DOI: 10.53469/jtpes.2024.04(06).02

Cloud Computing and Hyperscale Data Centers: A Comparative Study of Usage Patterns

Yu Ma[1], Zhixuan Shen[2], Jinnian Shen[3]

[1] Northeastern University, Computer Science, [2] Northeastern University, Computer Science, [3] Northeastern University, Information Systems,
[1]Portland, Main, 04101, USA [2]Portland, Maine, 04101, USA. [3] Oakland, California, 94613, USA
[1]writingcstech@gmail.com, [2]szxuan114@gmail.com, [3]shenjinnian@126.com

Abstract: ***The rapid growth of cloud computing and hyperscale data centers has transformed the way we store, process, and manage data. Understanding usage patterns is crucial for optimizing resource allocation, performance, security, and costs in these environments. This paper explores the similarities and differences between cloud computing and hyperscale data centers, highlighting their implications for users and providers. We discuss the importance of understanding.***

Keywords: Cloud Computing, Hyperscale Data Centers, Usage Patterns, Resource Utilization, Scalability, Cost Models.

1. INTRODUCTION

In recent years, cloud computing and hyperscale data centers have revolutionized the way businesses and organizations manage their computing resources. Cloud computing offers on-demand access to a shared pool of configurable computing resources, while hyperscale data centers provide the infrastructure necessary to support massive amounts of data processing and storage.

Understanding the usage patterns within these environments is crucial for several reasons. Firstly, it provides insights into how organizations leverage these technologies to meet their computing needs efficiently. Secondly, it helps identify trends and preferences among users, which can inform the development of new services and technologies. Lastly, studying usage patterns allows for comparisons between different approaches to computing, such as traditional data centers versus cloud-based solutions.

The objectives of this study are twofold. Firstly, we aim to provide a comprehensive overview of usage patterns in both cloud computing and hyperscale data centers. This includes examining common applications, user demographics, and adoption trends within each environment. Secondly, we seek to conduct a comparative analysis to highlight similarities and differences in usage patterns between the two technologies.
By achieving these objectives, this study aims to contribute to a better understanding of how organizations utilize cloud computing and hyperscale data centers to meet their computing needs. Additionally, it seeks to provide insights that can inform decision-making processes related to the adoption and implementation of these technologies.

2. CLOUD COMPUTING SERVIES

Cloud Computing is the delivery of computing where massively scalable IT-based capabilities are provided, as a service across the internet clients. This term effectively focuses on the different aspects of the Cloud Computing paradigm which can be found at different levels of infrastructure. There are three types of services provided by cloud computing architecture namely; SaaS, PaaS, and IaaS. Each type of service serves different purposes and different customers, they rent out the use of their computing resources such as services, applications, infrastructures, and platforms to customers.

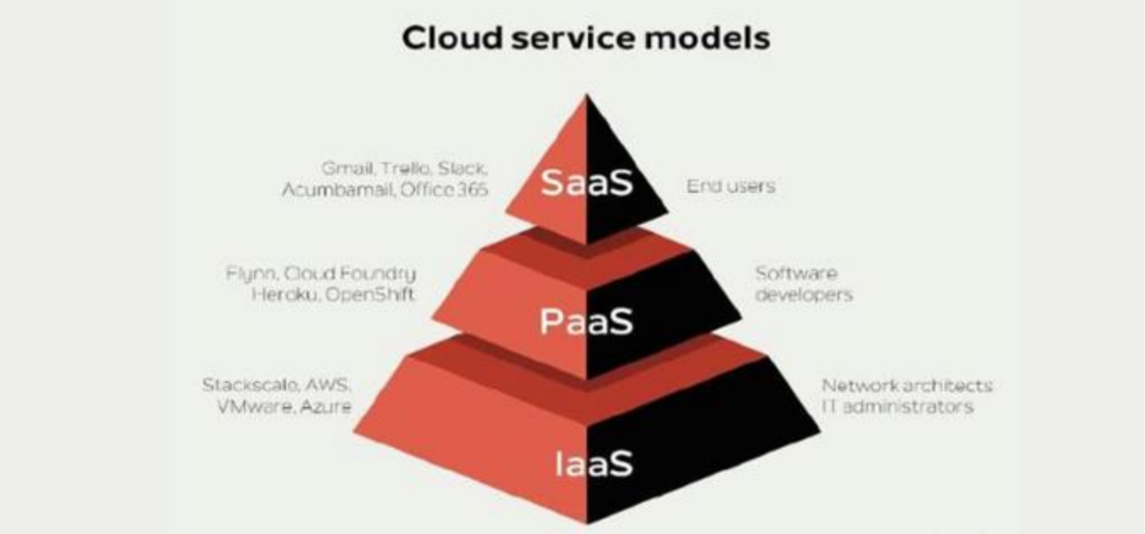
Cloud service models
Gmail, Trello, Slack, Acumbamail, Office365
SaaS
End users
Flynn, Cloud Foundry Heroku, OpenShift
PaaS
Software developers
Stackscale, AWS, VMware, Azure
IaaS
Network architects IT administrators

