SCIENTIFIC CULTURE
www.sci-cult.com
SCIENTIFIC CULTURE, Vol. 12, No. 4, (2026), pp. 3533-3543
Open Access. Online & Print

DOI: 10.5281/zenodo.12426412

A CLOUD-IOT ENABLED, GPU ACCELERATED HPC ARCHITECTURE FOR ENERGY OPTIMIZED ROBOTIC FLEETS WITH MULTI-SENSOR 3D VISION AND AUTONOMOUS DECISION-MAKING IN SMART MANUFACTURING

Prabakar D[1], Sudhindra B Deshpande[2*], Dr. Kallur V Vijayakumar[3], Dr. Ganesh Kumar R[4], Gangu Rama Naidu[5], Vijay Kumar Gowda B N[6]

[1]Computer Science and Engineering, Karpagam College of Engineering, Coimbatore, Tamil Nadu Professor,
[2]Department of Artificial Intelligence and Machine Learning (AIML) Anuvartik Mirji Bharatesh Institute of Technology Bharatesh,Chandragiri campus, Basavan kudachi, Belagavi - 591124
[3]Mathematics BMS Institute of Technology and Management, Ahlahalli, Bangalore, Karnataka, India (560064)
[4]Department of Computer Science and Engineering, CHRIST (Deemed to be University), Mysore Road, Kumbalgodu. Bangalore 560074.
[5]Department of Electronics and Communication Engineering, Aditya University, Surampalem, Andhra Pradesh, India
[6]Department of Electronics and Communication Engineering, BGS Institute of Technology, Faculty of Engineering Management and Technology, Adichunchanagiri University, Mandya, B G Nagar, Karnataka

Received: 29/09/2025
Accepted: 16/02/2026

Corresponding Author: Sudhindra B Deshpande (sbdsudhi@gmail.com)

ABSTRACT

Smart manufacturing according to Industry 4.0 requires intelligent robotic fleets with the ability to perceive, coordinate and operate with minimal energy consumption in real time. Nevertheless, state-of-the-art cloud-based systems have low bandwidth, high power usage, and low scalability to handle multi-sensor 3D vision images. This research suggests development of a Cloud-IoT equipped, GPU-enhanced High-Performance Computing (HPC) platform of energy-efficient robotic fleet with autonomous decision-making in an attempt to overcome these challenges. The suggested methodology consists of a four-layer system that incorporates robotic fleet, edge computing, cloud HPC, and IoT management levels. Based on which formulations are introduced on the basis of energy-aware task scheduling, which is a multi-objective optimization problem, a greater role is played by the comprehension of the perception and coordination with the assistance of the perception and decision models based on multi-sensor fusion and reinforcement learning. Through the experimental performance, energy consumption decreases by 820-520 J at 30 robots (approximately 37 increases per cent), latency decreases by 1450 ms to 920 ms (approximately 36 improvement), the mAP accuracy rises by 0.72 to 0.92, reconstruction error drops by 0.18 to 0.08, and the task success rate rises by 78 to 96 which has proved the scale up and scalability.

KEYWORDS: Cloud–IoT Architecture, GPU-Accelerated HPC, Energy Optimization, Robotic Fleets, Multi-Sensor 3D Vision, Autonomous Decision-Making

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License. (https://creativecommons.org/licenses/by/4.0/).

