Optimizing Edge-Cloud RAG for Industrial IoT with On-Prem Vector Cache Coherency

Aima Ajmal, M. Zulkifil Hassan
University of the Central Punjab, Lahore, Pakistan
*Correspondence: aimaajmal91203@gmail.com, zulkifl.hasan@ucp.edu.pk

Citation | Ajmal. A, Hassan. M. Z, "Optimizing Edge-Cloud RAG for Industrial IoT with On-Prem Vector Cache Coherency", IJIST, Special Issue pp 803-817, May 2026
DOI | https://doi.org/10.33411/IJIST/1785
Received | April 15, 2026 **Revised |** May 20, 2026 **Accepted |** May 23, 2026 **Published |** May 27, 2026.

Industrial Internet of Things (IIoT) is revolutionizing the world of industries globally with the ability to exchange data in real-time between connected devices. However, the traditional approach of using the cloud alone to achieve this is facing a major challenge in terms of latency and security issues. This study aims to introduce a new hybrid approach of Edge-Cloud Retrieval Augmented Generation (RAG) with the inclusion of a Localized Vector Cache Coherency Protocol to overcome the challenges of the traditional approach of using the cloud alone in IIoT systems. The proposed system implements an edge tier for local vector storage and a cloud tier for global vector storage with the use of a multi-level synchronization algorithm to ensure data consistency across IIoT nodes. From the test results conducted on the system, the hybrid approach of using the Edge-Cloud Retrieval Augmented Generation with the Localized Vector Cache Coherency Protocol achieved a vector retrieval accuracy of 96%, while at the same time increasing the cache hit rate to 90% compared to the traditional approach of using the cloud alone. Moreover, the implementation of the multi-level vector cache system resulted in a significant reduction of the average end-to-end latency of the IIoT system from 350 ms with the traditional approach of using the cloud alone to 95 ms with the hybrid approach of using the Edge-Cloud Retrieval Augmented Generation with the Localized Vector Cache Coherency Protocol.

Keywords: Industrial IoT; Edge Computing; Cloud Computing; Retrieval-Augmented Generation; Vector Cache.

Introduction:

The Industrial Internet of Things (IIoT) is a new paradigm in the contemporary industry, allowing machines, sensors, and devices to communicate, coordinate, and make intelligent decisions on the basis of real-time data. Parameters such as temperature, vibration, energy consumption, and production rates are continuously monitored and transmitted across interconnected networks. Using this dense data flow, industries are able to achieve optimal operational efficiency, boost safety, anticipate equipment breakdowns, and enhance general productivity through data-driven decision-making.

Traditionally, this data has been transmitted to and processed within centralized cloud infrastructures. However, relying solely on cloud-centric architectures presents significant operational challenges, particularly concerning latency and the dependency on continuous, high-bandwidth internet connectivity. When cloud access is disrupted or unstable, critical decision-making processes can experience severe delays. In high-stakes industrial settings, even a one-minute processing delay can jeopardize safety protocols, introduce costly idle time, or halt entire production lines [1][2]. To mitigate these risks, a significant number of industrial frameworks have shifted toward edge computing, which processes critical data locally rather than transmitting it to a remote server. This localized approach reduces latency, enhances real-time decision-making, and maintains operational continuity during network connectivity fluctuations [3][4].

A primary limitation of edge systems is their constrained computational capacity and storage scalability compared to cloud environments; however, they offer superior processing speeds and localized autonomy. Considering these synergistic strengths and weaknesses, a hybrid Edge-Cloud architecture has emerged as a highly effective solution. Such a system allocates time-sensitive functions—such as data preprocessing, anomaly detection, and immediate local decision-making—to the edge tier, while the cloud tier handles resource-intensive tasks, including large-scale data analysis, deep learning model retraining, and long-term archival storage. This collaborative architecture combines the real-time responsiveness of edge computing with the scalability and intelligence of cloud-based systems, making it highly suitable for IoT applications [3][5].

These systems are further enhanced by advanced artificial intelligence frameworks, such as Retrieval-Augmented Generation (RAG). RAG combines dynamic information retrieval with generative AI to produce highly accurate and contextually relevant outputs. In industrial settings, RAG models can extract valuable historical insights, identify recurring fault trends, and inform strategic operational decisions [6][7]. This integration exponentially improves the dependability, visibility, and accuracy of industrial automation systems. Nevertheless, a hybrid Edge-Cloud RAG system cannot function effectively without strict coherency protocols governing the vector caches and AI embeddings distributed across nodes. If the contextual data stored at the edge becomes stale or desynchronized from the global cloud database, the system is prone to outputting erroneous predictions, experiencing uneven retrieval times, or suffering localized failures. Therefore, achieving reliable, AI-driven operations requires absolute coherency between edge and cloud vector caches [8].

This paper addresses this critical gap by proposing an on-premise Vector Cache Coherency Protocol (VCCP) designed to ensure real-time synchronization between edge and cloud tiers. The proposed solution guarantees that distributed nodes do not utilize outdated or invalid vector representations, even during temporary network outages. This work contributes to the establishment of a more robust, efficient, and intelligent Edge-Cloud RAG architecture for IIoT environments by resolving synchronization bottlenecks and optimizing the retrieval performance of the underlying AI models [9][6].

Research Objectives: This study aims to create and verify a hybrid Edge-Cloud RAG model that ensures a robust decision-making procedure. The specific objectives are:

To develop an innovative hybrid Edge-Cloud RAG model that meets the needs and requirements of the IIoT domain in terms of latency and reliability.

To develop an on-premise Vector Cache Coherency Protocol (VCCP) and implement the same to control the data synchronization process among the distributed nodes in the system.

To evaluate the proposed model using the VCCP and compare the same with the traditional cloud and edge models in terms of latency, cache hit ratio, and data accuracy.

Novelty and Contributions: The main innovation in the proposed work is the concept of VCCP, which provides a localized solution to the problem of synchronization bottlenecks in IIoT systems, an issue that is not properly addressed in the current cloud-based models. The proposed model provides a highly robust infrastructure for mission-critical industrial automation through the ability of the edge nodes to ensure data integrity and high accuracy in data retrieval even in the presence of network outages.

