Journal of Information Systems Engineering and Management
2025, 10(16s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Microservices-Based Hybrid Cloud-Edge Architecture for Real-Time IIoT Analytics

Venkata Srinivas Kompally
Northeastern University, kompally.v@northeastern.edu

ARTICLE INFO

Received: 04 Dec 2024

Revised: 25 Jan 2025

Accepted: 12 Feb 2025

ABSTRACT

Hybrid and multi-cloud computing have become very important for enterprises and companies managing large-scale, real-time data streaming and analytics. At the same time, edge computing has come into the limelight as a key approach to reducing latency and optimizing resource usage at the network's outermost layer.

This paper bridges the gap between scalable multi-cloud architectures and modern microservices-driven cloud-edge collaboration which solves real-time streaming, analytics, and condition monitoring. We first summarize the major limitations—such as data integration, latency, interoperability, and vendor lock-in—when designing solutions that span in different cloud environments.

We then provide a hybrid architecture that integrates edge computing to guarantee quick response and uses microservices for containerized deployments. For activities like feature extraction and AI-driven analytics, we will investigate cloud-edge collaboration solutions that will allow for instant decision-making at the edge while also shifting more complicated processing to the cloud.

Lastly, we present a case study on battery scanning and quality analysis in battery manufacturing. Findings show that integrating multi-cloud and edge computing can reduce operating expenses, cut latency by up to 30%, and greatly increase predictive accuracy, with a 90% success rate in real-time anomaly detection and a 50% reduction in predictive mistakes.

Finally, this hybrid cloud-edge strategy positions itself as a strong framework for the upcoming generation of intelligent applications by improving scalability, real-time efficiency, and cost-effectiveness.

Keywords: Hybrid Cloud, Microservices, Multi-Cloud, Industrial Internet of Things, Cloud-Edge Computing, Integration, Scalability.

1. Introduction

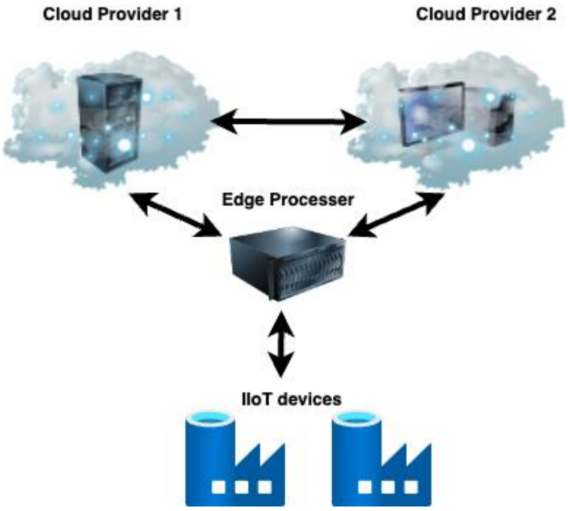

Figure 1: Microservices-Based Cloud-Edge Framework for IIoT

Digital strategies across most of the sectors have continuously changed over the last ten years because of advancements in analytics, scalable cloud computing solutions, and the Industrial Internet of Things (IIoT). To promote and increase innovation, major industries including manufacturing, shipping, finance, and healthcare today mostly depend on immediate access to massive data streams. The increasing amount of data and the demand for ultra-low-latency applications, such as condition monitoring and real-time control, are moving toward edge computing and hybrid or multi-cloud architectures, even though cloud computing has first made large-scale, centralized data processing is also feasible.

In order to balance data security, compliance, and scalable resource allocation, hybrid cloud combines private and public cloud infrastructures. Multi-cloud solutions, on the other hand, divide workloads between several public cloud providers to reduce vendor lock-in, maximize expenses, and benefit from specialized services offered by cloud providers, because each cloud provider might specialize in a particular type of service.

In contrast, edge computing reduces latency and network costs by processing and storing resources closer to the location of data generation near IIoT devices. At the same time, the desired architecture for creating modifiable, portable applications in edge and cloud contexts is microservices, which are small, containerized software components.

