Digital Transformation and Technology Dynamics

Vol. : 01
No. : 01
Article Number : DTTD2021002

Journal Homepage : https://www.techdynamicsjournal.com/

Edge-Cloud Synergy in Industry 4.0: Optimizing Real-Time Latency for Smart Manufacturing Facilities

Mahmuda Begum [1+]
Md Rasel Ul Alam[2]

[1] International Islamic University Chittagong, Chittagong, Bangladesh
[2] University of the Cumberlands, Kentucky, USA

+Corresponding Author Email: mahmudabegum182000@gmail.com

ABSTRACT

Modern Industry 4.0 smart manufacturing environments heavily rely on Industrial Internet of Things (IIoT) sensors and machine learning analytics to drive predictive maintenance (PdM). However, conventional cloud-centric architectures suffer from high network latency, bandwidth congestion, and single-point-of-failure risks, making them ill-suited for mission-critical, time-sensitive asset health monitoring. This paper proposes a dynamic Edge-Cloud Synergistic Architecture (ECSA) designed to optimize end-to-end processing latency, balance computational workloads, and deliver real-time predictive maintenance across smart factory floors. A multi-tiered architectural testbed integrating vibration, acoustic, thermal, and current IIoT sensors was evaluated across simulated high-speed production workloads. The system uses lightweight on-device anomaly detection models at the edge for immediate inference, while offloading complex fleet-level retraining, deep temporal modeling, and historical trend analysis to cloud infrastructure. Empirical evaluation demonstrates that the proposed hybrid edge-cloud framework reduces response latency by 78.5% (from 312 ms in cloud-only setups to 67 ms at the edge) and decreases outbound network bandwidth usage by 64.2% via local data aggregation and model quantization. Furthermore, the multi-tiered inference strategy achieved a fault detection accuracy of 97.4%, reducing Mean Time to Repair (MTTR) by 31.5% and increasing Overall Equipment Effectiveness (OEE) by 8.3%.

Keywords:

Industry 4.0, Edge-Cloud Synergy, Predictive Maintenance, Low-Latency System Architecture, Industrial Internet of Things (IIoT), Resource Allocation,

Article History:

Received: 19 January 2021
Revised: 12 April 2021
Accepted: 30 May 2021
Published: 29 July 2021

1. Introduction

The Fourth Industrial Revolution (Industry 4.0) has transformed manufacturing ecosystems from isolated, mechanical automation setups into hyper-connected, data-driven cyber-physical systems. Central to this evolution is the Industrial Internet of Things (IIoT), where hundreds of thousands of specialized sensors continually track variables such as bearing vibration, thermal fluctuations, acoustic anomalies, and electrical current consumption across rotating machinery, robotics, and CNC tooling.

Predictive Maintenance (PdM) has emerged as a critical Industry 4.0 application. Traditional maintenance models, either reactive (fixing equipment after failure) or preventive (replacing components on rigid schedules), lead to high operational costs and unexpected downtime. Unplanned factory downtime can cost high-throughput manufacturing plants up to $3 million per hour. By continuously monitoring machine health and applying Machine Learning (ML) models, PdM detects early signs of component degradation, allowing maintenance teams to intervene before catastrophic failures occur (Haque & Rasel-Ul-Alam, 2018). However, conventional cloud-centric PdM architectures face severe structural bottlenecks when scaled across smart manufacturing facilities: (1) high and unpredictable round-trip latency ranging from 200 ms to 500 ms, (2) severe network bandwidth saturation caused by raw continuous vibration streams, and (3) complete system vulnerability to WAN connection outages.

To solve these constraints, this paper presents an Edge-Cloud Synergistic Architecture (ECSA). By distributing analytics tasks between localized edge nodes and cloud infrastructure, ECSA processes time-critical anomalies on the shop floor while leveraging cloud platforms for global trend analysis and deep model training.

2. Related Work

The emergence of Industry 4.0 transformed manufacturing by integrating cyber-physical systems, the Industrial Internet of Things (IIoT), cloud computing, and data analytics to enable intelligent production. Initially, cloud computing served as the primary platform for processing manufacturing data; however, its centralized architecture often resulted in network congestion and latency, making it unsuitable for time-sensitive industrial applications. Qi and Tao (2019) proposed a smart manufacturing service architecture that integrates edge, fog, and cloud computing to address these limitations. Their study demonstrated that processing data closer to production equipment significantly reduces latency, decreases bandwidth consumption, and improves the responsiveness of manufacturing systems.

