Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 07| 2025

Research Article **Received:** 03-06-2025| **Accepted:** 25-06-2025 | **Published:** 12-07-2025

AI-Powered Infrastructure Management: From Reactive Operations to Predictive Intelligence in Cloud Computing

Nirup Baer
Independent Researcher, USA

Abstract: The exponential expansion of digital ecosystems has presented hitherto unheard-of problems with infrastructure maintenance, therefore driving artificial intelligence to maximize cloud activities. A paradigm shift marks AIOps from conventional reactive monitoring to proactive, intelligent system management. Using machine learning models now allows predictive autoscaling through reinforcement learning techniques that project workload patterns and dynamically change compute resources. While sophisticated root cause analysis offers contextual insights from performance metrics and log data, advanced anomaly detection systems filter monitoring noise to spot real service degradations. By dynamically balancing cost, performance, and availability concerns, AI-driven capacity planning maximizes resource consumption throughout hybrid and multicloud settings. Open source systems like Kubeflow and exclusive innovations from major cloud providers propel invention in this industry. As AIOps evolves from experimental technology to operational need, companies employing AI-powered infrastructure have great advantages in resilience, cost effectiveness, and scalability. This transformation reflects the incorporation of adaptive intelligence rather than only automation, matching infrastructure performance with shifting company objectives.

Keywords: AIOps, infrastructure management, machine learning, cloud computing, predictive autoscaling.

INTRODUCTION

The Evolution of Infrastructure Management In The Ai Era

The growing complexity of digital ecosystems and cloud operations

Digital ecosystems have changed drastically, therefore presenting operational difficulties beyond traditional management skills. Multi-tenant cloud systems now run complex service meshes wherein separate components scale independently across availability zones. For temporary job management, container orchestration provides abstraction levels that require sophisticated coordination tools. Geographical dispersion of computing resources presents data sovereignty demands and latency issues that complicate architectural choices. Network topologies change dynamically as edge nodes enter and exit clusters, therefore necessitating adaptable management strategies. Managing this maze of linked systems, where a single configuration change can cascade over dozens of dependent services, infrastructure teams are under increasing pressure to keep service continuity.

The meaning and application of AI-Powered Operations, or AIOps

Combining computational intelligecen with infrastructure management to produce self-governing systems (Appio, F. P. *et al.*,2023), AIOps signifies a radical change in operational philosophy. This fusion combines cognitive skills to decode sophisticated operational signals, hence going beyond simple automation. Heterogeneous data streams, including metrics, logs, traces, and events, are processed by machine learning algorithms to build thorough operational views. Pattern identification skills find faint connections between different system parts that human operators could miss. Predictive fault detection, automatic root cause analysis, and intelligent remedy orchestration are all included in the technical scope. To support optimal decision-making, these platforms integrate data from historical occurrences, current system conditions, and projected future circumstances. Infrastructure teams use these abilities to move from reactive firefighting to proactive system optimization.

Table 1: Evolution of Infrastructure Management Approaches (Appio, F. P. *et al.*,2023; Gula, A. *et al.*,2020)

Management Aspect	Traditional Approach	AIOps-Enabled Approach
Incident Detection	Threshold-based alerts	Pattern recognition & anomaly detection
Resource Scaling	Reactive (post-event)	Predictive (pre-event)
Root Cause Analysis	Manual log investigation	Automated correlation analysis
Decision Making	Rule-based policies	ML-driven optimization
Operational Mode	Human-initiated	Autonomous/Self-healing
Data Processing	Siloed monitoring tools	Unified data analytics

***Corresponding Author:** Nirup Baer

DOI- https://doi.org/10.5281/zenodo.15870342

The purpose of the study and the importance of integrating AI into infrastructure management

Ambitious targets related to building really autonomous operating settings (Gula, A. *et al.*,2020) drive present infrastructure management research. Researchers create algorithms able to learn optimum settings via experiential feedback instead of fixed guidelines. Considerable work goes toward developing models that keep performance stable despite ever-changing workload features. The actual consequences include organizational structures, as smaller teams can control enormously bigger infrastructure footprints. Through exact resource allocation approaches, intelligent systems detect and eradicate garbage and, so naturally, evolve cost optimization. Decreasing incident frequencies and faster recovery times following outages show how service quality has improved; this helps companies to be competitive in markets when digital performance and corporate success are directly related.