1 INTRODUCTION

The fast evolution of manufacturing systems according to the paradigm Industry 4.0 has resulted in the integration of high-level digital technologies: the Internet of Things (IoT), cloud computing, high-performance computing (HPC), robotics, artificial intelligence (AI) [1]. The concept of modern smart manufacturing settings is no more than a separated automated machinery, but an interdependent, intelligent, and responsive environment, which is able to sense, analyze, and make decisions in real-time. The crucial form of these systems is the robot fleets with multi-sensor perception and autonomous control, which ensure high productivity, flexibility and efficiency of operations [2, 3]. Smart robot fleets are also finding applications in factories to carry out sophisticated duties working on material management, production, inspection, and coordination with human workers. These activities require finely tuned sense of the surrounding situation, the ability to plan dynamic paths and quick reaction to the environment conditions [4]. The ability of the robots to build an accurate spatial representation of their environment is made possible by multi-sensor 3D vision systems, which combine information in cameras, LiDAR, depth sensors, and inertial measurement sensors [5]. Nevertheless, real time processing of high volume, high dimensional sensor data is computationally expensive and especially when many robots work in a common workspace [6] it requires substantial computational resources. It has appeared that cloud-IoT architectures could become a promising solution in these issues because it can ensure a smooth data exchange, central coordination, and scalable computing resources by design [7]. The IoTs enable real-time communication between robots and sensors with manufacturing infrastructure and provide elastic storage and computing resources on cloud computing infrastructure to process and analyze extensive data sets [8]. However, latency, bandwidth, and the lack of energy efficiency of traditional cloud-centric solutions is frequently a problem, and could negatively impact its use in time sensitive robotic systems. This requires incorporation of new computing paradigms that have the ability to provide high performance and energy efficiency [9]. High-Performance Computing (HPC) especially when powered by Graphics Processing Unit (GPUs) is now an influential enabler to computationally intensive operations like 3D vision processing, inference with deep learning and massively scalable optimization. HPC systems, based on GPUs, provide massive parallelism, enabling a sensor data stream feed to be processed in real-time, and complex AI models to be executed, which would otherwise be impossible on a CPU-based system [10]. Combined with cloud and edge computing architectures, HPC based on the use of GPUs can serve to facilitate distributed intelligence in real-time, whereby the computational workloads are dynamically offloaded according to latency, power consumption, and available resources [11]. One of the essential issues of smart manufacturing is energy optimization, in particular, the massive implementation of robot fleets and data-oriented computing systems. The pressure mounts on the manufacturing facilities to decrease the consumption of energy and exhaust emission but ensure high throughput and reliability [12]. Robotic systems are also important in energy consumption whether in mechanical actuation or in sensing, communication and computation. Data processing and organisation of tasks may be inefficientness, which may cause unnecessary energy overheads, lower battery life of mobile robots, and increased cost of operating the robot [13]. Thus, sustainable and cost-effective smart manufacturing system requires the use of energy-aware computing and control measures. Also, autonomous decision-making contributes to the efficiency of robotic fleets as it will allow them to work with minimum human guidance. With the help of AI-based algorithms, it is possible to represent sensor data and predict states of the system, as well as give optimal decisions about the navigation, allocation of tasks, and collaboration by robots [14]. This is because the combination of autonomous decision-making with cloud-IoT and HPC infrastructures permits both local intelligence at the robot level and global intelligence at the fleet level. This mixed method allows centralized action among the robots, responsive attitude to demands during production, and robustness against failures or environmental disruptions [15]. The use of multi-sensor 3D vision is one of the pillars of autonomous robot operation in manufacturing rooms. Object recognition, pose estimation, obstacle avoidance, and quality inspection need to be recognized correctly. But combining the heterogeneous sensor data to a coherent representation in 3D is computationally costly and latency and noise-sensitive [16].

HPC architectures that are implemented with G acquiring the use of a GPU can serve as a critical factor in enhancing the efficiency and precision of sensor fusion and 3D reconstruction algorithms, thus simplifying real-time perception even in a complex and cluttered factory environments [17]. As they are coupled with cloud-based analytics and IoT connection, these features promote a never-ending learning and system refinement of the whole robotic army [18]. Figure 1 shows the overall system architecture including the interaction of the IoT sensors with the GPUs based on the edge nodes, cloud based HRC resources and autonomous robot fleets. Despite the significant advancements in individual technologies, the holistic integration of cloud-IoT frameworks, GPU-accelerated HPC, energy-optimized computing, and autonomous robotic fleets remains a challenging research problem. Issues such as scalable architecture design, efficient workload distribution, energy-aware scheduling, real-time communication, and system reliability must be addressed in a unified manner. Existing approaches often focus on isolated components, lacking a comprehensive framework that simultaneously optimizes performance, energy efficiency, and autonomy in multi-robot manufacturing systems.

