REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

A STOCHASTIC MODEL FOR THE ENERGY-EFFICIENT SERVER CONSOLIDATION

Özge ERDEN

MASTER OF SCIENCE THESIS

Department of Industrial Engineering

Industrial Engineering Program

Supervisor

Assoc. Prof. Dr. Nezir AYDIN

June, 2022

REPUBLIC OF TÜRKİYE

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

A STOCHASTIC MODEL FOR THE ENERGY-EFFICIENT SERVER CONSOLIDATION

A thesis submitted by Özge ERDEN in partial fulfillment of the requirements for the degree of **Master of Science Thesis** is approved by the committee on 30/06/2022 in the Department of Industrial Engineering, Industrial Engineering Program.

Assoc Prof. Dr. Nezir AYDIN
Yıldız Technical University
Supervisor

Approved By the Examining Committee

Assoc Prof. Dr. Nezir AYDIN, Supervisor
Yildiz Technical University ____________________

Prof. Dr. Ali Fuat GÜNERİ, Member
Yildiz Technical University ____________________

Assoc Prof. Dr. Berk AYVAZ, Member
Istanbul Ticaret University ____________________

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of A Stochastic Model for The Energy-Efficient Server Consolidation supervised by my supervisor, Assoc. Prof. Dr. Nezir AYDIN. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Özge ERDEN

Signature

Dedicated to my family

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Nezir Aydın for their support and valuable opinions during the thesis.

I would like to thank my dear friend Ebru Geçici for her comments on the thesis to rectify it.

I would like to thank Mustafa Kemal Bulut for all his support and encouragement when starting the master's program.

Last but not least, I would like to express my endless gratitude to my mother Hayriye Erden, my father Ali Erden, and my brother Kadir Erden, who always support and show patience.

Özge ERDEN

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BRSA	Banking Regulation and Supervision Agency
CBRT	Central Bank of The Republic of Turkey
CPU	Central Process Unit
EFT	Electronic Funds Transfer
Hz	Hertz
IEC	International Electrotechnical Commission
IP	Internet Protocol
ISO	International Organization for Standardization
ISP	Internet Service Provider
IT	Information Technologies
ITIL	Information Technologies Infrastructure
PM	Physical Server
SAN	Storage Area Network
SLA	Service Level Agreement
TL	Turkish Liras
VM	Virtual Server

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

A Stochastic Model for the Energy-Efficient Server Consolidation

Özge ERDEN

Department of Industrial Engineering

Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Nezir AYDIN

Data centers form the infrastructure of information technologies. Most of the Internet Protocol addresses traffic passes through these data centers. Moreover, Internet Protocol addresses' related data is stored at these data centers. Many processing and storage servers are needed to cope with increasing flow. However, data centers have significant investment costs and high energy use. The most critical data center cost factors are investment and energy cost. Here, many factors and different resource types like CPU, bandwidth, and memory are considered for energy usage in servers. The CPU, on the other hand, is generally the most energy-consuming part of servers. Virtualization is the most critical technology to save energy in physical servers. Virtual servers are installed on physical servers with virtualization technology. Thus, server consolidation is achieved. This study aims to meet virtual servers' resource needs by minimizing costs and considering energy efficiency in a data center. In other words, here, we mention how much costs physical servers consume according to the energy cost and investment cost. It is assumed that a linear relationship between CPU and energy usage for workloads of the services running on virtual servers change in time. For this purpose, we proposed a stochastic model for server consolidation under resource demand uncertainty. This model, which presents the ambiguous workload, is a scenario-based stochastic programming model. According to the result, annual costs are minimized in the consolidation of virtual servers under uncertain resource demand.

Keywords: Server consolidation, stochastic optimization, modeling, data center.

ÖZET

Enerji Verimli Sunucu Konsolidasyonu için Stokastik Bir Model

Özge ERDEN

Endüstri Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Doç. Dr. Nezir AYDIN

Danışman

Veri merkezleri bilgi teknolojilerinin altyapısını oluşturur. İnternet Protokol trafiği bu veri merkezlerinden geçer. Ayrıca İnternet Protokol adreslerine ait veriler de bu tür veri merkezlerinde saklanmaktadır. Artan akışla başa çıkmak için birçok işleme ve depolama sunucusuna ihtiyaç vardır. Ancak veri merkezlerinin önemli yatırım maliyetleri ve yüksek enerji kullanımı vardır. En kritik veri merkezi maliyet faktörleri yatırım ve enerji maliyetidir. Sunucularda enerji kullanımı için CPU, bant genişliği ve bellek gibi birçok faktör ve farklı kaynak türleri göz önünde bulundurulur. CPU ise genellikle sunucularda en çok enerji tüketimi ile ilişkili kısımdır. Sanallaştırma, fiziksel sunucularda enerji tasarrufu için en kritik teknolojidir. Sanallaştırma teknolojisi ile fiziksel sunucular üzerinde sanal sunucular kurulur. Böylece sunucu konsolidasyonu sağlanır. Bu çalışma, bir veri merkezinde maliyetleri en aza indirerek ve enerji verimliliğini göz önünde bulundurarak sanal sunucuların kaynak ihtiyaçlarını karşılamayı amaçlamaktadır. Yani fiziksel sunucuların enerji maliyetine ve yatırım maliyetine göre ne kadar maliyet tükettiği göz önünde bulundurulmuştur. Sanal sunucular üzerinde çalışan servislerin iş yüklerinin zaman içindeki değişimi için CPU ile enerji kullanımı arasında doğrusal bir ilişki olduğu varsayılmaktadır. Bu amaçla, kaynak talebi belirsizliği altında sunucu konsolidasyonu için stokastik bir model önerilmiştir. Belirsiz iş yükünü sunan bu model, senaryo tabanlı bir stokastik programlama modelidir. Sonuca göre, belirsiz kaynak talebi altında sanal sunucuların konsolidasyonunda yıllık maliyetler minimize edilmektedir.

Anahtar Kelimeler: Sunucu konsolidasyonu, stokastik optimizasyon, modelleme, veri merkezi

1 INTRODUCTION

Data centers form a reliable and scalable IT infrastructure. They perform many different tasks and functions. Security services include packet filtering and inspection, intrusion detection, and intrusion prevention systems. Storage services that perform SAN architecture, backup, and archival are in data centers. It also provides business continuance with SAN extension, site selection, and data center interconnectivity [1]. Most of the IP address traffic passes through these data centers.

Moreover, the IP address data is also stored at data centers. For this reason, data centers are essential elements of the information technology industry [2]. Many processing and storage servers are needed to cope with this increasing flow in data centers. This need is increasing day by day [3].

There is a need for a rapid response to the high demand for information technologies. For this reason, management of data centers is an important and challenging issue. Data centers are typically provisioned for peak workloads. However, it runs well below capacity most of the time. Data centers have significant investment costs and high energy use [4]. For this reason, capacity planning of data centers is significant.

Furthermore, data centers consist of physical servers, storage, and air conditioning tools. Physical servers are components in data centers where IP address traffic passes. Investment costs and energy costs are primarily associated with physical servers. Capacity management in physical servers is the most significant factor in order not to create new investment costs and to reduce energy use. One of the techniques used in capacity management is server consolidation. It is a practical and widely used approach to reduce the total energy consumption in data centers [5]. Capacity management in physical servers is also essential for compliance with ITIL ISO / IEC standards. The reliability of overloaded and under-loaded physical servers is low. Server consolidation can be done in various ways and considers various parameters and effects.

