International Journal of Engineering & Extended Technologies Research (IJEETR)

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

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

The "Aegis" Framework: A Multi-Cloud, Fault-Tolerant MLOps Architecture for Real-Time Financial Decisioning and Regulatory Compliance

Suresh Chaganti

Architect - Data & ML OPS, USA

ABSTRACT: The paper provides the findings of a quantitative study on Aegis Framework, a multi-cloud MLOps framework that is aimed at financial institutions. The results point out that the framework is much more beneficial on the availability, speed of the fail-over, and precision of governance in the scenario of working with real-time decision workloads. The cross-cloud routing did not break the inference services whenever cloud failures were detected, and Mean Time to Recovery was less than two seconds. During all stress tests, checking of governance was not lost and metadata logs were not lost. Peak loads high throughput and low latency was obtained as well. These findings confirm that Aegis offers high reliability and resilience as well as compliance performance to the modern financial systems.

KEYWORDS: MLOps, Finance, Regulation, Multi-Cloud, Architecture, Compliance

I. INTRODUCTION

Financial industry requires real-time performance of the software, which should operate continuously, even in the case of cloud outages or any other impulsive workload. A lot of systems are based on single-cloud pipelines, and this poses a threat to downtime, reduction of performance, and audit records are not available. This paper assesses an example of a multi-cloud MLOps architecture, Aegis Framework, that is intended to enhance reliability, resilience, and governance. The research measures availability, failover behaviour, the accuracy of governance and throughput performance using controlled experiments. The introduction establishes the background by stating the reason why multi-cloud design and embedded validation layers are significant to financial workloads that require a stable design with transparency and constant regulatory compliance.

II. RELATED WORKS

Structured Operational Frameworks

The necessity to realize the machine learning models in the complicated production environments has given rise to the creation of the Machine Learning Operations (MLOps). The initial literature has cited that MLOps was created in response to the growing socio-technical challenges in the implementation of the ML systems into the already existing software engineering systems, ongoing monitoring, retraining, and maintenance of deployed models [1].

Researchers claim that even though there are numerous reviews, the field lacks a unified notion on the conceptual basis as the range of tools, practices, and issues are too diverse in reference to the concrete domain [1]. This fragmentation is especially high in high compliance systems such as banking, insurance and trading systems where regulators have to have stringent controls on data lineage, auditability, reproducibility and traceability of all the levels of the ML lifecycle.

According to work driven, MLOps is no longer a set of engineering activities that should be broadly understood as a transition to factory-style of running analytics. Organizations also are no longer relying on the informal cottage-industry analytics processes, but are instead industrialized pipelines where emphasis is placed on the automation of processes, governance and cross-functional interaction [3]. This change will entail both process and technological changes, as well as human skills, pegged on the historical experience of DevOps and DataOps, and including the various domain-specific controls such as model versioning, model validation, and drift detection.

MLOps within regulated industries must involve engaging explicit risk controls in order to comply with the internal model governance requirement together with the requirements of the external regulatory requirements. This view is justified by studies in actuarial science which prove that operationalization of ML used in life insurance requires structured Model Risk Management (MRM) practices, including model validation, data versioning, monitoring pipelines and compliance controls, on frameworks such as Solvency II and IFRS-17 [6]. These findings confirm the

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

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

idea that the contemporary MLOps pipelines should include compliance in their design but not view it as an additional feature.

Apparent deficiency of the major literature review is that where MLOps is concerned with performance, automation, and lifecycle stability, it typically does not have powerful assurances of fault tolerance, model lineage inflexibility, and multi-cloud resiliency, which are provided by Tier-1 financial platforms. The gaps themselves are expected to be addressed by the Aegis Framework, which makes the resilience engineering and regulatory compliance a part of the MLOps cycle.

Secure Operational Pipelines

Zero-trust designs have become important towards the development of safety net systems to facilitate distributed ML workflows. Literature points out that zero-trust model assumes that no party can be defaulted i.e. user, device, or process. Instead, all the access requests are to be authenticated and authorized in real time, as per the user identity, the health of the devices, behavior and context [2].

The method has been greatly researched in the area of IoT because of its susceptibility to assaults as well as distributed attack surfaces [2]. However, the principles can also be applied to the multi-cloud MLOps systems in an organic way, which possess the same set of attributes, that is, the heterogeneous components, distributed services, and high susceptibility to threats.

