POLITECNICO DI TORINO

Department of Electronics and Telecommunication (DET)

Master Degree Thesis

Energy efficient Data center operation: Measurement and Analysis

Computer and Communication Network Engineering

Supervisors:

Prof. Marco Mellia

Lisette Cupelli

Candidate:

Peyman Iravani

June 2018

Acknowledgements

I would like to express my deepest appreciation to my supervisor at the university of RWTH, Lisette Cupelli, who has been my guide and mentor through my journey at RWTH university and my professor at Polytechnic university of Turin, Professor Marco Mellia. Without their president, help and guidance this dissertation would not have been possible. Furthermore, I would like to thanks to Prof. Dr.-Ing. Antonello Monti to give me the amazing experience to work on my thesis at Institute for Automation of Complex Power Systems at the university of RWTH. I would also like to thanks the whole TNG (Telecommunication Network Group) of Polytechnic of Turin for teaching wonderful courses during the period of my master study. Finally, I would to thanks all my family especially my mother, father and my brother as well as my friends who have been a support and source of motivation pushing me to excel in my undertaking.

June 4, 2018

Abstract

In the contemporary world, the vast majority of organizations throughout the world manipulate the information systems to their businesses. Hence, data centers which fall into the category of information systems play the supreme role in the most organizational operations for the purpose of ensuring the business continuity. Causing the interrupt in the running status of the system would create the Irreparable impact on businesses. Therefore, optimizing the surrounding that servers and storage devices are located become extremely challenging. Almost all design the data center infrastructures for the sake of handling the peak load conditions in order to make sure that the hosted environment works almost continuously but the maintenance of such systems is very costly. This fact elaborates the necessity of having the durable system that regulates the energy consumption as well as reduce the maintenance cost of such environments and simultaneously ensuring the optimal performance of the system. In the classical method, a Fluid dynamic model is implemented in order to properly model the dynamic and complex environment of a datacentre. But the traditional approach, substantially long time to yield the steady state as the result it causes the loss of productive time and resources. We proposed the ARIMA time-series forecasting in order to predict the behavior of whole data center system but meanwhile promise a faster convergence and therefore a higher performance. By having the datacenter which have changing server heats and fans speed, our system has an objective to accurately predict the power which is used by IT loads that passing through the datacentre. Furthermore, in this thesis, we take the predicted values which were generated by predictor and use them as the input to the optimizer in order to reduce the whole power consumption of the datacentre.

Keywords: ARIMA time-series forecasting- Machine Learning-Energy optimization

Contents

Indices and Abbreviations

Symbol	Description
AR	Auto regression
MA	Moving-Average
DR	Demand-response
ANN	Artificial Neural Network
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error
MSE	Mean Square Error
AE	Absolute Error
ARMA	Auto-Regressive Moving-Average
ARIMA	Auto-Regressive Integrated Moving-Average
STLF	Short-term load forecasting
DSO	Distributed service organization
BESS	Battery Energy Storage System
HVAC	Heating Ventilation Air Conditioning
DR	Demand-Response
CDF	Cumulative Distribution Function
FFNN	Feed-Forward Neural Network
CFD	Computational Fluid Dynamics
API	Application Programming Interface
MPC	Model Predictive Control
HPC	High Performance Computing
GA	Genetic Algorithm

List of figures:

List of tables:

1. Introduction

In the modern world, the number of packets that transferred over the Internet due to the increase every day usage of people through the social media, APIs which exists in the web-applications that are using for businesses, datacenters are become a crucial component in the modern IT infrastructure. Datacenters are store vast majority of ICT equipment and related accessories that are used by the modern communication systems. These days every large IT organization have been equipped with the datacenter either in their place or outsourced to vendors. These elements have cause to enormous increase in size, number as well as consumed power in the datacenters. With significantly aggrandizement energy requests for high-performance computing architecture and associated facility, data centers have become a large consumer of electricity. [1] The U.S Environmental Protection Agency (EPA) estimated that energy that consumed by this section was around 61 billion kWh(kilowatt-hours) in 2006 which is 1.5% of total U.S. electricity consumption. This institute also showed that amount of energy that consumed by the nation's server and data centers in 2006 was estimated to be more than double the electricity that consumed at this section in 2000. Furthermore, EPA estimated that energy consumed nationally by servers and data centers could nearly double again in another 5 years to be more than 100 billion KWh, representing $7.4 billion annual electricity cost. Also, in the circumstances that power distribution as well as cooling systems have reached the peak capacity, data centers continue to install high-density servers in order to deal with the ever-increasing load.Power consumed by the data center is the result of several components that already installed in the data center, CPU activity, memory, disk drives and even mix of instructions that is executed, therefore it makes datacenter research an outstanding focus for optimized operation and design of a data center given a lot of importance.

Due to the fact that data centres are crucial components of the state of art IT infrastructure, it is vitally important that power demands which come through the data centre environment as well energy efficiency must be controlled effectively. At the modern data centres, we have the immensity of control regarding cooling that exists at the room level and the power which consumed at the server level. Commonly, the average capacity regarding the cooling which exists in the data centers is 3kW per cabinet with the maximum of 10-15KW per cabinet, although the common CRAC airflow supply to a cabinet is roughly 200-500CFM [22].

In the Data center environment, Airflow is the complex process and it is hard to be mapped accurately. Therefore, the environment in the data center is highly dynamic and needs to design the complex thermal modeling. Traditionally, in order to model and simulate the Data center environment, we use the Computational Fluid Dynamics (CFD). This task not only is difficult to be implemented in terms of software license but also it also computational heavy to be effectively run the software. Usually, in order to properly implement CFD we need to have the well-trained user, Furthermore, CFD takes the huge amount of time in order to provide the steady output. The environment of the data center is highly volatile and dynamic; therefore, continuous simulation is time-consuming.

One of the most important factors in data centers is reliability, which according to [2] it must be in availability range of 99.671% to 99.995% that highly depends on the class of data center. In order to have the reliable operation of data centers, they have the possibility to use backup power generators such as Uninterruptible Power Supplies (UPS) and a diesel generator sets. These systems are able to generate power for data centers; UPS is used for short transition periods during power failure (typically up to 15 minutes), whereas the Diesel generator is cable of powering the data center for longer periods.

The aim of this thesis is to develop and design machine learning methodologies and tools to perform real-time energy efficient optimization of Datacenter (DC) operation, leveraging on aggregated power consumption data to predict the behavior of different equipment in DC subsystems.

According to the definition of Tom Mitchell "Machine learning is a computer program is said to learn from experience with respect to some class of Tasks T and performance measure P if its performance at tasks in T, as measured by Improves with experience E" [23]. Machine learning has the wide range of applications such as: Time-series forecasting, Bioinformatics, Internet fraud detection, marketing, Economics. In this thesis, we use one of the applications of Machine learning which is Time-series forecasting in order to predict the future power consumption of Datacenter subsystems. Time series forecasting is an important statistical data analysis technique used as a basis for manual and automatic planning in many application domains such as sales, traffic control, and energy management [24]. We use this application of machine learning due to the fact that power consumption in the Datacenter that we need to deal with has the very slow change in the behavior in terms of power consumption therefore with helping the Time-series forecasting we are able to predict the future power consumption of Datacenter subsystems. ARIMA (Autoregressive Integrated Moving Average) which falls under the category of time-series forecasting have been implemented in this thesis. We used Rolling horizon forecasting method that has prediction horizon of 1440 minute which equals to one day and it will go forward on the general horizon. After one prediction horizon was completed the results will enter the optimizer that is implemented by Gurobi solver in Python programming language environment. Data center components that are involved in the model are consist of HVAC (Heating Ventilation Air Conditioner) and BESS (Battery Energy Storage System). In order to the better understanding of the system, the following figure has been produced.

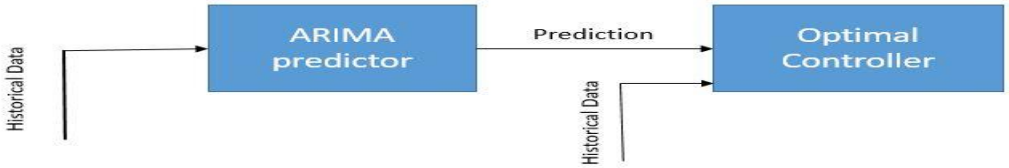


Figure 1 General presentation of work

This thesis begins with a review of the existing studies related to this topic, moving to the prediction of behavior regarding different subsystems of Datacenter. Furthermore, the focus is on implementing and evaluating possible techniques associated to reduce the power consumption in data centers as well as provide the flexibility to the data center. For evaluating the flexibilities, the data center model is simulated to incorporate various methods focusing on demand response.

2. Problem statement

In the modern world, due to the fact, people are using smartphones and smart devices such as IOT (Internet of Things) and Wireless sensor networks in their daily life, therefore, the amount of data that are producing are highly increased, so it causes to the dramatic increase of large-scale data centers. Because of the significant enhancement in hardware affordability as well as the dramatic growth of bid data, the state of the art Internet companies need to deal with wide range of personalized user experiences and minimal downtime.

Large energy consumers such as Industries, commercial complexes, and data centers are ideal to participate in demand response techniques. These consumers are ideal since they have large demands and generally have flexible loading capabilities. According to the report that recently published by Lawrence Berkeley National Institute in June 2016, data centers that located in the US are estimated to consume around 73 billion KWh in 2020.This growing power consumption also can be seen in Europe. Figure 1 shows the trend of energy costs and energy consumption from 2000 to 2015 with the prediction of 2016 and 2017 which evaluated by LRZ institute that located in Germany, as we could see the energy cost increased from around 0.5 million € to 7 million € in 2016. [14]

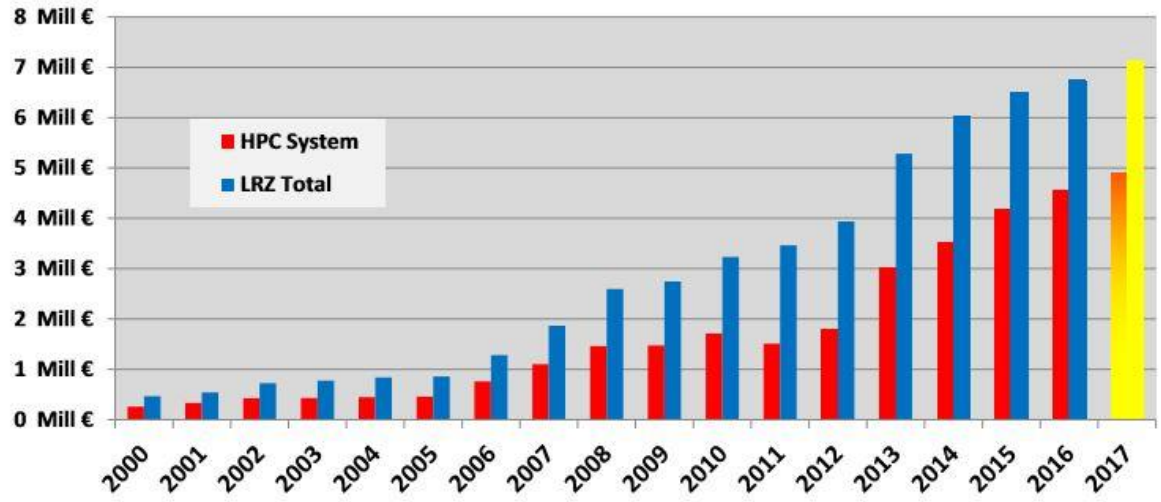


Figure 2 trend of energy cost from 2000 to 2015 with prediction of 2016 and 2017

This increasing trend happens due to the continuous increase in system size and computational performance. From figure 2 we could easily see that the costs increased from 0.5M € in 2000 to 6.8M € in 2016, as well as the rise in power consumption of deployed HPC systems. The reason is that of continued the increase in system size and computational performance.

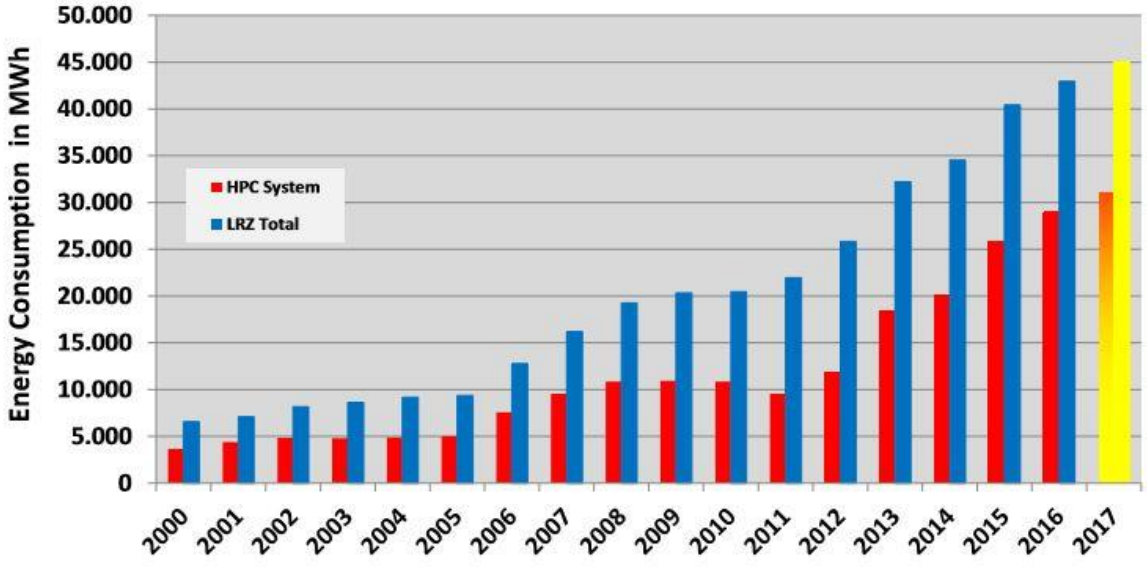


Figure 3 trend of energy consumption from 2000 to 2015 with prediction of 2016 and 2017

Not only data centers could handle flexible loads, but also they equipped with backup power systems such as UPS and diesel generators which are capable of powering the entire data center in power outage scenarios.

Depending on the class of data center, there have redundancies in place. This infrastructure can be utilized to provide ancillary services to the grid as required. Data center UPS can take part in grid stabilization whenever it is in the idle state and is required by the grid. As long as the grid is operational, the UPS batteries will always be in the idle state.

The aim of this thesis is to use time-series forecasting for modeling and predicting the DC dynamic behavior and its energy consumption. The First part of the thesis will focus on the prediction of DC dynamic subsystems. The focus of second part the thesis is on dynamic of data center environment as well as associated sub-systems.

3. Objectives and Layout of thesis

Section 4 dedicated to the literature review, the purpose of this section is to study and summarize the current development in this field.

Section 5 gives an overview of the state of Data centers, Grids and measures taken to stabilize the grid.

Section 6 gives an overview regarding chosen platform in order to predict the behavior of Datacenter subsystems. The modeling platform for predicting these subsystems is Python programming language.

Section 7 covers the implementation and modeling of the data center and its components in Python environment. The model is deconstructed to component level in the scope of this section. This section lays the groundwork for section 8, which present the tests and simulations carried out in Python environment based on the developed model to support the thesis and its findings.

This thesis concludes with the summary of the findings and listing possible future research possibilities on the topic.

4. Literature review

Due to the dramatic increase in the popularity of online services such as social networks and cloud services it requires to have high-performance computing [3]. Therefore, the datacenter power consumption has become the complex problem and various studied especially in terms of green-datacenter have been performed. As an example, Study [4] shows that by designing cooling-aware job management in the datacenter we have the possibility of reducing energy consumption. One of the possibilities for reducing the energy that consumed by the datacenter is using Thermal-aware job scheduling, the results of the study [5] which used Thermal-aware job scheduling showed that the energy savings achieved are greater than traditional job scheduling. The other possibilities for reducing the energy consumption in the datacenter is to assign all tasks to a small number of servers and remain the rest of server as shut down [6]. Another approach for improving the energy efficiency at the datacenter is configuring air conditioner settings based on temperature distribution in the datacenter. Because lowering the power consumption of some equipment in the datacenter might enhance the consumed power by the other equipment and the overall power which consumed by the data center must be examined, hence, cooperative control is crucial for enhancing the energy efficient datacenter.

Data centers are CPS (Cyber-Physical) systems, where servers offer numerous computational resources and in the meantime produce much related to their utilization, therefore the design of a proper cooling system is challenging task. Many approaches have been implemented to tackle this challenge, the study [7] implemented the heat flow model which uses temperature information which generated by onboard and ambient sensors, evaluate hot air recirculation according to this information and accelerates the thermal evaluation process for high-performance data centers. The Study [8] is introduced new cyber-physical index (CPI) a measure of the combined distribution of cyber and physical effects in the given data center, in this study the author implemented Model predictive control (MPC) method for enhancing the energy efficiency of data center. The amount of power that consumed by the server when it is fully loaded is almost twice the idle state. In sleep mode, the consumption is also reduced, and it is zero when the server is shut down. The problem of shutting down the servers is that there is the startup time that we need to take into consideration while shutting it down, it is not possible to immediately load the server when it is shut down and we need to consider the delay when we shutting the server down. On the other hand, putting the server in idle mode helps us to save energy for up to 50% compared to the fully loaded scenario.

Another important factor that plays a crucial role in the data center in order to reduce the operational cost is UPS and battery storage system (BESS). The storage system could be used for storing the energy during the low-cost period and use the stored energy during peak periods [9]. Data centers are facilitated with large battery storage systems in place, typically with additional systems for redundancies. These large systems would be sufficient in implementation ancillary services in the future grid. And along with the battery systems, the backup generators can also be utilized if required to provide power to the grid. The primary task for BESS is to provide backup power to IT load of the datacenter in case of power outage. Therefore, batteries

should always have sufficient amount of charge in order to maintain power in case of power failure. Hence, it is necessary to set boundaries on State of charge (SOC) of battery in order to have sufficient charge left at any point in time to provide the necessary backup power to the data center in case of power failure.

We have the category of datacentres which known as colocation datacentre. This type of datacentres has the ability to help to assure your day-to-day IT functions which are crucial for the running business, this category of data centers have several advantages, as an example colocation datacenters could use to store the IT equipment at the case that datacenter located in the case that datacenter positioned in very harsh weather condition, it can to have the separate location for data storage, backup and power redundancy as well as bandwidth redundancy which cause to continuously running the business in the case of major catastrophe. Further benefit of having colocation facility is associated to the server storage equipment, in the case of power outage, due to the fact that we have the backup storage the network traffic will not be affected. Any colocation facility is equipped with the backup power supply in the case that if the direct power interrupted, the power suppliers are routed to UPS which contain the large bank of powerful batteries. Many facilities have been equipped with at least two separate units of UPS and multiple generators in order to provide the power in the case of occurrence of the power outage at the facility. Colocation is very popular in the market which contains large bandwidth circuits such as Ethernet and Metro Ethernet and these systems are widely available. Another benefit of using colocation datacenter is bandwidth redundancy, depending on the exact location of colocation facilities, it allows to access to wide range of bandwidth types such as Fast Ethernet and Gigabit Ethernet from several providers. If the bandwidth at the business is disabled at the location which business located at, the traffic of business will not be interrupted. Also, colocation facilities have redundancy by themselves, due to the fact that the bandwidth is provided by multiple carriers. Colocation facilities provide the opportunity to remotely access to the servers either through the Internet connection or maintenance that could be done on-site and this access will be for the whole day/week/month/year. Additionally, we have the alternative of permitting a clerk which working inside the datacentre in order to execute the maintenance on the provisional or frequent timetable. In order to access to the facilities which, exist in colocation data centers, providers offer multiple layers of security for entering to the facilities such as 24-hour video surveillance, biometric security measures such as iris scan for avoiding access to the facilities by the unauthorized person. With the aid of collocation facilities, we could be assured that the crucial business is carry on regardless of whether interference, power outage or bandwidth disruption.