Physical servers are multidimensional resources. Resources such as CPU, memory, network bandwidth, and disk size are considered in server consolidation. Chase et al. aim to consider CPU parameters with negotiated SLAs [6]. Pinheiro et al. have achieved power consumption with load balancing considering CPU, disk, and network parameters [7]. Different methods can be used to measure energy use. Studies show that energy usage is directly related to CPU usage [8]. Many studies show that CPU usage and energy usage are almost linear [9], [10]. Therefore, reducing CPU usage increases energy efficiency, and this minimizes data center costs.

Real-life problems do not always fit deterministic models. Stochastic programming is used in real-life optimization problems involving uncertainty. Stochastic programming deals with solving decision problems under uncertainty [11], [12]. The problem parameters are considered as random or stochastic variables. The effect of uncertainty on strategic decisions can be explained by the two-stage stochastic programming model. [13],[14]. In this thesis, a real-life problem will be modeled with stochastic programming.

The rest of the thesis is planned as follows. In this section, we briefly reviewed the relevant literature related to the work done in this study. The objective of the study and the hypothesis is explained. Section 2 explains server consolidation. In Chapter 3, the methodology is explained, and the stochastic mathematical model is proposed. In Section 4, information about the application is given, and the results are explained. Section 5 presents the conclusion.

1.1 Literature Review

In the literature, there are two main goals tried to be achieved for data centers. These goals are overload avoidance and energy efficiency. To avoid overloading, the requested resources of the virtual servers should not exceed the available resources of the physical server. Otherwise, the physical server will be overloaded, resulting in SLA violations. For energy efficiency, the number of active physical servers should be minimized. Therefore, insufficient use of physical servers should be avoided. Many studies have been carried out in the literature to achieve these two primary goals. In the studies (Table 1.1), server consolidation problems and related virtual server allocation and load balancing problems were solved with different methods.

Table 1.1 Classification of the studies according to the method, problem and purpose

Reference	Year	Method	Problem	Purpose
[15]	2022	Consolidated index for CPU- server saturation	Server consolidation	Increase energy efficiency
[16]	2022	Locust-Inspired Algorithm	Server consolidation	Reduce energy consumption
[17]	2022	Consolidation-aware migration avoidance job scheduling	Server consolidation	Reducing VM migration
[18]	2021	Virtualized task scheduling algorithm	Server consolidation	Enhance power efficiency
[19]	2021	Stochastic bin packing programming	Server consolidation	Reduce energy consumption
[20]	2019	Server consolidation energy saving algorithm	Server consolidation	Reduce energy consumption
[21]	2018	Power-aware VM placement	Bin packing problem	Minimize power consumption
[22]	2017	Snapshot-based solution	Server consolidation	Reduce energy consumption
[23]	2017	DNA based scheduling strategy	Server consolidation	Reduce energy consumption
[24]	2017	Virtual resource scheduling method	VM allocation	Reduce energy consumption
[25]	2017	Multiple-objective optimization approach	Load balancing	Balancing resource load and VMs between PMs
[26]	2013	Power aware Genetic Algorithm	VM allocation	Reduce energy consumption
[27]	2013	Multi-dimensional online bin packing heuristics	VM allocation	Minimize total energy cost and meeting the specified client-level SLAs

Table 1.1 Classification of the studies according to the method, problem and purpose (continued)

Reference	Year	Method	Problem	Purpose
[28]	2013	CPU reallocation algorithm	Energy management efficiency	Reduce energy consumption
[29]	2012	Heuristics for dynamic reallocation of VMs	VM allocation	Reduce energy consumption
[30]	2012	Mixed integer linear programming	VM allocation	Reduce energy consumption
[32]	2012	Heuristics for assign task	Task consolidation	Reduce energy consumption
[33]	2011	Control Theory	Dynamic resource allocation	Maximize the revenue while minimizing energy cost
[34]	2010	Heuristics for dynamic reallocation of VMs	VM allocation	Reduce energy consumption
[35]	2009	Min-max and shares aware allocation problem	VM allocation	Reduce energy consumption
[36]	2009	Queuing theoretic model	Power allocation	Maximum performance by power allocation
[37]	2009	Queuing theoretic model	Job scheduling	Maximize the resource profit by minimizing cost
[38]	2008	Multidimensional bin packing problem	Workload consolidation	Reduce energy consumption
[39]	2008	Control Theory	Power management	Creating a power management solution that coordinates individual approaches
[40]	2008	Dynamic resource allocation	Power management	Power consumption and SLA violation.
[41]	2007	Resource provisioning and allocation algorithm	Virtual power management	Reduce energy consumption

The stochastic programming model created in this thesis solves the server consolidation problem. In other words, an energy-efficient model has been established to minimize the cost. Unlike the studies in the literature, the investment cost is also considered in addition to the energy cost.

1.2 Objective of the Thesis

The thesis aims to meet the resource needs of virtual servers by minimizing annual cost. By creating a stochastic model based on uncertain resource demand, the cost of the data center is minimized.

There are two significant factors affecting data center cost. These are energy costs and investment costs. The investment cost is related to the purchase cost of the new physical server. We can divide the energy cost into fixed cost and variable cost. Fixed cost is calculated by considering the base energy that a physical server will use when it is open, even if no virtual server is running. Cores are active if at least one workload is performed. On the other hand, variable cost is calculated by considering the amount of CPU used by the virtual server running on physical servers.

The resource constraints in the problem are the CPU capacity of physical servers. The resource needs of virtual servers are determined by their workloads.

1.3 Hypothesis

Uncertainties are not taken into account in traditional deterministic models. Therefore, they cannot accurately reflect the dynamic of real-life practices[42]. Due to the stochastic nature of workloads, resource requests from virtual servers follow a stochastic process [22]. The ambiguous workload is created as a scenario-based stochastic programming model.

The stochastic model decides whether there is a need to purchase a physical server in the first stage as a decision variable. In the second stage, virtual servers are assigned over physical servers under uncertain scenarios.

2
SERVER CONSOLIDATION

Server consolidation is an approach that minimizes the total number of physical servers by using virtual servers. It provides a balanced distribution of workload running on physical servers. It ensures efficient use of capacity in physical servers, minimizing the number of physical servers and minimizing costs by providing energy efficiency [43]. This technique makes better use of the processing capabilities of existing physical servers. For this, newly configured virtual servers are created on physical servers. This structure reduces the investment cost of the new physical server and provides lower energy consumption.

2.1 Data Center and The Importance of Data Centers

Data centers are systems that contain equipment used for data processing, data storage, and communication networks. Data centers are essential in almost all sectors where information technologies are used. Its importance is increasing daily due to the need for data processing, data storage, and communication networks.

The demand for data centers is very high, and the responsiveness and performance must be very high. Data centers carry and respond to enormous demand load. [44].

All devices in data centers interact with each other. The network structure determines the way the data center works. Figure 2.1 shows an example of a bank's network topology. Requests from internet service providers first come on hardware called switches. Switches are the hardware that provides communication between computers. Switches increase processing speed and ensure that all connected devices operate at the same processing speed. There are integrations between banks and other institutions to make transactions quickly and through a single system. For example, the bank must be integrated into the CBRT EFT line to carry out EFT transactions. The hardware that enables this integration to be done securely is routers. Routers also work connected to switches. All requests to go outside the routers go through security hardware called a firewall. This hardware provides cyber security in organizations. Before requests go to physical servers, they pass over hardware called backbones that provide load balancing. Physical servers are also connected to storage via switches. The storage task is done on this hardware.

