



Article Efficient Resource Utilization in IoT and Cloud Computing

Vivek Kumar Prasad ¹, Debabrata Dansana ²⁽¹⁾, Madhuri D. Bhavsar ¹, Biswaranjan Acharya ^{3,*}, Vassilis C. Gerogiannis ⁴ and Andreas Kanavos ^{5,*}

- ¹ Department of CSE, Nirma University, Ahmedabad 382481, India; vivek.prasad@nirmauni.ac.in (V.K.P.); madhuri.bhavsar@nirmauni.ac.in (M.D.B.)
- ² Department of Computer Science, Rajendra University, Balangir 767002, India; ddansana@rajendrauniversity.ac.in
- ³ Department of Computer Engineering—AI and BDA, Marwadi University, Rajkot 360003, India
- ⁴ Department of Digital Systems, University of Thessaly, 41500 Larissa, Greece; vgerogian@uth.gr
- ⁵ Department of Informatics, Ionian University, 49100 Corfu, Greece
- Correspondence: biswaranjan.acharya@marwadieducation.edu.in (B.A.); akanavos@ionio.gr (A.K.)

Abstract: With the proliferation of IoT devices, there has been exponential growth in data generation, placing substantial demands on both cloud computing (CC) and internet infrastructure. CC, renowned for its scalability and virtual resource provisioning, is of paramount importance in e-commerce applications. However, the dynamic nature of IoT and cloud services introduces unique challenges, notably in the establishment of service-level agreements (SLAs) and the continuous monitoring of compliance. This paper presents a versatile framework for the adaptation of e-commerce applications to IoT and CC environments. It introduces a comprehensive set of metrics designed to support SLAs by enabling periodic resource assessments, ensuring alignment with service-level objectives (SLOs). This policy-driven approach seeks to automate resource management in the era of CC, thereby reducing the dependency on extensive human intervention in e-commerce applications. This paper culminates with a case study that demonstrates the practical utilization of metrics and policies in the management of cloud resources. Furthermore, it provides valuable insights into the resource requisites for deploying e-commerce applications within the realms of the IoT and CC. This holistic approach holds the potential to streamline the monitoring and administration of CC services, ultimately enhancing their efficiency and reliability.

Keywords: cloud computing; cloud service provider; compliance monitoring; e-commerce; efficient resource utilization; IaaS; Internet of Things (IoT); policy-driven approach; service-level agreements (SLAs)

1. Introduction

In today's world, Internet of Things (IoT) devices have become increasingly pervasive, finding applications across various domains. These devices, equipped with sensors and communication tools, gather and transmit vast amounts of data from the physical environment to digital networks. Their uses span industrial automation, healthcare monitoring, smart home systems, and environmental sensing. However, managing and processing the immense data streams generated by IoT devices pose significant challenges. This is where the integration of cloud computing (CC) comes into play. CC offers a scalable and adaptable platform for handling and storing IoT data. By leveraging the capabilities of the cloud, organizations can analyze these data in real time, extract actionable insights, and make data-driven decisions. This symbiotic relationship between IoT devices and CC underscores the synergy between cutting-edge technologies, paving the way for a plethora of innovative solutions with substantial potential for research and development in security, communication, and AI.

The transformative impact of Internet of Things (IoT) technologies on consumer behaviors and enterprise operational models is increasingly evident. This phenomenon is



Citation: Prasad, V.K.; Dansana, D.; Bhavsar, M.D.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A. Efficient Resource Utilization in IoT and Cloud Computing. *Information* 2023, *14*, 619. https://doi.org/ 10.3390/info14110619

Academic Editor: Vladimír Bureš

Received: 28 September 2023 Revised: 8 November 2023 Accepted: 15 November 2023 Published: 19 November 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). catalyzed by the reduction in device deployment costs and the surging consumer demand, as illustrated in Figure 1. According to Gartner [1], a renowned advisory and research entity, the number of connected device installations surged from 23.14 billion units in 2018 to a projected 30.73 billion units in 2020. Such exponential growth offers an opportune landscape for various stakeholders, including investors and corporations, to amass extensive data.



Figure 1. IoT device adoption is expected to expand.

Financial projections indicate that businesses could invest nearly 5 trillion USD in expanding the IoT market and developing new applications by the end of 2021. Moreover, long-term investments in this sector are expected to surpass 100 billion USD by the mid-21st century. As the volume of devices and associated data continues to soar, the significance of sophisticated data management infrastructures, such as CC, becomes increasingly pivotal. The efficient orchestration of dynamic resource allocation within the domain of CC's Infrastructure as a Service (IaaS) is crucial for ensuring the prudent utilization of computational assets. The continuous oversight of these assets and adherence to service-level agreements (SLAs), evaluated through a set of defined quality indicators, will be instrumental in realizing the potential of adaptive resource governance.

The convergence of the IoT and CC has revolutionized the accessibility and management of resources for end users, providing unprecedented convenience and flexibility [2]. However, from the standpoint of cloud service providers (CSPs), meeting these demands necessitates robust resource management capabilities to accommodate dynamic workloads and evolving tasks. Consequently, contemporary CC systems must embody intelligence and resource abundance.

IoT devices and cloud systems play a pivotal role in managing peak workloads and facilitating the design and implementation of enterprise systems, empowering businesses to achieve their objectives. Cloud computing (CC), in particular, fosters the creation of an IT utilities marketplace commonly known as market-oriented cloud computing.

From an end user's viewpoint, CC presents an illusion of infinite resource availability, while CSPs are tasked with efficiently managing these resources while optimizing energy consumption [3]. Achieving this balance is challenging and requires the utilization of cloud monitoring and prediction techniques.

Cloud monitoring is critical to the reliability and performance of cloud-based infrastructures. It entails systematic data gathering, analysis, and visualization for the numerous elements of cloud services, such as resource use, network latency, and security events. From a third-party standpoint, cloud monitoring solutions are clearly vital for enterprises that rely on CC, as they provide critical insights into the health and effectiveness of their cloud-based applications and services. Businesses can use this technology to proactively identify and prevent problems, improve resource allocation, and maintain a high degree of service availability. Third-party observers identify cloud monitoring as an essential tool in guaranteeing the flawless functioning of cloud environments while improving the overall security and performance in an era where digital transformation is a primary goal. Public CC environments like Amazon Web Services (AWS) provide organizations with the resources to host critical services and applications [4,5]. The continuous monitoring of these cloud-hosted services is essential to ensure consistent performance throughout their operational lifespan [6].

Cloud resource prediction is an important feature of CC, as it ensures the optimal allocation of compute, storage, and network resources inside cloud settings. Based on a third party, precise resource prediction enables cloud service providers and users to optimize their infrastructure, reduce costs, and improve cloud-based applications and services' overall performance and dependability. Forecasting resource demand is an important aspect of cloud resource prediction. This includes forecasting the future resource requirements of cloud workloads based on user traffic patterns, data volumes, and application performance metrics. Advanced machine learning and data analysis techniques are frequently used to anticipate and predict these resource demands precisely. This proactive strategy enables cloud providers to dynamically assign resources, scaling up or down as needed, avoiding under- or over-provisioning, which can lead to inefficiencies and increased costs. Additionally, cloud resource prediction includes the prediction of potential resource deviations and failures. Deviations from expected resource use patterns can be recognized by continuously monitoring and evaluating system data. These variations may suggest potential breakdowns or bottlenecks in performance. Third-party observers understand the importance of these predictive skills in reducing service disruptions and assuring cloud service availability. In essence, cloud resource prediction is a critical component of intelligent cloud management, allowing both providers and customers to make informed decisions and optimize their cloud infrastructures for increased efficiency and reliability.

The forthcoming era of CC holds promise for the technology industry, as it paves the way for autonomous cloud infrastructure management, reducing the need for manual intervention [7]. The properties associated with CC will accelerate future technologies, enabling faster operations than the immediate environment.

Dynamic allocation mechanisms, such as auto-scaling techniques widely adopted by AWS, allow resources to be provisioned and de-provisioned based on current and future resource demands [8]. Quality of service (QoS) and service-level agreements (SLAs) vary for different cloud environments [9]. The challenge lies in scaling resources for distributed computational workloads worldwide.

Resource provisioning can be categorized into predictive and reactive tactics [10]. Reactive techniques respond to the system's current state, considering VM utilization and client requests. Predictive approaches, on the other hand, forecast future resource requirements, leading to better resource utilization and accurate response time estimates.

1.1. Metrics and Policies in CC

Metrics and policies in CC are critical components of effectively managing and administering cloud resources. These factors are crucial in ensuring the proper operation of cloud environments, allowing firms to align their cloud usage with business goals, security requirements, and cost efficiency.

Metrics are necessary for evaluating the performance of cloud resources. These metrics include a variety of factors, such as response times, throughput, latency, and availability.

According to third-party observers, these indicators provide a comprehensive picture of how well cloud services meet their service-level agreements (SLAs). Organizations may detect bottlenecks, optimize resource allocation, and guarantee that cloud services offer the desired level of performance to fulfill business goals by regularly monitoring and evaluating key performance data.

Cost is an important factor in CC, and cost indicators are critical for keeping track of cloud spending. These metrics monitor resource utilization, pricing structures, and usage trends. According to third-party experts, cost optimization policies driven by these indicators enable firms to decrease wasteful spending by detecting idle resources, setting budget constraints, and choosing cost-effective cloud service models. Businesses may make educated decisions regarding resource provisioning and consumption by matching cost indicators with cloud regulations.

Cloud security is critical, and security metrics are used to assess the effectiveness of security measures. These metrics include intrusion detection, access controls, and vulnerability evaluations. According to third-party assessments, security policies specify the rules and processes for protecting data and applications in the cloud. Organizations can assure compliance with industry rules and best practices by aligning security metrics with security policies, reducing the risks associated with data breaches and cyber-attacks.

Scalability is a crucial feature of cloud computing, and resource scaling measures are critical for reacting to changing workloads. Resource utilization, auto-scaling triggers, and capacity planning are examples of these measures. According to third-party experts, scalability rules drive resource allocation decisions, dictating when and how resources can be scaled up or down to meet demand while controlling costs. Properly aligned policies ensure that cloud resources can handle fluctuating workloads efficiently and without service interruptions.

Monitoring and enforcing compliance with organizational policies, industry standards, and legal requirements is part of CC governance. Governance metrics evaluate adherence to these laws and regulations, ensuring accountability and openness. Thirdparty viewpoints emphasize the significance of governance policies, which offer guidelines for data access, data preservation, and auditing methods. Organizations may maintain control over their cloud resources, enforce compliance, and show stakeholders and regulators their commitment to responsible cloud usage by aligning governance metrics with governance principles.

Furthermore, measurements and policies in CC are inextricably linked components of good cloud administration. These components enable businesses to assess and manage cloud performance, costs, security, scalability, and governance. Businesses can employ cloud resources strategically by aligning these KPIs with well-defined rules, ensuring that cloud computing corresponds with their objectives, regulatory requirements, and best practices.

Implementing auto-scaling techniques in the cloud involves using various metrics alongside policies that align with QoS parameters and SLAs, including performance metrics and thresholds [11]. Defining these parameters without human intervention presents challenges in comprehending their impact on cloud utility performance. Autonomic techniques, requiring minimal human intervention, are essential in such environments, enabling the system to make decisions based on specified metrics and policies.

The failure to define metrics can result in several issues, including the following:

- An inability to measure client resource requirements [12].
- The over- or under-provisioning of resources [13].
- Ambiguity in describing delivered work [14].
- Tedious resource monitoring and management [15].
- An inability to impose penalties for non-compliance [16].

1.2. Motivation

Efficient resource allocation and management in dynamic IoT and cloud environments are essential for optimizing system performance and minimizing resource wastage. With the proliferation of IoT devices and data, scalability becomes a critical factor, leading to the development of scalable architectures and advanced load-balancing techniques. This research article aims to address the exponential growth of IoT devices and data, ensuring optimal resource utilization while preventing performance bottlenecks. The contributions of this work provide valuable insights and solutions for researchers and practitioners focused on enhancing resource efficiency in the IoT and cloud computing.

