International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

OVERVIEW OF RESOURCE ALLOCATION TECHNIQUES IN CLOUD COMPUTING

R.SANDHIYA
MSc Student,
Department of Computer Science,
Rajeswari College of Arts and Science for Women,
Bommayapalayam, Tamil Nadu, India.
Email ID: rgsandhiya2002@gmail.com

Dr.A.SARANYA
Assistant Professor,
Department of Computer Application,
Rajeswari College of Arts and Science for Women,
Bommayapalayam, Tamil Nadu, India.
Email ID: drsaranyarcw@gmail.com

Abstract

An appealing processing approach is cloud computing, which enables users to alter data, run apps, and access personal files from any computer by using the internet and centrally located distant servers without the need to install additional software. Through the centralization of storage, memory, computation, and bandwidth, this technology enables more efficient computing.

One major advantage of cloud computing is resource optimization; cost reduction and client satisfaction are the key objectives. This study presents a number of resource allocation options together with their associated issues. This work aims to provide academics and cloud users with an awareness of various implemented resource allocation algorithms.

Keywords: Cloud Computing, Resource Allocation Strategies (RAS), Virtual Machine, Service Level Agreement (SLA), CloudSim.

I. Introduction

The technology known as "cloud computing" keeps data and apps up to date by using central, distant servers. "Cloud

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources, e.g., networks, servers, storage, applications, and any other available resources that can be rapidly provisioned with minimal management effort or additional provider interaction," according to the NIST definition.[1]

The phrase "cloud computing" was first used in 2006 or 2007. It refers to a technological advancement that focuses on the methods of developing software, creating applications, and constructing computer systems. Dynamic provisioning, which is applied to services as well as compute capacity, storage, networking, and IT infrastructure overall, forms its basis. Development models with resources are provided; section 3 offers several ways for allocating resources that are put into practice; and section 4 wraps up the paper's contribution.

2. Cloud Computing Architecture

Reliability, organization, and remote access to hardware and software resources are the objectives of cloud computing. Online apps, infrastructure, and data storage are all provided by it. Since the program does not need to be installed on a PC, it offers an independent platform. Cloud computing thereby facilitates the portability and collaboration of corporate applications. A technological framework that allows for the production of virtual instances of IT resources, virtualization is the secret to the cloud. In order to enable numerous users to share physical IT resources' underlying processing capabilities, a layer of virtualization software enables those resources to provide several virtual images [2]. Three service models, four deployment methods, and five characteristics are how the National Institute of Standards and Technology (NIST) defines cloud computing (see figure).

Essential Characteristics

As cloud computing advances technologically and commercially, businesses are utilizing its numerous advantages. To enhance and expand your company, make the most of these advantages by being acquainted with the fundamental features of cloud computing.

Self-Service Available on Demand

The term "on-demand self-service" describes the service offered by cloud computing providers that makes it possible to provide cloud resources whenever needed. With on-demand self-service, a user's online control panel provides access to cloud services.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

Metrical Service

To determine when to scale up or down, one must have knowledge of the cloud's present demand. Stated differently, monitoring CPU, memory, and network bandwidth utilization is necessary to ensure that cloud users never experience resource depletion. The kinds of services that the cloud system provides determine in part the kinds of resources to measure.

Pooling of Resource

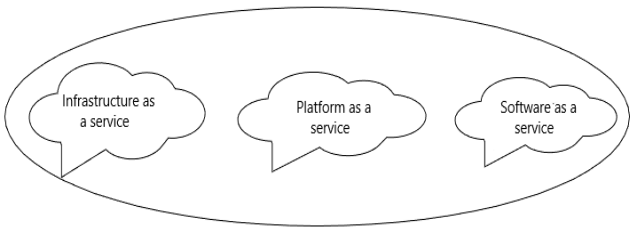
When suppliers supply temporary and scalable services to several clients, customers, or "tenants," the arrangement is referred to as resource pooling in IT contexts. Without affecting the client or end user, these services can be modified to meet the demands of any individual customer.