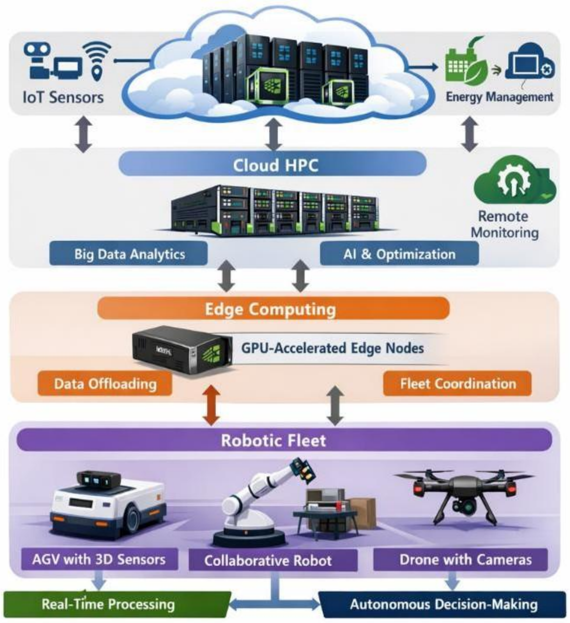

Figure 1: Cloud–IoT enabled GPU-accelerated HPC architecture for energy-optimized robotic fleets

In this context, the present work proposes a Cloud-IoT enabled, GPU-accelerated HPC architecture specifically designed for energy-optimized robotic fleets in smart manufacturing environments. The proposed architecture aims to seamlessly integrate multi-sensor 3D vision, high-performance computation, and autonomous decision-making within a scalable and energy-aware framework. By leveraging the strengths of cloud computing, IoT connectivity, and GPU-based acceleration, the system seeks to enhance real-time perception, collaborative intelligence, and sustainable operation of robotic fleets. Such an integrated approach has the potential to significantly improve manufacturing efficiency, adaptability, and resilience, contributing to the realization of next-generation smart factories. The research objectives of this Study are as follow:

1. To design a Cloud-IoT enabled, GPU-accelerated HPC architecture for coordinated and scalable operation of robotic fleets in smart manufacturing environments.
2. To develop energy-optimized computing and task-scheduling strategies that minimize power consumption while maintaining high computational performance and real-time responsiveness.
3. To implement multi-sensor 3D vision processing and sensor fusion mechanisms using GPU-accelerated HPC for accurate and real-time environmental perception.
4. To enable autonomous decision-making and collaborative intelligence among robotic fleets through AI-driven algorithms integrated with cloud and edge computing resources.

2 REVIEW OF LITERATURE

An increasing amount of smart manufacturing infrastructures is developed based on the utilization of IoT, cloud, and edge technologies to address efficiency and sustainability. The study by Alex et al., (2025) [19] reflected that a multi-layer IoT system incorporating sensors, edge computing, and cloud computing has a deep reduction in energy usage, machine unavailability and wastage of resources; hence, sustained operations of industries. Similarly, Soret et al., (2021) [20] also accentuated that in heterogeneous intelligent IoT spaces, comprising of robots, drones, and automated vehicles, there must be cautious trade-offs between latency, privacy, and energy efficiency. Salhaoui et al., (2021) [21] suggested using intelligence at the edge of the fog in order to address interoperability and latency issues and enhance real-time decision making and reliability in industrial systems. In addition, Ramareddy et al., (2025) [22] proposed a hybrid edge cloud architecture that can distribute the workload in an adaptive way and minimize latency time and energy usage and enhance the accuracy of decisions. In line with these results, Lilhore et al., (2025) [23] used resource scheduling, which relies on reinforcement learning, to maximally utilize the IoT workloads, enhancing response time and cost of operation. Altogether, these works define that scalable and energy-sensitive smart manufacturing systems cannot be accomplished without cloud-IoT convergence. High-performance computing and hardware acceleration is very important to make autonomous fleets of robots operating in that kind of environment possible. Liu et al., (2018) [24] developed a multi-accelerator runtime model enabling robots to concurrently handle navigation, mapping and scene understanding with severe power limits. In the same manner, Chen et al., (2021) [25] created a smoke robotics software architecture which redirects computation to cloud GPUs and dedicated hardware with a significant increase in processing speed but with manageable latency. Contributing to computer computational needs in industrial validation, Sanc Jose et al., (2024) [26] indicated the relevance of the cloud offloading to the use of the GPU to facilitate workload processing workloads due to computer vision and CAD/CAM operations. Also, Ahmad et al., (2024) [27] offered a hybrid deep-learning intrusion detection algorithm that uses the services of HPC clouds to offer real-time monitoring and safe communication in the IoT-connected setting. The combined works of these two ensure that the use of HPC infrastructure accelerated by GPUs and distributed computing can help in allowing real time robotic intelligence without compromising on the energy efficiency and the security of the system. Independent decision-making also relies on the credible perception and adjustive control concerning the industrialization. It has been demonstrated that the use of multi-sensor fusion and virtual sensing is highly useful in terms of providing accuracy and robustness to autonomous systems, (Sahoo et al., 2024) [28]. Within a manufacturing setup, Lu et al., (2024) [29] suggested multi sensory monitoring combined with adaptive control to increase the quality of additive manufacturing by using real-time adjustment of parameters. Also, the work by Yousif et al., (2025) [30] combined computer-vision-based digital twins to identify defects in the assembly and instruct robotic correction to minimize human intervention. All these among the perception-based research, coupled with safe and effective computing systems, points out to the fact that next generation manufacturing heavily relies on intelligent sensing and adaptive decision-making. Thus, the combination of cloud-IoT platforms, HPC processing, and multi- sensor perception can be considered a unified basis to build energy-efficient robotic fleets that are able to work autonomously in intelligent manufacturing facilities. Despite the results of the existing research showing the advantages of integrating IoT and clouds, fog–edge computing, accelerated with GPUs, and scheduling with reinforcement learning, multiple weaknesses still exist. Majority of the works cover each part alone instead of offering a single Cloud-IoT-GPU-HPC architecture specifically in coordinated robotic fleets. The approach of limited research optimizes at any given time energy, end-to-end latency, as well as autonomous decision-making when using large volumes of robots. Moreover, there is a lack of literature that incorporates multi-sensor 3D perception, dynamically distributed workloads, and fleet coordination with the help of reinforcement learning into a single scalable system. Thus, the design of collaborative robot fleets in smart manufacturing areas is still underdeveloped with a hermitically detailed, energy-conscious, and accelerated on GPUs architecture.