Colocation data centers have the potential to handle the workload of multiple customers as they rent out the IT equipment to these customers for their own use. Colocation data centers use the higher degree of redundancies. Colocation data centers are not flexible enough in terms of managing IT workload, Hence, they are reducing the flexibility possibilities. One of the flexibility options which mainly used in the industry could be implemented with backup and storage systems with cooling techniques such as varying temperature set point, but the HPC datacenters have the high degree of redundancies in place. In the case of HPC, it is possible to change the temperature setpoint due to the fact the performance of the server highly depends

on temperature, Therefore, in this case we use BESS and Diesel Gensets for the flexibility in the datacenter.

In data centers, the most important factor that we need to consider is QOS (Quality of Service). The data centers are bounded by Service Level Agreements (SLA). The SLAs need to be considered before going ahead with any type of flexibilities that mentioned by Geyser. SLAs define the quality of service that needs to provide and ramifications for not satisfying the conditions. As an example, in terms of IT workload, colocation data centers are bounded by SLA but they can have flexibility by using Diesel Genset and UPS that does not ignore the SLA requirements. Enterprise data centers have full control over the IT workload as well as backup systems. Therefore, they can provide the higher level of flexibilities without ignoring the quality of service. All systems that use in the data center for purpose of backup and providing redundancy they must meet the SLA requirements and maintain the desired QOS.

Recently demand response has been increasingly popular during Datacenter industry due to the fact that Datacenters are consuming huge amount of power and this power mainly consume for cooling down the components which exist in this industry and the data centers are fall into the category of energy-intense buildings, therefore datacenters are able to provide enormous for energy efficiency improvements, Although the demand response capabilities of this equipment have not been sufficiently investigated.

In the modern electrical industry, recently the new term which known as DR became popular. DR programs yield vast majority of opportunities for the customer to participate in the operation of the electrical grid and play the significant role by diminishing or shifting their electricity consumption throughout peak-periods according to the time-based rates. DR programs are used by some electrical planners and operators as resource options in order to balance demand and supply in the electrical market. DR programs have the potential to reduce the electricity cost at associated markets and as the result, they cause to reduce retail rates. They are the wide range of methods that customers able to participate in the DR programs such as: time-of-use pricing, critical peak pricing, variable peak pricing, real-time pricing and critical peak rebates. DR programs also include direct load control programs that supply the capability for power companies in order to cycle air conditioner and put water heater as the suspend or running state through off-peak-demand periods in exchange for a financial encouragement as well as reducing electrical invoices.

The electrical power industry assume demand respond programs as progressively precious resource that has potentiality and prospective impacts are cause to develop the modern grid system. As an example, sensors have the ability realize peak load problems and use automatic switching in order to either divert or reduce power in strategic places. Datacenters have the potential to reduce the power consumption by the help of demand response program, to be more precise it could happen by using load-peak shifting. As an example, authors [30] implemented the methodology for avoiding the coincident peak via workload shifting, and they evaluate the implemented algorithm by numerical simulation which was performed according to the real world traces, results of their test showed that their algorithm able to save 40% of energy cost.

Demand response programs not only have the ability to reduce the energy cost in the datacenter by the help of load-shifting methodology but also implementing demand response program on subsystems which exists in the datacenter environment (HVAC, UPS, BESS, etc.) could cause to reduce the cost of the energy that consumed in the datacenter. As an example, authors at study [31] developed the technique in order to calculate the optimal control of building's Heating-Ventilation-Air Conditioning (HVAC) system as DR tool by taking into the consideration of changing DR signal, energy which stored in the site (Known as on-site energy storage) and energy which generated in the site (Known as on-site energy generation). Furthermore, they implemented the model for the purpose of reducing the problem size and discover the optimal solution. Another subsystem of the datacenter that we could have energy cost reduction by using the Demand-response is BESS. At BESS we could use load shedding methodology in demand-response programs at Datacenter level, which means if the generated power at peak hours is higher than the pre-defined threshold it must be discarded but the problem of this methodology is QOS degradation. As an example, Authors at study [32] have been designed an online optimization framework in order to do the peak shaving, this model considers conditional value at risk and allows to navigate cost-risk-offs of datacenters according to their infrastructure which they used to handle the energy consumption and their workload characterization. They used the framework that they were designed in order to perform online peak shaving by taking into account the battery degradation costs under peak-based pricing. The study [33] that have been performed by their authors, they implemented BESS contemporaneously in order to peak shaving and frequency regulation by considering joint optimization framework that able to capture battery degradation, operational constraints, and uncertainties in customer load and regulation signals. Within the framework which implemented at mentioned research paper, they could reduce the electricity bills of customers up to 12%.

In modern world optimizing the consumed power by data centers became the very demanding topic due to the fact that the cost of energy will increase dramatically, according to the report that provided by LRZ institute [14] the energy consumption at data centers (to be more precise High-performance computers) will be around 45000 MWh also the energy cost is increasing, according to LRZ report at 2017 Datacenters need to pay 7,000,000 €. Hence, machine algorithms are using in this industry for reducing energy bills as well as minimizing the energy consumption. As an example, Google used ANN (Artificial Neural Network) in their Datacenters and by this way, they could reduce 40% of the energy required for cooling purposes in the datacenter. One possibility to reduce the power consumption in the data center is Predicting the temperature distribution through the servers that are located in Datacenter. In study [20] author apply the machine learning technique for predicting the temperature distribution in a data center and by applying this method they were able to reduce 30% of power consumption of air conditioners. The study [21] which performed by authors, they implemented neural network which integrated with Genetic Algorithm optimization, results that generated by the model represent an effective real-time design and control tool in order to have the energy efficient thermal management in the data center environment.

4.1 Assessment of flexibility techniques for data centers:

From 2004 to 2016 the global installed capacity of RES (Renewable Energy storage) has dramatically increased from 800GW to 2017GW [10] [11]. The total installed capacity was 25GW in 2004 which increased to 93GW in 2014 almost quadrupling in a decade. [17]. This rise in the RES has given the rise to additional stability issues in the grid, mainly due to the fact that generation from RES cannot be controlled and is totally dependent on the weather conditions.

Conventionally, the power plants are operated at the point lower the maximum capacity. Thus allowing for compensating for any discrepancies in the demand-supply by adjusting the operating point of the power plants. But with the increase in RES, these conventional power plants are falling short of compensating the RES generation. [18]. Therefore, additional services are called ancillary services.

Apart from the fact that Ancillary services help maintain and stabilize the grid, the service provider is also able to generate the revenue. If it is a generating unit already connected to the grid, it will be in addition to the revenue generated by selling the energy. Alternatively, Energy storage/generation devices are employed for the sole purpose of providing Ancillary services to the grid as and when required. The revenue in such case is only that of provided ancillary service. The generated revenue is dependent on the type of the Ancillary service provided.

4.1.1 Review of Flexibility techniques in the Data center environment

The various flexibility options that are introduced at Section 4 of this thesis, could be summarized in tabular format. Table 1 represents the possible flexibility options in the typical datacenter (According to GEYSER) [13].

Table 1 Flexibility options

Flexibility Component	Flexibility Action
IT Workload	Workload Consolidation
	Time Shifting the workload to achieve higher overall server utilization.

	Workload relocation to a different DC
Electrical Cooling Devices	Adjust the cooling intensity for certain time periods.
	Pre-/Post-Cooling of the Server Room depending on the Energy Price.
Thermal Storage	Dynamic usage of the thermal storages.
Electrical Storage Devices	Charge and Store energy in Batteries for later use (when available).
	Provide energy from Batteries to DC to reduce Grid consumption
	Provide Auxiliary Services to the grid when requested.
PV / Diesel Gen.	Feed energy to the grid when requested.

Although multiple solutions are available for having the flexible data center, as mentioned at above table, it is not possible to implement every option for every type of Data Center. The Five types are Public Cloud Providers, Colocation Data Centers, Enterprise Data Centers, High-Performance Computing (HPC) Data Centers and in-house (small) Data Centers.

The public cloud providers have already managed to reach PUE close to 1.0, therefore the considered the most efficient Data centers in operation. The chief culprits are the large number of Smaller Data Centers such as Colocation DC, Enterprise DC and in-house DC. These Data Centers operate at much higher PUE thus are highly inefficient in comparison to Public Cloud DC. The inefficiency of these Data Centers marks them as the ideal customers for implementing Flexibility Options in order to improve the efficiency, in turn improving the PUE, which will ultimately result in Economical benefit to the Owner/Operator of the DC.

The IT workload in a data center is varying throughout the day, as a result, servers are usually working underutilize . Depending on the workload on each server, the power consumption can greatly vary. A server at full load consumes twice as much power as the one in the idle state. But the power consumption does not change linearly with the workload, Therefore it is inefficient to use the server at lower workload. Ideally, the workload should be distributed such that all servers are running at the optimal operating point, alternatively only utilizing the neccsary servers to carry out the tasks while the rest are either at idle or off state. This falls under the First flexibility category that mentioned in Table 1. Instead of distributing the workload on the available servers, consolidating it over a few servers in order to achieve higher utilization while remaining servers at idle state, due to the fact that reducing the power consumption and achieving higher efficiency is the ideal strategy.

In circumstances that servers are underutilized, time-independent tasks could be shifted to the next time step in order to consolidate the workload at that particular time, Hence allowing the servers to operate at the higher operating point. In ideal DC in terms of energy, implementing these options will be easier but in the reality, this is not possible as there are different types of data centers. To control the workload, it is necessary that the DC operator is in control of the workload. But in the case of colocation DC, the operator has no control over the workload as the servers are rented to third parties, making it impossible to implement mentioned flexibility options. Similarly, HPC (High-Performance Computing) DC are typically always operating at high loads and the tasks are generally time dependent. But, the enterprise and in-house DC are in control of workloads and generally are not operating at high loads, therefore are ideal for implementing these flexibility options.

Another possibility is workload relocation, By relocating tasks from DC to other DC at other location. One of the DC has the possibility of working at high load while the other can remain at lower or idle state. This can be achieved in case of tasks that are not affected by the latency in shifting between the Datacenters. Implementation is possible only in the case that same entity is in control of more than one DC and they are in nearby locations. This is not possible in the case of the enterprise, HPC, and in-house Datacenters. Therefore, it is not possible to be implemented. It may be possible with colocation DC, as the entity may own and operate more than one DC. But, also,in this case, it is not possible to implement because they are rent out the servers, it is not possible for them to relocate workload in between servers.

Under operating conditions, servers generate and radiate heat and it causes raising the temperature of the servers and server room. The generated heat by servers is directly proportional to the IT load, Higher the IT load, higher is the heat. The Efficiency of servers is highly correlated with the operating temperature. Higher the temperature, lower the efficiency. There it is crucial to implement proper cooling infrastructure in the server room in order to maintain the temperature at the boundary of 18 to 27 °C that defined by ASHRAE standards [19]. In order to maintain the temperature, it is necessary to remove heat from the servers and the surrounding space of server i.e. the server room.

Majority of small data centers are equipped with air cooling systems. Liquid cooling in combination with air cooling is utilized in modern and large data centers, as the heat transfer is more efficient in case of liquid as compared to air. The energy consumed for cooling the data centers is the largest non-IT component hampering the overall PUE of the Datacenter.

For a certain amount of IT load, the energy consumption by cooling equipment will approximately remain the same, irrespective of the workload shifting. But, it is possible to adjust the cooling intensity at certain periods of time, when the IT load os lower or higher. This would result in the comparatively lower consumption of energy as it would be possible to cool only as to maintain the temperature between the limits rather than always maintainting at the low value. But at the low value. But, the rise in server temperature will ultimately result in inefficient operation of the operation of servers, therefore counteracting the energy consumption.

Alternatively, the cooling intensity could be controlled depending on the energy prices. During the peak load, energy prices are higher compared to off-peak periods. The server room can either be pre-cooled or past-cooled to a lower temperature when the energy prices are lower. Pre-cooled server room would take longer to heat up to the limits while under high load, As a result reducing the required energy during the peak periods. Even though the overall energy consumption would approximately be the same, most of the energy used was at a lower price than usual. Thus the ultimate cost of energy would be lower than usual. The result of implementation will not be in terms of improved PUE but would reduce the operating costs. This type of flexibility can be implemented in every type of Data centers except HPC. HPC servers are almost always under high load, so it is necessary to remove the heat and maintain the temperature to a set point. Degrading in server performance is not acceptable in HPC data centers. Public cloud providers already have a PUE of almost 1.0, so the actual energy consumption for cooling equipment is much lower and the benefits in terms of operating costs would also be lower as compared to any other DC with lower PUE.

Every data center is bounded by QOS agreement which specifies the availability i.e. the Uptime of the data centers. According to the standard of TIA (Telecommunication Industry Association) [12], Data centers are classified into four different tiers: Tier 1, Tier 2, Tier 3, Tier 4. These tiers are used to define the uptime of Datacenter. The required Uptimes will be in the following table:

Table 2 Required Uptime

Tier1: Basic	99.671% Availability
Tier2: Redundant Components	99.741% Availability
Tier3: Concurrently Maintainable	99.982% Availability
Tier4: Fault Tolerant	99.995% Availability

In order to achieve the necessary uptimes, Data centers are required to have redundancies in place.Higher Tiers of Data center needs to have higher requirements . Data centers are equipped with power backup option. When power failure happens IT components cannot work and they need to turn off, to make sure of this, UPS with battery storage systems are installed in conjunction with Diesel Gensets. Depending on the Tier higher degree of redundancies are required to be provided. Therefore, the higher number of backup systems are in place in a Datacenter. Irrespective of the Type of the DC, they have to follow the Tier protocol and install necessary Backup systems. The most common are UPS and Diesel Gensets. UPS are required to provide supply to the IT equipment in case of power failure till the Diesel Genset starts. UPS is utilized for a short duration usually less than 15 minutes. With the high degree of

redundancies, these systems remain underutilized in most cases. Thus are ideal for consideration as a flex option in a DC.

Few modern Data Centers are equipped with their own Photovoltaic Power Generating Units. The excess energy generated from the PV units can be fed to the grid, to earn benefit from bidirectional metering mechanisms in place.

Diesel Gensets are capable of providing power to the entire DC i.e. are typically in MW scale. This can be used in the case at any point in the day the energy price is higher than the cost of running a Diesel Genset. Although the cost of operation of Diesel Genset is generally higher, thus it is not the best solution.

Independent of the Type of the Data Center, every Data Center has UPS and Battery Storage. For higher Tiers, even multiples of these systems are in place.

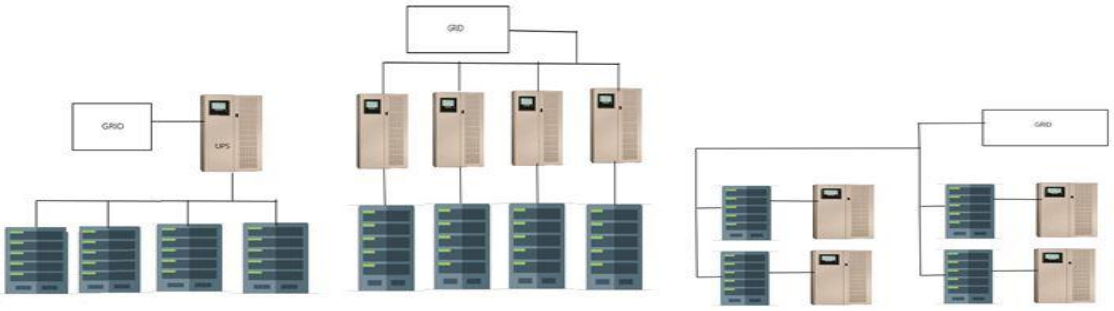


Figure 4 UPS Topologies - (a) Centralized, (b) Distributed per Rack, (c) Distributed per Server

In most of the DC, a central UPS system is installed which supplies the power to the entire IT load in case of power failure. In modern DC such as Google, Facebook, a distributed UPS solutions are installed to provide the backup power. Figure 4 represents the various topologies that are currently implemented in Datacenters. Distributed systems are either on per rack or even per server (Google) basis, thus each UPS only provides power to a particular rack or a server. Multiple smaller UPS are installed instead of a single large one. The advantage of using distributed UPS over a centralized one it that, due to the higher number of UPS, the failure in one only affects that particular rack or server rather than the entire IT load. A centralized system with larger UPS and large BESS are preferable for flexibility options over the smaller ones.

The UPS can be utilized in order to reduce the operating costs by implementing the possible flexible options. The cost of energy is changing throughout the day, it is higher during the peak periods, and it is cheaper during off-peak periods. Energy from the grid is currently used as required by the IT equipment. Therefore, the operating expenditure is as per the energy costs at the moment. Instead, the battery storage system can be used to store the energy whenever the energy is at a cheaper rate (i.e during off-peak periods) and this stored energy can be used to power the IT equipment when the energy price is higher (i.e peak load periods). The result of this action would be reduced operating expenses as more energy will be used when it is cheaper.

The benefit from this method would be minimized if these systems are employed to provide Ancillary services to the grid as and when required. The UPS and BESS are always connected to the grid. As long as the grid is operational, UPS is not required for the operation of the DC. Additionally, if the DC has the higher degree of redundancies in places (i.e. N, N+1 etc), there are additional UPS and BESS, which are only used in case the first one fails. These additional systems can also be used for the purpose of ancillary services as well. Figure 5 shows the basic N-type arrangement of the UPS. In this type, if N UPS is required for normal operation, only N is equipped without redundancies. For N+1 type, as shown in figure 4, an additional UPS is installed. In case of any of first N unit fails, the additional unit takes its place and the supply is maintained to the IT load.

