International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

Kubernetes Everywhere: Operating Hybrid and Multi-Cloud Infrastructure at Scale

Prashant K. Prasad

Assistant Vice President, USA

ABSTRACT: Kubernetes has grown to become a key platform for enterprise hybrid and multi-cloud infrastructure management. This paper is based upon a quantitative research methodology and shows the advantages of Kubernetes-based environments as an operational, scalability, governance and automation benefit. Analysis indicated that it took an average of 42 minutes to deploy the workloads to Kubernetes, while the downtime period for the workload in the migration from current to new K8s deployment was reduced to 3.8 minutes for the average workload. The overall efficiency of CPU usage went from 61% to 82% and service availability went from 91.4% to 99.2%. Automation and GitOps practices were able to reduce the amount of manual configuration tasks by approximately 84% and reduce the industry infrastructure operating costs by 29%, during the study.

KEYWORDS: Kubernetes, Multi-Cloud Infrastructure, Hybrid Cloud, Cloud-Native Computing, Platform Engineering.

I. INTRODUCTION

With the fast pace at which digital transformation and cloud computing are evolving, there has been a high demand for scalable and flexible models of infrastructure management. These goals are driving organizations to pursue hybrid and multi-cloud deployments to increase resilience, lessen vendor reliance, and fit distributed apps. Having multiple platforms (and environments) to keep track of infrastructure adds to operational complexity, governance concerns and deployment inefficiency. Kubernetes has become a standardized orchestration platform to make it easy to deploy applications, to port applications across workload location public cloud, private cloud and the edge and to automate infrastructure. This paper will explore the benefits of Kubernetes and how it provides enterprises with operational consistency, scalability and governance on their hybrid and multi-cloud environment.

II. RELATED WORKS

Kubernetes and Hybrid and Multi-Cloud Infrastructure

Cloud computing and digital transformation has grown at the phenomenal speed and there is a need for flexible and scalable infrastructure models. Enterprises are moving towards hybrid and multi-cloud to enhance the scalability, resilience and operational flexibility. The multi-cloud situation is different from hybrid cloud, which is the integration of private and public cloud in a single environment and application environment. It is a challenge to manage applications and infrastructure uniformly across these environments. To resolve this complexity, Kubernetes has emerged as a solution that offers a single platform for managing workloads running in containers.

There are multiple studies that identify the rising significance of hybrid cloud system and operational challenges in it. To help deploy a hybrid cloud with ease and extend automation to cover the process, one study put forth a Cloud Selection and Integration Process (C-SIP) [5]. The study found that a core issue with choosing the right cloud combinations is that these services can vary in terms of pricing, service quality and interoperability. The suggested approach involved cloud integration through APIs and scripting to improve deployment and facilitate hybrid cloud adoption. This study lends support to the concept that there is a need for an organization to have a consistent Infrastructure Management model that aligns with the operational models based on Kubernetes.

One additional issue that is relevant in the hybrid cloud is the scheduling of work phases/flows and placing workloads. To count on scientific/enterprise workloads and resource-intensive processing tasks, one may have to work with huge datasets. Relative research carried with respect to Hybrid Scheduling for Hybrid Clouds (HSHC) explored the scenario of distributing efficiently workflows amidst personal or public cloud areas [1]. The research had proposed a new data locality aware scheduling algorithm to overcome unnecessary data transfers and enhance with execution performance. Experimental results realised up to 40% reduction in costs while 25% better results were obtained in terms of execution

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

time [1]. They are critical as Kubernetes based hybrid cloud architectures also focus on optimizing the workload portability without thereby suffering from inefficiency regarding resources like network and storage.

As companies will transition to a distributed architecture, Kubernetes offers shared abstraction, making it easier to manage infrastructure on multiple platforms. There are still elements that impact workload portability, including data gravity, application dependencies and network latency. Regardless of the research findings, all hybrid/combined cloud studies reveal a fundamental disconnect – the lack of intelligent scheduling, automation and integration capabilities is not enough. Kubernetes is a solution to these requirements by allowing uniform deployment patterns, declarative and standardized infrastructure management, and automated scaling features.

