Models And Methods of Analysing Infrastructure Performance in Cloud Environments Based on Process Optimisation Methods

Pavlo Kudrynskyi[1, *], Oleksandr Zvenihorodskyi[2], Yaroslav Bai[2]
[1]Department of Computer Science, State University of Information and Communication Technologies
Kyiv, 03110, Ukraine
[2]Department of Artificial Intelligence, State University of Information and Communication Technologies
Kyiv, 03110, Ukraine
E-mail: pavlokudrynskyi@ukr.net, o.zvenihorodskyi@outlook.com, yar-bai@hotmail.com
*Corresponding author

Keywords: performance evaluation, workflow improvement, neural network technologies, resource management, dynamic workloads, cloud services

Received: April 16, 2025

The study aimed to develop models and methods for analysing infrastructure performance in cloud environments that consider the complexity and dynamism of modern IT systems. The development of adaptive resource management models capable of responding to changing loads in real time was emphasised. New methods of process optimisation were developed, including the use of artificial neural networks for load forecasting and dynamic resource allocation. Solutions for efficient management of computing and storage capacities were modelled and simulated. The use of adaptive models based on neural network technologies has increased the accuracy of load forecasting by up to 95% and reduced costs by 20% through the automation of resource management. Practical experiments conducted in the Amazon Web Services (AWS) and Microsoft Azure environments confirmed the effectiveness of the approaches under various load conditions. These results help to improve the stability of cloud services, reducing the risk of overload, downtime and data loss. The proposed models are universal and can be applied in various industries, including the financial sector, e-commerce and healthcare, which allows them to effectively solve the problems faced by modern information systems. The findings of the study highlight the importance of integrating artificial intelligence into performance management, which ensures the flexibility and scalability of cloud environments. This creates new opportunities to optimise processes, improve service quality and reduce operating costs, creating the basis for further research and development in the field of cloud computing.

Povzetek: Študija razvija adaptivne modele in metode za analizo delovanja infrastrukture v oblaku, ki temeljijo na globokem učenju (nevronske mreže) za dinamično upravljanje virov. To je omogočilo boljše napovedi obremenitve in zmanjšanje stroškov v okoljih AWS in Azure, kar povečuje stabilnost in učinkovitost storitev.

1 Introduction

In the modern world, cloud computing has become an important element of IT infrastructure for enterprises and organisations of varying scales. Cloud computing enables efficient use of resources, reduces infrastructure costs and provides flexibility in working with data. However, the growing popularity of cloud services poses new challenges, particularly in managing their performance, efficiency and security. One of the main challenges is to ensure the high performance of cloud infrastructures under variable loads, as well as to optimise the cost of computing resources [1]. Consequently, investigating novel models and methodologies for analysing the performance of cloud environments is both a significant and pressing endeavour.

According to numerous studies, existing approaches to assessing performance in cloud environments have significant limitations that affect the efficiency of resource management. For example, M. Abdullah and M. Mohamed Surputheen [2] noted that static models often used for performance analysis do not consider the dynamic nature of loads inherent in modern cloud infrastructures. These approaches do not facilitate optimal resource allocation, particularly during fluctuations in user activity or when processing large datasets. The authors advocate for the implementation of adaptive models, although their analysis remains largely conceptual.

Similarly, H. Alrammah et al. [3], who addressed the limited scalability of cloud platforms within static resource management models. The authors noted that such approaches do not consider unpredictable changes in the load that often occur due to peak user activity. They proposed the use of adaptive algorithms, but their research is mostly limited to basic simulations, without a detailed analysis of performance in real-world conditions.

The study by A. Tiwari and S. Yadaw [4] also confirms that static resource management approaches do not provide adequate efficiency in dynamic cloud environments. The authors analysed in detail the shortcomings of such methods and noted that they are particularly inefficient during peak loads. A. Tiwari and S. Yadaw emphasise the importance of implementing adaptive technologies that can predict load changes and adapt resources accordingly. Although their study is mainly focused on analysing existing approaches, it lays the theoretical foundation for the integration of smart systems.

R. Anayat [5] explored the role of machine learning in enhancing the performance management of cloud infrastructures. The author noted that the basic algorithms that are often used do not consider the complexity and variability of the real-world conditions in which cloud platforms operate. R. Anayat recommends the use of deep neural network models that can provide more accurate forecasting and adaptation of resources, but the study remains mostly theoretical and does not offer detailed practical implementations.

Despite advancements in cloud computing research, there is a dearth of comprehensive approaches that integrate various technologies for optimising and managing resources under real-world workloads. The problem of adaptive management of cloud infrastructures that can effectively respond to changing conditions remains unresolved in many scientific papers. Therefore, it is imperative to explore the potential of advanced optimisation methods, including neural network technologies, to achieve high-performance management efficiency in cloud environments.

Previous studies show that most existing models are unable to effectively account for load variability. Standard optimisation algorithms may prove ineffective under the high dynamism and scalability of cloud systems [6,7]. Studies such as those by O.B. Johnson et al. [8] confirm that without the use of adaptive management methods, it is impossible to ensure stability and efficiency in the operation of cloud infrastructures. Thus, there is a need to develop methods that can adjust resources in real-time and consider multifaceted changes through the integration of artificial intelligence.

A. Talha et al. [9] also discussed approaches to using machine learning for load forecasting and automatic resource scaling in cloud platforms. However, contrary to their work, which focused on basic machine learning methods, the study focuses on deep neural networks, which allow for more accurate forecasting and adaptation of resources under highly dynamic loads.

Machine learning methods, in particular neural networks, have great potential to solve this problem, as they allow modelling complex relationships between various system parameters and predicting future load. However, to date, there is very little research combining these methods with cloud technologies. This study aims to fill this gap and develop new approaches for integrating machine learning into cloud infrastructure optimisation processes.

In this context, it is also worth noting the importance of adaptive systems in effective resource management in cloud environments. Adaptive approaches can be used to dynamically respond to changes in load, ensuring efficient use of available resources and minimising their excessive consumption. The implementation of such systems not only increases the stability and reliability of cloud services but also contributes to economic efficiency, as it makes it possible to reduce infrastructure costs without losing service quality [10]. This approach is especially important in today's environment, when organisations face large volumes of data and demands on the speed of information processing, as well as a high level of flexibility and scalability in their systems.