Expansive Network

When resources housed in a private cloud network (managed within an organization's firewall) are accessible from a variety of devices, including tablets, PCs, Macs, and smartphones, it is referred to as broad network access. These resources can also be accessed from any location with internet connectivity.

Stretchability

Elasticity in cloud computing refers to the capacity to modify resources in response to demand, enabling organizations to effortlessly manage fluctuating workloads. Because it only costs for what is used, this economical solution works for all kinds of enterprises.



Service Models

Infrastructure as a Service (IaaS)

Hardware as a Service (HaaS) is another name for IaaS. It is a layer on the platform for cloud computing. Customers can use it to outsource various parts of their IT infrastructure, including virtual machines, servers, networking, processing, and storage. Pay-per-use access is available to users of these resources via the Internet.

Platform as a Service (PaaS)

A sort of cloud computing service model called Platform as a Service, or PaaS, provides a scalable and adaptable cloud platform for creating, deploying, operating, and managing applications. Without having to

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

worry about maintaining hardware or updating the operating system and development tools, PaaS gives developers everything they need to create applications. Instead, a third-party service provider uses the cloud to supply the complete PaaS environment, or platform.

Software as a Service (SaaS)

Application delivery via the Internet is accomplished through software as a service, or SaaS. You remove the need for complicated software and device maintenance by just accessing the content online, eliminating the need for installation and upkeep.

Cloud-based software, on-demand software, or hosted software are other terms for SaaS applications. A SaaS provider's servers host SaaS apps, regardless of their name. Application performance, security, and availability are all under the provider's control.

A cloud's second tier comprises three fundamentally different cloud computing service models:

- Infrastructure-as-a-Service (IaaS).
- Cloud-as-a-Service (CaaS).
- Software-as-a-Service (SaaS)

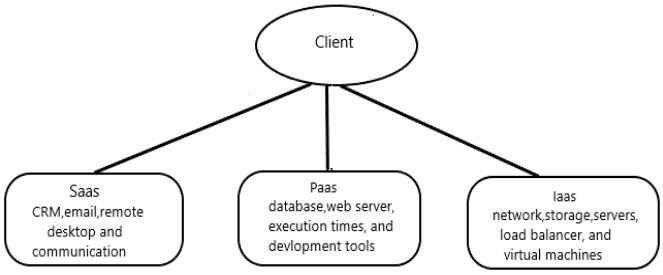


Figure 2 Cloud Computing Service Models

The ability to provision processing, storage, networks, and other essential computer resources is given to the user by Infrastructure-as-a-Service (IaaS). Consumers can install and run software, including operating systems, through administrative access to virtual machines and apps, even though they are unable to manage or control the underlying cloud infrastructure [3].

With Platform as-a-Service (PaaS), independent developers can create and launch online applications without having to purchase and configure physical servers, as seen in figure 2.

Software-as-a-service (SaaS) enables a software distribution paradigm where a third-party provider hosts applications and makes them available to clients over the internet. It can lower the overall cost of operations, maintenance, and software and hardware development. These days, companies like Google, Salesforce, Microsoft, and Zoho offer SaaS [4].

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

The last layer in cloud computing, known as development models, defines how a client and cloud provider will interact in a community, hybrid, private, or public cloud.

A public cloud provider, such as Microsoft, Amazon, or Google, owns the cloud infrastructure that is made available to the general public or a large industry group.

Implementing the public cloud model has several advantages, including low security, but it also has several drawbacks, including cost effectiveness, dependability, flexibility, and high scalability [5].

In a private cloud, a particular organization uses the infrastructure that is deployed, managed, and run. The procedure could take place on the property with a third party or internally. The primary advantages of a private cloud are its high security, privacy, and total control; nevertheless, its high cost and constrained scalability are regarded as drawbacks.