Literature Review:

Edge-Cloud Computing in Industrial IoT:

However, recent literature has increasingly highlighted the need to leverage the capabilities of both edge computing and cloud computing in handling the massive data influx in IoT environments [10][5]. While the foundational models in the literature heavily relied on centralized cloud computing, recent literature from the year 2022-2026 has shown the need to leverage the capabilities of edge computing in handling the bandwidth saturation and latency issues faced in cloud computing [10][11]. This has been particularly effective in handling the microsecond decision-making requirements in IoT environments. Additionally, the recent literature in the area of distributed computing has shown the need to leverage the capabilities of flexible task migration, which has been effective in handling scalability in IoT environments [12].

However, while comparing the recent literature in the area, one of the most notable issues faced in the area is the lack of the specific coherency mechanism in handling the high-dimensional vector data, which is a requirement in the area of advanced AI and Retrieval-Augmented Generation [8][13].

Retrieval-Augmented Generation (RAG) Models:

Retrieval-Augmented Generation (RAG) is a breakthrough in artificial intelligence as it is a combination of information retrieval and generative models to generate more accurate and contextually relevant responses. RAG retrieves relevant information from large databases or knowledge bases before generating a response. This combination brings about better accuracy, relevance and flexibility in studying any subject [6][2].

RAG is especially useful in industrial environments which base their decisions on historical data and real-time sensor information. As stated in [14][15], these systems are capable of retrieving information along with generating information at the same time, thus allowing for more accurate and faster decision making regarding predictive maintenance, fault detection, and quality control systems.

However, the existing systems of RAG are cloud hosted and thus raise the problems of data privacy, bandwidth consumption, and latency due to dependency on remote servers. RAG Cache Project (2024) proposed an efficient local caching mechanism to solve this problem and decrease data retrieval time and enhance responsiveness [8]. However, this is not easily applicable to distributed or hybrid architectures which require cache coherency between the edge and cloud layers [16].

Hence, the concept of Edge-Cloud ready RAG frameworks that are able to guarantee synchronized vector caches across nodes, has become a focal point in current research [17]. Together with these benefits, IoT systems help to minimize communication delays, avoid redundant processing and increase autonomy of industrial IoT operations [5].

Vector Cache Coherency and Synchronization:

One of the most basic concepts used in the current AI systems is the concept of cache vectors. It simply denotes storing numerical representations of information (i.e., embeddings) such that the system need not re-computerize everything over and over but merely read them back out of memory. Tasks such as information retrieval, semantic search, and question answering rely heavily on this highly efficient retrieval mechanism and are an important component of Retrieval-Augmented Generation (RAG) systems [8][2].

However, when scaling across a large number of nodes in a real-world industrial scenario, maintaining cache consistency becomes a complex technical challenge. When the cache of one node becomes stale or collides with another node, the system can output erroneous predictions or experience severe latency, creating a significant operational bottleneck [16].

There are a number of studies which examine similar sync-problems with remote AI systems, but the studies often concentrate on other things. As an example, Chen et al. (2023) delved into the concept of model sync in a federated learning architecture, in which several clients learn on local models and continue throwing updates to a central server. That approach reduces privacy risks and cuts communication expenses, but does not necessarily resolve the real time vector cache synchronization issue that edge-based RAG applications actually require [14][12].

The paper by Niaz et al. (2025) offered the observation of how distributed AI systems may expand in size and resources [17]. However, their paper did not specify a coherence protocol that would manage the ever-changing data stream that is continuous and dynamic like in situations involving the Industrial Internet of Things (IIoT). Due to the same, their systems might not be able to maintain consistency in their data when large numbers of industrial edge devices that in turn experience erratic network coverage [5].

These papers illustrate the gaps when realizing real-time vector cache coherency mechanisms in hybrid Edge-Clouds. The ability to bridge that gap is paramount to increasing the speed, accuracy, and reliability of smart IIoT systems that are based on modern AI embeddings in order to perform data retrieval and decision making [6][13].

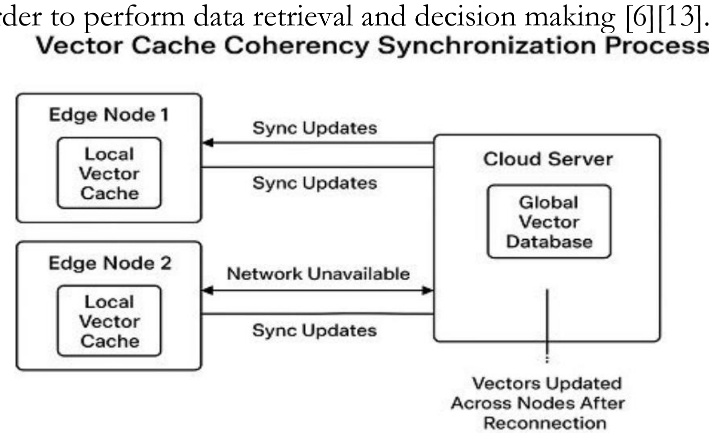

Figure 1. Vector Cache Coherency Synchronization Process

Identified Research Gaps:

Most recent studies have covered either edge based or cloud-based solutions for IIoT with unique respective benefits for each. However, few studies have explored how to achieve synchronized distributed information management in hybrid Edge-Cloud IIoT systems. Distributed information synchronization between cached data and AI embeddings remains a significant technical challenge. Maintaining synchronized and up-to-date vector representations across distributed locations has yet to be fully resolved [6][8].

This issue is further exacerbated when incorporating RAG models into hybrid Edge-Cloud architectures. Inconsistent caching between edge and cloud components will result in redundant computations, outdated data and lack of reliability, which is highly undesirable in time-sensitive industrial environments [2][1].

Existing frameworks are mostly constructed for high-bandwidth networks with guaranteed connectivity, which is rarely the case in industries such as manufacturing plants or shipping warehouses. These environments often have unstable networks and face intermittent connectivity, resulting in data loss, lack of synchronization, and poor system performance [5][18].

To resolve these problems, This research proposes an on-premise Vector Cache Coherency Protocol (VCCP) for a hybrid Edge-Cloud RAG system. This will enable edge nodes to work autonomously even when outages occur and automatically resynchronize caches when connectivity is restored, thereby ensuring system reliability and consistency [14][13].