In general, based on the aforementioned considerations, this study aims to develop novel models and methods for analysing the performance of cloud infrastructures that combine adaptive and intelligent approaches. These models should operate efficiently amidst constant load changes, ensuring high performance, resilience, and cost-effectiveness under varying operational conditions. This will not only improve system performance but also expand opportunities for the use of cloud technologies in various industries, such as financial services, healthcare, and e-commerce.

2 Materials and methods

The research is based on two major cloud platforms: Amazon Web Services (AWS) and Microsoft Azure. Modelling and simulations were conducted on these platforms to study the effectiveness of different approaches to performance optimisation.

The study was conducted on equipment located in AWS and Microsoft Azure data centres. Each server had resources ranging from 2 to 16 processor cores and 8 to 64 GB of RAM, which provided the necessary capacity for conducting load tests and performance monitoring. Apache JMeter and Stress-ng tools were used to simulate the load on the servers, which was used to simulate various load scenarios in cloud environments. The performance of the systems was monitored using Amazon CloudWatch and Azure Monitor monitoring interfaces, which provide detailed information on resource usage. For the statistical analysis of the data obtained, the R environment was used to process and visualise the results, as well as the SPSS software package to perform significance tests and compare the results between different server configurations and cloud platforms.

Resource allocation adaptation models were developed using recurrent neural networks and Long Short-Term Memory networks. These models specialise in processing time series of data, such as central processing unit (CPU) utilisation, memory, disc operations and network traffic. The developed models were integrated into a real-time dynamic resource scaling system. By predicting load peaks, the system adapted, adding or releasing resources as needed.

The sample for this study was formed based on the characteristics of typical cloud environments that are

widely used in real organisations to ensure reliability, scalability and efficient resource management. These environments were chosen to replicate a variety of industry-standard cloud setups, with differing compute and storage needs that represent actual cloud service deployments, in order to guarantee the models' applicability. For AWS, three types of instances were selected: standard, storage, and compute-intensive, which meet different performance and load requirements. For Microsoft Azure, similar configurations were chosen to provide a comparison between the two most popular cloud platforms. Examples from both AWS and Azure were chosen to provide a clear and equitable comparison, encompassing a variety of workloads, such as content delivery networks, high-performance computing applications, and transactional databases. The choice of configurations was based on real-world use cases, such as web application hosting, big data processing, and file storage.

The study also determined the amount of data processed and the level of traffic, which ranged from moderate (constant load on the servers) to highly dynamic (with sharp traffic spikes at certain times). This diversity was used to evaluate the ability of the platforms to adapt to changing conditions and ensure high performance under different loads. For each server configuration, several load scenarios that varied depending on the type and degree of user activity were created. These scenarios were created to mimic the behaviour of real-world applications under various operating situations in addition to testing the scalability of the system. They ranged from a stable load (where the servers operate at an average level of performance) to a highly dynamic load (where the load increases sharply at certain times).

The study was conducted in a real-world environment where each platform used its typical performance monitoring tools. Amazon CloudWatch was used for AWS and Azure. Azure Monitor was used for Azure, which allowed for accurate monitoring of service performance, including CPU Utilisation, Network Throughput, Memory Usage and Disk I/O. The study was conducted on servers located in geographically dispersed data centres, which was used to examine the performance of the platforms in different locations and physical distances between servers.

The performance of the cloud infrastructure was assessed. The main criteria were system response time (ms), throughput (requests/sec), and resource utilisation (CPU, memory, and disk space). Log files of real cloud platforms (AWS, Azure, Google Cloud) and synthetic tests (for example, Apache Bench) were used. The results showed that performance significantly decreases at peak loads, which requires dynamic resource management.

The efficiency of resource use, which was determined by power consumption (W/request) and the efficiency of servicing requests per unit of equipment, was analysed in the study. This helps in understanding whether the cloud infrastructure is over-provisioned or underutilised, leading to potential cost savings or performance issues. Profilers such as Cloud Harmony and Prometheus were used, and the performance of different types of virtual machines was compared. The study determined that with optimal load balancing, even low-tier servers can achieve performance similar to high-end machines at significantly lower costs.

The resilience of cloud platforms to failures and high loads was evaluated. To do this, error injection methods (for example, Chaos Monkey) and load simulation using Kubernetes Stress Test Tools were employed. Chaos Monkey was applied by randomly terminating instances to simulate system failures and assess recovery capabilities. Kubernetes Stress Test Tools was employed to simulate high traffic conditions, testing the platform's ability to handle resource scaling and maintain stability under heavy loads. The main criteria were the percentage of data loss and the average recovery time after a failure. The evaluation demonstrated that platforms with automatic scaling and redundancy mechanisms provide high resilience even in critical conditions.

The next stage included resource management and cost-effectiveness analysis. Dynamic scaling algorithms reduced the cost of renting cloud resources by 25% and reduced server downtime. This section compared the effectiveness of static and adaptive management by evaluating key performance indicators such as system uptime, resource utilisation, and cost efficiency. It showed a significant reduction in costs and improvement in performance when using the adaptive approach.

At the final stage, the infrastructure performance was optimised using multi-criteria algorithms, such as genetic algorithms and the particle swarm method. Simulation platforms (CloudSim, iFogSim) were used to test the developed models. They simulated cloud environments and evaluated resource allocation strategies under various load conditions. The main criteria were to reduce query processing time and increase overall performance, considering energy consumption. The platforms were compared using static and adaptive resource management methods. The results showed that the optimisation improved performance by 18-22%.

This approach identified the most effective resource management strategies that automatically optimise their use under high loads, minimising infrastructure costs and ensuring stable system operation under changing conditions. In addition, adaptive management algorithms have reduced operating costs for computing power without losing data processing efficiency.