Overview of Machine Learning Applications in Resource Optimization and System Reliability

Resource optimization uses complex mathematical models that adjust to infrastructure behavior patterns over long operational periods (Gula, A. *et al.*,2020). Gradient boosting techniques examine capacity usage trends to forecast future demand increases with great accuracy. Anomaly detection systems spot deviations from typical operational baselines using isolation forests and local outlier variables. Through simulated natural selection techniques, genetic algorithms develop best-fit resource allocation plans. Deep learning architectures analyze unorganized log data to find significant operational insights that are concealed in verbose output streams. Through trial and error, reinforcement learning systems use infrastructure APIs to find effective operating rules. They all aim to create infrastructure that anticipates needs, avoids malfunctions, and maximizes performance without continual human supervision, even though they employ various approaches.

USING PREDICTIVE AUTOSCALING AND MACHINE LEARNING FOR DYNAMIC RESOURCE MANAGEMENT

Fundamentals of Predictive Autoscaling in Cloud Environments

Predictive autoscaling changes resource allocation from threshold-based responses to anticipatory allocations powered by machine learning algorithms. Complex patterns of workload in cloud environments are affected by temporal elements, user activity, and outside events that conventional reactive scaling systems capture insufficiently. Modern autoscale solutions predict future resource needs by means of multi-dimensional feature sets that comprise CPU consumption, memory footprint, network throughput, and application-specific indicators. To differentiate between short peaks and persistent load increases, these predictive models include seasonality detection, trend analysis, and anomaly recognition. The basic difficulty resides in striking a compromise between prediction accuracy and computational overhead, since too sophisticated models may use resources similar to those for the workloads they intend to improve. Directly integrated into their orchestration systems, cloud providers include these features to provide smooth transitions between expected and actual resource needs while still meeting service-level objectives.

Reinforcement Learning Techniques for Kubernetes Pod Scheduling

Reinforcement learning techniques that maximize pod placement and resource allocation choices (Cheng, W. *et al.*,2023) greatly help Kubernetes orchestration. These methods portray the scheduling issue as a Markov decision process in which agents learn best policies by interacting with cluster environments. While actions include positioning decisions and resource allotments, state representations include node capacities, pod requirements, affinity rules, and quality of service classes. Reward mechanisms trade several goals, including application performance, operational costs, and resource usage efficiency. Particularly in heterogeneous environments with different hardware capabilities and job characteristics, trial-and-error learning allows nonobvious scheduling methods to be discovered that surpass heuristic methods (Cheng, W. *et al.*,2023). The computing power scheduling paradigm shows how reinforcement learning agents adapt to dynamic cluster settings without clear coding of scheduling rules.

Publisher: SARC Publisher

Table 2: Machine Learning Techniques for Infrastructure Optimization (Cheng, W. *et al.*,2023; Suksriupatham, N., & Hoonlor, A. 2020)

ML Technique	Infrastructure Application	Key Benefits
Reinforcement Learning	Pod scheduling, resource allocation	Adaptive policy learning
Time-Series Forecasting	Workload prediction, capacity planning	Proactive scaling
Regression Models	Resource demand estimation	Cost optimization
Ensemble Methods	Multi-metric analysis	Improved accuracy
LSTM Networks	Temporal pattern recognition	Long-term dependencies
Gradient Boosting	Performance prediction	Non-linear relationships