A specific group with common issues is created by sharing infrastructure among a few groups in a community cloud. The associations or a third party may be in charge of managing it [6].

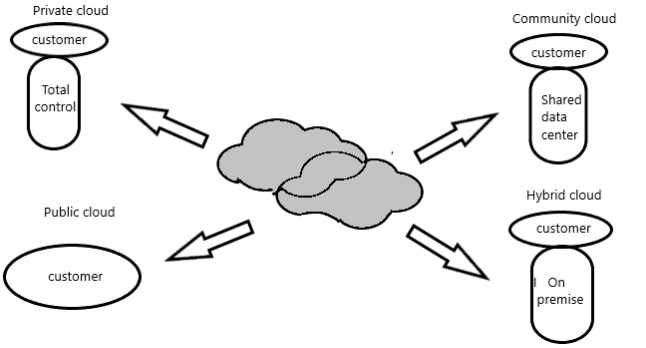


Figure 3 Cloud Computing Development Models

Many clouds of any kind make up the infrastructure of a hybrid cloud, but the clouds can transfer data and/or apps between them thanks to their interfaces. A hybrid of public and private clouds may be used to accommodate both the need to provide cloud services and the need for an organization to keep some data on file. (for instance, blasting clouds to distribute burden among them) [7].

3. Resource Allocation Strategies

Any system that seeks to ensure that physical and/or virtual resources are appropriately distributed to cloud users is referred to as a resource allocation strategy, or RAS. Reductions in resource conflict, scarcity, fragmentation, over- and under-provisioning, and resource instability result from this. The following section provides an overview of the many parameters that impact the resource allocation method that is being used.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

3.1 Methods of Linear Scheduling

The two most important scheduling techniques in linear scheduling are FIFO and LIFO. Abirami and Ramanathan [8] introduce Linear Scheduling for Tasks and Resources (LSTR), a scheduling technique that uses the application of scheduling for task and resource processing, respectively. It creates an IaaS cloud environment by attaching Cumulus [9] and Nimbus [10] services to a server node. KVM/Xen virtualization is used, and LSTR scheduling is used to distribute resources. The scheduler might perform dynamic allocation in response to requests for extra resources while continuously assessing the threshold value. A threshold value is used to gather the resource requests and classify them into distinct queues. The VMs then fulfill the requests.

3.2 Digital Assistant

Virtualization is the act of creating a simulation of a physical component, like a CPU, software program, or storage device. The resource is partitioned by the cloud into one or more execution environments [11]. The virtual system-based framework is prepared for development as a tangible component. Numerous benefits come with virtual machine migration, including proactive failure prevention, load balancing, server solidification, and online maintenance.

A plan that advocates for the use of virtualization technology to support green computing, maximize the amount of servers in use, and dynamically distribute available resources is put out [12]. To understand the variation in a server's multi-dimensional resource use, the authors present the notion of "skewness." In order to mix various workload types and enhance the overall usage of server resources, they attempt to reduce the skewness value. A collection of heuristics has been created to save energy consumption and avoid system overload.

An algorithm is proposed by Mofolo and Suchithra [13] to minimize the number of migrations and the time it takes to place VMs. The method treats the VM placement problem as a bin packing problem, where the actual servers are represented by the bins, the items by the virtual machines to be assigned, and the size of the bins by the available CPU power of those nodes. The virtual machine is implemented using the CloudSim [14]. In order to schedule virtual machines (VMs), Virtual Machine Manager (VMM) [15] was developed. By submitting the task requirements, it requests resources from the resource provider. The resource owner then gives the resource provider authorization to use the resources after the resource provider confirms with them that they are available. Additional resources for virtual machine

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

construction are made available by the resource supplier. By taking performance factors into account, tasks are completed faster and with less expense. The study to enhance CloudSim established the new resource allocation (ERA) approach in the VMM allocation policy and provider policies [16] and aid resource providers in making decisions.