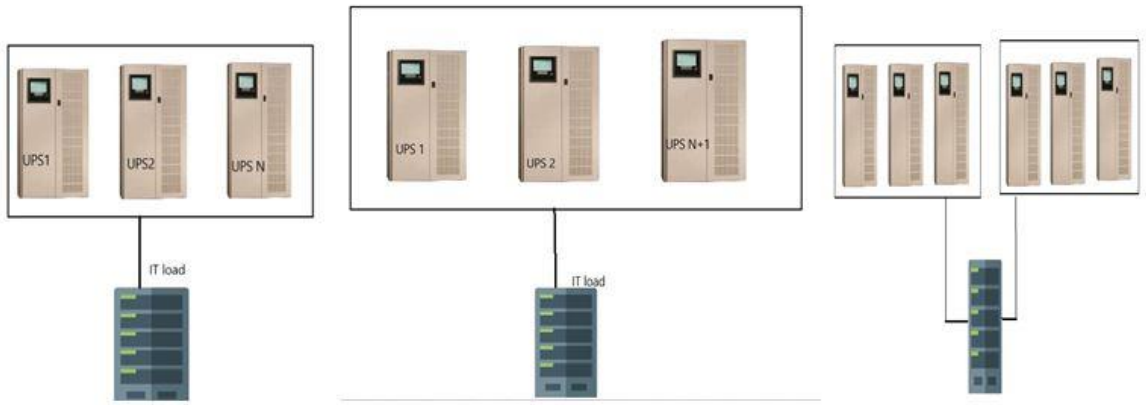


Figure 5 Redundancies - (a) N type, (b) N + 1 type, (c) 2N type

5. General Data center model

A datacenter model incorporating all the necessary components of the data center are implemented in the model. It is a prerequisite as the entire simulations and evaluation carried out during the thesis required a standardized model, so as to study the effects of implementing the flexibility options on the normal operation of the DC.

In the model of the datacenter, certain assumptions were made in order to obtain comparable results regarding BESS and HVAC. Following subsections describe the model and the components in the data center as well as the design process, assumptions and the design choices that were essential to obtain the final model.

5.1 Design of the Data center

A typical block diagram of the data center is required in Figure 5

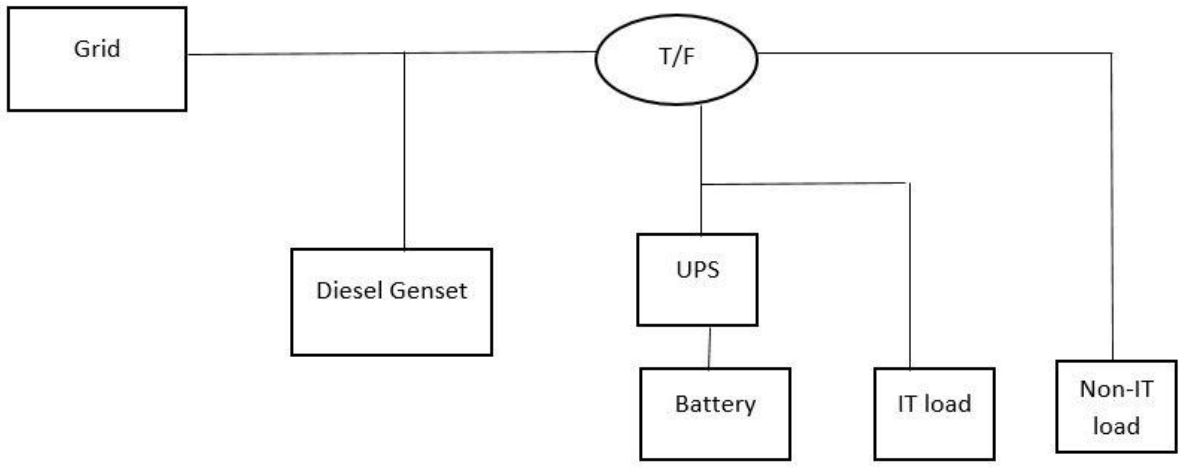


Figure 6 Block Diagram of a Data Center

The data center is connected to the grid through a step-down transformer to obtain 3-phase power supply at the rated voltage. A Diesel Genset is connected in parallel to the grid. The

Diesel Genset operates in tandem with the UPS in case of power failure (Grid failure). The Diesel Genset is capable of providing the necessary power to the entire DC. The stepped down voltage supply is connected directly to the non-IT load in the datacenter. The IT load is connected to the grid, through a UPS and BESS. The UPS is sized to support the IT load in case of power failure until the GenSet takes over. The reason to only provide UPS to IT load is that the inertia of the cooling system(non-IT-load) is sufficient to maintain the temperature until the GenSet is online.

5.1.1 Grid/Power system

Every data center is connected to the Grid, Therefore they highly depend on it for its energy needs. The energy demand of the datacenter (or most of the energy consumers) is insignificant compared to the actual size of the Grid. When considering a single datacenter connected to the Grid, the effect of the load of the DC is negligible on the Grid. Unless the number of DC is taken into consideration, then only the cumulative effect will be visible in the datacenter.

A large amount of power is required to maintain the balance of the Grid via the Ancillary services. Thus, while implementing the flexibility options, specifically, providing Auxilary services with the BESS of the data center, the influence of providing Ancillary service of 1MW/Hour (the minimum required bid to participate in ancillary service) on the Grid would be minor and could be neglected for the simulation purposes.

But in the large scheme of things, participating in the ancillary service market would help stabilize the grid even if the effect would not be observable on its own. Additionally, implementing the flexibility techniques will be beneficial to the datacenter, as it would provide the monetary benefit to the operator.

5.1.2 Transformers

A transformer is connected in between the Grid and the connection of the Loads. In the model, a separate transformer is connected to IT load and another one for other loads. The transformers are required to step down the voltage from 20000V to 400V which required for the load. Both the transformers. The parameters of the transformers are as follows:

Type of Transformers	Star – Star, 3 Phase, Ideal
Nominal Voltage	20000 V/ 400 V
Nominal frequency	50 Hz
Nominal power	1 MW

Alternatively, to participate in ancillary services, it is necessary to set up the voltage from nominal voltage of the inverter i.e 400V to the grid voltage of 20000V.

5.1.3 Diesel Generator Set

The purpose of the Diesel GenSet is to supply electricity during the emergency situations (such as power outage) conditions to the entire datacenter. The Diesel Genset is rated at 2MVA and 20000V. It is connected to the datacenter before the transformers. The Diesel Genset is capable of altering the output depending on the set points chosen as per load requirements. The output can be changed by changing the amount of Fuel injected into the Diesel motor. Which would be utilized for providing variable power demands of the datacenter.

It is capable electricity source, which potentially can be utilized for providing ancilary services to the grid. The startup times of a Diesel Genset typically lies between 30 seconds up to a couple of minutes. The startup time of Diesel Genset could be a major hurdle for participating in Ancillary services.

Although theoretically, it seems possible to combine the UPS and Diesel Genset to provide ancillary services, The major obstacle in utilizing Diesel Genset for a longer period of time, is the cost of fuel. Typically, cost of electricity from a generator is considerably higher than the electricity available from the Grid.

5.1.4 IT load

The power consumption in a datacenter is Distributed as shown in figure 6,In a typical datacenter (with a PUE of 2.0), the IT load consumes about 50% of total energy consumption of the datacenter.. This may vary depending on the quality of the equipment and their efficiencies. The power consumption of IT load in DC can be distributed as shown in figure 7.

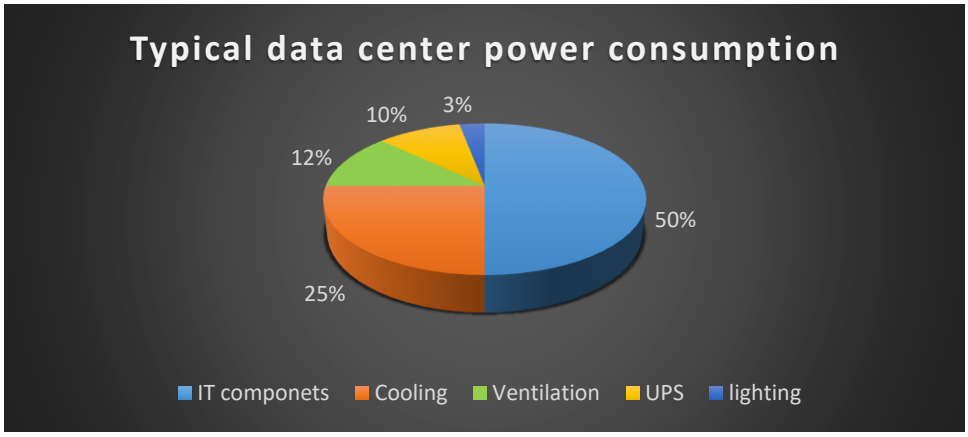


Figure 7 Typical data center power consumption

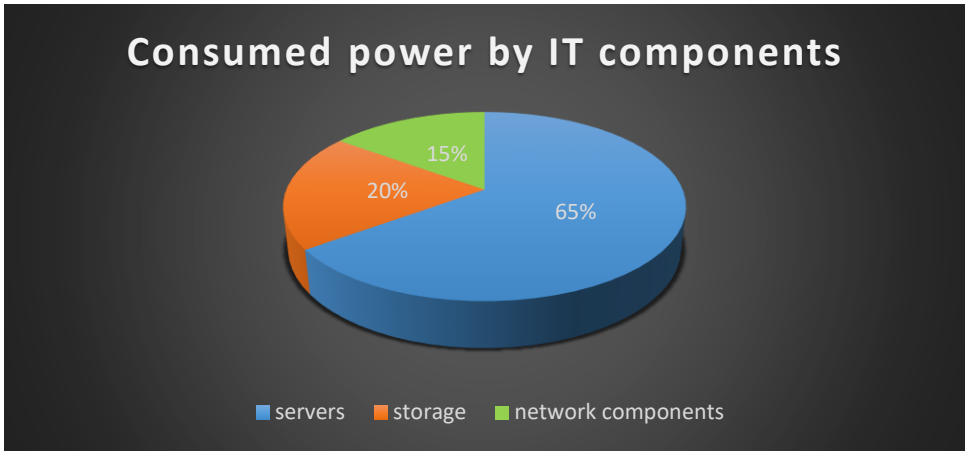


Figure 8 Consumed power by IT components

From the total IT power consumption, 65% percent of the power consumed by servers, The storage consumes about 20% while the network devices consume 15% of the power.[15][16]

The power consumed by the servers varies greatly depending on the load on the Servers. Power consumed by a server at full load is almost twice as much as consumed in the idle state.Similar to the Server, the power consumption by the storage devices varies depending on the state of the device i.e. whether it is active or is it idle. The consumption also depends on the amount of Data stored in the devices. The power consumption of storage devices is related to the server

loads. The network devices consume a considerable amount of power and depends on the status of various switches employed in the network and cabling.

5.1.5 Non-IT load

Apart from the power consumption of the IT components, the non-IT load consumes the almost equal amount of power in the datacenter (Considering a PUE of 2.0). The non-IT load comprises the majority of cooling equipment required for maintaining the temperature of the servers and the server room within the requried temperature range. A small portion of power is consumed by the lighting equipment installed in the Datacenter. The distribution system and UPS operation efficiency are responsible for the remaining power consumption.

The power requirement of the cooling equipment depends directly on the heat generated by the servers in the server rooms. The servers generate more heat under high load and vice versa. Thus the power consumed by cooling equipment varies as the load on servers varies throughout the day.

As specified in section 7.1.4, the IT load in the datacenter, is considered to be stable/constant throughout the operation of the DC. As server load is kept server is kept constant, the heat generated by these servers, will also more or less be the same. The power consumed by the cooling equipment will in turn remain constant (without considering the weather changes).

5.1.6 Battery Energy Storage System (BESS)

The Battery Energy Storage System is necessary along with the UPS, to provide the power to the IT load, as when required. The BESS is kept in the charged condition when the Grid is available, and in case of power failure, this stored energy is supplied to the IT load through the UPS.

Traditionally, Lead Acid (LA) batteries have been used for the storage purposes. But with the advancements in the Lithium Ion technologies over the last decade, Li-Ion batteries can now be considered for the application instead of Lead Acid batteries.

The main drawback of the Li-Ion batteries is the Capital Cost of the batteries as compared to LA batteries. But the price of Li-Ion batteries has gone down considerably over the last past decade. Apart from the price, Li-Ion batteries provide many advantages over the LA batteries which are described the following table as the comparison between these two technologies:

	Lead Acid battery	Lithium Ion Battery
Energy Density	35-40 Wh/Kg	140-150 Wh/Kg
Weight	100%	Approx.25% of LA
Overall Efficiency	80-85%	90-95%
CapEx Cost	Approx. euro 70/kWh	Approx. euro 150/kWh
Life cycle	Around 4 years	Around 12 years

The longer battery life and the high number of charge-discharge cycles make it an ideal alternative to LA batteries.

Many of the newer Data centers are opting for Li-ion batteries instead of LA batteries in light of advantages that they provide.

The sizing of the BESS is dependent on the IT load and the amount of power they need to provide in case of power outage. Typically, The BESS is sized in such a way, that it can supply the IT load until the Diesel Generator is online, but is not limited to just provide it for a single run. They are sized comparatively larger than what is required for a single cycle. As in case of the consecutive power failure, it needs to handle the load for a longer duration and multiple times.

6. Prediction

The energy market is facing the paradigm shift. Increasing shares of fluctuating renewables and decentralized power generation represent a challenge to power networks, which were originally designed for centralized production. One approach to meet this challenge is to introduce flexibility on the demand side of the power network, as known as demand-side management. In this context, smart homes, which can match their electricity demand with available supply, become increasingly crucial. An *Energy Management System (EMS)* provides the ground for a home being called "smart" since the EMS can control selected electrical and thermal devices. The performance of the EMS can be enhanced by employing electrical load forecasts, such that the EMS can predict electricity consumption of the smart home and optimize consumption under different objectives. However, for this optimization, an accurate forecast of the electrical load at the household level is required.

This section of master thesis starts with a literature review is conducted into already existing methods of electrical load forecasting and then we select the method and do the implementation and the results from the implemented method will be evaluated.

6.1 Literature review

The following sub-chapter is devoted to related work in electrical STLF(Short Term Load Forecasting) at Data center. Datacenter. Different approaches for STLF exist in the literature. Without decent experience in each proposed methods, it is difficult to judge which forecasting method is best suited for the implementation within an EMS. Hence, a literature review on STLF at the datacenter is conducted, targeting the identification of the most promising forecasting methods.

Throughout the paper, the term *method* refers to a general concept, such as Artificial Neural Networks (ANN). ANNs can be employed in many different forms, such as a Feed Forward Neural Network (FFNN), Echo State Network or Recurrent Neural Network. Accordingly, the term *model* is used for a specific application of a method, for instance, an FFNN with specified model parameters.

The objective of this chapter is not only to find the most promising forecasting method(s), but also to elaborate on specific models and *input variables* used in the literature, wherein input variables refer to the variables being used as input to the model. This will be particularly useful for the implementation part of the thesis. In the following, the term *lagged electrical load* will be frequently used. "Lagged" means the same as "past" and is usually used when a variable y_t shall be forecasted based on its past values, for instance, one hour before ($y_{t-1}h$). The term lagged variable is commonly used in statistics and shall also be used within this thesis.

First, a brief overview of applied methods in considered literature is given. Second, the findings of the literature on STLF are summarized by means of tables. Specifically, the research papers are examined using the following approach:

1. Identification of a subset of promising forecasting *methods* by comparing forecasting

errors of different methods reported in research papers.

2. Examination of the data sets underlying each research paper.

3. Identification of promising forecasting *models* in the literature.

4. Investigation of commonly used input variables particularly lagged electrical loads and exogenous variables.

In order to have the clear idea of different forecasting methods in the literature, a short overview is given.

At this point, it should be mentioned that only research papers that are dealing with electrical STLF at a datacenter level are considered within this literature review. Electrical STLF on aggerations of Datacenters, such as distribution or transmission level, might use the variety of methods for forecasting the energy consumption. In order to indicate which method it yields the best prediction over the horizon, we need use methods for indicating the forecast accuracy such as MAPE(Mean Absolute Percentage Error), RMSE(Root Mean Absolute square Error) and MSE(Mean Square Error) which are described in detail at the following subsection.

For having STLF, we could use two types of machine learning algorithms which falls into categories of ANN (Artificial Neural Networks) and Time-series forecasting. As an example regarding researchers have been done on ANN category,At study [25] authors are proposed a new network model for predicting the distribution of the temperature in the datacenter based on the real-time senario in order to admit energy-efficient task allocation and facility management at the datacenter level,this study shoes the model which ables to predict the distribution of future temprature that might happen 10-minute in 60 places in 3.3 ms with an RMSE of 0.49 degrees.A study [26] authors introduced a novel method for predicting the hourly energy consumption in the buildings, At the suggested approach they used nonlinear time-series analysis techniques in order to the reconstruction of energy consumption time series. In this model, they used window-size and the sampling lags for data as the factor that affect time-series prediction with the collaboration of neural network systems. A study [27] authors have implemented a General Regression neural networks(GRNN) in order to examine the practibility of applying this technology to optimize HVAC thermal energy storage in public buildings and office buildings.The results of this experiment show that properly designed neural network is a powerful instrument for optimizing thermal energy storage in buildings based on only external temperature records. The other method that we could use in order to predict the Time-series is using Time-series forecasting methods such as ARMA, ARIMA, and ARIMAX. A study [28] authors have estabilished an ARIMAX model in order to predict the power demand of building which a measure of building occupancy was a significant independent variable and enhanced

the accuracy of the model. The results of this experiments represent that implemented model could be applicable and more beneficial on buildings which more occupancy. One possibility to use Time-series forecasting methods is integrating it with ANN(Artificial Neural Networks), As an example in study [29] authors are used hybrid adaptive techniques in order to forecast the electrical load by using ARIMA and ANN and results of the implemented methodology shows that adaption of this work has increased the efficiency of the forecast and managed to reduce the forecasting errors.

6.2 Method Fundamentals

This chapter is devoted to the forecasting method from sub-chapter 6.1, which were identified as being most promising from STLF on Datacenter level. The objective of this chapter is to give an outright of each method. To do so, first, a basic understanding of general time series analysis is given, as all the forecasting methods are based on time series data. Second, the method fundamentals and important model parameters that impact model outcome will be set forth.

At this point, the following result can participate: considering the parameters influencing model results, Artificial Neural Networks (ANN) seems to be a most difficult model to set up. Selecting proper model parameters for ANN requires decent expertise with regards to implementation. Although a genetic algorithm could be used as an alternative approach in order to select model parameters, setting up a genetic algorithm is expected to be the most time consuming with regards to implementation. Therefore, the implementation of ANN shall be passed to future work in this area.Hence, from this point on, this thesis in terms of prediction focuses on Autoregressive-Integrated-Moving average (ARIMA).

6.2.1 About Time series Analysis

Electrical load of a Datacenter that is tracked over time represents a time series. This section shall introduce time series data and general concepts of time series analysis, particularly stationarity and differencing.

A time-series is usually defined as a set of quantitative observations arranged in chronological order. Based on this definition, time-series is the classification of a certain type of dataset, where observations of one or more variables are tracked over time. Time series could be considered ad a realization of the stochastic process. As an example, the electrical load which consumed by the datacentre over the time can be considered as time-series dataset. On the contrary, a

cross-sectional dataset consists of the sample of individuals such as households, individuals, firms, etc. which sampled at a given point in time. Therefore, cross-sectional dataset mostly consists observations of many individuals which collected at a fixed point in time.

For cross-sectional data, it can usually assume that the data obtained by random sampling. Random sampling means that a group of individuals (the sample) is chosen from a larger group (the population). Each individual has a certain chance of being chosen and is selected independently from other subjects of the population. The random sampling assumption simplifies the analysis of cross-sectional data.

