International Journal of Science, Strategic Management and Technology
Volume **02** Issue **03** March-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Resource Provisioning Strategies in Hybrid Cloud Infrastructure

Ajay. R , Vijay Anand. R

Undergraduate Student, Assistant Professor, Department of Computer Technology, Dr. N.G.P. Arts and Science College, Coimbatore, Tamil Nadu, India

https://doi.org/10.55041/ijsmt.v2i3.075

Cite this Article: R, A. (2026). Resource Provisioning Strategies in Hybrid Cloud Infrastructure. International Journal of Science, Strategic Management and Technology, *02*(03). https://doi.org/10.55041/ijsmt.v2i3.075

Abstract

The Hybrid Cloud Infrastructure has been adopted as a paradigm of modern enterprise computing that unites the scalability of a public cloud service environment with the security and control of a chosen on-premises enterprise resources. Resource provisioning in such environments that are part hybrid is essential in ensuring the best performance, cost-effectiveness, and service-level agreement (SLA) are met. The paper provides a detailed study on the resource provisioning strategies such as the static, dynamic, reactive, and proactive methods of hybrid cloud systems. The predictive provisioning, workload-aware scheduling, and cost-optimization frameworks, which are machine-learned, are assessed in the context of simulation analysis. We show that when used in proactive provisioning, ML significantly reduces resource over-provisioning (by 34 percent), minimizes the average response latency (by 28 percent), and also is much cost-effective (up to 41 percent) as compared to traditional threshold-based strategies. Further challenges that we find to be open include federated resource orchestration, multi-tenant isolation, and green cloud provisioning. The work adds a single taxonomy of provisioning strategies and performance benchmarking framework of hybrid cloud environments

Keywords:Hybrid Cloud, Resource Provisioning, ML-Driven Provisioning,

I. INTRODUCTION

The high adoption of digital services, real-time analytics, and edge-based applications, has become a radical change to the IT infrastructure requirements among enterprises. As organizations expand their workforce, hybrid cloud services, in which the private data center and public clouds like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform are combined, have become the way to manage peak workloads and ensure data sovereignty, regulatory compliance [1].

Hybrid clouds have their unique benefits over a public or private deployment such as provisioning of burst capacity, disaster recovery failover and workload portability. These benefits however present a lot of complexity in the management of resources. The delivery of the appropriate quantity of compute, memory and network resources at the correct time, SLA Optimization, CloudSim Plus, LSTM Forecasting, Static Provisioning,

Reactive Provisioning, Proactive Provisioning with each level of heterogeneous infrastructure is not a trivial optimization issue [2].Resource provisioning in hybrid clouds involves making the decisions associated with the allocation of virtual machines (VMs), container scheduling, automatic scaling policies, and placement plans that help to trade performance goals and cost limits. Under-provisioning causes the violation of the SLA as well as the abacus user experience, whereas over-provisioning generates the superfluous capital and operational spending [3].

The literature has already described the issues of provisioning when considering a purely public or purely private cloud setting; but has not offered detailed models that take into consideration the special characteristics of hybrid environments, such as inter-cloud latency, data transfer costs, workloadburstiness, and multi-tenancy. This gap is the reason why the current investigation is motivated.

The following are the main contributions of this paper:

(1) An in-depth taxonomy of resource provisioning strategies in hybrid cloud infrastructure, categorizingstrategies based on decision horizon, degree of automation and optimization goal.

(2) Comparative performance analysis of static, reactive and ML-driven proactive provisioning in realistic hybrid workload traces with the help of simulation.

II. LITERATURE REVIEW

A. Cloud Resource Management Development.

The initial cloud computing paradigms assumed the use of a static provisioning model whereby the resources were provisioned in fixed amounts depending on the maximum load assumptions. The original NIST definition of cloud computing by Mell and Grance [4] formed the basis of the service models (IaaS, PaaS, SaaS) and deployment models (public, private, hybrid) which form the backbone of the current provisioning research. The introduction of the Amazon EC2 service in 2006 brought with it the concept of on-demand provisioning, where an organization is able to spin up compute capacity on-demand and can only pay based on what it consumes

The history of virtualization technologies, especially VMware vSphere and open sources, including KVM and Xen, made it possible to abstruct and multiplex resources at a fine granularity.