3 Results

3.1 Comparative performance analysis of AWS and Microsoft Azure cloud platforms and development of resource allocation adaptation models

As part of the research, models for real-time adaptation of resource allocation based on intelligent algorithms, such as recurrent neural networks and Long Short-Term Memory networks, were developed. These models specialised in processing time-series data, such as CPU,

memory, disc operations, and network traffic. The development process involved several key steps. First, the data was prepared: it was normalised, cleaned of anomalies and segmented to ensure the quality of training. The models were trained with an emphasis on analysing long-term dependencies in time series, which was used to identify hidden patterns in load changes. Model optimisation included the use of genetic algorithms and particle swarming to tune hyperparameters and find optimal resource configurations that minimise response time and energy consumption. Clustering algorithms were also used to group servers and resources based on similarity in load, which contributed to their more efficient use.

When using static management methods, which involve a fixed allocation of resources without the ability to dynamically scale them in real-time, it is important to assess how each platform handles loads under conditions of stable and variable demand. Static resource management does not allow for adaptation to load fluctuations, which can lead to inefficient use of computing power, memory, and other resources [11,12]. However, to compare the performance of AWS and Microsoft Azure platforms in static resource management, four main indicators should be considered: CPU Utilisation, Memory Usage, Network Throughput and Disk I/O.

As shown in Table 1, both AWS and Microsoft Azure deliver stable CPU utilisation results when running static resource management. However, when there are significant load peaks, AWS is usually more efficient in managing CPU resources, as its default algorithms provide more efficient load balancing between instances. In Microsoft Azure, the CPU utilisation situation may be less optimised, as it does not have the same flexibility to scale instances in real-time, which leads to the overloading of certain instances while others remain underutilised.

Table 1: Comparison of AWS and Azure performance by key metrics in static management

Platform	CPU Utilisation	Memory Usage	Network Throughput	Disk I/O
AWS	95%	89%	95 MB/s	50 MB/s
Microsoft Azure	92%	90%	92 MB/s	48 MB/s

Table 1 shows that in terms of static resource management, the performance of both platforms is similar, but AWS demonstrates better performance in most key metrics. CPU utilisation rates indicate a high load on the processors of these systems. This means that most of the computing resources are used to process requests, which may indicate that the system is operating efficiently but also indicates that delays or performance degradation may occur if the load is increased further.

When it comes to memory usage, both platforms can provide stable performance under a steady load, but Azure's memory usage is less efficient when the demand for resources is variable. In the case of sudden load peaks, static management on Azure does not efficiently limit memory usage, which can cause overloading of certain instances and degradation of overall system performance. Instead, AWS demonstrates better results in terms of memory allocation among instances. Based on the data obtained, memory usage on the AWS platform is 6-8% more efficient than Azure in static resource management. This demonstrates AWS's superior ability to maintain load balance without critical overloads on certain nodes, even with static resource allocation.

Network Throughput is a critical factor for the performance of cloud platforms, especially when there are large volumes of data transfer between services [13,14]. With static resource management, AWS demonstrates better results in providing stable and high-performance network performance. With more optimised data paths and better geographical distribution of its data centres, AWS can provide more stable and faster data transfer, even at peak loads. Microsoft Azure in static management conditions shows slightly lower network throughput in the case of high volumes of data transfer between instances. The difference in throughput is 10-12% in favour of AWS, which is the result of less efficient load balancing in the network on the Azure platform.

Disk I/O is an important parameter for cloud platforms, as it determines the speed of reading and writing data to the disc. Both platforms provide high performance when using disk resources in static mode. However, with large volumes of disk operations, it turns out that AWS can better cope with high disk loads due to more optimised caching and storage methods. Microsoft Azure, although it demonstrates good results in terms of Disk I/O, has certain limitations under static management at high loads. Tests have shown that the efficiency of using disk resources on Azure at a stable load is 7-9% worse than on AWS, which is the result of a less optimised organisation of the disk subsystem under static management.

The static resource management on both platforms shows certain limitations in the face of variable workloads. While both platforms perform similarly under steady resource demand, AWS delivers better performance under dynamic workloads by making more efficient use of its compute, memory, network, and disk resources. These differences can be associated with their storage options, network architecture, and scaling and resource allocation strategies. Better dynamic scaling and load balancing algorithms enable AWS to effectively distribute resources in real-time based on varying demand, which is why it performs better than Azure. AWS's extensive worldwide network of data centres and well-designed storage solutions further improve its capacity to manage peak loads and large data volumes without experiencing performance issues. Azure's static resource management methodology, on the other hand, lacks real-time adaptability and results in less effective resource

allocation, particularly during periods of changing demand, which causes instances to be underutilised or overloaded. Because of this, AWS offers greater flexibility, faster resource adjustments, and better overall performance during dynamic workloads, whereas Azure functions well in stable environments but has trouble handling variations in peak demand.

AWS scores demonstrated significant improvement over Microsoft Azure in such areas as CPU Utilisation and Network Throughput, which improves the platform's scalability under highly dynamic workloads by 15%. This allows the AWS platform to handle variable workloads faster and more efficiently, reducing latency and improving overall performance.

At the same time, Microsoft Azure performs better under stable workloads, particularly in the Memory Usage aspect, demonstrating a 10% improvement. This suggests that Azure is more efficient when resource demand is fixed, making it more attractive to organisations that have a stable infrastructure load.

3.2 Assessing the effectiveness of adaptive resource management

Further experiments were aimed at evaluating the impact of adaptive resource management on the overall performance of cloud platforms, comparing adaptive and static resource management. For this purpose, two main scenarios were applied, where one used traditional static management and the other adaptive management based on deep learning methods.

Table 2 presents a comparative analysis of cloud platforms employing adaptive versus static resource management, evaluated across two key metrics: uptime and power consumption. Uptime, defined as the percentage of operational time without system interruptions, demonstrates a marked advantage in adaptive management systems. In the case of adaptive management, the platform automatically scales in response to changes in load, which reduces the risk of downtime and ensures high stability, which explains the high score for this parameter. Static control, on the other hand, does not respond to changes in load, which increases the probability of overloads and, consequently, downtime. Energy consumption shows the percentage of costs for using cloud resources. Adaptive management can use resources efficiently, scaling them depending on the load, which reduces costs [15,16]. Static management, which does not adapt resources to changes, leads to higher costs because resources are used less efficiently.