- Efficient resource allocation in dynamic IoT and cloud environments by SLA management optimization.
- The primary aim is minimizing resource wastage while enhancing system performance.
- The need for scalability in IoT and cloud systems has spurred the development of scalable architectures and advanced load-balancing techniques.
- This research article discusses a significant role in addressing the exponential surge of IoT devices and data, guaranteeing optimal resource utilization while preventing performance bottlenecks.

1.3. Contribution

Service-level agreements (SLAs) play a pivotal role in cloud computing, shaping contract terms, negotiations, and performance metrics. This article makes several significant contributions:

- Discussion of Various SLAs: We provide a comprehensive exploration of diverse service-level agreements (SLAs) and their associated parameters. These discussions shed light on the intricate aspects of SLAs and their integral role in cloud service contracts and negotiations.
- Linking SLAs to Quality of Service (QoS): Recognizing the crucial relationship between SLAs and quality of service (QoS), we emphasize how SLAs directly impact the quality of the services provided. This linkage underscores the paramount importance of SLAs in delivering satisfactory user experiences.
- Exploration of SLA Metrics: We conduct an in-depth examination of SLA metrics and their profound significance in the realm of IT resource management. These metrics serve as indispensable tools for assessing service quality, enabling providers and users to maintain agreed-upon service standards.
- Utilization of Metrics for CC Monitoring and Management: We shed light on the practical applications of metrics in cloud computing (CC) monitoring and management techniques. These metrics play a pivotal role in ensuring the efficient utilization of resources and the fulfillment of SLAs.
- Case Study on IoT-based Cloud Resource Utilization: This article culminates with a
 detailed case study showcasing the application of metrics to maintain CPU utilization
 in an Internet of Things (IoT)-based cloud environment. This real-world example
 highlights the practical relevance of the concepts discussed throughout this article.

Cloud monitoring and prediction are fundamental components of modern cloud computing (CC), providing crucial insights into the performance, availability, and resource utilization of cloud-based services and infrastructures. These practices are essential not only for optimizing cloud operations but also for improving security and ensuring cost efficiency.

Cloud monitoring involves the continuous collection and analysis of various data points within a cloud environment, including system performance metrics, application logs, network traffic, and security events. Real-time visibility into these aspects is vital for detecting abnormalities, identifying performance bottlenecks, and proactively addressing issues that may affect service quality and availability. Predictive analytics in cloud monitoring goes beyond real-time insights, utilizing historical data and complex algorithms to estimate future patterns and potential issues. This predictive capability is critical in cloud management, enabling organizations to anticipate resource requirements, prepare for scalability, and minimize security threats before they manifest. Predictive analytics empowers cloud providers and consumers to optimize resource allocation and reduce the risk of service outages.

Cloud security monitoring is essential for detecting and preventing security risks and breaches. Security Information and Event Management (SIEM) solutions correlate security data across cloud services and applications. Predictive analytics can help identify suspicious trends and predict potential security attacks, allowing for timely actions and an overall improvement in cloud security posture.

Given the pay-as-you-go model of CC, effective cost control is crucial. Cloud cost monitoring and forecasting track resource usage and project future cost trends. Predictive insights enable organizations to make informed decisions about resource provisioning, scalability, and consumption, thereby reducing wasteful costs. These practices also assist in capacity planning, ensuring that cloud resources can meet rising demand by forecasting future resource requirements based on past consumption trends.

Cloud monitoring and prediction are invaluable tools for modern cloud management, offering real-time insights, enabling a proactive issue response, improving security, lowering costs, and facilitating effective capacity planning. Organizations can ensure the reliable and cost-effective delivery of cloud-based services by integrating monitoring and predictive analytics into cloud operations, aligning their cloud resources with business objectives and user expectations.

2. Relationship between Monitoring, Prediction, and Policies

Monitoring plays a crucial role in identifying the current status of the cloud, encompassing metrics, such as CPU utilization in MHz and disk read throughput in KB/s. The application of policies [17] becomes apparent as monitored values cross predefined thresholds. With access to monitored data logs and insights into task behavior affecting cloud resource utilization, predictive techniques, such as supervised learning, become indispensable for managing such scenarios.

As depicted in Figure 2, the monitoring mechanism interacts with the rule engine, facilitated by the optimization engine. The primary role of the optimization system lies in determining when migration scenarios should be initiated based on policies and their associated activation functions.

The optimization objective may involve minimizing virtual machine migrations [18] or mitigating the impact on physical machines during migration. With the data finalized by the optimization engine, the provisioning engine assigns cloud resources to ensure optimized allocation. It is worth noting that the performance of the cloud system can be significantly affected by the selection of policies and their associated threshold settings, potentially leading to service degradation in the cloud computing (CC) environment [19–22].



Figure 2. Policy-Based System.

3. Policies and SLA Management

Let us delve into how policies play a pivotal role in SLA management. Figure 3 illustrates the four phases through which SLAs govern applications hosted in the cloud: feasibility, on-boarding, pre-production/production, and termination [23].





As previously discussed, policies are instrumental in making auto-scaling decisions, and it becomes evident that efficient resource utilization hinges on the criteria set by policies [24], which are chosen based on the metrics in use.

In the realm of cloud computing (CC), service-level agreements (SLAs) and policy management are closely related and play pivotal roles in ensuring the effective and reliable delivery of cloud services. Firstly, SLAs are the written contracts that specify the terms and conditions under which cloud services are provided. These agreements cover a wide range of topics, including data security, response times, performance, and availability. In contrast, policy management is responsible for establishing and upholding the guidelines

that control how cloud resources are used. SLAs frequently include policies that specify how services should be provided, what resources can be assigned, and when specific actions should be taken in a CC environment. For instance, to ensure the service complies with the established performance standards, a cloud provider may set up a policy that directs resource allocation based on particular SLA parameters.

Second, to guarantee that cloud services comply with customer expectations and legal requirements, SLAs and policy management work closely together. SLAs establish performance standards, and policies direct cloud infrastructure behavior to achieve those standards. For example, a policy can specify that when the system performance drops below a certain SLA-specified level, more resources have to be allocated automatically. This proactive resource management, based on defined policies, ensures that the SLAs are satisfied when conditions change, like abrupt spikes in user demand.

Finally, due to the dynamic nature of CC, both SLAs and policies must be continuously monitored and adjusted. Policies must adapt to these changing requirements, and SLAs may change as a result of changing client needs. Together, the two offer the responsiveness and flexibility needed in a cloud setting. Effective policy management guarantees resource allocation in accordance with SLAs, and the feedback loop between SLAs and policies enables the continuous optimization of cloud services to meet changing demands while upholding compliance and service quality. As a result, in cloud computing, SLAs and policies direct resource allocation and service behavior to fulfill those standards. When combined, they empower cloud service providers to offer excellent, adaptable, and flexible services while maintaining compliance with industry norms and client demands.

In Figure 4, we provide a detailed breakdown of the four phases of the SLA and policy management:

- Feasibility Analysis: This phase involves three types of feasibility analysis: technical, infrastructure, and financial. It aims to determine the suitability of resources to ensure that the projected demands of the applications can be met.
- On-boarding: On-boarding refers to the process of migrating an application to the cloud, accompanied by the use of corresponding SLAs. This phase also involves the creation of the policies (comprising various rules and operational policies) necessary to ensure the fulfillment of service-level objectives (SLOs) specified in the application's SLAs.
- Pre-Production and Production: In the pre-production phase, the application operates in a simulated environment to test its adherence to the specified SLAs. If this phase proceeds smoothly, the application moves on to the production phase, where it runs in the actual cloud environment.
- Termination: When a customer decides to withdraw an application running in the cloud, the termination phase is initiated, leading to the cessation of the application.



Figure 4. SLA and policy management.

4. Metrics Identified in Cloud Computing and Policy-Making Criteria

The approach for extracting data is designed to be a systematic framework, a detailed guide, that aids in identifying, categorizing, and sorting various metrics or measurements. This is meticulously laid out in Table 1, where all the criteria that are used for this data extraction are compiled and summarized. The overarching aim of this methodological setup is to ensure a consistent and uniform way of analyzing, evaluating, and comparing the existing quality indicators, specifically those related to the quality of service (QoS) in the domain of cloud services. By doing so, the strategy aspires to provide a comprehensive and authoritative snapshot of the current advancements and standards in the field.

Different Criteria	Possible Outcomes Related to QoS	Reference
Characteristics of QoS	Usability, maintainability, reliability, compatibility, suitability, security	[25]
Type of Metric	Indicator—analysis of the model, base—baseline measurement method, derived—functions of the various measurements	[26]
Measurement Unit	The corresponding metric unit	[27]
Associated Cloud Lifecycle Phases	 Step 1: Requirements gathering Step 2: Acquisition Step 3: Development process Step 4: Integration Step 5: Operation Step 6: Termination 	[28]
Cloud Artifact and Its Measurement	Specifications of the cloud services, the cloud design and architecture, various types of cloud services	[29]
Three Main Services of the Cloud	IaaS—Infrastructure as a Service, PaaS—Platform as a Service, SaaS—Software as a Service	[30]
Viewpoints of Various Users/Stakeholders of the Cloud	Cloud user, broker, developer, service provider, service request brokers	[31]
Support-based Tools	Automated and manual tools	[32]
Results of the Measurement	Quantitative, qualitative, hybrid	[33]
Function of the Measurement	Formula for calculation and explanation of how the metrics are calculated	[34]

Table 1. Different cloud criteria and their QoS.

Table 2 provides an overview of the different quality attributes as defined by the ISO/IEC 25010 [35] standard and the corresponding references from the primary studies. Each quality attribute, such as performance efficiency, reliability, security, operational policybased functions, maintainability, usability, portability, and compatibility, is associated with multiple references indicating the focus of the research in the respective areas. These quality attributes are crucial for assessing the effectiveness and reliability of cloud services. Additionally, the references showcase the diverse perspectives and approaches adopted by researchers to address various aspects of quality in cloud computing. The comprehensive exploration of these quality attributes and their associated references provides valuable insights into the current trends and advancements in the field, offering a holistic view of the multifaceted nature of cloud service quality evaluation.

Quality Attributes	References
Performance Efficiency	[34,34,36,37,37–46]
Reliability	[47-50,50-57]
Security	[58–73]
Operational Policy-Based Functions	[74–90]
Maintainability	[91–100]
Usability	[101–112]
Portability	[113–121]
Compatibility	[122–126]

Table 2. Quality attributes of cloud.

In the domain of cloud computing (CC), a diverse set of metrics assumes critical roles in assessing performance, optimizing resource utilization, and guiding policy-making decisions. These metrics collectively ensure the effective and seamless operation of cloudbased systems and services. Notably, the key metrics identified in CC include the following:

- The performance gain in scheduling techniques is typically quantified as the disparity between the current execution time and the baseline execution time, with the latter computed through task execution during idle scenarios [127,128].
- Computing performance often centers on response time, a crucial factor in determining system efficiency and user satisfaction.
- Quality of service (QoS) is upheld when the resources consumed remain below the total available resources in the computing environment, ensuring optimal service delivery and user experience [129].
- Cost efficiency, specifically in terms of energy consumption, significantly impacts overall performance, with an emphasis on maintaining lower operational costs and environmental impact [130].
- The overall effectiveness of a task in the cloud environment is evaluated based on the lowest total execution time, a metric that reflects the system's responsiveness and efficiency.

A variety of other circumstances in the cloud computing (CC) environment and their associated usages of the metrics [131] are presented in Table 3.

Scheme	Circumstances	Usages of Metrics	
	New services deployment	Percentage of the average time the request servicing thread has been found busy	
Microservices	The percentage of the time the service will be reachable	Enqueued requests number	
	The number of requests that are enqueued	Percentage of time the services were reachable	
	Databases quick response, the messages queues are faster	The frequency of query execution, failure rate, response time	

Table 3. Metrics and their usages.

Scheme	Circumstances	Usages of Metrics	
	Responsiveness of the processes in the container	Time for CPU throttled	
Container	The images that have been deployed	Disk I/O of container, memory usages	
-	Did containers are associated with over-utilization because of the hosts	Network (dropped packets and its volume)	
Host Changes in the utilization and Memory problem with the application or usage), CPU process		Memory capacity (percentage of usage), CPU utilization (percentage of usage)	
Infractructure	Cost of running services or deployments	The traffic of the network	
mastructure	The ratio of microservices and/or	Database utilization, shared services,	

Table 3. Cont.