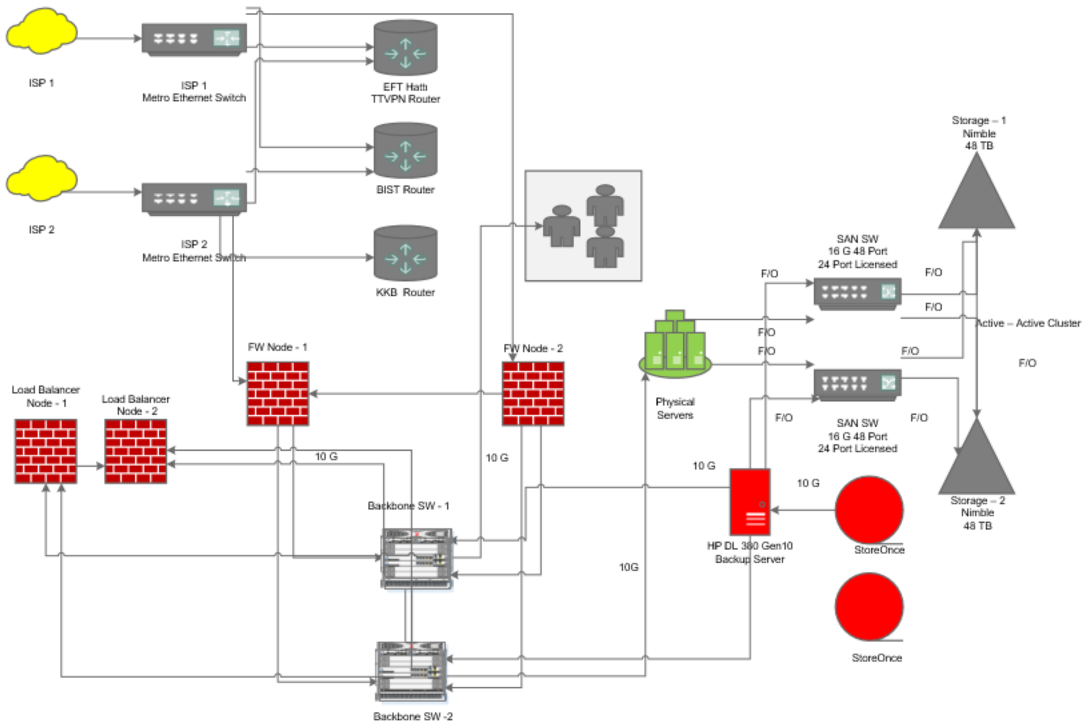

Figure 2.1 An example of a bank data center network topology

Data centers are divided into two ways of use, internal data centers and shared data centers.

A private facility is needed for the internal data center. Data center location selection is a critical issue. The location should be chosen considering issues such as power supply and security. The purchase and placement of all equipment belonging to the data center are made by considering the internal needs. In addition, electricity, cooling systems, fire extinguishing, and alarm systems are also important issues. The layout of all these hardware and systems is also an important issue. Figure 2.2 shows an example of an internal data center. Internal data centers are the type of use in which the operation processes are completed internally. Data center management is done internally. One of the essential advantages of the internal data center is confidentiality. The data is not shared with third parties. Another significant advantage is the direct connection to a data center located within the company. It reduces the delay time and provides a quicker response to the physical problems that may occur. The disadvantage is high costs. Internal data centers require a high investment cost and additional workforce.

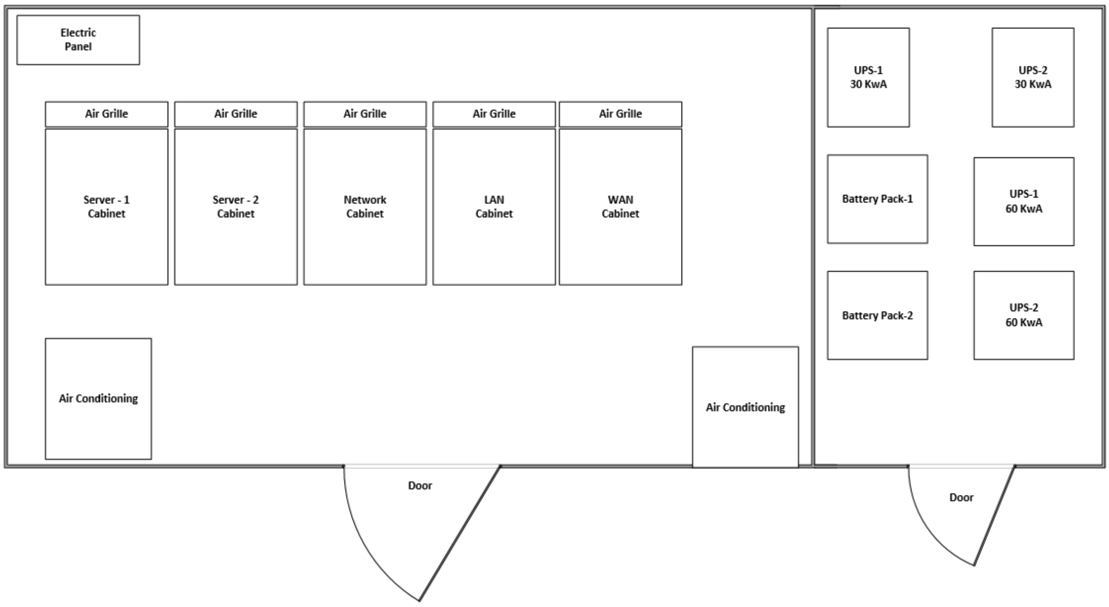

Figure 2.2 An example of an internal data center layout

In shared data centers, users rent racks and share a common data center with many other users. Users do not have any hardware. Infrastructure services, operational processes, hardware, and data center management belong to the company providing data center services. The most important advantage is cost. A cost is incurred according to the leased area and usage capacity. Outsourcing is high, but the need for labor is low. As the distance to the data center increases, it can cause delays in communication. Too much data center traffic may affect other users who share this data center.

2.1.1 Data Centers in the Banking Sector

The banking sector has an essential share in our country. The BRSA makes legal regulations pertaining to banks. Banks' data centers must comply with the Bank's Regulation on Information Systems and Electronic Banking Services. According to this regulation, banks cannot keep customer data in shared data centers. For this reason, the use of an internal data center is mandatory. Critical information systems are kept in internal data centers, and physical security measures must be taken. Access to these areas is restricted to authorized personnel only. Card systems should prevent unauthorized people from entering, and entrances and exits should be monitored with security cameras. Natural and environmental risks should be taken into account, and the location of data centers should not be marked from the outside. There should be a power cut, fire, smoke, and temperature sensors, and their maintenance should be done regularly. Compliance of data centers with necessary conditions is audited in periodic ITIL, and COBIT audits.

In addition, banks should have a secondary data center in case of a disaster in the primary data center. When choosing the secondary data center location, it should be so that it will not be affected by a regional disaster that will occur in the primary data center location. If necessary, business continuity should be provided with these data centers. There should be no data loss between primary and secondary data centers. For this, data is transferred at specific intervals. In today's technology, systems with instant data transfer can also be used.

2.2 Capacity Management

With the increasing need for data centers, capacity management in data centers has also become important. Capacity management is essential both for cost and sustainability by ensuring energy efficiency. Efficient capacity management must be done to ensure a cost-effective IT process [45].

The hardware used in data centers is costly. For this reason, capacity management should be done well to meet the increasing need by using the limited hardware available efficiently. In addition, energy consumption constitutes a considerable part of operational expenses. Energy efficiency can be achieved with capacity management.

One of the main goals of capacity management is to ensure that sufficient total capacity is available when needed. For this reason, it is necessary to manage the uncertain server workload [46].

The most critical technology that enables capacity management is virtualization. With the increase in virtualization technology, efficient capacity management can be done—virtual servers running on physical servers and virtualization technology that allows the management of these virtual servers. Many workloads are hosted by a limited number of servers with limited capacity. The capacity requirement of each workload, which can change frequently, can be met with effective capacity management. Adaptation to the frequently changing capacity needs is achieved with virtualization technologies.

