the model has to be re-trained under new environment. We propose to realize the model in a different way. The model has been built from thermal analysis of the whole cluster, it takes in the environmental conditions and system configurations as variables. It is expected to estimate the global power consumption under different environment and/or configuration situations.

Perspectives

During this thesis research, encouraged by research spirit, we've explored a lot the subject in different ways. However, there is still room to enrich the research in several aspects. Constrained by the time and original research scope, we didn't go much further for every aspect that worth digging deeper. In this section, we provide three feasible directions in pursuit of the current work. One of them is being proposed as a postdoc research in Orange Labs.

Static and dynamic power modeling of servers

In reality, due to imperfection fabrication processing, the supposed identical processors are not totally equal to each other in terms of energy efficiency. The evidence indicates that fabrication processing can make the static power differ from one sample to another, concretes in the parameters obtained from the fitting the power with temperature data are different. On the other hand, the same concept can be applied to estimate the static power of the processor from the total server power. When being stressed by a stable IT load, the power increment caused by temperature rise are brought principally by static power from processors (refer to 2.4.2, 2.4.3 and 3.3.2). Besides temperature, static power depends as well on frequency, supply voltage, size and number of transistor. Therefore, modeling the static and dynamic power of a server under different load and environmental temperatures can be realized, by taking the above aspects into consideration. Much work can be done based on the result of the model. For example, operator will be able to locate the servers with higher and lower static power in a homogeneous cluster, and set different priorities for servers in a VM migration design. Moreover, such a model can also provide a different way to predict the power consumption of a VM.

This subject is being proposed as a postdoc research in Orange Labs.

Workload classification based, single indicator power modeling for servers

In chapter 4, we evaluate the accuracy of the power model designed for servers, built by single indicator CPU utilization. The evaluation result show that, server power can be well fitted with CPU utilization by a nonlinear model during the execution of single workload or workloads with similar functionalities. Accuracy can be improved by applying different models for each workload type. This observation provides a possible way to optimize the server power models based on CPU utilization. The steps of realizing the model can be summarized as follows:

— Executing different type of workload on the SUT, while collecting CPU utilisation, power consumption and selective PMCs during execution.

— Classifying the workloads in using PMCs according to *CPU Utilization & Power* curves (can be expressed with same or similar nonlinear regression formula).

— Building the table of classification based on workload type, (CPU types: FLP/INT calculating, sort, compress; Memory & Hard Drive: Read/Write (RW) operations; Network, etc) and each type corresponding to one *CPU Utilization & Power* model.

When new workload coming, determine its type in real-time according to PMC values, then the power can be calculated by specific model in the classification table.

In our point of view, accuracy of power model based on CPU-utilization could be well improved by applying classification during the workload execution.

Real time and global power consumption modeling of cluster

We expect to realize a global power consumption model for physical cluster, as the final step of the work presented in chapter 5. A model estimating the real time global power consumption of a physical cluster is very helpful, once realized, much work can be realized in using the model. The final model is expected to estimate the global power consumption $P_{total}(t)$ by providing operating variables $P_{server}(t)$, $T_r(t)$, and the cooling system configuration parameters: $T_{i,high}$, $T_{i,low}$, $P_{cooling,high}$ and $P_{cooling,low}$. We have ready realized the the estimations of T_i and $P_{cooling}$ in chapter 5. The final model is expected to generate the T_i according to the other working and configuration variables. Steps of building the model has been detailed in section 5.5.

Once realized, operators can preview the further power consumption by simply varying working and operating conditions. Further more, cooling power consumption of the SCS (air conditioning) in the room is also interesting. The cooling model of SCS allows us simulating the seasonal cooling consumption, according to outdoor temperature.

RÉSUMÉ EN FRANÇAIS

Contexte et objectif de la recherche

L'internet des premiers jours est bien loin de celui de nos jours. De quelques messages échangés entre deux ordinateurs, ce dernier traite aujourd'hui des flux d'information toujours plus nombreux et volumineux. Mais au-delà des flux réseaux ce sont bien les services proposés qui ont largement évolués. Le Cloud computing et ses applications en ligne, la video à la demande, les réseaux sociaux, le streaming de jeux ont fait explosé tant les besoins en réseau que ceux en calcul. Ces nouveaux services sont principalement exécutés dans des salles informatiques dédiées, nommées centre de données, composées de dizaine, centaine voir plusieurs milliers d'ordinateurs. Un des sujets de préoccupation industriel, societal et scientifique concerne la consommation électrique mondiale de ces salles.

Cette thèse s'inscrit dans ce sujet et aborde le défi scientifique de la modélisation énergétique d'un centre de données. L'objectif de cette modélisation étant de pouvoir estimer et prédire la consommation de la salle en fonction des paramètres les plus importants. Disposant d'une telle modélisation, un opérateur comme orange aurait la possibilité de mieux repenser/concevoir ses actuels/futurs centre de données.

De nombreux travaux scientifiques s'inscrivent dans ce défi. Cependant, les paramètres importants retenus dans leurs modélisation sont souvent liés aux paramètres technique d'un serveur (activité cpu, ram etc. . .) et ne prennent que rarement en compte des paramètres externes comme la température. Dans la première partie de la thèse, nous avons réaliser un grand nombre d'expérience pour déterminer quels étaient les paramètres importants a prendre en compte dans la modélisation. Une fois déterminé les paramètres, la seconde partie de la thèse propose une modélisation d'un centre de données basée sur les résultats de la première (variables de charge des serveurs, conditions thermiques internes et externes, et variables opérationnelles liés aux configurations du système de refroidissement).

Organisations du manuscrit

Les travaux de recherche présentés dans ce manuscrit se composent de trois parties principales : une dédiée à l'état de l'art, une partie focalisée sur des expérimentations en vu de définir les principaux paramètre a prendre en compte pour la modélisation, cette dernière étant détaillée dans la troisième partie du manuscrit.

— L'état de l'art est présenté en chapitre 1.

— La partie évaluation se compose des chapitres 2, 3 et 4. Dans ces chapitres, les impacts énergétiques des éléments hors charge informatique classique sont identifiés, suite aux nombreuses expérimentations physiques. Les éléments étudiées comprennent : l'environnement physique et thermique associé, la variation entre les processeurs identiques amenés par l'imperfection de la fabrication, les défauts possibles causés par les outils de mesures, etc.

— La partie consacrée à la modélisation est décrite au chapitre 5. Cette étude a pour objectif d'estimer la consommation global d'un ensemble de serveurs physiques, composé dans le cadre d'expérimentation à 48 serveurs refroidi dans un système étanche à l'air (InRow).

Dans la suite de ce résumé, nous décrivons succinctement les travaux réalisés pendant cette thèse.

Résumes des recherches et contributions scientifiques

Etant donné que la consommation énergétique mondiale des centres de donnée augmentent chaque année et atteint des valeurs importantes, ce sujet de thèse fut un des sujets de recherches proposés par Orange Labs en 2016 en écho avec les objectifs de la COP21 : réaliser une réduction de 20% sur l'émission CO2 et de 15% de la consommation énergétique global. L'objectif de cette thèse consiste a permettre l'optimisation de l'efficacité énergétique globale du data center, en proposant une modélisation de la consommation électrique la plus fiable possible. Les défis scientifiques initiaux sont liés à la complexité des interactions entre l'architecture matérielle et l'architecture logicielle ainsi que les éléments externes tel que la température. Pour se faire, nos travaux scientifiques ont commencé par une série d'évaluation pour identifier et caractériser les impacts énergétiques sur la consommation des systèmes informatiques, y compris les impacts matériels et environnementaux tout en analysant les sources d'erreurs issues des instruments de mesure. Nous avons terminé cette étude scientifique par le développement d'une modélisation

global pour un cluster physique donné.

Tout d'abord, nous avons réalisé plusieurs expérimentations physiques et mis en évidence les impacts énergétiques externes sur la consommation des serveurs. Dans un première temps, les tests ont été effectués sur 12 serveurs identiques du site G5K de l'IMT Atlantique à Nantes. Nous avons comparé la variation de leurs consommations en exécutant la même suite de test. Les résultats montrent qu'il a y au maximum 7.8% de différence de consommation entre des serveurs pourtant identiques. Cela démontre que, en plus de la charge informatique, la consommation d'un serveur varie sous certains impacts externes. Les tests complémentaires ont été effectué afin d'identifier ces impacts externes importants. Plusieurs candidats sont évaluées: la construction intrinsèque du serveur, la température de l'environnement (ambiance et sources de chaleur alentour), les positions et agencement des serveurs (espacés ou accolés) et la variation de la tension d'alimentation. Les résultats expérimentaux ont montré que la position, l'agencement et la tension d'alimentation du serveur n'a que peu d'impact sur leur consommation. La variation observée entre les serveurs est liée à la température ambiante et/ou leur construction. Malheureusement, les conditions expérimentales ne nous permettent pas d'étudier la variation liée à la construction (par exemple échanger les processeurs de deux serveurs physiques dont la consommation électrique varie beaucoup). Nous nous sommes dans un premier temps concentré sur l'impact thermique sur la consommation.