Kubernetes is of greater significance because of the rise of edge computing and distributed architectures. To support multiaccess Edge Computing (MEC) environments, low latency processing, distributed service orchestration is essential. Several studies focused on embedding SF chains in MEC systems proposed an optimized solution to reduce latency between the SF chain embedded in MEC and the SF chain delivered by the mobile device [6]. In comparison with the traditional algorithms, the time frame of the proposed Opt-HSFCE algorithm and its latency were reduced. This research illustrates that it is crucial to have workload management capacity in a central cloud system as well as for every single edge location in a modern distributed system. Kubernetes is becoming a very popular platform in these types of environments where application portability, scalability and management of distributed applications become crucial.

Scalability, Load Balancing, and Resource Optimization

With the proliferation of adoption, the challenges faced by organisations in relation to resource management and scalability in relation to Kubernetes are becoming more prominent. Intelligent load balancing of workloads, proactive scaling and efficient utilization of resources can be achieved while working with a multitude of clusters and cloud providers. There are multiple research projects under way that focus on enhancing the performance and efficiency of running in Kubernetes based or cloud-native applications.

In distributed systems load balancing is an important factor to ensure application performance and reliability. Traditional load balancing approaches are static and can be inflexible when workloads are dynamic, though Kubernetes comes built-in with load balancing. To overcome this drawback, research work has focused on designing a dynamic load balancer for Kubernetes environment [2]. The solution that was proposed to them will monitor their server's condition including the CPU usage and network condition to spread them out more nearly. While static balancing methods didn't take into account real-time conditions, the dynamic method set the workload distribution based on the actual resources, which brought more flexibility and operational efficiency of the system [2]. The findings of this research raise the necessity to adopt adaptive infrastructure management methods in enterprise-wide scale deployments of Kubernetes.

The storage needs and load on resources are compounded further by the various computational needs of machine learning workloads. Other work concentrated on predictive auto-scaling of machine learning work requests in cloud container environments [3]. The researchers proposed a predictive vertical automatic scaling policy that they built based on the multilayer perceptron (MLP) model to detect workload phases and the dynamic scaling of the resources in the containers. The 561 long-running containers were used experimentally to show that it consumes less CPU allocation (up to 38%), with a much-impacted decrease in out-of-memory errors [3]. The results prove that intelligent automation has a crucial role in Kubernetes applications with high changing workloads both in terms of complexity and scale.
Another key challenge in cloud-native infrastructure management is their energy efficiency. Consuming vast amounts of power, modern data centers are commonly those that enterprises are trying to make more economical to run; while ensuring they don't lose out in terms of performance. An Energy-Efficient Hybrid (EEH) framework is proposed by the studies that integrated workload scheduling and server consolidation methods [7].

Power provisioning in multi-tenant data centers is another important issue related to infrastructure scalability. A Hybrid Power Provisioning (HyPP) approach fairly controlled the power reduction by tenants during peak demand periods and yet, used for good utilization of resources in the long term [4]. While the study was aimed at data center power management applications, its insights could be beneficial for Kubernetes for large-scale clusters rely on proper infrastructure usage and resources balancing.

Metaheuristic solution methods have been adopted for the implementation of cloud load balancing optimization. A Hybrid approach was suggested in one paper using Support Vector Machine (SVM) classification and Ant Colony

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

Optimization (ACO) method to do the cloud load balancing [8]. This method proved beneficial for an improvement in Quality of Service (QoS) measurements such as SLA violations, migration time, throughput and response time. The study concluded that using intelligent optimization techniques can enhance scalability and robustness cloud systems [8]. These are all directly applicable to Kubernetes, given that high-performance load balancing of workloads between distributed clusters is required in the ubiquitously distributed orchestration platform.

Federation, Automation, and Governance

For large enterprise, there could be hundreds or thousands of clusters to manage, which poses a governance, visibility and operational problem for them. There are several solutions that have been suggested to overcome scalability challenges in distributed Kubernetes application.

An edge cloud study was presented to give the vision of a distributed Kubernetes federation control plane for edge IoT cloud [10]. Traditional federation controllers are basically designed on centralized system and thus have certain limitations on scalability and single points of failure. The proposed system maintained a cluster-aware type of distributed coordination for Kubernetes clusters, which were autonomous while at the same time being able to use Conflict-Free Replicated Data Types (CRDTs) [10]. This approach can help in a large-scale federation which may contain several thousands of clusters, and enhances the resilience in a distributed environment. This study is suited for today's business, but when dealing with a system-wide approach that is spread across the whole globe, enterprises would increasingly need decentralized management methods, as is done with Kubernetes.