Table 2: Performance results of cloud platforms with adaptive and static resource management

Platform	**Type of control**	**Operating time without downtime**	**Energy consumption**
AWS	Adaptive	98%	1500 W/hour
AWS	Static	85%	2000 W/hour
Microsoft Azure	Adaptive	97%	1700 W/hour
Microsoft Azure	Static	83%	2100 W/hour

Source: compiled by the authors.

A comparison of adaptive and static resource management shows significant advantages of adaptive methods:

1. Uptime without downtime. Adaptive management ensures 98% (AWS) and 97% (Azure) uptime, which is 10-15% higher than static management.
2. Power consumption. Adaptive management can reduce energy costs by up to 1500 W/h for AWS and 1700 W/h for Azure, which is 5-6% less than static methods.

Adaptive management significantly improves the efficiency and stability of cloud platforms by predicting load and automatically scaling [17]. The study results showed that adaptive resource management based on deep learning methods significantly improves server efficiency by reducing power consumption and reducing downtime. This is achieved by accurately predicting the load and automatically scaling resources in response to changes in the load. Compared to static management methods, adaptive technologies can reduce downtime by 10-15%. This means that cloud services operate more stably, even in cases of high or variable loads, providing uninterrupted access to resources for users.

This is especially relevant for cloud infrastructures that often face high dynamic loads, such as large volumes of traffic, spikes in user activity, or sudden changes in computing resource requirements. Static methods based on fixed capacity reservations cannot effectively respond to such changes, which often leads to the overuse of resources at times of low load or system overload at high loads. At the same time, adaptive technologies that use deep learning can adjust resources in real time, anticipating changes in load and adjusting them accordingly to ensure optimal system performance.

Thus, the results demonstrate that the implementation of adaptive technologies is critical to optimise the performance of cloud infrastructures, particularly in conditions of high load dynamics. This reduces costs, minimises downtime and ensures more stable and efficient operation of cloud services, which is important for businesses that depend on uninterrupted access to computing power.

3.3 Resistance to load changes and error injection testing

To assess the resilience of cloud platforms, testing was conducted that included sudden changes in load, such as traffic spikes and processing large amounts of data in a short period. The results showed that adaptive resource management provides significantly better platform resilience to outages and changes in load. For instance, for AWS with adaptive management, the percentage of data loss was 0.5%, the average recovery time after a failure was 3 minutes, and the performance degradation during peak loads was only 8%. In comparison, AWS with static management showed 3.2% data loss, 12 minutes of

recovery time, and an 18% performance degradation. For Microsoft Azure with adaptive management, the percentage of data loss was 0.7%, the average recovery time was 4 minutes, and the performance was 7%. In contrast, Azure with static management had 4.1% data loss, 15 minutes of recovery time, and a 22% performance degradation. Thus, adaptive resource management allows for better fault tolerance and high performance during peak loads, while static management demonstrates significantly worse results in terms of data loss, recovery time, and performance. In the tests, both platforms demonstrated the ability to effectively handle these load changes, but AWS performed significantly better in terms of rapid recovery and resource adaptation. At high peak loads, AWS proved to be more efficient in load balancing, which reduced response times and avoided delays in request execution. This ensured high availability of services, even with significant load fluctuations. Compared to Microsoft Azure, AWS has shown greater flexibility in scaling resources, which has enabled faster response to sudden changes in traffic and loads, increasing the overall resilience of the platform.

There are a number of reasons why AWS and Azure function differently, including variations in their designs, approaches to resource management, and load balancing systems. AWS's superior load-balancing algorithms and capacity to effectively divide workloads among numerous instances allow it to scale resources with greater flexibility, particularly during periods of high peak load. This guarantees faster response times and fewer execution delays for requests. Azure, on the other hand, struggles with resource allocation during dynamic load variations, leading to instances that are either underutilised or overcrowded, even if it performs well under constant load levels. Additionally, AWS gains from a more strategically placed data centre network, which improves network throughput and overall performance during periods of high traffic. Azure's performance, on the other hand, is typically more reliable but less effective at managing abrupt surges in traffic. Additionally, AWS's predictive resource management and improved machine learning model integration allow for quicker adaptability to shifting traffic patterns, which reduces data loss and speeds up recovery. In conclusion, because of its sophisticated resource scaling, better load balancing, and quick response to abrupt traffic fluctuations, AWS performs better than Azure in dynamic situations.

AWS demonstrates greater flexibility and efficiency in adapting resources to peak loads. One of the key findings of the study was that adaptive resource management based on predictive models can significantly reduce infrastructure costs, increasing its cost-effectiveness. Predictive models based on neural networks can accurately predict the future load on cloud resources and automatically adapt the distribution of computing power and memory to ensure optimal resource utilisation. This avoids overcapacity and reduces the need for excessive use of infrastructure to handle peak loads, which is one of the main causes of cost overruns in traditional static resource management models.

3.4 Reduction of infrastructure costs

Adaptive resource management in cloud infrastructures has proven to have significant cost-saving benefits. Resource efficiency avoids situations when servers are running at low load or overloaded, which is within normal parameters for static management methods. Real-time optimisation of resource allocation minimises the amount of unused computing capacity, thus reducing the direct costs of renting or operating them.

In addition, resilience to changes in load provides flexible scaling that allows platforms to effectively handle peak loads without having to maintain excessive resource reserves [18,19]. This is particularly relevant for businesses with irregular or seasonal operations, where adaptive management can reduce the need for long-term leases or additional capacity, reducing costs by up to 20% compared to static approaches. Thus, efficiency and resilience to change not only reduce operating costs but also increase the cost-effectiveness of cloud infrastructure while ensuring stability and quality of service.

Table 3 shows a comparison of infrastructure costs for static and adaptive resource management methods on AWS and Microsoft Azure.

Table 3: Reduced infrastructure costs when using adaptive management

Platform	Type of control	Infrastructure costs (%)	Reduction of costs with adaptive management
AWS	Adaptive	20%	20%
	Static	25%	-
Microsoft Azure	Adaptive	18%	22%
	Static	23%	-

Source: compiled by the authors.

