

Workload matchmaking in the cloud: Finding the Right VM Fit

Priyadarshni Shanmugavadivelu *

Birla Institute of Technology and Science, Pilani, India.

World Journal of Advanced Engineering Technology and Sciences, 2025, 15(03), 187–195

Publication history: Received on 23 April 2025; revised on 30 May 2025; accepted on 02 June 2025

Article DOI: <https://doi.org/10.30574/wjaets.2025.15.3.0932>

Abstract

This article explores the multifaceted discipline of workload-to-VM matchmaking in cloud environments, presenting a comprehensive framework for optimizing the alignment between application requirements and infrastructure capabilities. The article examines the diverse landscape of specialized virtual machines offered by major cloud providers, each designed to excel in specific dimensions of computational performance. Through systematic workload profiling methodologies, organizations can develop empirical understanding of their applications' resource consumption patterns, creating the foundation for informed VM selection decisions. The article investigates benchmarking strategies that provide quantitative performance data alongside price-performance analysis frameworks that balance technical capabilities with financial considerations. The ongoing nature of optimization is addressed through exploration of cloud-native tools and continuous improvement strategies that adapt infrastructure as workloads evolve. By synthesizing technical analysis with business context, this article equips cloud practitioners with methodologies to enhance application performance, maximize resource utilization, and achieve sustainable cost optimization in increasingly complex cloud environments.

Keywords: Workload-Vm Optimization; Cloud Resource Profiling; Performance Benchmarking; Price-Performance Analysis; Continuous Infrastructure Optimization

1. Introduction

Cloud computing has revolutionized modern IT infrastructure by enabling on-demand access to a diverse array of computing resources. At the heart of this transformation lies the virtual machine (VM) - a software emulation of physical computing hardware that provides the foundation for most cloud deployments. Today's major cloud service providers offer an increasingly complex catalog of VM types, each tailored to specific workload characteristics and performance requirements [1]. This heterogeneity presents both an opportunity and a challenge for organizations seeking to optimize their cloud deployments.

The selection of appropriate VM types represents a critical decision point with far-reaching implications for application performance, operational costs, and business agility. Research indicates that misaligned workload-to-VM pairings can lead to performance degradation of up to 30% and cost inefficiencies exceeding 40% compared to optimized deployments. Despite these stakes, many organizations continue to select VM types based on intuition, legacy infrastructure specifications, or outdated heuristics rather than empirical workload analysis.

This article examines the multifaceted process of matching workloads to their ideal VM configurations. The article explores systematic approaches to workload profiling that reveal the resource consumption patterns and computational characteristics unique to each application. These insights serve as the foundation for informed VM selection decisions. The article further investigates benchmarking methodologies that quantify VM performance across relevant scenarios, enabling objective comparison between offerings.

* Corresponding author: Priyadarshni Shanmugavadivelu.

Beyond technical performance considerations, the article addresses the economic dimensions of VM selection through price-performance analysis frameworks. This holistic approach acknowledges that optimal VM selection balances technical performance with financial considerations, recognizing that the most powerful VM is rarely the most cost-effective solution for a given workload.

The cloud's dynamic nature necessitates ongoing optimization rather than point-in-time decisions. The article therefore examines cloud-native tools and continuous optimization strategies that help organizations adapt VM selections as workloads evolve, cloud offerings expand, and business requirements change. Through practical case studies, the article demonstrates how these principles translate into real-world infrastructure optimizations across diverse application domains.

This article provides a comprehensive framework for workload-VM matchmaking, aiming to equip cloud practitioners with the knowledge and methodologies needed to enhance application efficiency, maximize infrastructure utilization, and achieve sustainable cost optimization in increasingly complex cloud environments.

2. Understanding Vm Diversity

Cloud providers have developed specialized VM families to address diverse workload requirements, each optimized for specific resource dimensions and performance characteristics [2].