In contrast, time series data is harder to analyze, since the random sampling assumption no longer holds true. The reason is that observations can hardly be assumed to be independent of the other. In our electrical load example from above, the electrical load of a data center highly depends on the power that consumed by different subsections of the data center such as HVAC system, BESS etc. Hence, observations across time can rarely be independent as all observations depend on the same, specific behavior of components. Specifically, the electrical load of Datacenter at the time (t) will likely be dependent on the load at the time (t-1). However, the missing random sampling assumption can be cured if the time series data can be considered as a weakly stationary stochastic process. The formal definition of stationary data is defined as follows: the term stationarity means that the probability distribution of a stochastic process does not change upon shifting time (e.g. from t to t +1). However, at time-series there is the possibility of executing the trend exist, it means that the behavior, in this case, will be non-stationary. At time-series data, stationarity is a crucial concept which causes the stability over the time horizon.Otherwise, if the relationship between two or more variables randomly changes over time, it is difficult to gain any knowledge about the underlying process. However, time series often exhibit non-stationarity. Particularly, non-stationarity is indicated in time plots that exhibit a trend, seasonality and/or non-constant variance.

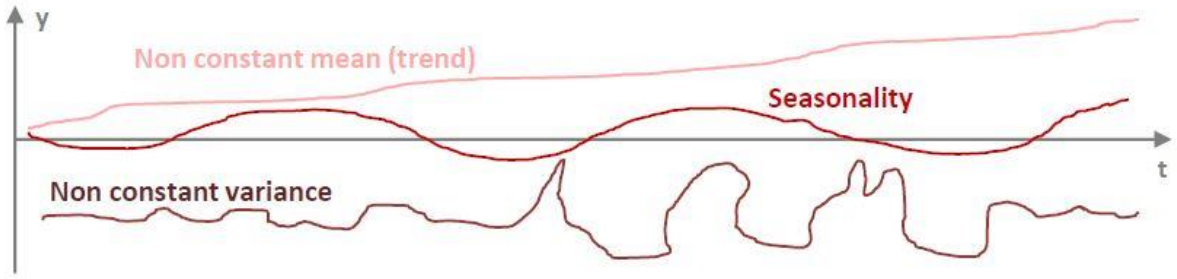


Figure 9

If stationarity is not at hand, the model results are likely to be biased: The results obtained may be spurious, in that they indicate a relationship between two variables, although none exists. Nevertheless, non-stationarity can be cured to a certain degree by either using filters or taking non-stationarity explicitly into account in the model. At this point, only the most widely used techniques for trend and stationarity elimination shall be introduced, namely the method of differencing.

A stochastic or deterministic trend of a non-stationary process can be eliminated by taking the first order differences of the random variable.

$$y'_t = y_t - y_{t-1}$$

By using the method of first differencing, a new stationarity process of the random variable y'_t is obtained. The models are then built upon this new process. If y'_t still exhibits non-stationarity, the second order differences can be taken: $y''_t = y'_t - y'_{t-1}$. Usually, the first order differences are sufficient in order to maintain a stationary process. In other words, if a time series exhibits a trend, it is non-stationary. However, if the difference between two consecutive time steps $y_t - y_{t-1}$ has a constant mean, a new stochastic process y'_t can be built which consists of the differences between consecutive time steps. The same concept can be applied to a stochastic process that exhibits seasonality, by taking the differences of seasonal cycles.

In conclusion, the performance of statistical and machine-learning models is highly dependent on the underlying data set. In the context of electrical load forecasting at a Datacentre level, stationarity becomes crucial, since the electrical load of Datacentres is likely to be non-stationary (seasonality, non-constant variance). Further explanations shall follow within this thesis. This section was devoted to time series data in order to get an intuition of stationarity. In the following sections, the aforementioned forecasting methods of time series data are examined, starting with the ARIMA model.

6.2.2 ARMA

The Autoregressive Moving Average (ARMA) method was introduced by was introduced by George E.P. Box and Gwilym M. Jenkins in 1970 and has become a very popular tool for short-term forecasting. The basic idea of ARMA is to forecast random variable as the linear function of its past values. ARMA can be classified as a special type of linear regression for time series data. ARMA consists of two parts, namely Autoregressive (AR) and Moving Average (MA), which are described in following sections.

Autoregressive process (AR):

An autoregressive model expresses a time series as a linear function of its past values. A formal definition of an AR process is:

$$y_t = p_0 + p_1 y_{t-1} + p_2 y_{t-2} + \cdots + p_p y_{t-p} + \epsilon_t$$

The order p denotes the number of lagged variables that are included in the model. AR models can be used for forecasting, wherein y_t would be considered as a one-step-ahead forecast based on its past values y_{t-1} , y_{t-2} , y_{t-p} .

The coefficients p_0 , p_1 , p_p of an autoregressive model indicate the strength of the linear relationship between y_t and its lagged values. Most commonly, the coefficients are estimated by Ordinary Least squares (OLS). Alternatively, Maximum Likelihood (ML) methods could be employed.

An AR(p) process the exhibits one or more coefficients $|\rho| \geq 1$ can never be mean stationary, as the process exhibits a trend. Hence, ρ is restricted to $|\rho| < 1$. This condition called *invertibility condition*. Invertibility and stationarity may not be used interchangeably. Invertibility is specifically referring to AR, Moving Average (MA) and ARMA processes, whereas stationarity is independent of the model used and directed towards the general underlying stochastic process. Non-invertibility always indicates non-stationarity, as the underlying process exhibits a trend

6.2.3 Moving-Average process (MA)

A moving-average model is an (unevenly) weighted average of the historic model error. Specifically, an MA process is a linear function of on y_t its past forecasting errors. MA(q) can be formally described as:

$$y_t = a_0 + a_1\varepsilon_{t-1} + a_2\varepsilon_{t-2} + \cdots + a_q\varepsilon_{t-q} + \varepsilon_t$$

The order q denotes the number of the lagged error, that are included in the model. In order to build a MA(q) model, the errors of MA(q) process have to be estimated. Error estimates are called residuals and can be obtained by building an AR (∞) model which is used to forecast y_t. The difference between the forecasted values of the AR (∞) model and the real values (observations) yield the residuals. Consequently, estimating the coefficients $a_0, a_1, \ldots, a_q$ requires that the MA(q) process can be expressed as a stationary AR (∞) process. Therefore, the invertibility condition of AR (∞) process is also required for MA models.

6.2.3 Autoregressive Moving-Average Process (ARIMA)

A model that contains both AR(p) and MA(q) is called ARIMA (p, q) and it will be defined according to the following equation:

$$y_t = p_0 + p_1y_{t-1} + p_1y_{t-2} + p_2y_{t-3} + \cdots + p_py_{t-p} + \alpha_0 + \alpha_1\varepsilon_{t-1} + \cdots + \alpha_q\varepsilon_{t-q}$$

Estimating the coefficients of an ARMA (p, q) model requires again the invertibility condition. The conditions can consistently be estimated by using maximum likelihood methods. Alternatively, estimating coefficients by OLS is also possible. With regards to predicting electrical loads of Datacentres, the strength of an ARMA model lies within its caption of both, routine consumer activity (AR) and stochastic activities (MA).

Several extensions of the classical ARMA model exist, namely:

- **ARMAX:** exogenous variables are added to the ARMA model Exogenous variables $x_{1,t}, x_{2,t}, \ldots, x_{h,t}$ can help to explain the dependent variable y_t. In case that represents the electrical load of a household, exogenous variables could include outside temperature, rain, outside brightness or occupancy. A linear relationship between the exogenous variables and $y_t$is assumed and OLS or ML can be used in order to estimate their coefficients.

- **ARIMA:** an ARIMA model is like an ARMA model, just that, before the ARMA model is applied, the method of differencing is used upon the underlying stochastic process.

Hence, the “I” in ARIMA stands for “differencing”. An ARIMA model is expressed as ARIMA (p, d, q), where “d” denotes the order of differencing.

All in all, ARIMA is based on the idea of the linear relationship between random number y_t and its past values (AR) and errors (MA). Given the formal definition of an ARMA (p, q) process, there are three model parameters that can influence model results and these need to be set in advance. Choosing degrees (p and q) is part of the model definition as opposed to determine the coefficients as part of the estimation (fitting the defined model to the data). There three model parameters include the order of autoregressive process p, the order of the moving-average q and the method for estimating the model coefficients.

6.3 Implementation

The following chapter lays the foundation for practical part of this master thesis. The preceding chapters concluded that ARIMA shall be implemented within this thesis and already suggested specific models and input parameters. The objective of this chapter is to illustrate how these recommendations are turned into effective practice. Special attention shall be paid to: first, the data sets used for implementation, second, the experimental setup and third, the specific models and model parameters being employed.

6.3.1 Data Sets

Two different datasets are used for modelling, namely the consumed power by Datacentre components in order to cool down the datacenter environment (Pcooling) and IT power. Both data sets indicate the significant linear correlation between electrical loads on consecutive hours and days. These data sets are taken from the experiment that was done by RWTH university by using Modelica simulation software. The duration of each data set is 5334 minutes which is equal to 3.7 days and the granularity of data set is 1min. The cooling power consumption and IT load are illustrated at figure 10 and figure 11.

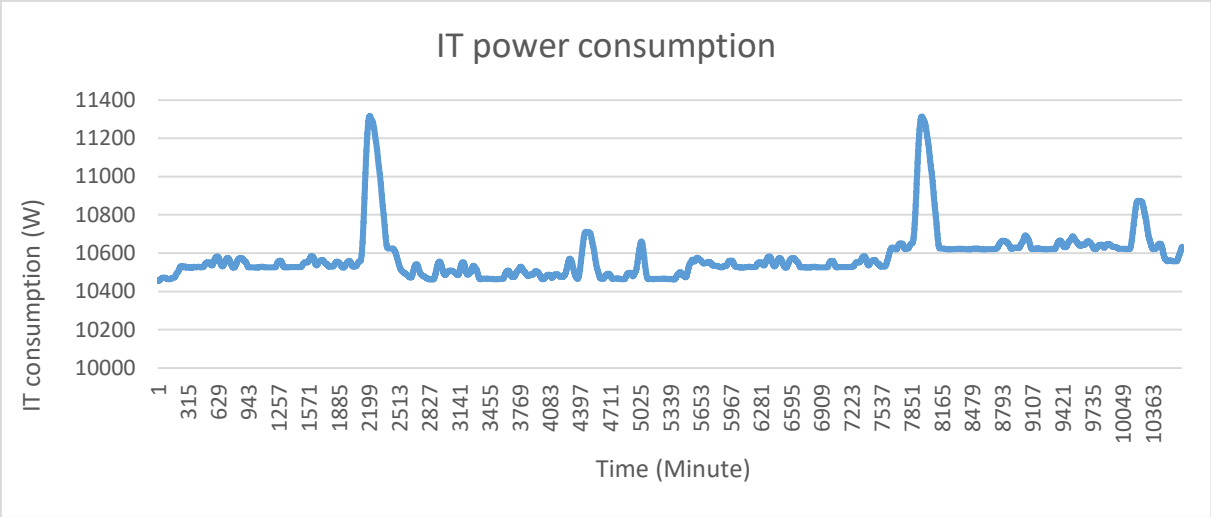


Figure 10 IT power consumption Dataset

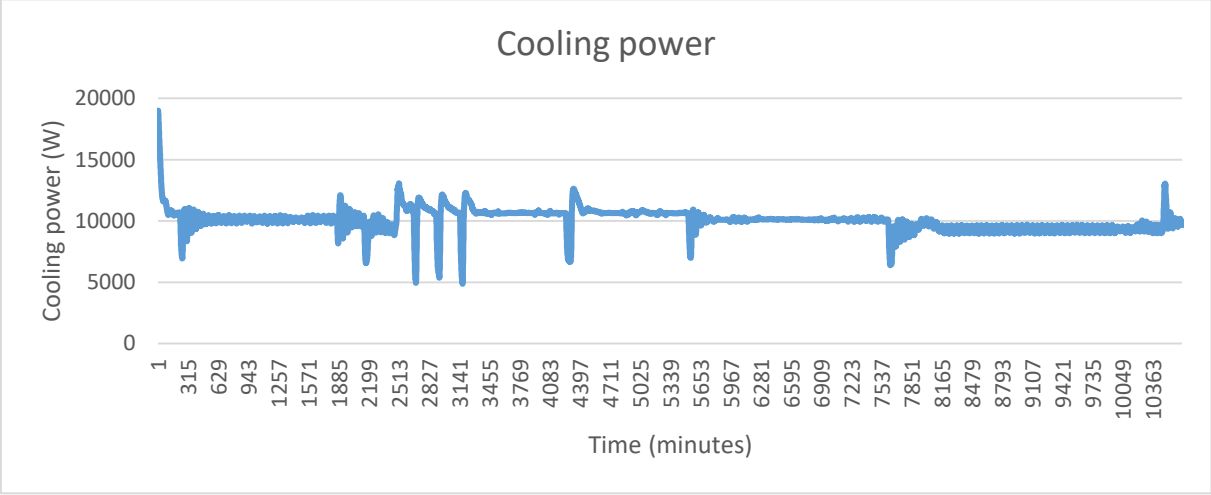


Figure 11 consumed power for cooling purpose Dataset

A popular way for identification of linear relationships between a variable and its lagged values is an autocorrelation plot. Autocorrelation is referring to linear relationship between a variable y_t and its lagged value y_{t-h}, where h denotes the time lag. The autocorrelation plot shows the estimated autocorrelation coefficients p_h of y_t and y_{t-h} for an increasing h. The autocorrelation coefficient p_h can be interpreted as the coefficient of a linear regression of y_t and y_{t-h}.

Figure 11 and Figure 12 show the autocorrelation plot of cooling power and IT load which has the granularity of 1 minute.

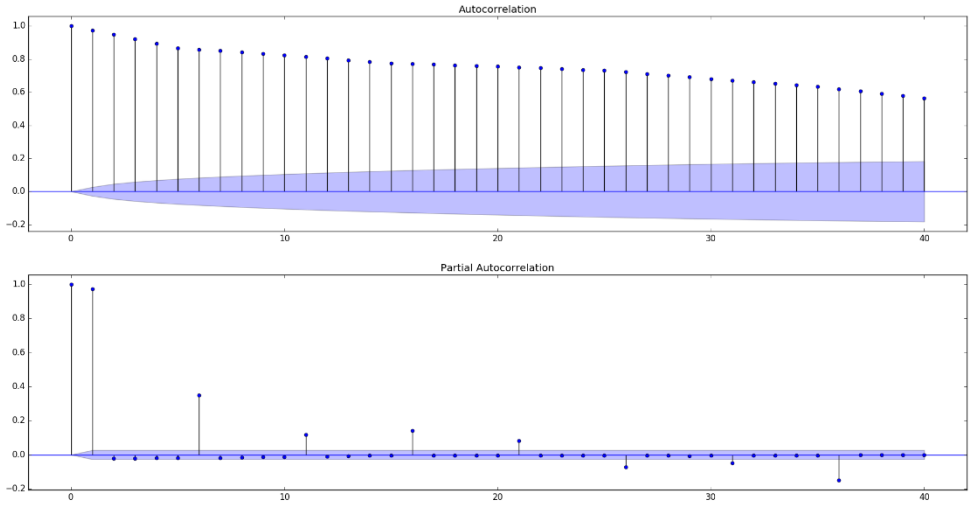


Figure 12 ACF/PACF test result_IT power

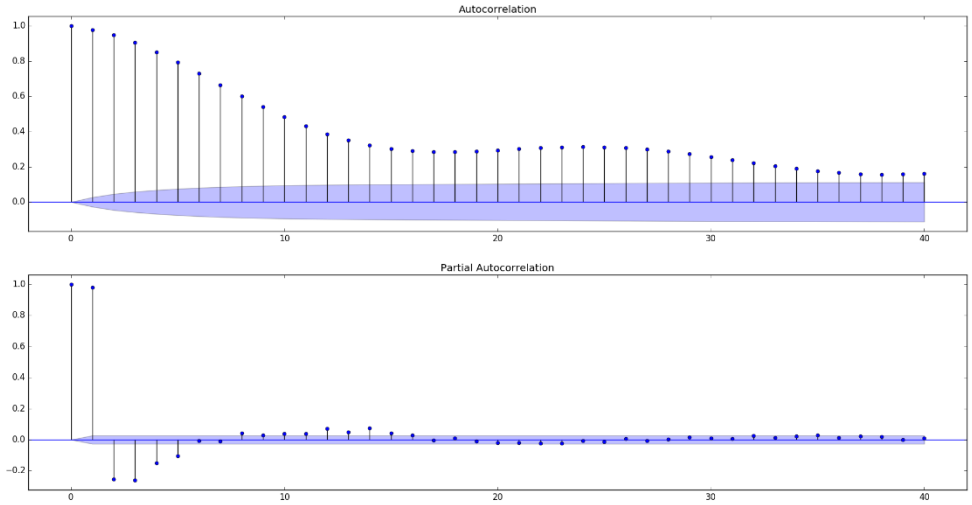


Figure 13 ACF/PACF test result _ Cooling power

6.4 Experimental Setup

The autoregressive section of the ARIMA model which defines as AR indicated to previous values of dependent-variable time-series. Furthermore, the moving average part of ARIMA model which defines by MA symbol refers to lagged error terms which yielded by ARIMA model's in order to generate precise estimates. Hence, ARMA model (ARIMA without considering integrated section) have the same behaviour as the regression model via all RHS (Right-hand-side) variables which been lagged version of the independently y_t as well as the ε_t that related to the lagged version of error terms.

A common order ARMA(p, q) with p autoregressive terms (y_t's) as well as q moving average terms (ϵ_t's) could express by the following equation:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \varepsilon_t - \theta_2 \varepsilon_{t-2} - \theta_3 \varepsilon_{t-3} - \cdots - \theta_q \varepsilon_{t-q}$$

Regrading to the structure, ARIMA (p, d, q) models are the equal to ARMA (p, q) models where time series has first been transformed by differencing. The d defines the order of differencing. As an example, in figure 10 and figure 11 shows the original time-series (d=0) of consumed power for cooling purpose and consumed power by IT components. Then figure 14 and figure 15 shows the differenced dataset by one order of difference which is equal to d=1.Due to the fact that time-series need to converged into the stable situation before they could be modelled with Autoregressive and Moving average, one common method is to use the differencing in order to convert a nonstationary time-series into a stationary time-series that the mean and variance are statistically concluded to be stable.

