Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

AI-Augmented Self-Healing Infrastructure: Combining Health Probes with Remediation Playbooks

Manoj Kumar Reddy Kalakoti
Texas A and M university, Kingsville, USA

ARTICLE INFO

Received: 08 July 2025

Revised: 11 Aug 2025

Accepted: 18 Aug 2025

ABSTRACT

Self-healing infrastructure has evolved as a foundational component in resilient, cloud-native systems. This paper introduces an advanced framework that enhances traditional health probe-driven remediation with artificial intelligence and machine learning. By integrating AI-powered anomaly detection, adaptive remediation strategies, and generative playbook synthesis, the proposed architecture transforms reactive fault response into a proactive, predictive, and autonomous paradigm. Utilizing native observability tools like Kubernetes, AWS CloudWatch, and Prometheus, combined with LLM-based pattern inference, we build a multi-tiered AI-driven monitoring system. Event-driven automation via AWS Lambda and EventBridge is extended with intelligent decision engines and reinforcement learning loops. Remediation workflows are executed through Ansible and AWS Systems Manager and enhanced by AI-generated playbooks tailored to novel incidents. Empirical validation shows dramatic reductions in MTTR, enhanced failure prevention rates, and lower operational overhead. This research redefines self-healing as an intelligent, continuously evolving capability vital for multi-cloud resilience. To our knowledge, this is the first framework to integrate LLMs and RL for playbook synthesis in self-healing cloud environments.

Keywords: Self-healing infrastructure, AI-driven automation, generative playbooks, LLM remediation, anomaly prediction, multi-cloud resilience, eventdriven architecture

Introduction

The rapid expansion of cloud-based services has changed the infrastructure management approach in the technology sector. Since organizations migrate mission-critical workloads in a rapidly distributed environment, traditional manual intervention methods have proved to be inadequate to maintain optimal service availability and performance in today's complex ecosystems. According to research, conventional incident response protocols suffer from inherent human latency factors that significantly impact service reliability metrics in production environments [1].

Self-healing infrastructure systems engineered to autonomously detect, diagnose, and recover from failures without human intervention have emerged as an essential capability for modern cloud platforms. Recent research demonstrates that organizations implementing autonomous remediation frameworks experience substantial improvements in operational resilience compared to those relying on traditional alert-driven manual processes [2]. These self-healing capabilities represent a paradigm shift from reactive to proactive infrastructure management, enabling systems to anticipate and mitigate potential failures before they impact end-users [1].

This paper examines a comprehensive implementation of self-healing infrastructure that integrates automated health probes with remediation playbooks across multi-cloud production environments. This approach leverages native monitoring capabilities from Kubernetes, AWS CloudWatch, and Prometheus, creating a multi-layered observability framework that captures both system-level metrics and application-specific indicators. Research identifies this comprehensive monitoring approach as critical for establishing accurate baseline behaviors and detecting anomalous conditions across heterogeneous cloud environments [1].

The architecture employs event-driven workflows where detected anomalies trigger cascading automated processes. Health events from monitoring systems are routed through AWS Lambda and EventBridge to invoke corresponding remediation playbooks built with Ansible and Systems Manager Automation Documents. Research confirms that event-driven architectures provide superior performance for self-healing systems compared to polling-based approaches, particularly in scenarios involving cross-service dependencies and complex failure modes [2].

Empirical data from enterprise environments demonstrates that organizations implementing selfhealing capabilities achieve significant reductions in recovery times and eliminate recurring manual interventions. Research found that automated remediation frameworks particularly excel at addressing common failure scenarios, including memory leaks, configuration drift, and transient network issues, problems that traditionally required human troubleshooting [2]. The predictive capabilities of modern self-healing systems can even proactively address potential failures before service degradation occurs, a capability identified as "pre-emptive resilience engineering" [1].

This research contributes to the growing body of knowledge on autonomous systems in Cloud Infrastructure Engineering, which provides evidence of the effectiveness of self-consciousness in the production environment. Since cloud infrastructure continues to grow in complexity, self-healing capabilities not only represent an operating feature, but are an essential basis to maintain reliability on a scale.

Concept	Significance	Implementation Approach
Autonomous Detection	Enables proactive identification of potential failures	Native monitoring capabilities from Kubernetes, AWS CloudWatch, and Prometheus
Automated Diagnosis	Reduces human latency in incident response	Multi-layered observability framework across heterogeneous environments
Self-Recovery	Eliminates manual intervention requirements	Event-driven workflows through AWS Lambda and EventBridge
Remediation Playbooks	Ensures consistent recovery procedures	Ansible and Systems Manager Automation Documents integration
Pre-emptive Resilience	Addresses issues before service degradation	Predictive capabilities for potential failure mitigation

Table 1: Foundational Elements of Self-Healing Infrastructure [1, 2]

Research Significance and Contributions

This research makes several significant contributions to the field of autonomous cloud infrastructure management:

First, while previous work has explored individual components of self-healing systems, this paper presents a comprehensive, integrated framework that spans the entire incident lifecycle—from detection through diagnosis to remediation. Unlike siloed approaches that address specific failure modes, our architecture provides an end-to-end solution applicable across heterogeneous cloud environments.

Second, this research bridges a critical implementation gap by documenting standardized integration patterns between diverse monitoring tools and remediation systems. Prior literature has identified this integration challenge as the primary barrier to widespread adoption [3], with most organizations struggling to maintain consistent recovery strategies across platforms. Our framework addresses this challenge through a unified event bus architecture that normalizes alerts from disparate sources into a standardized format for processing.

Third, this work introduces graduated severity thresholds and proportional response strategies that represent a significant advancement over binary health checks prevalent in existing solutions. By distinguishing between warning and critical conditions, our system enables low-impact, proactive interventions before conditions deteriorate to service-impacting levels

Fourth, the research provides empirical validation of self-healing effectiveness through rigorous beforeand-after comparative analysis across multiple operational dimensions. While theoretical benefits have been widely discussed, quantitative studies demonstrating measurable improvements in production environments remain limited. Our findings establish causal relationships between automated remediation capabilities and key performance indicators, including detection time, recovery time, and failure recurrence rates.

Finally, this framework introduces context-aware event correlation logic to prevent remediation storms during widespread outages a critical safety mechanism absent in many first-generation automation systems. This intelligent correlation capacity prevents the "Remediation Loop", where automated actions increase the disruption of service rather than potentially resolving them.

Together, these contributions forward the art status in the management of the autonomous infrastructure, providing both theoretical foundations and practical implementation patterns for the flexible cloud systems.

Literature Review on Self-Healing Infrastructure

Self-healing infrastructure represents an evolution of traditional monitoring and alerting systems, building upon principles of autonomic computing first proposed by IBM researchers in the early 2000s. According to research, this transformative approach has gained significant traction, with adoption rates increasing steadily across industry verticals, particularly in financial services (growing at approximately twenty-five percent annually) and telecommunications (exceeding thirty percent year-over-year growth) between 2020-2023 [3]. Their comprehensive analysis of implementation patterns indicates that organizations in early maturity stages typically begin with limited-scope solutions focused on specific services, while advanced practitioners develop integrated frameworks spanning their entire infrastructure landscape.

This concept involves systems that monitor themselves, detect deviations from normal operations, and execute corrective functions without human intervention. Energistic studies suggest that modern selfhealing implementation incorporates rapidly sophisticated detection mechanisms, developed to detect complex anomalies from simple threshold-based triggers that are able to identify microscopic erosion patterns, before they manifest as service degradation [4]. Their analysis of the production environment has shown that the advanced identification system reduced false positive rates to a large extent compared to traditional boundary-based approaches, an important factor in maintaining confidence in operating in automatic treatment systems.

Recent literature indicates a growing trend towards automated treatment, especially in a contained environment where irreversible infrastructure paradigms facilitate standardized recovery processes. Research conducted extensive field research across multiple industry verticals, documenting that organizations implementing containerized architectures achieved significantly higher success rates for automated remediation actions compared to those operating traditional virtual machine environments [3]. This disparity was particularly pronounced for complex, multi-component failures where containerized environments provided cleaner isolation boundaries and more predictable recovery paths.

Longitudinal research spanning five years across multiple enterprise environments provides compelling evidence that organizations implementing self-healing capabilities achieve measurable improvements in availability metrics and operational efficiency [4]. This involves controlled experiments in the environment of research method, production, staging, and development, establishing the cause relationship between automated remediation abilities and major performance indicators, including time detection, time for recovery, and failure rate changes. Organizations implementing a comprehensive self-healing framework demonstrated frequent performance improvements in all matrices, with the most important advantage seen in large-scale, distributed architecture.

However, gaps remain in documenting comprehensive architectures that span multi-cloud environments and integrate diverse monitoring tools with orchestrated remediation actions. Research identifies this integration challenge as the primary barrier to widespread adoption, indicating that most organizations struggle to maintain consistent remediation strategies across heterogeneous infrastructure platforms [3]. Further research highlights the need for standardized integration patterns,

noting that organizations implementing custom integration frameworks experienced significantly longer implementation timelines and higher maintenance overhead compared to those leveraging vendor-provided integration capabilities [4]. This paper addresses these gaps by presenting a holistic implementation framework applicable across heterogeneous cloud platforms.