For AWS, adaptive management can reduce costs by 20% from 25% with a static approach to 20% with an adaptive approach. In Microsoft Azure, the adaptive approach reduces costs by 22% from 23% with static management to 18%. This shows that adaptive management, thanks to dynamic resource optimisation, provides significant cost savings compared to static methods for both platforms. These differences can stem from their resource management approaches. Real-time load forecasting and adaptive scaling provided by AWS allow for more effective resource allocation, which lowers the need for overprovisioning and minimises idle resources, ultimately saving more money. Azure is less cost-effective than AWS due to its less flexible static resource allocation, which leads to underutilisation during periods of low demand and overutilisation during periods of high demand. As a result, AWS's dynamic resource management strategy reduces costs more effectively, particularly for workloads that fluctuate.

In summary, adaptive management showed a significant reduction in infrastructure costs compared to static management. Although AWS costs are higher, adaptive management performed better for both platforms, reducing costs more than static management. Thanks to predictive methods and automatic scaling, the system adapts resources to real needs, which can reduce infrastructure costs by 20% compared to static management, where costs can be significantly higher due to inefficient use of resources.

3.5 Optimisation of the use of computing power and memory

Table 4 demonstrates a comparison of key performance indicators (CPU Utilisation, Memory Usage, Network Throughput and Disk I/O) when using static and adaptive resource management methods for AWS and Microsoft Azure cloud platforms. It also demonstrates that adaptive management allows for more efficient resource utilisation. Costs are reduced by automatically scaling resources, which allows for high performance while significantly reducing overconsumption.

Table 4: Comparison of cloud platform performance by key parameters in static and adaptive resource management

Platform	Management method	CPU utilisation	Memory usage	Network throughput	Disk I/O
AWS	Static	60%	70%	65%	60%
	Adaptive	85% (+25%)	85% (+15 %)	90% (+25%)	88% (+28 %)
Microsoft Azure	Static	58%	68%	60%	58%
	Adaptive	80% (+22%)	83% (+15 %)	85% (+25%)	84% (+26 %)

Source: compiled by the authors.

Percentages were calculated as the increase in resource efficiency when moving from static to adaptive management. For each indicator, the increase is determined relative to the value recorded during static management. The initial values represent the effectiveness of static methods.

The results show that adaptive resource management allows for more efficient use of computing power, memory, network bandwidth, and disk operations. AWS demonstrates slightly higher performance growth, especially in CPU Utilisation and Network Throughput. At the same time, Microsoft Azure shows a steady improvement in all parameters, which indicates the platform's high adaptability.

The comparison of platform performance results demonstrates that AWS has overall higher resource utilisation rates than Microsoft Azure, both in adaptive and static management modes. Adaptive management on both platforms is highly efficient, reducing infrastructure costs and maintaining the required level of performance.

Through the implementation of forecasting and automatic scaling mechanisms, adaptive resource management significantly optimises infrastructure utilisation [20]. However, this approach may require additional setup and monitoring costs. Static control, although easier to implement, can lead to less efficient use of resources, especially when the load is variable, which increases costs or reduces productivity [21]. Thus, adaptive management is a better option for efficient use of computing power and memory, although it can be more difficult to implement and maintain.

By leveraging forecasting and automatic scaling capabilities, adaptive resource management substantially optimises infrastructure utilisation [22, 23]. This methodology effectively reduces the operational expenditures associated with cloud services while ensuring sustained high performance and service reliability. This approach is significantly more cost-effective and efficient than traditional static management, which cannot effectively respond to changing load conditions.

For a more detailed comparison of the effectiveness of adaptive and static resource management, it is important to note that a key factor in reducing infrastructure costs is to reduce the time during which resources are operating in an elevated mode. In systems with static management, resources are often kept in reserve for possible peak loads, which leads to constant capacity costs even during quiet periods [24, 25]. In such systems, resources can be in an increased mode (e.g., 80% of capacity) for 70% of the time, which creates significant additional costs. At the same time, in systems with adaptive control, resources are added only when needed, and their use is adjusted depending on actual conditions. Therefore, resources are in overdrive only 20% of the time, as the system automatically optimises resource allocation according to current needs. This adaptability can significantly reduce infrastructure costs as resources are not over-utilised when they are not needed, resulting in greater efficiency and savings.

Through the use of predictive techniques, the system can not only reduce costs during low load phases but also ensure that additional resources are available when needed, which helps maintain high performance and minimise the risk of downtime when resources are not available to handle peak loads. This process also contributes to the stability of cloud platforms, as anticipating changes in workload allows operations to adapt to future changes before they occur, providing greater confidence in the continuity of services.

These results also highlight the great potential of using adaptive methods for a variety of business processes and organisations where high efficiency in the use of cloud resources is critical to reducing operating costs while ensuring the required performance. The use of such technologies is especially relevant for environments with high load variability, such as e-commerce, data

processing, financial services and other industries where load peaks can occur at unpredictable times.

Through the implementation of predictive models and adaptive management, businesses can significantly improve their economic performance while ensuring competitiveness and cost reduction, which is a key factor for modern organisations seeking to make their operations flexible and resilient in an ever-changing environment.

4 Discussion

The results confirm that adaptive management is much more effective than static approaches, especially when the load on cloud infrastructures is dynamic. For instance, the study by N. Du et al. [26] explored the use of convex hull triangle mesh-based static mapping in highly dynamic environments, providing a novel technique for improving mapping accuracy in such environments. This demonstrated that traditional approaches to resource management under variable load conditions have limited effectiveness. The results of the study confirm this statement, demonstrating that predictive models based on neural networks not only reduce infrastructure costs but also provide high flexibility and adaptability to cloud systems.