2.3 Energy in Data Centers

With the increase in the need for data centers, the use of electricity by data centers has also increased. While data centers consumed 1.5% of the world's electricity in 2011, the consumption rate increased by approximately 10-12% annually. This growth indicates that the importance of efficiency in energy use will increase in the coming years [47].

With the developing server consolidation methods, energy efficiency is also observed. Power usage effectiveness ratio increased by 36.4% from 2007 to 2020 (Figure 2.3) [48]. The hardware with the highest energy consumption in data centers is physical servers. Energy consumption in physical servers can be achieved with effective capacity management.

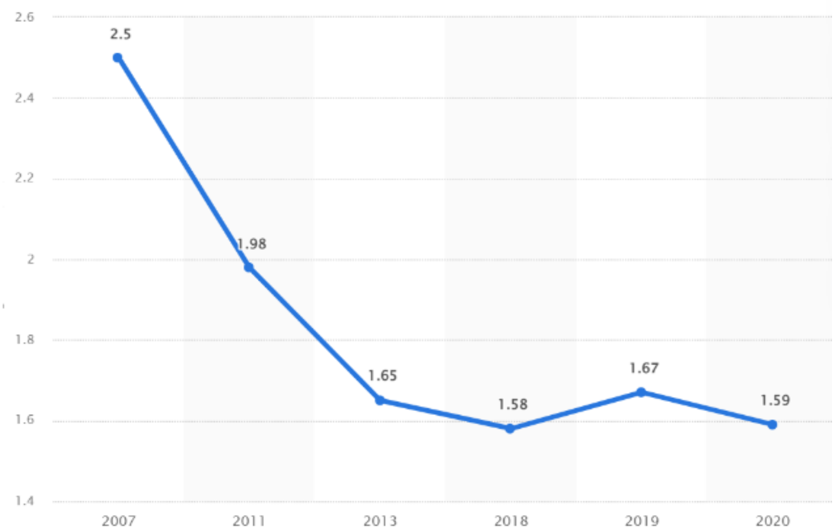

Figure 2.3 Power usage effectiveness ratio [48]

2.4 Physical and Virtual Servers

Physical servers are high-cost hardware with limited capacity. Services running on physical servers have different software and varying resource needs. Virtual servers are installed on physical servers so that multiple services with different software and varying resource needs can run on the same physical server. Virtual server capacity can be increased or decreased according to the service needs. This technology, which allows the installation and management of virtual servers, is called virtualization technology.

2.4.1 Physical Servers

- It is the hardware that can run many virtual servers on it.
- Has the limited capacity
- High operating cost
- Energy consumption increases as the total CPU used by virtual servers running on it increases.
- Performance issues occur if CPU usage exceeds a critical level.
- Migrating the workload from one physical server to another is complex and cannot be done online. It is also complicated to move a workload to another data center.
- No additional skills are required to manage each physical server. Managing the overall system is more difficult because the workload cannot be moved online.
- Security applications are challenging to deploy and manage.

2.4.2 Virtual Servers

- It is the server on which the services run.
- Resources are allocated according to the needs of the work to be done.
- Workload may change over time. A workload is all of the tasks running on a server, and each workload consumes computing resources at a variable rate over time. The resource to be allocated is determined by monitoring the workload.
- Virtual servers allow resources to increase and decrease.

- It allows workload migration between physical servers and even online from one data center to another.
- It is easier to manage the overall system as it reduces the need for physical servers and allows online switching between physical servers. However, additional skills are required to manage virtual servers.
- An error in an application does not affect the whole system; it can be solved within the virtual server.

2.5 Virtualization

Virtualization technologies are widely used to simplify the management of data centers and reduce energy consumption. There are five areas where virtualization is used. These are network virtualization, hardware virtualization, software virtualization, storage virtualization, and server virtualization.

- Network virtualization is the virtualization of networks that provide communication by connecting virtual environments and physical computers.
- Hardware virtualization is the virtual creation of hardware on a physical system more than once. This virtual hardware can be installed and run-on operating systems running on real hardware.
- Software virtualization is divided into operating system virtualization and application virtualization. In virtualization, depending on the operating system, these virtual environments are used by isolating them from each other with containers on a single operating system. In application virtualization, it is the operation of the software on a server without installation. It forms the basis of cloud software.
- Storage virtualization is the technology where the capacities of data storage centers are pooled, and all of them can be managed from a single interface. It enables operations such as backup, archiving and restoration performed in storage to be done more efficiently and in a short time.
- Server virtualization is the process that enables the resources of a single server to be used by multiple virtual servers. In this thesis, server virtualization is used.

Virtual server migration technology allows virtual machines to move between physical servers. Their resource usage reduces energy consumption by consolidating virtual servers on a minimum number of physical servers [24]. Virtualization enables many virtual servers to run on a single physical server. This technique is called Server Consolidation. One of the most important reasons for the spread of virtualization is server consolidation. Many applications and services run securely across multiple operating systems on the same physical server. Virtual servers enable live migration and allow virtual servers to be moved freely between physical servers with negligible downtime. It prevents many underutilized servers from taking up more space and consuming more resources than necessary for workloads.

Virtualization runs more applications and services on the less physical server. It maximizes the usage efficiency of physical servers. Another benefit of virtualization is that it can efficiently perform operations such as restore and failover on a virtual server. Virtualization allows workloads to be migrated live to another server. Thus, service interruption is prevented. Live migration without service interruption enables workload balancing and prevents bottlenecks in servers.

The main benefits of virtualization are,

- Provides resource management by consolidating workloads using virtual servers
- Reduces data center cost
- It reduces carbon emissions by providing energy efficiency. It increases sustainability.

Different companies have developed different software to enable the virtualization and management of servers. This software varies according to user needs. They can be used in different systems and infrastructures.

2.6 Central Process Unit (CPU)

CPU is the data processing part of servers. The number of cores in the CPU determines its performance. The CPU performs operations by converting logical operations to electrical frequencies. The data flow within the servers is carried out using arithmetic codes. This flow is provided by semiconductor elements called transistors. Electricity is produced by this system and produced according to incoming 0 and 1 commands. The clock beats determine the electricity produced according to these commands. The beat rate determines the processor's cycle rate per second. The cycle rate is measured in hertz. As the usage of an application increases, the number of instructions coming to the CPU also increases. This number of commands is measured in hertz: the larger the hertz, the more electricity produced.

Because of the electricity produced on the CPU, it needs a cooling system. Overheated CPUs cannot work correctly and cannot transmit data. In order to meet this cooling need, fans are working simultaneously with the processor in the server. These fans also consume electricity depending on the CPU operating speed.

One of the parameters that most affect the energy consumption in physical servers is the CPU. The uncertain workload on virtual servers consumes CPU resources on physical servers. For this reason, capacity planning should be done by considering the CPU parameter for efficient server consolidation.

3 PROBLEM STATEMENT AND MATHEMATICAL MODEL

The workload on virtual servers is uncertain due to the nature of the work. For this reason, resource usage, which is one of the parameters of the proposed mathematical model, is uncertain. Stochastic modeling is used to deal with uncertainty [43].

The proposed model is a two-stage stochastic programming model. First-stage decisions are made in the presence of uncertainty for future scenarios. The first step in the model is the decision to purchase a new physical server. The second stage decision depends on the first stage decisions. After purchasing a physical server, assigning virtual servers to existing physical servers is the second stage decision in the proposed model.

The model aims to minimize annual costs. The cost is associated with purchasing a physical server and the cost of energy use.