Selon nos résultats, la puissance moyenne d'un serveur peut être varier de 5.6% juste en fonction de différente température de fonctionnement et pour une même charge de travail. Les études précédentes ont des observations similaires 2.2.3. Etant donné que la puissance augmente avec la charge du processeur, la chaleur dissipée augmente et la température du composant également. Par ailleurs, cette augmentation de la température induit une augmentation des courants de fuite [KC09] [MB09], qui contribue aussi à une augmentation de la consommation énergétique. Cet impact n'a pas fait l'objet d'études suffisamment vastes (plusieurs types de CPU) ni d'une caractérisation précise, ce que nous proposons de compléter dans cette thèse. Pour cela nous avons étudié la variation de la puissance liée à deux sous-parties d'un serveur : le CPU et les autres composants. Nous avons proposé une méthode pour faire varier la température d'une partie tout en gardant une même température pour l'autre partie. Trois serveurs équipés de différentes générations de CPU ont été retenus pour cette évaluation. Cette étude démontre par des mesures expérimentales que la température du CPU peut induire une variation importante dans la consommation électrique des serveurs. Par exemple, pour des serveurs basés sur un CPU Intel Xeon v3 (E5-2609v3), la puissance électrique augmente de 16% lorsque nous augmentons seulement la température du CPU. Des expérimentation supplémentaires montrent

que l'influence de la température des autres composants sur la consommation du serveur peut être négligée. De plus, nous avons corrélé la puissance du serveur avec la puissance statique du processeur. La puissance statique dominée par le courant de fuite augmente considérablement et rapidement avec l'élévation de la température ambiante. Il apparaît donc que la température du milieu ambiant impacte sur la consommation des serveurs via deux composants : la consommation des ventilateurs et la consommation du CPU et des courants de fuite. Ces résultats sont importants pour le développement de modèles de prédiction de la consommation énergétique car ils indiquent que pour établir un modèle précis, tenir compte de la charge informatique du CPU ne suffit pas ; la connaissance de la température des cœurs doit aussi être prise en compte. Ils permettent également d'insister sur l'intérêt des méthodes de refroidissement liquides qui permettent de maintenir des températures de CPU plus basses, et dont l'impact sur la consommation du serveur a été sous-estimé.

Dans un second temps, nous avons complèté l'étude de la variation de la consommation entre serveurs identiques en évaluant l'impact de leur fabrication. Dans cette étude nous nous sommes concentré sur l'impact de la fabrication du composant le plus consommateur : le processeur. Ces études démontrant que deux processeurs identiques peuvent consommer différemment sont détaillées dans la section 3.2. Cependant, due au nombre d'échantillons processeur limité, il manque encore quelques expérimentations physiques et une exploration plus approfondie pour expliquer le pourquoi de cette différence de consommation.

Dans cette étude détaillée en chapitre 3, nous avons élargi la variabilité entre des processeurs identiques en imposant un environnement thermique. Deux types de processeur d'Intel de générations différentes ont participé à cette évaluation. Nous avons testé 30 échantillons pour chaqu'un des types. Le principe fût de définir un banc de test commun a toutes les expérimentations (température, carte mère, voltage ...), celles consistant a réaliser la même charge processeur sur les 30 échantillons. Nous avons montré que les processeurs moderne, pour une même génération, possédaient plus de variabilité dans leurs consommations électriques qu'en les anciens modèles.

Inspiré par les études précédentes, nous avons analysé deux hypothèses pour expliquer cette variation : le TIM (Thermal Interface Material) et les courants de fuite. Cependant, la suppression du TIM n'a pas aidé à réduire la variation. Avec la diminution de la taille de la lithographie dans les processeurs modernes, les courants de fuite devienent de plus en plus important mais surtout leurs quantités peuvent varier d'un processeur à l'autre. Ces courants de fuite varient sensiblement en fonction de la température. Dans cette étude, nous avons proposé une méthode pour caractériser le paramètre de courant de fuite du processeur, et nous considérons qu'il est

la raison principale de la variation de puissance entre les processeurs identiques et ceci due à l'imperfection de la fabrication.

Avant d'adresser la modélisation d'une infrastructure complète, nous avons évalué la modélisation de la consommation du serveur basé sur l'utilisation du processeur. Nous avons évalué sa fiabilité sous différentes température ambiante. En termes d'effet thermique, nous constatons que ce type de modèle peut perdre de sa précision si le serveur fonctionne sous différente température ambiante. Afin de corriger ce perte de précision, nous proposons de prendre la température ambiante comme une autre variable du modèle : l'augmentation de la consommation à cause d'un changement de condition thermique est modélisé en fonction de la température ambiante. La précision de ce nouveau modèle de consommation au niveau serveur prenant en compte la température ambiante permet d'avoir une estimation bien plus précise.

Cependant, toute la difficulté dans cette modélisation concernait le degré polynomial a étudier dans la régression linéaire. Augmenter le degré augmente la précision pour les valeurs données mais peut entraîner un effet de « sur-estimation ». Afin d'éviter ce problème, nous avons proposé un algorithme pour déterminer le meilleur degré polynomial pour les modèles à générer. Enfin, dans le même chapitre, nous avons également évalué la fiabilité des données obtenues par trois outils de mesure de puissance, qui sont largement utilisé dans un environnement de data center : IPMI (carte mère), Redfish et au niveau des PDU. Ces outils permettent d'obtenir les données de consommations en utilisant les wattmètres intégrés soit dans les serveurs (IPMI, Redfish) ou soit dans la « multi-prises » PDU. Si ces outils sont devenus populaires, peu d'études ont été réalisées sur la fiabilité des outils. L'idée principale de notre étude est de comparer les mesures de puissance fournit par ces outils avec un analyseur de puissance de haute précision. Les résultats de l'expérience montrent que la précision de l'IPMI et du Redfish peut être dégradés lors de mesure de consommation faible. Après analyse, on pense que la perte de précision est provoquée par le temps de latence des requêtes entre le contrôleur et les capteurs, car il y a souvent plus de variation sur des serveurs fonctionnant à faible puissance. On constate également que Redfish a moins de latence qu'IPMI. Outre la latence, les outils n'ont parfois pas été bien calibrés avant installation. La précision a été grandement améliorée après le calibrage. Finalement, nous présentons notre étude sur la modélisation de la consommation énergétique globale d'un cluster physique. La modélisation permet d'estimer la consommation énergétique globale du cluster en fonction des configurations opérationnelles et des données relatives à l'activité informatique, telles que la température ambiante, les configurations du système de refroidissement et la charge des serveurs. Dans cette étude, la consommation d'énergie globale à estimer inclut l'énergie consommée par les serveurs et le système de refroidissement.

Le cluster adopte le modèle de refroidissement nommé "In-Row" proposé par Schneider, qui consiste au refroidissement direct avec ventilateurs et contrôleurs au sein d'un ou plusieurs racks étanches à l'air. Ce type de système de refroidissement est largement adopté aujourd'hui dans les centres de données à grande échelle. L'ensemble du cluster est étudié est disposé dans un conteneur fermé, il comprend quatre baies de serveurs avec 48 serveurs et un baie d'In-Row servant de système de refroidissement principal du cluster. La modélisation de la consommation globale finale a été planifié sur deux étapes. La première étape concerne l'établissement d'une modélisation du système thermique pour le cluster, elle prend en entrée trois variables en temps réel : la puissance totale des serveurs, la puissance de refroidissement et la température en dehors du cluster (température ambiante de la salle), et estime les températures aux différentes positions du cluster. Après avoir analysé les données en temps réel du cluster, nous proposons un système thermique simplifié pour représenter le transfert de chaleur entre les éléments du système. Le système thermique a été présenté par des circuits électroniques de Résistance-Conduance équivalents, des flux de chaleur en quatre points ont été établis avec des équations différentielles ordinaires (ODE) : les températures à l'entrée et à la sortie des serveurs, les températures des serveurs et du conteneur. Les ODEs concernent le bilan thermique au sein du cluster. Comme le système thermique a été très simplifié, les paramètres présentés dans le modèle sont mixés à des propriétés physiques et thermiques complexes liés aux différents dimentions (volumes et masse). Il reste difficile d'estimer leurs valeurs, conformément aux règles fondamentales de la physique ou de la thermique. Pour ce cas, nous proposons une méthode d'optimisation globale basée sur l'évolution différentielle (DE) pour trouver des solutions aux paramètres du système proposé. L'idée principale est d'approcher les températures estimées à l'entrée et à la sortie des serveurs de la simulation au plus près possible des mesures physiques. La précision de cette simulation a été validée par plusieurs jeux de données avec des caractéristiques différentes. Les résultats de la validation indiquent que les valeurs MAPE (Maximum Average Percentage Error) sont inférieures à 4% pour les estimations des températures à l'entrée et 3% pour celles de la sortie. En ce qui concerne la deuxième étape, nous proposons une modélisation de consommation d'énergie du système de refroidissement basée sur l'évolution de la température à l'entrée des serveurs. On observe que la puissance du système de refroidissement bascule entre les puissances haute et basse dépend des états de fonctionnement du compresseur. La modélisation se compose donc de quatre paramètres de configuration du système de refroidissement: les seuils haut et bas de la température d'entrée, qui déterminent l'activation et la désactivation du compresseur du système de refroidissement, et les deux puissances haute et basse du système de refroidissement avec et sans fonctionnement du compresseur. Cependant,