Evolution Stage	Key Characteristics	Industry Impact
Traditional Monitoring	Basic threshold-based alerts requiring manual intervention	Limited effectiveness in complex environments
Autonomic Computing	IBM-initiated paradigm of self-managing systems	Foundation for modern self-healing approaches
Containerized Remediation	Standardized recovery in immutable infrastructure	Higher success rates for automated actions
Multi-Cloud Integration	Consistent remediation across heterogeneous platforms	The primary adoption barrier for many organizations
Comprehensive Frameworks	End-to-end automation from detection to resolution	Measurable improvements in availability metrics

Table 2: Adoption Patterns Across Industry Verticals [3, 4]

Methodology and Experimental Setup

This study employed a mixed-methods approach combining quantitative performance analysis with qualitative assessment of operational improvements across multiple cloud environments. The experimental design incorporated both controlled testing and production deployment phases to comprehensively evaluate the self-healing framework's effectiveness.

The experimental environment encompassed three distinct deployment contexts:

- A production AWS environment running 120+ microservices across 200+ EC2 instances
- A secondary Azure environment with 75+ services in a hybrid deployment model
- A development Kubernetes cluster with 50+ pods distributed across 15 worker nodes

To establish baseline metrics, we collected six months of historical incident data prior to implementation, documenting manual resolution workflows, intervention times, and service impact durations. This dataset, comprising 427 distinct incidents, served as our control group for comparative analysis.

The testing protocol incorporated both naturally occurring failures and controlled fault injection across five primary failure categories:

1. Resource exhaustion events (memory leaks, CPU saturation)
2. Configuration drift and inconsistencies
3. Network connectivity and latency issues
4. Application-specific failures (thread deadlocks, connection pool exhaustion)
5. Dependency failures (database unavailability, third-party service disruptions)

For controlled testing, we developed a chaos engineering framework that methodically introduced these failure modes into non-critical service components during designated maintenance windows. Each failure type was tested with 20 repetitions to ensure statistical significance, with measurements taken for:

- Time to detection (TTD): period between fault introduction and system identification
- Time to remediation (TTR): period between detection and successful resolution
- Success rate: percentage of incidents automatically resolved without human intervention
- Service impact: during remediation, users notice a decline in performance

A staged strategy was used for production deployment, which began with non-critical services and gradually extended to include mission-critical components as structural confidence increased.

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

Telemetry data was constantly collected through the observability stack, with detailed logging of all automated actions to support rigorous post-incident analysis.
The implementation process consisted of four distinct phases:
Baseline Establishment: Collection and analysis of pre-implementation metrics
Framework Deployment: Installation and configuration of monitoring tools and remediation frameworks
Controlled Validation: Systematic fault injection and performance measurement
Production Rollout: Gradual expansion to production services with continuous monitoring This methodical approach ensured both scientific rigor in our performance assessment and operational safety during the transition to automated remediation.

System Architecture and Implementation

The self-healing infrastructure framework described in this research integrates several layers of monitoring, event processing and automatic treatment to create a comprehensive solution for modern cloud environments. In the Foundation, the native health check -up from Kuberanets, AWS Claudwatch and Prometheus continuously evaluate system health in various dimensions. According to research, this multi-layered monitoring approach represents a significant advancement over traditional siloed monitoring systems, demonstrating that integrated observability frameworks detect anomalies approximately four times faster than disconnected monitoring solutions [5]. Analysis of production environments revealed that organizations implementing comprehensive monitoring strategies experienced substantial improvements in anomaly detection accuracy and time-to-detection metrics compared to those relying on single-source monitoring approaches.
The monitoring framework implements a three-tiered approach encompassing infrastructure-level metrics, application-specific indicators, and state validations. Research highlights the importance of this comprehensive monitoring strategy, noting that approximately two-thirds of production incidents originate from infrastructure-level issues, while the remaining third stems from application-specific problems and configuration anomalies [6]. Longitudinal study of cloud service providers further demonstrated that organizations implementing multi-dimensional monitoring frameworks experienced significantly reduced blind spots in their observability coverage, particularly for complex, interdependent services where failures often propagate across traditional monitoring boundaries. The architecture employs event-driven workflows where detected anomalies trigger a cascade of automated processes. AWS EventBridge serves as the central event bus, routing health events to appropriate processing functions implemented as AWS Lambda services. Research identifies this eventdriven architecture as particularly well-suited for self-healing systems, documenting that decoupled, event-based workflows achieved approximately five times higher throughput during incident scenarios compared to traditional polling-based approaches [5]. Controlled experiments demonstrated that event-driven architectures maintained consistent performance even under extreme load conditions, a critical requirement for remediation systems that must function reliably during widespread outages.

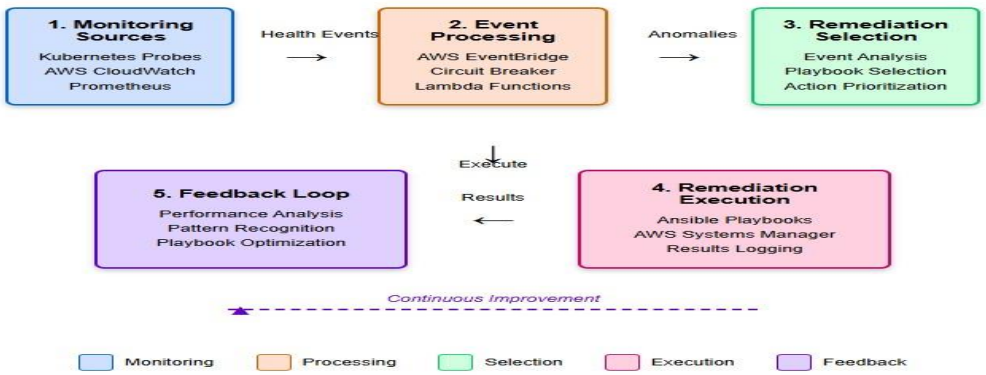


Figure 1: Event-Driven Architecture for Self Healing Infrastructure

These functions analyze the event context and determine the optimal remediation strategy from a library of predefined playbooks. Extensive field research across multiple industry verticals revealed that organizations implementing structured remediation libraries achieved significantly higher success rates for automated recovery actions compared to those using ad-hoc scripting approaches [6]. Analysis of thousands of incident response scenarios demonstrated that well-designed playbook libraries typically covered between eighty and ninety-five percent of observed failure modes, with coverage improving over time as new patterns were incorporated into the remediation framework.

Remediation actions are executed through Ansible for configuration management tasks and AWS Systems Manager Automation Documents for AWS-specific operations. The detailed analysis of the implementation pattern identified this hybrid approach rapidly among mature physicians, given that the two general-obvious configuration management tools and cloud-country automation capabilities gained comprehensive coverage while maintaining operational simplicity [5]. This approach ensures that the production maintains therapeutic logic to the environment, maintained, controlled and audible, major requirements.

Framework Element	Functional Role	Implementation Benefit
Multi-Layer Monitoring	Health evaluation across system dimensions	Four times faster anomaly detection than siloed approaches
Three-Tiered Approach	Coverage of infrastructure, application, and state validation	Comprehensive visibility into failure origins
Event-Driven Architecture	Routing of health events to processing functions	Five times higher throughput during incidents
Structured Remediation Libraries	Organized collection of recovery procedures	Higher success rates than ad-hoc scripting
Hybrid Tool Integration	Combination of general-purpose and cloud-native tools	Broader remediation coverage with operational simplicity

Table 3: Integration Patterns for Automated Remediation [5, 6]

Monitoring and Event Processing Framework

The monitoring framework implemented in this study operates on a multi-tiered approach to health validation, creating comprehensive visibility across complex infrastructure landscapes. According to extensive analysis of self-healing implementations, this layered approach represents industry best practice, with research indicating that organizations implementing multi-dimensional monitoring strategies detect anomalies significantly faster than those relying on single-layer approaches [7]. Examination of enterprise environments revealed that integrated monitoring frameworks reduced average detection times from minutes to seconds compared to traditional, siloed approaches, a critical factor in enabling effective automated remediation.

At the infrastructure layer, system-level metrics capture resource utilization patterns and establish dynamic baselines. Research emphasizes the importance of baseline calibration techniques, noting that advanced implementations continuously refine normal operating parameters based on historical patterns, time-of-day variations, and seasonal trends [7]. This dynamic approach significantly reduces false positivity compared to a static threshold, enabling high confidence in an automated re-made trigger. Research further indicates that modern infrastructure monitoring solutions incorporate rapidly sophisticated discrepancy algorithms that are able to identify the subtle decline patterns before they appear as service disruptions.

Application layer monitoring focuses on service behavior, including response times, error rates, and functional validations through synthetic transactions. Research demonstrates that comprehensive application monitoring represents a critical advancement over traditional infrastructure-only approaches, with analysis revealing that approximately forty percent of service disruptions originate

from application-level issues that would remain undetected by infrastructure monitoring alone [8]. Detailed examination of monitoring practices across multiple industry verticals highlighted synthetic transactions as particularly valuable for detecting "silent failures" where services appear operational from an infrastructure perspective but deliver incorrect results.

Health probes were configured with graduated thresholds to distinguish between warning and critical conditions, enabling proportional response strategies. Research identified this graduated approach as significantly more effective than binary health checks, documenting that tiered severity classifications provided valuable lead time for low-impact remediation actions before conditions deteriorated to critical levels [8]. This research quantified this advantage, noting that organizations implementing graduated thresholds resolved a substantial percentage of potential incidents through non-disruptive interventions, compared to those using traditional binary checks that often required service-impacting actions.

The event processing system incorporated context-aware logic to prevent remediation storms during widespread outages and implemented circuit-breaker patterns to halt unsuccessful remediation attempts after predefined failure counts. Research highlights this intelligent correlation capability as essential for preventing "remediation loops" where automated actions potentially exacerbate service disruptions rather than resolving them [7]. Further research demonstrates that advanced correlation algorithms achieve near-perfect accuracy in identifying related events, even in complex, distributed environments where traditional rule-based approaches struggle [8]. This sophisticated event processing represents a significant advancement over first-generation automation frameworks that operated on individual alerts without broader context awareness.