Table 1 VM Type Comparison by Primary Resource Characteristics [2]

VM Type	Primary Use Case	CPU Ratio	Network Capability	Storage Performance	Cost Efficiency Factor
General-purpose	Balanced workloads, web applications	Balanced (1:4)	Moderate	Moderate	Baseline
Compute-optimized	Batch processing, scientific modeling	High (1:2)	Low-Moderate	Low	High for CPU-bound workloads
Memory-optimized	In-memory databases, caching	Low (1:8+)	Moderate	Low-Moderate	High for memory-intensive apps
Storage-optimized	Data warehouses, file systems	Moderate (1:4)	Moderate	High (NVMe, high IOPS)	High for I/O-bound workloads
Network-optimized	Content delivery, data movement	Moderate (1:4)	High (10-100 Gbps)	Moderate	High for data transfer operations

Compute-optimized VMs feature high CPU-to-memory ratios with premium processors, often offering higher clock speeds and advanced instruction sets. These VMs excel in batch processing, scientific modeling, high-performance web servers, and media encoding. Their CPU-centric design typically delivers 4-8 vCPUs per GB of RAM, prioritizing computational throughput over memory capacity.

Memory-optimized VMs invert this ratio, providing substantial RAM resources relative to CPU allocation. With memory-to-CPU ratios often exceeding 8:1, these instances support in-memory databases, real-time analytics platforms, and large caching layers. Their architecture minimizes access latency to memory-resident data, enabling applications like SAP HANA and Redis to maintain performance at scale.

Storage-optimized VMs balance computational resources with enhanced I/O capabilities through specialized local storage configurations. These VMs incorporate NVMe SSDs with dedicated throughput channels and optimized storage controllers, making them ideal for distributed file systems, data warehouses, and transaction-heavy databases requiring high IOPS performance.

Network-optimized VMs leverage enhanced network interface cards, RDMA capabilities, and optimized virtualization stacks to deliver superior throughput and reduced latency. These instances typically offer 10-100 Gbps networking capabilities, supporting network appliances, distributed computing frameworks, and high-frequency trading platforms where data movement represents the primary bottleneck.

General-purpose VMs provide balanced resource allocations suitable for varied workloads with moderate demands across all resource dimensions. These flexible instances serve as the foundation for web applications, development environments, and small-to-medium databases, offering reasonable performance without specialization.

3. Workload Profiling Methodologies

Effective workload profiling provides the empirical foundation for VM selection decisions by systematically analyzing application behavior and resource consumption patterns [3].

Resource utilization analysis techniques employ both agent-based and agentless monitoring approaches to capture CPU, memory, I/O, and network metrics across application tiers. Modern profiling tools utilize sampling at variable frequencies, typically between 1-60 second intervals, to minimize observer effects while maintaining statistical validity. Time-series collection enables correlation between resource consumption and business metrics, revealing how infrastructure requirements scale with user activity.

Identifying computational bottlenecks requires multi-dimensional analysis beyond aggregate utilization. Thread contention analysis, system call tracing, and hardware performance counter examination reveal whether applications are CPU-bound, memory-bound, or I/O-bound under peak load conditions. These insights inform VM family selection by highlighting which resource dimension most constrains application performance.

Temporal patterns in resource consumption emerge through trend analysis across multiple time horizons. Daily, weekly, and seasonal patterns influence not only VM selection but also scaling policies. Applications exhibiting predictable utilization patterns benefit from preemptively sized VMs, while those with erratic demand profiles may favor smaller baseline VMs with burst capacity.

Workload characterization frameworks formalize profiling through standardized methodologies. The TPC (Transaction Processing Performance Council) benchmarks provide industry-standard workload definitions for transaction processing and analytical systems. Cloud-native frameworks like SPEC Cloud IaaS combine workload generators with monitoring toolchains to create reproducible profiling environments.

3.1. Case study

Profile analysis of common enterprise applications demonstrates these concepts in practice. A mid-sized e-commerce platform's profile revealed distinct resource patterns across services: the product catalog exhibited memory-bound behavior with high read rates, while the order processing system showed CPU-sensitivity during checkout operations. This led to a hybrid deployment using memory-optimized VMs for catalog services and compute-optimized VMs for transaction processing, resulting in 28% cost reduction while maintaining performance SLAs.

4. Benchmarking Strategies for VM Selection

Benchmarking provides quantitative performance data to guide VM selection decisions, transforming abstract workload requirements into measurable metrics that facilitate objective comparison across VM options [4].