End User

5. Relationship between the Metrics and Policies

container per instance

Average web response time practiced

by the end user per region

The relationship between the metrics and policies in cloud computing (CC) is depicted in Figure 5. This relationship is a critical aspect of the effective management and optimization of cloud resources [132]. The monitoring mechanism in the cloud environment is classified into proactive, reactive, and contractual methods:

- Proactive: Proactive monitoring involves making decisions based on predefined rules before tasks are allocated to the cloud environment.
- Reactive: Reactive monitoring entails making decisions by observing the current requests and their response parameters.
- Contractual: Contractual monitoring relies on decisions based on service-level agreements (SLAs).

In the monitoring process, the current state is observed, and subsequently, the metrics collector is triggered, which then activates the metrics analyzer. The metrics analyzer identifies the appropriate policies and parameters necessary to fulfill the requirements of the end users' requests. Finally, the request is transmitted to the resource manager, as illustrated in Figure 5.

Tables 4–6 present the various metrics [133] used in CC environments, along with the associated policies and descriptions for each metric. The terminology for identifying threshold values and creating policies is crucial, and dynamic threshold mechanisms are utilized to adapt to varying workloads.

The formulation of these policies hinges upon the precise determination of threshold values. This determination process, involving conditions such as "X should be less than or equal to" or "greater than," assumes a pivotal role in policy creation. The complexity of this task necessitates the use of dynamic threshold mechanisms.

Notably, cloud environments are dynamic and subject to evolving workloads. Consequently, the application of static thresholds, which remain constant over time, may prove to be inadequate in effectively managing the performance of cloud resources. To address this challenge, dynamic or adaptive thresholds are introduced. These adaptive thresholds are established based on the observed behavior of the cloud environment and the specific metrics under consideration. This dynamic approach ensures that performance policies remain relevant and responsive to the ever-changing demands and conditions within the cloud infrastructure.

storage

Response time, percentage of user

actions failed



Figure 5. The relationship between monitoring, prediction, and policies.

Table 4. Various metrics of cloud computing.

Measurable Quantities	Circumstances	Usages of Metrics
Communication	Data communication in the cloud environment	 The frequency of packet loss The rate of connection error Bit transfer speed (MPI) Delay in MPI transfer
Computation	Computing data or job processing in the cloud environment	 CPU load (%) OP benchmark (FLOP) rate Instance efficiency (% peak CPU)

Table 4. Cont.

Measurable Quantities	Circumstances	Usages of Metrics
Memory	Memory management-related	 Average hit timing (sec) Memory bit and byte speed (MB/s, GB/s) Rate of updated random memory Time of response (ms)
Time	Task completion time	Time of computationTime of communication

Table 5. Economic features and policies.

Features	s Description		Policies	
Elasticity	Addition and removal of cloud resources automatically	•	Task size (n) and resource level (X) Boot time (in seconds) Suspend time (in seconds)	

Table 6. Economic features of cloud computing.

Features	Description	Related Policies	Metrics
Elasticity	The addition and removal of cloud resources automatically	Task size (n) and the level of resources (X) required at the IaaS level	Boot time (in seconds)
		Depends on the downtime of the cloud, mean time to failure, mean time to repair	Suspend time (in seconds)
		Percentage of the availability of the resources (server, CPU, memory, etc.) on an hourly basis and provisioning time (in seconds) or uptime for an instance of the virtual server; virtual infrastructure server starts and stops date; cumulative and continuous frequency over a predefined period	USD0.15/hour small instances, USD0.90/hour large instances, USD0.20/hour medium instances

Tabl	le	6.	Cont.

Features	Description	Related Policies	Metrics
		Percentage of the availability of the resources, for example, network usage	Total acquisition time (in seconds); the outbound network traffic in terms of bytes, cumulative and continuous frequency over a pre-specified period for the cloud service, such as IaaS, PaaS, SaaS; example: up to 400 MB free daily and USD0.02/GB thereafter, USD0.005/GB after the 1TB per month

Dynamic thresholds indeed play a crucial role in adjusting to the changing conditions and demands within a cloud environment. These thresholds are designed based on the statistical analyses of the goal line metrics, which are akin to the benchmarks established during the baseline period. The baseline period is determined under ideal environmental conditions and serves as the reference point for evaluating the system's performance over a specific time frame. The concept of moving the window baseline phases involves assessing the performance based on the variance from a certain number of days preceding the present date. This approach enables a more responsive and adaptive mechanism for regulating the system's performance in dynamic cloud environments.

Not all metrics within the context of cloud computing require the implementation of dynamic thresholds; rather, their necessity is contingent upon the specific criteria outlined in the research. These criteria include factors such as the magnitude of the load the system is handling, the specific types of load being processed, the overall utilization of system resources, and the responsiveness of the system, as indicated by its response time. These key considerations play a crucial role in determining whether a particular metric would benefit from the adoption of dynamic thresholds, thereby ensuring an adaptive and responsive approach to managing the performance of cloud resources. Such insights are discussed in detail in various research papers within the domain.

Figure 6 visually presents the diverse kinds of threshold values employed within a cloud environment. The significance level of these thresholds is crucial in determining the statistical implications, enabling the identification of present values that deviate significantly from the norm. Additionally, the percentage of the maximum threshold is utilized to gauge the proportion of the highest practical value attainable within a specified time frame, thereby aiding in the assessment of the performance bounds.



Figure 6. Types of thresholds.

Clear thresholds indicate a state where no alert will be generated, and historical data are not removed. Occurrences thresholds check the successive quantity of occurrences

before raising an alert. Based on the discussions above, threshold values can be generated and used for setting policies for the respective metrics. These metrics and policies can serve market-oriented cloud computing.

Tables 4–7 provide a comprehensive overview of the various metrics and policies, and their descriptions, encompassing a wide array of features and aspects in cloud computing environments.

Features Description Policies Metrics Flexibility: percentage of Quantifiable and its Anywhere and anytime access uptime of the service. Total Availability performance at an average to services provided uptime/total time. Example: load 99% uptime (minimum) Data rate X at which the data Accuracy are being transferred Normal operational threshold Response time The expansion of the Average resources assigned Scalability infrastructure to handle the Normal operational threshold and requested resources amplified load The accuracy of the services. Under predefined conditions, identify the percentage of The services should be successful service outcomes, Reliability functional with time and no Normal operational threshold i.e., operational (normal) cases of malfunction period duration/failures number. Example: average 90 days (with the frequency: yearly or monthly) Calculation: (date/time of recovery-date/time of failure)/sum of number of Fault tolerance: mean time failures. Another calculation: between failures, for monthly identify the normal period or yearly duration of the operational/numbers of failures. Example: an average of 90 days, 120 min average Recoverability: Average time in the ideal (date/switchover completion scenario for repairing the time-date/failure time)/total failure, to reduce the failures number. Example: 10 downtime min average

Table 7. Other general features of cloud services.

Features	Description	Policies	Metrics
Efficiency and achieving maximum productivity and average utilization	Utilization of the resources, such as measurable characteristics, capacity of the storage device with continuous frequency	Assume the threshold is 60 GB, and if the demand rises and crosses the 60 GB of the utility, then add another 80 GB of storage from the resource pool	80 GB storage max
		The total percentage of successful services outcomes under pre-specified conditions	Downtime management: calculation—successful responses (total)/number of requests; with the frequency as yearly, monthly, and weekly. Example: minimum downtime acceptable 98%
Sustainability	Not be detrimental to the environment	Average performance in peak and non-peak hour	Data center performance: calculation—date/time of request-date/time of response/number of requests (total), with the frequency of monthly, weekly, and daily. Example: -5 milliseconds average
		Average power consumption in the ideal scenario	Power usage efficiency; power usages effectiveness (PUE) = total power of data center/power required or used by the IT equipment

Table 7. Cont.

6. Market-Oriented Architecture for the Data Centers

An application of SLA management and policies is the implementation of a marketoriented architecture (MOA). MOA is a pioneering approach to data center management that incorporates service-level agreements (SLAs) and operational standards. MOA emerges as a fundamental paradigm in the constantly evolving world of cloud computing (CC) and data center management, supporting the optimization of resource allocation, cost effectiveness, and customer satisfaction.

MOA makes use of a market-based system in which resources are treated as commodities, and allocation is governed by dynamic pricing models. By including SLAs in this architecture, data center operators may provide predictable performance assurances to clients, increasing trust and reliability. Operational policies are critical components of this architecture because they specify the rules that govern resource allocation, provisioning, and de-provisioning. These policies are crucial in balancing cost optimization and achieving SLAs, ensuring that the data center works in accordance with the business objectives.

MOA is further supported by significant data analytics and machine learning algorithms that examine historical data as well as real-time performance measurements. These analytics help not only estimate resource demands but also fine-tune pricing techniques to optimize resource utilization. MOA emerges as a powerful framework to address the complexities of modern data center management, contributing significantly to both cost efficiency and customer satisfaction, with its emphasis on market-driven resource allocation, SLA adherence, and data-driven decision making.

Data centers serve as the foundational infrastructure for cloud computing (CC) services. They are the backbone that supports the delivery of cloud services to users. Figure 7 provides an overview of the key components supporting MOA (market-oriented architecture) in the context of CC data center management [134]. These components work together to optimize resource allocation, ensure adherence to service-level agreements

(SLAs), and enhance overall data center efficiency. This reference architecture illustrates how MOA integrates SLAs, operational policies, and dynamic resource allocation into the data center environment, contributing to the effective and market-driven management of cloud resources.



Figure 7. Cloud data center: reference architecture.

Here are the descriptions of the significant components within this architecture:

- Users and Brokers: These entities play a crucial role in initiating workloads that the data center will manage. They are responsible for interacting with the data center and making requests for various cloud services.
- SLA Resource Allocation Mechanism: This component serves as the vital interface between the cloud service provider and the data center [135]. Its primary objective is to ensure that the services provided align with the service-level agreements (SLAs) agreed upon with the clients. It facilitates the allocation of resources in accordance with these SLAs.
- Admission Control Module and Service Request Examiner: This module evaluates the current state of the data center, including the availability of resources. It is responsible for scheduling and allocating requests for execution based on the available resources and the defined SLAs.

- Module for Pricing: This component is responsible for determining the charges for users based on the terms specified in their SLAs. It considers parameters, such as virtual machines, memory, computing capacity, disk size, and usage time.
- Accounting Module: This module generates billing data based on the actual resource usage by the users. It plays a critical role in maintaining transparency and accuracy in billing processes.
- Dispatcher: The dispatcher is responsible for instructing the infrastructure to deploy the necessary machines to fulfill user requests. It plays a significant role, particularly in the case of Infrastructure as a Service (IaaS), by managing the allocation of resources.
- Resource Monitor: This component is continuously engaged in monitoring the status of computing resources, including both physical and virtual resources. It plays a critical role in ensuring the optimal utilization and performance of the available resources.
- Services of Request Monitor: This component tracks the progress of service requests, providing valuable insights into the system's performance and offering quality feedback on the provider's capabilities. It helps in maintaining a high level of service quality and user satisfaction.
- Virtual Machines (VMs): VMs are fundamental units within the cloud computing (CC) infrastructure. They serve as the building blocks for addressing various user requirements and enabling the provisioning of different cloud services.
- Physical Machines: At the lowest level of the architecture, the physical machines constitute the core physical infrastructure, which can encompass one or more data centers. This layer provides the necessary physical resources required to meet the demands of the users and the services they request.

In [23], an analysis and the taxonomy of the schedulers were presented, as depicted in Figure 8. These schedulers were classified based on their allocation decisions, market models, objectives, participant focus, and application models. Notably, the market model plays a critical role in facilitating trade between providers and users within the cloud computing environment. The classification of market models was outlined as follows:

- Game Theory: Users engage in a provision game with various payoffs based on specific actions and different strategies. Game theory provides a framework for understanding strategic interactions among rational decision-makers.
- Proportional Share: This approach aims to allocate tasks fairly across a set of resources, with shares directly related to the user's bid. It ensures proportional distribution based on user demands and resource availability.
- Market Commodity: Cloud data center providers charge consumers based on their resource usage, and these charges may vary over time. This model allows for flexible pricing that can adapt to changes in demand and resource availability.
- Posted Price: Similar to the market commodity model, the posted price approach may include special discounts and offers for specific users. It offers transparency in pricing and allows users to make informed decisions based on the available options.
- Contract Net: End users advertise their requirements and invite resource owners to submit bids. Resource owners respond based on their resource availability and capabilities. The end user then consolidates the bids and selects the most favorable one, creating a contractual agreement.
- Bargaining: Negotiations between providers and resource consumers determine the final resource price. This model allows for flexibility and mutual agreement between the parties involved, ensuring that both parties benefit from the transaction.
- Auction: Initially, resource prices are unknown, and competitive bids, regulated by a third party (the auctioneer), determine the final price. Auctions provide a competitive environment where users can bid based on their willingness to pay, resulting in optimal resource allocation and fair pricing.