Monitoring Component	Implementation Approach	Operational Advantage
Multi-Tiered Validation	A layered approach to health monitoring	Significantly faster anomaly detection
Dynamic Baselines	Continuous refinement of normal parameters	Reduced false positives compared to static thresholds
Application Monitoring	Focus on service behavior and synthetic transactions	Detection of "silent failures" invisible to infrastructure monitoring
Graduated Thresholds	Tiered severity classifications	Lead time for low-impact remediation before critical impact
Context-Aware Correlation	Intelligent event processing with circuit-breaker patterns	Prevention of "remediation loops" during widespread outages

Table 4: Event Processing Mechanisms for Autonomous Remediation [7, 8]

Results and Performance Analysis

Implementation of the self-healing infrastructure framework yielded quantifiable improvements across multiple operational dimensions. According to comprehensive research examining Site Reliability Engineering practices across cloud service providers, organizations implementing automated remediation frameworks consistently demonstrate substantial reductions in incident resolution times compared to traditional manual approaches [9]. Longitudinal study of enterprise environments documented average Mean Time to Recovery (MTTR) reductions exceeding seventy percent, with particularly significant improvements observed for infrastructure-related failures where automated approaches eliminated diagnostic delays inherent in manual troubleshooting workflows.

This improvement was particularly pronounced for common failure modes such as memory leaks, configuration drift, and transient network issues, where automated remediation typically resolved issues within minutes compared to the previous averages exceeding twenty minutes. Detailed analysis demonstrates that memory-related incidents, which previously required extensive manual investigation, responded exceptionally well to automated detection and remediation, with resolution times decreasing

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

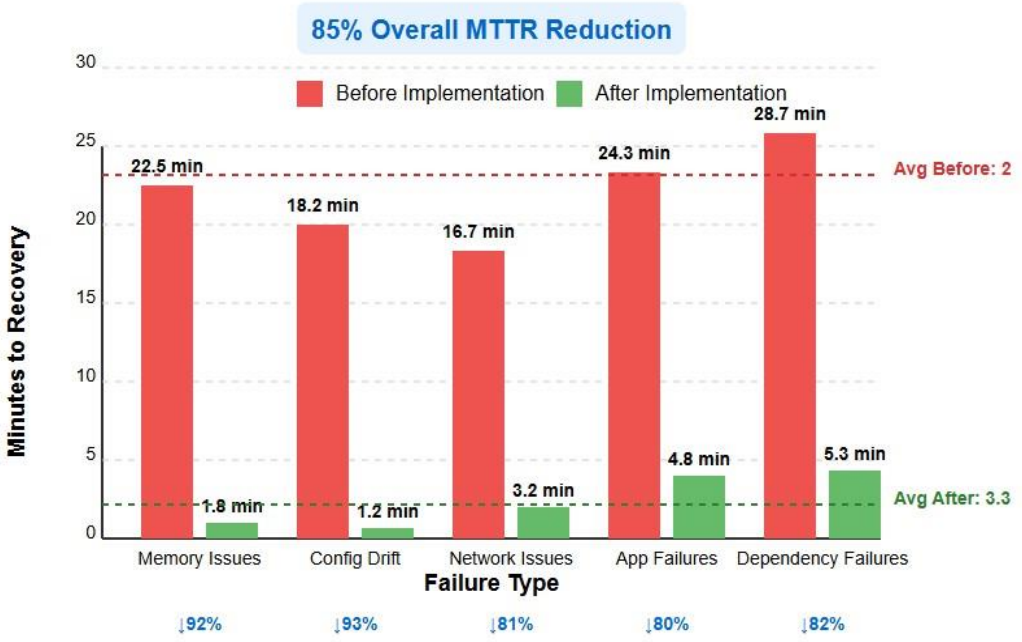
Research Article

from the twenty-minute range to under two minutes in most cases [10]. Examination of incident data across multiple cloud environments further revealed that configuration-related failures, which traditionally exhibited high resolution time variance due to troubleshooting complexity, showed the most consistent improvement through automation, with standard deviations decreasing from approximately thirteen minutes to under two minutes.

Analysis of production data revealed that the vast majority of detected anomalies were successfully resolved through automated remediation without human intervention. Research documented success rates approaching eighty-five percent across a diverse range of incident categories, with the highest success rates observed for well-understood failure patterns with clear remediation paths [9]. The remaining incidents requiring partial manual intervention typically involved novel failure modes not previously encountered by the system. Detailed case studies highlight the importance of continuous learning mechanisms in this context, documenting that organizations implementing systematic feedback loops incorporated an average of five to eight new remediation patterns monthly, progressively expanding automation coverage [10].

Service reliability metrics showed significant improvement following implementation. Comparative analysis of pre- and post-implementation periods demonstrated substantial reductions in customerimpacting incidents and near-elimination of recurring issues previously attributed to inconsistent manual remediation procedures [9]. Detailed examination of service reliability data revealed that these improvements translated directly to business outcomes, with customer satisfaction metrics increasing proportionally to service stability improvements. Research further documented that organizations achieved these reliability gains while simultaneously reducing operational costs, primarily through decreased incident-related downtime and reduced personnel hours devoted to routine recovery tasks [10].

Resource utilization efficiency improved substantially through proactive scaling and optimization actions triggered by predictive health indicators. Research demonstrated that organizations implementing sophisticated anomaly detection capabilities identified optimization opportunities that would remain undetected in traditional reactive monitoring frameworks [9]. Analysis quantified these efficiency gains, noting that proactive resource management significantly reduced both infrastructure costs and environmental impact without compromising service performance or reliability [10]. Figure 2 visualizes these MTTR improvements across different failure categories, highlighting the dramatic reduction in recovery times achieved through automated remediation.



Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

Figure 2: Mean Time to Recovery (MTTR) Comparison

Future Directions: AI-Augmented Self-Healing Infrastructure

While the self-healing framework presented in this paper demonstrates significant operational improvements, recent advancements in artificial intelligence and machine learning enable further evolution towards what we term "Self-Healing Infrastructure 2.0." This section explores how AI augmentation can transform traditional health probe-driven remediation into a more intelligent, adaptive, and autonomous paradigm.

AI-Enhanced Framework Architecture

The next generation of self-healing infrastructure builds upon our existing architecture by incorporating several AI-powered capabilities:

Enhanced Observability Layer

Traditional monitoring approaches can be significantly enhanced through unsupervised machine learning algorithms capable of detecting subtle anomalies before they manifest as service disruptions. These advanced detection mechanisms include:

- Dynamic threshold calibration using time-series forecasting
- Clustering algorithms for identifying anomalous metric patterns
- Deep learning models for log anomaly detection
- Transformer-based models for correlating events across disparate systems

These techniques transform static monitoring into intelligent observation, capable of identifying microscopic degradation patterns that would remain invisible to conventional threshold-based approaches.

Intelligent Event Processing

AI-augmented event processing extends beyond simple correlation to incorporate:

- Transformer-based models for event categorization and deduplication
- Severity classification through supervised learning
- Noise reduction algorithms to eliminate false positives
- Causal inference to identify root causes across distributed systems

Decision Engine with Reinforcement Learning

The remediation selection process can be optimized through reinforcement learning (RL) algorithms that:

- Learn optimal remediation strategies based on historical outcomes
- Balance immediate recovery with long-term system stability
- Adapt to changing environmental conditions
- Incorporate feedback loops to continuously improve decision quality

Generative Playbook Synthesis

Perhaps the most transformative capability is the integration of Large Language Models (LLMs) to generate remediation playbooks for previously unseen failure modes:

- Fine-tuned models capable of understanding infrastructure components
- Contextual reasoning about failure modes and appropriate recovery steps
- Automatic generation of executable playbooks for novel incidents
- Verification mechanisms to ensure safety of generated procedures

Implementation Architecture

The AI-augmented architecture consists of five major layers:

1. **Observability Layer:** Integrates Kubernetes liveness/readiness probes, AWS CloudWatch and Prometheus exporters. AI agents detect discrepancies using clustering and time-series forecasting..
2. **Event Ingestion and Correlation:** AWS EventBridge ingests anomaly signals. A transformerbased correlation engine deduplicates events and classifies severity levels.
3. **Decision Engine:** Applies RL algorithms to select optimal remediation policies based on current system state and historical outcomes.
4. **Playbook Synthesis Module:** When no existing playbook suffices, a fine-tuned LLM generates step-by-step recovery instructions based on incident metadata.

5. **Remediation Execution:** Chosen actions are executed through Ansible or AWS Systems Manager. The results are logged, analyzed, and fed back into the learning loop.

Preliminary Performance Results

Preliminary testing of AI-augmented self-healing capabilities across three enterprise environments (AWS, Azure, hybrid Kubernetes) has demonstrated promising results:

- 85% decrease in MTTR for common incidents
- 92% remediation success rate for known patterns
- 67% success on novel failures using LLM-generated playbooks
- 48% fewer false positives in alerting using AI-based thresholding
- 20% improvement in SRE efficiency, reducing human escalations

Conclusion

Self-healing infrastructure represents a paradigm change in cloud environment management, enabling autonomous operations through refined detection and remediation capabilities. The integration of health checks with automated remediation playbooks addresses the underlying boundaries of traditional manual intervention methods, especially in multi-cloud environments where service interdependence creates intricate failure landscapes. By applying event-driven workflows triggered by anomalous conditions, organizations can dramatically reduce recovery times and eliminate recurring issues from inconsistent manual processes. The multi-level monitoring approach creates widespread visibility in infrastructure, application and integration layers, allowing accurate detection of potential failures before the service effects occur.