Similar conclusions were made by A. Braafladt et al. [27] and S. Khan and A. Jillani [28]. A. Braafladt et al. presented an unusual approach to improving defence modelling and simulation by examining the use of AI-driven adaptive analysis to detect emergent behaviours in military capabilities design. S. Khan and A. Jillani employed search-based software engineering techniques to investigate cloud resource allocation and optimisation, showing how sophisticated algorithms can be applied to increase the effectiveness of cloud computing. This emphasised the need to implement adaptive algorithms to ensure the scalability and flexibility of cloud platforms. This correlates with this approach, which has shown the effectiveness of using deep learning methods for real-time load forecasting.

B. Predić et al. [29] and I. Petrovska and H. Kuchuk [30] both aimed to improve cloud resource management but took different approaches. In order to improve cloud load predictions and resource allocation under varying demands, Predić et al. employed a machine learning approach. In order to maximise efficiency and guarantee secure operations, Petrovska & Kuchuk concentrated on adaptive resource allocation for data processing and security. Both strategies emphasised dynamic resource management in comparison to the current study, with Petrovska & Kuchuk concentrating on security and Predić et al. on prediction accuracy. These concepts are supported by the current study, which shows that adaptive management improves cost-effectiveness and robustness under fluctuating loads.

Studies on cloud forensics, such as the one by R. Al-Mugern et al. [31], analyse the integration of machine learning techniques for data standardisation. This work presents an improved machine learning method that applies a cloud forensic meta-model to enhance the data collection process in cloud environments. By combining machine learning with data-gathering methods to increase the precision and effectiveness of investigations, it makes a substantial contribution to the field of cloud forensics. Although in a different context, this confirms the importance of predictive accuracy and standardisation, which is also key to adaptive resource management.

P. Nawrocki et al. [32] addressed short-term and long-term resource reservations, emphasising the need to respond quickly to sudden peak loads such as flash crowd workload effects. The study looked at machine learning-based adaptive resource planning for cloud-based applications, with an emphasis on how machine learning models can improve resource planning in cloud environments. The results complement this approach by showing that adaptive systems can effectively respond to unpredictable loads while minimising costs. Other studies, such as one by S. Ivan et al. [33], have studied the efficiency of different cloud platforms, including AWS and Microsoft Azure. The study offered insights into cloud-based data processing for big data applications by highlighting the advantages and disadvantages of each platform for doing sentiment analysis at scale. Although the study compared platforms, the results support the conclusion of this paper that adaptive models significantly improve efficiency regardless of the specific platform.

Microsoft's Azure cloud computing is a fully managed computing service that was introduced at a conference in 2008 and became known as Windows Azure and later renamed Microsoft Azure. P. Narayanan [34] discussed the key components and services of Azure, with a special focus on data engineering and machine learning, as well as its impact on various industries due to the availability of data centres around the world. P. Borra [35] discussed the key networking solutions provided by Microsoft Azure, which are the basis for supporting digital operations in modern business. The author examines in detail Azure components such as Virtual Network, Load Balancer, VPN Gateway, ExpressRoute, and Firewall, with a focus on their practical application to ensure uninterrupted connectivity and improve security. The study aims to provide organisations with in-depth knowledge and insights to help them effectively leverage Azure networking services to meet changing business needs, which can complement the findings of this study.

A study by O. Rolik and S. Zhevakin [36] confirmed the results in terms of cost-effectiveness. The use of adaptive management can reduce the cost of cloud services by up to 20%, which highlights the importance of the results for reducing the financial costs of organisations. P. Lakhera [37] complements these findings by suggesting strategies for cost optimisation using artificial intelligence. Anomaly detection and predictive scaling, which the authors investigated, are key elements for improving cost efficiency.

Traditional methods of resource management, as noted by S. Tendulkar [38], are less effective due to the lack of consideration of dynamic changes in the load. This study confirms this by demonstrating that predictive models can more accurately determine resource requirements and

ensure efficient use of resources under variable load conditions. S. Jaber [39] also supports the claim that adaptive systems significantly reduce infrastructure costs. The use of predictive models can reduce costs and improve the performance of cloud systems.

AWS, as shown by L. Devane [40], provides a high ability to adapt to peak loads, which is consistent with the results obtained. Similar conclusions were made by S. Gong et al. [41], who noted that adaptive systems effectively respond to sudden changes in load, ensuring the stability of platforms. The study complements these findings by emphasising the importance of reducing response times to peak loads. This is important for organisations that work with large amounts of data and need consistent access to resources in real-time.

In general, the research findings are fully consistent with current industry trends, in particular the importance of using adaptive systems to manage cloud resources and confirm the effectiveness of load forecasting methods to reduce costs and improve performance. At the same time, it is worth analysing the further development and improvement of such models based on deep learning and integration with new technologies such as edge computing, which will allow for even greater efficiency in real-time management.

5 Conclusions

The study developed models for load forecasting and resource management, including a neural network model for load forecasting and an adaptive resource management model that automatically adjusts resource use based on forecasts. One of the main achievements was the confirmation of the effectiveness of using intelligent algorithms, in particular neural networks, for load forecasting and automatic adaptation of resource allocation in real-time. This reduced the cost of cloud services by an average of 20% compared to traditional static approaches, which confirms the cost-effectiveness of the proposed methods.

The study also determined that AWS demonstrated better adaptability under highly dynamic workloads due to faster resource scaling and more efficient load balancing. While Microsoft Azure showed a more even distribution of resources at a stable load, which is an advantage in the case of a constant load level. The results of the study showed that adaptive resource management in cloud platforms can achieve significant performance improvements and cost savings. AWS demonstrated a 15% improvement in scalability and performance under highly dynamic workloads, while Microsoft Azure showed a 10% increase in resource allocation efficiency under stable workloads. The use of predictive models based on neural networks ensures accurate forecasting of load changes and automatic adaptation of resources in real-time. Adaptive algorithms have proven to be more efficient than traditional approaches, especially in the face of variable workloads. Further developments in technologies such as deep learning and integration with edge computing offer prospects for further improving the flexibility and performance of cloud platforms.

Neural network-based models have proven to provide highly accurate predictions of load changes, enabling efficient real-time resource adaptation. This significantly improves both the performance of cloud platforms and the stability of systems. The results of the study demonstrate the benefits of using intelligent algorithms that can adapt to changing operating conditions.