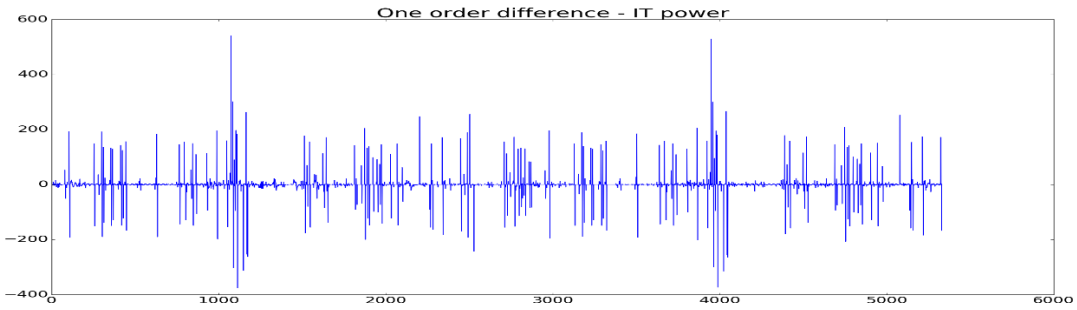


Figure 14 one order difference_ IT power

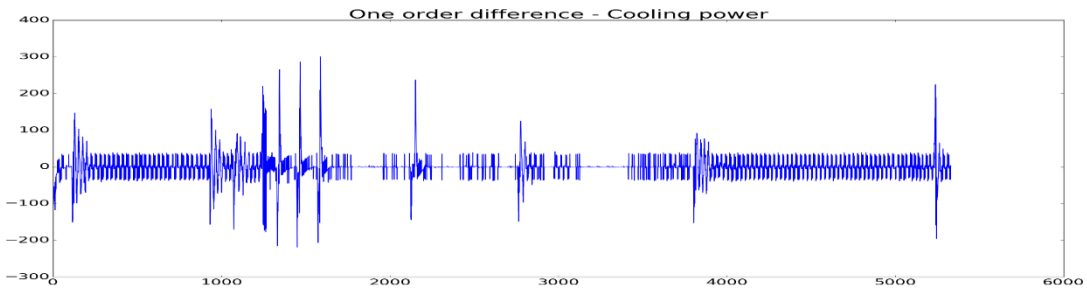


Figure 15 One order of difference_ Cooling power

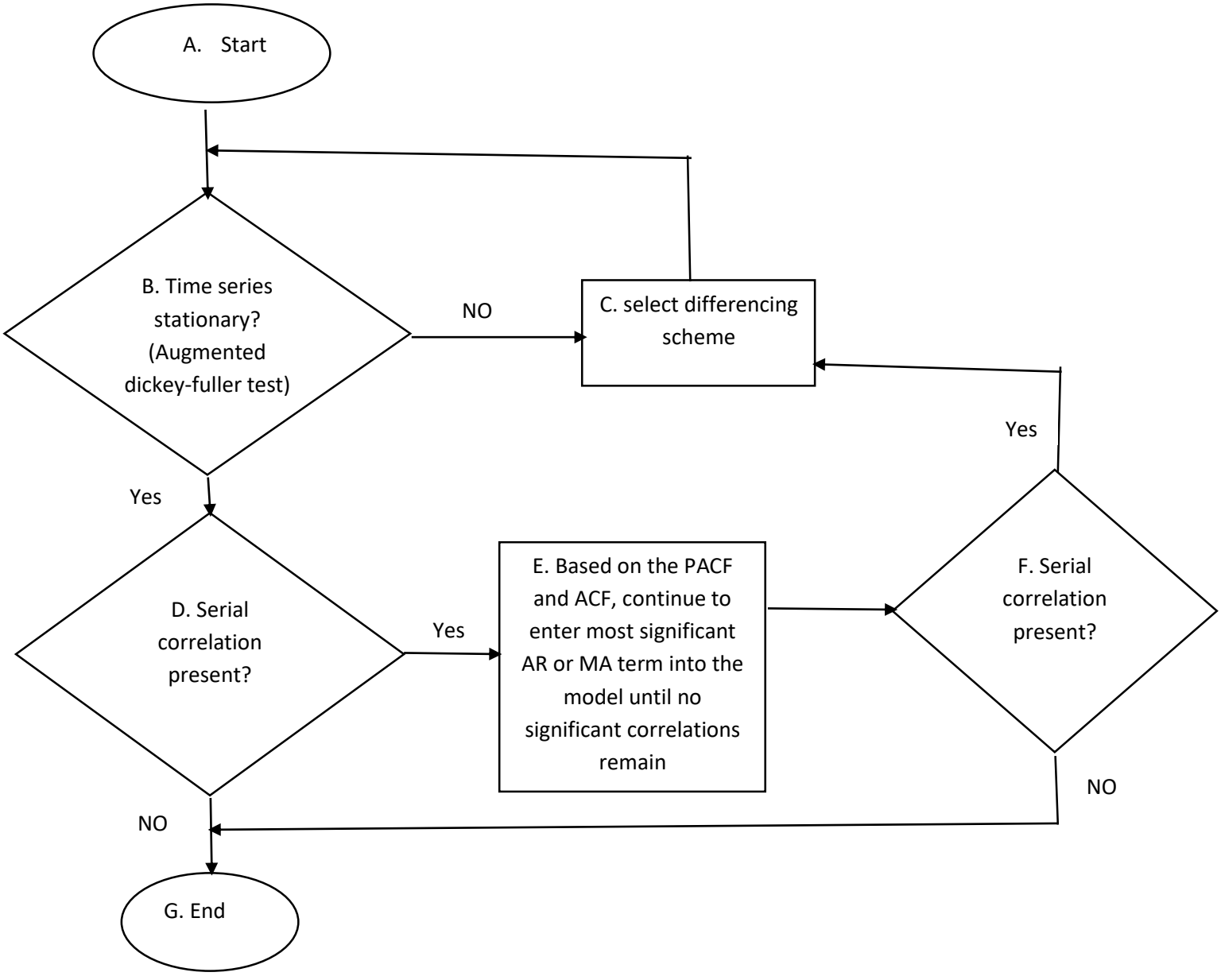
As an example, if we have the time-series dataset which is affected with the days of the week we could perform the differencing by seven for the purpose of removing the day-of-the-week effect. When we take the differenced time-series, this result could show the weekly variation in the daily data and the we have the potential to remove the variance which generated by the day-of-the-week effect. Meanwhile, if we do not have the weekly trend in the original dataset, therefore, the transformed data would possibly show to be stationary with mean value close to zero. Nevertheless, if at the same original time-series we have upward trend in the quadratic fashion, the time-series which were already differenced by seven would perform a positive linear trend and the mean of the dataset will not be constant over the time horizon. In order to deal with the absence of stationarity at time-series, if we difference the outcome of time-series by one, we could get rid of upward trend and the result of this action will be having the relatively constant mean at double-transformed time-series. We could evaluate the stationarity degree of the transformed time-series by using the augmented Dicky-Fuller test which is described in subsection 6.4.1.

Creating ARIMA/ARMA model could initiate when a time-series is statistically analysed in terms of stationarity. In order to recognize the AR and MA terms we need to create the model builder that able to investigate the autocorrelation coefficient function (ACF) and partial autocorrelation coefficient function (PACF). For having the clear image regarding the ACF and PACF test, results of these two tests were represented in figure 12 and figure 13.

At the fundamental level, we have two types of ARIMA/ARMA models namely subset and order. An order model for both ARIMA (p, d, q) and ARMA (p, q) in order to estimate y_t is consist of p terms which include of $y_{t-1}, y_{t-2}, y_{t-3}, \dots, y_{t-p}$ and q terms involve $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-n}$. Particularly, AR terms would comprise lags of 1 to p and MA would consist of lags from 1 to q. On the contrary, a subset model consists of only particularized lags for AR terms and MA terms.

In order to recognize the form of ARMA or ARIMA model, we need to the reiterative process which needs to select proper differencing scheme for reaching stationarity as indicated by ACF test. After that, suitable lagged AR and MA terms are proposed according to the notable patterns displayed by their correlation functions. When we defined each MA and AR term, the residuals are reassessing for significant applying of ACF as well as PACF. This procedure will continue until ACF and PACF functions generate no more statistical sign to demonstrate that any AR or MA terms are missing.

The flowchart that presented at the following represented the series of phases that should be pursued to generate a reliable ARIMA model. At this flowchart we have two nested loop structure, hence this procedure may take a while to be completed.



Creating an ARIMA model for energy-consumption time series needs to execute the following mandatory five steps which are labelled as B, C, D, E and F at mentioned flowchart.

1. The row time series which represent the energy consumed for cooling purpose and consumed power by IT components which exist in the Datacenter must be evaluated for stationarity by using methodologies which are designed for this purpose such as KPSS

and ADF-test (Augmented Dickey Fuller test), more detail about these stationarity test are described at following sub-section. The output of Python regarding stationarity test shows that time-series datasets are weakly stationary, the P-value is very small and do not perfectly patronage rejection of the null hypothesis.

2. In Figure 9 and 10, we could clearly see that there is neither upward nor downward trend exist in the dataset. Therefore, we could easily skip the removing seasonality in order to have the stable pattern (Step C in the flowchart).

3. Since our dataset has the constant mean the best differencing transformation in order to provide minimum variance is one order of differencing. Moreover, this is substituted by evaluating the five highly significant p-values for lags from 0 to 4 by using the ACF test (step B). The five positive p-values regarding the single-mean model via lags 0-4 support rejection of the null hypothesis claiming a distinctive mean at every lag values. This results shows that the data that we got after doing the differencing is stationary and at this dataset we have the single mean.

4. When we take the time-series which is stationary, we could start to build the ARIMA model. Inspection of the ACF plot of residuals (Step E) demonstrate that an MA3 for dataset associated with IT power and MA 2 regarding dataset of cooling power was needed according to their highly significant negative correlation. The PACF accommodated proof that AR 5 term was significant therefore we select 5 as the AR(Autoregressive) order.

5. At the next step (Step E), the ACF and PACF regarding the first-stage were re-evaluated in order to recognize the subsequent Moving Average or Auto Regression applicants that have the capability to generate the numerically considerable descriptive contribution to model the first stage residuals. The autocorrelation function at lag 1 and partial autocorrelation function at lag 1 both have particularly substantial similar correlations. Although, the approximation regarding to the standard error for the ACF and PACF applied for calculating the 95% confidence interval is slightly different. In order to compute PACF, we could compute the approximate standard error by $\sqrt{\frac{1}{n}}$ and n refers to the number of data points which exists in the time series. In terms of ACF, the standard error is computed by this equation $\sqrt{\frac{(1+2\sum_{q=1}^{k-1} r_q^2)}{n}}$. at the mentioned equation, k represents the ACF lag which been examined and r_q^2 refers to the autocorrelation of lag q, that lessens to the standard error of ACF.

6. The autocorrelation evaluates the spans of lags which goes from 1 to 24, gave us the small indication that autocorrelation stayed in the residuals. With p-values which spanned from 0.0920 to 0.4341, there was not enough proof to support rejection of null hypothesis.

7. As well as applying the diagnostic Ljung-Box test for evaluating the independence of residuals, there are two further presumption regrading to residuals need to be evaluated namely: normality and homoscedasticity.

- Normality: usually the residuals shall be ordinarily spread therefore the t-statistics test is applied to evaluate the importance of Auto regression and Moving Average terms are valid. K-S test is frequently performing in order to check the normality, that excellence of fit among the detected cumulative distribution function (CDF) and fully specified hypothesized cumulative distribution. If the perpendicular interval among two CDF increase, it causes to enlargement of the results which generated by K-S test, that will reduce the chance of acceptance the null hypothesis of normally distributed errors.

  According to the test that has been performed by using Python programming language on the datasets, the K-S statistic is 0.0214 with a noteworthy predicted p-value>0.150 for the N suited distribution. This does not support rejection of the null hypothesis that the model residuals are normally distributed. The average is near to zero and the standard error is 0.0138, Furthermore, the mean if residuals are not statistically different from zero.

- Homoscedasticity: In the ARIMA model in order to support the suitable calculation of impartial standard error which are part of t-statistics as well as F-statistics we need to have the residuals which represent steady variance. In the case that we have the biased standard error, it would propel to aberrant denial of the null hypothesis. In order to see that the variance is fix or varying over the time we could perform the White test, the result of this test which generated by Python indicates that the model residuals do not exhibit homoscedasticity at the p= 0.0500 level.

The experimental setup describes the procedure of generating a forecast. The raw data is seasonally decomposed before being fed to the specific forecasting model. A rolling horizon approach is used in order to continuously generate forecasts for each time step.

Python is used for the implementation part of the forecast, due to reusability reasons within the E. ON Energy Research Center. This part of the thesis made use of the project-oriented Python by creating two programming classes: Masterclass and ARIMA class. The Master Class provides a basic construction in which the forecast is embedded, whereas the actual forecast is conducted by ARIMA Class. Python provides already implemented ARIMA packages. However, each package requires different kinds of data pre-processing before being employed. Consequently, using ARIMA class, which summarize specific data pre-processing and respective ARIMA packages, was found to be more convenient in order to keep overall clearness.

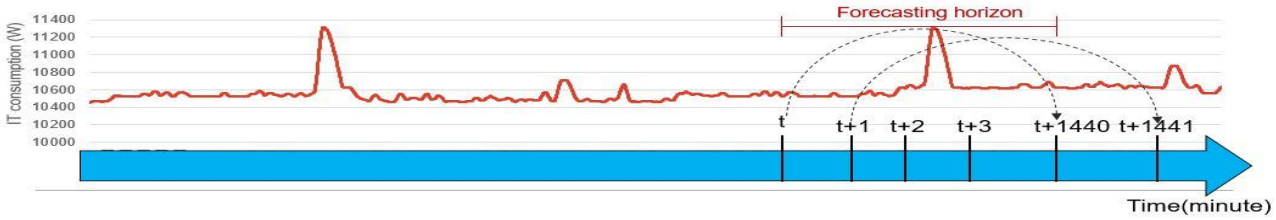


Figure 16 Forecasting process applied over the horizon

The forecasting procedure is designed such that it approximates the real application of Energy management system (EMS). The model user has to define five model-independent parameters in order to generate forecasts: starting point (t), forecasting horizon, the granularity of the data set, size of the training data set and time until specific model parameters are recalculated. After these parameters are set, the program will generate forecasts, starting from time t, for each point of time, it considers the desired granularity. Accordingly, this method of forecasting is called rolling horizon, since the forecasting horizon is fixed, for instance 24h-ahead, but moves ahead in time. After each time step, the underlying input data is updated by the most recent observation and moves ahead by one-time step. Specific model parameters, such as the coefficients of an ARIMA model and the order of p/q are estimated by using the predefined training window. These coefficients are reused for the progressive forecasts until the predefined recalculation window is reached and coefficients are recalculated.

6.4.1 ARIMA model

An ARIMA model is identified by the order p of the autocorrelation process, the order q of the Moving average and I for Integrated. The best order that given by ARIMA package in Python according to lowest MSE (Mean Square Error) is used in the implementation. In order to choose the proper p and q, a grid search is employed: for a predefined range of p, for instance [1,2,3,4,5,6,7], and range of q, for instance [1,2,3,4,5,6,7], the grid search will find the best combination of p and q such that ARIMA model maximise a certain objective function. The most commonly used objective function is Akaike Information Criterion (AIC).

Finally, the orders p, q and d of the ARIMA model are automatically determined by the mean of the grid search. According to the grid search that has implemented at Python programming language with helping mean-square-error module of Sklearn package the best order for IT load Dataset was ARIMA (5,1,3) and the best order for Cooling Power Dataset was ARIMA (5,1,2).

In this section of the thesis, we only consider lagged variables and we do not consider exogenous variables. Exogenous variables such as calendar features or weather data might slightly improve the model. Consequently, input variables being fed to the ARIMA model are lagged electrical load values $yt-h$ of the AR process and lagged estimated model errors $\epsilon t-h$ of the MA process. The specific time lag does not only depend on the order of p/q, which is determined by the grid search, but also by the forecasting horizon.

The first test that needs to be done on the Dataset is stationarity test. In order to do the stationarity, the test we have 2 options: ADF test (augmented Dickey–Fuller test) or KPSS(Kwiatkowski–Phillips–Schmidt–Shin) test.

Here is the brief introduction of these two test:

KPSS: KPSS test determines that the time-series dataset is approximately close to mean or it has the linear trend or it is non-stationary because of unit root that might exists in the dataset. A stationary time series dataset could be indicated by having statistical properties such as mean and variance are stable over the time horizon. KPSS test uses the null hypothesis in order to evaluate whether the dataset that we have is stationary or not. If KPSS reject the null hypothesis it means the time series is not stationary otherwise it means the Time series dataset is stationary.

ADF test: ADF is unit root test for testing the stationarity of the Time series dataset. Unit roots could reason unreliable results when we are performing the analysis on the time-series dataset. ADF test the null hypothesis and if the null hypothesis may not be rejected, it will show the strong evidence of stationarity and vice versa.

In this thesis, we use KPSS test and it shows that the actual Dataset does not shows the strong stationarity.

```
    KPSS Stationarity Test Results
=====================================
Test Statistic                  1.763
P-value                         0.000
Lags                               33
-------------------------------------

Trend: Constant
Critical Values: 0.74 (1%), 0.46 (5%), 0.35 (10%)
Null Hypothesis: The process is weakly stationary.
Alternative Hypothesis: The process contains a unit root.
```

Figure 17 KPSS result for IT power consumption Dataset

```
    KPSS Stationarity Test Results
=====================================
Test Statistic                  4.261
P-value                         0.000
Lags                               33
-------------------------------------

Trend: Constant
Critical Values: 0.74 (1%), 0.46 (5%), 0.35 (10%)
Null Hypothesis: The process is weakly stationary.
Alternative Hypothesis: The process contains a unit root.
```

Figure 18 KPSS result for Cooling power consumption Dataset

Therefore, in order to have the stationary dataset, in order to do it we have two option either performing log transform or perform differencing on the time series Dataset.

After we transfer non-stationary dataset to stationary ones, we could use the best model that was detected by Sklearn package of python according to the mean-square-error which was

ARIMA (5,1,3) for IT power consumption Dataset and ARIMA (5,1,2) for Cooling power consumption Dataset.

6.4.1.1 Accuracy of prediction

When the prediction procedure is done we need to evaluate how well is the prediction, in order to do this, we need the use methods to determine accuracy. The main algorithms that are mainly using for determining the accuracy of prediction are described as following:

MSE (Mean squared Error): MSE is an estimator that evaluates the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss.

The MSE is a measure of the quality of an estimator, it is always non-negative, and values which are closer to zero are better.

$$\frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}|$$

MAPE (Mean Absolute Percentage Error): MAPE is a measure of prediction accuracy of forecasting method. It usually expresses accuracy as a percentage and defined as the following formula:

$$\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y_i}}{y_i}\right|$$

Where y_t is the actual value and ŷ is the forecast value. The difference between y_t and ŷ is divided by the actual value y_t again. The absolute value in this calculation is summed for every forecast point in time and divided by the number of fitted points n. By multiplying 100 makes it a percentage error.

In this master thesis we have tested some possible scenarios and determine which one is more accurate it means that it has lowest MSE, RMSE and MAPE. Results are representing the following tables.