3.3 Nature Inspired Optimization Methods

Approaches that draw inspiration from biology use animal behavior modeling to solve optimization challenges. It covers firefly and eagle tactics, as well as ant and bee colonies. An investigation utilizing ant colony optimization has been suggested [16]. The network bandwidth, service completion time, system dependability, and expenses are used by the authors to classify tasks based on quality of service (QoS). The experimenters varied the job count from 20 to 100, the eight is the calculated number of nodes. They created a wide gap in the QoS attribute of the node setup, mostly including the CPU, memory, and network bandwidth, in order to demonstrate differentiation.The ant colony optimization methodology completed all of the tasks more quickly than the general method as the number of tasks increased. The random distribution algorithm was applied ten times, while the ant colony optimization algorithm was applied ten times. When the task amount is less, such as twenty, the implementation effect of this methodology is less obvious because to the ant colony method, which chooses the target path based on pheromone strength. However, the execution times of two algorithms were close to one second when the job quantity reached 80.

In the study ABC Bee's Algorithm [17], Hussain and Mishra define a scheduler which selects the task with the lowest memory, input-output, and CPU needs.. Finding the right location is the equivalent of this job being a scout bee. The location of the task's current resource requirement is where the scout job is despatched. Utilizing a fitness function, the scout task locates the spot. This fitness function does the task in a specific instance to determine if it depends on memory or a processor. Fit is the performance of a given task using the resources allotted to it. After locating and identifying the resource, the search job returns to the scheduler and runs the waggle algorithm. An analysis of the Waggle function's planned tasks shows that

3.4 Methods with a Priority

A resource allocation paradigm that assigns a priority to various user requests is presented by Gouda, Radhika, and Akshatha [18]. Different duties are included in each request. Different characteristics, including

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

time, processor request, importance, and price, are taken into account for each task. The term "time" describes the amount of computational time required to finish a specific job. The number of processors required to complete the operation is referred to as the processor request. A task will be completed faster the more processors there are. Regardless of the user's age or status as a new customer, importance describes how significant they are to a cloud administrator. The cost incurred by cloud administrators on behalf of cloud consumers is the final pricing parameter. Considering all of the previously mentioned factors and the bin packing algorithm, priority.

An additional approach that is prioritized according to the type of application in Truong Huu and Montagnat's study . Strategies for allocating virtual infrastructure are designed for workflow-based applications, in which resources are assigned according to the application's workflow representation. An execution schedule estimate for workflow-based applications can be generated by interpreting and using the application logic. This makes it easier for the user to project how many resources will be used exactly when the application runs. There are four methodologies that are used to plan and distribute computing tasks: Naive, FIFO, Optimized, and Services Group Optimization.

3.5 Gossip Protocol Based Methods

Wuhib and Stadler [20] have developed a gossip-based paradigm for resource allocation in big cloud settings. The proposed model has been described as a dynamic assembly of nodes that gathers the physical nodes, or machines, of the cloud environment. Every node has a fixed amount of RAM and CPU capacity. The model uses a distributed technique for splitting cloud resources to a group of apps with time-dependent memory demands in order to dynamically maximize a global cloud usefulness. As memory demand is less than cloud memory available, the model generates optimal allocation, according to the experimental simulated results, and the quality of the allocation remains constant across machine and application counts. To make the resource allocation strategy resilient to machine failure across several clusters and datacenters, however, this effort needs more functionality. With the middleware layer components operating on each cloud environment server, Yanggratoke, Wuhib, and Stadler [19] suggest a decentralized architecture. In order to reduce power consumption through server integration and satisfy a shifting demand pattern, they design an instantiation of the generic gossip protocol.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

3.6 Service Level Agreement (SLA)

A contract outlining the services that a service provider will offer to a customer is called a service-level agreement (SLA). Although they were first used by network service providers, SLAs are now frequently utilized by cloud computing and telecommunications service providers [20]. a suggested mechanism for allocating resources to SaaS providers in order to reduce infrastructure costs and SLA violations. By mapping user requests to infrastructure level attributes, they verify that SaaS providers can manage the variety of virtual machines and the dynamic change of clients. CloudSim [21] is used to simulate the cloud computing environment that implements the recommended resource allocation procedures. Customers' and SaaS providers' perspectives are taken into account while measuring performance. Noted the number of SLA violations from the consumers' point of view.