les seuils de température d'entrée varient en fonction de la puissance totale des serveurs. Afin de tenir compte de ce phénomène, nous proposons de définir différents seuils de température en fonction de la plage de puissance totale du serveur. Les résultats de la validation de la modélisation finale présente un pourcentage d'erreur de moins de 1,2% pour les jeux de donnée avec la puissance totale des serveurs inférieure à 4,7 kW, et moins de 3,3% pour la puissance totale supérieure à 4,7 kW. Les erreurs sont dues à certaines exceptions difficiles à estimer, issue du fonctionnent du compresseur, et se produisant lorsque la consomation des serveurs est élevée. Le modèle peut être amélioré si ces cas particulier peuvent être correctement modélisé, sujet à nos travaux futurs. Pour nos travaux futurs, nous adresserons La modélisation de la consommation énergétique globale du cluster, ce qui inclut à la fois la consommation total des serveurs et la consomation du système de refroidissement. Ces futurs travaux sont détaillés en section 5.6.

Conclusion

Estimer la consommation énergétique d'un data center est un sujet très complexe. Durant toute cette recherche, nous avons volontairement mettre un accent important sur des évaluations expérimentales, à partir d'éléments individuels tels que processeurs, serveurs, outil de mesure ou cluster physique. Ces expérimentions ont pris un temps très important pendant la durée de la thèse mais ont permis d'ouvrir de nouvelles pistes sur la moddélisation de la consomation des salles informatiques. En plus de se concentrer sur l'analyse des données relatives à l'activité informatique des serveurs, qui a été beaucoup explorée au cours des dernières années, nous nous sommes concentrés sur les incertitudes de la consommation énergétique induite par les variables externes: effets thermiques induits par variation de la température ambiante, différences d'efficacité entre des processeurs identiques causées par un processus de fabrication imparfait, problèmes de précision issus de choix d'outil de mesure de la puissance. Chaque facteur a été profondément étudié et évalué dans notre recherche. Outre les évaluations expérimentales, nous avons également proposé une méthode pourt estimer la consommation énergétique globale d'un cluster physique, ce cluster est composé par 48 serveurs identiques et équipé d'un système de refroidissement à expansion à direct, classiquement utilisé de nos jours pour les data centers modernes. Généralement, les méthodes traditionnelles consistent à entraîner un modèle en utilisant des outils de type « intelligence artificielle » avec des données représentatives (big data). Le modèle obtenu est vu comme une "boîte noire". Si les conditions initiales changent, le modèle doit être redéveloppé pour le nouvel environnement. Nous proposons de réaliser un modèle d'une manière différente. Le modèle a été construit à partir d'une analyse thermique de

l'ensemble du cluster. Il prend en compte les conditions environnementales et les configurations du système en tant que variables. Il permet de prévoir la consommation énergétique globale du cluster en changeant les conditions environnemental et opérationnelles.

BIBLIOGRAPHY

[AA66] Vedat S Arpaci and Vedat S Arpaci. *Conduction heat transfer*, volume 237. Addison-Wesley Reading, MA, 1966.

[AAF12] Victor Avelar, Dan Azevedo, and Alan French. Pue: A comprehensive examination of the metric. Technical report, The Green Grid, 2012.

[ABC17] Maria Avgerinou, Paolo Bertoldi, and Luca Castellazzi. Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency. *Energies*, 10(10):1470, 2017.

[Aga18] Animesh Agarwal. Polynomial regression, 2018. retrieved: May 8, 2018.

[Alc] Alciom. Powerspy2: An advanced power analyzer. retrieved: July 08, 2019.

[AMK16] Bilge Acun, Phil Miller, and Laxmikant V. Kale. Variation among processors under turbo boost in hpc systems. In *Proceedings of the 2016 International Conference on Supercomputing*, ICS '16, pages 6:1–6:12, New York, NY, USA, 2016. ACM.

[And13] Cunningham Andrew. The technical details behind intel's 7 watt ivy bridge cpus, 2013.

[And17] Anders Andrae. Total consumer power consumption forecast. In *Nordic Digital Business Summit*, pages 1–6, October 2017.

[AQ14] ARM and QUALCOMM. White paper: Enabling the next mobile computing revolution with highly integrated armv8-a based socs. Technical report, 2014.

[AQ18] Susanne Albers and Jens Quedenfeld. Optimal algorithms for right-sizing data centers - extended version. *CoRR*, abs/1807.05112, 2018.

[AR08] Shanmuga Sundaram Anandan and Velraj Ramalingam. Thermal management of electronics: A review of literature. *Thermal science*, 12(2):5–26, 2008.

[Avi19] Geographic load balancing definition, 2019.

[BA11] Hüsamettin Bulut and Mehmet Azmi Aktacir. Determination of free cooling potential: A case study for İstanbul, turkey. *Applied Energy*, 88(3):680 – 689, 2011.

[BC11] Shekhar Borkar and Andrew A. Chien. The future of microprocessors. *Communications of the ACM*, 54(5):67–77, May 2011.

[BCAC+13] Daniel Balouek, Alexandra Carpen Amarie, Ghislain Charrier, Frédéric Desprez, Emmanuel Jeannot, Emmanuel Jeanvoine, Adrien Lèbre, David Margery, Nicolas Niclausse, Lucas Nussbaum, Olivier Richard, Christian Pérez, Flavien Quesnel, Cyril Rohr, and Luc Sarzyniec. Adding virtualization capabilities to the Grid'5000 testbed. In Ivan I. Ivanov, Marten van Sinderen, Frank Leymann, and Tony Shan, editors, *Cloud Computing and Services Science*, volume 367 of *Communications in Computer and Information Science*, pages 3–20. Springer International Publishing, 2013.

[BCH09] Luiz André Barroso, Jimmy Clidaras, and Urs Hölzle. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second Edition*, volume 4. Morgan & Claypool Publishers, 2009.

[BdM12] Robert Basmadjian and Hermann de Meer. Evaluating and modeling power consumption of multi-core processors. In *Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet*, e-Energy '12, pages 1–10, New York, NY, USA, 2012. ACM.

[Bea13] Donald L Beaty. Internal it load profile variability. *ASHRAE Journal*, 55(2):72–75, 2013.

[BEK+02] Pat Bohrer, Elmootazbellah N. Elnozahy, Tom Keller, Michael Kistler, Charles Lefurgy, Chandler McDowell, and Ram Rajamony. *The Case for Power Management in Web Servers*, pages 261–289. Springer US, Boston, MA, 2002.

[BFG+17] Nicolas Beldiceanu, Bárbara Dumas Feris, Philippe Gravey, Sabbir Hasan, Claude Jard, Thomas Ledoux, Yunbo Li, Didier Lime, Gilles Madi-Wamba, Jean-Marc Menaud, Pascal Morel, Michel Morvan, Marie-Laure Moulinard, Anne-Cécile Orgerie, Jean-Louis Pazat, Olivier Roux, and Ammar Sharaiha. Towards energy-proportional clouds partially powered by renewable energy. *Computing*, 99(1):3–22, Jan 2017.

[BH07] Luiz André Barroso and Urs Hölzle. The case for energy-proportional computing. *IEEE Computer*, 40, 2007.

[BHR89] T. L. Borden, J. P. Hennessy, and J. W. Rymarczyk. Multiple operating systems on one processor complex. *IBM Systems Journal*, 28(1):104–123, 1989.

[BJ08] W. Lloyd Bircher and Lizy K. John. Analysis of dynamic power management on multi-core processors. In *Proceedings of the 22Nd Annual International Conference on Supercomputing*, ICS '08, pages 327–338, New York, NY, USA, 2008. ACM.

[BJ12] W. L. Bircher and L. K. John. Complete system power estimation using processor performance events. *IEEE Transactions on Computers*, 61(4):563–577, April 2012.

[BK03] Adrian Bejan and Allan D Kraus. *Heat transfer handbook*, volume 1. John Wiley & Sons, 2003.

[BKST13] Gemma A Brady, Nikil Kapur, Jonathan L Summers, and Harvey M Thompson. A case study and critical assessment in calculating power usage effectiveness for a data centre. *Energy Conversion and Management*, 76:155–161, 2013.

[BLFP10] Daniel Bedard, Min Yeol Lim, Robert Fowler, and Allan Porterfield. Powermon: Fine-grained and integrated power monitoring for commodity computer systems. In *Proceedings of the IEEE SoutheastCon 2010 (SoutheastCon)*, pages 479–484, March 2010.