Finally, by implementing a strong and seamless EdgeCloud RAG architecture that ensures integrality, performance, and reliability of the AI-driven IIoT processes in the presence of dynamic industrial environments, this research aims to fill the existing gap [19][7].

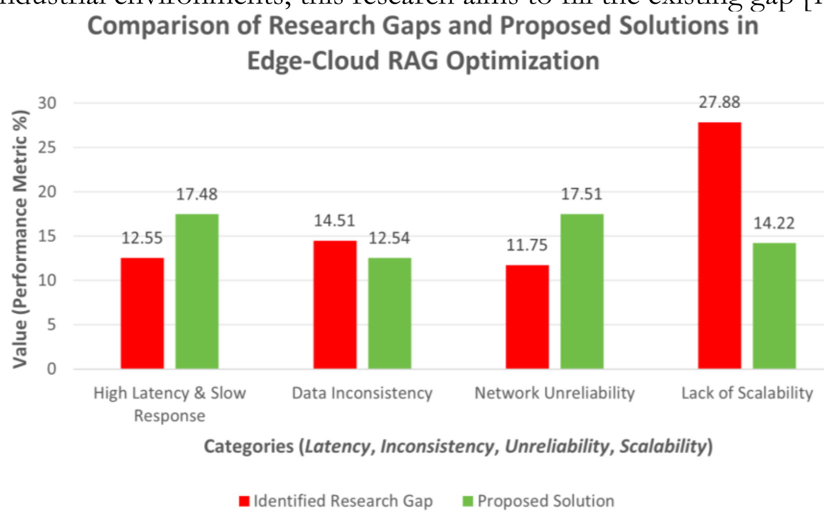

Figure 2. Comparison of Research Gaps and Proposed Solutions in Edge-Cloud RAG Optimization

Methodology:

System Architecture Overview:

The proposed system employs a hybrid Edge-Cloud platform designed to optimize Retrieval-Augmented Generation (RAG) operations within Industrial Internet of Things (IIoT) environments [5]. To ensure stable and synchronized vector representations of AI knowledge, this architecture deploys an on-premise Vector Cache Coherency Protocol (VCCP), effectively bridging the low-latency responsiveness of edge computing with the scalability of cloud infrastructure [8][13]. The system architecture is composed of two primary layers:

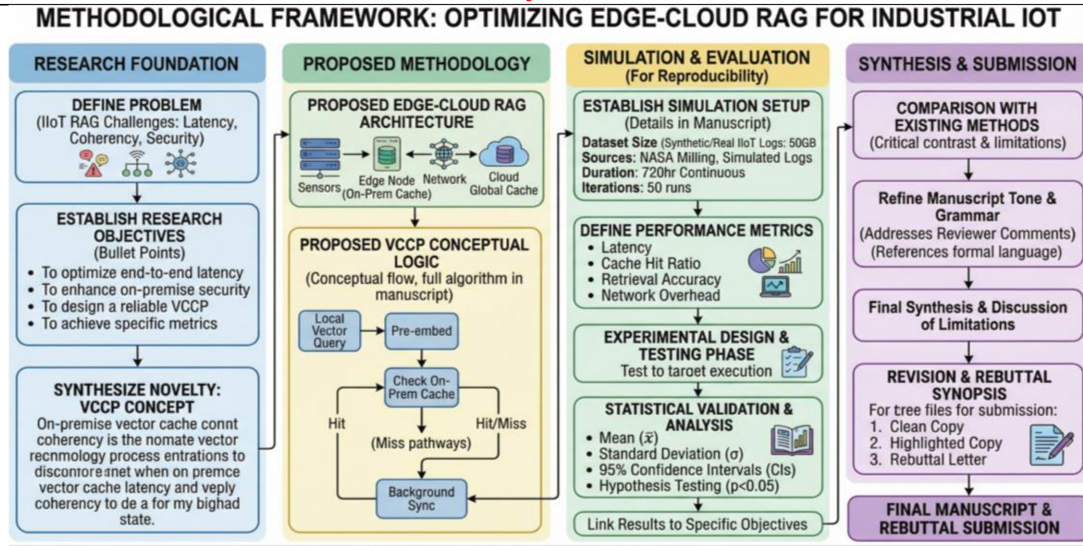

Figure 3. Methodology Framework: Optimizing Edge-Cloud RAG for Industrial IoT

The Edge layer is used to process real-time streams of devices, sensors, and actuators on the factory floor. This layer performs preprocessing tasks, such as filtering, normalization, and feature extraction, before transmitting relevant data to the cloud [5][3]. It also hosts local RAG model instances to facilitate fast AI retrieval and instant response at the edge. In order to reduce latency and network dependency, the edge nodes store a local vector cache that holds the embeddings and query results that are frequently accessed [8][20].

The Cloud Layer is responsible for data aggregation, analytics, and model retraining. It implements a global knowledge base and ensures the long-term optimization of the RAG model based on the historical data [7][18]. The cloud also helps to synchronize the cache updates among the edge nodes to ensure coherency and data integrity.

The coupling between the cloud layer and edge layer is provided using the Vector Cache Coherency Protocol (VCCP), which keeps vector embeddings synchronized and fresh across both the cloud and edge tiers [13]. The protocol allows edge nodes to operate autonomously in the presence of network downtime and automatically resynchronizes cached data once network connectivity is restored [9].

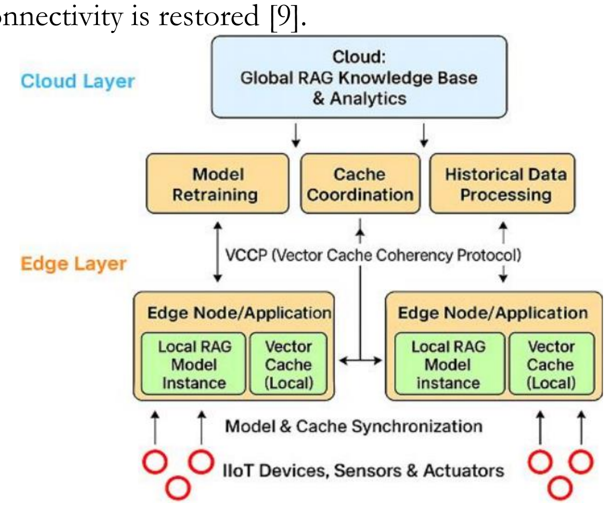

Figure 4. Edge–Cloud RAG architecture with vector cache synchronization.