While determining the purchase cost, a server whose features are specified in Table 3.1 and the sales price of the hardware required for the server to work in January 2022 were taken into account. The annual cost of purchased physical servers has been determined by considering the useful life stated in the tax procedure law. The specified useful life for physical servers is four years.

Studies showing that energy usage is directly related to CPU usage are considered. In this model, energy usage cost is analyzed according to CPU usage. We measured the energy usage every 100 [msec] using the power meter Uwmeter (Figure 3.1). The characteristics of the measured physical server are given in Table 3.1. In this model, it is assumed that there is a linear relationship between CPU and energy usage. The measurement was made by monitoring five identical servers. The measured power usage of these servers is averaged. Observed results support previous studies [22], [49].

Table 3.1 Physical server features

	CPU Cores	Memory
Physical Server	16	511.66 GB

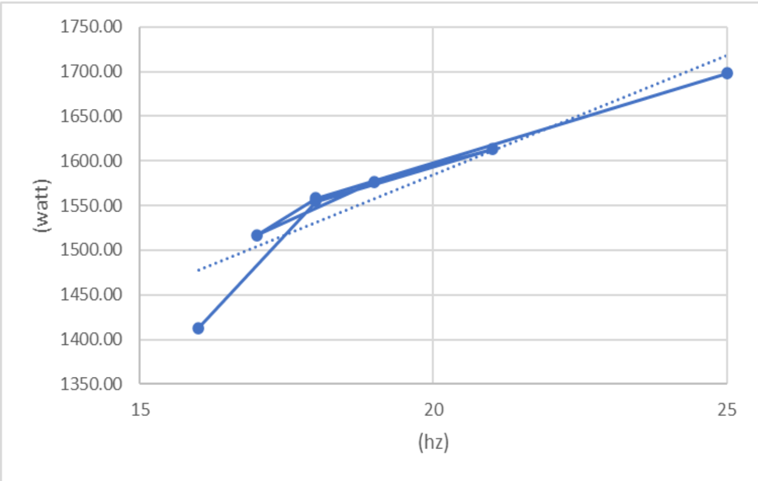

Figure 3.1 Relationship between CPU and energy usage

A linear line equation was created according to Figure 3.1. According to equation (3.1), the CPU energy consumption relationship is established. It has been determined how much watts of energy the physical servers should consume for each Hz.

$$y(watt) = ax(Hz) + b \tag{3.1}$$

The cost is calculated by multiplying the power consumed in watts by the electricity cost and time.

Total energy usage is the sum of the energy consumed per CPU unit used and the constant energy usage due to physical servers being turned on. The workloads of the services running on virtual servers change in specific periods. With the change in these workloads, the amount of CPU used by virtual servers fluctuates.

3.1 Model

In the study, the following assumptions have been made in developing the mathematical model for minimizing cost in a data center:

- The most essential and costly equipment of data centers is physical servers.
- The most important factor affecting energy usage in physical servers is the CPU parameter.
- In the model, the capacity constraint affecting the assignment of virtual servers to physical servers is the CPU parameter.
- The workloads of the services running on virtual servers change in specific periods.
- The amount of CPU used by virtual servers fluctuates. Therefore, the parameter b_{it}^{s} is a stochastic parameter in the model.

3.1.1 Sets

$i: virtual\ server \qquad i \in I$

$j: physical\ server \qquad j \in J$

$t: period \qquad t \in T$

$s: scenario \qquad s \in S$

3.1.2 Parameters

$c_j: CPU\ resource\ capacity\ of\ the\ physical\ server\ j\ (Hz)$

$b_{it}^{s}: CPU\ resource\ used\ by\ the\ virtual\ server\ i\ in\ the\ period\ t\ and\ scenario\ s\ (Hz)$

$FC_t: The\ fixed\ annual\ cost\ of\ the\ physical\ server\ for\ period\ t\ (TL)$

$VC_t: The\ variable\ annual\ cost\ per\ Hz\ \ that\ depending\ on\ the\ resource\ used\ on\ the$ $physical\ server\ for\ period\ t\ (TL/Hz)$

$PC_j: The\ annual\ purchase\ cost\ of\ physical\ server$

$\pi_s: Probability\ of\ scenario\ s$

$M: A\ Sufficiently\ large\ positive\ number$

3.1.3 Decision Variables

$$x_{jt}^{s} = \begin{cases} 1, & \textit{if physical server } j \textit{ works in period } t \textit{ at scenario } s \\ 0, & \textit{otherwise} \end{cases} \tag{3.2}$$

$$y_{j} = \begin{cases} 1, & \textit{if physical server } j \textit{ is purchased} \\ 0, & \textit{otherwise} \end{cases} \tag{3.3}$$

$$w_{ijt}^{s} = \begin{cases} 1, & \textit{if virtual server } i \textit{ is assigned to physical server } j \textit{ in period } t \\ & \textit{at scenario } s \\ 0, & \textit{otherwise} \end{cases} \tag{3.4}$$

3.1.4 Mathematical Model

Objective function

$$min\ z = \sum_{j} PC_j y_j + \sum_{j}\sum_{t}\sum_{s} \pi_s FC_t x_{jt}^{s} + \sum_{i}\sum_{j}\sum_{t}\sum_{s} \pi_s VC_t b_{it}^{s}\, w_{ijt}^{s} \tag{3.5}$$

Constraints

$$b_{it}^{s} \le \sum_{j} w_{ijt}^{s}\, M \qquad \forall i, t, s \tag{3.6}$$

$$x_{jt}^{s} \le y_j \qquad \forall\, j, t, s \tag{3.7}$$

$$\sum_{j} w_{ijt}^{s} \le 1 \qquad \forall\, i, t, s \tag{3.8}$$

$$\sum_{i} w_{ijt}^{s} \le x_{jt}^{s} M \qquad \forall\, j, t, s \tag{3.9}$$

$$\sum_{i} b_{it}^{s} w_{ijt}^{s} \le c_j \qquad \forall\, j, t, s \tag{3.10}$$

$$x_{jt}^{s}, y_j, w_{ijt}^{s} \in \{0,1\} \qquad \forall\, i, j, t, s \tag{3.11}$$

The model aims to minimize annual costs related to the energy consumption caused by physical servers, the CPU resource that the virtual servers run on, and the purchase of physical servers. The objective function (3.5) minimizes annual cost use by considering these factors. The first stage of the equation is associated with the purchase cost. The other two sums are related to the energy cost. FC_t is the annual fixed cost. The energy a physical server consumes when it is turned on. VC_t is the annual variable cost per Hz. It is associated with the CPU resource used by the virtual server running on a physical server. If a virtual server consumes a resource, it must be assigned to a physical server (3.6). Virtual servers provide their resource needs from physical servers; software and security settings are made independently of physical servers. If our physical servers are insufficient, a new physical server must be purchased (3.7). Each virtual server is assigned to only one physical server (3.8). Even if no virtual server is running on a physical server, there is energy consumption when idle. For this reason, physical servers that do not have virtual servers are turned off (3.9). Physical servers have limited resources. The number of cores determines the CPU resource that can work on a physical server. For this reason, the total CPU resource consumed by virtual servers running on a physical server cannot exceed the capacity of the physical server (3.10). Equation (3.11) is a restriction on the type of decision variables.

4
IMPLEMENTATION

The proposed model is applied to a bank's internal data center. The Bank is a medium-sized investment bank operating in Turkey. Bank provides services to entrepreneurs through its corporate and investment banking products. Operating at the head office, the Bank has 60 employees. Significant IT investments are made due to the digitalization strategy. 146 virtual servers are installed on 5 physical servers in the bank's data center. VMware is used as virtualization software. This software is the most frequently used software in the world in virtualization. The software has different products that are used for different purposes. The product used in the bank for server virtualization is the VMware Server product. CPU resource demand varies according to the needs of the services running on these virtual servers.