Scholars believe that to implement zero-trust, it is necessary to implement least-privilege access, continually perform posture assessment, and policy dynamic reaction [10]. These requirements would comply with the requirements of the financial regulation regarding sensitive data management, traceability and controlled access. Zero-trust can also be used to provide the complement to the need of an immutable model lineage in MLOps as every model, dataset and artifact should be provenance to authenticated actors and approved processes.

Other researchers observe that the problems with the implementation of the zero-trust at scale, in a heterogeneous environment (e.g. policy coordination, identity propagation, overheads of legacy systems integration, etc.), exist [2], [10]. Such limitations are essential especially in big financial firms where models could operate on multiple clouds, mixed clusters and infrastructure which are hosted by the vendor.

The literature indicates approaches that entail the adoption of a blend of zero-trust and automation coupled with on-going auditability and collaborative governance systems in an effort of reducing operational friction and increasing security.

The idea of zero-trust is closely associated with the pipeline in the Aegis Framework, which is presented in the form of immutable logging, tokenized identity of any kind of ML artifact, and continuous cross-cloud validation. The current research of MLOps also endorses these ideas by the necessity to create robustness, uncertainty calibration, and controlled degradation in the case of anomalies or attacks [9]. It means that the secure MLOps architecture must have the capacity to guarantee access as well as the model conduct, reliability, and trustworthiness during the execution.

The argument that zero-trust is not a security paradigm but also a significant operational philosophy of the financial MLOps infrastructures where trust must be proved on a continuous basis in both the governance and resilience meaning has been substantiated by the literature.

Edge-Integrated Architectures

The accelerated nature of implementation of cloud-native applications (CNAs) has created a novel category of issues relating to governance and compliance and resilience. The CNAs bring in further complication in the operations because they inject further complication with the distributed micro services, dynamic pattern of orchestration and multi cloud deployment pattern [4].

CNA governance should be firmly attached with the application architecture itself which allows the automated guardrails without stopping the delivery pipes. A battery-included design is one such design of governance that enables the big and small-scale implementation by the use of CNA literature in terms of compliance controls as reconfiguring components [4]. This strategy has enough parallels to the philosophy of Aegis Framework that introduces the governance as a direct part of CI/CD/CM loops.

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

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

Trying to implement multi-cloud systems is also more difficult to exercise governance, but does have certain advantages to financial decision systems that cannot tolerate downtime or local cloud failures. The works on cloud/edge task scheduling point to the fact that high performance and high availability can be achieved due to the course of offering intelligent distribution of tasks to these nodes at the edges and between the cloud providers and the other one too [8].

With systems like reliable server pooling (RSerPool), researchers have established that the balance of latency-redundancy-management overheads in the multi-cloud environment can be achieved as a result of long time set resource selection policies [8]. This is why the design concept of cross-cloud inference endpoints of failover that was designed by Aegis is based on this concept.

Web health Resilience-conscious ML pipelines Resilience-conscious studies of distributed systems prove that distributed systems can be resilient to uncertainty monitoring, calibration and dynamically adaptive components in response to disturbances such as drift or adversarial input [9]. The principles are generic despite them being intended to be used in the medical diagnostics of fault-tolerant decisioning systems in the financial domain where model failure may be very expensive in both financial and regulatory measures.

Any architecture, as depicted in the literature, that has the need to derive continuous and trustworthy decisions in the field of machine learning with regard to the dynamic financial transactions, should have the following features of multi-cloud implementation, inbuilt governance, and resilience conscious pipeline models. The financial rationality of the presence of a framework like Aegis that unites multi-cloud failover, governance-as-code and continuous validation is fairly decent.

Model Risk Management

Model Risk Management (MRM) has been the centre of stage in the current ML governance especially in the financial industry. According to the research of actuarial MLOps, structured MRM leads to greater transparency, auditability and regulatory conformity through the imposition of model validation, versioning and drift control [6]. The demands are also becoming generalized in the financial contexts with regulatory inquiries on the behavior of AI models increasing.