In smart manufacturing, microservices-based cloud-edge collaboration has recently accelerated because of the requirement for reliable defect detection, low-latency analytics, and real-time process management. By breaking down the traditional monolithic or virtualized services into containerized microservices, manufacturers and major enterprises can achieve seamless updates, faster fault recovery, and more efficient use of distributed computing resources.

This paper bridges links the best practices in multi-cloud adoption with new frameworks that integrate edge computing and microservices-based architectures, emphasizing on real-time data streaming and analytics in intelligent environments.

1.1. Problem statement

Enterprise companies are collecting massive amounts of data streams from a wide range of sources, including industrial sensors, IoT devices, and user-facing applications. However, when deploying real-time analytics pipelines, some challenges are:

1. **Data Integration:** Inconsistent data formats and APIs across multiple cloud platforms which makes it difficult to simplify analytics pipelines, often delaying real-time data processing
2. **Scalability & Latency:** Without adding additional network latency, compute and storage resources must scale well across geographically dispersed locations
3. **Interoperability & Vendor Lock-in:** Workload portability and smooth integration are limited by proprietary API, special cloud features, and the complicated nature of multi-cloud orchestration
4. **Edge Responsiveness:** Ultra-low-latency applications—especially in manufacturing and other mission-critical domains—cannot depend on only remote cloud architectures.
5. **Cost Optimization:** Dynamic workload arrangement is essential to balance costs across different providers while accounting for edge resource limitations and inconsistent pricing models

This paper proposes a microservices-based cloud-edge collaborative platform for real-time data streaming and analytics, demonstrated through a battery scanning and quality analysis use case in battery manufacturing,

2. Literature Review

Adoption of multi-cloud strategies results from the necessity to diversify the service capabilities, lower the reliance on a single cloud provider, and maximize compute storage or specialist artificial intelligence services. In parallel, edge computing has also been applied for its ability to reduce end-to-end latencies. Many manufacturing or IIoT-based applications are now using microservice-based architecture at the edge to preprocess the data and then offload complex analytics to the cloud.

2.1. Microservices architecture

Traditional monolithic deployments and even virtual machine-based services come with some operational overhead due to heavy operating systems and complex dependencies. On the other hand, containerized microservices include only the essential components needed for each independent service, improving portability and reducing downtime.

Kubernetes is widely used in the industry for automating container orchestration and handling tasks including container formation, scaling, and load balancing, simplifying deployment and management of containers

2.2. Cloud-edge synergy

Cloud-based services offer massive computing power but often struggle with round-trip latency and bandwidth limitations. On the other hand, edge computing provides real-time responsiveness but often comes with storage and processing constraints. To tackle the strengths of both, a cloud-edge combination is important:

- Edge: Enables real-time streaming analytics, can make immediate local decision-making, reduces network load, and enhances data privacy.
- Cloud: Supports large-scale data processing, long-term model training, vast storage capacity, and seamless global collaboration among domain experts.

2.3. Condition monitoring and AI

As device connectivity expands, AI-driven fault detection and diagnosis are becoming increasingly valuable in manufacturing and other IIoT applications. By adopting a microservices-based cloud-edge collaboration model, machine learning algorithms—such as deep neural networks, long short-term memory (LSTM), and particle filtering—can be trained and updated in the cloud, while inference and partial analytics are performed at the edge.

3. Methodology

3.1. Microservices-based cloud-edge collaborative architecture

To show a practical real-time analytics framework, we propose a three-layer architecture::