[BLOR17] Anne Benoit, Laurent Lefèvre, Anne-Cécile Orgerie, and Issam Raïs. Shutdown policies with power capping for large scale computing systems. In Francisco F. Rivera, Tomás F. Pena, and José C. Cabaleiro, editors, *Euro-Par 2017: Parallel Processing*, pages 134–146, Cham, 2017. Springer International Publishing.

[BMGA12] Bharathan Balaji, John McCullough, Rajesh K. Gupta, and Yuvraj Agarwal. Accurate characterization of the variability in power consumption in modern mobile processors. In *Proceedings of the 2012 USENIX Conference on Power-Aware Computing and Systems*, HotPower'12, pages 8–8, Berkeley, CA, USA, 2012. USENIX Association.

[Boh07] Mark Bohr. A 30 year retrospective on dennard's mosfet scaling paper. *IEEE Solid-State Circuits Society Newsletter*, 12(1):11–13, 2007.

[BS00] J. A. Butts and G. S. Sohi. A static power model for architects. In *Proceedings 33rd Annual IEEE/ACM International Symposium on Microarchitecture. MICRO-33 2000*, pages 191–201, Dec 2000.

[C+] Gary Cook et al. Clicking clean: Who is winning the race to build a green internet. retrieved: December 16, 2019.

[CBB17] Daniele Cesarini, Andrea Bartolini, and Luca Benini. Benefits in relaxing the power capping constraint. In *Proceedings of the 1st Workshop on AutotuniNg and aDaptivity AppRoaches for Energy Efficient HPC Systems*, ANDARE '17, pages 3:1–3:6, New York, NY, USA, 2017. ACM.

[CDCP15] Leandro Fontoura Cupertino, Georges Da Costa, and Jean-Marc Pierson. Towards a generic power estimator. *Computer Science-Research and Development*, 30(2):145–153, 2015.

[CFY04] Yan Chen, Toni Farley, and Nong Ye. Qos requirements of network applications on the internet. *Information Knowledge Systems Management*, 4(1):55–76, 2004.

[Che] Erik Cheever. Elements of thermal systems. retrieved: Oct 23, 2019.

[CHP+11] Henry C Coles, Taewon Han, Phillip N Price, Ashok J Gadgil, and William F Tschudi. Air corrosivity in us outdoor-air-cooled data centers is similar to that in conventional data centers. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2011.

[CIGH17] Y. Cui, C. Ingalz, T. Gao, and A. Heydari. Total cost of ownership model for data center technology evaluation. In *2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 936–942, May 2017.

[CK16] Jinkyun Cho and Yundeok Kim. Improving energy efficiency of dedicated cooling system and its contribution towards meeting an energy-optimized data center. *Applied Energy*, 165:967 – 982, 2016.

[CM05] Gilberto Contreras and Margaret Martonosi. Power prediction for intel xscale processors using performance monitoring unit events. In *Proceedings of the 2005 International Symposium on Low Power Electronics and Design*, ISLPED '05, pages 221–226, New York, NY, USA, 2005. ACM.

[Com] European Commission. Energy efficiency. [Online; accessed 2018-04-01].

[Com11] ASHRAE Technical Committee. 2011 thermal guidelines for data processing environments – expanded data center classes and usage guidance. Technical report, ASHRAE, 2011.

[Com13] SPEC Power Committee. Server efficiency rating tool public design document (latest version), 2013. retrieved: May 2, 2016.

[Com14] SPECpower Committee. (spec) power and performance benchmark methodology v2.2. Technical report, Standard Performance Evaluation Corporation, 2014.

[Com16] ASHRAE Technical Committee. Data center power equipment thermal guidelines and best practices whitepaper. Technical report, ASHRAE, 2016.

[Com18] Wikimedia Commons. File:transistor count and moore's law - 2011.svg — wikimedia commons, the free media repository, 2018. [Online; accessed 22-March-2019].

[COM19] EUROPEAN COMMISSION. Commission regulation laying down ecodesign requirements for servers and data storage products pursuant to directive 2009/125/ec of the european parliament and of the council and amending commission regulation (eu) 617/2013, Mars 2019. [Online; accessed 02-December-2019].

[con] Criu.

[Cor] Standard Performance Evaluation Corporation. Spec ptdaemon tool: List of accepted power analyzers. retrieved: November 08, 2017.

[Cor13] Intel Corporation. Intel turbo boost technology 2.0, 2013. lastaccessed: December 13, 2019.

[CQP14] Henry C Coles, Yong Qin, and Phillip N Price. Comparing server energy use and efficiency using small sample sizes. Technical report, Lawrence Berkeley National Laboratory, 2014.

[CRGRS18] Stephane Caux, Paul Renaud-Goud, Gustavo Rostirolla, and Patricia Stolf. It optimization for datacenters under renewable power constraint. In Marco Aldinucci, Luca Padovani, and Massimo Torquati, editors, *Euro-Par 2018: Parallel Processing*, pages 339–351, Cham, 2018. Springer International Publishing.

[CSP05] Kihwan Choi, Ramakrishna Soma, and Massoud Pedram. Fine-grained dynamic voltage and frequency scaling for precise energy and performance tradeoff based on the ratio of off-chip access to on-chip computation times. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 24(1):18–28, Jan 2005.

[CYP14] Jinkyun Cho, Joonyoung Yang, and Woopyoung Park. Evaluation of air distribution system's airflow performance for cooling energy savings in high-density data centers. *Energy and buildings*, 68:270–279, 2014.

[Dat15] Itai Dattner. A model-based initial guess for estimating parameters in systems of ordinary differential equations. *Biometrics*, 71(4):1176–1184, 2015.

[Dav] Ziff Davis. Definition of power distribution unit. retrieved: July 10, 2019.

[Dav04] Morris G Davies. *Building heat transfer*. John Wiley & Sons, 2004.

[DBMS0s] Jack Dongarra, Jim Bunch, Cleve Moler, and Gilbert Stewart. Linpack, 1970s.

[DCH10] Georges Da Costa and Helmut Hlavacs. Methodology of measurement for energy consumption of applications. Technical report, 2010.

[DDG+13] Mohammed El Mehdi Diouri, Manuel F. Dolz, Olivier Glück, Laurent Lefèvre, Pedro Alonso, Sandra Catalán, Rafael Mayo, and Enrique S. Quintana-Ortí. Solving some mysteries in power monitoring of servers: Take care of your wattmeters! In Jean-Marc Pierson, Georges Da Costa, and Lars Dittmann, editors, *Energy Efficiency in Large Scale Distributed Systems*, pages 3–18, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[Del18] Dell. What is the c-state, May 2018. [Online; accessed June 25, 2019].

[Dem15] Dustin Demetriou. Effectively applying the expanded ashrae guidelines in your data center. Technical report, IBM Systems, 2015.

[Des18] Jeff Desjardins. What happens in an internet minute in 2018?, May 2018.

[DGLM13] M. El Mehdi Diouri, O. Glück, L. Lefévre, and J. Mignot. Your cluster is not power homogeneous: Take care when designing green schedulers! In *2013 International Green Computing Conference Proceedings*, pages 1–10, June 2013.

[DGR+74] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268, Oct 1974.

[DLLW12] Jack Dongarra, Hatem Ltaief, Piotr Luszczek, and Vincent M. Weaver. Energy footprint of advanced dense numerical linear algebra using tile algorithms on multicore architectures. In *Proceedings of the 2012 Second International Conference on Cloud and Green Computing*, CGC '12, pages 274–281, Washington, DC, USA, 2012. IEEE Computer Society.

[DMT] DMTF. Redfish api. retrieved: Jan 22, 2019.

[DR10] Kevin Dunlap and Neil Rasmussen. Choosing between room, row, and rack-based cooling for data centers. Technical report, Schneider Electric's Data Center Science Center, 2010. [White Paper 130].

[DS11] Swagatam Das and Ponnuthurai Nagaratnam Suganthan. Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation*, 15(1):4–31, Feb 2011.

[DS12] Waltenegus Dargie and Alexander Schill. Analysis of the power and hardware resource consumption of servers under different load balancing policies. In *2012 IEEE Fifth International Conference on Cloud Computing*, pages 772–778, June 2012.

[Dug18] John Sydney Dugdale. *Entropy and its physical meaning*. Taylor & Francis, 2018.

[DW17] Hafiz M. Daraghmeh and Chi-Chuan Wang. A review of current status of free cooling in datacenters. *Applied Thermal Engineering*, 114:1224 – 1239, 2017.

[EJF14] Khosrow Ebrahimi, Gerard F Jones, and Amy S Fleischer. A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. *Renewable and Sustainable Energy Reviews*, 31:622–638, 2014.

[ele19] Schneider electric. Inrow direct expansion - acrd602, 2019. accessed: December 15,2019.