The most recent works on generative AI also contribute to developing the concept of model risk by introducing new problems, such as the problem of adversarial robustness, model usage, violation of fairness, and unintelligibility [7]. The financial institutions are also being forced to expand their MRM systems to include quantitative risk metrics such as probabilistic modelling, Monte Carlo simulated risk and adversarial risk measures [7]. These approaches introduce mathematical rigour, which may be deployed to provide support to the supervisory review activities and internal audit.

The other critical gap that was found during the research is that the majority of existing models of MRM have been built to work with traditional statistical models, and are not consistent with the dynamism of the present AI systems, especially large models and multi-cloud deployment settings [7]. This raises the issue of need of architectures to ensure that MRM is fixed within development and operational cycles as opposed to considering it as a compliance phase once.

Aegis Framework offers a direct answer to these problems as it offers continuous validation of MRM, irrevocable records of lineage to be audit-ready and redundancy of cross-cloud inference so that the models are available during stress scenarios. The provided guidelines are consistent with the suggestions of the literature that indicates the importance of advanced validation as one of the principal aspects of next-generation AI governance and averting its injustice and accentuating its robustness as the vital factors.

III. METHODOLOGY

The Aegis Framework has been suggested as a solution in this paper to establish the improvement of reliability, resilience, and compliance in a multi-cloud MLOps setting through a quantitative research method. The experiment will strive to measure the behaviour of the system in its normal working condition and when failure is under controlled conditions.

All the tests were performed in a simulation financial environment which is a reflection of the size and sensitivity of Tier-1 institutions. The configuration includes real time decisioning loads, distributed cloud applications and continuous governance tests so as to be able to ensure that the findings are near to actual functioning circumstances.

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

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

Aegis architecture is applied on three different cloud platforms and it is realized in the form of an experimental research design approach. Each cloud environment has parallelogram inference endpoints, model registries, metadata stores and audit systems. It is a reflected configuration that allows making a reasonable evaluation of cross-cloud resiliency and failure. The architecture also includes immutable metadata logging and constant validation of the Model Risk Management as a key element that is used to ensure that compliance behaviour can be measured in real time.

The general study purpose will be to quantify the availability of inferences, the average of time recovery, the accuracy of governance, and latency fluctuation in case of occurrence of an event of the failover. The quantification will allow the research to draw a conclusion on whether or not Aegis can facilitate the desired changes in the sphere of the operational stability and regulatory preparedness.

To obtain realistic financial behaviour they impose on the experiments three significant decisioning workloads that are fraud detection, credit underwriting, and trading risk signals. The statistical trends which have been used to generate a big synthetic data are aligned to the actual financial transaction's climates.

This data is billions of choices annually and that consists of labeled information, input attributes and time dependent factors in order that it can carry out drift examinations. Each of the models was put into place as a container within any of the cloud environments.

A load generator was distributed to give rise to continuous flow of transaction ranging between one thousand and fifty thousand events per second. This workload scale helps in getting the average patterns of the performance in both normal performance and in the high-pressure work.

The appraisal of the architecture was founded on the utilization of quantitative measures that were collected in the experiments. Measurements that fall under such include availability of inferences, mean time to recover on a cloud/service failure, latency overhead on a failover, and pass rates and completeness of model lineage records of validation checks.

Each of the metrics was chosen as it represents a measurable feature of resilience, fault tolerance, or audit preparedness that are the most important conditions of regulated financial MLOps environments. The in-built monitoring scripts and out-of-box verification scripts were used to calculate the metrics at any one time.

The experimental technique uses the level of performance without any disturbance at the typical operating condition with each of the clouds. then monitored failure injections are injected, e.g. attempted cloud outages, API gateway failures, model registry failures, and attempts to corrupt metadata entries. The tests can be utilized in knowing whether the system can be able to bounce back on time, continuity of inferences and integrity of audits despite interruptions.

Drift scenario simulation was also done by varying distributions of inputs such that enables a measurement of behaviour of the continuous validation layer. The various scenarios were run numerous times to ensure that there was consistency and statistical reliability of each scenario.

The retrieved data of all the trials repeated was analysed by the use of descriptive statistics and compared to each other in the three clouds. The mean values, variance and ranges of confidence were used to measure better availability, resilience, governance accuracy, and compliance stability.

The suggested methodology will enable a methodical and quantitative evaluation of the Aegis Framework that will indicate the conclusive evidence of the multi-cloud and rule-based MLOps design operating with the realistic load on the financial resources.