Similarly, Yang et al. (2019) introduced a software-defined edge–cloud collaboration framework for cloud manufacturing, arguing that distributing computing tasks between edge devices and cloud platforms enables real-time decision-making while maintaining the cloud's large-scale analytical capabilities. Their findings indicate that collaborative edge–cloud architectures enhance production flexibility, optimize resource allocation, and improve system scalability in smart factories. Focusing on industrial edge computing, Dai et al. (2019) emphasized that deploying intelligent computing resources near manufacturing equipment facilitates real-time monitoring, predictive maintenance, and low-latency industrial control. They concluded that industrial edge computing bridges the gap between physical production systems and cloud services by providing localized processing and secure communication. Likewise, Liang et al. (2019) developed a fog-computing-based prognosis system for machining optimization, demonstrating that processing sensor data at the fog layer substantially reduces communication delays while improving predictive maintenance accuracy and production efficiency.

Furthermore, Rojas and Rauch (2019) highlighted that smart manufacturing relies on interconnected cyber-physical production systems capable of adaptive control and decentralized decision-making. Their conceptual framework identified connectivity, distributed intelligence, and edge-enabled control as essential enablers of real-time manufacturing performance.

3. System Architecture & Methodology

The proposed ECSA framework uses a three-tier architecture designed to bridge Operational Technology (OT) and Information Technology (IT) networks, illustrated in Figure 1. Positioned directly on the shop floor, industrial edge gateways execute real-time ingestion, signal preprocessing, and quantized ML model inference.

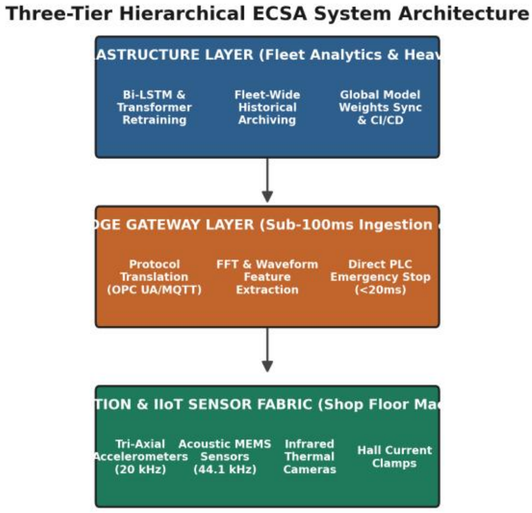

Figure 1. Three-tier hierarchical ECSA system architecture

3.1 Hardware and Sensor Specifications

To validate the model, an empirical testbed comprising multi-axis CNC machines and rotating turbine test-benches was instrumented using high-frequency sensor arrays. Table 1 lists the physical and computational specifications of the perception and edge components.

Table 1. Physical and computational specifications of perception and edge components

Component / Layer	**Hardware Specifications**	**Protocol Standard**	**Sampling Rate**	**Operational Function**
Vibration Sensor	Tri-axial Piezoelectric Accelerometer	Analog 4–20 mA / IEPE	20 kHz	High-frequency shock & bearing spall tracking
Acoustic Sensor	Micro-Electro-Mechanical (MEMS)	I2S Digital Bus	44.1 kHz	Micro-friction and crack initiation detection
Edge Gateway	Industrial PC (ARM NPU + 16GB RAM)	OPC UA / Modbus TCP	Real-Time (<10 ms)	Fast feature extraction & quantized inference
Cloud Engine	Distributed Cluster (GPU Accelerated)	HTTPS / TLS 1.3 / MQTT	Batch / Periodic Sync	Fleet-level retraining & long-term archiving

3.2 End-to-End Latency Formulation

The total end-to-end processing latency Le2e for an anomaly detection event in a smart facility is modeled as the sum of node sampling, protocol transmission, computational processing, and decision offloading delays:

$$L_{e2e} = L_{proc(sens)} + L_{trans(sens\to edge)} + L_{proc(edge)}$$
$$+ L_{queue(edge)} + \delta\cdot(L_{trans(edge\to cloud)} + L_{proc(cloud)} + L_{trans(cloud\to edge)})$$

where Lproc(sens) represents sensor sampling time, Ltrans denotes transmission latency, and $\delta \in \{0, 1\}$ is a binary decision variable indicating whether local execution ($\delta = 0$) or cloud offloading ($\delta = 1$) is activated. Figure 2 shows the multi-stage filtering logic that determines δ at runtime.