1. **Edge layer**:
    - Feature Extraction Service (FES): Captures and filters IIoT data, extracting important parameters such as sensor signals and health indicators (HIs).
    - Edge Data Storage Service (EDSS): Locally stores for critical or sensitive data to ensure quick access and security.
    - Edge Artificial Intelligence Service (EAIS): Conducts immediate analytics, including preliminary fault detection and rapid inference.
    - Information Model Service (IMS): Standardizes data into common formats (e.g., OPC UA) and publishes structured data to the cloud.
2. **Cloud layer**:
    - Message Delivery Service (MDS): Receives and decodes data from many edge nodes before sending it to cloud storage.
    - Cloud Data Storage Service (CDSS): Handles large-scale structured, semi-structured, and unstructured data with horizontal scalability.
    - Cloud Monitoring Service (CMS): Provides dashboards for real-time monitoring.
    - Cloud Artificial Intelligence Service (CAIS): Enable collaborative training of advanced AI models using aggregated data.
3. **Control plane** (Kubernetes cluster manager):
    - Oversees resource allocation and container lifecycle management. Key components include:
        - API Server: Manages communication between components.

- etcd Database: Stores cluster configuration and state.
- Controller Manager: Maintains system health and performance.
- Scheduler**:** Distributes workloads efficiently across available resources

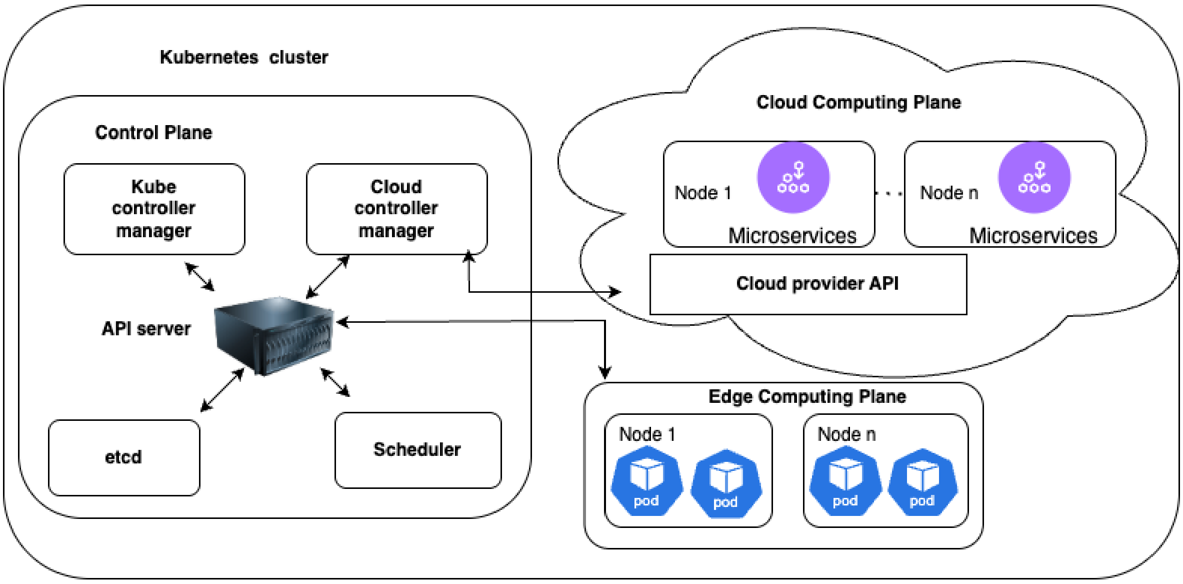

Figure 2: Microservices-Oriented Kubernetes Cluster for Cloud-Edge Computing

Deploying services as microservices makes updates more manageable and provides more control over CPU, memory, and network resources. Each service is packaged as a container image, which allows easy migration across nodes or data centers.

3.2. Cloud-edge collaboration mechanism

3.2.1. Real-time diagnostics

1. Local Edge Diagnostics: The Edge Artificial Intelligence Service (EAIS) runs AI models (e.g., LSTM, CNN) on time-domain features from the Feature Extraction Service (FES), which enables real-time anomaly detection and classification.
2. Cloud-Based Model Optimization: Historical data from many edge locations is stored in the Cloud Data Storage Service (CDSS). The Cloud Artificial Intelligence Service (CAIS) then runs advanced training of AI models using global datasets, which leverage distributed learning.

After improved model parameters are found in the cloud, they are deployed back to EAIS to continually refine real-time detection.