Figure 8. Classification of the market-oriented model.

Service-level agreements (SLAs) are integral to the operation of e-commerce cloudhosted applications and market-oriented cloud design architectures. These agreements, coupled with cloud metrics and policies, ensure that cloud services meet performance and reliability standards. Comprehensive details regarding SLAs, including specific metrics and economic considerations, are elaborated in Tables 3 and 6. These tables provide a structured framework for understanding the relationships between SLAs, metrics, and policies within the context of cloud computing.

7. Case Study: Utilizing Metrics, Policies, and Machine Learning for IoT-Based Cloud Monitoring

This case study outlines the pivotal role of metrics, policies, and machine learning in the context of IoT-based cloud monitoring. The integration of these elements is vital for ensuring the seamless operation, performance optimization, and reliability of IoT applications within cloud computing ecosystems. By leveraging metrics and servicelevel agreement (SLA) management, organizations can achieve comprehensive monitoring capabilities, enabling the real-time analysis of IoT device and cloud service performance. This capability is crucial for timely issue detection and resolution, preventing disruptions that could impact the functionality of IoT applications.

Furthermore, the dynamic scaling of cloud resources, facilitated by these systems, allows for efficient resource allocation and optimization in response to fluctuating demands from IoT devices. This dynamic resource allocation not only enhances the overall system efficiency but also contributes to cost effectiveness, a critical aspect of resource management. Additionally, metrics and SLA management serve as guardians of the quality of service (QoS) standards, defining, monitoring, and ensuring compliance with stringent performance and reliability criteria that are essential for high-quality IoT applications.

In terms of security and privacy, these systems play a significant role by integrating security and privacy provisions into SLAs, safeguarding sensitive IoT data from unauthorized access or breaches. They also facilitate cost control by providing accurate usage statistics and cost metrics, allowing organizations to monitor and regulate cloud expendi-

ture effectively. The architecture incorporates fault detection and recovery mechanisms, swiftly identifying performance deviations and implementing recovery protocols in the event of service outages, thereby minimizing downtime and interruptions.

Moreover, these systems facilitate continuous improvement by analyzing performance data, identifying areas of weakness, and enabling informed decisions and adjustments to enhance the scalability, reliability, and performance of IoT and cloud services over time. Lastly, metrics and SLA management support capacity planning by offering valuable insights into usage patterns and resource requirements, enabling organizations to ensure that their IoT applications and cloud services are equipped to handle future growth and evolving needs.

Cloud monitoring is a vital component in the effective management of cloud-based systems, enabling the efficient handling of dynamic scheduling, cross-layer monitoring, and the identification of diverse fault scenarios [136]. The case study presented here highlights the significant role of metrics and policies within cloud computing, specifically focusing on their application in addressing the challenge of monitoring overhead. By leveraging these metrics and implementing effective policies, organizations can enhance their ability to ensure optimal performance, reliability, and security within their cloud environments.

The visual depiction in Figure 9 effectively highlights critical open issues in cloud monitoring, such as pattern and root cause analysis, workload generation, intelligent agents, and the reduction in monitoring overhead. These challenges underscore the importance of implementing efficient metrics and policies to effectively address these concerns and optimize the overall performance of cloud-based systems.



Figure 9. Cloud monitoring open issues.

7.1. Dataset

The dataset mentioned represents a cloud environment comprising 750 virtual machines (VMs) that are utilized by a cloud service provider for hosting diverse analytical strategies. These strategies leverage data collected through IoT devices used by patients, with applications including the adjustment of medication dosages, monitoring recovery stages, and other health-related analyses. In this case study, resource utilization metrics such as CPU usage (as a percentage), memory usage (as a percentage), and networktransmitted throughput (measured in KB/s) are the primary focus. These metrics serve as essential indicators for evaluating the performance and efficiency of the cloud-based infrastructure, ensuring optimal service delivery and resource management.

7.2. Hardware Setup

The experimental configuration was established using the OpenStack private cloud, employing three virtual machines (VMs) dedicated to conducting forecasting analytics. The initial two VMs, M1 and M2, were primarily utilized for the modeling process, while the subsequent two VMs, M3, were employed for the modeling phase. The hardware setup incorporated specific components, including a PDL380 10th-generation SFF rackmount server, an MS Windows Server Standard Core 2019 with a single OLP 16 license-ae, an Intel XEON Silver 4110 processor, 128 GB of DDR4-2666 MHz memory (32 GB in each of the four modules), a 2.4 TB SAS 12 G 10 k SFF HDD, and an HPE Smart Array 8161a SR 10th Gen Controller. This robust hardware configuration facilitated the efficient and effective execution of the forecasting analytics tasks within the cloud environment, ensuring the timely and accurate processing of the collected data. This system has been implemented within the university campus laboratory and was sourced from a local vendor in Ahmedabad, India.

7.3. Monitoring IoT-Based Cloud Resources

Monitoring a real-world cloud environment is a complex yet crucial task. In this case study, the cloud ecosystem comprises 750 virtual machines (VMs) that are overseen and utilized by a cloud service provider. These VMs are instrumental in hosting various analytical strategies that leverage data collected from IoT devices employed for patient monitoring. The following resources are specifically targeted for monitoring and analysis within this context:

- CPU Usage: This metric reflects the percentage of CPU utilization, offering insights into the processing load and performance demands on the virtualized computing resources.
- Memory Usage: Representing the percentage of memory utilization, this metric provides essential information about the memory requirements and allocation efficiency within the cloud environment.
- Network-Transmitted Throughput: Measured in kilobytes per second (KB/s), this
 metric is indicative of the data transmission rate through the network, which is critical
 for evaluating the efficiency of data communication and network performance.

By closely monitoring and analyzing these key metrics, we can gain valuable insights into the performance and resource utilization of the cloud-based infrastructure, enabling effective decision-making and optimization strategies.

7.4. Solution Approach

The ubiquity of the Internet of Things (IoT) has revolutionized the way in which everyday activities are interconnected. IoT devices, equipped with sensors, software, and embedded electronics, seamlessly gather, transmit, and process large volumes of data, often referred to as "big data." However, this data deluge presents a significant challenge for both internet infrastructure and cloud computing (CC) systems. To effectively manage this surge in data, CC systems must navigate the complexities of handling substantial network traffic while upholding stringent quality of service (QoS) standards. Consequently, the efficient management of resources becomes a critical priority. In this context, the various parameters of Infrastructure as a Service (IaaS) cloud systems have been meticulously considered to devise a comprehensive solution approach.

7.4.1. Metrics and Policies

Metrics and policies play a crucial role in managing cloud resources effectively. In Figure 10, we provide an illustrative example of the metrics and policies related to CPU utilization in the cloud ecosystem, as detailed in the dataset content mentioned above. These metrics and policies play a crucial role in reducing the overhead of monitoring data.

23 of 39



Figure 10. Metrics and policies example.

The monitoring process retrieves data from the cloud data center, as depicted in Figure 11. The CPU utilization graph, based on the Public Cloud dataset, monitors the real-time status of cloud resources, specifically CPU utilization. When the CPU utilization falls below a certain threshold, such as 50%, monitoring mechanisms are triggered. Corresponding metrics, in this case, CPU utilization, and policies are applied to optimize the resource management.



Figure 11. CPU Utilization Graph: x-axis—time stamp in ms; y-axis—percentage of utilization.

7.4.2. Machine Learning Predictions

Workload Utility Levels and Metrics: The identification of workload utility levels, including low utility, moderate utility, or high utility, in the context of CC, involves a complex process influenced by various metrics and policies. These components form the basis for cloud resource allocation and optimization strategies, taking into consideration

resource utilization factors, such as CPU, network bandwidth, and other Infrastructure as a Service (IaaS) resources. Cloud providers rely on a wide range of resource utilization metrics to accurately classify workloads. CPU utilization serves as a critical metric in these measurements, delineating the differences between low-utility tasks with occasional, minor CPU demands and high-utility workloads that necessitate persistent and substantial CPU resources. Similarly, network bandwidth consumption is an essential indicator, with high-utility applications consistently requiring more network throughput than their low-utility counterparts.

Resource Allocation Policies: The thorough assessment of workload utility levels encompasses metrics for storage, memory, and I/O activities, providing a comprehensive understanding of resource requirements. Cloud resource allocation policies are meticulously designed to align with workload utility levels. Low-utility workloads often favor resource consolidation and cost reduction, promoting resource sharing and dynamic allocation. On the other hand, average-utility workloads are met with balanced resource allocations, ensuring optimal performance while optimizing costs. High-utility workloads, which demand continuous and high-performance delivery, typically receive dedicated and premium resource allocations.

Role of Machine Learning and Predictive Analytics: Machine learning and predictive analytics play a crucial role in forecasting workload utility levels with a degree of accuracy. By analyzing historical data, cloud providers can identify consumption trends, enabling automated resource allocation decisions. This data-driven approach ensures that the cloud ecosystem can swiftly respond to changes in workload utility levels. The inherent agility of the cloud allows for real-time adjustments in resource allocation, which is invaluable for adapting to fluctuations in workload utility levels.

Dynamic Resource Scaling: When workloads display indications of transitioning across utility categories, policies can be established to trigger resource scaling. For instance, if a typical utility application experiences sudden spikes in the CPU or network demand, automated scaling mechanisms are activated to ensure uninterrupted resource provision. These mechanisms guarantee that the application continues to receive the necessary resources without interruption, maintaining performance levels.

Customer-Centric Utility Levels: The classification of workload utility levels within a user-centric paradigm is closely linked to customer-defined service-level agreements (SLAs). Users define their desired utility levels based on resource performance and availability, directly influencing how the cloud manages workloads. This approach ensures that consumers receive the promised utility level, aligning with their operational requirements and resource investment preferences.

Optimizing Resource Allocation: To summarize, defining workload utility levels within the cloud ecosystem is a multifaceted process supported by measurements, policies, and advanced analytics. It represents an ongoing effort aimed at enhancing resource allocation while maintaining a balance between cost effectiveness and performance, all while remaining responsive to the evolving demands and expectations of cloud customers.

Incorporating a diverse range of machine learning (ML) algorithms is crucial for accurate workload prediction. The utilization of various algorithms such as Linear Regression (LiR), Support Vector Regression (SVR), Decision Tree (DT), Random Forest (RF), Logistic Regression (LoR), and Artificial Neural Network (ANN) enables the comprehensive analysis and forecasting of the workload. These algorithms, when applied to the dataset, facilitate precise and robust workload predictions, ensuring the effective management and allocation of resources within the cloud ecosystem.

Figure 12 illustrates the predictions for CPU utilization using a range of machine learning techniques, including Linear Regression (LiR), Support Vector Regression (SVR), Decision Tree (DT), Random Forest (RF), Logistic Regression (LoR), and Artificial Neural Network (ANN). These predictive models enable accurate forecasting of the CPU utilization, providing valuable insights into the resource demands and usage patterns within the cloud environment.



Figure 12. Prediction of CPU utilization.

Figure 13 demonstrates the predictions for memory usage using various machine learning techniques, including Linear Regression (LiR), Support Vector Regression (SVR), Decision Tree (DT), Random Forest (RF), Logistic Regression (LoR), and Artificial Neural Network (ANN). These predictive models enable the accurate forecasting of memory utilization, providing insights into the memory usage trends and patterns within the cloud infrastructure.



Figure 13. Memory utilization prediction.

Figure 14 presents the memory utilization data, demonstrating the patterns and trends in memory usage over a specific period. The visualization offers valuable insights into how memory resources are being utilized within the cloud environment, aiding in the assessment of memory allocation and requirements. Understanding memory utilization is critical for optimizing resource allocation and ensuring the efficient performance of cloud-based applications and services.