Figure 1: A comprehensive picture of data center energy usage modeling

2.1 Data center power modeling at the individual software level

RQ1 what are the various approaches used at the software level like operating systems, virtualization, and applications to reduce the usage of power by data centers?

2.2 Operating system level

The operating system is placed between two layers: the application and the hardware. The main role of applications is to create the resource demand and the OS job is to manage the resources for all these applications. The main component that consumes power is the physical hardware but it is very essential to keep a check on the events that consume power at the operating system level if energy usage optimization at data center is to be done at the software level too. The power usage breakdown of the operating system functions is shown in Figure 3. Data-path and pipeline topologies that allow for numerous problems and out-of-order execution were found to squander 50% of the total power of the OS processes investigated. Furthermore, the clock consumes 34% power and different levels of cache consume the remaining power.

Operating System Power Management (OSPM) is a mechanism utilized by OS to manage the power of the underlying platform and transition of it between different power modes. OSPM allows a platform or system to adopt the most efficient power mode and applies to all devices and components inside the platform/system. OSPM is also known as OS-directed configuration and Power Management.

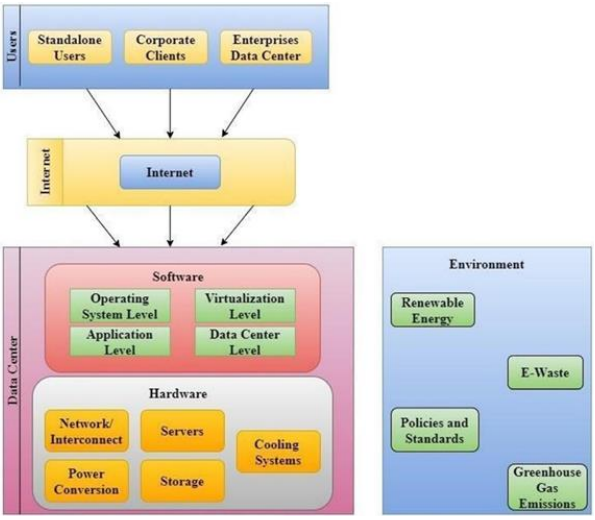

Figure 2: A comprehensive picture of data center energy usage modeling

The trade-off between quality and power efficiency has been intensively examined and analyzed since control over running voltage and energy management has been largely shifted from the hardware and firmware level to the operating system. Herzog et al. offer PolaR, a method for automatically determining energy-efficient setups, as well as a Linux implementation. PolaR proactively chooses optimal settings by integrating application profiles and system-level data, and no application adjustments are required. They take into account bank shots (configuration settings unrelated to power management) in addition to properly controlling the system. OS development teams recognized the value of energy as a resource on par with time. With energy seen as just another resource available to the operating system, operating system inter-nails (such as locking mechanisms) were changed to accommodate this new perspective to produce energy-aware operating systems. Scordino et al. illustrate how the deadline scheduler and the subsystem may be changed to relax the restriction that the frequency scaling technique is used only when no real-time processes are running and to create an energy-aware real-time scheduling approach. They described the architectural issues they encountered when trying to deploy the GRUB- PA algorithm on a real OS like Linux. Experiment findings on a multi-core ARM architecture demonstrated the efficacy of their suggested solution.