3.2.2. Prognostics with cloud-edge synergy

For Remaining Useful Life (RUL) estimation in machines, a hybrid edge-cloud solution makes sure both real-time inference and continuous model improvement:

1. Edge-Based Inference: A Particle Filter (PF) or LSTM-based prognosis model runs locally, analyzing real-time sensor data to estimate RUL on the fly.
2. Cloud-Based Parameter Updates: When condition drift is detected or at scheduled intervals, newly observed data is sent to the cloud. The Cloud Artificial Intelligence Service (CAIS) then fine-tunes or recalibrates the model parameters.
3. Bidirectional Model Deployment: The optimized parameters are pushed back to the edge, improving inference accuracy in real-time.

By offloading computationally intensive tasks—such as parameter estimation in large particle filters or neural network training—to the cloud, edge nodes can stay responsive even under resource constraints.

4. Implementation and Case Study

To demonstrate the proposed methodology in action, we built a small-scale proof-of-concept environment that integrates multi-cloud orchestration with edge computing for battery scanning and quality analysis. In this setup, an industrial battery production line is equipped with sensors and scanning devices (e.g., thermal, optical, or impedance-based). The machine connects to the network via a local gateway, which supports high bandwidth and low latency, enabling real-time data exchange between the edge node and the cloud cluster.

4.1. System Deployment

Edge Node Configurations

On edge servers with moderate CPU resources (4–8 cores) and adequate memory, the Feature Extraction Service (FES), Information Model Service (IMS), Edge Data Storage Service (EDSS), and Edge Artificial Intelligence Service (EAIS) each run as Kubernetes deployments to ensure scalability and efficient resource management.

- Feature Extraction Service (FES): Filters and aggregates sensor signals (spindle current, vibration, temperature), reducing data volume while preserving critical indicators.
- Edge Artificial Intelligence Service (EAIS): Hosts containerized AI models (e.g., JupyterLab-based) for rapid inference tasks like fault detection and tool wear monitoring.
- Edge Data Storage Service (EDSS): Maintains redundant storage so that key sensor data and inference outputs remain locally accessible, even if network connectivity is lost.
- Information Model Service (IMS): Structures edge-level data in a standardized format (e.g., OPC UA) and publishes it to the cloud using MQTT for seamlessly integrating

Cloud Cluster Configurations

- Message Delivery Service (MDS): Listens for incoming data streams from IMS, decodes them, and writes structured sensor data to cloud storage.
- Cloud Data Storage Service (CDSS): Provides highly scalable storage for large datasets, including raw sensor data and inference logs.
- Cloud Monitoring Service (CMS): Offers a web-based dashboard for remote monitoring of machine status and system health, accessible from anywhere.
- Cloud Artificial Intelligence Service (CAIS): Enables advanced analytics and model retraining at scale. It also facilitates collaboration by integrating expertise from distributed teams, allowing for the development of improved AI algorithms.

4.2. Quality Defect Detection Scenario

1. Data Extraction: The Feature Extraction Service (FES) constantly captures every scan output (e.g., thermal images, voltage curves) from each battery cell. It then computes key feature vectors—such as maximum temperature gradient, variance in voltage, or impedance factors—to summarize battery health characteristics.
2. Edge Inference: The Edge Artificial Intelligence Service (EAIS) runs a CNN-based defect detection model in real-time, classifying batteries into ‘OK’ or ‘defect-suspected’ with around 90% accuracy, outperforming baseline models like SVM and standard DNNs (which range from 65% to 80%).
3. Cloud Retraining: At periodic intervals, raw signals and extracted features are sent to Cloud Data Storage Service (CDSS). The Cloud Artificial Intelligence Service (CAIS) then uses this broader dataset to fine-tune LSTM hyperparameters. Once retrained, the updated model is then redeployed to EAIS, continuously improving edge-based diagnostics.

4.3. RUL Prediction Scenario

Even though the previous discussions focused on fault classification, many manufacturing processes also require prognostics, such as predicting the Remaining Useful Life (RUL) of battery cells and machines.