Analysis of Historical Usage Patterns and SLA-Driven Optimization

Historical workload analysis forms the foundation for accurate resource prediction models that prevent both over-provisioning and under-provisioning scenarios (Suksriupatham, N., & Hoonlor, A. 2020). Time-series decomposition techniques extract seasonal patterns, growth trends, and cyclical variations from usage data spanning multiple temporal granularities. Regression models incorporate these historical insights to project future resource requirements while accounting for business growth and changing usage patterns (Suksriupatham, N., & Hoonlor, A. 2020). Service-level agreements introduce additional constraints that prediction models must respect, transforming the optimization problem into a multi-objective challenge. In order to find minimal viable configurations that meet performance criteria, machine learning algorithms examine the link between resource allocation levels and SLA compliance measures. Advanced implementations incorporate cost functions that penalize SLA violations more heavily than excess capacity, reflecting business priorities in resource allocation decisions.

Case Studies of Workload Surge Prediction and Resource Adjustment

Real-world implementations demonstrate the practical effectiveness of predictive autoscaling across diverse application domains and traffic patterns. E-commerce platforms leverage ensemble models combining ARIMA, LSTM networks, and gradient boosting to anticipate flash sales and promotional event impacts on infrastructure demands. Streaming services employ predictive scaling to handle viewing pattern variations between weekdays and weekends, with models trained on years of consumption data. Gaming platforms face unique challenges with sudden player influxes following content releases, requiring models that rapidly adapt to new behavioral patterns. Financial services implement predictive scaling for trading systems where microsecond latencies during market opens demand preemptive resource allocation. These implementations reveal common patterns: successful deployments combine multiple prediction techniques, maintain fallback mechanisms for unexpected scenarios, and continuously retrain models as workload characteristics evolve. The integration of business context into technical predictions emerges as a critical success factor across all case studies.

INTELLIGENT MONITORING: ANOMALY DETECTION AND ROOT CAUSE ANALYSIS

AI-Driven Noise Reduction in Monitoring Data Streams

Monitoring systems create enormous volumes of data streams, including meaningless noise as well as significant signals that hide important operational insights. Drawing parallels from acoustic signal processing, where comparable approaches reduce undesired interference (Wang, Z. *et al.*,2022), neural network architectures show an amazing ability to filter noise from complex data streams. Convolutional neural networks are used by infrastructure monitoring systems to find and eliminate benign variances, seasonal changes, and recurring patterns that do not point to genuine concerns. Autoencoders filter variations inside predicted ranges, therefore lowering false positive rates significantly by learning normal operational baselines. The noise reduction process has several steps: initial signal decomposition, frequency domain analysis, pattern classification, and

Publisher: SARC Publisher

selective reconstruction of meaningful elements. Advanced solutions use attention mechanisms that dynamically modify filtering thresholds depending on system urgency and operational context, guaranteeing that small but significant anomalies stay visible while removing overpowering alert storms.

Machine Learning Models for Service Degradation Identification

Service degradation manifests through subtle performance shifts that precede complete failures, requiring sophisticated detection mechanisms beyond simple threshold monitoring. Ensemble learning approaches combine multiple weak classifiers to identify complex degradation patterns across heterogeneous service components. Random forests analyze feature interactions between latency metrics, error rates, and resource utilization to detect performance deterioration before user impact occurs. Support vector machines classify operational states into healthy, degrading, and critical categories based on multidimensional metric spaces. Long short-term memory networks capture temporal dependencies in service behavior, recognizing gradual performance erosion that develops over extended periods. These models distinguish between expected variations due to load changes and genuine degradation requiring intervention. The classification accuracy improves through continuous model retraining on newly labeled incidents, creating feedback loops that enhance detection sensitivity while maintaining acceptable false positive rates.