With the advancement of semiconductor and software technologies, the capabilities of an embedded system have grown by incorporating new features and performance. In recent years, the network has also advanced as communication infrastructure and contact with server systems have become essential. So far, TCP / IP connections between servers and embedded devices have been established by two methods. The first is a technique that includes a TCP/ IP stack in embedded devices. The second is a technique of communicating via a ‘‘gateway’’ (to translate end-device communications). There are several server system com- position options, such as putting a server in-house, establishing a server at a data center outside of town, and utilizing cloud computing. Smaller, more widespread, and less well-known ‘‘embedded data centers’’ consume half of all data center energy or about 1% of all energy generated in the United States. In general, embedded data centers are data center facilities that have less than 50 kW of IT demand. Server rooms, server closets, localized data centers, and several mid-tier data centers are among them. Energy harvesting technologies based on rechargeable batteries are a popular option for addressing the issue of delivering continuous power to deeply implanted devices such as wireless sensor nodes. However, if the use of a node is not carefully planned, the battery may be depleted too quickly, making the continuous operation of such a device unfeasible. To regulate the flow of energy, an energy-management solution is necessary. Presented an idea that enables the modeling of hardware energy usage and the creation of energy-aware device drivers for the embedded OS. Their drivers can account for the energy usage of each driver function call with greater than 90% accuracy. Similarly, presented Tock, a unique embedded OS for low-power systems that utilizes the limited hardware-protection processes accessible to the latest microcontrollers and type-safety functions- ties of the Rust programming language to offer a multiprogramming ecosystem that provides software fault separation, memory protection, and efficient memory governance for dynamic applications and services written in Rust. Low-power embedded operating systems frequently use the same memory areas for both applications and the operating system. Merging applications and the kernel allows them to easily exchange references and gives efficient procedure call access to low-level functionality. This monolithic method often necessitates building and installing or upgrading a device’s apps and operating system as a single unit.