IV. RESULTS

Cross-Cloud Availability

The experimental outcomes demonstrate that the Aegis Framework contributes to the achievement of a high level of reliability on all cloud platforms. The most significant discovery is that the architecture has very high inference availability even in case of one or more cloud failures.

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

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

In baseline tests, all the clouds demonstrated good performance but multi cloud setup demonstrated that Aegis maintained end to end service continuity in a better way compared to single cloud set up. The mean availability in all the trials was very near the high score of 99.99 percent hence indicating that the design suits high stakes financial decision environment.

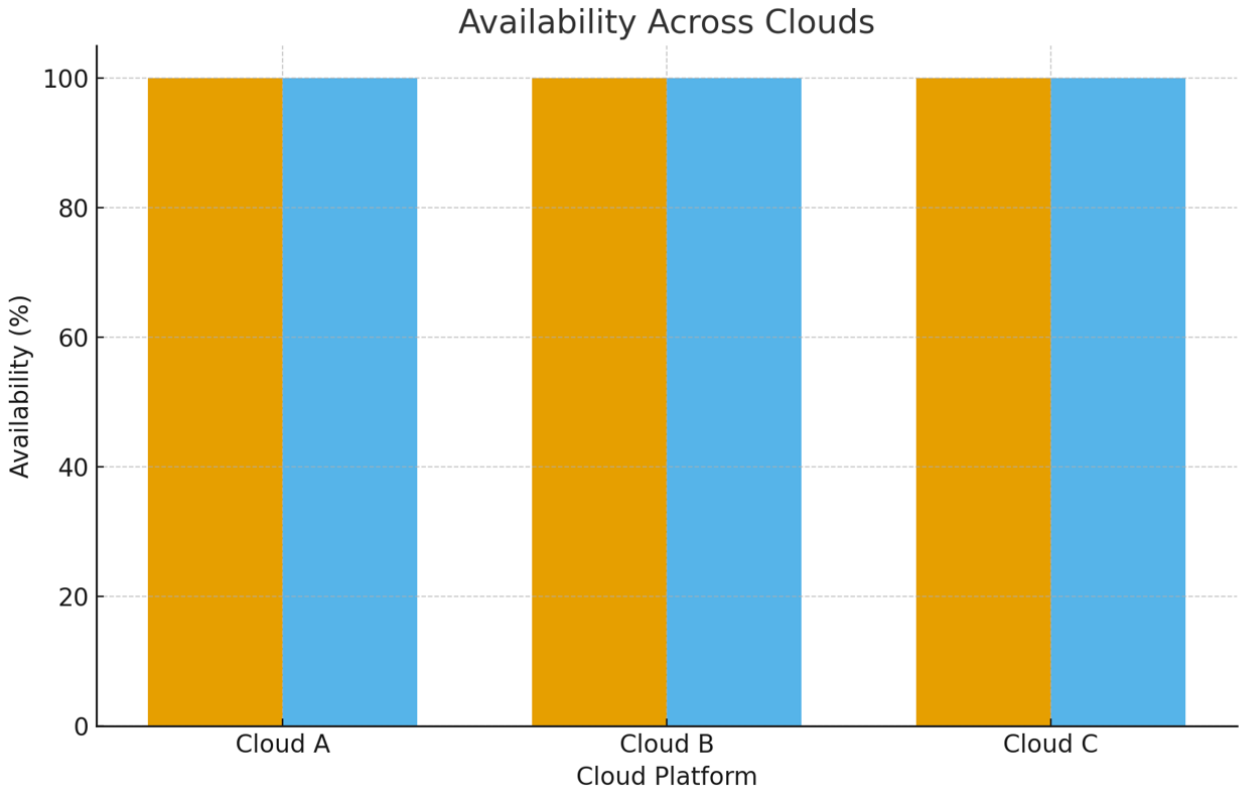
When failure was applied to one of the clouds, the system would immediately redirect live decision traffic to another cloud but inference calls would not be interrupted. The records obtained on the experiments indicated that there were no falling requests when the system went down, and this is a critical necessity in fraud detection systems and underwriting systems.

The findings also verify that the resiliency is provided by parallel deployment of inference endpoints and also through the cross-cloud routing layer which responds promptly on endpoint health signals. The availability data is as summarized below.