3.7 Auction System

The Fujiwara thesis [4] describes a combinatorial auction-based marketplace technique for cloud computing services that allows users to reserve any combination of services at preferred timeslots, pricing, and service quality. Combinatorial Simple Virtual Market Protocol, or CombiSVMP, is a unique protocol designed to make information sharing between the marketplace server and the participating agents easier. The seller agents represent cloud computing service providers, and the buyer agents represent cloud computing service users. To assess the marketplace designs, three experiments have been conducted.

The findings demonstrated that the forward/combinatorial design yields the highest worldwide usage together with the most suitable cost-performance ratio and completion rate for the clients. A dynamic transaction process for cloud resource distribution is described by Wei-Yu Lin, Guan-Yu Lin, and Hung-Yu Wei [22] by creating a real-time model with two periods, n cloud users, and a cloud service provider (CSP). The two jobs assigned to the CSP are dynamic resource distribution to cloud users and time-insensitive background computing. The CSP will gain a set amount of value if the overall input into the background task exceeds the threshold. Additionally, the cloud users will purchase the CSP's leftover resources.

Following the decision to assign resources to the background activity.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

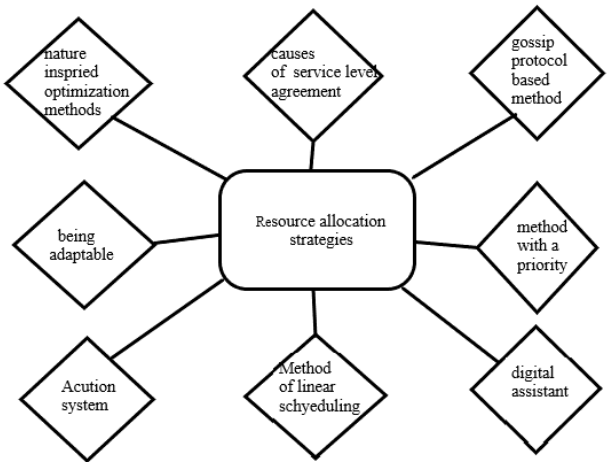


Figure 4. Applied RAS Summary

3.8 Other Methods

Rarely used techniques are included in this section. This method provided a procedure that divides cloud clusters according to the quantity and kind of computational, data storage, and communication resources they manage, such as a hardware resource dependency [23]. Each server allots these resources. Generalized Processor Sharing (GPS) is used to cluster and distribute different types of server resources, such as disk resources, based on the fixed utilization of each client. Using a greedy technique to determine the suitable starting solution and paralleling the solution, this procedure minimizes the decision time using distributed decision making. Allocating resources differently could enhance the solution. However, this approach is unable to accommodate significant parameter changes.

4. CONCLUSION

This summary highlights the diverse functions of RAS in the cloud. Cloud resource throughput, utilization, response time, and latency will all be impacted by the RAS cloud provider choice. Additionally, reducing response times, avoiding resource fragmentation, and avoiding both over and under provisioning are our top priorities.

References

[1] National Institute of Standards and Technology, "The NIST Definition of Cloud Computing", Computer Security Division, Information Technology Laboratory (2011).

[2] Linan Zhu, Qingshui Li, and Lingna He, " Study on Cloud Computing Resource Scheduling Strategy Based on the Ant Colony Optimization Algorithm", I International Journal of Computer Science, September 2012.