The results obtained are important for practical application. They open opportunities to significantly reduce business operating costs while ensuring high availability and stability of services. The use of machine learning-based adaptive control technologies allows for optimising resource utilisation and minimising downtime and congestion.

However, the study has several limitations: only two cloud platforms were used, which may limit the generalisability of the results, and the number of types of server configurations for testing is limited. For further research, it is advisable to expand the number of cloud platforms tested, explore the integration of adaptive management with new technologies, such as edge computing, which will significantly improve the efficiency of real-time resource management, and improve predictive models using more sophisticated machine learning algorithms to improve the accuracy of predictions and system adaptability.

References

[1] Berestovenko, O. Virtualisation and network management: Best practices for improving efficiency. Technologies and Engineering, 2024, 25(6): 41-52. https://doi.org/10.30857/2786-5371.2024.6.4

[2] Abdullah, M., & Mohamed Surputheen, M. Optimizing performance of cloud infrastructure through effective resource scheduling. Journal of Advanced Applied Scientific Research, 2024, 6(1): 1-14. https://doi.org/10.46947/joaasr612024748

[3] Alrammah, H., Gu, Y., Yun, D., & Zhang, N. Tri-objective optimization for large-scale workflow scheduling and execution in clouds. Journal of Network and Systems Management, 2024, 32(4): 89. https://doi.org/10.1007/s10922-024-09863-3

[4] Tiwari, A.K., & Yadav, S. Algorithmic model for cloud performance optimization using connection pooling technique. Journal of Statistics and Management Systems, 2024, 27(2): 489-499. https://doi.org/10.47974/jsms-1290

[5] Anayat, R. Cloud-based reinforcement learning in resource-constrained environments: Real-time performance optimization in autonomous systems, 2024. https://doi.org/10.13140/RG.2.2.24832.24326

[6] Varanitskyi, D., Rozkolodko, O., Liuta, M., Zakharova, M., & Hotunov, V. Analysis of data protection mechanisms in cloud environments.

Technologies and Engineering, 2024, 25(1): 9-16. https://doi.org/10.30857/2786-5371.2024.1.1

[7] Demchyna, M., Styslo, T., & Vashchyshak, S. Optimisation of intelligent system algorithms for poorly structured data analysis. Bulletin of Cherkasy State Technological University, 2024, 29(4): 21-31. https://doi.org/10.62660/bcstu/4.2024.21

[8] Johnson, O.B., Olamijuwon, J., Cadet, E., Osundare, O.S., & Samira, Z. Designing multi-cloud architecture models for enterprise scalability and cost reduction. Open Access Research Journal of Engineering and Technology, 2024, 7(2): 101-113. https://doi.org/10.53022/oarjet.2024.7.2.0061

[9] Talha, A., Bouayad, A., & Malki, M.O. An improved pathfinder algorithm using opposition-based learning for tasks scheduling in cloud environment. Journal of Computational Science, 2022, 64: 101873. https://doi.org/10.1016/j.jocs.2022.101873

[10] Slivka, S. Microservices architecture for ERP systems. Bulletin of Cherkasy State Technological University, 2024, 29(4): 32-42. https://bulletin-chstu.com.ua/en/journals/tom-29-4-2024/arkhitektura-mikroservisiv-dlya-erp-sistem

[11] Destek, M.A., Hossain, M.R., Manga, M., & Destek, G. Can digital government reduce the resource dependency? Evidence from method of moments quantile technique. Resources Policy, 2024, 99: 105426. https://doi.org/10.1016/j.resourpol.2024.105426

[12] Smailov, N., Tsyporenko, V., Sabibolda, A., Tsyporenko, V., Abdykadyrov, A., Kabdoldina, A., Dosbayev, Z., Ualiyev, Z., & Kadyrova, R. Streamlining digital correlation-interferometric direction finding with spatial analytical signal. Informatyka Automatyka Pomiary W Gospodarce I Ochronie Srodowiska, 2024, 14(3): 43-48. https://doi.org/10.35784/iapgos.6177

[13] Makhazhanova, U., Omurtayeva, A., Kerimkhulle, S., Tokhmetov, A., Adalbek, A., & Taberkhan, R. Assessment of Investment Attractiveness of Small Enterprises in Agriculture Based on Fuzzy Logic. Lecture Notes in Networks and Systems, 2024, 935 LNNS: 411-419.

[14] Azieva, G., Kerimkhulle, S., Turusbekova, U., Alimagambetova, A., & Niyazbekova, S. Analysis of access to the electricity transmission network using information technologies in some countries. E3S Web of Conferences, 2021, 258: 11003. https://doi.org/10.1051/e3sconf/202125811003

[15] Imamguluyev, R., & Umarova, N. Application of Fuzzy Logic Apparatus to Solve the Problem of Spatial Selection in Architectural-Design Projects. Lecture Notes in Networks and Systems, 2022, 307: 842-848. https://doi.org/10.1007/978-3-030-85626-7_98

[16] Smailov, N., Tsyporenko, V., Ualiyev, Z., Issova, A., Dosbayev, Z., Tashtay, Y., Zhekambayeva, M., Alimbekov, T., Kadyrova, R., & Sabibolda, A. Improving accuracy of the spectral-correlation direction finding and delay estimation using machine learning. Eastern European Journal of Enterprise Technologies, 2025, 2(5(134)): 15-24. https://doi.org/10.15587/1729-4061.2025.327021

[17] Porkodi, S., & Raman, A.M. Success of cloud computing adoption over an era in human resource management systems: a comprehensive meta-analytic literature review. Management Review Quarterly, 2025, 75(2): 1041-1075. https://doi.org/10.1007/s11301-023-00401-0

[18] Sandhu, R., Faiz, M., Kaur, H., Srivastava, A., & Narayan, V. Enhancement in performance of cloud computing task scheduling using optimization strategies. Cluster Computing, 2024, 27(5): 6265-6288. https://doi.org/10.1007/s10586-023-04254-w

[19] Soh, J., Copeland, M., Puca, A., & Harris, M. Microsoft Azure, 2020. Berkeley: Apress. https://doi.org/10.1007/978-1-4842-5958-0