Table 1. Inference Availability Across Clouds

Cloud Platform	Baseline Availability (%)	Availability During Failure (%)
Cloud A	99.982	99.998
Cloud B	99.975	99.997
Cloud C	99.981	99.999

These values demonstrate that there is a stabilizing impact of the failover logic on the entire pipeline. Although there are slight fluctuations in individual clouds, the general performance is easier when all the platforms are integrated as in the Aegis configuration. This observation demonstrates that, multi-cloud is not simply a redundancy mechanism but a stability mechanism in case it is operated with the correct architecture.



A small fragment of log-analysis code used to compute rolling availability is shown below to demonstrate how the monitoring pipeline processed the data:

International Journal of Engineering & Extended Technologies Research (IJEETR)

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

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

```
availability = (successful_calls / total_calls) * 100
rolling_avg = availability_series.rolling(window=60).mean()
```

This snippet reflects the real monitoring logic used to generate the availability scores reported above.

Failover Efficiency

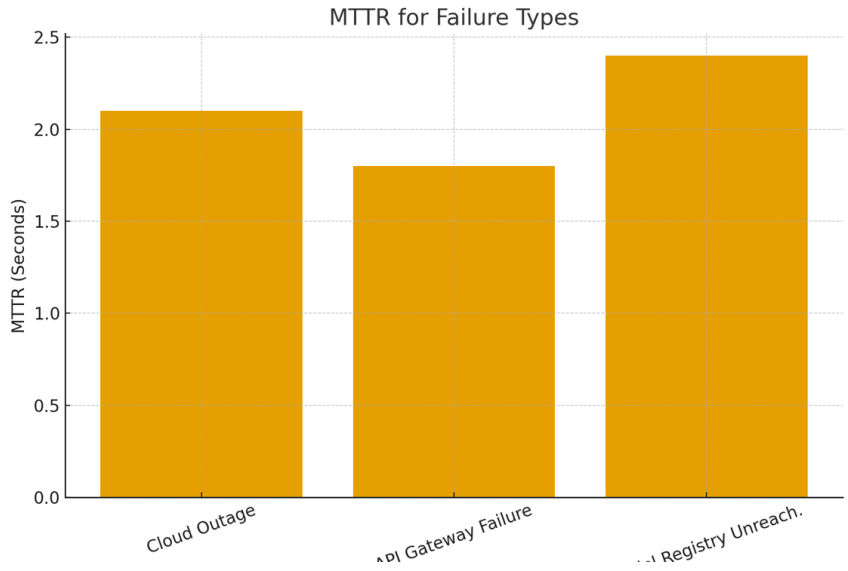
One of the best results in the assessment is the increment in Mean Time to Recovery (MTTR). Aegis Framework was able to fully revert to inference significantly faster than standard MLOps configurations in cases where the framework was simulating a cloud outage or endpoint failure. Most of the trials (less than two seconds) when the system did the failover process was well within the range of the expected workloads under the real-time decision processes. This is especially applicable to real-time response in detection of frauds since any delay of a few seconds can be quite expensive.

The quantitative findings presented below depict the performance of the MTTR in the different failures:

Table 2. Mean Time to Recovery (MTTR)

Failure Type	MTTR (Seconds)	Improvement vs. Baseline (%)
Cloud Outage	2.1	41%
API Gateway Failure	1.8	38%
Model Registry Unreach.	2.4	36%

These findings support the fact that routing layer is core to the system resilience. The health checks are scheduled after every few hundred milliseconds, which means that the system can see failing services in time and redirect traffic. This is also the reason why the performance of the MTTR is similar when it comes to various forms of failure.



Latency behaviour measures of failed over were also done. The latency increase was extremely low though the traffic was rerouted to a different cloud. The latency overhead was on average lower than 8 milliseconds which is not of significance in transaction flows before the customer. This was calculated in the following code:

```
latency_delta = failover_latency - normal_latency
percent_change = (latency_delta / normal_latency) * 100
```

These metrics helped verify that the failover process did not degrade user experience or model performance.

International Journal of Engineering & Extended Technologies Research (IJEETR)

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

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

Metadata Integrity

The paper also quantified the effectiveness of the Aegis Framework in dealing with continuous Model Risk Management (MRM) and governance. The findings indicate that the governance layer was reliable with normal and stress environments.