In the bank's current situation, there is an over-capacity problem of physical servers. When this problem is overloaded, it causes slowness in applications, stopping some virtual servers and service interruptions. Before making high-cost investments for solving this problem, servers were monitored, and cases of overcapacity were determined. To prevent sudden service interruptions in critical servers, non-critical servers are turned off and on by making manual interventions. These solutions are both temporary and cause a loss of workforce. While categorizing the collected data in this study, data about the moments of overload were also collected.

Service interruptions that may occur in critical services in the bank cause loss of customers and money. Problems such as not being able to make customer transactions and not making official and timely reports to legal authorities cause financial losses and penal actions. For this reason, it is one of the most important issues to ensure the continuity of critical services. With effective server consolidation, server-induced service interruptions should be prevented. Due to banking transactions and other services provided, the need for resources is changing during the day. After the banking transactions are completed, the workload of banking applications is reduced. The need for resources increases with backup operations and reporting applications in databases. For this reason, the resource needs of virtual servers change during the day. There are uncertainties in resources due to banking transaction intensities and the needs arising from the use of different products.

4.1 Data Collection

The appropriate distribution of the data collected under different scenarios was found with the Arena Input Analyzer tool, and the model was run considering this distribution. Data fit the beta distribution. The mean and the standard deviation parameters of the data were found for each scenario.

Data were collected under different scenarios for different periods. Periods are determined by considering the working hours of the bank, routine backup, and reporting processes.

- 1st Period: It is 8:00-17:00 hours covering the working hours of banking transactions.
- 2nd Period: The period after work until the end of day operations.
- 3rd Period: It is between 23:00-2:00, covering the end-of-day period in which different services covering the accounting processes of banking transactions are run.
- 4th Period: It is in which automatic reporting is made after the accounting process until the beginning of the working hours after the end of the day.

The accurate data in the bank is taken from the Microsoft System Center family product Operation Manager application, which is used to monitor and control. Through this application, the capacity status of physical servers and the instant or average resource consumption of running virtual servers can be monitored and reported.

After the model decides whether to purchase additional physical servers, it assigns virtual servers to physical servers. When calculating the data center energy cost, the fixed cost used by the physical servers and the variable cost that varies according to the CPU resource used by the virtual servers running on it are considered. The energy cost is directly proportional to the total CPU resource consumed. In practice, the cost was calculated by measuring the energy consumed by physical servers under different conditions. The costs are determined by considering the periods and the tariff in which banks are included in Turkey.

4.2 Results

This thesis made server consolidation in the data center using stochastic modeling. The purpose of the mathematical model is to minimize the cost. Cost minimization has been made by considering the investment cost and the energy usage arising from CPU usage. As a result of the thesis, a model that provides energy efficiency was proposed and, the model was solved for a bank. The model was solved by using the GAMS program. The computer on which the GAMS program is run is Intel(R) Core(TM) i7-8565U CPU , 1.80GHz 1.99GHz, 16GB.

The data are used without being divided into scenarios and if the mean and standard deviation values of all data are found, the results in Table 4.1 are obtained.

Table 4.1 Scenario independent solution

b(i,t)	Beta (0,92, 8,14)
Obj. Value	1776,72

According to the solution in Table 4.1, when we examine the results in this model, there is no need for a new physical server purchase. This scenario-independent model will not yield results for all real-life situations. Let us divide the data into scenarios and assume that only one of these scenarios occurs. If we only use the data of these scenarios in the calculation, we get the results in Table 4.2. When scenario 4 is realized, the scenario independent model will not meet the need.

Table 4.2 Solutions for each scenario

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**
b(i,t)	Beta (1,04, 6,49)	Beta (0,84, 5,74)	Beta (0.83, 5,44)	Beta (0.62, 4,98)
Obj. Value	1739,94	2745,44	2796,08	111.952,07

When we examine the results, we see that the scenario with the highest cost is Scenario 4, and the scenario with the lowest cost is scenario 1. Therefore, scenario 2 is the best, and scenario 4 is the worst.

Considering all scenarios, the probability of realization was assigned to these scenarios, and the model was solved. It is examined how the change in the probability of occurrence in each scenario affects the objective function. The change in objective functions is as in Table 4.3.

Table 4.3 The effect of the changes in scenario probabilities on the objective function

Probability				Objective Function
Scenario 1	Scenario 2	Scenario 3	Scenario 4	
0,25	0,25	0,25	0,25	109.608
0,4	0,2	0,2	0,2	109.314
0,55	0,15	0,15	0,15	109.021
0,7	0,1	0,1	0,1	108.727
0,85	0,05	0,05	0,05	108.433
1	0	0	0	108.139
0,2	0,4	0,2	0,2	109.515
0,15	0,55	0,15	0,15	109.423
0,1	0,7	0,1	0,1	109.330
0,05	0,85	0,05	0,05	109.238
0	1	0	0	109.145
0,2	0,2	0,4	0,2	109.525
0,15	0,15	0,55	0,15	109.443
0,1	0,1	0,7	0,1	109.361
0,05	0,05	0,85	0,05	109.278
0	0	1	0	109.196
0,2	0,2	0,2	0,4	110.077
0,15	0,15	0,15	0,55	110.545
0,1	0,1	0,1	0,7	111.014
0,05	0,05	0,05	0,85	111.483
0	0	0	1	111.952

It is observed that the objective function decreases as the probability of occurrence of scenario 1 increases. In this case, scenario 1 is the best, and the higher the probability of occurrence, the lower the cost.

It is observed that the objective function increases as the probability of realization of scenario 4 increases. In this case, scenario 4 is the worst-case scenario, and the increased probability of occurrence increases the cost.

When the results are examined, the most significant factor affecting the cost is the investment cost. The number of physical servers running was 6 for all scenario probabilities. Accordingly, there is a need to purchase a new server. All physical servers are operational in all scenarios. Virtual servers could not be transferred to other servers to free a physical server because there was no capacity gap.

Problems arising from overcapacity will be resolved by purchasing a new physical server. A server consolidation model has been proposed to avoid the need for manual interventions in case of overcapacity. There is also a risk cost, as service interruptions arising from overcapacity prevent critical services from being provided. The investment cost to be made will eliminate the risk.

4.3 Sensitivity Analysis

Sensitivity analysis was performed to examine the effect of the parameter change on the objective function in the model. The effect of the change in the physical server capacity and the change in the purchasing cost due to this change on the objective function has been examined.

The solution assumes that all servers are taken at equal capacity with the existing servers. The capacities of the new physical servers to be purchased may differ from the existing servers. It has been examined how the new servers' capacities to be purchased affect the objective function. The model is solved by assuming that each scenario occurs with equal probability.

- Assuming that purchasing a server with high capacity does not affect the cost of purchasing a server, the following results are obtained (Table 4.4).

Table 4.4 Effect of purchase cost independent server capacities on the decision variable

CPU Cores				Objective Function	Purchase Decision
Existing PMs ($y_{1,2,3,4,5}$)	Option 1 (y_6)	Option 2 (y_7)	Option 3 (y_8)		
16	16	16	16	109.608	Only one server is purchased.
16	32	32	32	108.918	Only one server is purchased.
16	64	64	64	108.057	Only one server is purchased.
16	16	32	64	108.057	Only Option 3 (y_8) server is purchased.

According to these results, increasing the CPU capacity will reduce the total workload on the servers and cause a decrease in energy usage.