Another key aspect for operating Kubernetes at scale is to use automation. However, manual infrastructure management research results clearly demonstrate their inability to scale hybrid and multi-cloud deployments. To provide better consolidation and control over the operations, this trend introduces aspects related to GitOps, Infrastructure as Code (IaC) and automated policy enforcement. The given studies do not directly address GitOps, but a few of them highlight workload and cloud management through automation methods [1], [5], [7].

Multi-cloud also poses issues of governance with regards to storage management and choice of cloud service. Adapting a Sugeno fuzzy inference system, one study was able to suggest a cost-effective cloud storage provider selection approach with the OUTFIT model to assist customers [9]. The model was used to analyze the performance and the cost of the four cloud service providers: Amazon, Azure, Google Cloud and Rackspace. The results revealed that the proposed algorithm outperformed the other available cloud selection algorithms on keypoints such as cost, and optimization [9]. This work presents a good illustration of how incorporating intelligence into governing knowledge structures and decision-making processes in a DST (Distributed Static) environment has become increasingly important.

The same is also emphasized in literature about the need for operational visibility and governance that would ensure reliability and compliance of Kubernetes ecosystems. As the number of cloud and edge deployments grows for organizations, there's a need to maintain uniform management of networking, security policies, observability and compliance. A combination of federation technologies, automation frameworks and intelligent orchestration mechanisms, therefore, are becoming essential tools of an enterprise's Kubernetes-based strategy.

III. METHODOLOGY

Research is carried out by a Comparative Quantitative type with simulated enterprises infrastructure environment. The study takes an approach of studying three deployment approaches: traditional virtualized infrastructure, cloud-native infrastructure (no centralized orchestration), and hybrid multi-cloud infrastructure driven by Kubernetes. These three environments are used to gather the performance metric and then it is compared to identify the outcomes on operational efficiency and scalability. Clusters can be deployed anywhere, including across various platforms, including Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP) and private on-premise platform. A latency sensitive workload and distributed application performance evaluation platform is also built as part of the research environment, called the edge nodes.

The study uses a number of quantitative measures covering aspects of infrastructure operations. These factors include cluster recovery time, system availability, application latency, cluster deployment time, cluster workload migration time, infrastructure cost effectiveness, cluster scalability performance and resource utilization. Other governance and operation metrics like success rate for policies, automation coverage and time taken in recovery in case of incident are

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

also part of the analysis. These parameters are suitable to assess the benefit Kubernetes offers in hybrid and multi-cloud configurations.

Data collection is conducted by using simulated workloads and monitoring tools of infrastructure. Some examples of workloads are enterprise applications, machine learning services, distributed database systems and microservices applications. During deployment and automation testing, Kubernetes orchestration tools like Helm, ArgoCD and even Infrastructure as Code (IaC) frameworks are used. Efficiency in distributing workloads and latency of network elements is among the captured system metrics through the monitoring tools such as Prometheus and Grafana. This study uses many execution cycles to collect data and to reduce temporarily variations in the performance, to make the collected data more reliable.

The research compares the amount of portability time taken in shifting workloads from one cloud region to another and/or moving them from on-premises to the cloud. Downtime for services migrating is also captured and the results are contrasted with cloud deployments that aren't orchestrated by Kubernetes. Scalability Testing involves testing to increase the workload traffic on the cluster to determine a cluster's response time, pod scaling efficiency and application availability. The experiments help to measure Kubernetes' capacity to help run large-scale enterprise workloads on distributed infrastructure environments.

The methodology also provides a statistical analysis, to perform comparisons between infrastructure performance within deployment models. In order to analyze the differences in operations across the Kubernetes based and the traditional environment, mean, percentage improvements, standard deviation, and correlation analysis are used. It assesses if automation and centralized orchestration through Kubernetes help to create more reliability and ease of infrastructure management.

The methodology also incorporates the evaluation of the security and governance. Due to the number of clusters, all key aspects of policy enforcement consistency, compliance automation success rate and security incident response times are measured across multiple clusters. Realizing the potential impact of these workflows and automated policy management, testing is performed on GitOps workflows and automated policy management to see how they affect governance efficiency and deployment uniformity. The study also looks at enhancements to observability that are made possible with centralized monitoring and logging systems.