Table 3: Anomaly Detection Methods in AIOps (Wang, Z. *et al.*,2022; Riedesel, J. 2021)

Detection Method	Data Source	Use Case
Autoencoders	Metrics & telemetry	Baseline deviation detection
Random Forests	Multi-dimensional metrics	Service degradation classification
Support Vector Machines	Performance indicators	State categorization
Neural Networks	Log streams	Noise filtering
Isolation Forests	Time-series data	Outlier identification
Graph Neural Networks	Service dependencies	Root cause propagation

Training Methodologies Using Log Data, Telemetry, and Performance Metrics

Effective model training calls for advanced methods to manage the varied nature of operational data sources (Riedesel, J. 2021). With unstructured text needing natural language processing methods to pull out significant features, log data presents particular problems. Performance measures call for normalization across several scales and units to facilitate meaningful comparisons. Semi-supervised learning techniques make use of the abundance of unlabeled operational data by using small sets of annotated events to direct model building. Transfer learning lets models trained on one system adapt quickly to new environments with little further training data. Active learning methods find the most instructive samples for human annotation, therefore maximizing model improvement while reducing annotation effort. Cross-validation guarantees model generalization across several operational situations, hence preventing overfitting to particular event types or system setups (Riedesel, J. 2021).

Contextual Insight Generation and Actionable Alert Prioritization

Contextualization that considers operational boundaries, system dependencies, and business ramifications is necessary when processing raw anomaly detection to produce actionable intelligence. Graph neural networks provide a mapping of service relationships to propagate anomaly signals through dependency chains and reveal root causes within cascading alerts. Natural language generation techniques are then leveraged to convert a technical anomaly into human-readable explanations along with any necessary historical context and remediation recommendations. Priority scoring algorithms evaluate multiple dimensions, including different populated affected, revenue implications, impact on the SLA, and remediation complexity, to rank alerts. Contextual bandits provide operator responses as feedback to move toward a preferred strategy, dynamically adapting to changes in their environment and changing business priorities. The insight generation also considers externalities, which help the operators develop awareness of their current situation, i.e., scheduled maintenance windows, known issues, and current incident load. Dynamic thresholding also serves an important function in helping reduce alert fatigue during

Publisher: SARC Publisher

periods of expected high activity while still providing the sensitivity required to stabilize critical operations.

CAPACITY PLANNING AND MULTI-CLOUD OPTIMIZATION

AI Integration in Infrastructure Orchestration Layers

Infrastructure orchestration platforms are beginning to add artificial intelligence directly into control planes, shifting resource management from static solutions to adapting systems that react to changing conditions. New orchestration and scheduling architectures include machine learning pipelines at a variety of abstraction layers, ranging from resource schedulers close to the metal to service mesh controllers at the application layer. These machine-intelligent layers can analyze the telemetry data that is exposing the control plane in real time, and make decisions that manage resource allocations, service placements, and even control traffic, quite human-like (Riedesel, J. 2021). Some collaborative scheduling methodologies exemplify how AI-enabled components share decision making and coordination across a set of independent infrastructure components to ensure performance optimization and not just locally optimize (Du, X., & Zhihui, L. 2022). This capability moves impasse decisions beyond static automation scripts toward complex decision engines that can factor multiple constraints at the same time, such as the capabilities of hardware planes, network topologies, and the requirements of the applications that are presented to the infrastructure. The orchestration platform can also leverage and consume predictive models of resource usage to match resource requirements before demand arises. As a result, the orchestration platform may schedule workloads for performance ahead of time while minimizing the overhead associated with moving workloads to a new location.

Dynamic Workload Placement Across Hybrid and Multi-Cloud Environments

In heterogeneous cloud settings, workload distribution poses intricate optimization challenges that are beyond the scope of conventional static strategies. Cloud-edge hybrid architecture demands intelligent scheduling algorithms to manage workloads based on latency requirements, data locality, and process capability on geographically distributed physical resources (Du, X., & Zhihui, L. 2022). Machine learning models can assess the technical characteristics of workloads in order to determine the optimal placement plan, given the characteristics of the workload, such as the amount of computational power that is needed, the amount of data that transmissions need, and the extent of communication between workloads/services. Dynamic placement algorithms are capable of continuously evaluating the performance of current workloads/deployments in light of changing conditions to determine whether a migration is worth the performance gain relative to the associated costs of movement. The content of the placement logic and algorithms will also include information related to the capabilities, pricing models, and service level agreements provided by providers, to inform choices regarding workload placement models. The most advanced implementations also utilize game theory principles and will perform resource allocations among workloads in a way that can negotiate the "best" result for system-wide collaboration of competing workloads while still acknowledging all applications are performing to their expectations.