Figure 14. Memory utilization.

Figure 15 presents the predictions for the network-transmitted throughput, utilizing various machine learning techniques, including Linear Regression (LiR), Support Vector Regression (SVR), Decision Tree (DT), Random Forest (RF), Logistic Regression (LoR), and Artificial Neural Network (ANN). These predictions offer valuable insights into the anticipated network throughput trends and patterns within the cloud infrastructure, aiding in the proactive management and optimization of network resources.



Figure 15. Cont.



Figure 15. Network-transmitted throughput prediction.

Figure 16 depicts the network-transmitted throughput within the cloud environment, illustrating the dynamic changes and fluctuations in the network data transmission rates over a specific time period. The graph serves as a visual representation of the actual network throughput data, providing insights into the overall network performance and data transmission patterns, which are crucial for understanding the network's efficiency and capacity utilization.



Figure 16. Network-transmitted throughput.

7.5. Algorithm: Effective Resource Monitoring Using Metrics and Policies

In this section, we introduce Algorithm 1, designed for efficient resource monitoring and management by utilizing metrics and policies. This algorithm systematically defines the steps involved in monitoring cloud resources while effectively mitigating the challenges posed by extensive monitoring logs. Additionally, it offers valuable insights into predicting the behavior of the cloud environment using diverse parameters. It is worth noting that this analytical approach can readily extend its applicability to address other crucial parameters, like disk read-and-write throughput and network-received throughput.

This algorithm showcases the effective utilization of metrics and policies to dynamically adapt the monitoring frequency in response to the prevailing performance parameter, such as CPU utilization. This intelligent adjustment minimizes the accumulation of superfluous monitoring data and logs, consequently enhancing resource management and optimization within the cloud environment.

Algorithm 1 Steps for effective resource monitoring using metrics and policies

- 1: Set a monitoring interval for any instance at which it is being monitored for any interval $= m_1$ seconds for CPU Utilization
- 2: Let performance parameters be CPU utilization, denoted as P_1
- 3: Define a policy: Let the upper threshold of the performance parameter for CPU utilization be *t*₁
- 4: Define a policy: Let the lower threshold of the performance parameter for CPU utilization be *t*₂
- 5: Create a log of performance parameter monitoring for every job in the cloud
- 6: Let nt_1 be the performance parameter (e.g., CPU utilization) of a new job
- 7: **if** $nt_1 > t_1$ **then**
- 8: Increase the frequency of monitoring
- 9: end if
- 10: **if** $t_1 > nt_1 > t_2$ **then**
- 11: Decrease the frequency of monitoring
- 12: end if

7.6. Evaluation of Machine Learning Predictions

Evaluating the prediction accuracy of machine learning models is essential in ensuring the reliability and effectiveness of the proposed approach. By assessing metrics such as the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), this study can effectively quantify the extent of the prediction errors, thus providing valuable insights into the performance of various machine learning algorithms in forecasting resource parameters. Lower values of the RMSE and MAE signify improved predictive accuracy and model performance, thereby establishing the credibility of the predictive models in the context of resource monitoring and management in cloud environments.

Table 8 evaluates the prediction accuracy of different ML approaches for CPU utilization. The evaluation metrics used are the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These metrics help us understand how close the predicted values are to the actual values, with lower values indicating better prediction accuracy. In this table, we observe that SVR (Support Vector Regression) has the lowest RMSE and MAE values compared to the other ML models. This indicates that SVR provides the most accurate predictions for CPU utilization among the models evaluated. Lower RMSE and MAE values mean that the predicted values closely match the actual CPU utilization, which is crucial for efficient resource management in a cloud environment.

ML Model	RMSE	MAE
LiR	2.63	1.43
SVR	0.99	0.80
DT	1.37	1.28
RF	1.53	1.14
LoR	39.42	32.36
ANN	2.15	1.36

Table 9 assesses the prediction accuracy of the different ML models for memory usage, similar to Table 8. Again, the RMSE and MAE are used as the evaluation metrics, with lower values indicating better prediction accuracy. In this table, we can see that SVR (Support Vector Regression) has the lowest RMSE and MAE values for the memory usage predictions.

This suggests that SVR is the most accurate model for predicting memory usage, providing predictions that closely align with the actual values.

ML Model	RMSE	MAE
LiR	2.01	1.42
SVR	3.65	2.79
DT	3.86	2.85
RF	1.58	1.12
LoR	73.67	56.56
ANN	2.53	1.55

Table 9. RMSE and MAE for memory usage predictions using various ML approaches.

Similar to Tables 8 and 9, Table 10 evaluates the prediction accuracy of the various ML models but this time for network-transmitted throughput. The RMSE and MAE are once again used as metrics to assess the accuracy. In Table 10, SVR (Support Vector Regression) consistently stands out as the model with the lowest RMSE and MAE values for the network-transmitted throughput predictions. This means that SVR excels in accurately predicting the network-transmitted throughput values, which is critical for maintaining efficient network resource management.

Table 10. RMSE and MAE for network-transmitted throughput predictions using variousML approaches.

ML Model	RMSE	MAE
LiR	0.48	0.28
SVR	0.52	0.29
DT	0.50	0.30
RF	0.47	0.29
LoR	5.66	3.64
ANN	0.49	0.30

In summary, across all three tables, SVR consistently demonstrates the highest prediction accuracy among the evaluated ML approaches. This indicates that SVR is a robust choice for predicting CPU utilization, memory usage, and network-transmitted throughput in cloud environments, making it a valuable tool for optimizing resource allocation and reducing monitoring overhead.

Figure 17 offers an overall comparison of the entire parameter set, encompassing CPU usage, memory usage, and network-transmitted throughput, across various ML techniques. It is evident that ANN, RF, and SVR consistently outperform the other ML techniques in terms of prediction accuracy. These predicted values, reflecting cloud resource parameters—CPU usage (in percentage), memory usage (in percentage), and network-transmitted throughput (in KB/s)—serve as input for the metrics, guiding resource management actions within the cloud.

Forecast accuracy and reliability are critical in the field of cloud workload prediction. To do this, we use a diverse set of machine learning (ML) algorithms, each chosen for its own capabilities and adaptability to distinct data features. Linear Regression (LiR), Support Vector Regression (SVR), Decision Tree (DT), Random Forest (RF), Logistic Regression (LoR), and Artificial Neural Network (ANN) are examples of these algorithms. Support Vector Regression, for example, captures non-linear patterns, while Linear Regression establishes a linear link between input characteristics and workload. Logistic Regression can be modified for probabilistic workload prediction, while Decision Trees provide interpretability. Random Forest mixes many trees for increased accuracy. Finally, Artificial Neural Networks are very good at capturing complex data patterns. Our research is built on the systematic application of these techniques to the workload dataset. The decision on

which algorithm to use for a certain prediction task is data-driven, informed by statistical analysis and insights gained from previous conversations and research in the fields of cloud computing and machine learning. This entire method seeks to provide a strong and scalable framework for workload prediction, guaranteeing that our conclusions are technically sound as well as statistically rigorous.



Figure 17. Comparative analysis of various ML approaches.

The reduction in the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values across these machine learning (ML) algorithms underscores their ability to achieve precise Infrastructure as a Service (IaaS) resource prediction. This crucial feature is aligned with the overarching objective of ensuring accurate workload predictions, ultimately enabling the provisioning of an optimal amount of resources. The effective alignment of workloads and resources is vital for sustaining the reliable availability of cloud services, contributing to the overall efficiency and effectiveness of cloud-based operations.

Consequently, a diverse array of ML models has been employed in this domain, trained to cater to a variety of scenarios pertinent to the utility of diverse cloud resources and their predictive applications. To further bolster the predictive capabilities and resource management, tailored to the specific workloads, the machine learning model that furnishes the most precise forecasts will be selected for deployment.

7.7. E-Commerce Benefits for Running in IoT and Cloud Computing

E-commerce businesses can reap various advantages by operating on Internet of Things (IoT) and cloud computing (CC) platforms [137]. Some of these benefits include the following:

- Reduced Investment Costs: Leveraging cloud infrastructure allows businesses to lower upfront investment costs by procuring IT resources in a cost-effective manner [138].
- Operational Cost Reduction: Cloud platforms enable businesses to scale IT resources, such as CPU, memory, and storage, according to demand, leading to cost savings over time.
- On-Demand Service Provisioning: Cloud services provide on-demand access and agility for end users, allowing businesses to quickly adapt to changing market demands [139].
- Improved Service Quality: Cloud-based e-commerce platforms can enforce critical service-level agreements (SLAs) and enhance computational resilience, resulting in heightened service quality for end users [140].

The interrelationship between cloud characteristics and cloud mechanisms is depicted in Figure 18, highlighting how the implementation of cloud computing mechanisms aligns with the achievement of the desired cloud characteristics. Each cloud mechanism can be thoroughly evaluated based on its specific policies and metrics, as discussed earlier.



Figure 18. Mapping of cloud mechanisms to cloud characteristics.

7.8. Proposals to Improve New Application Challenges for E-Commerce Deployment Using IoT in Cloud Computing

To navigate the challenges and leverage the opportunities of deploying e-commerce applications using the IoT and cloud computing, businesses can consider implementing the following proposals:

- Develop New IT Practices: Establish innovative IT practices that align with evolving market demands, focusing on IT earnings, technology lifecycle management, and data center management to adapt to changing business landscapes [23].
- ROI Identification and Planning: Invest in continuous training and monitoring to accurately identify the return on investment (ROI) and effectively plan the capacity to meet the demands of e-commerce applications powered by the IoT and cloud computing [11].
- Virtualization Platform Selection: Choose the most suitable virtualization platform to facilitate efficient provisioning and de-provisioning of IT resources, ensuring optimal SLA monitoring, billing, and resource management to support e-commerce operations [12].
- Governance and Resiliency: Implement governance and organizational strategies to
 effectively manage and control large-scale resiliency, negotiate cloud-based agreements
 with clients, and foster trust in cloud services, which are essential for a successful
 e-commerce ecosystem [13].

• Mobile Business Expansion: Embrace the growing influence of mobile access to cloud services and ensure that cloud offerings are well aligned with the evolving mobile business landscape to support e-commerce operations efficiently [14].

These proposals provide a framework for businesses to address the unique challenges and seize the opportunities presented by the integration of the IoT and cloud computing in e-commerce applications, ultimately contributing to their success in this dynamic environment.

8. Conclusions and Future Work

The convergence of the Internet of Things (IoT) and cloud computing (CC) has unlocked significant potential for advancements across various technical industries, promising a future characterized by autonomous adaptability and improved environmental sustainability. Within the dynamic cloud environment, characterized by uncertain workloads, the role of policy mechanisms in CC decision making is pivotal. Assessing available cloud capacity before deploying tasks in the CC environment is imperative, and policies can range from simple conditional statements to complex logical structures comprising multiple combinations of actions and triggers.

This paper has leveraged monitoring and prediction mechanisms to establish the current state of cloud infrastructure and anticipate future resource scenarios. This knowledge aids cloud service providers (CSPs) in effective resource management and in triggering various policies based on the relevant metrics.

As a direction for future research, we propose the implementation of intelligent agents, particularly Hierarchical Reinforcement Learning, to engage with cloud resource statuses. These agents can assign positive and negative rewards based on predefined metrics and policies, ultimately aiming to minimize negative rewards and maximize positive rewards. This approach could lead to the identification of optimal solutions that enhance cloud resource management and provide a solid foundation for continued exploration in this field.