[(EP19] UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (EPA). Energy star computer server sert v2.0.2 clarification, August 2019. [Online; accessed 02-December-2019].

[ERKR06] Dimitris Economou, Suzanne Rivoire, Christos Kozyrakis, and Partha Ranganathan. Full-system power analysis and modeling for server environments. In Workshop on Modeling Benchmarking and Simulation (MOBS), 2006.

[Ete07] Akın Burak Etemoglu. A brief survey and economical analysis of air cooling for electronic equipments. *International Communications in Heat and Mass Transfer*, 34(1):103 – 113, 2007.

[Eti18] Daniel Etiemble. 45-year CPU evolution: one law and two equations. In *Second Workshop on Pioneering Processor Paradigms*, Vienne, Austria, February 2018.

[FB18] IEA Fatih Birol. Renewables 2018, market analysis and forecast from 2018 to 2023, October 2018.

[FC07] Wu-chun Feng and Kirk Cameron. The green500 list: Encouraging sustainable supercomputing. *Computer*, 40(12):50–55, 2007.

[FGXR18] Zhang Fei, Liu Guangming, Fu Xiaoming, and Yahyapour Ramin. A survey on virtual machine migration: Challenges, techniques, and open issues. *IEEE Communications Surveys Tutorials*, 20(2):1206–1243, Secondquarter 2018.

[FNCR11] Tiago C. Ferreto, Marco A.S. Netto, Rodrigo N. Calheiros, and César A.F. De Rose. Server consolidation with migration control for virtualized data centers. *Future Generation Computer Systems*, 27(8):1027 – 1034, 2011.

[FVLA02] Gilles Fraisse, Christelle Viardot, Olivier Lafabrie, and Gilbert Achard. Development of a simplified and accurate building model based on electrical analogy. *Energy and buildings*, 34(10):1017–1031, 2002.

[FWB07] Xiaobo Fan, Wolf-Dietrich Weber, and Luiz Andre Barroso. Power provisioning for a warehouse-sized computer. In *Proceedings of the 34th Annual International Symposium on Computer Architecture*, ISCA '07, pages 13–23, New York, NY, USA, 2007. ACM.

[GAAS+16] Peter Garraghan, Yaser Al-Anii, Jon Summers, Harvey Thompson, Nik Kapur, and Karim Djemame. A unified model for holistic power usage in cloud datacenter servers. In *Utility and Cloud Computing (UCC), 2016 IEEE/ACM 9th International Conference on*, pages 11–19. IEEE, 2016.

[GC15] Douglas D. Gransberg and Edward Patrick O' Connor. Major equipment lifecycle cost analysis. Technical report, April 2015.

[GE19] OZAN GÖZCÜ and Hamza Salih Erden. Energy and economic assessment of major free cooling retrofits for data centers in turkey. *Turkish Journal of Electrical Engineering & Computer Sciences*, 27(3):2097–2212, 2019.

[GFS+10] Rong Ge, Xizhou Feng, Shuaiwen Song, Hung-Ching Chang, Dong Li, and Kirk W. Cameron. Powerpack: Energy profiling and analysis of high-performance systems and applications. *IEEE Transactions on Parallel and Distributed Systems*, 21(5):658–671, May 2010.

[GKL+13] Íñigo Goiri, William Katsak, Kien Le, Thu D. Nguyen, and Ricardo Bianchini. Parasol and greenswitch: Managing datacenters powered by renewable energy. *SIGARCH Comput. Archit. News*, 41(1):51–64, March 2013.

[GM16] B. Goel and S. A. McKee. A methodology for modeling dynamic and static power consumption for multicore processors. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 273–282, May 2016.

[GMG+10] Bhavishya Goel, Sally A. McKee, Roberto Gioiosa, Karan Singh, Major Bhadauria, and Marco Cesati. Portable, scalable, per-core power estimation for intelligent resource management. In *International Conference on Green Computing*, pages 135–146, Aug 2010.

[Goo13] Google. Google's green ppas: What, how, and why, 2013.

[Goo16] Google. Achieving our 100% renewable energy purchasing goal and going beyond. White paper, Google, 2016.

[Goo19] Google. Efficiency: How we do it, 2019.

[GP13] Hadi Goudarzi and Massoud Pedram. Geographical load balancing for online service applications in distributed datacenters. In *2013 IEEE Sixth International Conference on Cloud Computing*, pages 351–358, June 2013.

[Gre18] Andy Greenberg. A critical intel flaw breaks basic security for most computers, March 2018. [Online; accessed 02-December-2019].

[Gre19] Andy Greenberg. Meltdown redux: Intel flaw lets hackers siphon secrets from millions of pcs, May 2019. [Online; accessed 02-December-2019].

[Gro16] Orange Group. Marrakech cop22: reviewing our commitments, November 2016.

[Gro19] Orange Group. We help drive the environmental and energy transition, August 2019.

[Har19] Scharon Harding. What is a cpu's ihs? a basic definition, March 2019. lastaccessed: December 13, 2019.

[HCD+17a] Franz Christian Heinrich, Tom Cornebize, Augustin Degomme, Arnaud Legrand, Alexandra Carpen-Amarie, Sascha Hunold, Anne-Cécile Orgerie, and Martin Quinson. Predicting the energy-consumption of mpi applications at scale using only a single node. In *2017 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 92–102. IEEE, Sep. 2017.

[HCD+17b] Franz Christian Heinrich, Tom Cornebize, Augustin Degomme, Arnaud Legrand, Alexandra Carpen-Amarie, Sascha Hunold, Anne-Cécile Orgerie,

and Martin Quinson. Predicting the energy-consumption of mpi applications at scale using only a single node. In *2017 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 92–102, Sep. 2017.

[HHSS05] Wei Huang, Eric Humenay, Kevin Skadron, and Mircea R. Stan. The need for a full-chip and package thermal model for thermally optimized ic designs. In *Proceedings of the 2005 International Symposium on Low Power Electronics and Design*, ISLPED '05, pages 245–250, New York, NY, USA, 2005. ACM.

[HIG94] M. Horowitz, T. Indermaur, and R. Gonzalez. Low-power digital design. In *Proceedings of 1994 IEEE Symposium on Low Power Electronics*, pages 8–11, Oct 1994.

[HIS+13] Daniel Hackenberg, Thomas Ilsche, Robert Schöne, Daniel Molka, Maik Schmidt, and Wolfgang E. Nagel. Power measurement techniques on standard compute nodes: A quantitative comparison. In *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 194–204, April 2013.

[HIVK11] Hendrik F Hamann, Madhusudan K Iyengar, and Theodore G Van Kessel. Cooling infrastructure leveraging a combination of free and solar cooling, September 20 2011. US Patent 8,020,390.

[HKLP17] Md Sabbir Hasan, Yousri Kouki, Thomas Ledoux, and Jean-Louis Pazat. Exploiting renewable sources: When green sla becomes a possible reality in cloud computing. *IEEE Transactions on Cloud Computing*, 5(2):249–262, April 2017.

[HLM+09] Fabien Hermenier, Xavier Lorca, Jean-Marc Menaud, Gilles Muller, and Julia Lawall. Entropy: a consolidation manager for clusters. In *Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments*, pages 41–50. ACM, 2009.

[Hog04] Emma Jane Hogbin. *ACPI: Advanced Configuration and Power Interface*. July 2004.

[Hol16] ARM Holdings. big.little technology, July 2016. retrieved: December 6, 2016.

[HP11] John L Hennessy and David A Patterson. *Computer architecture: a quantitative approach*. Elsevier, 2011.

[HPB12] Tom Harvey, Michael Patterson, and John Bean. Updated air-side free cooling maps: The impact of ashrae 2011 allowable ranges. Technical report, The Green Grid, 2012.

[HSP+15] Anna M Haywood, Jon Sherbeck, Patrick Phelan, Georgios Varsamopoulos, and Sandeep KS Gupta. The relationship among cpu utilization, temperature, and thermal power for waste heat utilization. *Energy Conversion and Management*, 95:297–303, 2015.

[Hun07] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

[IA09] Ali Ipakchi and Farrokh Albuyeh. Grid of the future. *IEEE Power and Energy Magazine*, 7(2):52–62, March 2009.

[IBM10] IBM. Direct use of waste heat to minimize carbon-dioxide emissions, 2010. [Online; accessed 05-December-2019].

[ILA18] Kıvanc ILAL. Advantages of inrow cooling systems against perimeter cooling systems. Technical report, Canovate Group, 2018.

[Inc] Raritan Inc. Pdu metering at the inlet, outlet, and branch circuits. retrieved: July 10, 2019.

[Int] Intel. Intelligent platform management interface: Ipmi adopters list. retrieved: Jan 22, 2019.

[int15a] What exactly is a p-state?, January 2015.

[Int15b] Intel. *Intel Intelligent Power Node Manager 3.0: Specification*, March 2015. Version 3.0.

[IT07] Info-Tech. Top 10 energy-saving tips for a greener data center. Technical report, Info-Tech Research Group, 2007.