ML Models for Balancing Cost, Performance, and Availability

Multi-objective optimization in cloud environments involves advanced machine learning models that will drive the trade-offs between conflicting objectives. Particle swarm optimization algorithms were developed specifically for optimizing power grids but can be easily adapted to solve cloud resource allocation problems where a multitude of constraints must be satisfied at the same time (Pei, W. *et al.*,2014). Evolutionary algorithms search solution spaces for the best solution, revealing combinations that maximize economic costs versus performance, or other related measures like availability. Neural networks can learn about the relationship between resource allocation decisions and their cascading effects on application behavior and decision making, making predictive optimization possible through anticipation and reactivity to changes. Reinforcement Learning agents learn by trial and error, where their decision path can lead to certain configurations of resources that lead to impractical solutions, but still perform significantly better than voluntary, low-performance decision making from rules. The models utilize business measurements with technical parameters so that optimization decisions relate back to the organization, rather than simply being efficient in a technical sense (Pei, W. *et al.*,2014).

Publisher: SARC Publisher

Table 4: Multi-Cloud Optimization Strategies (Du, X., & Zhihui, L. 2022; Pei, W. *et al.*,2014)

Optimization Factor	Consideration	ML Approach
Cost Management	Spot instances, regional pricing	Particle swarm optimization
Performance	Latency, throughput requirements	Multi-objective neural networks
Availability	Redundancy, failover capabilities	Game theory models
Compliance	Data residency, regulatory constraints	Constraint-based algorithms
Workload Placement	Hardware affinity, provider features	Collaborative scheduling
Resource Utilization	Capacity efficiency	Evolutionary algorithms

Regional and Provider-Specific Optimization Strategies

The geographic distribution of cloud resources entails region-specific factors that are relevant to optimization. Regulatory compliance requirements oblige adherence to data residency limitations with respect to where workloads can be placed in a compliance context and/or a workload placement context across certain jurisdictions. Variations in regional network connectivity (i.e., distance with respect to latency in latency-sensitive applications, etc.) necessitate models that understand not only physical distance, but also network topology that dictates placement decisions. Provider-specific features, such as particular hardware accelerators, proprietary services, and pricing structures, all create optimization opportunities for workloads that can take advantage of them. Machine learning models extract provider affinity scores based on historical performance data for each workload type to determine which cloud platform excels the most for each. Cost optimization strategies include spot instance availability, reserved capacity discounts, and discounts on regionally priced capacity to try to keep costs lower while still maintaining an appropriate level of service. The optimization framework maintains provider-agnostic abstractions, yet takes advantage of platform-specific features, ensuring workloads remain portable while not encumbering the original conceptual purpose of maximizing workload performance post-cloud distribution and/or transformation.

OPEN-SOURCE ECOSYSTEM AND INDUSTRY IMPLEMENTATIONS

Overview of Leading AIOps Platforms

Modern AIOps (AI Operations) platforms, designed around context-specific use cases, are different models and implementations to support specific operational context use cases. Kubeflow supports scalable AI workloads as a fully functional machine learning operations platform, providing the ability to coordinate distributed training workflows across Kubernetes clusters. The complexities of infrastructure are abstracted through declarative pipelines, ranging from data processing to model and ultimately deployment, while being agnostic of deep Kubernetes experience. The OpenAI Triton Inference Server enables model serving infrastructure while supporting multiple deep learning frameworks simultaneously, as well as providing resource-efficient shared hardware resources on which dissimilar models can execute concurrently. DeepMind demonstrated distributed reinforcement learning scalability in cluster architectures, along with the specialized configuration and hardware advances that could accelerate executions of AI workloads, demonstrating organizational commitment. These operational platforms share many key design concepts, such as containerization for portability, declarative configuration management, and monitoring that provides visibility into AI pipeline performance. Each operational platform has its own area of focus regarding AI infrastructure lifecycle management, offering platform functionality in the areas of training, experimenting, production and continuous performance improvement, and IT management process improvement.