Number of iterations	MSE	RMSE	MAPE
100	0.942	0.97	0.009
500	0.093	0.30	0.002
1000	0.003	0.05	0.0005
1440	1.24	1.11	0.105773609661

Figure 19 Prediction accuracy of IT load

Number of Iteration	Mean square error	Root mean square error	Mean absolute percentage error
10	31253	176.7	1.80
100	5752	75.87	0.7
500	16250	127.47	1.33
1000	18.76	4.33	0.03
1440	11.55	3.39	0.1

Figure 20 Prediction accuracy of cooling power

When all stationarity tests and other data preparation procedure which shown at mentioned flowchart have been done we are able to predict the future behaviour of the power consumption in both of IT power and Cooling power. The results of these two predictions have been shown at following figures:

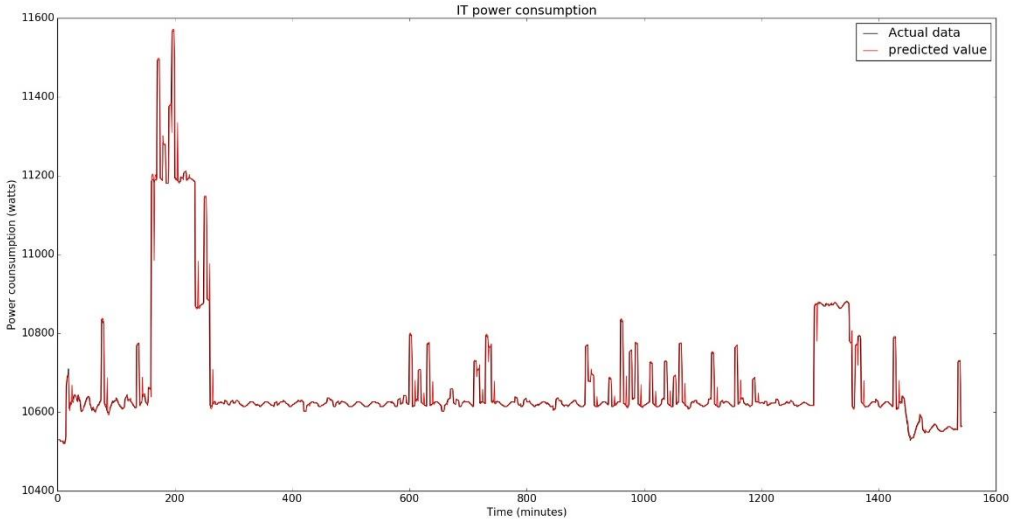


Figure 21 IT power consumption (Actual data and predicted value)

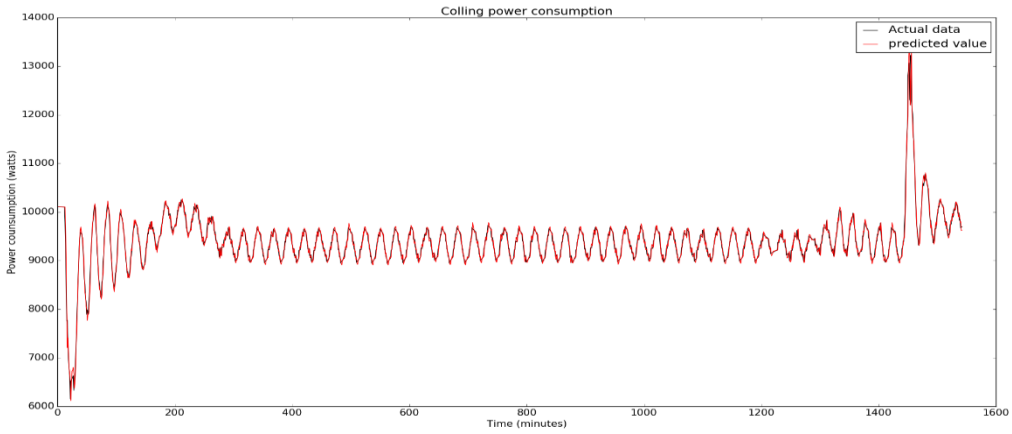


Figure 22 Cooling power consumption (Actual data and predicted value)

7. implementation and modeling of the data center

At this chapter of the master thesis, we start with study the actual data center specification and the overall infrastructure of this datacenter which used in this master thesis. The datacenter that considered at this thesis is located at Hermann-Rietschel-Institute (HRI) of the Technical University of Berlin, Further details are described in subchapter 7.1. the rest of chapter 7 devoted to optimizer model which designed in order to optimize the total power consumption of datacenter and its related information.

7.1 Data center Test Bed:

Datacenter model which is used in this master thesis is defined based on the actual datacenter that exists at TUB university in order to simplify the validation of the DC model components. The test-bed consists of 80 servers with the size which varies between 1U to 4U(unit) that is scattered between five racks with the size of 19 inches. In this test-bed air conditioning is designed according to the idea of "cold-hot-aisle" that racks are conducting the heat transfer from hot aisle to the cold aisle. The datacenter environment has been designed by using Google Sketchup software based on the plot that was provided at [34] and this document consists of Datacenter Test-Bed – Lateral View as well as Datacenter Test-Bed – Top View. The following figures are representing the sliding slice and top slice of the designed datacenter.

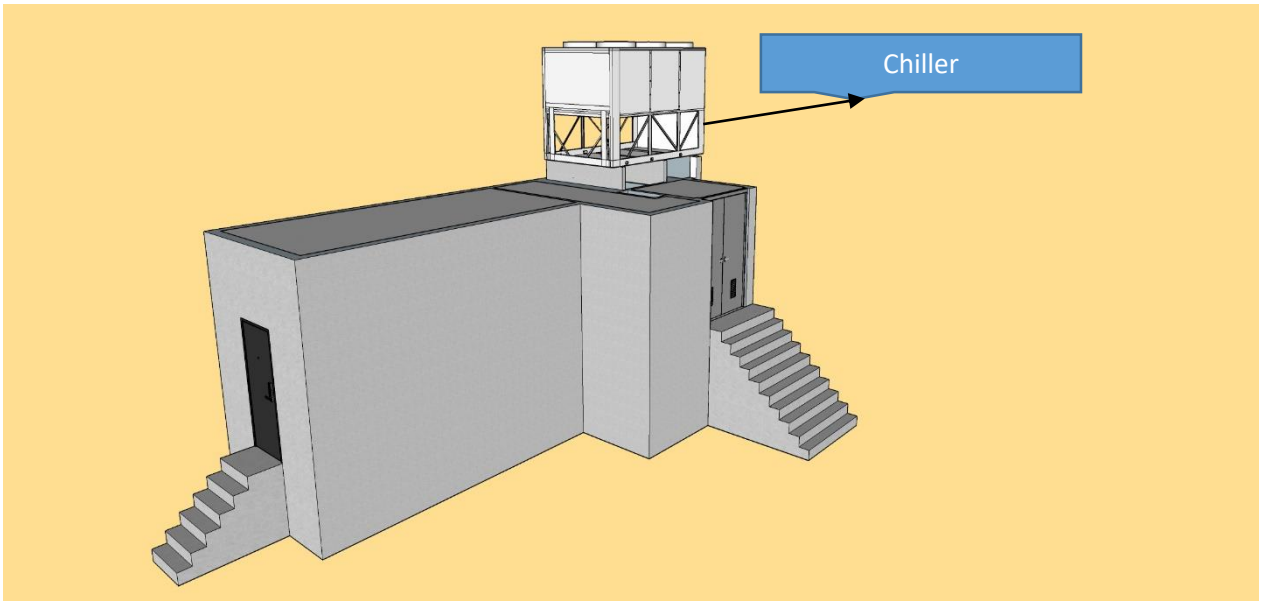


Figure 23 General presentation of Actual datacenter

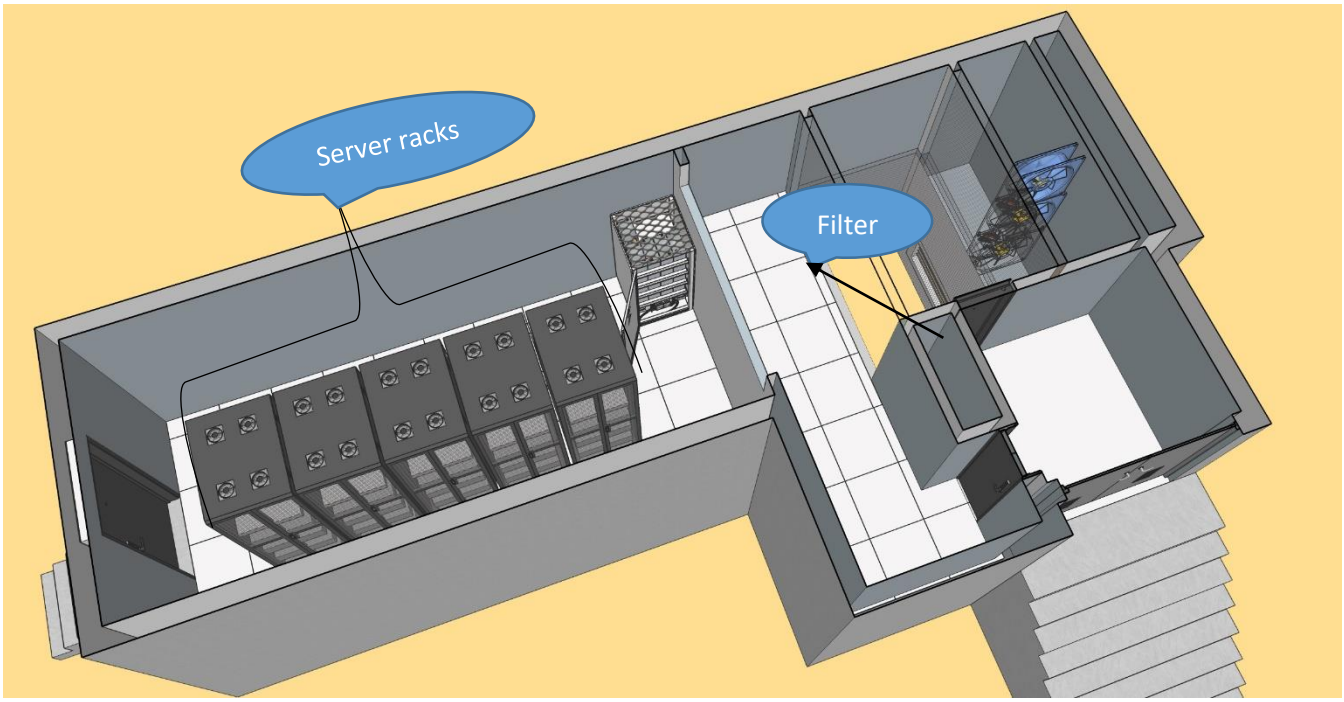


Figure 24 Top view of the datacenter

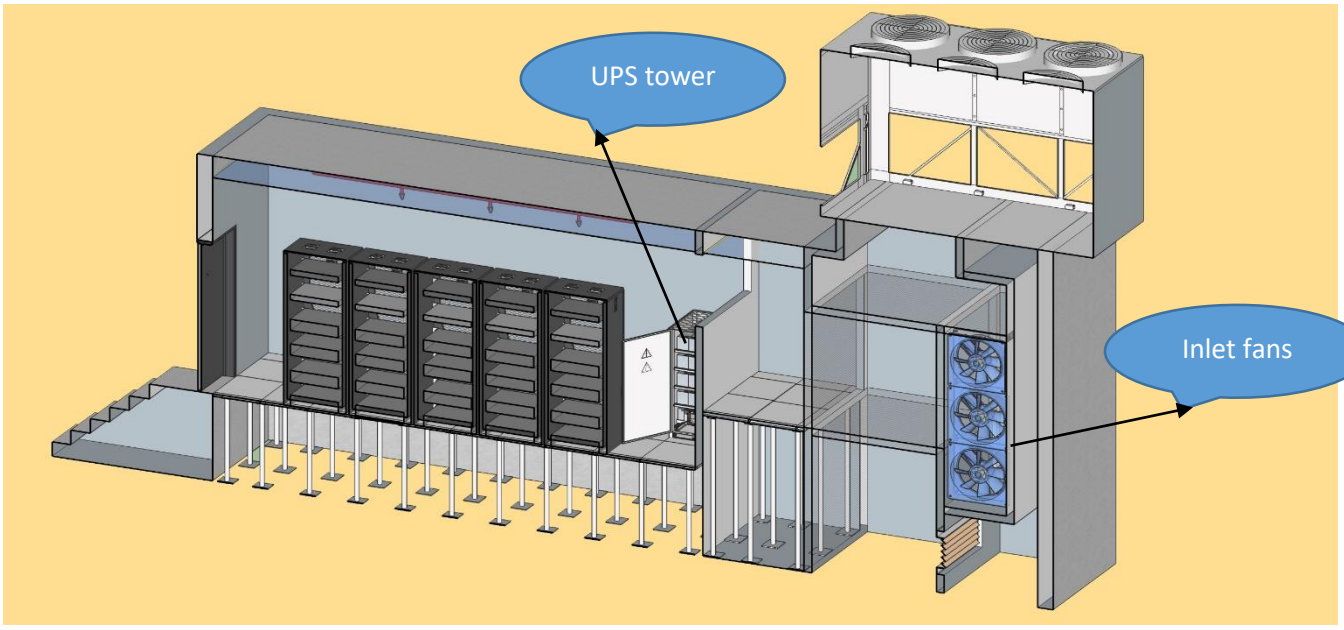


Figure 25 Side view of the datacenter

In this model three continuous fans inject cold air toward the data center via inlet louver by a raised floor. After inlet fans, three components namely as a humidification unit, a fan and a chiller are placed in order to make sure that all parameters which correspond to air are remained steady. After that, cold air will blow by fans within the server through the IT units toward the hot aisle

The size of the mentioned data center is 30 m² and it is designed according to the Oriented strand board (OSB). The temperature of the room that circumambient the test-bed is held at twenty-two °C in order to dispel any possible climate effects.

The model that used at this master thesis depends on genuine physical specifications of the system with certain simplifications and corrections for unknown variables. In this model, we need to handle two main components at the Data center in parallel, which is BESS (Battery Energy Storage System) and HVAC (Heating Ventilation Air Conditioning). The thermal model which correspond the HVAC section of the model mainly concentrated on power consumed by these devices and Temperature the air that generated by these devices. Server racks which considered in this model are sieged by cold and hot aisle. Two-aisle are structured in an identical way to each other from the thermodynamic point of view. The following figure is used in order to properly represent the schematic of Data center model.

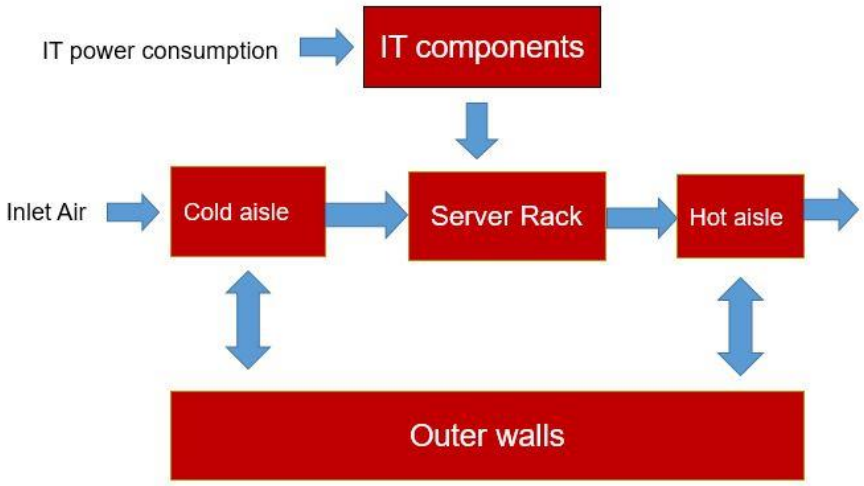


Figure 26 Data center test-bed heat flow

At this model, the heat flow which by servers that exists in the racks is planned according to the average thermal capacity and the heat transfer coefficient related to the servers and it applied for thermal energy propagated to the inlet air in every server rack. The model consists of the power supply, the thermal mass related to the IT components and the amount of air which resists inner the racks as well as connecting points in order to flow the air in and out the datacentre.

At the considered model for this master thesis, we assumed that the room that we are dealing with is the perfect virtual room which means all sections of the room namely: walls, celling and floors are built of similar materials that have identical thermal conductivity and they have same thickness. Furthermore, we assumed that energy transform rate is constant which indicate that when AC consumed on unit of energy, energy that perfused into the room is steady.

We assumed T as the indoor temperature and T_0 as the temperature which exists outside the considered room. Let's consider Q as the heat transfer rate that comes from outdoor to the room, then we consider K as the thermal conductivity of the material and the other import factor is A that corresponds to the total area of the room as well as L which related to the thickness of the

material. Before going into detail for computing heat transfer, we need to have the clear idea about the meaning of thermal conductivity and air mass, how they measured and so on.

The mass flow rate refers to the mass of the substance which passes through the pipe per unit of time. The unit which mainly uses for mass flow rate is kilogram per second in SI unit. The symbol of mass flow rate is ṁ and the dot that used at this symbol refers to Newton's notation. On account of that mass has the scalar quantity, the mass flow rate also has the scalar quantity. The adjustment of mass considers as the amount that flows after crossing the boundary for some duration of time.

According to Fourier's law we have : $Q = \frac{kA}{L}(T_0^{\sim} - T)$. Fundamentally, this equation tells us that relative to the thermal conductivity of the material, the size of walls, the difference in the temperature and it is reciprocally proportional to the thickness of the material. In the case that we have the fixed room which has the constant values for material, size, and the thickness of the walls, the mentioned law described that the rate which related to the amount heat that transferred in the datacenter is proportional to the temperature variation among the outdoor and indoor temperature. In the case that we have bigger difference in terms of temperature, we need to consume more energy in the time unit to rectify the heat transferred from outdoors.

We considered P_e as the affected energy which is interjected toward the air of the datacenter environment every second, and assume the m be the air mass which exists in datacenter environment, as well as C to be the thermal capacity of the air from inside the datacenter environment. Specifically, C is the required energy for one KG of a particular material, which in this case we are dealing with the air, to increment one Celsius. the fluctuation rate of the temperature $\frac{dT}{dt}$ of the datacenter environment could be computed by using the following equation:

$$\frac{dT}{dt} = \frac{Q + P_e}{mC}$$

Table 3 General notation

Notation	Definition	Unit
L	Thickness of material	Centimeter
λ	Conductivity of room	$J/(s \cdot K)$
r	Energy transformation ratio of the air conditioner	------
A	Total area of six walls	m^2
m	Air mass of the room	Kg
K	Thermal conductivity of a material	$W/(K \cdot m)$

Q	Heat transfer rate from outdoor to the room	*J/s*
c	Specific air heat capacity	$J/(kg \cdot K)$
P	Electrical power associated to air conditioner	*J/s*

Let $\lambda = \frac{K*A}{L}$,which λ is considered as the thermal conductivity of datacenter environment. By combining the equation of Fourier low and previously mentioned equation we could compute the temperature change in the datacenter environment.

$$T_{dc}(t) = T_0^{\sim} + \frac{P_e}{\lambda} + T(0) - T_0^{\sim} - \frac{P_e}{\lambda})P_e e^{-\frac{\lambda}{mC}t}$$

We need to take into the consideration that the mentioned equation will exist only in the case that $T_0^{\sim}$ and P_e could be assumed to be sustained at the interval of time from 0 and t. But at the real world, $T_0^{\sim}$ is altered by the exterior temperature. Furthermore, an air conditioner regulate its immediate power P based on the interior temperature. Therefore, P_e also fluctuate over the time horizon. However, interior and exterior temperature that related to the room do not varies unexpectedly, as the result, for the short period of time we consider $T_0^{\sim}$ and P_e as a constant. Due to the fact that just a couple of walls at the datacenter environment are In the vicinity to the open air, $T_0^{\sim}$ is partly associated to T. We assumed the association among $T_0^{\sim}$ and T as linear function $T_0^{\sim} = b_0 + b_1 T$,where a_0 and a_1 are constant and $b_1 \in [0,1]$. If we read the indoor temperature and outdoor temperature every few minutes the mentioned equation could be modified to the subsequent equation:

$$T_{dc}[n+1] = b_0 + b_1 T_0[n] + \frac{r}{\lambda} p[n] + (T_{dc}[n] - b_0 - b_1 T_0[n] - \frac{r}{\lambda} p[n]) e^{-\frac{r}{\lambda} \triangle t[n]}$$

At the mentioned equation, n shows n^{th} iteration that was done. $\triangle t[n]$ represents among n^{th} iteration and $n+1^{th}$ iteration. $T_{dc}[n]$ is related to the interior temperature of the n^{th} iteration and $T_0[n]$ is related to the exterior temperature, as well as p[n] is associated to the immediate power at n^{th} iteration.