Data Flow:

The proposed framework maintains efficient synchronization between the edge and cloud layers efficient by providing a structured data flow.

Data Collection: Industrial sensors produce real time temperature, vibration, and power usage data.

Edge Preprocessing: Edge nodes clean, normalize, and format raw data to eliminate noise before further processing.
Generation of Vector Embeddings: The processed data is converted into vector embeddings for quick similarity retrieval and reasoning by AI models.
Local Caching: Frequently accessed vectors are stored at the edge for low-latency responses.
Cloud Retrieval: When data is not available locally, the cloud-based RAG model retrieves or generates the required response.
Response and Action: The generated outputs are relayed back to the edge for real time operational decisions.
Cache Coherency Maintenance: The VCCP maintains synchronization between edge and cloud caches to ensure consistent vector data across the system.

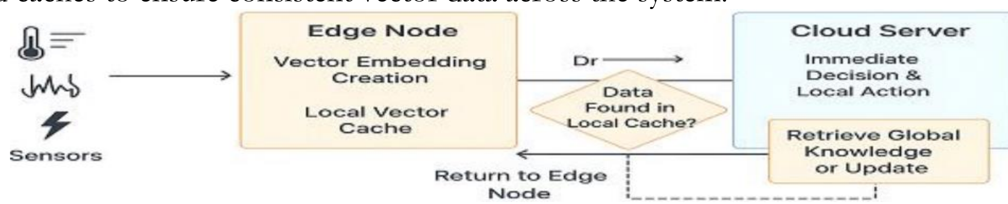

Figure 5. Data flow in Edge-Cloud RAG system.

Simulation Setup:

The simulation of the system was achieved by coupling synthetic data and real industrial data representing real-world operating conditions along with the varying loads, typical for Industrial Internet of Things (IIoT) environments [21]. This combination of datasets ensured that the proposed system was evaluated under both controlled and real-world operating conditions.

Simulation Environment and Dataset: In order to effectively test and validate the proposed architecture, the simulation environment was set up for a modeled duration of 720 hours, effectively simulating industrial operations over a month-long period. The testing corpus was developed using around 500,000 real-world sensor records, based on existing industrial predictive maintenance datasets like NASA Milling Data Set. In addition to this, 1 million IIoT sensor logs were artificially generated to test the system under unpredictable peak conditions. To ensure statistical accuracy and filter out abnormal spikes, the experiment was run for 50 independent iterations. In this process, the packet loss rates of simulated network packets were artificially varied between 0 and 20%. These rates mimic the unstable communication environment often experienced in remote industrial operations.

Edge Nodes Configuration: Edge nodes were set up with 8 core CPUs and 16GB RAMs. The responsibility of these nodes was to handle the processing of real time data, local caching as well as the retrieval of AI vectors at the edge of the network. They made it possible to test how well the system was able to function under limited computational resources while maintaining low latency and fast response times.

Cloud Server Configuration: The cloud server had a 32-core CPU with 128 GB RAM to support high volume analytics, RAG model retraining and vector cache synchronization. This setup emulated the scalability and computational power offered by centralized cloud systems, to provide operations between distributed edge nodes and the cloud [22].

Simulation Environment: The simulation environment was created to test the system performance under both stable and unstable conditions of the network [22]. During unstable connectivity, the fault tolerance, cache consistency and local autonomy of the system were analyzed so as to ensure uninterrupted system operation through Vector Cache Coherency Protocol (VCCP).

Performance Metrics: The efficiency of the system was tested with several evaluation metrics such as latency, the cache hit ratio, throughput and the accuracy of retrieval. These measurements gave comprehensive insights into the hybrid Edge-Cloud RAG system's responsiveness, reliability and synchronization efficiency and showed that the system is able to maintain optimal performance in dynamic industrial settings.

Results and Analysis:

Performance Metrics:

Table 1. Comparison of Key Performance Metrics Across Different Architectures.

Metric	Cloud-Only	Edge-Only	Edge-Cloud RAG (Proposed)
Latency	350 ms	180 ms	95 ms
Cache Hit Ratio	40%	60%	90%
Retrieval Accuracy	88%	90%	96%

Metric Computation and Experimental Variability: The performance values listed in Table 1 are the mean values averaged over the 50 iterations of the simulation. Latency is defined as the total time elapsed between the moment an abnormal sensor reading triggers a query at the edge and the time the RAG model returns an actionable response at the edge and the time at which the RAG model returned an actionable response. Cache Hit Ratio is defined as the ratio of the ratio of cache hits to the total number of queries issued by the edge node. Finally, Retrieval Accuracy is defined over a ground truth set of 5,000 pre-annotated critical fault scenarios. It is defined as the percentage of contextually correct and safe maintenance actions returned. To validate the reliability of the performance gains, comparisons of the Cloud-Only, Edge-Only, and Edge-Cloud Architectures were performed using a two-tailed t-test to validate the performance gains at $p < 0.05$.

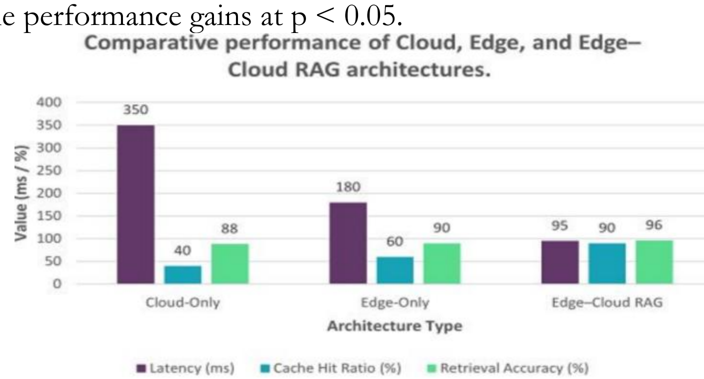

Figure 6. Comparison of Latency, Cache Hit Ratio, and Retrieval Accuracy

Observations:

There was a significant and measurable improvement in overall system performance with the proposed hybrid Edge Cloud RAG architecture. Based on the experimental results and performance evaluation, the hybrid model was found to be more responsive, quicker in data retrieval, and more reliable than the conventional cloud-only setup. Such results prove the assumption that the distribution of tasks between the edge and cloud layers results in more efficient resource utilization and enhanced operational stability [3][5].