[3] Bhaskar Prasad, Eunmi Choi and Ian Lumb, "A Taxonomy and Survey of Cloud Computing Systems", Fifth International Joint Conference on INC, IMS and IDC, 2009

[4] Ikki Fujiwara," Study on Combinatorial Auction Mechanism for Resource

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

Allocation in Cloud Computing Environment", Ph.D. thesis 2012.

[5] K C Gouda, Radhika T V and Akshatha M," Priority based resource allocation model for cloud computing", International Journal of Science, Engineering and Technology Research, January 2013.

[6] Tram Truong Huu and John Montagnat, "Virtual Resource Allocations distribution on a cloud infrastructure", IEEE, 2010.

[7] Wei-Yu Lin, Guan-Yu Lin and Hung-Yu Wei," Dynamic Auction Mechanism for Cloud Resource Allocation", IEEE/ACM 10th International Conference on Cluster, Cloud and Grid Computing.

[8] Linlin Wu, Saurabh Kumar Garg and Rajkumar Buyya, "SLA-based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments", IEEE Computer Society 2011.

[9] Abirami S.P. and Shalini Ramanathan, "Linear Scheduling Strategy for Resource Allocation in Cloud Environment", International Journal on Cloud Computing: Services and Architecture, February 2012.

[10] Javed Hussain and Durgesh Kumar Mishra," An Efficient Resource Scheduling In Cloud Using Abc Algorithm", International Journal of Computer Engineering and Applications, June 201

[11] Z.Gong, X. Gu, and J. Wilkes, "PRESS: PRedictive ElasticReSource Scaling for cloud systems," in CNSM 2010, October, pp. 9 –16.

[12] D. Meisner, B. T. Gold, and T. F. Wenisch, "PowerNap: elimi-nating server idle power," SIGPLAN Not., vol. 44, pp. 205–216, March 2009.

[13] Hewlett-Packard, Intel, Microsoft, Phoenix Technologies Ltd.,Toshiba Corporations, "Advanced configuration and power in-terface specification," 2010.

[14] D. Carrera, M. Steinder, I. Whalley, J. Torres, and E. Ayguade,"Utility-based placement of dynamic web applications with fair-ness goals," in IEEE NOMS, April 2008, pp. 9 –16.

[15] C. Tang, M. Steinder, M. Spreitzer, and G. Pacifici, "A scalable application placement controller for enterprise data centers," in WWW2007. New York, NY, USA: ACM, 2007, pp. 331–340.

[16] R. Yanggratoke, F. Wuhib, and R. Stadler, "Gossip-based resource allocation for green computing in large clouds (long version)," KTH Royal Institute of Technology,

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 1, No 1, 2024. ISSN: 2348-6600
Reference ID: IJCS-492 PAGE NO: 3377-3388

https://eeweb01.ee.kth.se/upload/publications/reports/2011/

[17] TRITA-EE 2011 036.pdf, Tech. Rep. TRITA-EE 2011:036,April 2011.

[18] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in INFOCOM, vol. 1, 1999, pp. 126 –134.

[19] G. Jung, M. Hiltunen, K. Joshi, R. Schlichting, and C. Pu, "Mis-tral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures," in ICDCS2010, 2010, pp. 62 –73.

[20] B. Speitkamp and M. Bichler, "A mathematical programmingapproach for server consolidation problems in virtualized data centers," IEEE TSC, vol. 3, no. 4, pp. 266 –278, 2010.

[21] N. Tolia, Z. Wang, P. Ranganathan, C. Bash, M. Marwah, and X. Zhu, "Unified thermal and power management in server en-closures," ASME Conference Proceedings, vol. 2009, no. 43604, pp. 721–730, 2009.

[22] M. Cardosa, M. Korupolu, and A. Singh, "Shares and utilities based power consolidation in virtualized server environments," in IM 2009, pp. 327 –334.

[23] F. Wuhib, R. Stadler, and M. Spreitzer, "Gossip-based resource management for cloud environments," in CNSM 2010, October, pp. 1 –8.