[Joh] McCalpin John. Stream: Sustainable memory bandwidth in high performance computers. [Online; accessed 2018-04-01].

[Jon18] Nicola Jones. How to stop data centres from gobbling up the world's electricity. *Nature*, 561(7722):163–166, 2018.

[JR05] James E. SmithSmith and Ravi Nair. The architecture of virtual machines. *Computer*, 38(5):32–38, May 2005.

[KAB+03] N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan. Leakage current: Moore's law meets static power. *Computer*, 36(12):68–75, Dec 2003.

[Kam19] George Kamiya. Data centres and data transmission networks: Tracking clean energy progress. Technical report, IEA, Northwestern University, May 2019.

[KC09] E. Kursun and C. Cher. Temperature variation characterization and thermal management of multicore architectures. *IEEE Micro*, 29(1):116–126, Jan 2009.

[KDLA07] Hyekseong Kweon, Younggu Do, Jaejeong Lee, and Byoungchul Ahn. An efficient power-aware scheduling algorithm in real time system. In *2007 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 350–353, Aug 2007.

[KGB13] Atefeh Khosravi, Saurabh Kumar Garg, and Rajkumar Buyya. Energy and carbon-efficient placement of virtual machines in distributed cloud data centers. In Felix Wolf, Bernd Mohr, and Dieter an Mey, editors, *Euro-Par 2013 Parallel Processing*, pages 317–328, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[KHF+19] Paul Kocher, Jann Horn, Anders Fogh, , Daniel Genkin, Daniel Gruss, Werner Haas, Mike Hamburg, Moritz Lipp, Stefan Mangard, Thomas Prescher, Michael Schwarz, and Yuval Yarom. Spectre attacks: Exploiting speculative execution. In *40th IEEE Symposium on Security and Privacy (S&P'19)*, 2019.

[Koo11] Jonathan Koomey. Growth in data center electricity use 2005 to 2010. *A report by Analytical Press, completed at the request of The New York Times*, 9, 2011.

[KPG+01] J Kosny, T Petrie, D Gawin, P Childs, A Desjarlais, and J Christian. Thermal mass-energy savings potential in residential buildings. *Report no., Buildings Technology Center, ORNL*, 2001.

[KRH97] A. Keshavarzi, K. Roy, and C. F. Hawkins. Intrinsic leakage in low power deep submicron cmos ics. In *Proceedings International Test Conference 1997*, pages 146–155, Nov 1997.

[KSSC17] Priyanka Kumari, Fatima Saleem, Alan Sill, and Yong Chen. Validation of redfish: The scalable platform management standard. In *Companion Proceedings of the10th International Conference on Utility and Cloud Computing*, UCC '17 Companion, pages 113–117, New York, NY, USA, 2017. ACM.

[Laz17] Lazard. Lazard's levelized cost of energy analysis—version 11.0, November 2017.

[LC13] Kuei-Peng Lee and Hsiang-Lun Chen. Analysis of energy saving potential of air-side free cooling for data centers in worldwide climate zones. *Energy and Buildings*, 64:103 – 112, 2013.

[Len19] Lenovo. The server management engine for future-defined data centers, 2019. retrieved: May 8, 2019.

[LGT08] Adam Lewis, Soumik Ghosh, and N.-F. Tzeng. Run-time energy consumption estimation based on workload in server systems. In *Proceedings of the 2008 Conference on Power Aware Computing and Systems*, HotPower'08, page 4, Berkeley, CA, USA, 2008. USENIX Association.

[Lin16] Isaac Lino. Python utility for logging data from a watts up pro power meter, 2016.

[LKM16] E. A. León, I. Karlin, and A. T. Moody. System noise revisited: Enabling application scalability and reproducibility with smt. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 596–607, May 2016.

[LLW+11a] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven H. Low, and Lachlan L.H. Andrew. Geographical load balancing with renewables. *SIGMETRICS Perform. Eval. Rev.*, 39(3):62–66, December 2011.

[LLW+11b] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven H. Low, and Lachlan L.H. Andrew. Greening geographical load balancing. In *Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '11, pages 233–244, New York, NY, USA, 2011. ACM.

[LLWA12] M. Lin, Z. Liu, A. Wierman, and L. L. H. Andrew. Online algorithms for geographical load balancing. In *2012 International Green Computing Conference (IGCC)*, pages 1–10, June 2012.

[LNL15] Paul Lin, John Niemann, and Leo Long. Choosing between direct and indirect air economization for data centers. *White Paper*, 215, 2015.

[LO10] Laurent Lefèvre and Anne-Cécile Orgerie. Designing and evaluating an energy efficient cloud. *The Journal of Supercomputing*, 51(3):352–373, Mar 2010.

[LPD13] James H. Laros, Phil Pokorny, and David DeBonis. Powerinsight - a commodity power measurement capability. In *2013 International Green Computing Conference Proceedings*, pages 1–6, June 2013.

[LSG+18] Moritz Lipp, Michael Schwarz, Daniel Gruss, Thomas Prescher, Werner Haas, Anders Fogh, Jann Horn, Stefan Mangard, Paul Kocher, Daniel Genkin, Yuval Yarom, and Mike Hamburg. Meltdown: Reading kernel memory from user space. In *27th USENIX Security Symposium (USENIX Security 18)*, 2018.

[LSH10] Etienne Le Sueur and Gernot Heiser. Dynamic voltage and frequency scaling: The laws of diminishing returns. In *Proceedings of the 2010 international conference on Power aware computing and systems*, pages 1–8, 2010.

[LT11] Klaus-Dieter Lange and Michael G. Tricker. The design and development of the server efficiency rating tool (sert). In *Proceedings of the 2Nd ACM/SPEC International Conference on Performance Engineering*, ICPE '11, pages 145–150, New York, NY, USA, 2011. ACM.

[Lui10] Yury Y Lui. Waterside and airside economizers design considerations for data center facilities. *ASHRAE Transactions*, 116(1), 2010.

[LWAT11] Minghong Lin, Adam Wierman, Lachlan L. H. Andrew, and Eno Thereska. Dynamic right-sizing for power-proportional data centers. In *2011 Proceedings IEEE INFOCOM*, pages 1098–1106, April 2011.

[LWW08] Charles Lefurgy, Xiaorui Wang, and Malcolm Ware. Power capping: A prelude to power shifting. *Cluster Computing*, 11(2):183–195, June 2008.

[LZY18] Hongjie Lu, Zhongbin Zhang, and Liu Yang. A review on airflow distribution and management in data center. *Energy and Buildings*, 179:264 – 277, 2018.

[MAC+11] John C McCullough, Yuvraj Agarwal, Jaideep Chandrashekar, Sathyanarayan Kuppuswamy, Alex C Snoeren, and Rajesh K Gupta. Evaluating the effectiveness of model-based power characterization. In *USENIX Annual Technical Conf*, volume 20, 2011.

[Mak93] Spyros Makridakis. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527 – 529, 1993.

[MB09] David Moss and John H Bean. Energy impact of increased server inlet temperature. *APC white paper*, 138, 2009.

[MBH+02] Deborah T Marr, Frank Binns, David L Hill, Glenn Hinton, David A Koufaty, J Alan Miller, and Michael Upton. Hyper-threading technology architecture and microarchitecture. *Intel Technology Journal*, 6(1-12), 2002.

[MCRS05] Justin D Moore, Jeffrey S Chase, Parthasarathy Ranganathan, and Ratnesh K Sharma. Making scheduling" cool": Temperature-aware workload placement in data centers. In *USENIX annual technical conference, General Track*, pages 61–75, 2005.

[Mem19] Uptime Institute Network Members. Annual data center survey results. Technical report, 2019.

[MGW09] David Meisner, Brian T. Gold, and Thomas F. Wenisch. Powernap: Eliminating server idle power. *SIGARCH Comput. Archit. News*, 37(1):205–216, March 2009.

[MHEZ13] Jason Mair, Zhiyi Huang, David Eyers, and Haibo Zhang. Myths in pmc-based power estimation. In *European Conference on Energy Efficiency in Large Scale Distributed Systems*, pages 35–50. Springer, 2013.

[MHEZ14] J. Mair, Z. Huang, D. Eyers, and H. Zhang. Pmc-based power modelling with workload classification on multicore systems. In *2014 43rd International Conference on Parallel Processing Workshops*, pages 129–138, Sep. 2014.

[Mic17] Bearchell Michael. How to determine the lifespan of your server, October 2017. [Online; accessed March 21, 2019].

[MKK08] Andrey Mirkin, Alexey Kuznetsov, and Kolyshkin Kir. Containers checkpointing and live migration. In *Linux Symposium*, volume 2, pages 85–90, 2008.

[Moo98] G. E. Moore. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, Jan 1998.

[MPS02] William Magro, Paul Petersen, and Sanjiv Shah. Hyper-threading technology: Impact on compute-intensive workloads. *Intel Technology Journal*, 6(1-9), 2002.