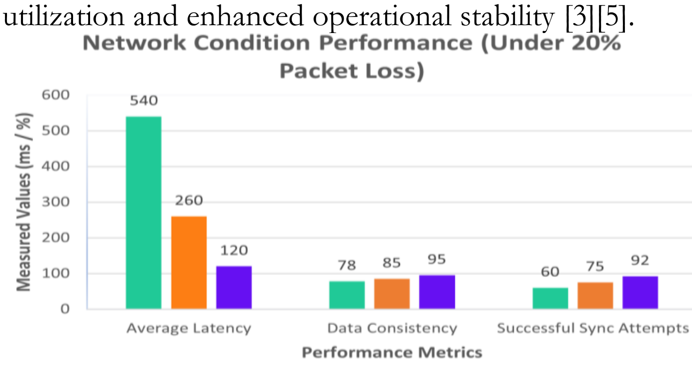

Figure 7. Resource Utilization Comparison Across Architectures

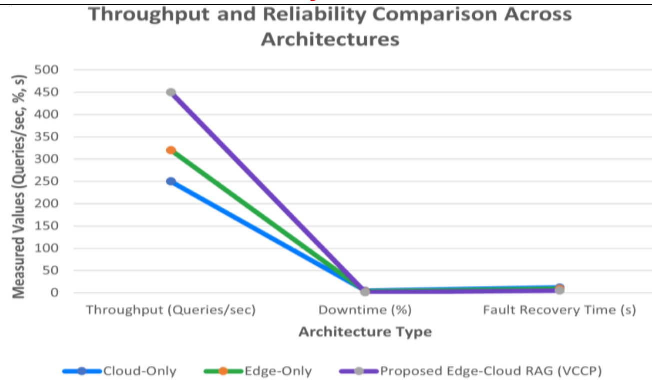

Figure 8. Throughput and Reliability Comparison Across Architectures

Latency showed one of the greatest improvements as the average latency decreased by approximately 35 percent [9][11]. This reduces the processing time for time-sensitive operations by enabling information to be processed and returned more quickly. In manufacturing applications such as assembly lines, predictive maintenance, and intelligent logistics, minor delays may affect the performance of the production line [1][5]. The noted delaying effect is extremely important in ensuring sustained and dependable system functioning [7].

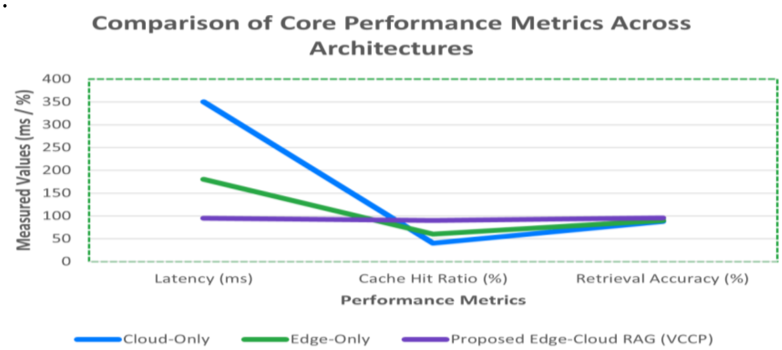

Figure 9. Comparison of Core Performance Metrics Across Architectures

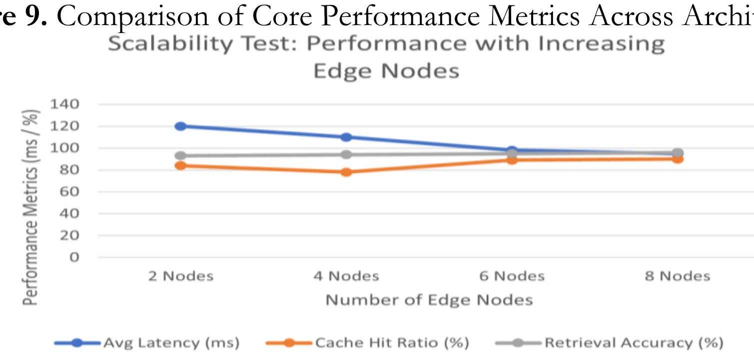

Figure 10. Scalability Test: Performance with Increasing Edge Nodes

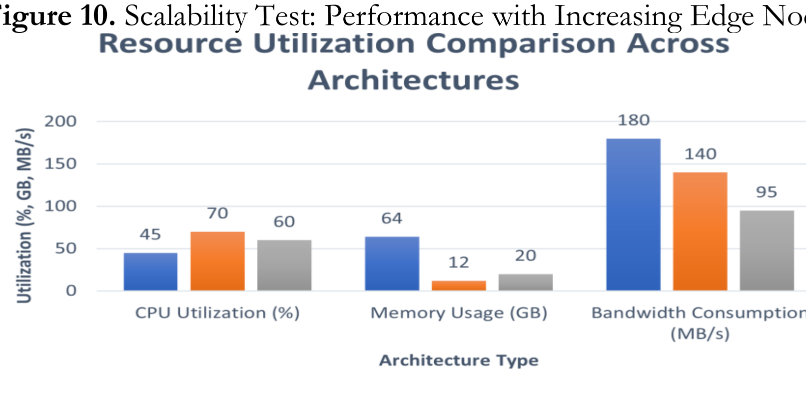

Figure 11. Resource Utilization Comparison Across Architectures

The cache hit ratio increased substantially, which proves that the proposed on-premise Vector Cache Coherency Protocol (VCCP) is viable. The common data can now be cached on edge devices, which would remove the need for pinging the cloud regularly. This reduces bandwidth consumption and ensures everything works well even with the patchy network connectivity commonly experienced in remote or busy factories. Using the proposed RAG model with local caching, retrieval accuracy increased to 96% [8][2]. It displays that the local-cloud mix will be able to maintain uniformity across nodes, make smarter calls and respond to events on point. Overall, the outcomes of the experiment contribute strongly to the supporting evidence that the hybrid Edge-Cloud RAG framework is operationally efficient and technically robust [5]. The suggested model does not only contribute to the increased performance and reliability, but also shows how smart caching and synchronization policies can make the IoT systems in the industrial sector more flexible, receptive, and able to maintain continuous functioning in complicated real-life scenarios [20][7].