Author Contributions: Conceptualization, V.K.P., D.D., M.D.B., B.A., V.C.G. and A.K.; Data Curation, V.K.P., D.D., M.D.B., B.A., V.C.G. and A.K.; Writing—Original Draft, V.K.P., D.D., M.D.B., B.A., V.C.G. and A.K.; Methodology, V.K.P., D.D., M.D.B., B.A., V.C.G. and A.K.; Review and Editing, V.K.P., D.D., M.D.B., B.A., V.C.G. and A.K.; Review and Editing, V.K.P., D.D., M.D.B., B.A., V.C.G. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Capra, M.; Peloso, R.; Masera, G.; Ruo Roch, M.; Martina, M. Edge computing: A survey on the hardware requirements in the internet of things world. *Future Internet* 2019, 11, 100. [CrossRef]
- Luong, N.C.; Wang, P.; Niyato, D.; Wen, Y.; Han, Z. Resource Management in Cloud Networking Using Economic Analysis and Pricing Models: A Survey. *IEEE Commun. Surv. Tutorials* 2017, 19, 954–1001. [CrossRef]
- Breitgand, D.; Silva, D.M.D.; Epstein, A.; Glikson, A.; Hines, M.R.; Ryu, K.D.; Silva, M.A. Dynamic Virtual Machine Resizing in a Cloud Computing Infrastructure. U.S. Patent 9,858,095, 1 February 2018.
- 4. Soumya, E.; Kumar, V.S.; Vineela, T.; Aishwarya, M. Conducive Tracking, Monitoring, and Managing of Cloud Resources. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2018**, *3*, 385–390.
- Tsai, W.; Sun, X.; Balasooriya, J. Service-Oriented Cloud Computing Architecture. In Proceedings of the 7th International Conference on Information Technology: New Generations (ITNG), IEEE Computer Society, Virtual Event, 10–13 April 2010; pp. 684–689.
- Alhamazani, K.; Ranjan, R.; Mitra, K.; Rabhi, F.A.; Jayaraman, P.P.; Khan, S.U.; Guabtni, A.; Bhatnagar, V. An Overview of the Commercial Cloud Monitoring Tools: Research Dimensions, Design Issues, and State-of-the-art. *Computing* 2015, 97, 357–377. [CrossRef]
- 7. Amiri, M.; Khanli, L.M. Survey on prediction models of applications for resources provisioning in cloud. *J. Netw. Comput. Appl.* **2017**, *82*, 93–113. [CrossRef]

- Chard, R.; Chard, K.; Wolski, R.; Madduri, R.K.; Ng, B.; Bubendorfer, K.; Foster, I.T. Cost-Aware Cloud Profiling, Prediction, and Provisioning as a Service. *IEEE Cloud Comput.* 2017, 4, 48–59. [CrossRef]
- 9. Garg, R.; Prasad, V. Survey Paper on Cloud Demand Prediction and QoS Prediction. Int. J. Adv. Res. Comput. Sci. 2017, 8, 794–799.
- Souza, V.B.; Masip-Bruin, X.; Marín-Tordera, E.; Ramírez, W.; Sánchez-López, S. Proactive vs reactive failure recovery assessment in combined Fog-to-Cloud (F2C) systems. In Proceedings of the 22nd International IEEE Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Lund, Sweden, 19–21 June 2017; pp. 1–5.
- Kauffman, R.J.; Ma, D.; Yu, M. A Metrics Suite of Cloud Computing Adoption Readiness. *Electron. Mark.* 2018, 28, 11–37. [CrossRef]
- Prasad, V.K.; Shah, M.; Bhavsar, M.D. Trust Management and Monitoring at an IaaS Level of Cloud Computing. In Proceedings of the 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), Jaipur, India, 27–28 March 2018; pp. 26–27.
- 13. Singh, A.; Kinger, S. An Efficient Fault Tolerance Mechanism Based on Moving Averages Algorithm. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2013**, *3*, 937–942.
- Cai, H.; Gu, Y.; Vasilakos, A.V.; Xu, B.; Zhou, J. Model-Driven Development Patterns for Mobile Services in Cloud of Things. IEEE Trans. Cloud Comput. 2018, 6, 771–784. [CrossRef]
- Comuzzi, M.; Kotsokalis, C.; Spanoudakis, G.; Yahyapour, R. Establishing and Monitoring SLAs in Complex Service Based Systems. In Proceedings of the IEEE International Conference on Web Services (ICWS), IEEE Computer Society, San Francisco, CA, USA, 27 June–2 July 2009; pp. 783–790.
- Waldman, H.; Mello, D.A.A. On the Risk of non-compliance with some Plausible SLA Requirements. In Proceedings of the 11th International IEEE Conference on Transparent Optical Networks, Azores, Portugal, 28 Junue–2 July 2009; pp. 1–4.
- 17. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Obermeyer, Z. Prediction Policy Problems. *Am. Econ. Rev.* 2015, 105, 491–495. [CrossRef] [PubMed]
- 18. Noshy, M.; Ibrahim, A.; Ali, H.A. Optimization of live virtual machine migration in cloud computing: A survey and future directions. *J. Netw. Comput. Appl.* **2018**, *110*, 1–10. [CrossRef]
- 19. Liu, Y.; Daum, P.H.; McGraw, R.; Miller, M. Generalized Threshold Function Accounting for Effect of Relative Dispersion on Threshold Behavior of Autoconversion Process. *Geophys. Res. Lett.* **2006**, *33*, 11. [CrossRef]
- Rai, S.C.; Nayak, S.P.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A.; Panagiotakopoulos, T. ITSS: An Intelligent Traffic Signaling System Based on an IoT Infrastructure. *Electronics* 2023, 12, 1177. [CrossRef]
- Somani, G.; Gaur, M.S.; Sanghi, D.; Conti, M.; Buyya, R. DDoS Attacks in Cloud Computing: Issues, Taxonomy, and Future Directions. *Comput. Commun.* 2017, 107, 30–48. [CrossRef]
- 22. Wu, X.; Zhang, R.; Zeng, B.; Zhou, S. A Trust Evaluation Model for Cloud Computing. In Proceedings of the 1st International Conference on Information Technology and Quantitative Management (ITQM), Suzhou, China, 3 June 2013; Volume 17, pp. 1170–1177.
- 23. Buyya, R.; Broberg, J.; Goscinski, A.M. Cloud Computing: Principles and Paradigms; John Wiley & Sons: Hoboken, NJ, USA, 2010.
- 24. Jennings, B.; Stadler, R. Resource Management in Clouds: Survey and Research Challenges. J. Netw. Syst. Manag. 2015, 23, 567–619. [CrossRef]
- Ksentini, A.; Jebalia, M.; Tabbane, S. IoT/Cloud-enabled smart services: A review on QoS requirements in fog environment and a proposed approach based on priority classification technique. *Int. J. Commun. Syst.* 2021, 34, e4269. [CrossRef]
- 26. Ramaiah, N.S. Cloud-Based Software Development Lifecycle: A Simplified Algorithm for Cloud Service Provider Evaluation with Metric Analysis. *Big Data Min. Anal.* 2023, *6*, 127–138. [CrossRef]
- 27. Riekstin, A.C.; Rodrigues, B.B.; Nguyen, K.K.; de Brito Carvalho, T.C.M.; Meirosu, C.; Stiller, B.; Cheriet, M. A survey on metrics and measurement tools for sustainable distributed cloud networks. *IEEE Commun. Surv. Tutorials* 2017, 20, 1244–1270. [CrossRef]
- 28. Chen, J.; Hagos, S.; Feng, Z.; Fast, J.D.; Xiao, H. The Role of Cloud–Cloud Interactions in the Life Cycle of Shallow Cumulus Clouds. J. Atmos. Sci. 2023, 80, 671–686. [CrossRef]
- 29. Henning, S.; Hasselbring, W. A configurable method for benchmarking scalability of cloud-native applications. *Empir. Softw. Eng.* **2022**, *27*, 143. [CrossRef]
- 30. Tusa, F.; Clayman, S. End-to-end slices to orchestrate resources and services in the cloud-to-edge continuum. *Future Gener. Comput. Syst.* **2023**, 141, 473–488. [CrossRef]
- 31. Adane, M. Business-driven approach to cloud computing adoption by small businesses. *Afr. J. Sci. Technol. Innov. Dev.* **2023**, 15, 166–174. [CrossRef]
- Lagartinho-Oliveira, C.; Moutinho, F.; Gomes, L. Support Operation and Maintenance of Power Wheelchairs with Digital Twins: The IoT and Cloud-Based Data Exchange. In Proceedings of the Doctoral Conference on Computing, Electrical and Industrial Systems, Caparica, Portugal, 3–5 July 2023; pp. 191–202.
- 33. Park, J.; Han, K.; Lee, B. Green cloud? An empirical analysis of cloud computing and energy efficiency. *Manag. Sci.* 2023, 69, 1639–1664. [CrossRef]
- Seneviratne, S.; Levy, D.C.; De Silva, L.C. A Taxonomy of Performance Forecasting Systems in the Serverless Cloud Computing Environments. In Serverless Computing: Principles and Paradigms; Springer: Berlin/Heidelberg, Germany, 2023; pp. 79–120.
- ISO/IEC 25010. Available online: https://iso25000.com/index.php/en/iso-25000-standards/iso-25010 (accessed on 24 September 2023).

- Abd, S.K.; Al-Haddad, S.A.R.; Hashim, F.; Abdullah, A.B.; Yussof, S. An effective approach for managing power consumption in cloud computing infrastructure. J. Comput. Sci. 2017, 21, 349–360. [CrossRef]
- Al-Jawad, A.; Trestian, R.; Shah, P.; Gemikonakli, O. Baprobsdn: A probabilistic-based qos routing mechanism for software defined networks. In Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft), London, UK, 13–17 April 2015; pp. 1–5.
- de Oliveira, F.A., Jr.; Ledoux, T. Self-management of applications QoS for energy optimization in datacenters. In Proceedings of the Green Computing Middleware on Proceedings of the 2nd International Workshop, Lisbon, Portugal, 12 December 2011; pp. 1–6.
- Ezenwoke, A.; Daramola, O.; Adigun, M. QoS-based ranking and selection of SaaS applications using heterogeneous similarity metrics. J. Cloud Comput. 2018, 7, 1–12. [CrossRef]
- 40. Garg, S.K.; Versteeg, S.; Buyya, R. A framework for ranking of cloud computing services. *Future Gener. Comput. Syst.* 2013, 29, 1012–1023. [CrossRef]
- Ghahramani, M.H.; Zhou, M.; Hon, C.T. Toward cloud computing QoS architecture: Analysis of cloud systems and cloud services. *IEEE/CAA J. Autom. Sin.* 2017, 4, 6–18. [CrossRef]
- 42. Zheng, X.; Martin, P.; Brohman, K.; Da Xu, L. CLOUDQUAL: A quality model for cloud services. *IEEE Trans. Ind. Inform.* 2014, 10, 1527–1536. [CrossRef]
- 43. Prasad, V.K.; Bhavsar, M.D. SLAMMP framework for cloud resource management and its impact on healthcare computational techniques. *Int. J. Health Med. Commun.* **2021**, *12*, 1–31. [CrossRef]
- 44. Prasad, V.K.; Tanwar, S.; Bhavsar, M.D. Advance cloud data analytics for 5G enabled IoT. In *Blockchain for 5G-Enabled IoT: The New Wave for Industrial Automation*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 159–180.
- Didachos, C.; Kintos, D.P.; Fousteris, M.; Mylonas, P.; Kanavos, A. An optimized cloud computing method for extracting molecular descriptors. In Worldwide Congress on "Genetics, Geriatrics and Neurodegenerative Diseases Research"; Springer: Berlin/Heidelberg, Germany, 2022; pp. 247–254.
- Didachos, C.; Kintos, D.P.; Fousteris, M.; Gerogiannis, V.C.; Le Hoang, S.; Kanavos, A. A cloud-based distributed computing approach for extracting molecular descriptors. In Proceedings of the 6th International Conference on Algorithms, Computing and Systems (ICACS), Larissa, Greece, 16–18 September 2022; pp. 1–6.
- Zhu, L.; Zhuang, Q.; Jiang, H.; Liang, H.; Gao, X.; Wang, W. Reliability-aware failure recovery for cloud computing based automatic train supervision systems in urban rail transit using deep reinforcement learning. *J. Cloud Comput.* 2023, *12*, 147. [CrossRef]
- 48. Khurana, S.; Sharma, G.; Kumar, M.; Goyal, N.; Sharma, B. Reliability Based Workflow Scheduling on Cloud Computing with Deadline Constraint. *Wirel. Pers. Commun.* **2023**, *130*, 1417–1434. [CrossRef]
- 49. Qin, S.; Pi, D.; Shao, Z.; Xu, Y.; Chen, Y. Reliability-Aware Multi-Objective Memetic Algorithm for Workflow Scheduling Problem in Multi-Cloud System. *IEEE Trans. Parallel Distrib. Syst.* **2023**, *34*, 1343–1361. [CrossRef]
- 50. Khaleel, M.I. Hybrid cloud-fog computing workflow application placement: Joint consideration of reliability and time credibility. *Multimed. Tools Appl.* **2023**, *82*, 18185–18216. [CrossRef]
- Liang, J.; Ma, B.; Feng, Z.; Huang, J. Reliability-aware Task Processing and Offloading for Data-intensive Applications in Edge computing. *IEEE Trans. Netw. Serv. Manag.* 2023. [CrossRef]
- Ding, R.; Xu, Y.; Zhong, H.; Cui, J.; Sha, K. Towards Fully Anonymous Integrity Checking and Reliability Authentication for Cloud Data Sharing. *IEEE Trans. Serv. Comput.* 2023, 16, 3782–3795. [CrossRef]
- 53. Ma, H.; Li, R.; Zhang, X.; Zhou, Z.; Chen, X. Reliability-aware online scheduling for dnn inference tasks in mobile edge computing. *IEEE Internet Things J.* **2023**, *10*, 11453–11464. [CrossRef]
- Fesenko, H.; Illiashenko, O.; Kharchenko, V.; Kliushnikov, I.; Morozova, O.; Sachenko, A.; Skorobohatko, S. Flying Sensor and Edge Network-Based Advanced Air Mobility Systems: Reliability Analysis and Applications for Urban Monitoring. *Drones* 2023, 7, 409. [CrossRef]
- 55. Chamkoori, A.; Katebi, S. Security and storage improvement in distributed cloud data centers by increasing reliability based on particle swarm optimization and artificial immune system algorithms. *Concurr. Comput. Pract. Exp.* **2023**, *35*, 1. [CrossRef]
- 56. Taghavi, M.; Bentahar, J.; Otrok, H.; Bakhtiyari, K. A reinforcement learning model for the reliability of blockchain oracles. *Expert Syst. Appl.* **2023**, *214*, 119160. [CrossRef]
- 57. Xu, H.; Xu, S.; Wei, W.; Guo, N. Fault tolerance and quality of service aware virtual machine scheduling algorithm in cloud data centers. *J. Supercomput.* 2023, *79*, 2603–2625. [CrossRef]
- Zdun, U.; Queval, P.J.; Simhandl, G.; Scandariato, R.; Chakravarty, S.; Jelic, M.; Jovanovic, A. Microservice security metrics for secure communication, identity management, and observability. ACM Trans. Softw. Eng. Methodol. 2023, 32, 1–34. [CrossRef]
- 59. Ibnugraha, P.D.; Satria, A.; Nagari, F.S.; Rizal, M.F.; NonAlinsavath, K.N. The Reliability Analysis for Information Security Metrics in Academic Environment. *JOIV Int. J. Inform. Vis.* 2023, 7, 92–97. [CrossRef]
- Madavarapu, J.B.; Yalamanchili, R.K.; Mandhala, V.N. An Ensemble Data Security on Cloud Healthcare Systems. In Proceedings of the 2023 4th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 20–22 September 2023; pp. 680–686.
- Ali, M.; Jung, L.T.; Sodhro, A.H.; Laghari, A.A.; Belhaouari, S.B.; Gillani, Z. A Confidentiality-based data Classification-as-a-Service (C2aaS) for cloud security. *Alex. Eng. J.* 2023, *64*, 749–760. [CrossRef]