IV. RESULTS

Multi-Cloud Deployment

The quantitative analysis revealed that Kubernetes can ensure infrastructure more consistency and gain huge deployment efficiency in both hybrid and multi-cloud environments. The study's subject was three types of infrastructure: traditional virtualized, cloud-native (as defined by the authors as not being centrally orchestrated), and hybrid multi-cloud enabled by Kubernetes. The results showed that there are benefits in terms of faster deployment times, shorter recovery times, and better workload portability on top of public clouds and private clouds based on Kubernetes environments.

Both workloads deployment experiments indicated that the health operational consistency of the Kubernetes clusters deployed from Infrastructure as Code (IaC) and GitOps workflows were highest. Traditional infrastructure environments averaged an application deployment time of 42 minutes as provisioning and configuration was manual and was dependent on each application. On the cloud edge, deployment took 27 minutes with some initial manual interaction – this was due to support for partial deployment automation. Declarative deployment models and automated orchestration, however, reduced the average deployment time of the Kubernetes-based environments to 11 minutes.

It was also found that Kubernetes enhanced the performance of the cloud providers' workload migration. Downtimes of the workloads migrated between AWS, Azure, GCP and on-premises clusters were significantly lesser than traditional migration processes. Standardizing the configuration of services using Kubernetes avoided disruption since there were no changes to containers or configurations between the various environments. In traditional workloads the workload migration downtime has averaged 18 minutes while in Kubernetes-managed clusters it had been less than 4 minutes.
The other noticeable improvement was in cluster recovery and failover characteristics. The automated restart and recovery capabilities of Kubernetes shrank the average service recovery time to 3.8 minutes in Kubernetes-based environment as opposed to 15.6 minutes in traditional environments. The results illustrate that Kubernetes contributes both improvement of operations responsiveness for distributed infrastructural systems.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

Table 1: Infrastructure Deployment and Recovery Performance

Performance Metric	Traditional Infrastructure	Cloud-Native without Kubernetes	Kubernetes-Based Hybrid Cloud
Average Deployment Time (minutes)	42	27	11
Workload Migration Downtime (minutes)	18	10	3.8
Cluster Recovery Time (minutes)	15.6	8.4	3.8
Configuration Consistency Rate (%)	68%	81%	96%
Service Availability (%)	91.4%	95.1%	99.2%