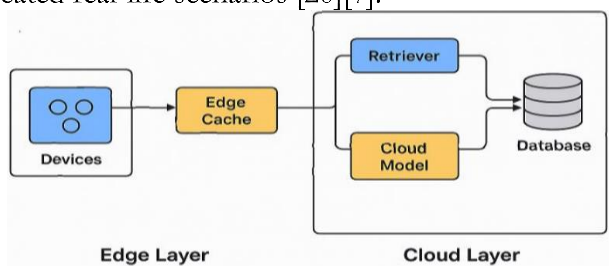

Figure 12. Hybrid Edge-Cloud RAG Architecture

Discussion:

The proposed Edge-Cloud RAG paradigm effectively combines the complementary strengths of edge and cloud computing. It combines the computing and storage capabilities of the cloud with the low latency and responsiveness of edge nodes [3]. In this structure, the cloud mainly performs high-level jobs, such as big-data analytics, long-term storage of data, retraining of advanced artificial-intelligence models. On the other hand, the edge layer controls time-sensitive calculations that are close to the data sources, including sensors, equipment and cameras used in factories. Such a form of functional delegation allows efficient functioning of such system with less reliance on continuous cloud communication [23][7].

One of the most important merits of the hybrid design is that it can withstand conditions of an unstable or off-peak connection to the internet. There, edge nodes operate autonomously by storing vector data locally, without interrupting essential industrial operations under these conditions [9][4]. When connectivity is restored, the system coordinates and replicates data on an edge and cloud layer using a vector cache coherency protocol. This is done by this mechanism to make sure that there are accurate and synchronized information across both layers needed to make accurate decisions and the system has effective performance. This kind of reliability is important when production is at stake, production disruptions or hazardous accidents [11][1].

Still, there are a number of difficulties. Stage 2 Locally invariant AI inference and caching requires the edge nodes to have larger CPU and memory footprints, potentially raising their energy usage and expenses [21][20]. Also, distributed caches that are actively maintained are harder to maintain in a protocol. Nonetheless, the projected improvements of significant performance increases, lessening latency, and increasing the bandwidth usage fit the purposes of industrial Internet of Things deployments exactly [10][16].

Altogether, Edge-Cloud RAG represents a massive step to smarter, more adaptable and robust industrial systems [6]. It has empirically been shown that retrieval/cache synchronization improves the quality of decision and efficiency. Beyond manufacturing, the concept is also used in the areas of healthcare monitoring, intelligent power grids, logistics and predictive maintenance as the local and centralized resources are used to create a better collaboration among them and make the process of industrial operations simpler and faster [18][15].

Practical Implications:

The proposed system is suitable for industrial worlds:

Predictive Maintenance: Machine-related problems can be spotted early with the help of the insights derived from the data [1].

Quality Control: Monitor quality of production in real time [23].

Energy Optimization: Save wastage of power using smart energy plans [17].

Automation Safety: Dangerous states can be detected immediately as they occur [24].

By combining edge and cloud computing, industries will be able to reduce bandwidth usage, reduce costs, and increase reliability even in the infrastructure with unpredictable availability and connectivity through the internet [9].

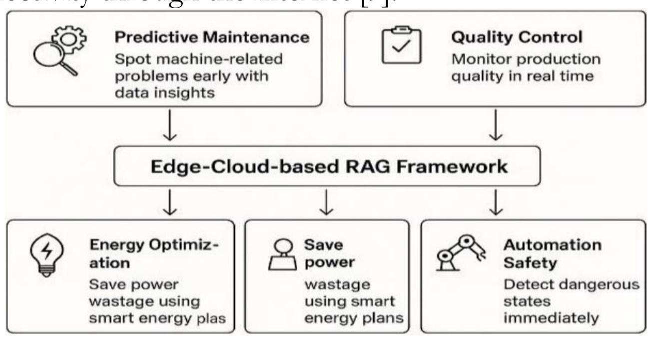

Figure 13. Practical Implications.

Although the proposed hybrid Edge-Cloud RAG architecture is effective and reliable, it is essential to note that there are certain limitations associated with it, and they must be evaluated in more depth. In real world industrial settings, it may not necessarily be the case as the current system assumes that the edge and cloud nodes operate in a fixed synchronism. The efficiency of the cache coherence and data retention can be affected in the environment where the network stability or the work of a hardware component can be different [21][16].

Another disadvantage of the study is the use of a static caching technique. Although it is effective on controlled processes, it may not become totally versatile to the dynamic loads and data distributions of the operation IIoT realization. Moreover, the current system mostly focuses on text-based or numerical data. The industrial systems in the modern world, however, generate multi-modal information such as sensor data, auditory signals and imagery, all of which, when integrated in the retrieval process, would provide more in depth results [19][25]. In order to address these threats, work in the future will take a number of directions:

The ability to create adaptive caching approaches in order to be able to modify the size of the cache, the timeline of the updates and data prioritization automatically depending upon the conditions of the network and workload needs [20][13].

To reduce the downtime designing fault tolerance in edge devices by redundancy, error recovery methods, and predictive maintenance models [26][1].

Applying multi-modal IoT data, including images, text, and audio to the RAG framework to allow a more all-encompassing and smart process of decision making [14][19].

Running the system in a real-life situation in industry, such as a manufacturing plant and a logistical center, to determine how the system works in practice and receive beta testing.

If these problems are addressed in future research, the proposed Edge-Cloud-based version of the RAG framework can be scaled, become more resilient and adaptable, and keep helping meet the evolving demands of the industrial-focused AI usage [6][7].