[MS+18] Newville Matthew, Till Stensitzki, et al. Lmfit: Non-linear least-squares minimization and curve-fitting for python, 2018. retrieved: March 8, 2019.

[MZB+17] Aniruddha Marathe, Yijia Zhang, Grayson Blanks, Nirmal Kumbhare, Ghaleb Abdulla, and Barry Rountree. An empirical survey of performance and energy efficiency variation on intel processors. In *Proceedings of the 5th International*

Workshop on Energy Efficient Supercomputing, E2SC'17, pages 9:1–9:8, New York, NY, USA, 2017. ACM.

[NBA11] John Niemann, Kevin Brown, and Victor Avelar. Impact of hot and cold aisle containment on data center temperature and efficiency. *Schneider Electric Data Center Science Center, White Paper*, 135:1–14, 2011.

[Nod16] Technology node, 2016.

[OAL14] Anne-Cecile Orgerie, Marcos Dias de Assuncao, and Laurent Lefevre. A survey on techniques for improving the energy efficiency of large-scale distributed systems. *ACM Comput. Surv.*, 46(4):47:1–47:31, March 2014.

[OLG10] A. Orgerie, L. Lefèvre, and J. Gelas. Demystifying energy consumption in grids and clouds. In *International Conference on Green Computing*, pages 335–342, Aug 2010.

[Org14] The Climate Change Organisation. The world's most influential companies, committed to 100% renewable power, 2014. [Online; accessed 26-March-2019].

[Pas42] Victor Paschkis. Periodic heat flow in building walls determined by electrical analogy method. *ASHVE Transactions*, 48(1):75–90, 1942.

[Pat08] Michael K Patterson. The effect of data center temperature on energy efficiency. In *Thermal and Thermomechanical Phenomena in Electronic Systems, 2008. ITHERM 2008. 11th Intersociety Conference on*, pages 1167–1174. IEEE, 2008.

[PK14] Jetsadaporn Priyadumkol and Chawalit Kittichaikarn. Application of the combined air-conditioning systems for energy conservation in data center. *Energy and Buildings*, 68:580 – 586, 2014.

[PM18] J. Pastor and J. M. Menaud. Seduce: a testbed for research on thermal and power management in datacenters. In *2018 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–6, Sep. 2018.

[Pot11] Zac Potts. Free cooling technologies in data centre applications. *SUDLOWS White Paper, Manchester*, 2011.

[Pra06] R. Prasher. Thermal interface materials: Historical perspective, status, and future directions. *Proceedings of the IEEE*, 94(8):1571–1586, Aug 2006.

[Psu09] Psutil. psutil 5.5.0 documentation, 2009.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PVM+19] Carlo Puliafito, Carlo Vallati, Enzo Mingozzi, Giovanni Merlino, Francesco Longo, and Antonio Puliafito. Container migration in the fog: A performance evaluation. *Sensors*, 19(7):1488, 2019.

[PW08] Changhai Peng and Zhishen Wu. Thermoelectricity analogy method for computing the periodic heat transfer in external building envelopes. *Applied energy*, 85(8):735–754, 2008.

[QWB+09] Asfandyar Qureshi, Rick Weber, Hari Balakrishnan, John Guttag, and Bruce Maggs. Cutting the electric bill for internet-scale systems. *SIGCOMM Comput. Commun. Rev.*, 39(4):123–134, August 2009.

[Rac17] C. Rachel. Intel now packs 100 million transistors in each square millimeter - ieee spectrum, 2017.

[REN18] REN21. Renewables 2018 global status repor, 2018.

[RLXL10] Lei Rao, Xue Liu, Le Xie, and Wenyu Liu. Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. In *2010 Proceedings IEEE INFOCOM*, pages 1–9, March 2010.

[RMPV06] D. Ashay R. Mukul P, P. Neha and W. Vijay. Material technology for environmentally micro-electronic packagingl perspective, status, and future directions. *Intel Technology Journal*, 2006.

[Rob11] Redelmeier Robert. Ubuntu manpage: cpuburn, burnbx, burnk6, burnk7, burnmmx, burnp5, burnp6 - a collection, 2011. Online; accessed 2018-05-28.

[RP18] Urs Hölzle Ruth Porat. Google environmental report 2018, 2018.

[Sad17] Richard Sadler. Video demand drives up global co2 emissions, 2017.

[Sam09] Ted Samson. Power capping yields savings and floor space, July 2009. retrieved: August 22, 2019.

[Sam12] SHREYAS Sampath. Thermal analysis of high end servers based on development of detail modeland experiments. Technical report, The university of texas at Arlington, 2012.

[Sch14] Mathijs Jeroen Scheepers. Virtualization and containerization of application infrastructure: A comparison. In *21st Twente Student Conference on IT*, volume 1, pages 1–7, 2014.

[SGT+08] S. R. Sarangi, B. Greskamp, R. Teodorescu, J. Nakano, A. Tiwari, and J. Torrellas. Varius: A model of process variation and resulting timing errors for microarchitects. *IEEE Transactions on Semiconductor Manufacturing*, 21(1):3–13, Feb 2008.

[SH18] Matthias Kimmel Amy Grace Jonas Rooze Sophie Lu Tom Harries Jenny Chase Seb Henbest, Elena Giannakopoulou. New energy outlook 2018. Report, Bloomberg New Energy Finance, 2018.

[Sha12] Niket Shah. Cfd analysis of direct evaporative cooling zone of air-side economizer for containerized data center. 2012.

[Sha18] Simon Sharwood. Openbsd disables intel's hyper-threading over cpu data leak fears, June 2018. [Online; accessed 02-December-2019].

[SLH+16] Faqiang Sun, Huawei Li, Yinhe Han, Guihai Yan, and Jun Ma. Powercap: Leverage performance-equivalent resource configurations for power capping. In *2016 Seventh International Green and Sustainable Computing Conference (IGSC)*, pages 1–8. IEEE, 2016.

[Smi18] Kirstine Smith. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12(1/2):1–85, 1918.

[SN90] T. Sakurai and A. R. Newton. Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas. *IEEE Journal of Solid-State Circuits*, 25(2):584–594, April 1990.

[SP97] Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, Dec 1997.

[(SP19] Standard Performance Evaluation Corporation (SPEC). Spec ptdaemon tool accepted devices list, August 2019. [Online; accessed 02-December-2019].

[SSS+16] Arman Shehabi, Sarah Smith, Dale Sartor, Richard Brown, Magnus Herrlin, Jonathan Koomey, Eric Masanet, Nathaniel Horner, Inês Azevedo, and

William Lintner. United states data center energy usage report. Technical Report LBNL-1005775, Lawrence Berkeley National Laboratory, Berkeley, June 2016.

[STD94] Ching-Long Su, Chi-Ying Tsui, and A. M. Despain. Low power architecture design and compilation techniques for high-performance processors. In *Proceedings of COMPCON '94*, pages 489–498, Feb 1994.

[STI74] Gergonne's 1815 paper on the design and analysis of polynomial regression experiments. *Historia Mathematica*, 1(4):431 – 439, 1974.

[STIH11] Robert Schöne, Ronny Tschüter, Thomas Ilsche, and Daniel Hackenberg. The vampirtrace plugin counter interface: Introduction and examples. In *Proceedings of the 2010 Conference on Parallel Processing*, Euro-Par 2010, pages 501–511, Berlin, Heidelberg, 2011. Springer-Verlag.

[Sys12] SysTutorials. ipmi-oem (8) - linux man pages, 2012. retrieved: December 14, 2019.

[Tau02] Yuan Taur. Cmos design near the limit of scaling. *IBM Journal of Research and Development*, 46(2.3):213–222, 2002.

[TAZ+17] Ozan Tuncer, Emre Ates, Yijia Zhang, Ata Turk, Jim Brandt, Vitus J. Leung, Manuel Egele, and Ayse K. Coskun. Diagnosing performance variations in hpc applications using machine learning. In Julian M. Kunkel, Rio Yokota, Pavan Balaji, and David Keyes, editors, *High Performance Computing*, pages 355–373, Cham, 2017. Springer International Publishing.

[Tea] FreeIPMI Core Team. Freeipmi - home. retrieved: Jan 22, 2019.

[Tec] Server Technology. Sentry power manager. retrieved: July 10, 2019.

[Tec15] Vernier Software Technology. Watts up pro user manuel, 2015.

[Top19] Top500.org. Green500 - the latest list - november 2019, November 2019. [Online; accessed 02-December-2019].

[TQdAB17] Adel Nadjaran Toosi, Chenhao Qu, Marcos Dias de Assunção, and Rajkumar Buyya. Renewable-aware geographical load balancing of web applications for sustainable data centers. *Journal of Network and Computer Applications*, 83:155 – 168, 2017.

[TWR02] W. H. Tang, Q. H. Wu, and Z. J. Richardson. Equivalent heat circuit based power transformer thermal model. *IEE Proceedings - Electric Power Applications*, 149(2):87–92, March 2002.