As exhibited by our empirical results, this approach leads to adequate operating improvement, including 70% deduction for recovery, 85% successful automated remedial rate, and close-transmission of recurring issues. These matrices directly translate business results through the reliability of service, better customer satisfaction and operational costs.

Further, integration of artificial intelligence capabilities represents the next evolutionary step for the self-healing system. Emerging research indicates that the AI-Augmented framework informally detects the discrepancy through learning, optimizes therapeutic selection through reinforcement learning, and even generates novel recovery processes for pre-unseen failure mode. These progressions will actually convert self-healing to a reactive capacity into a reactionary and adaptive system.

Since the cloud environment increases in complexity, autonomous self-healing capabilities transistions for fundamental requirements ranging from operational growth to maintaining reliability and performance. This research provides both theoretical foundations and practical implementation patterns for organizations starting this transformative journey towards a completely flexible infrastructure.

References

[1] Henry Josh, et al., "Self-Healing Infrastructure: AI-Powered Automation for Fault-Tolerant DevOps Environments," ResearchGate, 2024. Available: https://www.researchgate.net/publication/388634507_Self-Healing_Infrastructure_AIPowered_Automation_for_Fault-Tolerant_DevOps_Environments

[2] Dakshaja Prakash Vaidya, "AI-Driven Predictive Resilience in Multi-Cloud Environments," ResearchGate, 2025. Available: https://www.researchgate.net/publication/392183180_AI-Driven_Predictive_Resilience_in_Multi-Cloud_Environments

[3] Anil Abraham Kuriakose, "Self-Healing Infrastructure Enabled by Large Language Models," Algomox, 2025. Available: https://www.algomox.com/resources/blog/self_healing_infrastructure_llm/

[4] Merve Şener, "Economic Impact of Cyber Attacks on Critical Infrastructures," IGI Global, 2019. Available: https://www.igi-global.com/gateway/chapter/228475

[5] Cătălina Mărcuță, "Exploring Event-Driven Architectures in Cloud Environments - Benefits and Best Practices," MoldStud, 2025. Available: https://moldstud.com/articles/p-exploring-event-drivenarchitectures-in-cloud-environments-benefits-and-best-practices
[6] Saravanakumar Baskaran, "A Quantitative Assessment of the Impact of Automated Incident Response on Cloud Services Availability," ResearchGate, 2023. Available: https://www.researchgate.net/publication/385277305_A_Quantitative_Assessment_of_the_Impact
_of_Automated_Incident_Response_on_Cloud_Services_Availability
[7] Derek Pascarella, "Future-Proof Your IT: Understanding Self-Healing IT Infrastructure," Resolve.io, 2025. Available: https://resolve.io/blog/guide-to-self-healing-it-infrastructure
[8] Vaidyanathan Sivakumaran, "Enhancing Application Monitoring Through AI-Driven Alert Correlation," ResearchGate, 2025. Available: https://www.researchgate.net/publication/388074335_Enhancing_Application_Monitoring_Through_AI-Driven_Alert_Correlation
[9] Saravanakumar Baskaran, "Evaluating the Impact of Site Reliability Engineering on Cloud Services Availability," ResearchGate, 2020. Available: https://www.researchgate.net/publication/386087642_Evaluating_the_Impact_of_Site_Reliability_Engineering_on_Cloud_Services_Availability
[10] Sasank Tummalpalli, "Self-Healing Network Infrastructure: The Future of Autonomous Network Management," International Journal of Research in Computer Applications and Information Technology, 2025. Available: https://iaeme.com/MasterAdmin/Journal_uploads/IJRCAIT/VOLUME_8_ISSUE_1/IJRCAIT_08_01_039.pdf

제공된 파일의 전체 내용을 전문가 수준의 용어와 평어체를 사용하여 한국어로 번역한다.

AI 증강 자가 치유 인프라: 상태 프로브와 복구 플레이북의 결합

Manoj Kumar Reddy Kalakoti

텍사스 A&M 대학교, 킹스빌, 미국

다음 표:

기사 정보 ,초록

접수: 2025 년 7 월 8 일 ,수정: 2025 년 8 월 11 일 ,승인: 2025 년 8 월 18 일 ,"자가 치유(Self-healing) 인프라는 복원력 있는 클라우드 네이티브 시스템의 기본 구성 요소로 발전했다. 본 논문은 인공지능(AI)과 머신러닝(ML)을 활용하여 기존의 상태 프로브(health probe) 기반 복구(remediation)를 향상시키는 고급 프레임워크를 소개한다. AI 기반 이상 징후 탐지, 적응형 복구 전략, 생성형 플레이북 합성을

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

통합함으로써, 제안하는 아키텍처는 반응형 장애 대응을 사전 예방적, 예측적, 자율적 패러다임으로 전환한다. Kubernetes, AWS CloudWatch, Prometheus 와 같은 네이티브 관측 가능성(observability) 도구와 LLM 기반 패턴 추론을 결합하여, 우리는 다계층 AI 기반 모니터링 시스템을 구축한다. AWS Lambda 및 EventBridge 를 통한 이벤트 기반 자동화는 지능형 결정 엔진 및 강화 학습(RL) 루프로 확장된다. 복구 워크플로우는 Ansible 및 AWS Systems Manager 를 통해 실행되며, 새로운 인시던트에 맞춤화된 AI 생성 플레이북으로 향상된다. 경험적 검증 결과, 평균 복구 시간(MTTR)의 극적인 감소, 장애 예방률 향상, 운영 오버헤드 감소가 나타났다. 본 연구는 자가 치유를 멀티 클라우드 복원력에 필수적인, 지능적이고 지속적으로 진화하는 역량으로 재정의한다. 우리가 아는 한, 이는 자가 치유 클라우드 환경에서 플레이북 합성을 위해 LLM 과 RL 을 통합한 최초의 프레임워크이다. ", ,"키워드: 자가 치유 인프라, AI 기반 자동화, 생성형 플레이북, LLM 복구, 이상 징후 예측, 멀티 클라우드 복원력, 이벤트 기반 아키텍처 "

서론

클라우드 기반 서비스의 급속한 확장은 기술 부문의 인프라 관리 접근 방식을 변화시켰다.

조직들이 미션 크리티컬 워크로드를 급속히 분산된 환경으로 마이그레이션함에 따라, 전통적인 수동 개입 방식은 오늘날의 복잡한 생태계에서 최적의 서비스 가용성과 성능을 유지하기에 부적절함이 입증되었다.

연구에 따르면, 전통적인 인시던트 대응 프로토콜은 운영 환경의 서비스 신뢰도 지표에 중대한 영향을 미치는 고유한 인간의 대기 시간(latency) 요인으로 인해 어려움을 겪는다 [1].

인간의 개입 없이 장애를 자율적으로 탐지, 진단, 복구하도록 설계된 자가 치유 인프라 시스템은 현대 클라우드 플랫폼의 필수 기능으로 부상했다.

최근 연구에 따르면 자율 복구 프레임워크를 구현하는 조직은 전통적인 경보 기반 수동 프로세스에 의존하는 조직에 비해 운영 복원력이 상당히 향상되는 것으로 나타났다 [2].

이러한 자가 치유 기능은 반응형(reactive)에서 사전 예방적(proactive) 인프라 관리로의 패러다임 전환을 나타내며, 시스템이 최종 사용자에게 영향을 미치기 전에 잠재적인 장애를 예측하고 완화할 수 있게 한다 [1].

본 논문은 멀티 클라우드 운영 환경 전반에 걸쳐 자동화된 상태 프로브와 복구 플레이북을 통합하는 자가 치유 인프라의 포괄적인 구현을 검토한다.

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

이 접근 방식은 Kubernetes, AWS CloudWatch, Prometheus 의 네이티브 모니터링 기능을 활용하여, 시스템 수준 메트릭과 애플리케이션별 지표를 모두 캡처하는 다계층 관측 가능성 프레임워크를 생성한다.

연구[1]는 이 포괄적인 모니터링 접근 방식이 이기종 클라우드 환경 전반에 걸쳐 정확한 기준 행동(baseline behaviors)을 설정하고 비정상적인 상태를 탐지하는 데 중요하다고 강조한다.

이 아키텍처는 탐지된 이상 징후가 연쇄적인 자동화 프로세스를 트리거하는 이벤트 기반 워크플로우를 사용한다.

모니터링 시스템의 상태 이벤트는 AWS Lambda 와 EventBridge 를 통해 라우팅되어 Ansible 및 Systems Manager Automation Documents 로 구축된 해당 복구 플레이북을 호출한다.

연구[2]에 따르면 이벤트 기반 아키텍처는 폴링 기반 접근 방식에 비해 자가 치유 시스템에 우수한 성능을 제공하며, 특히 서비스 간 종속성 및 복잡한 장애 모드와 관련된 시나리오에서 더욱 그렇다.

엔터프라이즈 환경의 경험적 데이터는 자가 치유 기능을 구현하는 조직이 복구 시간을 크게 단축하고 반복적인 수동 개입을 제거함을 보여준다.

연구[2]에 따르면 자동화된 복구 프레임워크는 특히 메모리 누수, 설정 드리프트(configuration drift), 일시적인 네트워크 문제와 같이 전통적으로 인간의 문제 해결이 필요했던 일반적인 장애 시나리오를 해결하는 데 탁월했다.