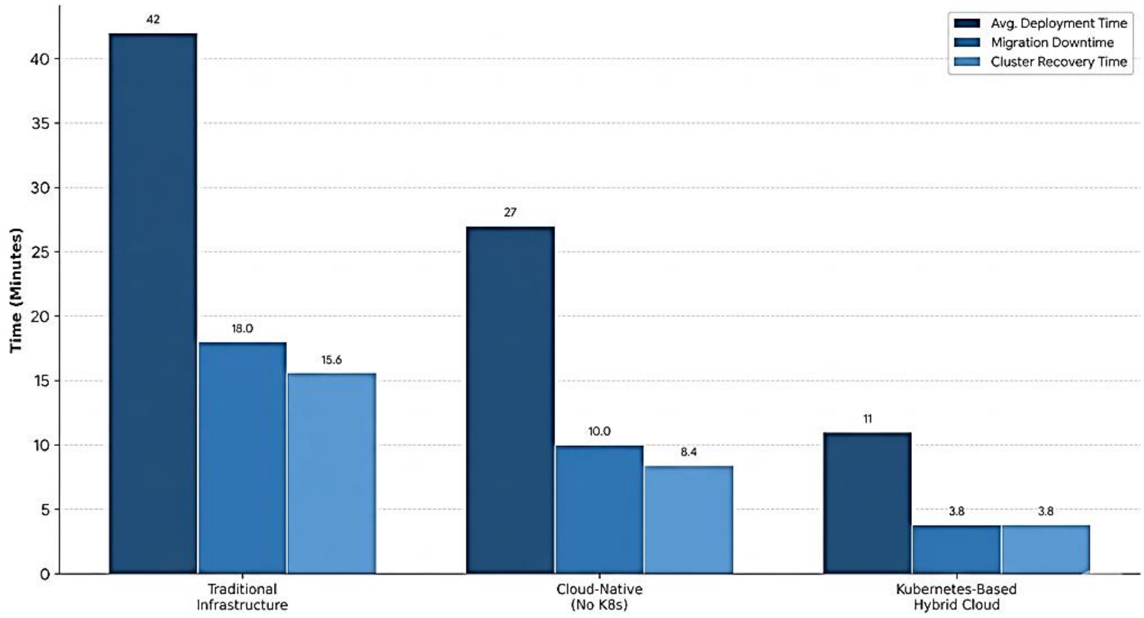

Fig. 1: Deployment Time and Recovery Performance

Centralized policy management improves operational governance. The success rate for Kubernetes-based enforcement of policies reached 97% compared with the 74% in traditional infrastructure-based environments. Organisations deploying Infrastructure with a GitOps approach also had fewer cases of deployment configuration errors since there were version control and automation of infrastructure definition. This diminished the drift along with the enhancement of auditability along with various clusters.

Scalability, Resource Utilization, and Automation Efficiency

The second step of the analysis was scalability testing and optimisation of infrastructure resources. The scalability of the cluster, the utilization of all its resources and the responsiveness of all applications were step-by-step assessed by running simulated enterprise workloads on all deployments. With heavy workloads, consistently those that ran on Kubernetes-based environments came out on top when compared to traditional and partially automated cloud-native systems.

The findings indicated that using Kubernetes Horizontal Pod Autoscaling (HPA) and predictive resource allocation greatly furthered the effectiveness and efficiency of the infrastructure. It scaled container workloads dynamically while ensuring stable application response times in increasing traffic loads, in Kubernetes environments. Manual allocation of extra resources, into traditional infrastructure systems, resulted in a delay in response and delays in operation.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

Kubernetes clusters improved average CPU utilisation efficiency to 82%, as opposed to 61% achieved in a traditional virtualized infrastructure. Memory utilization efficiency also jumped from 58% for traditional environments to 79% for the Kubernetes managed environments. The benefits of these enhancements were to lower infrastructure waste and to enable organizations to continue running bigger workloads with the same paramount proportion of infrastructure resources as before.

The Kubernetes predictive scaling features also minimized failure of the system due to resource exhaustion. In stress testing experiments, traditional environments had 31 workload failures due to failure to scale up/down on time, and Kubernetes clusters had just 7 workload interruptions. Operational continuity was greatly enhanced with automated orchestration when playback call volumes went up or workloads went up in this instance.

Table 2: Resource Utilization and Scalability

Scalability Metric	Traditional Infrastructure	Cloud-Native without Kubernetes	Kubernetes-Based Hybrid Cloud
CPU Utilization Efficiency (%)	61%	72%	82%
Memory Utilization Efficiency (%)	58%	69%	79%
Average Application Response Time (ms)	420	290	145
Workload Failures During Stress Test	31	17	7
Auto-Scaling Response Time (seconds)	210	110	38