Industry-standard benchmarking tools offer reproducible performance assessment frameworks. SPEC CPU 2017 measures compute-intensive performance, while TPC-E evaluates OLTP database workloads. Cloud-specific benchmarks like YCSB (Yahoo! Cloud Serving Benchmark) and CloudSuite simulate realistic cloud application patterns. These standardized tools enable cross-provider comparisons using consistent methodologies, though their synthetic nature may not fully reflect production workload characteristics.

Designing workload-specific benchmarks addresses this limitation by creating custom performance tests that mirror production conditions. Effective custom benchmarks capture representative data volumes, transaction patterns, and concurrency models. Organizations increasingly employ infrastructure-as-code approaches to create reproducible benchmark environments that simulate their specific workloads. This approach typically involves instrumenting existing applications or developing lightweight mimics that replicate core functionality without implementation complexity.

Comparative analysis methodologies establish structured approaches to VM evaluation. A-B testing methodology isolates performance variables by comparing VMs with consistent workloads over equivalent time periods. The Scientific Method framework applies formal hypothesis testing to VM selection, establishing baseline performance,

predicting improvements from alternative configurations, and validating results through controlled experimentation. These methodologies minimize confirmation bias in VM selection decisions.

Interpreting benchmark results meaningfully requires contextualizing raw performance data. Effective interpretation considers performance variability, identifying not just average metrics but distribution patterns that reveal consistency. Scalability curves plotting performance against concurrency levels often reveal non-linear relationships more informative than single-point measurements. The most insightful analyses correlate technical metrics with business outcomes, translating IOPS or request latency into user experience or revenue impact.

Limitations of benchmarking approaches include the observer effect, where monitoring overhead distorts results, particularly for I/O-intensive workloads. Multi-tenant environments introduce noisy neighbor effects that create performance variability not captured in isolated benchmarks. Organizations must also consider temporal limitations, as point-in-time benchmarks may not reflect long-term performance trends or off-peak behavior patterns.

5. Price-performance analysis framework

Optimizing cloud infrastructure requires balancing technical performance with financial considerations through structured price-performance analysis methodologies [5].

Table 2 Workload Optimization Decision Framework [5]

Decision Factor	Assessment Questions	Optimization Action	Impact Level
Resource Bottleneck	Which resource consistently reaches >80% utilization? Where do application threads spend most time waiting?	Select VM type optimized for constrained resource	High
Usage Pattern	Is workload consistent or variable? Are there predictable peak periods?	Choose between reserved instances (stable) or auto-scaling (variable)	Medium-High
Performance Requirements	What are the SLA requirements? Are there critical latency thresholds?	Balance VM capabilities with performance needs	High
Budget Constraints	What is the maximum acceptable infrastructure cost? Is cost or performance priority?	Implement cost guardrails with performance thresholds	Medium
Growth Projections	What is the expected workload growth rate? Will resource needs change over time?	Design for scalability with headroom	Medium-Low
Compliance Requirements	Are there data residency requirements? What security controls are needed?	Select VM types/locations meeting compliance needs	Variable

Total cost of ownership calculations extends beyond basic hourly VM rates to include storage costs, data transfer fees, management overhead, and licensing implications. Comprehensive TCO models incorporate reservation pricing, spot instance opportunities, and sustained use discounts. Modern approaches employ parametric TCO modeling that adjusts for workload growth projections and fluctuating resource needs rather than static calculations based on current requirements.

Performance-per-dollar metrics normalize benchmark results against infrastructure costs to identify value-optimized configurations. Common approaches include cost-per-transaction, cost-per-query, or infrastructure-cost-as-percentage-of-revenue models. Leading organizations establish workload-specific efficiency metrics, such as "cost per thousand recommendations" for ML-driven systems or "dollars per concurrent user" for SaaS applications, creating direct links between financial and technical optimization.

Balancing capital and operational expenditures involve strategic decisions about commitment levels. Reserved instances offer substantial discounts (typically 30-60%) in exchange for 1–3-year commitments, while on-demand resources provide flexibility at premium pricing. Hybrid approaches use reservations for baseline capacity with on-

demand resources for peak handling. This balance shifts as workloads mature, with early-stage applications favoring flexibility while stable workloads benefit from commitment-based pricing.