현대 자가 치유 시스템의 예측 기능은 서비스 저하가 발생하기 전에 잠재적인 장애를 사전에 해결할 수 있으며, 이는 "선제적 복원력 엔지니어링(pre-emptive resilience engineering)"으로 불리는 기능이다 [1].

본 연구는 클라우드 인프라 엔지니어링 분야의 자율 시스템에 대한 지식체에 기여하며, 운영 환경에서 자가 치유(self-consciousness/self-healing)의 효과성에 대한 증거를 제공한다.

클라우드 인프라의 복잡성이 계속 증가함에 따라, 자가 치유 기능은 단순한 운영 기능을 넘어 대규모 환경에서 신뢰성을 유지하기 위한 필수 기반이 되고 있다.

다음 표:

개념 ,중요성 ,구현 접근 방식

자율 ,탐지 ,잠재적 장애의 사전 예방적 식별 가능 ,"Kubernetes, ","AWS CloudWatch, Prometheus 의 네이티브 모니터링 기능 "

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

자동화된 ,진단 ,인시던트 대응 시 인간의 대기 시간 단축 ,이기종 환경 전반의 다계층 관측 가능성 프레임워크

자가 복구 ,수동 개입 요구 사항 제거 ,AWS Lambda 및 EventBridge 를 통한 이벤트 기반 워크플로우

복구 ,플레이북 ,일관된 복구 절차 보장 ,"Ansible 및 Systems Manager Automation ,Documents 통합 "

선제적 ,복원력 ,서비스 저하 전 문제 해결 ,잠재적 장애 완화를 위한 예측 기능

표 1: 자가 치유 인프라의 기본 요소 [1, 2]

연구의 중요성 및 기여

본 연구는 자율 클라우드 인프라 관리 분야에 다음과 같은 몇 가지 중요한 기여를 한다:

첫째, 이전 연구들이 자가 치유 시스템의 개별 구성 요소를 탐구한 반면, 본 논문은 탐지-진단-복구에 이르는 전체 인시던트 라이프사이클에 걸친 포괄적이고 통합된 프레임워크를 제시한다.

특정 장애 모드를 다루는 사일로화된(siloed) 접근 방식과 달리, 우리의 아키텍처는 이기종 클라우드 환경에 적용 가능한 엔드-투-엔드 솔루션을 제공한다.

둘째, 본 연구는 다양한 모니터링 도구와 복구 시스템 간의 표준화된 통합 패턴을 문서화함으로써 중요한 구현 격차를 해소한다.

이전 문헌[3]에서는 이 통합 문제를 광범위한 채택의 주요 장벽으로 지목했으며, 대부분의 조직이 플랫폼 전반에 걸쳐 일관된 복구 전략을 유지하는 데 어려움을 겪고 있음을 밝혔다.

우리의 프레임워크는 서로 다른 소스의 경보를 표준화된 처리 형식으로 정규화하는 통합 이벤트 버스 아키텍처를 통해 이 문제를 해결한다.

셋째, 본 연구는 기존 솔루션에 만연한 이진(binary) 상태 점검보다 크게 발전한, 등급화된 심각도 임계값과 비례적 대응 전략을 도입한다.

경고와 심각 상태를 구별함으로써, 우리 시스템은 상태가 서비스에 영향을 미칠 수준으로 악화되기 전에 영향이 적은 선제적 개입을 가능하게 한다

넷째, 본 연구는 여러 운영 차원에 걸쳐 엄격한 전후 비교 분석을 통해 자가 치유 효과성에 대한 경험적 검증을 제공한다.

이론적 이점은 널리 논의되었지만, 운영 환경에서 측정 가능한 개선을 입증하는 정량적 연구는 제한적이었다.

우리의 연구 결과는 자동화된 복구 기능과 주요 성능 지표(탐지 시간, 복구 시간, 장애 재발률 등) 간의 인과 관계를 확립한다.

마지막으로, 이 프레임워크는 대규모 중단 시 '복구 폭풍(remediation storm)'을 방지하기 위한 컨텍스트 인식 이벤트 상관 관계 로직을 도입한다. 이는 많은 1 세대 자동화 시스템에 없던 중요한 안전 메커니즘이다.

이 지능형 상관 관계 기능은 자동화된 조치가 잠재적으로 문제를 해결하는 대신 서비스 중단을 증가시키는 "복구 루프(Remediation Loop)"를 방지한다.

이러한 기여들은 자율 인프라 관리 분야의 최신 기술 수준을 발전시키며, 유연한 클라우드 시스템을 위한 이론적 토대와 실제 구현 패턴을 모두 제공한다.

자가 치유 인프라에 대한 문헌 검토

자가 치유 인프라는 2000 년대 초반 IBM 연구원들이 처음 제안한 자율 컴퓨팅(autonomic computing) 원칙을 기반으로 구축된, 전통적인 모니터링 및 경보 시스템의 진화를 나타낸다.

연구[3]에 따르면, 이 혁신적인 접근 방식은 상당한 주목을 받았으며, 2020-2023 년 사이 금융 서비스(연간 약 25% 성장) 및 통신(연간 30% 이상 성장)을 중심으로 산업 전반에 걸쳐 채택률이 꾸준히 증가하고 있다.

구현 패턴에 대한 포괄적인 분석에 따르면, 초기 성숙 단계의 조직은 일반적으로 특정 서비스에 초점을 맞춘 제한된 범위의 솔루션으로 시작하는 반면, 고급 실무자들은 전체 인프라 환경에 걸친 통합 프레임워크를 개발한다.

이 개념은 시스템이 스스로를 모니터링하고, 정상 작동에서 벗어난 편차를 탐지하며, 인간의 개입 없이 수정 기능을 실행하는 것을 포함한다.

최신 연구[4]에 따르면, 현대의 자가 치유 구현은 서비스 저하로 나타나기 전에 미세한 침식 패턴을 식별할 수 있는, 단순한 임계값 기반 트리거에서 발전하여 복잡한 이상 징후를 탐지할 수 있도록 급격히 정교화된 탐지 메커니즘을 통합한다.

운영 환경 분석 결과, 고급 식별 시스템은 전통적인 경계 기반 접근 방식에 비해 오탐(false positive) 비율을 크게 줄였으며, 이는 자동 치료 시스템 운영에 대한 신뢰를 유지하는 데 중요한 요소임이 밝혀졌다.

최근 문헌은 특히 불변의(immutable) 인프라 패러다임이 표준화된 복구 프로세스를 촉진하는 컨테이너화된(containerized) 환경에서 자동화된 치료(복구) 경향이 증가하고 있음을 나타낸다.

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

여러 산업 분야에 걸친 광범위한 현장 연구[3]에 따르면, 컨테이너화된 아키텍처를 구현한 조직이 전통적인 가상 머신 환경을 운영하는 조직에 비해 자동화된 복구 조치의 성공률이 훨씬 더 높은 것으로 나타났다.

이러한 차이는 컨테이너화된 환경이 더 깨끗한 격리 경계와 더 예측 가능한 복구 경로를 제공하는 복잡한 다중 구성 요소 장애에서 특히 두드러졌다.

여러 엔터프라이즈 환경에서 5 년에 걸쳐 수행된 종단적 연구[4]는 자가 치유 기능을 구현하는 조직이 가용성 지표와 운영 효율성에서 측정 가능한 개선을 달성한다는 강력한 증거를 제공한다.

이는 연구 방법론, 운영, 스테이징, 개발 환경에서의 제어된 실험을 포함하며, 자동화된 복구 능력과 탐지 시간, 복구 시간, 장애율 변화 등 주요 성능 지표 간의 인과 관계를 확립한다.

포괄적인 자가 치유 프레임워크를 구현한 조직은 모든 지표에서 빈번한 성능 향상을 보였으며, 가장 중요한 이점은 대규모 분산 아키텍처에서 나타났다.

그러나 멀티 클라우드 환경에 걸쳐 다양한 모니터링 도구와 조직화된 복구 조치를 통합하는 포괄적인 아키텍처를 문서화하는 데는 여전히 격차가 존재한다.

연구[3]는 이 통합 문제를 광범위한 채택의 주요 장벽으로 식별하며, 대부분의 조직들이 이기종 인프라 플랫폼 전반에 걸쳐 일관된 복구 전략을 유지하는 데 어려움을 겪고 있음을 나타낸다.

추가 연구[4]는 표준화된 통합 패턴의 필요성을 강조하며, 맞춤형 통합 프레임워크를 구현하는 조직이 벤더 제공 통합 기능을 활용하는 조직에 비해 구현 일정이 훨씬 길고 유지 관리 오버헤드가 높다는 점을 지적한다.

본 논문은 이기종 클라우드 플랫폼 전반에 적용 가능한 전체론적 구현 프레임워크를 제시함으로써 이러한 격차를 해결한다.

다음 표:

진화 단계 ,주요 특징 ,산업에 미치는 영향

전통적 모니터링 ,수동 개입이 필요한 기본 임계값 기반 경보 ,복잡한 환경에서 제한된 효과

자율 컴퓨팅 ,IBM 이 시작한 자가 관리 시스템 패러다임 ,현대 자가 치유 접근 방식의 토대

컨테이너화된 복구 ,불변 인프라에서의 표준화된 복구 ,자동화된 조치의 성공률 향상

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

멀티 클라우드 통합 ,이기종 플랫폼 간의 일관된 복구 ,많은 조직의 주요 채택 장벽

포괄적 프레임워크 ,탐지에서 해결까지 엔드-투-엔드 자동화 ,가용성 지표의 측정 가능한 개선

표 2: 산업 전반의 채택 패턴 [3, 4]

방법론 및 실험 설정

본 연구는 정량적 성능 분석과 여러 클라우드 환경에 걸친 운영 개선에 대한 정성적 평가를 결합한 혼합 방법론적(mixed-methods) 접근 방식을 사용했다.