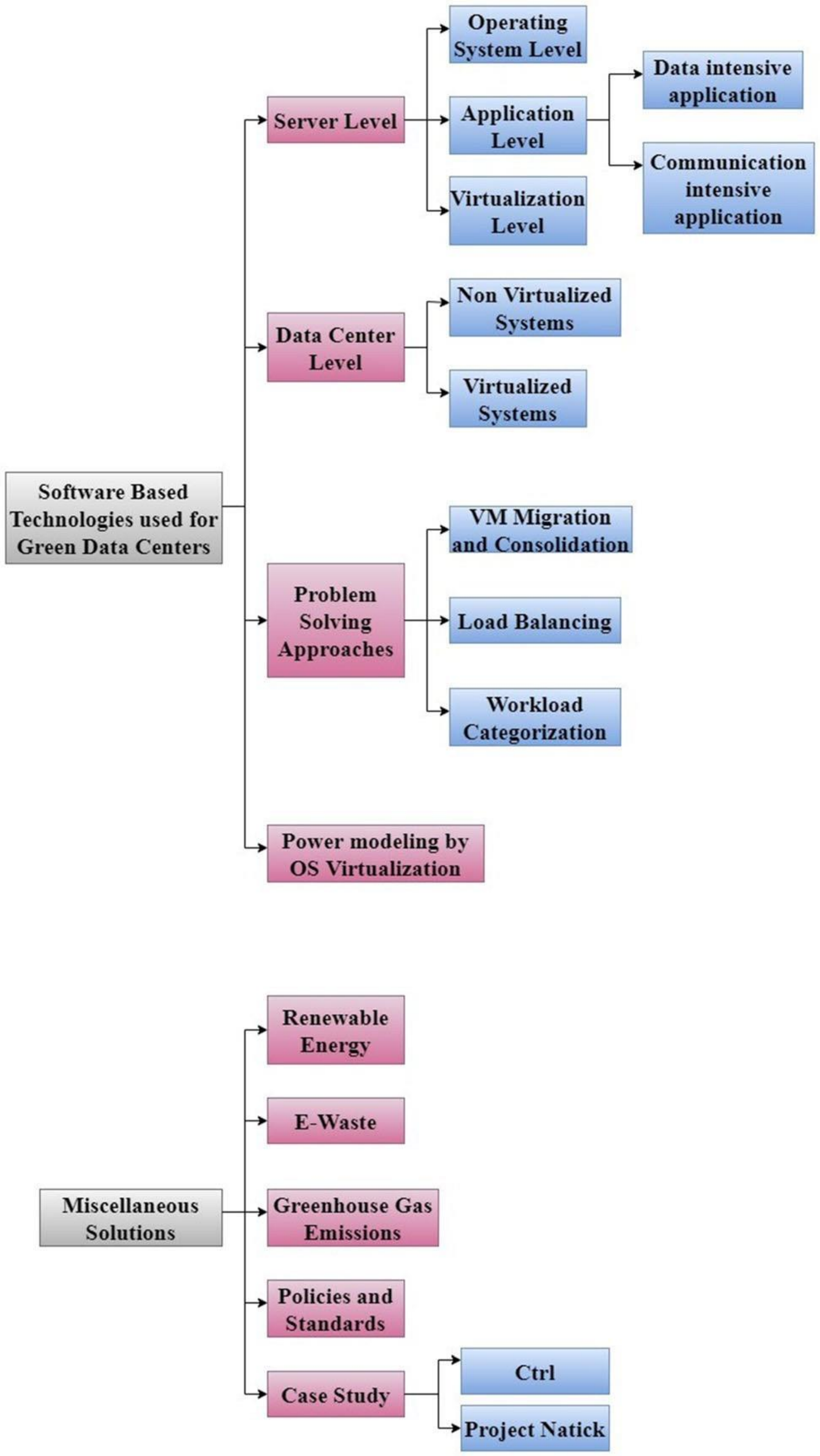

Figure 3: A systematic summary of data centre power demand prediction at the software level

2.3 Virtualization level

Virtualization uses software to construct a layer of abstraction above computer equipment, enabling the actual features of a single computer, storage, disk, and so on—to be separated into numerous virtual computers, also called as virtual machines (VMs). Each virtual machine created for a user can be allocated an individual operating system on a single physical machine that makes sure of the per- performance of the virtual machines and failure isolation among them. Hence, a Virtual Machine Monitor (VMM) / Hypervisor is responsible for multiplexing of resources to the virtual machine and helps in the management of the power to perform efficient operations. The two ways in which a virtual machine monitor can take part in the management of power:

- A VMM acts as a power-aware operating system. It verifies the entire performance of the system and applies the DVFS (Dynamic Voltage and Frequency Scaling) or any DCD (Dynamic Component Deactivation) techniques to the components of the system.
- The other way is to leverage the policies for the management of power and knowledge of applications at the OS level. Power management calls can be mapped from different virtual machines. In addition, coordinated system-wide limits on the power can be enforced.

Virtualization technology has regained prominence in computer system architecture during the last few years. Virtual machines (VMs) provide a development route for adding new capabilities—for example, server consolidation, transparent migration, and secure computing—into a system while maintaining compatibility with existing operating systems (OSs) and applications. Multiple VMs executing on the same core in contemporary virtualized settings must adhere to a single management of power controlled by the hypervisor. These settings have different limitations. It does not enable users to specify a desired power control scheme for each virtual machine (or client). Second, it frequently affects the energy efficacy of some or all VMs, particularly when the VMs need competing energy management strategies. To mitigate the above problems, a per-VM power control method that enables each VM's guest OS to utilize its chosen energy administration strategy and prevent similar VMs from competing with each other's energy control strategy. When compared to the Xen hypervisor's default on-demand governor, Virtual performance (VIP) minimizes power usage and enhances the completion time of CPU-intensive applications by up to 27% and 32%, respectively, without breaching the SLA of latency-sensitive implementations. Furthermore, Xiao et al. Examined the VM scheduling model and the I/O virtualization paradigm in terms of energy-efficiency optimization. They provided a power-fairness credit sequencing approach with a novel I/O offset method to achieve speedy I/O performance while simultaneously raising energy conservation. Apart from this, VM resource calibration. They created a system to reduce the energy usage of virtual servers by utilizing controlled feedback architecture as well as power monitoring services.

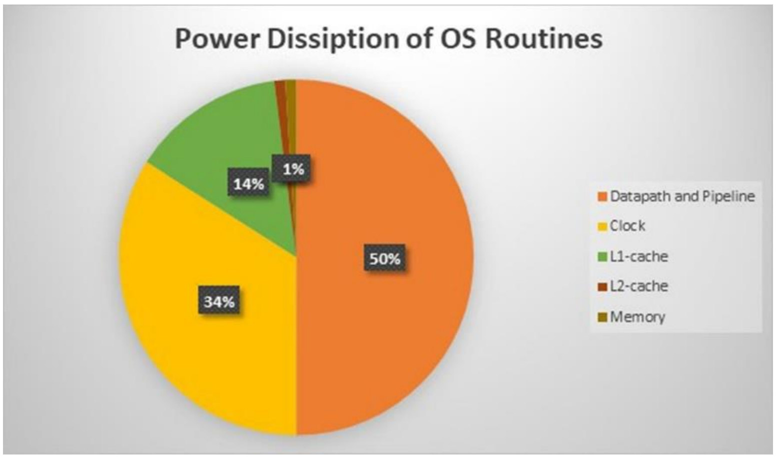

Figure 4: Power dissipation of OS routines

3. Comparative ANALYSIS

Comparison of Cloud Computing and Hyperscale Data Center Usage Patterns.