3 RESEARCH METHODOLOGY

The proposed study uses a model-based, system-focused, and experiment-based approach to the creation, deployment, and testing of an architecture of the Cloud-IoT-enabled, GPU-accelerated high-performance computing (HPC) in energy-efficient robots fleet in intelligent manufacturing settings. The process combines the architectural modelling, mathematical methods, implementation of algorithms on GPUs, modeling as well as performance analysis to guarantee scalability, real-time response, and energy efficiency. The entire working process is organized into integrative stages in accordance with the mentioned research goals.

3.1 System Architecture Design and Framework Development

In order to satisfy the goal of the first, a multi-layer Cloud-IoT-HPC architectural model is set up, whereby four functional layers are placed:

(i) Robotic Fleet Layer

The robotic layer of fleet includes a heterogeneous robot that are developed with multi-fatherly sensors that comprise RGB cameras, LiDAR, depth sensors, and inertial units. This layer is a layer of real-time sensing, local actuation, and implementation of local control activities with high latency requirements. It is the main source of data on which the process of perception and decision-making is based.

(ii) Edge Computing Layer

The edge computing layer is made up of GPU accelerated edge nodes that are placed in a manner that is close to the manufacturing floor in order to minimize latency in communication. It has real-time sensor fusion, preliminary 3D vision processing, and low-latency AI inference. This layer allows quick response at low cost to the cloud computing resources.

(iii) Cloud HPC Layer

The HPC layer uses processors with graphic cards to combine the power of computationally intensive and delay tolerant tasks. It promotes 3-dimensional reconstruction on a large scale, training of deep learning models, managing a fleet on a global scale and predictive analytics. This layer offers scalability, the high throughput and the location of intelligence.

(iv) IoT and Management Layer

The IoT and management layer facilitate a smooth connectivity, data interaction and coordination of robotic, edge and cloud components. It handles the authentication of devices, their monitoring and synchronization via standard communication protocols. This layer also allows a secure, reliable and scalable operation of the system.

Let the robotic fleet be represented as a set:

$$\mathcal{R} = \{R_1, R_2, \dots, R_N\},$$

where each robot $R_i$is equipped with a heterogeneous sensor suite

$$\mathcal{S}_i = \{s_i^{\text{RGB}}, s_i^{\text{LiDAR}}, s_i^{\text{Depth}}, s_i^{\text{IMU}}\}.$$

The robotic fleet layer performs local sensing and actuation, while latency-sensitive inference tasks are delegated to the edge computing layer, composed of GPU-accelerated edge nodes

$$\mathcal{E} = \{E_1, E_2, \dots, E_M\}$$

The cloud HPC layer contains GPU-enabled clusters that perform extensive computational work for deep learning model training and global fleet optimization and predictive analytics. The IoT and management layer establishes secure communication through lightweight protocols which enable device orchestration and synchronization using MQTT and RESTful APIs. The system achieves modularity and scalability together with its ability to manage fleet operations through its layered architectural design.