Comparative Analysis of Open-Source vs. Proprietary Solutions

Organizations that want to select AIOps platforms need to weigh the advantages of open source flexibility against the convenience of commercial solutions. There is no other technology that gives flexibility and transparency to open-source technologies. There is community-based development across the globe, improving features and fixing issues at an unprecedented pace, and not through the standard vendor timeframes. Also, organizations could modify the code directly and add custom functionality. A successful implementation of open source will require some technical resources to make modifications or build features into the software. Long-term financial commitment to maintain the software will also be required. Commercial AIOps products streamline deployment through integrated features and professional support services, particularly

Publisher: SARC Publisher

benefiting teams with limited technical depth. Integration presents another consideration: proprietary systems typically connect more readily with existing infrastructure but may create vendor dependencies, whereas open source alternatives preserve architectural freedom despite requiring additional integration effort. Security models differ in that commercial vendors have in-house compliance teams that focus on specific people working together, while open source relies on very detailed code quality audits. Either way, the approach an organization takes will ultimately be determined by technical mastery, budget, risk appetite, and strategic goals - neither model is inherently better than the other.

Industry trends and adoption-related concerns

Real-world AIOps installations revealed systematic patterns across industries, despite differences with respect to operational context and maturity levels (Dang, Y. *et al.*,2019). Financial services organizations lead the way in adoption levels, clearly influenced by vendor uptime requirements and large capital investments in technology and infrastructure. Depending on the level of assurance required, organizations cannot be satisfied with simpler management approaches. Technology firms are leveraging their internal capabilities to fully adopt open-source platforms and create a tailored experience. Ultimately, traditional enterprises are aware of the formal change management processes that must accompany implementation, which include realistic helpdesk adoption within organizations moving from manual management and operations to automated AIOps processes, which is crucial when going through large-scale internal changes. Separate from this, skills/role gaps continue to be barriers to adoption, where organizations are having challenges recruiting professionals who can combine infrastructure knowledge with machine learning skills (Dang, Y. *et al.*,2019). Data quality issues were a consistent problem that impacted all the initial deployments, because most operational data is historical and does not have the requisite level of consistency and completeness for model development and training. Integrating the AIOps platform with existing legacy monitoring tools, configuration management tools, and ticketing tools is much more complex than expected. Reputable AIOps implementation consultants advocate for a phased approach to implementation where the entire environment is eventually integrated, but if the initial work can be isolated, the teams can expand scope based on their confidence and experience.

Future Directions in AI-Powered Infrastructure Tools

Innovations that represent sea changes in the design of infrastructure tools driven by AI are beginning to emerge to meet future operational needs. For example, federated learning architectures will provide organizations with the ability to train models across distributed infrastructure while keeping highly sensitive operational data decentralized, eliminating privacy issues associated with blending private information using multi-tenancy. As these federated models become increasingly widespread, organizations will feel the urgency to incorporate more explainable AI techniques in the decision-making process because users of the technology will become more demanding about the transparency of decisions in operating essential infrastructure. Edge-native Application Network Operations (AIOps) platforms are emerging to operate across platforms, including cloud, centralized data centers, and distributed edge resources. Quantum computing systems are on the distant horizon and will provide unlimited computing speed and, therefore, unlimited speed in optimizing resource allocation for optimization problems. The use of natural language interfaces embedded in AIOps platforms, which allow operational workers to use conversational interfaces to manage complex infrastructure, means operational technology is becoming less of a technical fitness hurdle. Finally, the advancements of AI and ML will eventually lead to self-improving systems that can optimize or adapt their algorithms automatically based on operational results. This may be considered the last step toward completely autonomous infrastructure. As industry organizations start looking for standard AIOps interoperability frameworks for investment work and to avoid the siloing that harmed the previous generation of infrastructure management technologies, the standardization effort will pick up speed.