3.1 Cloud Computing

Cloud computing offers a range of services and deployment models, catering to diverse user needs and preferences. The following aspects highlight the usage patterns commonly observed in cloud computing environments:

Table 1: Highlight the usage patterns

Aspect	Description
common Application	Software as a Service (SaaS) applications for productivity, communication, and collaboration. Platform as a Service (PaaS) offerings for software development and deployment. Infrastructure as a Service (IaaS) for scalable computing resources.
User Demographic	Small to medium-sized enterprises (SMEs) leveraging cloud services for cost-effective IT solutions. Large enterprises utilizing cloud resources for agility and scalability. Individuals and startups benefiting from pay-as-you-go models.
Adoption Trends	Increasing adoption of cloud services across various industries due to cost savings and flexibility. Growth in hybrid cloud deployments, combining private and public cloud resources. Shift towards multi-cloud strategies to avoid vendor lock-in and enhance resilience.

3.2 Hyperscale Data Centers

Hyperscale data centers represent a distinct approach to data processing and storage, characterized by their massive scale and efficiency. The following aspects highlight the usage patterns commonly observed in hyperscale data center environments:

Table 2: Common Usage Patterns in Hyperscale Data Centers

Aspect	Description
Workload Types	High-performance computing (HPC) workloads for scientific research, simulations, and modeling. Big data analytics for processing and analyzing large datasets in real-time. Content delivery networks (CDNs) for efficient distribution of multimedia content.
User Demographic	Tech giants and large-scale enterprises rely on hyperscale data centers for their computational needs. Industry-specific applications in finance, healthcare, and gaming, require massive processing power and storage capacity. Government agencies and research institutions leverage hyperscale infrastructure for data-intensive projects.
Adoption Trends	Rapid scalability to accommodate fluctuating workloads and handle peak demands efficiently. Resource optimization to ensure cost-effectiveness while meeting performance requirements. Data center design and management strategies to achieve high-density computing and energy efficiency.

4. ANALYSIS OF DIFFERENCES

Infrastructure: Cloud computing relies on virtualized resources managed by external providers, offering flexibility and ease of use. In contrast, hyperscale data centers are owned and operated by large tech companies, and designed to handle vast amounts of data with high efficiency.

Scalability: Both environments provide excellent scalability, but cloud computing focuses on elastic scaling to quickly adjust to demand, while hyperscale data centers are built for massive, ongoing scalability to support extensive data processing needs.

Security: Security in cloud computing is a shared responsibility between the provider and the user, often depending on the service model (IaaS, PaaS, SaaS). Hyperscale data centers typically implement robust, centralized security measures, benefiting from economies of scale and proprietary technologies.

Cost: Cloud computing offers a pay-as-you-go model, making it accessible for smaller entities and those needing flexible resource management. Hyperscale data centers, while requiring significant upfront investments, achieve cost efficiency through economies of scale over time.

Common Applications: Cloud computing is versatile, supporting a wide range of applications from web hosting to big data analytics. Hyperscale data centers are optimized for high-performance computing, extensive web services, and supporting cloud platforms at a massive scale.

User Demographics: Cloud computing attracts SMEs, startups, and individual developers due to its accessibility and scalability. Hyperscale data centers primarily serve large enterprises and tech giants, providing the backbone for large-scale operations and cloud services.

Adoption Trends: Cloud computing continues to grow rapidly as organizations seek flexible, scalable solutions. Hyperscale data centers are expanding in response to the increasing demands of big data, artificial intelligence, and the Internet of Things.

Performance: Performance in cloud computing can vary widely based on the provider and service plan. Hyperscale data centers, by contrast, are designed to deliver consistently high performance, optimized for large-scale, high-demand environments.