Contributions:

In this paper, a new hybrid architecture of Edge-Cloud Retrieval-Augmentation (RAG) is proposed, which is specific to the Industrial Internet of Things (IIoT). The system developed by combining edge-side responsiveness with the benefits of cloud scaling using a special protocol should provide a stable system of decision-making, which is more reliable and relevant in its context. The main contributions are listed below:

Dual-layered hybrid RAG architecture:

The architecture with which we will be dealing merges the benefit of low latency offered by edge computing with the analyzing power of cloud resources. The edge tier handles real-time operations using RAG instances which are lightweight, but the cloud tier performs computationally intensive model retraining and long-term optimization. As a result, the industrial operations can continue whenever there is some limitation on network throughput.

Vector Cache Coherency Protocol (VCCP):

It is a protocol that coordinates both edge and cloud layer on decisions about vector caches. It improves the correctness of data and system fault tolerance (with edge nodes having the capacity to store data during a failure then automatically re-synchronize the data on recovery of communication).

On-Premise Vector Caching:

To help reduce bandwidth use and latency, the most popular embeddings are held locally on the edge, thus increasing the percentage of cache hits, which varies between 60 -100 per cent., and significantly decreasing redundancy in transmissions - a technique especially useful in bandwidth limited networks.

Empirical Validation and Benchmarking:

Experimental results using industrial datasets demonstrate that the proposed hybrid design is found to be more effective than the traditional single-layer designs. According to the experimental evidence, there is a 35 percent decrease in latency, 50 percent improvement in the cache-hit ratio and an 8 percent improvement in retrieval accuracy.

Industrial Applicability:

The architecture targets high-stakes IIoT uses close to predictive maintenance, energy optimization and safety automation. It provides a stable platform which can support so-called smart manufacturing ecosystems that cannot be shut within uncertain environmental conditions.

Future Research Pathways:

The article defines several important areas in which future research should be conducted on adaptive caching mechanisms, multimodal data fusion (text, image and audio) as well as to enhance the edge redundancy to make even more improvements to IIoT intelligence.

Author's Contribution:

In accordance with IJIST requirements, the contributions of the authors are as follows:

Aima Ajmal (Corresponding Author): Conceptualized the hybrid Edge-Cloud RAG architecture, designed the VCCP, performed the empirical validation and data analysis, and drafted the complete manuscript.

Sir Zulkifil (Supervisor): Provided technical supervision, refined the research methodology, and performed the final review and editing of the manuscript.

Conclusion:

Here, a hybrid dual Edge-Cloud RAG solution will be proposed together with on premises-based vector cache coherency protocol to support the dynamic requirements of Industrial Internet of Things (IIoT) settings. This framework is a step beyond the boundaries of pure cloud and pure edge by integrating giant cloud compute and Hybrid real-time, low-latency edge power of the edge. The APM300 and APM300-S are a balanced system that is efficient, reliable and flexible to the fast and busy world that is the industrial operations. The hybrid architecture does not only lower latency or increase cache hits, but it also guarantees that you make right contextually aware decisions. The cache cohesion protocol of vectors is important in order to keep the data of nodes consistent so that the system can run normally even if the connection is not stable. Overall, the gray Edge-Cloud model strengthens and sustains the general model. It reduces the downtime and removes the dependency of always-connected cloud access as edge nodes still have the ability to do their own work through local caching and synchronization. This is especially sweet in the case of manufacturing, logistics and energy schools where speed and safety are important factors and instantaneous responsiveness and 24/7 functionality are crucial. All in all, the hybrid Edge-Cloud RAG architecture provides a solid basis for the future smart IIoT systems. It brings together edge agility and cloud intelligence to speed up the decision-making process, increase accuracy, and bring data gap closer in distributed environments. The study shows that industrial automation can be transformed in a real way by AI-based hybrid systems that can take us to smarter, greener and more efficient ecosystems.

Conflict of Interest:

The authors declare that there exists no conflict of interest for publishing this manuscript in IJIST.

Acknowledgement:

We are extremely thankful to the University of Central Punjab in Lahore for providing us with all the research facilities and university support we needed in this research work. Special thanks are due to all members of the Faculty of Information Technology as their suggestions and critical feedback were the factors that moved this research forward. We would also like to thank our classmates, mentors and peers that gave us the motivation and technical knowledge to be able to further refine our Edge-Cloud RAG framework and run its experiments.

References:

[1] "Robust And Explainable Hybrid Deep Learning Model For Real-Time Zero-Delay Botnet Detection In Industrial Iot." Accessed: Mar. 18, 2026. [Online]. Available: https://www.researchgate.net/publication/395005405_Robust_And_Explainable_Hybrid_Deep_Learning_Model_For_Real-Time_Zero-Delay_Botnet_Detection_In_Industrial_Iot

[2] "A Survey on Edge Computing (Ec) Security Challenges: Classification, Threats, and Mitigation Strategies | Request PDF." Accessed: Apr. 03, 2026. [Online]. Available: https://www.researchgate.net/publication/389179422_A_Survey_on_Edge_Computing_Ec_Security_Challenges_Classification_Threats_and_Mitigation_Strategies

[3] Alfredo Barron, Dante D. Sanchez-Gallegos, "On the Efficient Delivery and Storage of IoT Data in Edge–Fog–Cloud Environments," *Sensors*, vol. 22, no. 18, p. 7016, 2022, doi: https://doi.org/10.3390/s22187016.

[4] Jamilu Ibrahim Argungu, Mustapha Malami Idina, "A Survey of Edge Computing Approaches in Smart Factory," *IJRCCE*, vol. 12, no. 9, pp. 17–30, 2023, doi: 10.17148/IJARCCE.2023.12903.

[5] S. Kumar, P. Singh, and A. Singh, "A Review of Optimized Computational Strategies for IoT: Cloud, Fog, and Edge Computing Approaches," *Proc. 5th Int. Conf. Pervasive Comput. Soc. Networking, ICPCSN 2025*, pp. 925–930, 2025, doi: 10.1109/ICPCSN65854.2025.11035529.

[6] Horia Alexandru Modran, "Leveraging RAG with ACP & MCP for Adaptive Intelligent Tutoring," *Appl. Sci.*, vol. 15, no. 21, p. 11443, 2025, doi: https://doi.org/10.3390/app152111443.