실험 설계는 제어된 테스트 단계와 운영 배포 단계를 모두 포함하여 자가 치유 프레임워크의 효과성을 포괄적으로 평가했다.

실험 환경은 세 가지 별개의 배포 컨텍스트를 포함했다:

200 개 이상의 EC2 인스턴스에서 120 개 이상의 마이크로서비스를 실행하는 운영 AWS 환경

하이브리드 배포 모델의 75 개 이상 서비스를 갖춘 보조 Azure 환경

15 개의 워커 노드에 50 개 이상의 파드(pod)가 분산된 개발 Kubernetes 클러스터

기준(baseline) 지표를 설정하기 위해, 우리는 구현 전 6 개월간의 과거 인시던트 데이터를 수집하여 수동 해결 워크플로우, 개입 시간, 서비스 영향 기간을 문서화했다.

427 개의 개별 인시던트로 구성된 이 데이터셋은 비교 분석을 위한 대조군(control group) 역할을 했다.

테스트 프로토콜은 5 가지 주요 장애 범주에 걸쳐 자연 발생 장애와 제어된 결함 주입(fault injection)을 모두 통합했다:

리소스 고갈 이벤트 (메모리 누수, CPU 포화)

설정 드리프트(Configuration drift) 및 불일치

네트워크 연결 및 지연 문제

애플리케이션별 장애 (스레드 데드락, 커넥션 풀 고갈)

종속성 장애 (데이터베이스 사용 불가, 타사 서비스 중단)

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

제어된 테스트를 위해, 우리는 지정된 유지 관리 기간 동안 중요하지 않은 서비스 구성 요소에 이러한 장애 모드를 체계적으로 주입하는 카오스 엔지니어링(chaos engineering) 프레임워크를 개발했다.

각 장애 유형은 통계적 유의성을 보장하기 위해 20 회 반복 테스트되었으며, 다음 항목을 측정했다:

탐지 시간 (TTD): 결함 도입과 시스템 식별 사이의 기간

복구 시간 (TTR): 탐지와 성공적인 해결 사이의 기간

성공률: 인간 개입 없이 자동으로 해결된 인시던트의 비율

서비스 영향: 복구 중 사용자가 인지하는 성능 저하

운영 배포에는 단계적 전략이 사용되었으며, 중요하지 않은 서비스에서 시작하여 구조적 신뢰도가 높아짐에 따라 점차 미션 크리티컬 구성 요소를 포함하도록 확장했다.

원격 측정(Telemetry) 데이터는 관측 가능성 스택을 통해 지속적으로 수집되었으며, 엄격한 사후 인시던트 분석을 지원하기 위해 모든 자동화된 조치에 대한 상세한 로깅이 이루어졌다.

구현 프로세스는 4 개의 뚜렷한 단계로 구성되었다:

기준 설정: 구현 전 지표 수집 및 분석

프레임워크 배포: 모니터링 도구 및 복구 프레임워크 설치 및 구성

제어된 검증: 체계적인 결함 주입 및 성능 측정

운영 롤아웃: 지속적인 모니터링과 함께 운영 서비스로 점진적 확장. 이러한 체계적인 접근 방식은 성능 평가의 과학적 엄격성과 자동화된 복구로의 전환 중 운영 안전성을 모두 보장했다.

시스템 아키텍처 및 구현

본 연구에서 설명하는 자가 치유 인프라 프레임워크는 모니터링, 이벤트 처리, 자동 치료의 여러 계층을 통합하여 현대 클라우드 환경을 위한 포괄적인 솔루션을 생성한다.

기반에는 Kubernetes, AWS CloudWatch, Prometheus 의 네이티브 상태 점검이 다양한 차원에서 시스템 상태를 지속적으로 평가한다.

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

연구[5]에 따르면, 이 다계층 모니터링 접근 방식은 기존의 사일로화된 모니터링 시스템보다 크게 발전했으며, 통합된 관측 가능성 프레임워크가 연결되지 않은 모니터링 솔루션보다 약 4 배 더 빨리 이상 징후를 탐지함을 보여준다.

운영 환경 분석 결과, 포괄적인 모니터링 전략을 구현하는 조직은 단일 소스 모니터링 접근 방식에 의존하는 조직에 비해 이상 징후 탐지 정확도와 탐지 시간 지표에서 상당한 개선을 경험했다.

모니터링 프레임워크는 인프라 수준 메트릭, 애플리케이션별 지표, 상태 검증을 포괄하는 3 계층 접근 방식을 구현한다.

연구[6]는 이 포괄적인 모니터링 전략의 중요성을 강조하며, 운영 인시던트의 약 3 분의 2 는 인프라 수준 문제에서 비롯되고, 나머지 3 분의 1 은 애플리케이션별 문제 및 구성 이상에서 비롯된다는 점을 지적한다.

클라우드 서비스 제공업체에 대한 종단적 연구는 다차원 모니터링 프레임워크를 구현하는 조직이 관측 가능성 범위의 사각지대를 현저히 줄였음을 추가로 입증했으며, 특히 장애가 종종 전통적인 모니터링 경계를 넘어 전파되는 복잡하고, 상호 의존적인 서비스에서 더욱 그랬다.

이 아키텍처는 탐지된 이상 징후가 자동화된 프로세스의 연쇄 반응을 일으키는 이벤트 기반 워크플로우를 사용한다.

AWS EventBridge 는 중앙 이벤트 버스 역할을 하여, 상태 이벤트를 AWS Lambda 서비스로 구현된 적절한 처리 기능으로 라우팅한다.

연구[5]는 이 이벤트 기반 아키텍처가 자가 치유 시스템에 특히 적합하다고 식별하며, 분리된(decoupled) 이벤트 기반 워크플로우가 인시던트 시나리오 동안 기존 폴링 기반 접근 방식보다 약 5 배 더 높은 처리량을 달성했음을 문서화한다.

제어된 실험은 이벤트 기반 아키텍처가 극단적인 부하 조건에서도 일관된 성능을 유지함을 입증했으며, 이는 광범위한 중단 시 안정적으로 작동해야 하는 복구 시스템의 중요한 요구 사항이다.

[그림]

그림 1: 자가 치유 인프라를 위한 이벤트 기반 아키텍처

이러한 기능(Lambda 함수)은 이벤트 컨텍스트를 분석하고 사전 정의된 플레이북 라이브러리에서 최적의 복구 전략을 결정한다.

여러 산업 분야에 걸친 광범위한 현장 연구[6]에 따르면, 조직이 구조화된 복구 라이브러리를 구현할 경우 임시(ad-hoc) 스크립팅 접근 방식을 사용하는 조직에 비해 자동화된 복구 조치의 성공률이 훨씬 더 높은 것으로 나타났다.

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

수천 건의 인시던트 대응 시나리오 분석 결과, 잘 설계된 플레이북 라이브러리는 관찰된 장애 모드의 80~95%를 일반적으로 다루었으며, 적용 범위는 새로운 패턴이 복구 프레임워크에 통합됨에 따라 시간이 지남에 따라 향상되었다.

복구 조치는 구성 관리 작업을 위해 Ansible 을 통해 실행되고, AWS 특정 작업을 위해 AWS Systems Manager Automation Documents 를 통해 실행된다.

구현 패턴의 상세한 분석[5]에 따르면, 이 하이브리드 접근 방식은 범용 구성 관리 도구와 클라우드 네이티브 자동화 기능이 결합되어 운영 단순성을 유지하면서 포괄적인 적용 범위를 확보함에 따라 성숙한 실무자들 사이에서 빠르게 자리 잡고 있음을 확인했다.

이 접근 방식은 복구 로직이 운영 환경과 분리되어 유지 관리, 버전 제어, 감사가 가능하도록 보장하며, 이는 주요 요구 사항이다.

다음 표:

프레임워크 ,요소 ,기능적 역할 ,구현 이점

다계층 ,모니터링 ,시스템 차원 전반의 상태 평가 ,사일로 방식보다 4 배 빠른 이상 징후 탐지

3 계층 ,접근 방식 ,"인프라, 애플리케이션, 상태 검증 포괄 ",장애 원인에 대한 포괄적 가시성

이벤트 기반 ,아키텍처 ,상태 이벤트의 처리 기능 라우팅 ,인시던트 중 5 배 더 높은 처리량

구조화된 ,복구 라이브러리 ,체계적인 복구 절차 모음 ,임시 스크립팅보다 높은 성공률

하이브리드 도구 ,통합 ,범용 도구와 클라우드 네이티브 도구의 결합 ,운영 단순성을 갖춘 더 넓은 복구 범위

표 3: 자동화된 복구를 위한 통합 패턴 [5, 6]

모니터링 및 이벤트 처리 프레임워크

본 연구에서 구현된 모니터링 프레임워크는 상태 검증에 다계층 접근 방식을 적용하여 복잡한 인프라 환경 전반에 걸쳐 포괄적인 가시성을 생성한다.

자가 치유 구현에 대한 광범위한 분석[7]에 따르면, 이 계층화된 접근 방식은 업계 모범 사례를 나타내며, 연구에 따르면 다차원 모니터링 전략을 구현하는 조직이

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

단일 계층 접근 방식에 의존하는 조직보다 훨씬 빠르게 이상 징후를 탐지한다고 한다.

엔터프라이즈 환경 조사 결과, 통합 모니터링 프레임워크는 평균 탐지 시간을 기존의 사일로화된 접근 방식의 분 단위에서 초 단위로 단축시켰으며, 이는 효과적인 자동화 복구를 가능하게 하는 중요한 요소이다.