- Edge-Based Particle Filter (PF): The Edge Artificial Intelligence Service (EAIS) initially runs a PF-based approach to estimate battery degradation using real-time voltage, current, and thermal measurements. Early predictions may have higher uncertainty due to limited available data.
- Periodic Parameter Updates: As new measurements are collected, the Information Model Service (IMS) publishes updated data to the cloud. The Cloud Artificial Intelligence Service (CAIS) then refines the PF model parameters using a larger historical dataset, before pushing the recalibrated values back to EAIS.
- Improved Accuracy: This continuous feedback loop dramatically reduces RUL prediction error—for example, improving accuracy from 200% error down to 50%. This hybrid approach outperforms purely data-driven methods (e.g., Gaussian Process Regression (GPR)) and purely physics-based models (e.g., exponential degradation models).

4.4. Discussion of Results

This case study demonstrates how a microservices-based hybrid cloud-edge framework can significantly enhance real-time analytics and fault detection—without the need for an extensive physical setup. By strategically balancing edge inference and cloud retraining, organizations can unlock several advantages:

- Lower Latency: Running AI inference at the edge lowers round-trip time and network overhead, therefore decreasing latency by up to 30% compared to depending only on the cloud.
- Seamless Scalability: Kubernetes' orchestration dynamically adjusts microservice replicas based on workload demands, scaling resources automatically as sensor data volume fluctuates—without manual intervention.
- Cost Efficiency: Filtering data locally and executing rapid inferences at the edge reduces unnecessary cloud data transfers and compute expenses.
- Built-in Resilience: Important services like Edge Data Storage Service (EDSS) and Information Model Services (IMS) operate in replicated modes, which guarantees continuous operation even if an instance fails.

5. Conclusion and Future Work

This paper has introduced an integrated hybrid and multi-cloud architecture enhanced by edge computing to enable real-time data streaming, analytics, and condition monitoring. The proposed microservices-based framework brings together:

- Lightweight containerization for efficient feature extraction, AI inference, data storage, and monitoring.
- Cloud-edge collaboration, leveraging the low-latency responsiveness of edge computing while utilizing the scalability and storage power of the cloud.
- Modular, scalable orchestration with Kubernetes, allowing seamless deployment, migration, and scaling across heterogeneous infrastructures.

Through a battery scanning and quality analysis case study, this approach has demonstrated robust real-time defect detection (~90% accuracy) and a 50% improvement in RUL prediction accuracy. accuracy by iteratively retraining models in the cloud. Overall, the system highlights key advantages such as cost efficiency, reduced latency, higher availability, and dynamic resource optimization.

Future Work

In future work, we aim to improve this framework by extending to larger, varied datasets, adding audio and video processing for richer analytics, integrating digital twin technology to enable real-time simulation and what-if analyses for predictive maintenance. developing advanced scheduling algorithms to dynamically allocate microservices between edge and cloud environments, optimizing both resource costs and performance. Through the improvement of these features, this hybrid cloud-edge architecture may extend the limits of real-time artificial intelligence-driven decision-making in industrial sectors.

References

[1] George, Jobin, Optimizing hybrid and multi-cloud architectures for real-time data streaming and analytics: Strategies for scalability and integration (October 29, 2022). World Journal of Advanced Engineering Technology and Sciences, volume 7, issue 1, 2022[10.30574/wjaets.2022.7.1.0087], Available at SSRN: https://ssrn.com/abstract=4963389

[2] Yang, H., Ong, S. K., Nee, A. Y. C., Jiang, G., & Mei, X. (2022). Microservices-based cloud-edge collaborative condition monitoring platform for smart manufacturing systems. *International Journal of Production Research*, *60*(24), 7492–7501. https://doi.org/10.1080/00207543.2022.2098075

[3] Hybrid cloud vs. multi-cloud: Exploring pros and cons. (n.d.). https://reolink.com/blog/hybrid-cloud-vs-multicloud/

[4] Singh, G. (2024, August 19). Hybrid multi-cloud - Management and strategies. XenonStack. https://www.xenonstack.com/blog/hybrid-multi-cloud

[5] Bayya, A. K. (2025). Leveraging Advanced Cloud Computing Paradigms to Revolutionize Enterprise Application Infrastructure. Asian Journal of Mathematics and Computer Research, 32(1), 133–154. https://doi.org/10.56557/ajomcor/2025/v32i19067

[6] Zhu, Yuhan, et al. "Root Cause Localization for Microservice Systems in Cloud-edge Collaborative Environments." arXiv preprint arXiv:2406.13604 (2024).