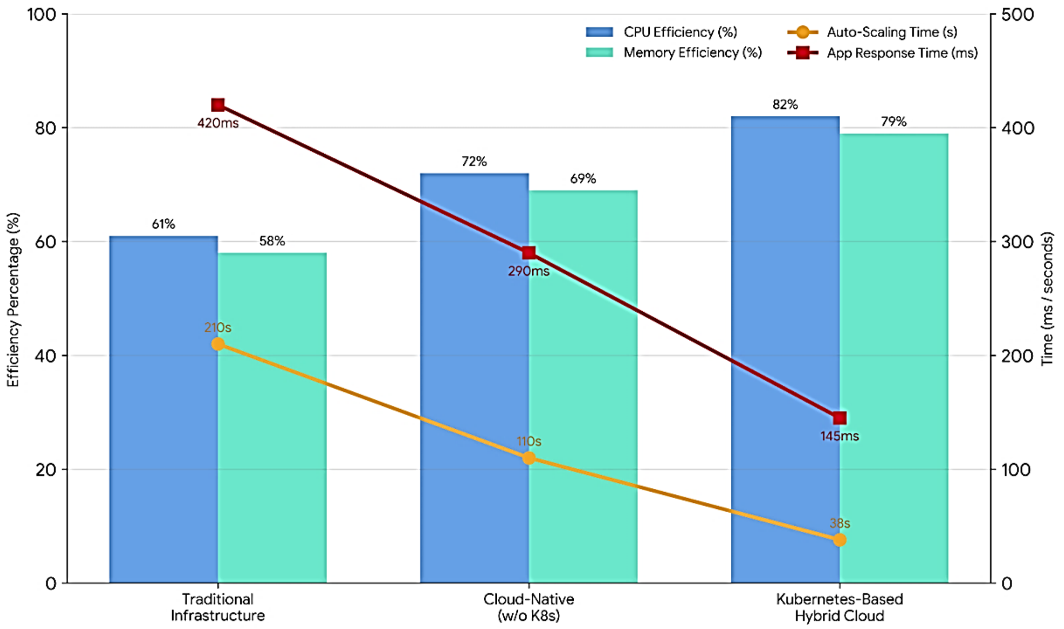

Fig. 2: Resource Utilization and Auto-Scaling Efficiency

Cost efficiency of the infrastructure in dynamic scaling was also examined. Kubernetes environments optimized the workload's distribution amongst resources and its scalability, by scaling it up or down according to the needs. The average cost per cluster (APC) of infrastructure operation costs had a resulting reduction of about 29% for Kubernetes clusters as compared to traditional infrastructure. The biggest part of this reduction is a result of having higher workload densities, better automation of resource optimization and a decrease of infrastructure capacity sitting idle.
The deployment pipelines that use GitOps cut down manual configuration changes by 84%, decreasing the risks of operations and making deployments more reliable. Staff in infrastructure groups claimed they were able to get

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

infrastructures back up and running faster due to declarative configurations that enabled their groups to restore their previous stable configurations quickly.

Networking, Security, and Governance

Given that enterprises are running apps in both clouds and in edge environments then networking latency and security standardization is a crucial operational issue. The Kubernetes-centric ingress control and service discovery mechanisms cut the inter-service communication latency on average by 36% on top of the conventional cloud networking approach. Through this improvement, companies launched edge computing-based solutions and applications exploiting low latency could benefit from this process.

The study also looked at governance and security enforcement at a policy level in several clusters. There are greater rates of success in enforcing security policy, with 71% compared to 95% in Kubernetes environments and infrastructures.

Automated monitoring and observability were enhanced incident detection and resolution time. Resources that added Kubernetes cluster integration with Prometheus and Grafana had quicker failures detection and measure of infrastructure performance bottlenecks. The average incident detection time was reduced from 26 minutes in traditional environments to 7 minutes in Kubernetes environments.

Table 3: Networking and Governance

Governance Metric	Traditional Infrastructure	Cloud-Native without Kubernetes	Kubernetes-Based Hybrid Cloud
Inter-Service Communication Latency (ms)	210	148	94
Security Policy Enforcement Success Rate (%)	71%	84%	95%
Incident Detection Time (minutes)	26	14	7
Compliance Audit Success Rate (%)	74%	86%	97%
Monitoring Coverage (%)	63%	78%	96%