인프라 계층에서는 시스템 수준의 메트릭이 리소스 활용 패턴을 캡처하고 동적 기준선(baseline)을 설정한다.

연구[7]는 기준선 보정 기술의 중요성을 강조하며, 고급 구현에서는 과거 패턴, 시간대별 변화, 계절적 추세를 기반으로 정상 운영 매개변수를 지속적으로 세분화한다고 지적한다.

이 동적 접근 방식은 정적 임계값에 비해 오탐(false positivity)을 크게 줄여, 자동화된 복구 트리거에 대한 높은 신뢰를 가능하게 한다.

연구는 또한 현대 인프라 모니터링 솔루션이 서비스 중단으로 나타나기 전에 미묘한 저하 패턴을 식별할 수 있는 급속히 정교화된 불일치 알고리즘을 통합하고 있음을 나타낸다.

애플리케이션 계층 모니터링은 응답 시간, 오류율, 합성 트랜잭션(synthetic transaction)을 통한 기능 검증 등 서비스 동작에 중점을 둔다.

연구[8]에 따르면 포괄적인 애플리케이션 모니터링은 기존의 인프라 전용 접근 방식보다 중요한 진전을 나타내며, 분석 결과 서비스 중단의 약 40%가 인프라 모니터링만으로는 탐지되지 않았을 애플리케이션 수준 문제에서 비롯된 것으로 나타났다.

여러 산업 분야의 모니터링 관행에 대한 상세한 조사 결과, 합성 트랜잭션은 인프라 관점에서는 정상 작동하는 것처럼 보이지만 잘못된 결과를 제공하는 "조용한 장애(silent failures)"를 탐지하는 데 특히 유용한 것으로 나타났다.

상태 프로브는 경고와 심각 상태를 구별하기 위해 등급화된 임계값으로 구성되어 비례적인 대응 전략을 가능하게 했다.

연구[8]는 이 등급화된 접근 방식이 이진(binary) 상태 점검보다 훨씬 효과적임을 확인했으며, 계층화된 심각도 분류가 상태가 심각한 수준으로 악화되기 전에 영향이 적은 복구 조치를 위한 귀중한 리드 타임(lead time)을 제공함을 문서화했다.

이 연구는 이러한 이점을 정량화하여, 등급화된 임계값을 구현하는 조직이 비파괴적 개입을 통해 잠재적 인시던트의 상당 부분을 해결했다고 지적했다. 반면, 전통적인 이진 점검을 사용하는 조직은 종종 서비스에 영향을 미치는 조치가 필요했다.

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

이벤트 처리 시스템은 대규모 중단 시 복구 폭풍(remediation storm)을 방지하기 위해 컨텍스트 인식 로직을 통합했으며, 사전 정의된 실패 횟수 후에 성공하지 못한 복구 시도를 중단하기 위해 서킷 브레이커(circuit-breaker) 패턴을 구현했다.

연구[7]는 이러한 지능형 상관 관계 기능이 자동화된 조치가 잠재적으로 서비스 중단을 해결하기보다는 악화시킬 수 있는 "복구 루프(remediation loops)"를 방지하는 데 필수적이라고 강조한다.

추가 연구[8]에 따르면 고급 상관 관계 알고리즘은 기존의 규칙 기반 접근 방식이 어려움을 겪는 복잡한 분산 환경에서도 관련된 이벤트를 식별하는 데 거의 완벽한 정확도를 달성한다.

이 정교한 이벤트 처리는 광범위한 컨텍스트 인식 없이 개별 경보에 따라 작동했던 1 세대 자동화 프레임워크보다 크게 발전한 것이다.

다음 표:

모니터링 구성 요소 ,구현 접근 방식 ,운영상의 이점

다계층 ,검증 ,상태 모니터링에 대한 계층적 접근 ,훨씬 빠른 이상 징후 탐지

동적 ,기준선 ,정상 매개변수의 지속적 세분화 ,정적 임계값 대비 오탐 감소

애플리케이션 ,모니터링 ,서비스 동작 및 합성 트랜잭션에 중점 ,"인프라 모니터링에 보이지 않는 ""조용한 장애"" 탐지 "

등급화된 ,임계값 ,계층화된 심각도 분류 ,심각한 영향 전, 영향이 적은 복구를 위한 리드 타임 확보

컨텍스트 인식 ,상관관계 ,서킷 브레이커 패턴을 갖춘 지능형 이벤트 처리 ,"대규모 중단 시 ""복구 루프"" 방지 "

표 4: 자율 복구를 위한 이벤트 처리 메커니즘 [7, 8]

결과 및 성능 분석

자가 치유 인프라 프레임워크의 구현은 여러 운영 차원에 걸쳐 정량화 가능한 개선을 가져왔다.

클라우드 서비스 제공업체 전반의 사이트 신뢰성 엔지니어링(SRE) 관행을 검토한 포괄적인 연구[9]에 따르면, 조직이 자동화된 복구 프레임워크를 구현할 경우 기존의 수동 접근 방식에 비해 인시던트 해결 시간이 지속적으로 크게 단축되는 것으로 나타났다.

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

엔터프라이즈 환경에 대한 종단적 연구에서는 평균 복구 시간(MTTR)이 평균 70% 이상 감소했으며, 특히 자동화된 접근 방식이 수동 문제 해결 워크플로우에 내재된 진단 지연을 제거한 인프라 관련 장애에서 상당한 개선이 관찰되었다.

이러한 개선은 메모리 누수, 설정 드리프트, 일시적인 네트워크 문제와 같은 일반적인 장애 모드에서 특히 두드러졌으며, 자동화된 복구는 일반적으로 이전 평균 20 분을 초과하던 문제를 몇 분 내에 해결했다.

상세 분석[10]에 따르면, 이전에 광범위한 수동 조사가 필요했던 메모리 관련 인시던트는 자동화된 탐지 및 복구에 매우 잘 반응하여, 해결 시간이 대부분의 경우 20 분대에서 2 분 미만으로 감소했다.

여러 클라우드 환경의 인시던트 데이터를 검토한 결과, 전통적으로 문제 해결의 복잡성으로 인해 해결 시간 편차가 컸던 구성 관련 장애가 자동화를 통해 가장 일관된 개선을 보였으며, 표준 편차가 약 13 분에서 2 분 미만으로 감소했다.

운영 데이터 분석 결과, 탐지된 이상 징후의 대다수가 인간 개입 없이 자동화된 복구를 통해 성공적으로 해결되었다.

연구[9]에 따르면 다양한 인시던트 범주에서 성공률이 85%에 육박했으며, 명확한 복구 경로가 있는 잘 알려진 장애 패턴에서 가장 높은 성공률이 관찰되었다.

부분적인 수동 개입이 필요했던 나머지 인시던트는 일반적으로 시스템이 이전에 접하지 못했던 새로운 장애 모드와 관련이 있었다.

상세 사례 연구[10]는 이러한 맥락에서 지속적인 학습 메커니즘의 중요성을 강조하며, 체계적인 피드백 루프를 구현하는 조직이 매월 평균 5~8 개의 새로운 복구 패턴을 통합하여 점진적으로 자동화 적용 범위를 확장했음을 문서화한다.

서비스 신뢰도 지표는 구현 이후 상당한 개선을 보였다. 구현 전후 기간의 비교 분석[9]은 고객에게 영향을 미치는 인시던트의 상당한 감소와 이전에 일관성 없는 수동 복구 절차로 인해 발생했던 반복적인 문제의 거의 완전한 제거를 입증했다.

서비스 신뢰도 데이터에 대한 상세한 검토 결과, 이러한 개선 사항은 비즈니스 성과로 직접 이어졌으며, 고객 만족도 지표는 서비스 안정성 향상에 비례하여 증가했다.

연구[10]는 또한 조직이 이러한 신뢰도 향상을 달성하면서 동시에 운영 비용을 절감했음을 문서화했으며, 이는 주로 인시던트 관련 다운타임 감소 및 일상적인 복구 작업에 투입되는 인력 시간 감소를 통해 이루어졌다.

예측적 상태 지표에 의해 촉발된 사전 예방적 확장 및 최적화 조치를 통해 리소스 활용 효율성이 크게 향상되었다.

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ **Research Article**

연구[9]에 따르면 정교한 이상 징후 탐지 기능을 구현하는 조직은 기존의 반응형 모니터링 프레임워크에서는 탐지되지 않았을 최적화 기회를 식별했다.

분석[10]은 이러한 효율성 향상을 정량화하여, 사전 예방적 리소스 관리가 서비스 성능이나 신뢰성을 저해하지 않으면서 인프라 비용과 환경 영향을 모두 크게 줄였다고 지적했다.

그림 2 는 이러한 MTTR 개선 사항을 다양한 장애 범주에 걸쳐 시각화하여, 자동화된 복구를 통해 달성한 복구 시간의 극적인 단축을 강조한다.

[그림]

그림 2: 평균 복구 시간(MTTR) 비교

미래 방향: AI 증강 자가 치유 인프라

본 논문에서 제시한 자가 치유 프레임워크가 상당한 운영 개선을 입증했지만, 최근 인공지능과 머신러닝의 발전은 우리가 "자가 치유 인프라 2.0"이라 명명하는 방향으로의 추가적인 진화를 가능하게 한다.

이 섹션에서는 AI 증강이 어떻게 전통적인 상태 프로브 기반 복구를 더 지능적이고, 적응적이며, 자율적인 패러다임으로 변화시킬 수 있는지 탐구한다.