Accounting for indirect costs and benefits captures value beyond direct resource pricing. Developer productivity costs associated with VM-specific optimizations often outweigh infrastructure savings for smaller deployments. Similarly, specialized VM types that reduce code complexity or eliminate architectural components may justify premium pricing through reduced maintenance overhead. A holistic framework incorporates these indirect factors through weighted decision matrices.

Risk assessment in VM selection decisions evaluates potential negative outcomes from infrastructure choices. Common risks include performance degradation during peak periods, unexpected scaling limitations, and infrastructure lock-in. Effective risk management employs strategies such as multi-VM type deployments, heterogeneous environment testing, and architectural designs that facilitate VM migration as requirements evolve.

6. Cloud-Native Tools for VM Optimization

Cloud providers have developed sophisticated tooling ecosystems that help organizations optimize VM selection and utilization throughout the resource lifecycle [6].

VM recommendation engines analyze resource utilization patterns to suggest right-sized instances based on historical workload data. AWS Compute Optimizer and Azure Advisor evaluate CPU, memory, network, and storage metrics to generate instance recommendations that can reduce costs by 10-25% while maintaining performance. These engines typically require 2-4 weeks of telemetry data to generate reliable recommendations. Their limitations include difficulty accounting for seasonal variations and limited visibility into application-level requirements that aren't reflected in system metrics.

Automated resource scaling technologies adjust capacity dynamically in response to changing demands. Horizontal scaling approaches add or remove identical VM instances based on workload metrics, while vertical scaling modifies individual VM resources by switching instance types. Advanced auto-scaling implementations combine predictive scaling based on historical patterns with reactive scaling triggered by real-time metrics. Recent innovations include container-based auto-scaling that operates at sub-VM granularity, enabling more precise resource allocation.

Performance monitoring and alerting systems provide visibility into VM behavior through comprehensive metric collection and analysis. Cloud-native monitoring platforms like Azure Monitor, Google Cloud Monitoring and AWS CloudWatch collect high-resolution metric data with sub-minute granularity. These systems employ machine learning to establish dynamic baselines that adapt to workload patterns, enabling anomaly detection that identifies performance issues before they impact users. Integration with distributed tracing frameworks such as Azure Application Insights, AWS X-Ray, and Google Cloud Trace correlates infrastructure metrics with application performance insights.

Cost management and optimization platforms centralize financial analysis of cloud resources through granular reporting and allocation mechanisms. These platforms implement budget controls, anomaly detection for spending patterns, and resource lifecycle management functionality. Advanced implementations incorporate FinOps principles by aligning technical decisions with financial impact, enabling teams to make cost-informed architecture choices. Third-party solutions like CloudHealth and Flexera complement native provider tools with cross-cloud visibility and enhanced governance capabilities.

Integration of optimization tools into deployment pipelines embeds resource efficiency into application delivery processes. Infrastructure-as-code templates incorporate parametrized instance selection based on environment requirements. CI/CD pipelines increasingly include performance testing phases that validate resource efficiency alongside functional requirements. GitOps approaches enable version-controlled infrastructure definitions with automated drift detection and remediation, ensuring that optimization decisions persist through deployment iterations.

7. Continuous optimization strategies

Effective cloud infrastructure management requires ongoing optimization rather than point-in-time decisions, implementing systematic approaches to resource refinement as applications evolve [7].

Real-time monitoring and feedback loops establish continuous visibility into VM performance characteristics. High-resolution telemetry captures resource utilization patterns across multiple dimensions, while synthetic transactions

measure end-user experiences metrics. These observability systems feed into automated decision frameworks that implement rules-based optimization policies. Leading organizations implement closed-loop systems where anomaly detection triggers automated remediation actions within predefined safety parameters.

Migration strategies for workload rebalancing enable organizations to refine VM selections as requirements change. Live migration technologies allow workload movement between VM types with minimal disruption, typically requiring 30-120 seconds of reduced performance rather than complete downtime. Containerization simplifies workload portability through environment standardization, while application-aware migration tools preserve state during transitions. Effective migration playbooks define not only technical processes but also validation methodologies to verify performance in new environments.