- Alam, M.; Shahid, M.; Mustajab, S. Security prioritized multiple workflow allocation model under precedence constraints in cloud computing environment. *Clust. Comput.* 2023, 2023, 1–36. [CrossRef]
- 63. Singh, K.K.; Jha, V.K. Security enhancement of the cloud paradigm using a novel optimized crypto mechanism. *Multimed. Tools Appl.* **2023**, *82*, 15983–16007. [CrossRef]
- Prasad, V.K.; Tanwar, S.; Bhavsar, M. C2B-SCHMS: Cloud computing and bots security for COVID-19 data and healthcare management systems. In Proceedings of the Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020, Delhi, India, 18–20 September 2021; pp. 787–797.
- 65. Chudasama, V.; Mewada, A.; Prasad, V.K.; Shah, A.; Bhavasar, M. CS2M: Cloud security and SLA management. Ann. Rom. Soc. Cell Biol. 2021, 2021, 4459–4465.
- 66. Dansana, D.; Prasad, V.K.; Bhavsar, M.; Mishra, B.K. Intensify Cloud Security and Privacy Against Phishing Attacks. *SPAST Abstr.* **2021**, *1*, 12.
- Bakshi, M.S.; Banker, D.; Prasad, V.; Bhavsar, M. SMLHADC: Security Model for Load Harmonization and Anomaly Detection in Cloud. In *Internet of Things and Its Applications: Select Proceedings of ICIA 2020*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 407–418.
- Pratyush, K.; Prasad, V.K.; Mehta, R.; Bhavsar, M. A Secure Mechanism for Safeguarding Cloud Infrastructure. In Proceedings of the International Conference on Advancements in Smart Computing and Information Security, Bhubaneswar, India, 19–20 November 2022; pp. 144–158.
- Verma, A.; Bhattacharya, P.; Prasad, V.K.; Datt, R.; Tanwar, S. AutoBots: A Botnet Intrusion Detection Scheme Using Deep Autoencoders. In Proceedings of the International Conference on Computing, Communications, and Cyber-Security, Virtual Event, 17–19 October 2022; pp. 873–886.
- 70. Abbas, Z.; Myeong, S. Enhancing Industrial Cyber Security, Focusing on Formulating a Practical Strategy for Making Predictions Through Machine Learning Tools in Cloud Computing Environment. *Electronics* **2023**, *12*, 2650. [CrossRef]
- Chahin, N.; Mansour, A. Improving the IoT and Cloud Computing integration using Hybrid Encryption. WSEAS Trans. Des. Constr. Maint. 2023, 3, 1–6. [CrossRef]
- 72. Badri, S.; Alghazzawi, D.M.; Hasan, S.H.; Alfayez, F.; Hasan, S.H.; Rahman, M.; Bhatia, S. An Efficient and Secure Model Using Adaptive Optimal Deep Learning for Task Scheduling in Cloud Computing. *Electronics* **2023**, *12*, 1441. [CrossRef]
- 73. Karamitsos, I.; Papadaki, M.; Al-Hussaeni, K.; Kanavos, A. Transforming Airport Security: Enhancing Efficiency through Blockchain Smart Contracts. *Electronics* 2023, 12, 4492. [CrossRef]
- 74. Herbst, N.; Bauer, A.; Kounev, S.; Oikonomou, G.; Eyk, E.V.; Kousiouris, G.; Evangelinou, A.; Krebs, R.; Brecht, T.; Abad, C.L.; et al. Quantifying cloud performance and dependability: Taxonomy, metric design, and emerging challenges. ACM Trans. Model. Perform. Eval. Comput. Syst. 2018, 3, 1–36. [CrossRef]
- 75. Herbst, N.; Krebs, R.; Oikonomou, G.; Kousiouris, G.; Evangelinou, A.; Iosup, A.; Kounev, S. Ready for rain? A view from SPEC research on the future of cloud metrics. *arXiv* 2016, arXiv:1604.03470.
- 76. Jangra, A.; Mangla, N. An efficient load balancing framework for deploying resource schedulingin cloud based communication in healthcare. *Meas. Sensors* 2023, 25, 100584. [CrossRef]
- 77. Cheng, Q.; Sahoo, D.; Saha, A.; Yang, W.; Liu, C.; Woo, G.; Singh, M.; Saverese, S.; Hoi, S.C. AI for IT Operations (AIOps) on Cloud Platforms: Reviews, Opportunities and Challenges. *arXiv* **2023**, arXiv:2304.04661.
- Bian, H.; Sha, T.; Ailamaki, A. Using Cloud Functions as Accelerator for Elastic Data Analytics. Proc. ACM Manag. Data 2023, 1, 1–27. [CrossRef]
- Shahid, M.A.; Alam, M.M.; Su'ud, M.M. Performance Evaluation of Load-Balancing Algorithms with Different Service Broker Policies for Cloud Computing. *Appl. Sci.* 2023, 13, 1586. [CrossRef]
- Russo, G.R.; Mannucci, T.; Cardellini, V.; Presti, F.L. Serverledge: Decentralized function-as-a-service for the edge-cloud continuum. In Proceedings of the 2023 IEEE International Conference on Pervasive Computing and Communications (PerCom), Atlanta, GA, USA, 13–17 March 2023; pp. 131–140.
- Iftikhar, S.; Ahmad, M.M.M.; Tuli, S.; Chowdhury, D.; Xu, M.; Gill, S.S.; Uhlig, S. HunterPlus: AI based energy-efficient task scheduling for cloud–fog computing environments. *Internet Things* 2023, 21, 100667. [CrossRef]
- 82. Costa, J.; Matos, R.; Araujo, J.; Li, J.; Choi, E.; Nguyen, T.A.; Lee, J.W.; Min, D. Software aging effects on kubernetes in container orchestration systems for digital twin cloud infrastructures of urban air mobility. *Drones* 2023, 7, 35. [CrossRef]
- Mahalingam, H.; Velupillai Meikandan, P.; Thenmozhi, K.; Moria, K.M.; Lakshmi, C.; Chidambaram, N.; Amirtharajan, R. Neural Attractor-Based Adaptive Key Generator with DNA-Coded Security and Privacy Framework for Multimedia Data in Cloud Environments. *Mathematics* 2023, 11, 1769. [CrossRef]
- 84. Deepika, T.; Dhanya, N. Multi-objective prediction-based optimization of power consumption for cloud data centers. *Arab. J. Sci. Eng.* **2023**, *48*, 1173–1191. [CrossRef]
- 85. Adeppady, M.; Giaccone, P.; Karl, H.; Chiasserini, C.F. Reducing microservices interference and deployment time in resourceconstrained cloud systems. *IEEE Trans. Netw. Serv. Manag.* 2023, 20, 3135–3147. [CrossRef]
- 86. Wu, W.; Lu, J.; Zhang, H. A fractal-theory-based multi-agent model of the cyber physical production system for customized products. *J. Manuf. Syst.* 2023, 67, 143–154. [CrossRef]