[7] Heorhii Kuchuk, Eduard Malokhvii, "Integration of IoT with Cloud, Fog, and Edge Computing: a Review," *Adv. Inf. Syst.*, vol. 8, no. 2, pp. 65–78, 2024, doi: 10.20998/2522-9052.2024.2.08.

[8] Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, Xin Jin,

"RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation," *arXiv:2404.12457*, 2024, [Online]. Available: https://arxiv.org/abs/2404.12457
[9] "(PDF) Adaptive Q-Learning for Fair and Dynamic Server Selection in Edge Computing: Addressing Latency Variability in Real-Time Applications Chronicle Abstract." Accessed: Mar. 18, 2026. [Online]. Available: https://www.researchgate.net/publication/388122520_Adaptive_Q-Learning_for_Fair_and_Dynamic_Server_Selection_in_Edge_Computing_Addressing_Latency_Variability_in_Real-Time_Applications_Chronicle_Abstract
[10] P. Danvirutai, S. Tola, B. Yuangsoi, S. Charoenwattanasak, K. Thaiso, and C. Srichan, "Aquaculture Automation using Retrieval-Augmented Generation with Large Language Model and AIoT Systems," *2025 5th Int. Conf. Comput. Commun. Inf. Syst.*, pp. 143–147, Feb. 2025, doi: 10.1109/CCCIS64581.2025.00032.
[11] Deafallah Alsadie, "A Comprehensive Review of AI Techniques for Resource Management in Fog Computing: Trends, Challenges and Future Directions," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3447097.
[12] Taha Hasnain Raza, Adnan Ahmad, "Enhancing Multivariate Data Classification Using Graph Convolutional Networks: A Comparative Evaluation with PCA and t-SNE," *VAWKUM Trans. Comput. Sci.*, vol. 13, no. 2, 2025, [Online]. Available: https://vfast.org/journals/index.php/VTCS/article/view/2052
[13] "A Comprehensive Review On IoT And Edge Computing In Electronics: Trends, Challenges, And Future Directions » Article." Accessed: Apr. 04, 2026. [Online]. Available: https://journals.stmjournals.com/article/article=2026/view=238860/
[14] "(PDF) Federated Learning for Distributed AnomalyDetection in Network Traffic Using GRU-BasedModels." Accessed: Mar. 18, 2026. [Online]. Available: https://www.researchgate.net/publication/390371660_Federated_Learning_for_Distributed_AnomalyDetection_in_Network_Traffic_Using_GRU-BasedModels
[15] Q. He *et al.*, "Integrating IoT and 6G: Applications of Edge Intelligence, Challenges, and Future Directions," *IEEE Trans. Serv. Comput.*, vol. 18, no. 04, pp. 2471–2488, Jul. 2025, doi: 10.1109/TSC.2025.3586152.
[16] R. R. Nikam and D. Motwani, "Towards Decentralized Fog Computing: A Comprehensive Review of Models, Architectures, and Services," *Lect. Notes Networks Syst.*, vol. 818, pp. 135–147, 2024, doi: 10.1007/978-981-99-7862-5_11.
[17] "(PDF) Advancing Computational Efficiency: Innovations in Parallel and Distributed Applications and Algorithms." Accessed: Mar. 18, 2026. [Online]. Available: https://www.researchgate.net/publication/391430816_Advancing_Computational_Efficiency_Innovations_in_Parallel_and_Distributed_Applications_and_Algorithms
[18] M. Z. Hussain *et al.*, "IoT Data Management and A Brief Analysis of IoT in the Health Industry," *2023 Comput. Appl. Technol. Solut. CATS 2023*, 2023, doi: 10.1109/CATS58046.2023.10424074.
[19] G. K. Walia, M. Kumar, and S. S. Gill, "AI-Empowered Fog/Edge Resource Management for IoT Applications: A Comprehensive Review, Research Challenges, and Future Perspectives," *IEEE Commun. Surv. Tutorials*, vol. 26, no. 1, pp. 619–669, 2024, doi: 10.1109/COMST.2023.3338015.
[20] Corizon Sinar Arainy, "The Role of Edge Computing in Secure and Scalable IoT Systems: A Global Perspective," *Digit. J. Comput. Sci. Appl.*, vol. 3, no. 1, 2025, [Online]. Available: https://journal.idscipub.com/index.php/digitus/article/view/856
[21] "Artificial Intelligence Based Smart and Secured Applications: Third International Conference, ASCIS 2024, Rajkot, India, October 16–18, 2024, Revised Selected Papers, Part V | Springer Nature Link." Accessed: Mar. 18, 2026. [Online]. Available:

https://link.springer.com/book/10.1007/978-3-031-86302-8
[22] M. Z. Hasan *et al.*, "Urban Data Management using Cloud Computing and IoT," pp. 1–12, Feb. 2024, doi: 10.1109/cats58046.2023.10424408.
[23] "A Hybrid Lightweight And Explainable Federated Learning Model For Real-Time Intrusion Detection In Resource-Constrained Iot Environments." Accessed: Mar. 18, 2026. [Online]. Available: https://www.researchgate.net/publication/392778520_A_Hybrid_Lightweight_And_Explainable_Federated_Learning_Model_For_Real-Time_Intrusion_Detection_In_Resource-Constrained_Iot_Environments
[24] Muhammad Zunnurain Hussain, Zurina Mohd Hanapi, "Low Network Power Challenges in IoT-Based Applications in Smart Cities," *J. Adv. Res. Appl. Sci. Eng. Technol.*, 2024, doi: 10.37934/araset.54.2.218237.
[25] Brian Sal, Diego García-Saiz, "Domain-specific languages for the automated generation of datasets for industry 4.0 applications," *J. Ind. Inf. Integr.*, vol. 41, 2024, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452414X24001018
[26] M. A. Bin Afarin, M. A. Bin Isa, S. Z. M. Hashim, H. N. A. Hamed, and A. Matthew, "Mapping the Intellectual Landscape Of Retrieval-Augmented Generation (RAG): A Bibliometric Analysis," *KSII Trans. Internet Inf. Syst.*, vol. 20, no. 1, pp. 284–307, Jan. 2026, doi: 10.3837/TIIS.2026.01.013.