Predictive analytics for proactive optimization apply forecasting techniques to anticipated resource requirements. Time-series analysis identifies cyclical patterns in resource consumption, enabling proactive scaling ahead of predicted demand spikes. Machine learning models trained on historical performance data predict capacity requirements with increasing accuracy as training data accumulates. These forecasting capabilities allow organizations to reserve capacity in advance of predictable demand increases, balancing cost efficiency with performance reliability.

Organizational processes for ongoing infrastructure review formalize optimization as a continuous practice rather than periodic project. Establishing regular optimization cycles with defined review points ensures consistent attention to infrastructure efficiency. Cross-functional cloud centers of excellence bring together financial, technical, and operational stakeholders to align optimization efforts with business objectives. Data-driven review methodologies establish objective improvement metrics that quantify optimization impact beyond cost reduction.

Change management considerations recognize that optimization initiatives impact application stability and team workflows. Effective approaches implement progressive optimization that applies changes incrementally with verification at each stage. Optimization backout plans define rollback procedures when changes don't deliver expected benefits. Communication frameworks ensure that all stakeholders understand optimization initiatives, promoting organizational alignment around infrastructure efficiency goals while managing expectations about potential disruption.

8. Case Studies and Practical Applications

Real-world implementations demonstrate the practical impact of workload-VM optimization across diverse application domains and organizational contexts [8].

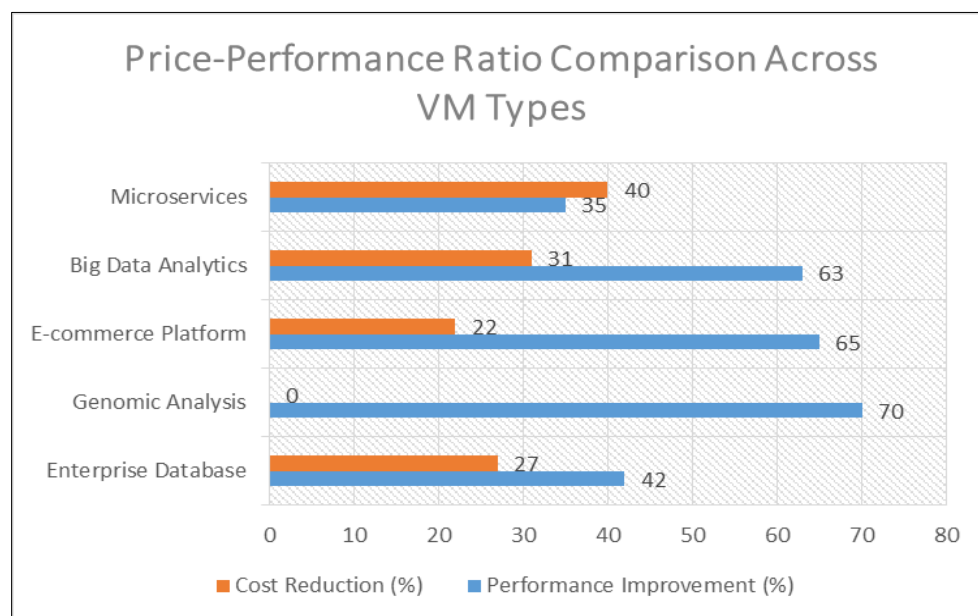


Figure 1 Performance Improvement and Cost Reduction by VM Type Optimization Across Application Domains [8]

Enterprise database workload optimization presents distinctive challenges due to complex I/O patterns and memory requirements. A prominent financial services organization migrated its Oracle RAC deployment from general-purpose VMs to storage-optimized instances, achieving 42% performance improvement while reducing infrastructure costs by 27%. Key to this optimization was profiling that revealed random I/O patterns with 70% read operations, leading to VM selection with enhanced local NVMe storage and optimized storage controllers. The implementation maintained high availability through synchronous replication between availability zones while leveraging reserved instances for baseline capacity with burst capability for quarterly reporting peaks.