- The high capacity of the new server to be purchased has reduced the server load running on other servers. Reducing the server load reduces the variable cost dependent on the CPU used. Since it is assumed that the purchase cost of the new server to be purchased does not change depending on the server capacity, only the energy cost has affected the objective function. This result highlights the importance of load distribution for energy-efficient operation. The following results are obtained if we include the cost of purchasing a server in purchasing a high-capacity server (Table 4.5).

Table 4.5 Effect of cost dependent server capacities on the decision variable

CPU Cores				Objective Function	Purchase Decision
Existing PMs ($y_{1,2,3,4,5}$)	Option 1 (y_6)	Option 2 (y_7)	Option 3 (y_8)		
16	16	16	16	109.608	Only one server is purchased.
16	32	32	32	129.918	Only one server is purchased.
16	64	64	64	150.057	Only one server is purchased.
16	16	32	64	109.608	Only Option 1 (y_6) server is purchased.

Servers are high-cost hardware. Increasing the CPU capacity increases the purchasing cost of the server. Since the purchase cost is much higher than the energy cost, it was decided to purchase the y_6 physical server by bearing the energy cost in the last option.

In the solution, where we did not consider the purchase cost, the energy cost was reduced by distributing the resources used by the virtual servers. When the purchasing cost is taken into account, it is observed that the cost increases as the capacity increases. The purchase cost is much more than the energy cost savings that high-capacity servers will provide. For this reason, a low-capacity server with low purchasing cost is selected in the model.

5
RESULTS AND DISCUSSION

With the developing technology, the importance of data centers has increased. Data centers form the infrastructure of information technologies. Therefore, providing energy efficiency in data centers has also been a significant problem. In data centers, the hardware that IP traffic passes through and makes a significant part of energy consumption is servers. With the development of virtualization technology, many virtual servers can be installed on physical servers. Server consolidation is one of the techniques that provides capacity management in physical servers.

The thesis proposes a stochastic model for the server consolidation problem. A real-life problem has been handled, and an application has been made for a bank data center. With server consolidation in data centers where high-cost hardware is used, fewer physical servers can operate with less energy consumption. The proposed stochastic mathematical model aims to minimize the annual cost.

In servers, the cost is the cost of purchasing and energy use. Physical servers are hardware with high investment costs. Therefore, making the purchasing decision is an important issue. The efficient use of fewer physical servers is an essential factor affecting the investment cost. The constraint that affects the energy usage cost the most is the CPU. For this reason, the relationship between the CPU resource usage of the servers and the consumed power was examined, and the energy usage cost was determined. CPU, which is one of the many parameters of the server, has been considered a capacity constraint. The model is solved for different scenarios under this constraint.

As a result of this study, cost-minimizing was provided for a bank's data center using two-stage scenario-based stochastic modeling. Cost minimization was achieved by providing energy efficiency. The need for new servers was determined in the 1st stage, and in the 2nd stage, the energy cost was calculated by assigning virtual servers to physical servers. The model offered a solution to the servers' exceeding capacity. With previous manual interventions, service interruptions in critical services were prevented. Manual intervention cannot prevent instant service interruptions. It could not be a permanent solution and caused the loss of the workforce. The proposed model prevents service interruptions without the need for manual intervention. It has eliminated the risk of loss of earnings and legal penalties caused by the overcapacity problem.

Furthermore, the model is tested for different scenarios. The effect of the change in the probability of realization of the scenarios on the result and the effect of the new server capacity on the cost were also examined. Since the purchase cost is much higher than the energy cost, the energy cost was incurred instead of purchasing a high-capacity server.

The study established a model for the servers by considering the CPU parameter. In future studies, the model's reliability can be increased by including other parameters such as memory and bandwidth.

REFERENCES

[1] A. Mauricio ve P. Maurizio, *Data Center Fundamentals*. Cisco Press, 2003.

[2] G. Scheithauer *vd.*, "A Stochastic Bin Packing Approach for Server Consolidation with Conflicts A Stochastic Bin Packing Approach for Server Consolidation with Conflicts Highly Adaptive Energy-Efficient Computing View project The Skiving Stock Problem View project A Stochastic", 2020.

[3] J. Arjona Aroca, A. Chatzipapas, A. Fernández Anta, ve V. Mancuso, "A measurement-based analysis of the energy consumption of data center servers", *e-Energy 2014 - Proc. 5th ACM Int. Conf. Futur. Energy Syst.*, ss. 63–74, 2014,.

[4] B. Heller *vd.*, "Elastictree: Saving energy in data center networks", içinde *Proceedings of NSDI 2010: 7th USENIX Symposium on Networked Systems Design and Implementation*, 2010, ss. 249–264.

[5] A. Zomaya, Y. C. Lee, A. Y. Zomaya, Y. C. Lee, · A Y Zomaya, ve A. Y. Zomaya, "Energy efficient utilization of resources in Cloud computing systems", *Artic. J. Supercomput.*, 2010.

[6] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, ve R. P. Doyle, "Managing energy and server resources in hosting centers", s. 103, 2001.

[7] E. Pinheiro, R. Bianchini, E. V Carrera, ve T. Heath, "Load Balancing and Unbalancing for Power and Performance in Cluster-Based Systems", *Colp*, ss. 4.1-4.8, 2001.

[8] and R. B. Beloglazov, Anton, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers", *Concurr. Comput. Pract. Exp.*, c. 24, sayı 13, ss. 1397–1420, 2012.

[9] X. Fan, W. D. Weber, ve L. A. Barroso, "Power provisioning for a warehouse-sized computer", içinde *Proceedings - International Symposium on Computer Architecture*, 2007, ss. 13–23.

[10] J. Martinovic, M. Hähnel, W. Dargie, ve G. Scheithauer, "A Stochastic Bin Packing Approach for Server Consolidation with Conflicts", sayı March, ss. 159–165, 2020.

[11] M. Kafali, N. Aydin, Y. Genç, ve U. B. Çelebi, "A two-stage stochastic model for workforce capacity requirement in shipbuilding", *J. Mar. Eng. Technol.*, 2019.

[12] N. Aydin, "A stochastic mathematical model to locate field hospitals under disruption uncertainty for large-scale disaster preparedness", *An Int. J. Optim. Control Theor. Appl.*, c. 6, sayı 2, ss. 85–102, Mar. 2016.

[13] E. Celik, N. Aydin, ve A. Taskin Gumus, "A stochastic location and allocation model for critical items to response large-scale emergencies: A case of Turkey", *An Int. J. Optim. Control Theor. Appl.*, c. 7, sayı 1, ss. 1–15, 2016.

[14] N. Aydin, “Designing reverse logistics network of end-of-life-buildings as preparedness to disasters under uncertainty”, *J. Clean. Prod.*, c. 256, May. 2020,

[15] B. Bermejo ve C. Juiz, “PERFORMANCE AND ENERGY CONSUMPTION TRADEOFF in Server Consolidation”, 2014.

[16] M. A. Ala’anzy ve M. Othman, “Mapping and Consolidation of VMs Using Locust-Inspired Algorithms for Green Cloud Computing”, *Neural Process. Lett.*, c. 54, sayı 1, ss. 405–421, 2022.

[17] S. Padhy, M. H. Tsai, S. Sharma, ve J. Chou, “CAMIRA: a consolidation-aware migration avoidance job scheduling strategy for virtualized parallel computing clusters”, *J. Supercomput.*, c. 78, sayı 9, ss. 11921–11948, Şub. 2022.

[18] M. Uddin *vd.*, “Server consolidation: A technique to enhance cloud data center power efficiency and overall cost of ownership”, *Int. J. Distrib. Sens. Networks*, c. 17, sayı 3, 2021.