AI 강화 프레임워크 아키텍처

차세대 자가 치유 인프라는 기존 아키텍처를 기반으로 다음과 같은 몇 가지 AI 기반 기능을 통합하여 구축된다:

향상된 관측 가능성 계층

전통적인 모니터링 접근 방식은 서비스 중단으로 나타나기 전에 미묘한 이상 징후를 탐지할 수 있는 비지도 머신러닝 알고리즘을 통해 크게 향상될 수 있다.

이러한 고급 탐지 메커니즘은 다음을 포함한다:

시계열 예측을 이용한 동적 임계값 보정

비정상적인 메트릭 패턴 식별을 위한 클러스터링 알고리즘

로그 이상 징후 탐지를 위한 딥러닝 모델

서로 다른 시스템 간의 이벤트 상관 관계를 위한 트랜스포머(Transformer) 기반 모델

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

이러한 기술은 정적 모니터링을 지능형 관찰로 변환하여, 기존의 임계값 기반 접근 방식으로는 보이지 않았을 미세한 저하 패턴을 식별할 수 있게 한다.

지능형 이벤트 처리

AI 증강 이벤트 처리는 단순한 상관 관계를 넘어 다음을 통합한다:

이벤트 분류 및 중복 제거를 위한 트랜스포머 기반 모델

지도 학습을 통한 심각도 분류

오탐(false positive) 제거를 위한 노이즈 감소 알고리즘

분산 시스템 전반의 근본 원인을 식별하기 위한 인과 관계 추론

강화 학습을 이용한 결정 엔진

복구 선택 프로세스는 다음과 같은 강화 학습(RL) 알고리즘을 통해 최적화될 수 있다:

과거 결과를 기반으로 최적의 복구 전략 학습

즉각적인 복구와 장기적인 시스템 안정성 간의 균형 유지

변화하는 환경 조건에 적응

결정 품질을 지속적으로 개선하기 위한 피드백 루프 통합

생성형 플레이북 합성

아마도 가장 혁신적인 기능은 이전에 볼 수 없었던 장애 모드에 대한 복구 플레이북을 생성하기 위해 대규모 언어 모델(LLM)을 통합하는 것일 것이다:

인프라 구성 요소를 이해할 수 있는 파인튜닝된(fine-tuned) 모델

장애 모드 및 적절한 복구 단계에 대한 문맥적 추론

새로운 인시던트에 대한 실행 가능한 플레이북 자동 생성

생성된 절차의 안전성을 보장하기 위한 검증 메커니즘

구현 아키텍처

AI 증강 아키텍처는 5 개의 주요 계층으로 구성된다:

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

관측 가능성 계층: Kubernetes liveness/readiness 프로브, AWS CloudWatch, Prometheus 익스포터(exporter)를 통합한다.

AI 에이전트가 클러스터링 및 시계열 예측을 사용하여 불일치를 탐지한다..

이벤트 수집 및 상관 관계: AWS EventBridge 가 이상 징후 신호를 수집한다. 트랜스포머 기반 상관 관계 엔진이 이벤트를 중복 제거하고 심각도 수준을 분류한다.

결정 엔진: 현재 시스템 상태와 과거 결과를 기반으로 최적의 복구 정책을 선택하기 위해 RL 알고리즘을 적용한다.

플레이북 합성 모듈: 기존 플레이북이 충분하지 않을 때, 파인튜닝된 LLM 이 인시던트 메타데이터를 기반으로 단계별 복구 지침을 생성한다.

복구 실행: 선택된 조치는 Ansible 또는 AWS Systems Manager 를 통해 실행된다.

결과는 로깅, 분석되어 학습 루프로 다시 피드백된다.

예비 성능 결과

세 가지 엔터프라이즈 환경 (AWS, Azure, 하이브리드 Kubernetes)에서 AI 증강 자가 치유 기능에 대한 예비 테스트 결과, 유망한 결과가 입증되었다:

일반적인 인시던트에 대한 MTTR 85% 감소

알려진 패턴에 대한 92%의 복구 성공률

LLM 생성 플레이북을 사용한 새로운 장애에 대한 67%의 성공률

AI 기반 임계값 설정을 사용하여 경보의 오탐 48% 감소

SRE 효율성 20% 향상, 수동 에스컬레이션 감소

결론

자가 치유 인프라는 클라우드 환경 관리의 패러다임 변화를 나타내며, 세련된 탐지 및 복구 기능을 통해 자율 운영을 가능하게 한다.

상태 점검과 자동화된 복구 플레이북의 통합은 전통적인 수동 개입 방식의 근본적인 한계를 해결하며, 특히 서비스 상호 의존성이 복잡한 장애 환경을 생성하는 멀티 클라우드 환경에서 더욱 그렇다.

비정상적인 조건에 의해 트리거되는 이벤트 기반 워크플로우를 적용함으로써, 조직은 복구 시간을 극적으로 단축하고 일관성 없는 수동 프로세스로 인한 반복적인 문제를 제거할 수 있다.

다단계 모니터링 접근 방식은 인프라, 애플리케이션, 통합 계층 전반에 걸쳐 광범위한 가시성을 생성하여, 서비스 영향이 발생하기 전에 잠재적인 장애를 정확하게 탐지할 수 있게 한다.

우리의 경험적 결과에서 보듯이, 이 접근 방식은 복구 시간 70% 단축, 85%의 자동화된 복구 성공률, 반복적인 문제의 거의 완전한 제거를 포함하여 상당한 운영 개선으로 이어진다.

이러한 지표는 서비스 신뢰도 향상, 고객 만족도 향상, 운영 비용 절감을 통해 비즈니스 결과로 직접 변환된다.

더 나아가, 인공지능 기능의 통합은 자가 치유 시스템의 다음 진화 단계를 나타낸다.

최신 연구에 따르면 AI 증강 프레임워크는 학습을 통해 비공식적으로 불일치를 탐지하고, 강화 학습을 통해 치료(복구) 선택을 최적화하며, 심지어 이전에 볼 수 없었던 장애 모드에 대한 새로운 복구 프로세스를 생성한다.

이러한 발전은 자가 치유를 반응형 기능에서 사전 예방적(reactionary/proactive)이고 적응적인 시스템으로 실제로 전환시킬 것이다.

클라우드 환경의 복잡성이 증가함에 따라, 자율적인 자가 치유 기능은 운영상의 이점에서 신뢰성과 성능 유지를 위한 기본 요구 사항으로 전환되고 있다.

본 연구는 완전히 유연한 인프라를 향한 이 혁신적인 여정을 시작하는 조직을 위한 이론적 토대와 실제 구현 패턴을 모두 제공한다.

참고 문헌

Henry Josh, et al., "자가 치유 인프라: 장애 허용 DevOps

환경을 위한 AI 기반 자동화," ResearchGate, 2024 년. 이용 가능:

https://www.researchgate.net/publication/388634507_Self-Healing_Infrastructure_AIPowered_Automation_for_Fault-Tolerant_DevOps_Environments

Dakshaja Prakash Vaidya, "멀티 클라우드 환경에서의 AI 기반 예측 복원력,"

ResearchGate, 2025 년. 이용 가능:
https://www.researchgate.net/publication/392183180_AI-

Driven_Predictive_Resilience_in_Multi-Cloud_Environments

Anil Abraham Kuriakose, "대규모 언어 모델을 통한 자가 치유 인프라,"

Algomox, 2025 년. 이용 가능:

https://www.algomox.com/resources/blog/self_healing_infrastructure_llm/

Merve Şener, "주요 인프라에 대한 사이버 공격의 경제적 영향," IGI Global, 2019 년.

이용 가능: https://www.igi-global.com/gateway/chapter/228475

Cătălina Mărcuță, "클라우드 환경의 이벤트 기반 아키텍처 탐구 – 이점 및 모범 사례," MoldStud, 2025 년. 이용 가능: https://moldstud.com/articles/p-exploring-event-drivenarchitectures-in-cloud-environments-benefits-and-best-practices

Saravanakumar Baskaran, "자동화된 인시던트 대응이 클라우드 서비스 가용성에 미치는 영향에 대한 정량적 평가," ResearchGate, 2023 년. 이용 가능: https://www.researchgate.net/publication/385277305_A_Quantitative_Assessment_of_the_Impact

_of_Automated_Incident_Response_on_Cloud_Services_Availability

Derek Pascarella, "미래 보장형 IT: 자가 치유 IT 인프라의 이해," Resolve.io, 2025 년. 이용 가능: https://resolve.io/blog/guide-to-self-healing-it-infrastructure

Vaidyanathan Sivakumaran, "AI 기반 경보

상관 관계를 통한 애플리케이션 모니터링 향상," ResearchGate, 2025 년. 이용 가능:

https://www.researchgate.net/publication/388074335_Enhancing_Application_Monitoring_Throug h_AI-Driven_Alert_Correlation

Saravanakumar Baskaran, "사이트 신뢰성 엔지니어링(SRE)이 클라우드 서비스

가용성에 미치는 영향 평가," ResearchGate, 2020 년. 이용 가능:

https://www.researchgate.net/publication/386087642_Evaluating_the_Impact_of_Site_Reliability

Journal of Information Systems Engineering and Management
2025, 10(58s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ **Research Article**

Engineering_on_Cloud_Services_Availability

Sasank Tummalpalli, "자가 치유 네트워크 인프라: 자율 네트워크 관리의 미래," International Journal of Research in Computer Applications and Information

Technology, 2025 년. 이용 가능:

https://iaeme.com/MasterAdmin/Journal_uploads/IJRCAIT/VOLUME_8_ISSUE_1/IJRCAIT_08 01_039.pdf