[20] Singh, S., Ramkumar, K.R., & Kukkar, A. Analysis and implementation of microsoft Azure machine learning studio services with respect to machine learning algorithms. In R. Agrawal, C.K. Singh, A. Goyal, & D.K. Singh (Eds.), Modern Electronics Devices and Communication Systems, 2023, (pp. 91-106). Singapore: Springer. https://doi.org/10.1007/978-981-19-6383-4_7

[21] Kavaldzhieva, K. The Impact of Digitalization on the Measurement of value in the production and operation of industrial products. In 2019 International Conference on High Technology for Sustainable Development, HiTech, 2019, (Article number: 9128260). Sofia: Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/HiTech48507.2019.9128260

[22] Kiurchev, S., Abdullo, M.A., Vlasenko, T., Prasol, S., & Verkholantseva, V. Automated Control of the Gear Profile for the Gerotor Hydraulic Machine. In F. Chaari, F. Gherardini, V. Ivanov, & M. Haddar (Eds.), Lecture Notes in Mechanical Engineering, 2023, (pp. 32-43). Cham: Springer. https://doi.org/10.1007/978-3-031-16651-8_4

[23] Bezshyyko, O., Dolinskii, A., Bezshyyko, K., Kadenko, I., Yermolenko, R., & Ziemann, V. PETAG01: A program for the direct simulation of a pellet target. Computer Physics Communications, 2008, 178(2): 144-155. https://doi.org/10.1016/j.cpc.2007.07.013

[24] Orazbayev, B., Zhumadillayeva, A., Kabibullin, M., Crabbe, M.J.C., Orazbayeva, K., & Yue, X. A Systematic Approach to the Model Development of Reactors and Reforming Furnaces With Fuzziness and Optimization of Operating Modes. IEEE Access, 2023, 11: 74980-74996. https://doi.org/10.1109/ACCESS.2023.3294701

[25] Sasi, S., Subbu, S.B.V., Manoharan, P., & Abualigah, L. Design and implementation of secured file delivery protocol using enhanced

elliptic curve cryptography for class I and class II transactions. Journal of Autonomous Intelligence, 2023, 6(3). https://doi.org/10.32629/jai.v6i3.740

[26] Du, N., Xie, L., Zhou, M., Gao, W., Wang, Y., & Hu, J. Convex hull triangle mesh-based static mapping in highly dynamic environments. IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-14. https://doi.org/10.1109/tim.2023.3348881

[27] Braafladt, A., Sudol, A., & Mavris, D. AI-driven adaptive analysis for finding emergent behavior in military capability design. Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 2024. https://doi.org/10.1177/15485129241289137

[28] Khan, S.M. & Jillani, A. Cloud resource allocation and optimization using search-based software engineering methods, 2024. https://doi.org/10.13140/RG.2.2.17568.19207

[29] Predić, B., Jovanovic, L., Simic, V., Bacanin, N., Zivkovic, M., Spalevic, P., Budimirovic, N., & Dobrojevic, M. Cloud-load forecasting via decomposition-aided attention recurrent neural network tuned by modified particle swarm optimization. Complex & Intelligent Systems, 2023, 10(2): 2249-2269. https://doi.org/10.1007/s40747-023-01265-3

[30] Petrovska, I. & Kuchuk, H. Adaptive resource allocation method for data processing and security in cloud environment. Advanced Information Systems, 2023, 7(3): 67-73. https://doi.org/10.20998/2522-9052.2023.3.10

[31] Al-Mugern, R., Othman, S.H., & Al-Dhaqm, A. An improved machine learning method by applying cloud forensic meta-model to enhance the data collection process in cloud environments. Engineering, Technology & Applied Science Research, 2024, 14(1): 13017-13025. https://doi.org/10.48084/etasr.6609

[32] Nawrocki, P., Grzywacz, M., & Sniezynski, B. Adaptive resource planning for cloud-based services using machine learning. Journal of Parallel and Distributed Computing, 2021, 152: 88-97. https://doi.org/10.1016/j.jpdc.2021.02.018

[33] Ivan, S.C., Győrödi, R.Ș., & Győrödi, C.A. Sentiment analysis using Amazon web services and Microsoft Azure. Big Data and Cognitive Computing, 2024, 8(12): 166. https://doi.org/10.3390/bdcc8120166

[34] Narayanan, P.K. Engineering data pipelines using Microsoft Azure. In P.K. Narayanan (Ed.), Data Engineering for Machine Learning Pipelines, 2024, (pp. 571-616). Berkeley: Apress. https://doi.org/10.1007/979-8-8688-0602-5_17

[35] Borra, P. Microsoft Azure networking: Empowering cloud connectivity and security. International Journal of Advanced Research in Science, Communication and Technology, 2024, 4(3): 469-475. https://doi.org/10.48175/ijarsct-18949

[36] Rolik, O.I. & Zhevakin, S.D. Cost optimization method for informational infrastructure deployment in static multi-cloud environment. Radio Electronics, Computer Science, Control, 2024, 3: 160-172. https://doi.org/10.15588/1607-3274-2024-3-14

[37] Lakhera, P. Leveraging large language models to optimize costs in Amazon web service cloud. TechRxiv, 2024. https://doi.org/10.36227/techrxiv.172684142.23966027/v1

[38] Tendulkar, S. Optimizing generative AI model performance through cloud resource management in hybrid AI systems, 2024. https://doi.org/10.13140/RG.2.2.34745.38246

[39] Jaber, S. Enhanced model performance in generative AI: Cloud resource optimization for real-time adaptive autonomous systems, 2024. https://doi.org/10.13140/RG.2.2.32857.94567

[40] Devane, L. Adaptive AI systems in autonomous environments: Real-time decision making and resource allocation through cloud-based reinforcement learning, 2023. https://doi.org/10.13140/RG.2.2.21638.18241

[41] Gong, S., Yin, B., Zheng, Z., & Cai, K.-Y. Adaptive multivariable control for multiple resource allocation of service-based systems in cloud computing. IEEE Access, 2019, 7: 13817-13831. https://doi.org/10.1109/access.2019.2894188