[19] J. Martinovic, M. Hähnel, G. Scheithauer, ve W. Dargie, “An introduction to stochastic bin packing-based server consolidation with conflicts”, *TOP*, ss. 1–36, 2021.

[20] et al Song, Tao, “Server Consolidation Energy-Saving Algorithm Based on Resource Reservation and Resource Alloacation Strategy”, *Proc. 2011 IEEE Int. Conf. Robot. Biomimetics*, c. 7, ss. 1714452–171452, 2019.

[21] A. Verma, P. Ahuja, ve A. Neogi, “pMapper: Power and Migration Cost Aware Application Placement in Virtualized Systems”.

[22] S. Mazumdar ve M. Pranzo, “Power efficient server consolidation for Cloud data center”, *Futur. Gener. Comput. Syst.*, c. 70, ss. 4–16, 2017.

[23] S. K. Abd, S. A. R. Al-Haddad, F. Hashim, A. B. H. J. Abdullah, ve S. Yussof, “An effective approach for managing power consumption in cloud computing infrastructure”, *J. Comput. Sci.*, c. 21, ss. 349–360, 2017.

[24] W. Zhu, Y. Zhuang, ve L. Zhang, “A three-dimensional virtual resource scheduling method for energy saving in cloud computing”, *Futur. Gener. Comput. Syst.*, c. 69, ss. 66–74, 2017.

[25] R. Li, Q. Zheng, X. Li, ve J. Wu, “A novel multi-objective optimization scheme for rebalancing virtual machine placement”, içinde *IEEE International Conference on Cloud Computing, CLOUD*, 2017, ss. 710–717.

[26] Q. H. Nguyen, P. D. Nien, N. H. Nam, N. Huynh Tuong, ve N. Thoai, “A genetic algorithm for power-aware virtual machine allocation in private cloud”, içinde *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, c. 7804 LNCS, ss. 183–191.

[27] A. Gupta, L. V Kalé, D. Milojicic, P. Faraboschi, ve S. M. Balle, “HPC-aware VM placement in infrastructure clouds”, içinde *Proceedings of the IEEE International Conference on Cloud Engineering, IC2E 2013*, 2013, ss. 11–20.

[28] W. Chawarut ve L. Woraphon, “Energy-aware and real-time service management in cloud computing”, *2013 10th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. ECTI-CON 2013*, ss. 1–5, 2013.

[29] A. Beloglazov, J. Abawajy, ve R. Buyya, “Future Generation Computer Systems Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing”, *Futur. Gener. Comput. Syst.*, c. 28, ss. 755–768, 2012.

[30] B. Kantarci, L. Foschini, A. Corradi, ve H. T. Mouftah, “Inter-and-intra data center VM-placement for energy-efficient large-Scale cloud systems”, içinde *2012 IEEE Globecom Workshops, GC Wkshps 2012*, 2012, ss. 708–713.

[31] H. Goudarzi, M. Ghasemazar, ve M. Pedram, “SLA-based optimization of power and migration cost in cloud computing”, içinde *Proceedings - 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGrid 2012*, 2012, ss. 172–179.

[32] Y. C. Lee ve A. Y. Zomaya, “Energy efficient utilization of resources in cloud computing systems”, *J. Supercomput.*, c. 60, sayı 2, ss. 268–280, 2012.

[33] Q. Zhang, E. Gürses, R. Boutaba, ve J. Xiao, “Dynamic resource allocation for spot markets in clouds”, içinde *USENIX Workshop on Hot Topics in Management of Internet, Cloud, and Enterprise Networks and Services, Hot-ICE 2011*, 2011.

[34] A. Beloglazov ve R. Buyya, “Energy Efficient Allocation of Virtual Machines in Cloud Data Centers Geographical load balancing View project Data-intensive application resource provisionig in clouds View project Energy Efficient Allocation of Virtual Machines in Cloud Data Centers”, 2010.

[35] M. Cardosa, M. R. Korupolu, ve A. Singh, “Shares and utilities based power consolidation in virtualized server environments”, içinde *2009 IFIP/IEEE International Symposium on Integrated Network Management, IM 2009*, 2009, ss. 327–334.

[36] A. Gandhi, M. Harchol-Balter, R. Das, ve C. Lefurgy, “Optimal Power Allocation in Server Farms”, 2009.

[37] L. Li, “An optimistic differentiated service job scheduling system for Cloud Computing service users and providers”, içinde *3rd International Conference on Multimedia and Ubiquitous Engineering, MUE 2009*, 2009, ss. 295–299.

[38] S. Srikantaiah, A. Kansal, ve F. Zhao, “Energy aware consolidation for cloud computing”, içinde *Workshop on Power Aware Computing and Systems, HotPower 2008*, 2008.

[39] R. Raghavendra, P. Ranganathan, Z. Wang, ve X. Zhu, *No “Power” Struggles: Coordinated Multi-level Power Management for the Data Center*. 2008.

[40] D. Kusic, J. O. Kephart, J. E. Hanson, N. Kandasamy, ve G. Jiang, “Power and performance management of virtualized computing environments via lookahead control”, içinde *5th International Conference on Autonomic*

Computing, ICAC 2008, 2008, ss. 3–12.

[41] R. Nathuji ve K. Schwan, “VirtualPower: Coordinated power management in virtualized enterprise systems”, içinde *SOSP’07 - Proceedings of 21st ACM SIGOPS Symposium on Operating Systems Principles*, 2007, ss. 265–278.

[42] G. B. Dantzig ve G. Infanger, “A probabilistic lower bound for two-stage stochastic programs”, içinde *International Series in Operations Research and Management Science*, c. 150, Springer New York LLC, 2011, ss. 13–35.

[43] B. Speitkamp ve M. Bichler, “A mathematical programming approach for server consolidation problems in virtualized data centers”, *IEEE Trans. Serv. Comput.*, c. 3, sayı 4, ss. 266–278, 2010.

[44] A. Wierman, Z. Liu, I. Liu, ve H. Mohsenian-Rad, “Opportunities and challenges for data center demand response”, içinde *2014 International Green Computing Conference, IGCC 2014*, 2015.

[45] H. Müller, A. Kharitonov, A. Nahhas, S. Bosse, ve K. Turowski, “Addressing IT Capacity Management Concerns Using Machine Learning Techniques”, *SN Comput. Sci.*, c. 3, sayı 1, ss. 1–15, Eki. 2022.

[46] D. Gmach, D. Gmach, J. Rolia, L. Cherkasova, ve A. Kemper, “Capacity Management and Demand Prediction for Next Generation Data Centers XenQoS View project Optimizing Backup Solution View project Capacity Management and Demand Prediction for Next Generation Data Centers Capacity Management and Demand Prediction for Next Generation Data Centers”, ss. 9–13, 2007.

[47] T. Dreibholz, M. Becke, ve H. Adhari, “Report to Congress on Server and Data Center Energy Efficiency Public Law 109-431”, *Tdr.Wiwi.Uni-Due.De*, c. 109, sayı August, s. 431, 2007.

[48] “• Data center average annual PUE worldwide 2020 | Statista”. Available at: https://www.statista.com/statistics/1229367/data-center-average-annual-pue-worldwide.

[49] H. Kataoka, D. Duolikun, T. Enokido, ve M. Takizawa, “Power Consumption and Computation Models of a Server with a Multi-core CPU and Experiments”, içinde *Proceedings - IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2015*, 2015, ss. 217–222.

PUBLICATIONS FROM THE THESIS

Conference Papers

1. N. Aydin, Ö. Erden., "A Stochastic Model For Energy-Efficient Server Consolidation", 9. International Congress on Engineering, Architecture and Design, 2022, 1059-1068.