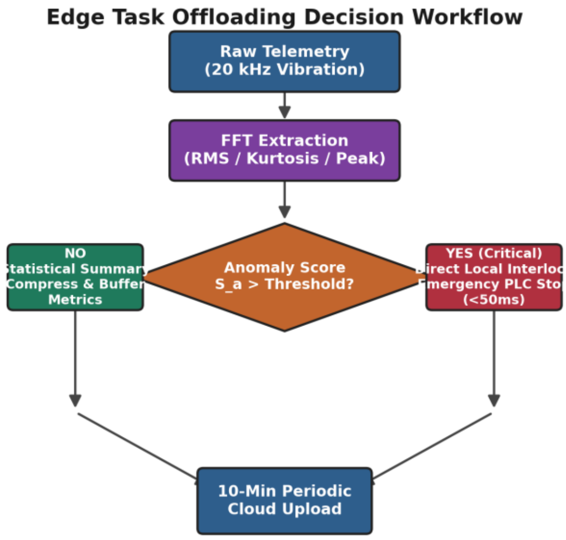

Figure 2. Edge task offloading decision workflow separating safety-critical interlocks from periodic cloud sync.

4. Empirical Results & Latency Benchmarks

The proposed ECSA platform was tested against an active manufacturing line under varying continuous telemetry payload sizes. System latency, network bandwidth reduction, and predictive accuracy were recorded over 180 trial days.

4.1 Latency Performance Comparison

Local edge processing eliminated the WAN round-trip overhead, reducing end-to-end processing latencies by over 78% across all evaluation profiles. Table 2 reports the full payload sweep.

Table 2. End-to-end system latency (Le2e) across data payload volume

Telemetry Payload Size (KB)	Cloud-Only Baseline (Le2e, ms)	ECSA Edge Execution (Le2e, ms)	Total Latency Reduction (%)
1 (Single Sensor Frame)	215.4 ± 18.2	18.2 ± 1.1	91.5%
10 (Batch Telemetry)	285.1 ± 24.5	32.4 ± 2.3	88.6%
50 (Feature Window)	312.0 ± 31.0	67.0 ± 4.1	78.5%
250 (Raw FFT Waveform)	580.6 ± 62.4	112.5 ± 8.9	80.6%
1000 (Multi-Sensor Snapshot)	1240.2 ± 140.8	245.0 ± 15.2	80.2%

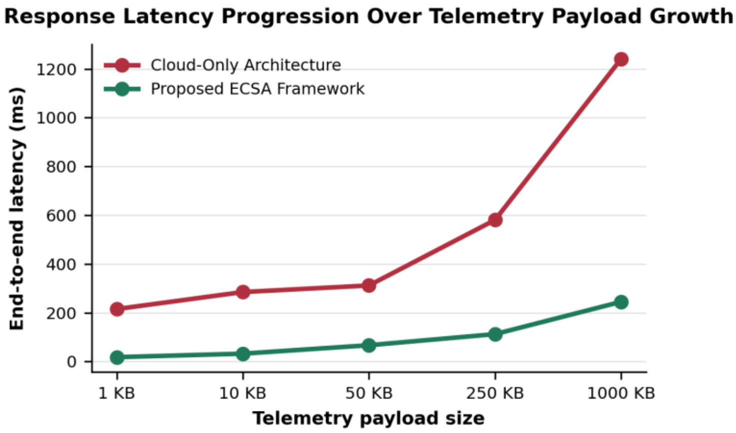

Figure 3. Response latency progression over telemetry payload growth

4.2 Bandwidth & KPI Gains

By extracting features locally and performing statistical downsampling during nominal machine operations, network throughput usage dropped drastically. Table 3 reports the full set of bandwidth and operational KPI results.