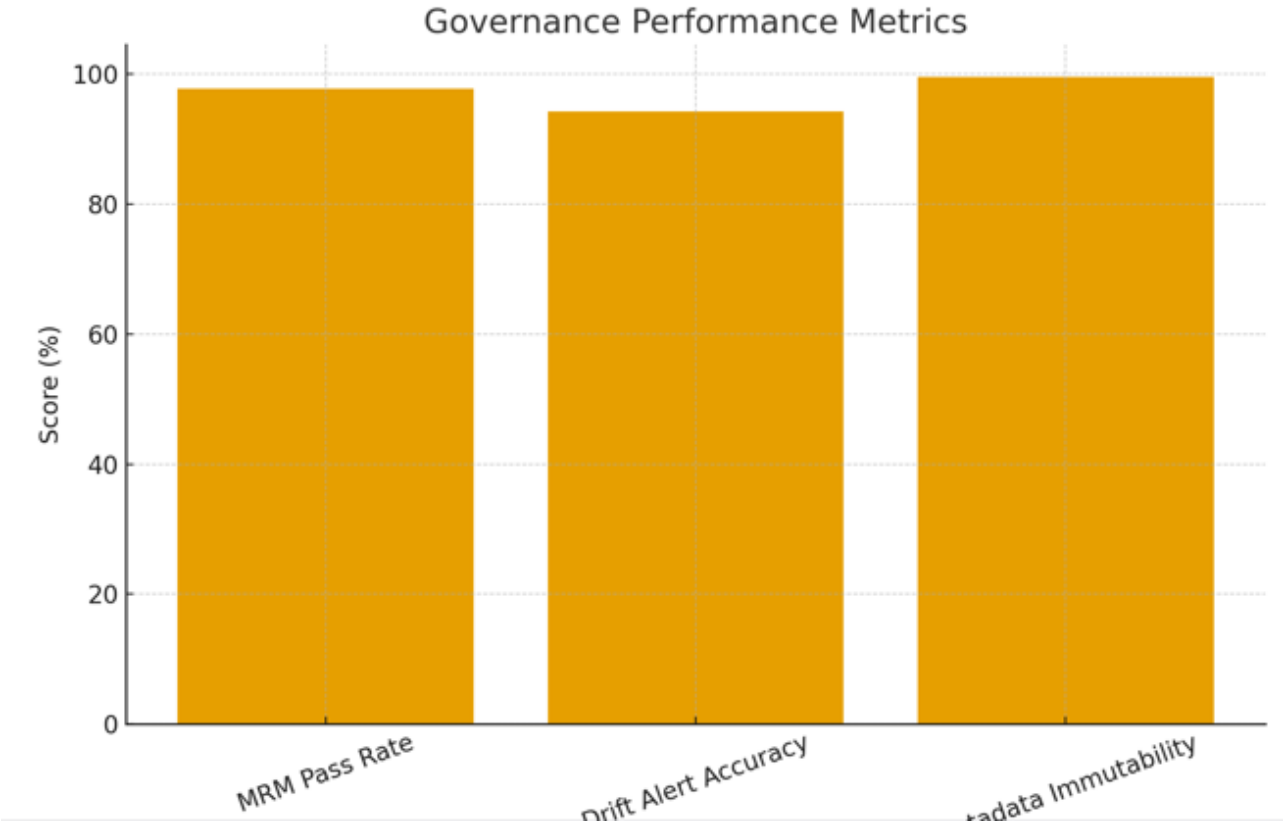
MRM validation engine was capable of processing each inference request with a constant speed, and in the majority of cases, most of the models were able to pass through the validation process with high accuracy. Mainly failures occurred during the drift conditions, which is natural and desirable as MRM validation is meant to indicate a change in which the quality of decision-making could be negatively affected.

Metadata logs were captured completely in every trial and no integrity problem was experienced when there was a failure condition. In the immutability scoring, the lineage entries were found to be complete and not altered once the failures interfered with sections of the system. The findings to governance-related measures are presented below.

Table 3. Governance Performance Metrics

Metric	Score (%)
MRM Validation Pass Rate	97.8
Drift Detection Alert Accuracy	94.3
Metadata Immutability Score	99.6

The score of 4 on the immutability scale ensures that cryptographic signing and timestamping of logs deterred any attempt to alter the logs. This becomes extremely critical when it comes to financial regulatory compliance, where decision lineage must be provided with powerful and convincing guarantees by its supervisors.