We observed that in previously mentioned equation T[n+1] is the linear function that consists of T[n], $T_0[n]$ and p[n] if we consider $e^{-\frac{r}{\lambda} \triangle t[n]}$ as the constant. We assumed λ to be the fixed value due to the fact that, this value is highly associated to the physical property of the materials. Additionally, we supposed mC as fixed value which defines by the size of the room, in our case the size of the room is $30m^2$.As the result, if we presumed all $\triangle t[n]$ will be fixed over the time horizon, we have the possibility to simplify the previously mentioned equation into the following ones:

$$T_{dc}[n+1] = U_i T_{dc}[n] + U_c P_{it} + U_o T_0[n] - U_p P[n]$$

At this equation U_i is the fixed value corresponds to the characteristics of the room, in order to compute it we need to consider air mass that represented by m and specific heat capacity of air which considered as 718 (*Joules per Kg-Celsius*) which represented by C , As well as this, we need to compute the conductivity of the room which shows by λ. Conductivity of room was computed with take into consideration of thermal conductivity of materials, as we already discussed, in this master thesis the material which considered was cement at it has thermal conductivity of 0.7 W/(m k), the other two factors that playing crucial role for conductivity of room are area of the datacenter which could easily compute by multiplying length, width and height of the room and the other import factor is thickness of walls which considered as 33 centimeters. U_c is a constant value which associated to linear function of $T_0^{\sim} = b_0 + b_1 T_0$. U_o is a fixed regarding to interior temperature that varies over the time based on the exterior temperature in degree Celsius. U_p is fixed for interior temperature change yielded by HVAC system over the time horizon. The relationship among (U_i, U_o, U_c, U_p) and (λ, r, b_0, b_1, Δt) are represented at consecutive equations:

$$U_i = e^{-\frac{\lambda * \Delta T}{mC}}$$

$$U_c = (1 - U_i) b_0$$

$$U_o = (1 - U_i) b_1$$

$$U_p = (1 - U_i)\ \frac{r}{\lambda}$$

Furthermore, we must take into the consideration the temperature boundary which defined by ASHRAE Thermal Guidance, according to this standard Datacenters that falls into the first category (Known as class A) must be in the temperature interval of 18 °C and 27 °C. In order to satisfy this boundary, we need to design the following constraint: $T_{min} \leq T_{dc}[n] \leq T_{max}$ as we could see from the following figure the generated results show that minimum generated result is 21.5 °C and maximum temperature which was generated was 26 °C.

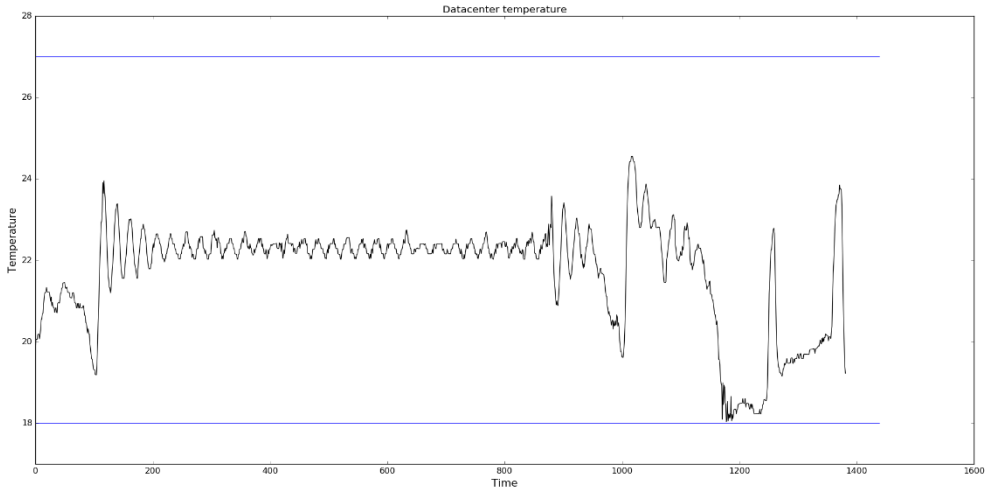


Figure 27 Temperature of datacenter

After we properly designed the temperature behaviour of datacentre, now we need to concentrate on the other important and one of the most power consuming sub-systems which exist in the Datacenter environment, and it is HVAC subsystem. This subsystem plays the crucial rule in the Datacenter environment. HVAC systems mainly devoted to ventilating the air in this environment and take the hot air which was generated by servers and transfer this air to the cooling compartment, then take the fresh and inject the cold air by passing through the raised floor to the server racks and so on and so forth. We need to take into consideration the fact that the temperature and the consumed power for cooling purpose proportional to each other it means that when the temperature rises also the consumed power is increasing.

In order to properly design the method to compute the consumed power we need to consider the following constraints:

$$P_{hvac\_min} \leq P_{hvac} \leq P_{hvac\_max}$$

$$P_{hvac} = P^{+}_{hvac} + P^{-}_{hvac}$$

$$P^{+}_{hvac} = \min(P_{hvac-up\_}, (P_{hvac-max} - P_{cooling}))$$

$$P^{-}_{hvac} = \max(P_{cooling}, (P_{hvac\_down\_}))$$

As we could see from mentioned constraint, the power will be consuming by HVAC compartment inside the datacenter must be inside the boundary with lower bound of P_{hvac_min} which defined in the optimizer as 0 KW and upper bound which represented by P_{hvac_max} was considered as 300 KW. After setting the constraint for bounding the consumed power by HVAC system, now we need to define constraints for the purpose of computing amount of power consumed by this sub-system of datacenter environment, the power is separated into two main part P^{+}_{hvac} and P^{-}_{hvac}. P^{-}_{hvac} is computed as the minimum consumed power among the subtraction of maximum consumed power which defined as the upper-bound for HVAC power consumption which represented by $P_{hvac-max}$ and $P_{cooling}$ which related to the dataset that taken from the simulation done by Dymola on the actual datacenter, As well as Power to be consumed from DSO for flexibility provision. P^{+}_{hvac} is computed as maximum power among the $P_{cooling}$ which related to the dataset that taken from the simulation done by Modelica which performed remotely on the actual datacenter that located at Berlin, Germany and the power to be provided to DSO for flexibility provision.

After the implementation of this section of optimizer have been done, we could obtain the result of optimizer which represented at the figure 28. as we could see from results the minimum power which consumed is during the minimum of DSO that refers to the off-peak time and maximum power that consumed power by this section of Datacentre peak of DSO signal.

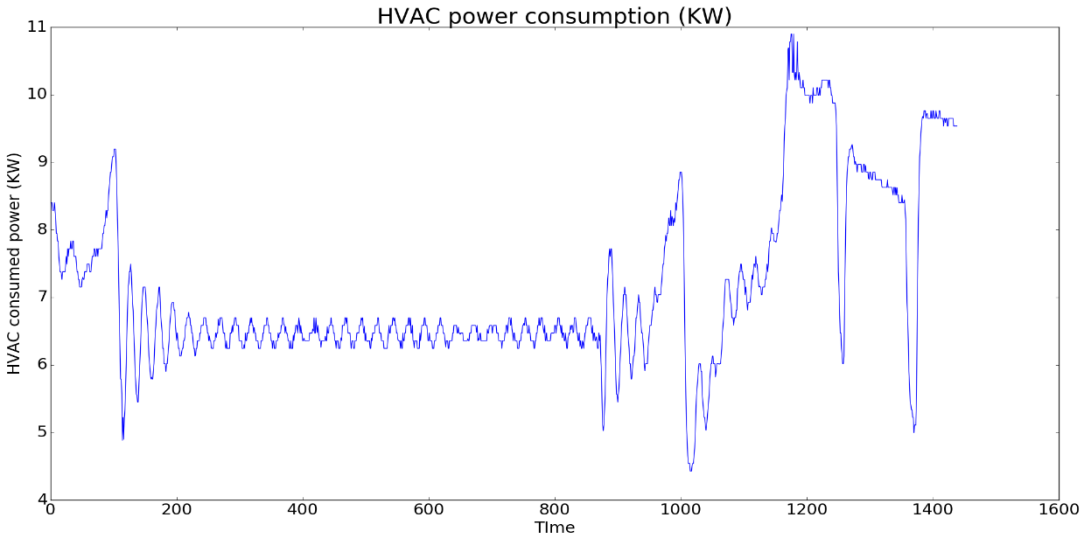


Figure 28 HVAC power consumption (KW)

If we compare the results of the datacenter temperature and HVAC power consumption, we could clearly observe that during the high period of the HVAC power consumption we have the lower temperature at the datacenter environment, although at the period that we dealing with the reduction in the HVAC power consumption the temperature which exists in the datacenter environment will be increase and it reaches to 24.5 which still below the allowable range that defined by ASHRAEE association.

After defining the model for power which consumed by HVAC subsystem that located in datacenter environment, now we need to design the model regarding the power which consuming/generating by BESS (Battery Energy Storage System). BESS compartment mainly uses as the backup power in the case power failure. In order to properly design the behavior of BESS we need to define the following constraints:

$$-p_{bess-rated} \leq p_{bess} \leq p_{bess-rated}$$

$$P_{BESS} = P^{+}_{BESS} + P^{-}_{BESS}$$

$$E^{t}_{BESS} = E^{t-1}_{BESS} + T_s\left(\frac{p^{+}_{out}}{\eta} + p^{-}_{out} * \eta\right)$$

$$E_{bess-min} \leq E_{BESS} \leq E_{bess-max}$$

As we could see from mentioned constraints, the design of BESS consists of two categories of power generated/consumed by this compartment and energy that stored/released by this section of Datacenter environment. At the beginning we concentrate on the power generated/consumed by BESS that known as P_{BESS}, this power must be bounded into the boundary of negative and positive rated power that defined as $P_{bess-rated}$ and it has the value of $\pm$15.462 KW ,if this power is less than lower bound or more than upper bound the gurobi solver reach to infeasibility. Furthermore, this power the summation generated(charging) power that defined as P^{+}_{BESS} and consumed(discharging) power P^{-}_{BESS}.

If we retain the state of the charge (SOC) of the BESS at the maximum (fully charged) or minimum (fully discharged) is not ideal due to the fact that it could cause to provide BESS futile in the situation that we need to have opposite regulation from the grid. Therefore, BESS is maintained in between the maximum and minimum SOC states.

After computing the power either consumed or generated by BESS, now we need to compute the Energy that stored in this compartment of Datacenter. The energy is measured as Kwh (Kilowatt hour). In order to properly design the method for computing the energy first and foremost we need to set the upper bound and lower bound that defined as $E_{bess-max}$ and $E_{bess-min}$, Furthermore, the value for the upper bound set to 30.462 KWh and lower upper is set to zero KWh. Then, we also need to consider the battery efficiency which in this thesis is considered as 0.85 or 85% and it has shown as η, this percentage of efficiency refers to LA (Lead-Acid) battery that inserted into the datacenter, After that, we compute the E_{BESS}^{t} that refers to energy of BESS at time step t and is the summation of energy stored at time t-1 with the summation of generated power and consumed power, the summation of powers need to be transferred into the energy and in order to do it we need to multiply them with the time step that we have.

In order to have better overview regarding parameters that used in the optimization, following table have been presented.

Table 4 Parameters used at optimizer

Parameters	Values
Minimum Temperature	18℃
Maximum Temperature	27℃
$\boldsymbol{P_{bess-rated}}$	15.462KW
$\boldsymbol{-P_{bess-rated}}$	-15.462 KW
A	320m
M	420 Kg
$\boldsymbol{P_{hvac-max}}$	300 KW
$\boldsymbol{P_{hvac-min}}$	0 KW
$\boldsymbol{E_{bess-max}}$	30.462 KWh
$\boldsymbol{E_{bess-min}}$	0 KWh
Minimum SOC	40%

Maximum SOC	80%
Coefficient of performance(COP)	3.4
L	33cm
$\boldsymbol{\eta}$	85%

SOC (State-Of-Charge) for the battery is computed by taking the Energy of BESS and divide it to the maximum energy that capable to store at the battery. The change in BESS SOC (State of Charge) is shown in figure 25 during the simulations. As we could see at the figure 29 SOC (State of Charge) regarding the battery remaining between 45% and 65% due to the fact that if the we put the battery at the maximum or minimum SOC it will be reducing the life of the battery, Therefore, we consider the BESS is maintained between at the boundary of 45 and 65 percent.

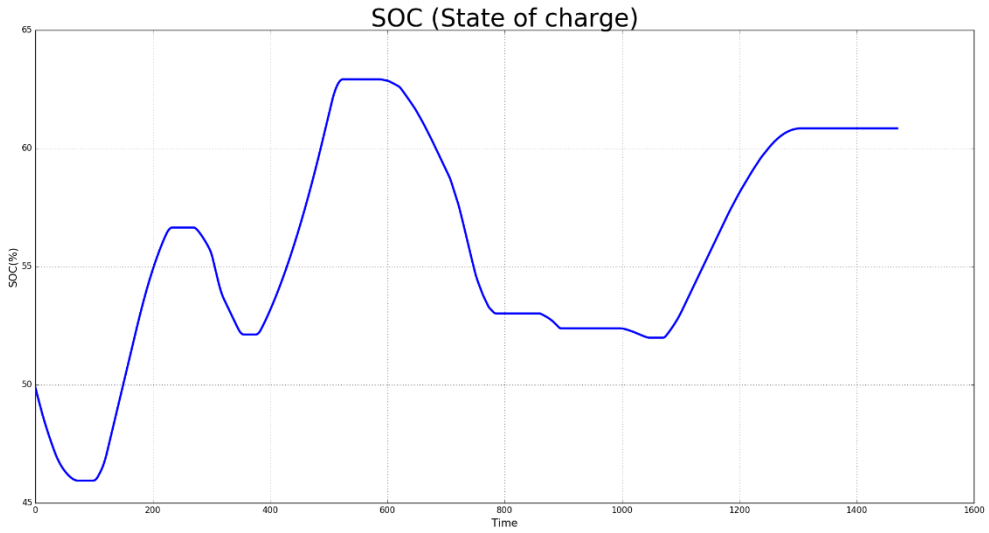


Figure 29 State of charge (SOC)

The figure 30 shows the power of Battery Energy Storage system (BESS), the activity that we have at the battery would provide the flexibility by charging and discharge which highly depends on the signal that arrived from the DSO and during the high power consumption the battery will be discharging and when we have the low power consumption, the battery will be charging.

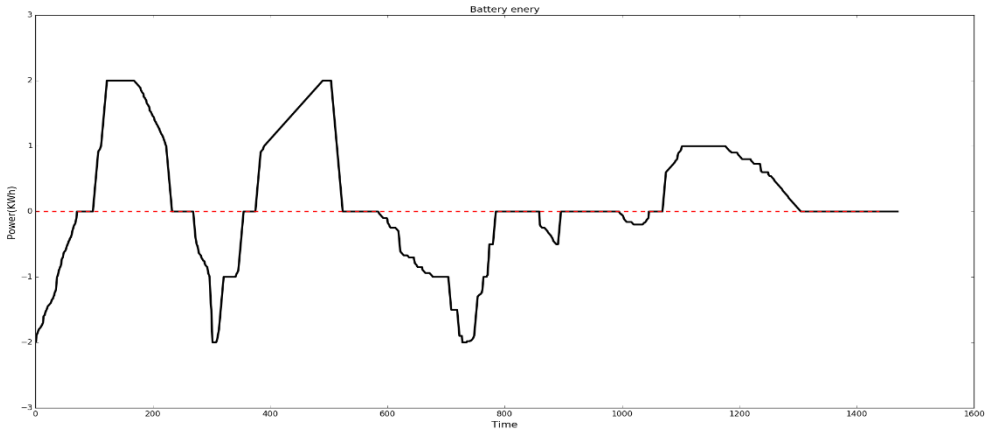


Figure 30 Power of BESS

The other vital factor that we need to consider besides consideration of HVAC power and BESS power, is the power which consumed due to IT load that passing through the servers. The IT workload in a datacenter is varying throughout the day, therefore the servers usually are not running at the optimal working point, Additionally, we need to take into consideration the power consumption is tightly associated to the workload on each server.

In order to compute the IT power load, we need to use the concept of deferrable load which described as following: Deferrable load refers to the type of load which is flexible over time, the deferrable load is using in many cases at the modern electricity market such as washer/dryer, air conditioner, etc. The real-time pricing attribute supplied by DR (Demand Response) program that enables the deferrable load scheduler in order to periodically receive real-time price information. Based on real-time price information the scheduler can make optimal decisions on power consumption when the price is too high.

In this master thesis, we considered the deferrable IT load that is working based on the general idea of the deferrable load as well as the signal which provided by the DSO. We used the following constraint $p_{it} = p_{undeferrable} + p_{deferrable}$, As we could see at this constraint we have two main part $p_{undeferrable}$ and $p_{deferrable}$. At the peak hours which the energy price is too high in order to reduce the cost we have the potential to defer this power to the time that energy price is at the mean price or below it, the amount of power that deferred to the next time step is denoted as $p_{deferrable}$ and the rest of the IT load that could remain is defined as $p_{undeferrable}$.