B. Static and Threshold-Based Provisioning.

There are two different types of provisioning based on a static or threshold value. In statistic provisioning, the resources are assigned according to the predetermined capacity plans that are simple and predictable but do not provide the flexibility to adapt to the dynamic changes in the workload. Zhang et al.
[7] established that in the case of enterprise setting, an average of 25-40 percent of the distributed CPU is going to waste due to the existence of fixed over-provisioning. Auto-scaling based on threshold, which is an operation of AWS Auto Scaling Groups and Azure VMSS, activates a scale-out or scale-in processes when resource utilization exceeds established threshold. Although they are responsive compared to strictly static schemes, these reactive

systems have serious scaling delays as well as oscillation to variable workloads

C. Dynamic and predictive Provisioning.

Dynamic provisioning strategies vary the resource assignments dynamically to be adjusted on the basis of observed workload indicators. Lorido-Botran et al.
[9] make a thorough survey of the auto-scalingmethods of cloud applications and categorize the approaches into reactive, proactive and hybrid. Proactive provisioning is based on workload forecasting to supply resources before the predicted spikes in demand, and removes the latency involved with reactive scaling. Cloud workload prediction has been conducted by use of time-series forecasting ARIMA, exponential smoothing and recurrent neural networks (RNNs) in different levels of accuracy.

Li et al. [11] suggested a reinforcement based learning framework of dynamically consolidated VM in private clouds with 18 percent energy saving at the expense of SLA compliance. On the same note, it was shown by Islam et al. [12] that the neural network-based workload prediction saves 22% of the provisioning cost relative to threshold-based baselines using Google Cloud Trace data.

D. Cloud-Specific Provisioning of Hybrid Cloud.

The workload placement between the private and public tiers, the minimization of the cost of inter-cloud data transfer, and the routing of tasks depending on latency are the additional dimensions of decisions that the provisioning in the hybrid environment brings as compared with the single-cloud case. Garg et al. Cost optimization in hybrid cloud environments it is necessary to consider the price of spot/preemptible

instances, data egress fees and the price of Reserved Instance. Xu et al. [15] designed the hybrid cloud cost minimization problem as a mixed-integer linear program (MILP), which was solved using approximation algorithms that attain a solution within

5 percent of the optimal solution in a time of polynomials.

E. Research Gap.

III. RESEARCH METHODOLOGY

A. Research Design

This paper has a mixed-method research design, which involves systematic literature analysis and quantitative simulation-based experiments. Systematic review was conducted in accordance with the PRISMA principles, and 912 papers published since 2015 in IEEE Xplore, ACM Digital Library, costs and private-to-public burst provisioning logic Springer, and arXiv were searched. The 138 articles were used in the complete analysis after the introduction of inclusion criteria that included focus on the hybrid cloud, peer review, English language, and quantitative outcomes.Although each of the three has been studied individually in the context of dynamic provisioning, ML-based prediction, and hybrid cloud scheduling, a single framework that would compare the performance of these strategies, given the communication advances presented by as context of next generation network-aware provisioning requirements have not been established.

B. Simulation Framework

CloudSim Plus v7.3 extended with a custom Hybrid Cloud Module that introduced inter-tier data transfer costs and private-to-public burst provisioning logic was used to do simulation experiments. The simulated hybrid infrastructure included: (1) a private environment with resources consisting of 50 physical hosts with 32 vCPU and 128 GB of RAM; (2) a public environment based on the AWS us-east-1 pricing with the on-demand, spot, and reserved instance types; and (3) a 10 Gbps inter-cloud connection with egress cost adjustable. Main simulation parameters are summarised in Table I.

TABLE I — Simulation Parameters

Parameter	Private Cloud	Public Cloud (AWS)
Physical Hosts / Instances	50 hosts	Unlimited (on-demand)
CPU per Node	32 vCPU	2–96 vCPU (instance type)
Memory per Node	128 GB RAM	4–1,152 GB RAM
Storage	2 TB SSD/host	EBS gp3 (variable)
Network Bandwidth	10 Gbps internal	Up to 100 Gbps
Inter-Cloud Egress Cost	—	$0.09/GB
Workload Trace Duration	72 hours	72 hours
Simulation Tool	CloudSim Plus v7.3	CloudSim Plus v7.3

C. The workload is characters.

The workload traces were obtained based on the Google Cluster Workload Trace v3 (2019) that contained about 2.6 million events of tasks on 12,500 machines. Traces were scaled and mapped to thesimulation environment, and retaining the patterns of temporal burstiness. Three workloads were modeled:
(1) steady-state base loads at 45 per cent CPU utilization; (2) diurnal patterns 2025 of IEEE Xplore, ACM Digital Library, costs and private-to-public burst provisioning logic with morning/evening peaksEvaluated Strategies of Provisioning.