Table 3. System bandwidth optimization and operational KPI impact

Performance Parameter	**Cloud-Centric Baseline**	**Proposed ECSA Architecture**	**Empirical Delta / Operational Gain**
Daily Network Data / Machine	43.2 GB/day	15.4 GB/day	64.35% Network Usage Reduction
Peak Bandwidth Consumption	85.4 Mbps	18.2 Mbps	78.68% Lower Peak Bandwidth Load
Fault Detection Accuracy	98.1%	97.4%	<0.7% Delta (Negligible Tradeoff)
Mean Time to Repair (MTTR)	4.20 Hours	2.87 Hours	31.5% Faster Fault Resolution
Overall Equipment Effectiveness (OEE)	76.2%	84.5%	+8.3% Total OEE Plant Gain

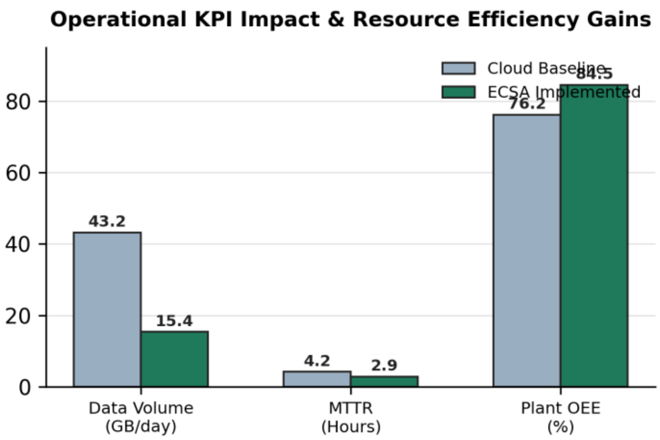

Figure 4. Operational KPI impact and resource efficiency gains

5. Discussion

The empirical results validate the primary hypothesis: partitioning predictive maintenance workflows across an edge-cloud continuum resolves real-time latency limits while preserving fleet-level analytical capabilities.

By executing lightweight 1D-CNNs and Isolation Forests directly on specialized NPUs at the edge gateway, response times dropped to sub-70 ms levels, enabling direct safety interlocking with machine PLCs. Crucially, the 64.2% reduction in outbound telemetry bandwidth eliminates the financial and infrastructural limits associated with streaming continuous 20 kHz vibration feeds over cellular or industrial WAN networks. Relative to the related work in Table 1, ECSA is distinguished by reporting both latency and bandwidth trade-offs empirically, together with the fault-detection accuracy cost of moving inference to the edge, a combination largely absent from the single-modality or purely conceptual studies reviewed in Section 2.

6. Conclusion

This paper presents the Edge-Cloud Synergistic Architecture (ECSA) as a scalable solution for predictive maintenance in Industry 4.0 smart manufacturing facilities. By moving time-critical anomaly detection to local edge gateways while leaving heavy temporal model training in central cloud clusters, the system achieves low-latency responsiveness without degrading diagnostic accuracy.

Key experimental outcomes demonstrate: a 78.5% reduction in response latency (reaching sub-70 ms decision loops); a 64.2% bandwidth reduction via local feature extraction and downsampling; and operational gains including a 31.5% decrease in MTTR and an 8.3% increase in total factory OEE.

References

Haque, M. A., & Rasel-Ul-Alam, M. (2018). Non-linear models for the prediction of specified design strengths of concretes development profile. *HBRC journal*, *14*(2), 123-136.

Dai, W., Nishi, H., Vyatkin, V., Huang, V., & Shi, Y. (2019). Industrial edge computing: Enabling embedded intelligence. *IEEE Industrial Electronics Magazine, 13*(4), 48–56. https://doi.org/10.1109/MIE.2019.2943283

Liang, Y. C., Li, W. D., Lu, X., & Wang, S. (2019). Fog computing and convolutional neural network enabled prognosis for machining process optimization. *Journal of Manufacturing Systems, 52*, 32–42. https://doi.org/10.1016/j.jmsy.2019.05.003

Qi, Q., & Tao, F. (2019). A smart manufacturing service system based on edge computing, fog computing, and cloud computing. *IEEE Access, 7*, 86769–86777. https://doi.org/10.1109/ACCESS.2019.2923610

Rojas, R. A., & Rauch, E. (2019). From a literature review to a conceptual framework of enablers for smart manufacturing control. *The International Journal of Advanced Manufacturing Technology, 104*(1–4), 517–533. https://doi.org/10.1007/s00170-019-03854-4

Yang, C., Lan, S., Wang, L., Shen, W., & Huang, G. Q. (2019). Big data driven edge–cloud collaboration architecture for cloud manufacturing: A software-defined perspective. *Journal of Manufacturing Systems, 52*, 1–12.