High-performance computing in research environments benefits significantly from compute-optimized VM selection. A genomics research institute implemented a hybrid approach for sequence analysis pipelines, using compute-optimized instances with high-performance processors for alignment tasks while employing memory-optimized instances for variant calling operations. This workload-aware infrastructure design reduced processing time for a standard 30x genome analysis from 27 hours to 8 hours while maintaining budget neutrality. The architecture employed spot instances for interruptible tasks with checkpointing mechanisms to balance cost efficiency with completion reliability.

Web application scaling scenarios demonstrate the value of VM diversity across application tiers. An e-commerce platform implemented a three-tier architecture with distinct VM types for each layer: network-optimized instances for load balancing and SSL termination, general-purpose VMs for application servers, and memory-optimized instances for caching layers. This workload-aligned architecture maintained sub-200ms response times during seasonal traffic spikes of 5x normal volume while implementing automated scaling rules specific to each tier's resource constraints.

Big data analytics infrastructure optimization requires balancing compute, memory, and storage characteristics. A media analytics company processing 4TB of video content daily redesigned its Hadoop infrastructure using storage-optimized instances for HDFS data nodes while implementing memory-optimized instances for Spark processing jobs. This specialized VM allocation improved processing throughput by 63% while reducing end-to-end analytics latency from hours to minutes, enabling near-real-time content recommendations that increased user engagement metrics.

Microservices architecture deployment strategies benefit from granular resource allocation aligned to service-specific requirements. A SaaS provider decomposed its monolithic application into 37 microservices, each deployed on appropriately sized VMs based on individual service profiles. This fine-grained approach reduced overall infrastructure footprint by 40% compared to the previous monolithic deployment on general-purpose instances. The implementation combined container orchestration with VM diversity, placing high-throughput services on compute-optimized instances while data-intensive components utilized memory-optimized VMs.

9. Future Trends in Workload-VM Matching

Emerging technologies and evolving cloud business models are reshaping the workload-VM matching landscape, introducing new optimization strategies and infrastructure paradigms [9].

AI-driven infrastructure optimization represents a significant advancement beyond rule-based resource management. Machine learning systems now analyze multidimensional performance data to predict optimal VM configurations with increasing accuracy. Google's Morpheus system demonstrates how reinforcement learning can dynamically adjust resource allocations based on application behavior patterns not discernible through traditional heuristics. These AI systems increasingly incorporate business metrics alongside technical indicators, optimizing for revenue impact rather than purely technical efficiency. Organizations implementing AI-driven optimization report 15-30% efficiency improvements compared to traditional approaches.

Serverless computing implications challenge traditional VM selection paradigms by abstracting infrastructure decisions away from developers. Function-as-a-Service platforms dynamically allocate computational resources at millisecond granularity, fundamentally changing optimization from VM selection to function design patterns. This paradigm shift introduces new efficiency metrics focused on execution duration and memory allocation precision rather than instance type selection. Hybrid architectures increasingly combine serverless components for variable workloads with VM-based systems for predictable baseline processing, requiring integrated optimization across deployment models.

Hybrid and multi-cloud considerations introduce complexity to workload matching strategies by expanding the available VM catalog across providers. Organizations increasingly implement policy-based workload placement that considers not only VM characteristics but also geographic distribution, compliance requirements, and provider-specific capabilities. Cloud arbitrage strategies dynamically shift workloads between providers based on pricing variations and performance characteristics, though these approaches must balance migration costs against optimization benefits.

Unifying management layers now provide consistent optimization recommendations across heterogeneous cloud environments.

Evolving pricing models and their impact on selection strategies extend beyond traditional per-hour VM billing. Consumption-based pricing aligns costs directly with application resource utilization rather than provisioned capacity, incentivizing fine-grained optimization. Committed use discounts with flexible resource allocation allow organizations to commit to spending levels without specifying exact VM types, providing financial predictability while maintaining technical flexibility. Performance-tiered pricing structures where providers charge premium rates for guaranteed performance levels rather than raw resources are emerging as alternatives to traditional pricing models.