The following figure is representing IT load when the deferrable load is considered:

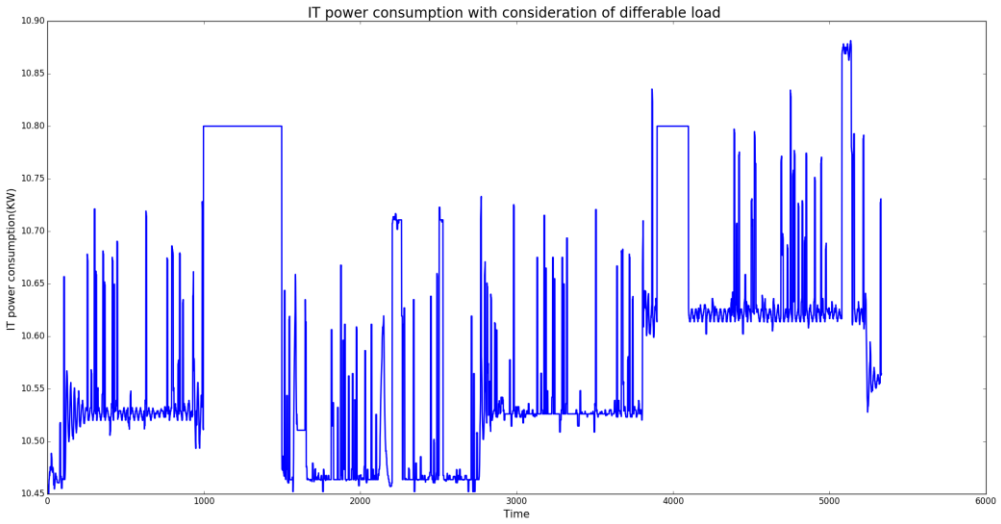


Figure 31 deferrable IT load

When all constraints that have been mentioned before are satisfied we need to introduce the objective function in order to optimize the energy that consumed by all sub-systems that exist in the datacentre environment. The objective that has been considered in this thesis is consists of power consumed by datacentre which defines as p_{dc} that is the summation of power consumed for cooling purpose and power of BESS as well as consumed power by IT which demonstrated at the following equation $P_{dc}^{t} = p_{cooling} + p_{Bess} + p_{IT}$. P_{dr}^{t} is the amount of power that Distribution System Operators (DSO) is able to generate. The figure 32 clearly represents the overall power that consumed inside the datacenter (P_{dc}^{t}) with considering all compartments which exists inside this environment. If we compare the overall consumed power inside the datacenter with the signal of demand-response (mentioned at appendix), we could clearly see that the general consumed power in the datacenter is below the power which came from DSO.

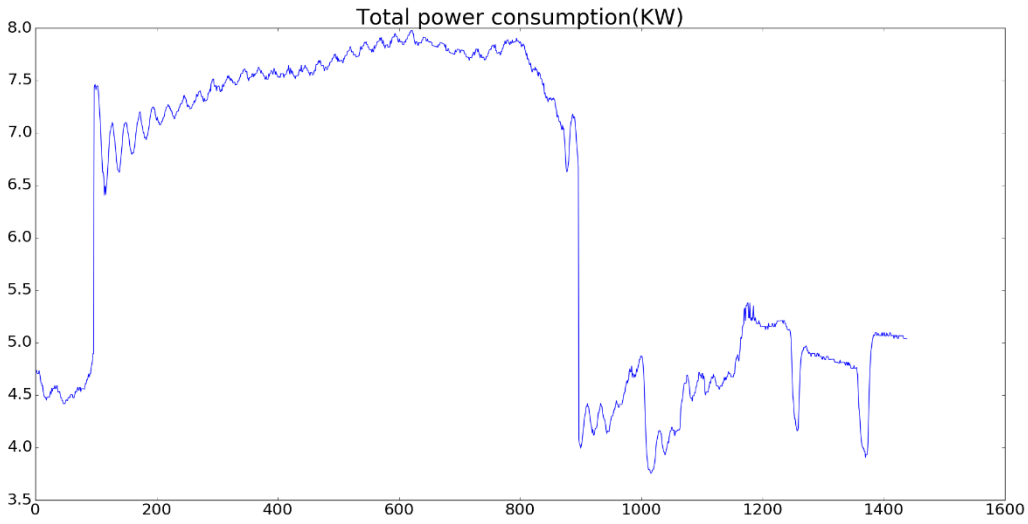


Figure 32 Total consumed power inside the data center

The last component which considered the objective function equation is predicted values of energy consumption of cooling compartment and BESS which denoted as $P^t_{dc\_base}$. In order to minimize the overall power which consumed in the datacenter environment, the linear objective function has been provided according to the following equation $\min(\sum_{t=0}^{1440}(P^t_{dr} - P^t_{dc} - P^t_{dc\_base}))$.

After the optimization was done the results shows that this optimization has been able to reduce 13.51 KW and the overall run time of Gurobi optimizer in python for each iteration is 0.048 second and we have 1440 iterations hence the computation takes 69.12 seconds equals to 1.152 minutes.

8. Conclusion and future work:

As the conclusion, the primary objective of this master thesis was to create the model for forecasting behavior of power which will consume in the future. The secondary objective was to design the optimizer in order to optimize the consumed power in the data center.

Towards these objectives, in order to satisfy the first objective of the thesis the time-series forecasting was designed by using ARIMA (Autoregressive Integrated Moving Average) which predicting the future behavior of power will be consumed for cooling purposes and the power will be used by servers in order to process IT tasks.

The adaptive time-series has been implemented by using ARIMA that is shown to accurately predict the power consumption by cooling compartment which exists inside the data center and the power which consumed by IT loads which passing through servers inside the data center. The time-series which used in this master thesis has been adaptively trained by existed dataset in order to predict the future behavior of data center. The adaptive training is beneficial as the time-series model no more requires to be trained on plenty of datasets. The forecasting performance enhances when increasing granularity of training data. The reason for that seems to be the availability of lagged inputs closer to the forecast horizon rather than using the higher order patterns exhibited by a smaller granularity. By properly setting the Time-series model, we could predict the future with the 98% accuracy.

Additionally, as this adaptive time-series has the adaptability to frequently learn even after development. Therefore, it has the capability of receiving the real-time data from the system, then if there is the need it will tuning itself. This capability is advantageous when the time-series forecasting is used in a real-time datacenter.

After satisfying the first objective of this master thesis, then so as to fulfil the second objective of this master thesis, the optimizer which has the input signal of predicted values that generated by time-series model and the DR signal which was provided by DSO company, designed by Python programming language and using the Gurobi solver for optimizing the power consumed power in the whole data center environment.

There are practical issues which need to be addressed before developing the methods which used in this thesis to the real-world applications. These include:

1. The trade-off between accuracy of the length and the complexity of model must be determined in order to maximize the economic viability.

2. Investigating the influence of the length and characteristics of the training data on the forecast results; More training data prevents overfitting and increasing accuracy, but at the same time consumption patterns may change during longer periods. The strategy for dealing with permanent behaviour changed by new devices must be developed.

As the future work, the experiment can be expanded by integrating the created optimizer with the real-time system that is capable to take the real-time signal and at the first step predict the

future behavior of the generated signal then use it as the input of the MPC (Model Predictive Control) that is using by the industry; Furthermore, at the MPC which is applicable to the industry we need also take into consideration the uncertainty disturbance that exists in the system. In this master thesis regarding the prediction part, we only used the autoregressive methods in order to predict the future behavior since the input signal is varying very few but if we have the input signal that changing very fast we need to have different methodologies such as SVM or Neural networks.

At the second part of this master thesis, the first element (power of HVAC and temperature of datacenter) which was used at the optimizer has many barriers. First, in the current model, we did not consider some elements like movements which generated by the human as well as heat which generated by electronic devices, hence; we have the potential to improve this model in the future and add mentioned factors to this model. Second, in the current model, we consider that capacity of the room and facilities are the constraints which need to be satisfied, In the future model, we have the possibility to consider further constraints like the distance among the rooms as the resulting personnel which working in the datacenter has adequate time to walk from one room to the other.

Furthermore, at this part, we did not consider the cost of energy that is widely varying over the time horizon, which could be added to the optimizer as the extra parameter in order to reduce the general power consumption. The other factor that could be added to the model in the future for the aim of energy reduction will be the real-time DR (Demand Response) signal which will be high impact on the energy shaving.

Additionally, regarding the second part of this master thesis at the section of computing the power of BESS we only considered the efficiency of this compartment as the fixed value, in the future work since the battery efficiency highly depends on auxiliary load, we could consider this factor in the computation of consumed power by BESS.

9. Bibliography

[1] Report to congress on server and data center energy efficiency, August 2007

[2] Telecommunications industry Association, TIA-912 Data center standards overview

[3] J. Koomey, “Growth in data center electricity use 2005 to 2010,” A report by Analytical Press, completed at the request of The New York Times, Aug. 2011.

[4] A. Banerjee, T. Mukherjee, G. Varsamopoulos, and S. K. Gupta, “Cooling-aware and thermal-aware workload placement for green HPC data centers.,” in Proceedings of Green Computing Conference 2010, pp. 245–256, Aug. 2010.

[5] Georgios Varsamopoulos, Ayan Banerjee, and Sandeep K.S. Gupta “Energy Efficiency of Thermal-Aware Job Scheduling Algorithms under Various Cooling Models”

[6] S. Khuller, J. Li, and B. Saha, “Energy efficient scheduling via partial shutdown,” in Proceedings of SODA 2010, pp. 1360–1372, Jan. 2010.

[7] Q. Tang, T. Mukherjee, S. K. Gupta, and P. Cayton, “Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters,” in Fourth International Conference on Intelligent Sensing and Information Processing, 2006., pp. 203–208, IEEE, 2006.

[8] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. Wang, “A cyber–physical systems approach to data center modeling and control for energy efficiency,” Proceedings of the IEEE, vol. 100, no. 1, pp. 254–268, 2012.

[9] S. Govindan, “Optimizing Power Delivery Cost In Datacenters,” *PennState Computer* Science and Engineering*,* 17 May 2011.

[10] Renewable Energy Policy Network for 21st Century, “The First Decade : 10 years of renewable energy progress,” REN21, 2015.

[11] Renewable 2016 Global status report, REN21, 2016.

[12] Telecommunications Industry Association, “TIA-942 Data center standards overview” ADC Krone

[13] E.on Energy Research center (GEYSER), ‘Energy consumption and production monitoring subsystem design’

[14] Data Center Monitoring and Analysis at LRZ

[15] Open data center Alliance Usage model: Carbon footprint and energy efficiency Rev 2.0 2013

[16] Seidl, H.; Noster, R.; Blank, S.. "Leistung steigern, Kosten senken: Energieeffizienz im Rechenzentrum”. Deutsche Energie-Agentur (DENA). (2012, February).

[17] Bundesministerium für Wirtschaft und Energie ,“Development of renewable energy sources in Germany 2014”, BMWI, 2015.

[18] K. Das, M. Litong-Palima, P. Maule und P. E. Sørensen, “Adequacy of operating reserves for power systems in future european wind power scenarios” *2015 IEEE Power & Energy Society General Meeting,* pp. 1-5, 2015.

[19] ASHRAE TC9.9 , “Data Center Networking Equipment – Issues and Best Practices Whitepaper”, prepared by ASHRAE Technical Committee (TC) 9.9

[20] Temperature Distribution Prediction in Data Centers for Decreasing Power Consumption by Machine Learning

[21] Zhihang Song, Bruce T. Murray, Bahgat Sammakia, “A dynamic compact thermal model for data center analysis and control using the zonal method and artificial neural network”

[22] Y. J. Emad Samadiani and F. Mistree, “The Thermal Design of a Next Generation Data Center: A Conceptual Exposition,” Journal of Electronic Packaging, November 2008.

[23] Mitchell, T. (1997). Machine Learning. McGraw Hill. p. 2. ISBN 0-07-042807-7.

[24] Gooijer, J.G.D., Hyndman, R.J.: 25 years of time series forecasting. International Journal of Forecasting22(3), 443–473 (2006)

[25] Shinya Tashiro, Yuya Tarutani, Go Hasegawa, Yutaka Nakamura, Kazuhiro Matsuda and Morito Matsuoka, “ A Network Model for Prediction of Temperature Distribution in Data Centers ”

[26] S. Karatasou, M. Santamouris, and V. Geros,“Prediction of energy consumption in buildings with artificial intelligent techniques and chaos time series analysis”

[27] Abdullatif E.Ben-Nakhil,Mohammed A.Mahmoud “Cooling load prediction for buildings using general regression neural networks”

[28] Guy R.Newsham,Benjamin J.Birt , “Building-level occupancy Data to improve ARIMA-based Electricty forecasts”

[29] A.A.EI Desouky ,M.M.EIKateb, “Hybrid adaptive techniques for electric-load forecast using ANN and ARIMA ”

[30] Zhenhua Liu, Adam Wierman, Yuan Chen, Benjamin Razon, Niangjun Chen, “Data Center Demand Response: Avoiding the Coincident Peak via Workload Shifting and Local Generation”

[31] Young M.Lee,Raya Horesh,Leo Liberti, “Optimal HVAC control as demand response with on-site energy storage and generation system”

[32] Neda Nasiriani, George Kesidis, Di Wang, “Optimal Peak Shaving Using Batteries at Datacenters: Characterizing the Risks and Benefits”

[33] Yuanyuan Shi, Bolun Xu, Di Wang, Baosen Zhang, “Using Battery Storage for Peak Shaving and Frequency Regulation: Joint Optimization for Superlinear Gains”

[34] M. Schaub, B. Zielke and M. Kriegel, “KLIMAZELLE Zur indirekten freien Kühlung eines Rechenzentrums“. Hermann-Rietschel-Institut, Technical University of Berlin. December 2015.

10. Appendexes

1. Code regarding the computing MSE (Mean Square Error), RMSE (Root Mean Absolute Error) and MAPE (Mean Absolute Percentage Error)

```
mse = np.mean((yhat - obs)**2)
def rmse(predictions, targets):
    return np.sqrt(((predictions - targets) ** 2).mean(axis=None))

#--------
#MAPE
def mean_absolute_percentage_error(y_true, y_pred):
    y_true, y_pred = np.array(y_true), np.array(y_pred)
    return np.mean(np.abs((y_true - y_pred) / y_true)) * 100
rmse_val = rmse(obs,yhat)
MAPE=mean_absolute_percentage_error(obs,yhat)
```

2. BIC test

```
# evaluate an ARIMA model for a given order (p,d,q)
def evaluate_arima_model(X, arima_order):
    # prepare training dataset
    train_size = int(len(X) * 0.66)
    train, test = X[0:train_size], X[train_size:]
    history = [x for x in train]
    # make predictions
    predictions = list()
    for t in range(len(test)):
        model = ARIMA(history, order=arima_order)
        model_fit = model.fit(disp=0)
        yhat = model_fit.forecast()[0]
        predictions.append(yhat)
        history.append(test[t])
    # calculate out of sample error
    error = mean_squared_error(test, predictions)
    return error

# evaluate combinations of p, d and q values for an ARIMA model
def evaluate_models(dataset, p_values, d_values, q_values):
    dataset = dataset.astype('float32')
    best_score, best_cfg = float("inf"), None
    for p in p_values:
        for d in d_values:
            for q in q_values:
                order = (p, d, q)
                try:
                    mse = evaluate_arima_model(dataset, order)
                    if mse < best_score:
                        best_score, best_cfg = mse, order
                    print('ARIMA%s MSE=%.3f' % (order, mse))
                except:
                    continue
    print('Best ARIMA%s MSE=%.3f' % (best_cfg, best_score))

# load dataset


series = read_csv('IIpower.csv')
# evaluate parameters
p_values = [0, 1, 2, 4, 6]
d_values = range(0, 3)
q_values = range(0, 3)
warnings.filterwarnings("ignore")
evaluate_models(series.values, p_values, d_values, q_values)

#print  model., model.hqic


residuals.plot()
plt.show()
```

3. ARIMA model

```
#ARIMA model
X = df.values
from pandas import Series
from statsmodels.tsa.stattools import adfuller
series = pd.read_csv('ITpower.csv', header=0)
X = series.values

size = int(len(X) *0.1)
train, test = X[0:3800], X[3800:5334]
history = [x for x in train]
predictions = list()
for t in range(10):
    model = ARIMA(history, order=(5,1,3))
    model_fit = model.fit(disp=0)
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    obs = test[t]
    history.append(obs)
    print('predicted=%f, expected=%f' % (yhat, obs))
```

4. Consumed power by HVAC compartment

```
ConHVAC2 = m.addConstr(Phvac_flex[i] == (Phvac_down[i]+Phvac_up[i]), name='ConHVAC1')
ConHVAC3 = m.addConstr(0 <= Phvac_flex[i], name='ConHVAC2')
ConHVAC4 = m.addConstr(Phvac_flex[i] <= PcoolingRated, name='ConHVAC3')
CONHVAC1 = m.addConstr(Phvac_up[i] == minimum((Phvac_up_[i]), (PcoolingRated-float(newList2[i]))))

if float((DR[i])) > 10:

    CON3=m.addConstr(Phvac_down_[i] == 0, name='CON3')
    CON4=m.addConstr(Phvac_up_[i] == float((DR[i])),name='CON4')
    CON5=m.addConstr(Pbess_out_[i] == float((DR[i])),name='CON5')
    CON6=m.addConstr(Pbess_in_[i] == float((DR[i])),name='CON6')
elif float((DR[i])) < 10:
    CON7=m.addConstr(Phvac_down_[i] == float((DR[i])),name='CON7')
    CON8=m.addConstr(Phvac_up_[i] == 0,name='CON8')

    CON9=m.addConstr(Pbess_out_[i] == float((DR[i])),name='CON9')
    CON10=m.addConstr(Pbess_in_[i] == 0,name='CON10')
else:
    CON11=m.addConstr(Phvac_down_[i] == 0,name='CON11')
    CON12=m.addConstr(Phvac_up_[i] == 0,name='CON12')
    CON14=m.addConstr(Pbess_out_[i] == 0)
    CON15=m.addConstr(Pbess_in_[i] == 0,name='CON15')

ConHVAC6 = m.addConstr(Phvac_down[i]== max((-newList2[i])),Phvac_down_[i]))
```

5. Computation of Temperature inside the data center

```
ConTemp1 = m.addConstr(18 <= Tdc_i[i] <= 27, name='ConTemp1')
ConTemp2 = m.addConstr(Tdc_i[i]==((ki*Tdc_i[i-1])+(ko*Tamb)+(kc*((float(ITload[i])))-(((float(newList2[i])+(Phvac_flex[i]))*kp)*COPhvac)))
```

6. Computation of consumed power for charging and discharging of BESS as well as stored energy inside the BESS

```
ConBESS1= m.addConstr(Pbess_flex[i]==Pbess_out[i]+Pbess_in[i])
ConBESS2 = m.addConstr(0<=Pbess_in[i]<=(max(Pbess_rated,Pbess_in_[i])) #positive

ConBESS3 = m.addConstr(min(Pbess_ratedN,Pbess_out_[i])<=Pbess_out[i]<=0) #negative
ConEbess1= m.addConstr(0<=Ebess_i[i]<=Ebess_max)
iter += 1
ConEbess2 = m.addConstr(Ebess_i[i]== Ebess_i[i-1]+((Pbess_in[i]*(1*0.85/60)) + (Pbess_out[i] * (1/0.85/60))))
```

7. Plotting the results

```
#plot results

plt.plot([Temp[t] for t in range(1,1440)], 'g-')
plt.title('Data center Temprature (Centigrad)')
plt.show()

plt.plot([(EB[t]/Ebess_max) for t in range(1,1440)], 'b-')
plt.title('Ebess (KWh) ')
plt.show()

plt.plot([(Pbess_Result[t]) for t in range(1,1440)], 'b-')
plt.show()

plt.plot([Phvac_result[t] for t in range(1,1440)], 'b-')
plt.title('HVAC power consumption (KW)')
plt.xlabel('TIme')
plt.ylabel('HVAC consumed power (KW)')

plt.show()
plt.plot([PDC[t] for t in range(1,1440)], 'b-')
plt.title('Total power consumption(KW)')
plt.show()
plt.plot([(PDC_BASE[t]) for t in range(1,1440)], 'b-')
plt.title('Base power(KW)')
plt.show()
plt.plot([DR[t] for t in range(1,1440)], 'b-')
plt.title('DR')
plt.show()
```

8. Demand-response signal