Three types of provisioning strategies have been tested under the same workload conditions: (1) Static Provisioning assigns a constant size of VMs with 80th percentile peak load; (2) Reactive Threshold-Based Provisioning scales out when the CPU utilization reaches 75% and scales in when it reaches 30% and does it with a 5-minute cooldown; and (3) ML-Driven Proactive Provisioning scales out on 15-minute predictions of workload by an LSTM neural network that has been trained on 30Performance Metrics.

It has assessed the following metrics: (1) Resource Utilization Rate (%) which is the average CPU utilization on provisioned VMs; (2) SLA Violation Rate (%) indicating the tasks that failed to meet response time goals; (3) Average Response Latency (ms) of user facing requests; (4) Provisioning Cost (USD/hour) which is the cost per hour of provisioning normalized to the 72 hour trace; and (5) Over-Provisioning Ratio which is the wasted capacity as a part of the overall provisioned resources.

IV. RESULTS AND DISCUSSION

A. Resource Utilization.

As shown in figure 1, the mean CPU usage of the three provisioning strategies during the 72 hours simulation was as follows. The conservative capacity allocation in the process of the static provisioning led to the only 48.2% mean utilization, whereas the reactive provisioning based on the threshold and reacting to the measured demand resulted in the 67.4% mean utilization. The best mean utilization was 79.6 per cent when produced proactively by ML that was a result of proper demand forecasting whichremoved the latency time between demand surges and the capacity being available.

In the case of bursty workload, static provisioning always ensured adequate capacity (no SLA violations as a result of over-provisioning) but reactive provisioning experienced a temporary under-provisioning during the 5-minute scaling lag leading to 4.2% SLA violations. The 15-minute look-ahead window of proactive provisioning was able to predict all burst events only and the SLA violation rate was recorded at 0.8-percent, which is 81-percent lower than the reactive methods.

TABLE II — Performance Comparison of Provisioning Strategies

Metric	Static	Reactive (Threshold)	ML Proactive
Avg. CPU Utilization (%)	48.2	67.4	79.6
SLA Violation Rate (%)	0.4	4.2	0.8
Avg. Response Latency (ms)	142	189	103
Provisioning Cost (USD/hr)	38.60	29.40	22.80
Over-Provisioning Ratio (%)	51.8	32.6	20.4
Scale-Out Events (72 hr)	—	84	31

B. Cost Analysis.

In Table II, we have the overall cost performance comparison. ML-based proactive provisioning resulted in a provisioning cost of 22.80/h, which is 41 percent lower than in the case of static provisioning (38.60/h) and 22.4 percent lower than in the case of reactive provisioning (29.40/h). The benefits of proactive provisioning are based on two mechanisms;
(1) less dependence on on-demand instances with pre-positioning Reserved Instance capacity, and (2) lower over-provisioning ratios, which removes idle computer charges.

The less expensive nature of reactive provisioning compared to the static approaches but through the higher rate of SLA violation (4.2 percent) is an indirect cost of the production environment .

V. Future Scope

A. Green Cloud Provisioning.

Incorporate carbon footprints into the ML-based provisioning model to plan data center workloads on to renewable-energy-based data centers to operate the cloud in a sustainable manner.

B. Federation of Multi-Domain Resources.

Federated learning and distributed optimization can be used to extend the provisioning framework across multiple organizations, cloud vendors and administrative domains.

C. 6G Network-Compute Co-Optimization.

Optimize network and compute resources together to achieve ultra-low latency networks, e.g. autonomous vehicles, real-time AR/VR and smart manufacturing as 6G infrastructure becomes a reality.

D. Reinforcement of Adaptive Provisioning.

Substitute LSTM-based forecasting with Deep Reinforcement Learning (DRL) agents which constantly modify provisioning policies in real time, even to unseen patterns of workload.

E. Real-Time Compulsory and regulatory

Develop a dynamic policy engine to automatically verify, in multi-jurisdiction cloud set-ups, that workloads are not violating policies such as GDPR and HIPAA by policy. driven provisioning model to schedule workloads on.

Expand the model to three-tier (edge private public) provisioning hierarchy of latency-sensitive IoT applications.

G. Learning Ensemble and Transfer of Learning to Workload Forecasting.

Implement LSTM, Transformer-based and classical models on a mixture with the transfer learning to enhance the accuracy of the prediction on new or sparse workload patterns.

H. Serverless and Micro services.

Design new scheduling schemes consisting of resource allocation frameworks that are fine-grained, event driven, and functional, and operate at millisecond time scales.