Management: Cloud services are managed by the providers, requiring minimal management from the user. Hyperscale data centers, however, demand in-house expertise for their setup and ongoing maintenance, reflecting their complexity and scale.

Compliance: Compliance offerings in cloud computing vary by provider, with many ensuring adherence to major standards. Hyperscale data centers, due to their role in critical operations and serving high-profile clients, maintain strict compliance with industry regulations.

This comparison highlights the distinct characteristics and usage patterns of cloud computing and hyperscale data centers, providing valuable insights for organizations considering these technologies.

4.1 Comparison of Cloud Computing and Hyperscale Data Center Usage Patterns

To comprehensively compare the usage patterns of cloud computing and hyperscale data centers, we will focus on key aspects such as infrastructure, scalability, security, cost, and common applications. Table 3 provides a structured comparison:

Table 3: provides a structured comparison:

Aspect	Cloud Computing	Hyperscale Data Centers
Infrastructure	Virtualized resources, managed by third-party	providers Large-scale, highly automated data centers owned by major tech firms
Scalability	Elastic scaling; resources can be quickly scaled up or down	Massive scalability to handle large volumes of data and traffic
Security	Shared responsibility model; provider and user share security duties	Robust security measures, often proprietary; centralized control

Cost	Pay-as-you-go pricing; operational expenses	Economies of scale; significant upfront investment, lower long-term cost per unit
Common Applications	Web hosting, SaaS, disaster recovery, big data analytics Web hosting, SaaS, disaster recovery, big data analytics	High-performance computing, large-scale web services, cloud services
User Demographics	SMEs, startups, individual developers	Large enterprises, tech giants, cloud service providers
Adoption Trends	Rapid adoption due to flexibility and lower entry barriers	Growing with the rise of big data, AI, and IoT
Performance	High variability depending on provider and plan	Consistently high performance, optimized for large-scale operations
Management	Managed by cloud service providers; minimal user management	Requires in-house expertise for setup and maintenance
Compliance	Varies by provider; many offer compliance with major standards	Strict compliance due to high-profile clients and critical operations

5. CONCLUSION

In conclusion, this study has explored the similarities and differences between cloud computing and hyperscale data centers, highlighting their implications for users and providers. The analysis revealed that while both share some similarities, they have distinct differences in workload distribution, resource utilization, scalability, and cost models. Understanding these differences is crucial for optimizing resource allocation, performance, security, and costs.

The study also emphasized the importance of understanding usage patterns in cloud computing and hyperscale data centers, enabling optimized resource allocation, improved performance, enhanced security, and better capacity planning. Furthermore, it highlighted future directions for research and development in these areas, including edge computing, artificial intelligence, machine learning, serverless computing, and quantum computing.

In summary, this study provides a comprehensive analysis of cloud computing and hyperscale data centers, offering valuable insights for users, providers, and researchers. Its findings and recommendations can help optimize the use of these technologies, driving innovation, efficiency, and sustainability in the digital landscape.

REFERENCES

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.

[2] Buyya, R., Broberg, J., & Goscinski, A. (Eds.). (2011). Cloud computing: principles and paradigms. John Wiley & Sons.

[3] Gartner. (2020). Gartner Forecasts Worldwide Public Cloud Revenue to Grow 6.3% in 2020. Retrieved from https://www.gartner.com/en/newsroom/press-releases/2020-04-02-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-6point3-percent-in-2020

[4] Hamilton, J. (2009). Internet-Scale Service Efficiency. Keynote presentation at the 8th USENIX Symposium on Operating Systems Design and Implementation (OSDI), San Diego, CA.

[5] IDC. (2020). Worldwide Quarterly Cloud IT Infrastructure Tracker. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS46779620

[6] Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. National Institute of Standards and Technology, 53(6), 50.

[7] Sharma, R., & Bhardwaj, S. (2018). Hyperscale Data Centers: A Review of Technical Challenges, Business Models, and Recent Developments. IEEE Access, 6, 65868-65881.

[8] Varia, J. (2010). Cloud architectures. In Amazon web services: Overview of security processes (pp. 1-10).

[9] Vaquero, L. M., Rodero-Merino, L., Caceres, J., & Lindner, M. (2009). A break in the clouds: towards a cloud definition. ACM SIGCOMM Computer Communication Review, 39(1), 50-55.

[10] Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications, 1(1), 7-18.