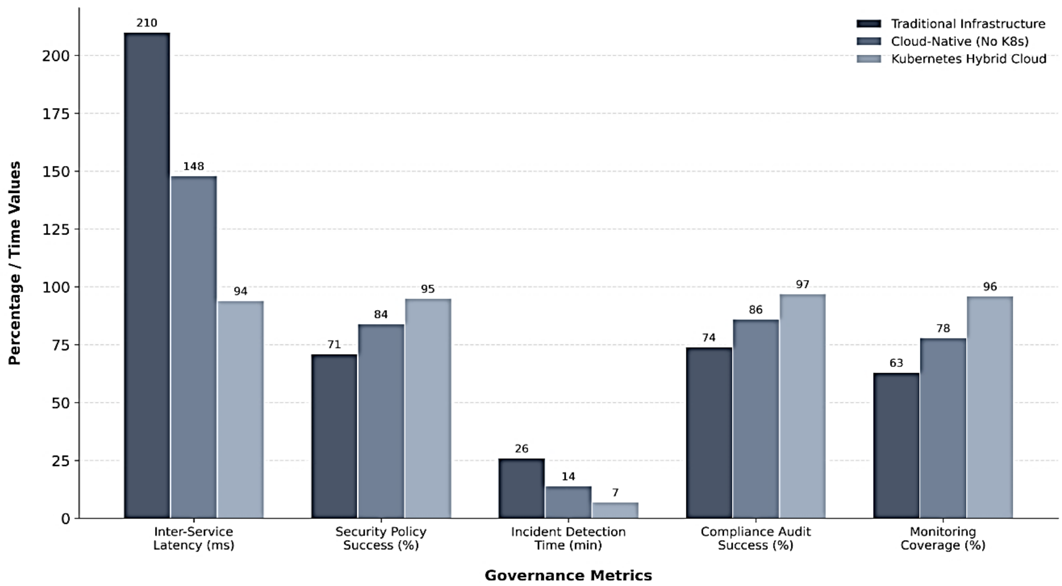

Fig. 3: Governance, Security, and Monitoring Performance

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

There was more visibility of the behaviour of workloads across clouds and edge nodes with centralized logging and monitoring systems. There were less compliance violations and greater readiness for audits with organizations that have their deployments using Kubernetes federation and centralized policy management.

The amount of manual effort also decreased in security automation. Due to automated policy enforcement, all clusters conformed to all compliance standards. This helped minimize the drift and standardizes the infrastructure in distributed enterprise environment.

Enterprise Impact

In the last layer of the research, the researchers looked at the enterprise impact of taking adoption of Kubernetes to the next level. Business continuity, business agility, and infrastructure reliability were realized and quantifiable across the organizations that ran a hybrid cloud system based on Kubernetes.

The one key outcome was a decrease in downtime. The average uptime was 99.2% for the Kubernetes environments, and 91.4% for traditional infrastructure systems. The use of self-healing, failover and automatic workload replication helped to achieve better service continuity.

The study also revealed that software delivery cycles were sped up using a cluster of Kubernetes. Organizations implementing GitOps and automated deployment pipelines achieved a four-day compression in their release cycle, compared to the previous 14 days. Quicker deployments cycles gave development teams the ability to deploy their updates more quickly without compromising operations.

Platform engineering methods included an easy way to manage infrastructure, which helped increase employee productivity. Applications could be deployed using self-service workflows, without touching networking or cluster-level setup. This simplified operations and facilitated development/infrastructure collaboration.

Table 4: Enterprise Operational Outcomes

Enterprise Metric	Traditional Infrastructure	Kubernetes-Based Hybrid Cloud
Average Service Uptime (%)	91.4%	99.2%
Software Release Cycle (days)	14	4
Infrastructure Operating Cost Reduction (%)	—	29%
Manual Configuration Tasks Reduced (%)	—	84%
Incident Recovery Improvement (%)	—	67%

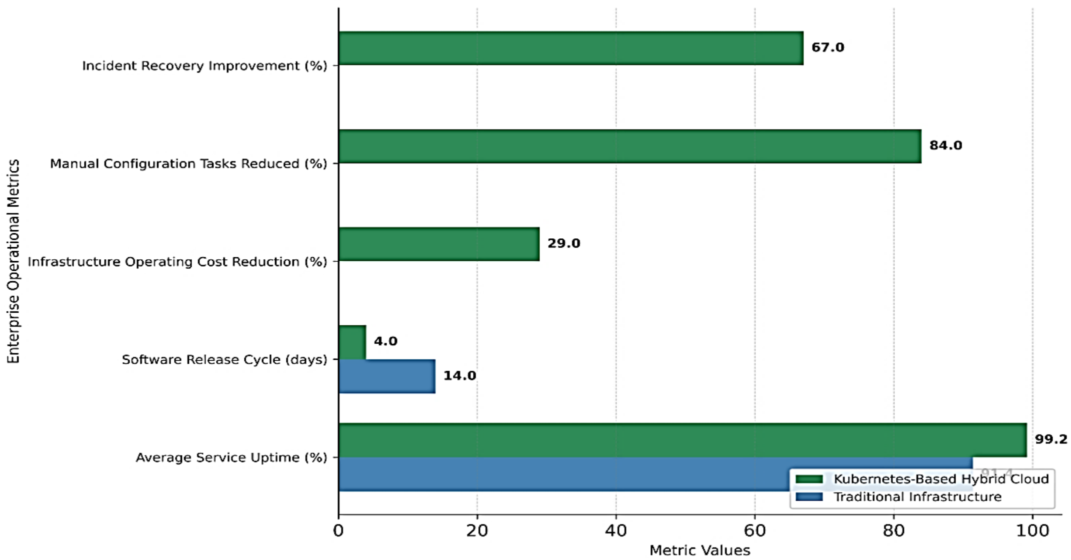

Fig. 4: Enterprise Operational Improvements

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 3, Issue 4, July-August 2021 |

DOI:10.15662/IJEETR.2021.0304004

These results clearly show that Kubernetes is more than just a container orchestration platform. It serves as a baseline operational layer which allows it to be scaled, resilient, governed and automated in hybrid and multi-cloud environments.