VI. CONCLUSION

The current paper has been able to provide an extensive study on resource provisioning approaches in hybrid cloud infrastructure, including those based on static, reactive, and proactive based on ML strategies. We have also shown that, with regard to all the examined dimensions, 79.6% average CPU usage, 0.8% SLA violation rate, 103 ms average response latency, and $22.80/hour provisioning cost, ML-driven proactive provisioning is significantly more

successful: 79.6% mean CPU utilization, 0.8% SLA violation rate, 103 ms average response latency and $22.80/hour provisioning cost.

The 41 percent reduction in costs and 28 percent reduction in latency of ML-driven solutions compared to their static counterparts along with the 81 percent reduction in SLA violations compared to reactive provisioning justify proactive provisioning as the solution of choice in a production hybrid cloud deployment. The technologies that have made such advances a reality, such as Kubernetes orchestration, LSTM-based workload forecasting, serverless FaaS integration, and service mesh traffic management, are all part of a fully developed provisioning technology stack that can be adopted by enterprises.

Federated multi-domain orchestration, compliance-aware placement, multi-tenant isolation and green provisioning are all open challenges. Future directions will involve carbon-conscious workload scheduling combined with ML-based provisioning, real-time enforcement of compliance policies and network-compute co-optimization systems in line with the 6G infrastructure necessities. These developments will make hybrid cloud provisioning one of the foundational providers of sustainable and high-performance digital infrastructure over the course of the next decade

VII. REFERENCES

[1] P. Mell and T. Grance, NIST Definition of Cloud Computing, NIST Special Publication 800-145, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2011.
[2] R. Buyya, C. S. Yeo, S. Venugipal, J. Broberg and
I. Brandic, Cloud computing and emerging IT platforms: Vision, hype and reality in providing computing as the 5th utility, Future Gener. Comput. Syst., vol. 25, no. 6, pp. 599–616, Jun. 2009.
[3] A. N. Toosi, R. N. Calheiros and R. Buyya, "Interconnected cloud computing environments: Challenges, taxonomy and survey, ACM Comput. Surv., vol. 47, no. 1, pp. 1–47, Jul. 2014.
[4] Amazon Web Services, "Amazon EC2: Elastic Compute Cloud — Developer Guide," Amazon Web Services, Inc., Seattle, WA, USA, 2006. [Online].
Available: https://aws.amazon.com/ec2/
[5] B. Burns, B. Grant, D. Oppenheimer, E. Brewer, and J. Wilkes, "Borg, Omega, and Kubernetes: 10 years of lessons learned involving container-management systems," ACM Queue, vol. 14, no. 1,
pp. 7093, Jan. 2016.
[6] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing:State-of-the-art and research challenges,"
J. Internet Serv. Appl., vol. 1, no. 1, pp. 718, May 2010.
[7] N. Roy, A. Dubey, and A. Gokhale, "Efficient Autoscaling in the Cloud with the Workload Forecasting based on Predictive Models," in Proc. IEEE 4th Int.
[8] T. Lorido-Botran, J. Miguel-Alonso, and J. A. Lozano, A Review of Auto-Scaling Techniques to Elastic Applications in Cloud Environments, J. Grid Comput., vol. 12, no. 4, pp. 559 592, Dec. 2014.
[9] A. Canziani, A. Paszke and E. Culurciello, Workload Prediction with ARIMA Model and its effects in Cloud Applications, in Proc. IEEE Int. Conf. Cloud Netw. CloudNet (CloudNet), Niagara Falls, ON, Canada, 2016, pp. 1-6.
[10] Z. Li, J. Ge, H. Hu, W. Song, H. Hu, and B. Luo, Cost and Energy Aware Scheduling Algorithm of Scientific Workflow with Deadline Constraint in Clouds, IEEE Trans. Serv. Comput., vol. 11, no. 4,
pp. 713726, Jul.Aug. 2018.
[11] S. Islam, J.Keung, K. Lee, and A.Liu, Empirical Prediction Models of Adaptive resource Provisioning in the Cloud, Future Gener. Comput. Syst., vol. 28, no. 1, pp. 155–162, Jan. 2012.
[12] S. K. Garg, S. Versteeg and R. Buyya, Ranking Framework of Cloud computing Service, Future Gener. Comput. Syst., vol. 29, no. 4, pp. 1012–1023,
Jun. 2013.
[13] L. Xu, C. Zheng, and Q. Wang, "Providing resources in a cost-effective and efficient way to a hybrid cloud computing environment with mixed-intensity computing through mixed-integer linear programming," IEEE Trans. Cloud Comput., vol. 9, no. 3, p. 11021115, Jul.Sep. 2021.