- Buttar, A.M.; Khalid, A.; Alenezi, M.; Akbar, M.A.; Rafi, S.; Gumaei, A.H.; Riaz, M.T. Optimization of DevOps Transformation for Cloud-Based Applications. *Electronics* 2023, 12, 357. [CrossRef]
- Golec, M.; Gill, S.S.; Parlikad, A.K.; Uhlig, S. HealthFaaS: AI based Smart Healthcare System for Heart Patients using Serverless Computing. *IEEE Internet Things J.* 2023, 10, 18469–18476. [CrossRef]
- Vonitsanos, G.; Panagiotakopoulos, T.; Kanavos, A. Issues and challenges of using blockchain for iot data management in smart healthcare. *Biomed. J. Sci. Tech. Res.* 2021, 40, 32052–32057.
- Krania, A.; Statiri, M.; Kanavos, A.; Tsakalidis, A. Internet of things services for healthcare systems. In Proceedings of the 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), Larnaca, Cyprus, 27–30 August 2017; pp. 1–6.
- 91. Hasan, M.H.; Osman, M.H.; Novia, I.A.; Muhammad, M.S. From Monolith to Microservice: Measuring Architecture Maintainability. Int. J. Adv. Comput. Sci. Appl. 2023, 14, 5. [CrossRef]
- 92. Sriraman, G.; Raghunathan, S. A Systems Thinking Approach to Improve Sustainability in Software Engineering—A Grounded Capability Maturity Framework. *Sustainability* 2023, *15*, 8766. [CrossRef]
- 93. Yu, Y.C. Smart Parking System Based on Edge-Cloud-Dew Computing Architecture. Electronics 2023, 12, 2801. [CrossRef]
- 94. Himayat, S.; Ahmad, D.J. Prediction systems for process understandability and software metrics. *SSRN* **2023**, *2023*, 4514290. [CrossRef]
- 95. Saini, M.; Choudhary, R.; Kumar, A.; Saini, D.K. Mathematical modeling and RAMD investigation of cloud infrastructure. *Int. J. Inf. Technol.* **2023**, 2023, 1–12. [CrossRef]
- Souza, L.; Camboim, K.; Araujo, J.; Alencar, F.; Maciel, P.; Ferreira, J. Dependability evaluation and sensitivity analysis of data center cooling systems. J. Supercomput. 2023, 2023, 1–29. [CrossRef]
- Pundir, M.; Sandhu, J.K.; Gupta, D.; Gupta, P.; Juneja, S.; Nauman, A.; Mahmoud, A. MD-MARS: Maintainability Framework Based on Data Flow Prediction Using Multivariate Adaptive Regression Splines Algorithm in Wireless Sensor Network. *IEEE Access* 2023, 11, 10604–10622. [CrossRef]
- 98. Mihai, I.S. A Systematic Evaluation of Microservice Architectures Resulting from Domain-Driven and Dataflow-Driven Decomposition. Bachelor's Thesis, University of Twente, Enschede, The Netherlands, 2023.
- 99. Hamid, K.; Iqbal, M.W.; Abbas, Q.; Arif, M.; Brezulianu, A.; Geman, O. Cloud Computing Network Empowered by Modern Topological Invariants. *Appl. Sci.* 2023, *13*, 1399. [CrossRef]
- Nikolaidis, N.; Arvanitou, E.M.; Volioti, C.; Maikantis, T.; Ampatzoglou, A.; Feitosa, D.; Chatzigeorgiou, A.; Krief, P. Eclipse Open SmartCLIDE: An End-to-End Framework for Facilitating Service Reuse in Cloud Development. J. Syst. Softw. 2023, 2023, 111877. [CrossRef]
- Abraham, A.; Yang, J. A Comparative Analysis of Performance and Usability on Serverless and Server-Based Google Cloud Services. In Proceedings of the International Conference on Advances in Computing Research, Orlando, FL, USA, 8–10 May 2023; pp. 408–422.
- Saleem, M.; Warsi, M.; Islam, S. Secure information processing for multimedia forensics using zero-trust security model for large scale data analytics in SaaS cloud computing environment. *J. Inf. Secur. Appl.* 2023, 72, 103389. [CrossRef]
- 103. Hong, F.; Wang, L.; Li, C.Z. Adaptive mobile cloud computing on college physical training education based on virtual reality. *Wirel. Netw.* **2023**, 2023, 1–24. [CrossRef]
- 104. Fazel, E.; Shayan, A.; Mahmoudi Maymand, M. Designing a model for the usability of fog computing on the internet of things. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 5193–5209. [CrossRef]
- 105. Spichkova, M.; Schmidt, H.W.; Yusuf, I.I.; Thomas, I.E.; Androulakis, S.; Meyer, G.R. Towards modelling and implementation of reliability and usability features for research-oriented cloud computing platforms. In *Proceedings of the Evaluation of Novel Approaches to Software Engineering: 11th International Conference, ENASE 2016, Rome, Italy, 27–28 April 2016, Revised Selected Papers 11*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 158–178.
- 106. Agarwal, R.; Dhingra, S. Factors influencing cloud service quality and their relationship with customer satisfaction and loyalty. *Heliyon* **2023**, *9*, 4. [CrossRef] [PubMed]
- Wu, X.; Jin, Z.; Zhou, J.; Duan, C. Quantum Walks-based Classification Model with Resistance for Cloud Computing Attacks. Expert Syst. Appl. 2023, 2023, 120894. [CrossRef]
- 108. Vlachogianni, P.; Tselios, N. Perceived Usability Evaluation of Educational Technology Using the Post-Study System Usability Questionnaire (PSSUQ): A Systematic Review. *Sustainability* **2023**, *15*, 12954. [CrossRef]
- 109. Tan, J.; Shao, L.; Lam, N.Y.K.; Toomey, A.; Chan, H.H.; Lee, C.; Feng, G.Y. Evaluating the usability of a prototype gesture-controlled illuminative textile. *J. Text. Inst.* 2023, 2023, 1–7. [CrossRef]
- 110. Monks, T.; Harper, A. Improving the usability of open health service delivery simulation models using Python and web apps. *NIHR Open Res.* **2023**, *3*, 48. [CrossRef]
- Vonitsanos, G.; Panagiotakopoulos, T.; Kanavos, A.; Kameas, A. An Apache Spark Framework for IoT-enabled Waste Management in Smart Cities. In Proceedings of the 12th Hellenic Conference on Artificial Intelligence, Corfu, Greece, 7–9 September 2022; pp. 1–7.
- 112. Lingaraju, A.K.; Niranjanamurthy, M.; Bose, P.; Acharya, B.; Gerogiannis, V.C.; Kanavos, A.; Manika, S. IoT-Based Waste Segregation with Location Tracking and Air Quality Monitoring for Smart Cities. *Smart Cities* 2023, 6, 1507–1522. [CrossRef]

- 113. Ennis, S.F.; Evans, B. Cloud Portability and Interoperability under the EU Data Act: Dynamism versus Equivalence. *SSRN* 2023. [CrossRef]
- 114. Olabanji, D.; Fitch, T.; Matthew, O. Cloud-native architecture Portability Framework Validation and Implementation using Expert System. *Int. J. Adv. Stud. Comput. Sci. Eng.* 2023, 12, 4.
- 115. Barnes, K.M.; Buyskikh, A.; Chen, N.Y.; Gallardo, G.; Ghibaudi, M.; Ruszala, M.J.; Underwood, D.S.; Agarwal, A.; Lall, D.; Runggar, I.; et al. Optimising the quantum/classical interface for efficiency and portability with a multi-level hardware abstraction layer for quantum computers. *EPJ Quantum Technol.* 2023, 10, 36. [CrossRef]
- 116. Malahleka, M. The right to data portability: A ghost in the protection of personal information. J. S. Afr. Law 2023, 2023, 1. [CrossRef]
- 117. Islam, R.; Patamsetti, V.; Gadhi, A.; Gondu, R.M.; Bandaru, C.M.; Kesani, S.C.; Abiona, O. The Future of Cloud Computing: Benefits and Challenges. *Int. J. Commun. Netw. Syst. Sci.* **2023**, *16*, 53–65. [CrossRef]
- Jeon, D.S.; Menicucci, D.; Nasr, N. Compatibility Choices, Switching Costs, and Data Portability. Am. Econ. J. Microeconomics 2023, 15, 30–73. [CrossRef]
- 119. Kaur, K.; Bharany, S.; Badotra, S.; Aggarwal, K.; Nayyar, A.; Sharma, S. Energy-efficient polyglot persistence database live migration among heterogeneous clouds. *J. Supercomput.* **2023**, *79*, 265–294. [CrossRef]
- Pacheco, J.A.; Rasmussen, L.V.; Wiley, K., Jr.; Person, T.N.; Cronkite, D.J.; Sohn, S.; Murphy, S.; Gundelach, J.H.; Gainer, V.; Castro, V.M.; et al. Evaluation of the portability of computable phenotypes with natural language processing in the eMERGE network. *Sci. Rep.* 2023, 13, 1971. [CrossRef] [PubMed]
- Mpofu, P.; Kembo, S.H.; Chimbwanda, M.; Jacques, S.; Chitiyo, N.; Zvarevashe, K. A privacy-preserving federated learning architecture implementing data ownership and portability on edge end-points. *Int. J. Ind. Eng. Oper. Manag.* 2023, *ahead-of-print*. [CrossRef]
- 122. Hosseini, L.; Kumar, S. Is Multi-Cloud the Future? Desirability of Compatibility in Cloud Computing Market. *Desirability Compat. Cloud Comput. Mark.* 2023, 5, 7.
- Mohiuddin, K.; Islam, A.; Islam, M.A.; Khaleel, M.; Shahwar, S.; Khan, S.A.; Yasmin, S.; Hussain, R. Component-centric mobile cloud architecture performance evaluation: An analytical approach for unified models and component compatibility with next generation evolving technologies. *Mob. Netw. Appl.* 2023, 28, 254–271. [CrossRef]
- 124. Yakubu, A.S.; Kassim, A.M.; Husin, M.H. Conceptualizing hybrid model for influencing intention to adopt cloud computing in North-Eastern Nigerian academic libraries. *J. Acad. Librariansh.* **2023**, *49*, 102747. [CrossRef]
- 125. Lall, A.; Tallur, S. Deep reinforcement learning-based pairwise DNA sequence alignment method compatible with embedded edge devices. *Sci. Rep.* **2023**, *13*, 2773. [CrossRef]
- 126. Chi, C.; Liu, Y.; Ma, B.; Chai, S.; Zhang, P.; Yin, Z. A compatible carbon efficiency information service framework based on the industrial internet identification. *Digit. Commun. Netw.* **2023**, *in press.* [CrossRef]
- 127. Bardsiri, A.K.; Hashemi, S.M. QoS Metrics for Cloud Computing Services Evaluation. Int. J. Intell. Syst. Appl. 2014, 6, 27. [CrossRef]
- 128. Orhean, A.I.; Pop, F.; Raicu, I. New Scheduling Approach using Reinforcement Learning for Heterogeneous Distributed Systems. *J. Parallel Distrib. Comput.* **2018**, 117, 292–302. [CrossRef]
- 129. Cui, D.; Peng, Z.; Xiong, J.; Xu, B.; Lin, W. A Reinforcement Learning-Based Mixed Job Scheduler Scheme for Grid or IaaS Cloud. *IEEE Trans. Cloud Comput.* 2020, *8*, 1030–1039. [CrossRef]
- Liu, N.; Li, Z.; Xu, J.; Xu, Z.; Lin, S.; Qiu, Q.; Tang, J.; Wang, Y. A Hierarchical Framework of Cloud Resource Allocation and Power Management Using Deep Reinforcement Learning. In Proceedings of the 37th IEEE International Conference on Distributed Computing Systems (ICDCS), IEEE Computer Society, Atlanta, GA, USA, 5–8 June 2017; pp. 372–382.
- 131. Rethinking Monitoring for Container Operations. Available online: https://thenewstack.io/monitoring-reset-containers/ (accessed on 24 September 2023).
- Kardani-Moghaddam, S.; Buyya, R.; Ramamohanarao, K. Performance-Aware Management of Cloud Resources: A Taxonomy and Future Directions. ACM Comput. Surv. 2019, 52, 1–37.
- 133. Erl, T.; Puttini, R.; Mahmood, Z. Cloud Computing: Concepts, Technology & Architecture; Pearson Education: London, UK, 2013.
- 134. Peiris, C.; Balachandran, B.; Sharma, D. Governance Framework for Cloud Computing. GSTF J. Comput. 2014, 1, 14. [CrossRef]
- 135. Anithakumari, S.; Chandrasekaran, K. Adaptive Resource Allocation in Interoperable Cloud Services. In *Proceedings of the Advances in Computer Communication and Computational Sciences*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 229–240.
- 136. Prasad, V.K.; Bhavsar, M. Efficient Resource Monitoring and Prediction Techniques in an IaaS Level of Cloud Computing: Survey. In Proceedings of the 1st International Conference on Future Internet Technologies and Trends (ICFITT), Surat, India, 31 August–2 September 2018; pp. 47–55.
- 137. de Farias, C.M.; Pirmez, L.; Delicato, F.C.; Pires, P.F.; Guerrieri, A.; Fortino, G.; Cauteruccio, F.; Terracina, G. A multisensor data fusion algorithm using the hidden correlations in Multiapplication Wireless Sensor data streams. In Proceedings of the 2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC), Falerna, Italy, 16–18 May 2017; pp. 96–102. [CrossRef]
- 138. McDonald, D.; Breslin, C.; MacDonald, A. Review of the Environmental and Organisational Implications of Cloud Computing: Final Report; University of Strathclyde: Glasgow, UK, 2010.

- Rimal, B.P.; Choi, E.; Lumb, I. A Taxonomy and Survey of Cloud Computing Systems. In Proceedings of the International Conference on Networked Computing and Advanced Information Management (NCM), IEEE Computer Society, Seoul, Republic of Korea, 25–27 August 2009; pp. 44–51.
- 140. Kim, H.A.H.H.; Barua, S. Service Level Agreement (SLA) for Cloud Computing Compilation with Common and New Formats. *Int. J. Sci. Res. Manag.* **2018**, *6*, 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.