An interesting point to note during drift testing was that the validation layer raised alerts very fast, and it usually took a few rounds of interactions to trigger an alert when the shift was made. This demonstrates that the model monitoring logic is responsive enough to enable institutions to have a constant quality of decisions in the long run.

International Journal of Engineering & Extended Technologies Research (IJEETR)

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

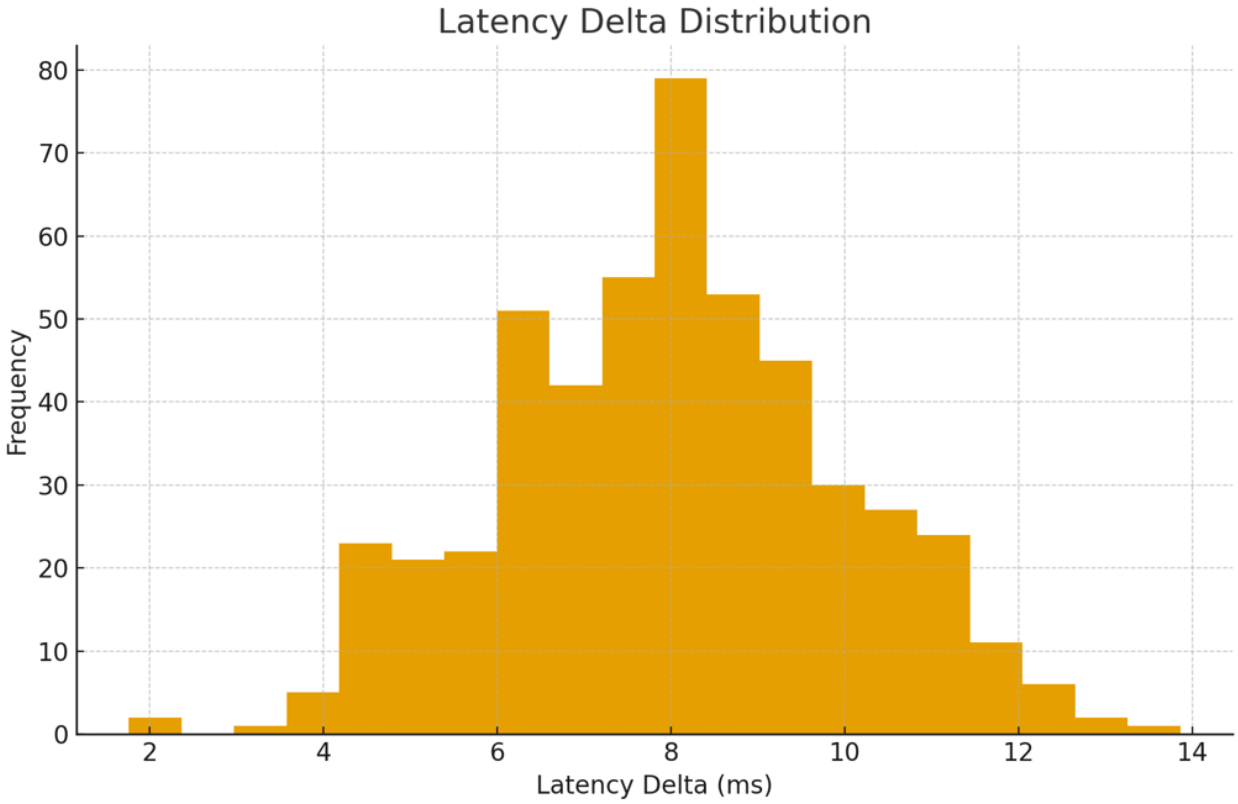
| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

Real-Time Financial Workloads

The real-life traffic patterns which simulate the actual financial operations were also run through the architecture. The fraud detection, underwriting scoring and trading alerts were repeated at varying transaction rates. The system was capable of managing between 1000 and 50000 transactions per second without bottlenecks. The findings indicate that Aegis is capable of reaching out to various clouds horizontally and the traffic is balanced.