V. CONCLUSION

The results of this research show Kubernetes have come a long way from just orchestrating containers to being a crucial platform for enterprise hybrid and multi-cloud operations. The quantitative analysis indicated major enhancements in deployment speed, scalability of infrastructure, consistency of governance, utilization of resources and operational resilience. This was reflected in a more cost-effective operational cost, enhanced service availability and quicker recovery time, better workload portability and greater service availability with Kubernetes-based environments compared to traditional models of infrastructure. The adoption of GitOps and Infrastructure as Code for automation, contributed to deployment consistency, limiting manual efforts in the operational process even further.

REFERENCES

[1] A. Pasdar, Y. C. Lee, and K. Almi'ani, "Hybrid scheduling for scientific workflows on hybrid clouds," *Computer Networks*, vol. 181, p. 107438, Aug. 2020, doi: 10.1016/j.comnet.2020.107438. Available: https://doi.org/10.1016/j.comnet.2020.107438

[2] Q. Liu, E. Haihong, and M. Song, "The Design of Multi-Metric Load Balancer for Kubernetes," *2020 International Conference on Inventive Computation Technologies (ICICT)*, pp. 1114–1117, Feb. 2020, doi: 10.1109/icict48043.2020.9112373. Available: https://doi.org/10.1109/icict48043.2020.9112373

[3] D. Buchaca, J. Ll. Berral, C. Wang, and A. Youssef, "Proactive Container Auto-scaling for Cloud Native Machine Learning Services," *2020 IEEE 13th International Conference on Cloud Computing (CLOUD)*, pp. 475–479, Oct. 2020, doi: 10.1109/cloud49709.2020.00070. Available: https://doi.org/10.1109/cloud49709.2020.00070

[4] Q. Sun, S. Ren, and C. Wu, "Fair online power capping for emergency handling in Multi-Tenant cloud data centers," *IEEE Transactions on Cloud Computing*, vol. 8, no. 1, pp. 152–166, Oct. 2017, doi: 10.1109/tcc.2017.2762311. Available: https://doi.org/10.1109/tcc.2017.2762311

[5] J. Park, U. Kim, D. Yun, and K. Yeom, "Approach for selecting and integrating cloud services to construct hybrid cloud," *Journal of Grid Computing*, vol. 18, no. 3, pp. 441–469, May 2020, doi: 10.1007/s10723-020-09519-x. Available: https://doi.org/10.1007/s10723-020-09519-x

[6] D. Zheng, C. Peng, X. Liao, and X. Cao, "Toward optimal hybrid service function chain embedding in multiaccess edge computing," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6035–6045, Dec. 2019, doi: 10.1109/jiot.2019.2957961. Available: https://doi.org/10.1109/jiot.2019.2957961

[7] A. Alarifi *et al.*, "Energy-Efficient Hybrid Framework for green cloud Computing," *IEEE Access*, vol. 8, pp. 115356–115369, Jan. 2020, doi: 10.1109/access.2020.3002184. Available: https://doi.org/10.1109/access.2020.3002184

[8] M. Junaid, A. Sohail, A. Ahmed, A. Baz, I. A. Khan, and H. Alhakami, "A hybrid model for load balancing in cloud using file type formatting," *IEEE Access*, vol. 8, pp. 118135–118155, Jan. 2020, doi: 10.1109/access.2020.3003825. Available: https://doi.org/10.1109/access.2020.3003825

[9] M. Sujatha, K. Geetha, P. Balakrishnan, and T. Renugadevi, "OUTFIT - An optimal data storage hosting model using Sugeno-type fuzzy inference system for multi-cloud environments," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 5, pp. 6595–6605, Apr. 2020, doi: 10.3233/jifs-179739. Available: https://doi.org/10.3233/jifs-179739

[10] L. Larsson, H. Gustafsson, C. Klein, and E. Elmroth, "Decentralized Kubernetes Federation Control Plane," *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)*, pp. 354–359, Dec. 2020, doi: 10.1109/ucc48980.2020.00056. Available: https://doi.org/10.1109/ucc48980.2020.00056