CONCLUSION

The advent of artificial intelligence in infrastructure management is a complete transition from maintenance to optimization through proactive means and will shift how organizations manage complex digital infrastructures. Machine learning algorithms will enable predictive autoscaling (before workload demand), smarter

Publisher: SARC Publisher

monitoring systems that remove noise but detect true anomalies, and advanced capacity planning across different clouds. The saddle between open-source and proprietary platforms allows everyone access to functionality, but there are hurdles to implementation regarding data quality, availability of skills, and organizational readiness. With the advancement of AIOps technologies, we see a move toward federated learning capabilities, explainable AI, and edge-native architectures that will give infrastructure even greater autonomy. Organizations will benefit from lower operational costs, better/higher reliability of services, and the ability to scale when adopting AI-enhanced tools. The shift to self-governing infrastructure is increasing as complexities continue to rise in systems and the acknowledgment of behavior (and results) executed by ML. Future infrastructure management techniques will have conversational user interfaces, independently self-improving algorithms, and improvements as a result of quantum computing, concluding a transition from systems governed by humans to self-learning intelligent platforms that operate continuously to be fit for purpose while maintaining a clear tie to organizational outcomes.

REFERENCES

1. Appio, F. P., La Torre, D., Masri, H., & Jayaraman, R. “AI-powered paradigms in RD&E management: Opportunities and challenges. *IEEE Transactions on Engineering Management*”. (2023) https://www.ieee-tems.org/special-issue-ai-powered-paradigms-in-rde-management-opportunities-and-challenges/
2. Gula, A., Ellis, C., Bhattacharya, S., & Fiondella, L. "Software and system reliability engineering for autonomous systems incorporating machine learning." (2020) *Annual Reliability and Maintainability Symposium (RAMS)*. IEEE, (2020).
3. Cheng, W., Xu, Y., Xu, Q., Zhang, H., Li, X., & Shao, X. "CSFRL: A Reinforcement Learning Technology Enabled Computing Power Scheduling Framework Based on Kubernetes." *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, (2023).
4. Suksriupatham, N., & Hoonlor, A. "Workload prediction with regression for over and under provisioning problems in multi-agent dynamic resource provisioning framework." *2020 17th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, (2020).
5. Wang, Z., Na, Y., Tian, B., & Fu, Q. "NN3A: Neural network supported acoustic echo cancellation, noise suppression and automatic gain control for real-time communications." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, (2022).
6. Riedesel, J. “Software Telemetry: Reliable logging and monitoring”. *Simon and Schuster*, (2021)
7. Du, X., & Zhihui, L. “Collaborative Scheduling of Workload Tasks in a Cloud-Edge Hybrid Architecture." (2022).
8. Pei, W., Dongmei, Z., & Xu, Z. "Reactive power optimization in regional grids by improved PSO algorithm." *2014 Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control*. IEEE, (2014).
9. Wong, W. K., Lin, W. C., & Chen, Y. B. "Research on Kubeflow Distributed Machine Learning." *2022 IET International Conference on Engineering Technologies and Applications (IET-ICETA)*. IEEE, (2022).
10. Dang, Y., Lin, Q., & Huang, P. "Aiops: real-world challenges and research innovations." *2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*. IEEE, (2019).

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:

Baer, N. “AI-Powered Infrastructure Management: From Reactive Operations to Predictive Intelligence in Cloud Computing” *Sarcouncil Journal of Engineering and Computer Sciences* 4.7 (2025): pp 515-522.

Publisher: SARC Publisher