[v3.] SimGrid v3.20. Energy plugin. retrieved: Sep 9, 2019.

[vCV11] S. van der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.

[VDBDCJ+14] Micha Vor Dem Berge, Georges Da Costa, Mateusz Jarus, Ariel Oleksiak, Wojciech Piatek, and Eugen Volk. Modeling data center building blocks for energy-efficiency and thermal simulations. In Heidelberg Springer, Berlin, editor, *Energy-Efficient Data Centers*, volume 8343, pages 66–82. Springer, 2014.

[VDCL+15] Violaine Villebonnet, Georges Da Costa, Laurent Lefevre, Jean-Marc Pierson, and Patricia Stolf. "big, medium, little": Reaching energy proportionality with heterogeneous computing scheduler. *Parallel Processing Letters*, 25(03):1541006, 2015.

[vKBB+16a] Jóakim von Kistowski, Hansfried Block, John Beckett, Cloyce Spradling, Klaus-Dieter Lange, and Samuel Kounev. Variations in cpu power consumption. In *Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering*, pages 147–158. ACM, 2016.

[vKBB+16b] Jóakim von Kistowski, Hansfried Block, John Beckett, Cloyce Spradling, Klaus-Dieter Lange, and Samuel Kounev. Variations in cpu power consumption. In *Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering*, ICPE '16, pages 147–158, New York, NY, USA, 2016. ACM.

[vKLA+17] Jóakim von Kistowski, Klaus-Dieter Lange, Jeremy A. Arnold, Hansfried Block, Greg Darnell, John Beckett, and Mike Tricker. The sert 2 metric and the impact of serve configuration. Technical report, September 2017.

[WA12] D. Wong and M. Annavaram. Knightshift: Scaling the energy proportionality wall through server-level heterogeneity. In *2012 45th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 119–130, Dec 2012.

[WASM14] Beth Whitehead, Deborah Andrews, Amip Shah, and Graeme Maidment. Assessing the environmental impact of data centres part 1: Background, energy use and metrics. *Building and Environment*, 82:151 – 159, 2014.

[WCLK12] Xiaorui Wang, Ming Chen, Charles Lefurgy, and Tom W. Keller. Ship: A scalable hierarchical power control architecture for large-scale data centers. *IEEE Transactions on Parallel and Distributed Systems*, 23(1):168–176, Jan 2012.

[WCS11] Shinan Wang, Hui Chen, and Weisong Shi. Span: A software power analyzer for multicore computer systems. *Sustainable Computing: Informatics and Systems*, 1(1):23–34, 2011.

[WDG+16] Qiang Wu, Qingyuan Deng, Lakshmi Ganesh, Chang-Hong Hsu, Yun Jin, Sanjeev Kumar, Bin Li, Justin Meza, and Yee Jiun Song. Dynamo: Facebook's data center-wide power management system. In *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, pages 469–480, June 2016.

[WKV11] Torell Wendy, Brown Kevin, and Avelar Victor. The unexpected impact of raising data center temperatures. Technical report, Schneider Electric, 2011.

[WNLMM18a] Yewan Wang, David Nörtershäuser, Stéphane Le Masson, and Jean-Marc Menaud. Etude de l'influence des aspects thermiques sur la consommation et l'efficacité énergétique des serveurs. In *SFT 2018 - 26ème Congrès Français de Thermique*, pages 1–8, Pau, France, May 2018.

[WNLMM18b] Yewan Wang, David Nörtershäuser, Stéphane Le Masson, and Jean-Marc Menaud. Potential effects on server power metering and modeling. *Wireless Networks*, pages 1–8, 2018.

[WOPW13] M. Witkowski, A. Oleksiak, T. Piontek, and J. Węglarz. Practical power consumption estimation for real life hpc applications. *Future Generation Computer Systems*, 29(1):208–217, 2013. Including Special section: AIRCC-NetCoM 2009 and Special section: Clouds and Service-Oriented Architectures.

[XWG17] Chenhan Xu, Kun Wang, and Mingyi Guo. Intelligent resource management in blockchain-based cloud datacenters. *IEEE Cloud Computing*, 4(6):50–59, November 2017.

[YAJ17] Yunbo, Li, Anne-Cécile, Orgerie, and Jean-Marc, Menaud. Balancing the use of batteries and opportunistic scheduling policies for maximizing renewable energy consumption in a cloud data center. In *2017 25th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*, pages 408–415, March 2017.

[YM13] Jumie Yuventi and Roshan Mehdizadeh. A critical analysis of power usage effectiveness and its use in communicating data center energy consumption. *Energy and Buildings*, 64:90–94, 2013.

[yyo15] yyongpil. Watts up pro meter logger, 2015.

[ZBS+04] Bo Zhai, David Blaauw, Dennis Sylvester, Dennis Sylvester, and Krisztian Flautner. Theoretical and practical limits of dynamic voltage scaling. In *Proceedings of the 41st Annual Design Automation Conference*, DAC '04, pages 868–873, New York, NY, USA, 2004. ACM.

[ZLQZ13] Xiao Zhang, Jianjun Lu, Xiao Qin, and Xiaonan Zhao. A high-level energy consumption model for heterogeneous data centers. *Simulation Modelling Practice and Theory*, 39:41–55, 2013. S.I.Energy efficiency in grids and clouds.

[ZMT+12] Severin Zimmermann, Ingmar Meijer, Manish K. Tiwari, Stephan Paredes, Bruno Michel, and Dimos Poulikakos. Aquasar: A hot water cooled data center with direct energy reuse. *Energy*, 43(1):237 – 245, 2012. 2nd International Meeting on Cleaner Combustion (CM0901-Detailed Chemical Models for Cleaner Combustion).

[ZSX+14] Hainan Zhang, Shuangquan Shao, Hongbo Xu, Huiming Zou, and Changqing Tian. Free cooling of data centers: A review. *Renewable and Sustainable Energy Reviews*, 35:171–182, 2014.

[ZWZ17] Yin Zhang, Zhiyuan Wei, and Mingshan Zhang. Free cooling technologies for data centers: energy saving mechanism and applications. *Energy Procedia*, 143:410–415, 2017.

[ZZS18] Ming Zhang, Zizhan Zheng, and Ness B. Shroff. An online algorithm for power-proportional data centers with switching cost. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6025–6032, Dec 2018.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Titre : Évaluation et modélisation de l'impact énergétique des centres de donnée en fonction de l'architecture matérielle/logicielle et de l'environnement associé

Mot clés : Modélisation du consommation énergétique, Efficacité énergétique, Variabilité des processeurs, Effect Thermique

Résumé : Depuis des années, la consommation énergétique du centre de donnée a pris une importance croissante suivant une explosion de demande dans cloud computing. Ce thèse aborde le défi scientifique de la modélisation énergétique d'un centre de données, en fonction des paramètres les plus importants. Disposant d'une telle modélisation, un opérateur pourrait mieux repenser/concevoir ses actuels/futurs centre de données.

Pour bien identifier les impacts énergétiques des matériels et logiciels utilisés dans les systèmes informatiques. Dans la première partie de la thèse, nous avons réaliser un grand nombre évaluations expérimentales pour identifier et caractériser les incertitudes de la consommation d'énergie induite par les éléments externes : effets thermiques, différences entre des processeurs identiques causées par un processus de fabrication imparfait, problèmes de précision issus d'outil de mesure de la puissance, etc. Nous avons terminé cette étude scientifique par le développement d'une modélisation global pour un cluster physique donné, ce cluster est composé par 48 serveurs identiques et équipé d'un systèmede refroidissement à expansion à direct, largement utilisé aujourd'hui pour les data centers modernes. La modélisation permet d'estimer la consommation énergétique globale en fonction des configurations opérationnelles et des données relatives à l'activité informatique, telles que la température ambiante, les configurations du système de refroidissement et la charge des serveurs.

Title: Evaluating and Modeling the Energy Impacts of Data centers, in terms of hardware/software architecture and associated environment

Keywords: Energy consumption modeling, Energy efficiency, Processor variability, Thermal Effect

Abstract: For years, the energy consumption of the data center has dramatically increased followed by the explosion of demand in cloud computing. This thesis addresses the scientific challenge of energy modeling of a data center, based on the most important variables. With such modeling, an data center operator will be able to better reallocate/design the current/future data centers.

In order to identify the energy impacts of hardware and software used in computer systems. In the first part of the thesis, to identify and characterize the uncertainties of energy consumption introduced by external elements: thermal effects, difference between identical processors caused by imperfect manufacturing process, precision problems resulting from power measurement tool, etc. We have completed this scientific study by developing a global power modeling for a given physical cluster, this cluster is composed by 48 identical servers and equipped with a direct expansion cooling system, conventionally used today for modern data centers. The modeling makes it possible to estimate the overall energy consumption of the cluster based on operational configurations and data relating to IT activity, such as ambient temperature, cooling system configurations and server load.