[7] Bottacci, F. (2023). The Role of Edge Computing in Industrial IoT: Real-Time Analytics Revolution. LinkedIn Pulse. https://www.linkedin.com/pulse/role-edge-computing-industrial-iot-real-time-fabio-bottacci

[8] Rolando Brondolin and Marco D. Santambrogio. 2020. A Black-box Monitoring Approach to Measure Microservices Runtime Performance. ACM Trans. Archit. Code Optim. 17, 4, Article 34 (November 2020), 26 pages. https://doi.org/10.1145/3418899

[9] Eliganti Ramalakshmi, Venkata Srinivas Kompally, Baddam Deepika Reddy. (2020). Solar Powered Smart Irrigation and Monitoring System for Greenhouse Farming using IoT. International Journal of Advanced Science and Technology, 29(04), 8239 -. Retrieved from http://sersc.org/journals/index.php/IJAST/article/view/30559

[10] Naik, S. (2025). Cloud-Based Data Governance: Ensuring Security, Compliance, and Privacy.The Eastasouth Journal of Information System and Computer Science, 1(01), 69–87. https://doi.org/10.58812/esiscs.v1i01.452

[11] Singh, G. (2024). Hybrid Multi-Cloud Management Strategies. XenonStack. https://www.xenonstack.com/blog/hybrid-multi-cloud

[12] Sekonya, N. & Sithungu, S. (2023). Edge Computing Impact on IIoT Responsiveness. International Conference on Cyber Warfare & Security. https://papers.academic-conferences.org/index.php/iccws/article/download/969/974/3430

[13] ClearScale (2024). Edge Computing's Data Analysis Transformation. https://blog.clearscale.com/bringing-iot-to-the-edge-how-edge-computing-is-transforming-data-analysis-on-the-cloud/

[14] Chouliaras, S., & Sotiriadis, S. (2023). An adaptive auto-scaling framework for cloud resource provisioning. Future Generation Computer Systems, 148, 173–183. https://doi.org/10.1016/j.future.2023.05.017

[15] Hanbo Yang & S. K. Ong & A. Y. C. Nee & Gedong Jiang & Xuesong Mei, 2022. "Microservices-based cloud-edge collaborative condition monitoring platform for smart manufacturing systems," International Journal of Production Research, Taylor & Francis Journals, vol. 60(24), pages 7492-7501, December.

[16] A. Ghosh, A. Mukherjee and S. Misra, "SEGA: Secured Edge Gateway Microservices Architecture for IIoT-based Machine Monitoring," in IEEE Transactions on Industrial Informatics, doi: 10.1109/TII.2021.3102158.

[17] Li Z, Fei F, Zhang G. Edge-to-Cloud IIoT for Condition Monitoring in Manufacturing Systems with Ubiquitous Smart Sensors. Sensors (Basel). 2022 Aug 7;22(15):5901. doi: 10.3390/s22155901. PMID: 35957460; PMCID: PMC9371406.

[18] Ouyang R, Wang J, Xu H, Chen S, Xiong X, Tolba A, Zhang X. A Microservice and Serverless Architecture for Secure IoT System. Sensors (Basel). 2023 May 18;23(10):4868. doi: 10.3390/s23104868. PMID: 37430781; PMCID: PMC10220873.

[19] https://www.cloudgeometry.com/case-studies/big-data-analytics-for-industrial-iot#:~:text=Asynchronous%20processing%20of%20IIoT%20event,and%20the%20range%20of%20devices