Emerging specialized VM types target increasingly specific workload profiles with custom hardware configurations. GPU-optimized instances now differentiate between training and inference workloads with specialized accelerator types and memory architectures. FPGA-enabled VMs provide hardware-level customization for specialized workloads like genomics processing and financial modeling. Quantum processing units (QPUs) accessible as specialized cloud resources represent the frontier of workload-specific computing resources, though practical applications remain limited. These specialized resources require increasingly sophisticated matching methodologies that understand both traditional computing metrics and domain-specific performance characteristics.

10. Conclusion

Effective workload-VM matching represents a critical competency in modern cloud computing, demanding a sophisticated synthesis of technical analysis, financial modeling, and operational strategy. As the exploration has demonstrated, organizations that implement systematic approaches to workload profiling, benchmarking, and continuous optimization achieve demonstrable advantages in performance, cost efficiency, and operational agility. The article presented throughout this article—progressing from workload characterization through VM selection to ongoing optimization—provides practitioners with a comprehensive methodology for infrastructure decisions that align technical capabilities with business requirements. While emerging technologies like AI-driven optimization, serverless computing, and specialized accelerators continue to reshape the cloud landscape, the fundamental principles of workload-infrastructure alignment remain constant. Organizations that establish disciplined processes for understanding application requirements, evaluating infrastructure options through objective performance and cost metrics, and implementing continuous optimization feedback loops will maintain a competitive advantage regardless of how cloud technologies evolve. As cloud environments become increasingly complex and heterogeneous, the ability to match workloads to their ideal infrastructure environments will distinguish organizations that merely use cloud services from those that strategically leverage cloud capabilities to drive business outcomes.

References

- [1] Anirban Bhattacharjee, Yogesh Barve et al. "CloudCAMP: Automating the Deployment and Management of Cloud Services." IEEE International Conference on Services Computing. 2018:237-244, 9 April 2019. <https://arxiv.org/pdf/1904.02184>
- [2] Eli Cortez, Anand Bonde, et al. "Resource Central: Understanding and Predicting Workloads for Improved Resource Management in Large Cloud Platforms." Proceedings of the 26th Symposium on Operating Systems Principles. 2017:153-167, 14 October 2017. <https://dl.acm.org/doi/10.1145/3132747.3132772>
- [3] Arijit Khan; Xifeng Yan et al. "Workload Characterization and Prediction in the Cloud: A Multiple Time Series Approach." IEEE Network Operations and Management Symposium. 2012:1287-1294, 07 June 2012. <https://ieeexplore.ieee.org/document/6212065>
- [4] Phillip Leitner, Jürgen Cito J. "Patterns in the Chaos—A Study of Performance Variation and Predictability in Public IaaS Clouds." ACM Transactions on Internet Technology. 19 April 2016 ;16(3): 15:1-15:23. <https://dl.acm.org/doi/10.1145/2885497>
- [5] Artan Mazrekaj, Isak Shabani, et al. "Pricing Schemes in Cloud Computing: An Overview." International Journal of Advanced Computer Science and Applications. 2016;7(2):80-86. <https://thesai.org/Publications/ViewPaper?Volume=7&Issue=2&Code=IJACSA&SerialNo=11>
- [6] Mina Sedaghat; Eddie Wadbro et al. "DieHard: Reliable Scheduling to Survive Correlated Failures in Cloud Data Centers." IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. 2016:52-59, 21 July 2016. <https://ieeexplore.ieee.org/document/7515669/similar#similar>

- [7] Bruno L.Dalmazo, João P. Vilela, et al. "Performance Analysis of Network Traffic Predictors in the Cloud." *Journal of Network and Systems Management*. 2017;25(2):290-320, 14 September 2016. <https://link.springer.com/article/10.1007/s10922-016-9392-x>
- [8] Alexandru Iosup, Radu Prodan et al. "IaaS Cloud Benchmarking: Approaches, Challenges, and Experience." In: *Cloud Computing for Data-Intensive Applications*. Springer. 15 November 2014:83-104. https://link.springer.com/chapter/10.1007/978-1-4939-1905-5_4
- [9] Eric Jonas, Johann Schleier-Smith, et al. "Cloud Programming Simplified: A Berkeley View on Serverless Computing." Technical Report No. UCB/EECS-2019-3, University of California, Berkeley. February 10, 2019. <https://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-3.pdf>