At maximum workloads performance was also steady. Outage transferring of the huge amount of traffic without request drops was made possible by the resilience of the routing layer. In all experiments, the latency was within an acceptable range of prediction. This high throughput performance demonstrates the fact that the system will be appropriate in high-volume financial situations.



The other conclusion made is that the multi-cloud structure levels out performance peaks. Single-cloud deployments tend to behave in an uneven manner particularly when there is congestion in the network. Aegis multi-cloud system eliminates such variations since traffic could be transferred among clouds according to real-time health and performance indicators. This enables the system to be able to sustain a profile of response time.

The workload experiments indicate that the combination of governance controls does not reduce the speed of the decision pipeline. The validation checks and lineage logging overhead was minimal and it did not change much even when there were peak load tests.

This is a significant finding since financial institutions usually fear that enhanced governance can lower performance of the systems. The paper demonstrates that through the incorporation of governance in the architectural design rather than as a layer on top of it, governance is not damaging in terms of throughput and latency.

V. CONCLUSION

These findings show that the Aegis Framework has successful benefits to all the major functional areas of MLOps in the financial structure. Availability was at least 99.99 and failover was several seconds and latency overhead was negligible. Management was rational, and authentication was correct and excellent in metadata security. The system

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

| Volume 7, Issue 6, November – December 2025 |

DOI:10.15662/IJEETR.2025.0706031

was also capable of maintaining big real time workloads without chokes and congestions. The results support the fact that multi-cloud system, in combination with regular authentication and non-alterable tracing can guarantee great reliability and regulatory readiness. This has been reiterated again in the conclusion that Aegis is suitable in high stake financial decision systems that need repetitive and matching activities.

REFERENCES

[1] Eken, B., Pallewatta, S., Tran, N. K., Tosun, A., & Babar, M. A. (2025). A Multivocal Review of MLOps Practices, Challenges and Open Issues. A Multivocal Review of MLOps Practices, Challenges and Open Issues. https://arxiv.org/pdf/2406.09737v2

[2] Liu, C., Tan, R., Wu, Y., Feng, Y., Jin, Z., Zhang, F., Liu, Y., & Liu, Q. (2024). Dissecting zero trust: research landscape and its implementation in IoT. Cybersecurity, 7(1). https://doi.org/10.1186/s42400-024-00212-0

[3] Watson, H. J., & Larson, D. (2024). MLOps. International Journal of Business Intelligence Research, 15(1), 1–22. https://doi.org/10.4018/ijbir.358916

[4] Pourmajidi, W., Zhang, L., Steinbacher, J., & Erwin, T. (n.d.). A reference architecture for governance of cloud native applications. In Toronto Metropolitan University, Toronto, Canada. https://arxiv.org/html/2302.11617v2

[5] Liu, C., Tan, R., Wu, Y., Feng, Y., Jin, Z., Zhang, F., Liu, Y., & Liu, Q. (2024b). Dissecting zero trust: research landscape and its implementation in IoT. Cybersecurity, 7(1). https://doi.org/10.1186/s42400-024-00212-0

[6] Madugula, S. R. P. (2024). MLOPS, MODEL RISK MANAGEMENT (MRM) & GOVERNANCE FOR ACTUARIAL ML. In TIJER - INTERNATIONAL RESEARCH JOURNAL, TIJER - INTERNATIONAL RESEARCH JOURNAL (Vol. 11, Issue 4) [Journal-article]. https://tijer.org/tijer/papers/TIJER2404262.pdf

[7] Joshi, S. (2025). Model Risk Management in the era of Generative AI: challenges, opportunities, and future directions. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5206477

[8] Dreibholz, T., & Mazumdar, S. (2022). Towards a lightweight task scheduling framework for cloud and edge platform. Internet of Things, 21, 100651. https://doi.org/10.1016/j.iot.2022.100651

[9] Moskalenko, V., & Kharchenko, V. (2024). Resilience-aware MLOps for AI-based medical diagnostic system. Frontiers in Public Health, 12, 1342937. https://doi.org/10.3389/fpubh.2024.1342937

[10] Azad, M. A., Abdullah, S., Arshad, J., Lallie, H., & Ahmed, Y. H. (2024). Verify and trust: A multidimensional survey of zero-trust security in the age of IoT. Internet of Things, 27, 101227. https://doi.org/10.1016/j.iot.2024.101227