3.2 Energy-Optimized Computing and Task Scheduling Strategy

To address the second objective, an energy-aware computation and task scheduling model is formulated. The total energy consumption of a robot–edge–cloud system is modeled as

$$E_{\text{total}} = E_{\text{comp}} + E_{\text{comm}} + E_{\text{act}},$$

where E_{comp}, E_{comm}, and E_{act}represent computation, communication, and actuation energy, respectively. For GPU-based computation, energy consumption is estimated as

$$E_{\text{comp}} = \sum_k P_k \cdot T_k,$$

where $P_k$is the power consumption of computing node k(robot, edge, or cloud GPU), and $T_k$is the task execution time. Task scheduling is formulated as a multi-objective optimization problem:

$$\min_{\mathcal{A}} (\alpha E_{\text{total}} + \beta L_{\text{total}}),$$

subject to

$$L_{\text{total}} \le L_{\max}, U_{\text{GPU}} \le U_{\max},$$

where $\mathcal{A}$denotes task allocation decisions, $L_{\text{total}}$is end-to-end latency, and α, $\beta$are weighting coefficients. Latency-critical tasks (e.g., obstacle detection) are executed at the robot or edge layer, while delay-tolerant tasks (e.g., model retraining) are offloaded to the cloud HPC layer.

3.3 Multi-Sensor 3D Vision Processing and Sensor Fusion Implementation

To achieve the third objective, a GPU-accelerated multi-sensor 3D perception framework is implemented. Sensor measurements from heterogeneous sources are fused to estimate the environment state $X_t$as

$$X_t = f(Z_t^{\text{RGB}}, Z_t^{\text{LiDAR}}, Z_t^{\text{Depth}}, Z_t^{\text{IMU}}),$$

where $Z_t$represents sensor observations at time t. Probabilistic sensor fusion is performed using a Bayesian framework:

$$p(X_t \mid Z_{1:t}) \propto p(Z_t \mid X_t)\, p(X_t \mid Z_{1:t-1}).$$

GPU acceleration is applied to parallelize feature extraction, point cloud registration, and 3D reconstruction operations. Deep learning–based perception models are trained and deployed on GPU-enabled edge and cloud platforms to enhance object detection accuracy and reduce inference latency. Performance is evaluated using metrics such as mean average precision (mAP), reconstruction error, and processing latency.

3.4 Autonomous Decision-Making and Collaborative Intelligence

To satisfy the fourth objective, AI-driven autonomous decision-making mechanisms are integrated into the proposed architecture. The decision-making process is modeled as a Markov Decision Process (MDP):

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle,$$

where $\mathcal{S}$is the state space derived from 3D perception, $\mathcal{A}$is the action space, $P$is the state transition probability, $R$is the reward function, and $\gamma$is the discount factor. Reinforcement learning algorithms are used to optimize the policy π^*:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right].$$

Local decision-making is executed at the robot or edge level, while global coordination and task allocation are performed in the cloud HPC layer using shared fleet knowledge models.

3.5 Performance Evaluation Metrics

The performance of the proposed Cloud–IoT enabled, GPU-accelerated HPC architecture is quantitatively evaluated using the following metrics.

- **Energy Consumption**

The total energy consumption of the system is computed as the sum of computation, communication, and actuation energy:

$$E_{\text{total}} = E_{\text{comp}} + E_{\text{comm}} + E_{\text{act}}$$

where the computational energy is given by

$$E_{\text{comp}} = \sum_{i=1}^{N} P_i \cdot T_i$$

Here, P_i denotes the average power consumption of node i(robot, edge GPU, or cloud GPU), and $T_i$represents the execution time. Lower values of E_{total}indicate improved energy efficiency.

- **End-to-End Latency**

End-to-end latency measures the total delay experienced by a task from sensing to actuation:

$$L_{\text{total}} = L_{\text{sense}} + L_{\text{proc}} + L_{\text{comm}} + L_{\text{decision}} + L_{\text{act}}$$

where $L_{\text{sense}}$is sensor acquisition delay, $L_{\text{proc}}$is computation delay, $L_{\text{comm}}$is communication delay, $L_{\text{decision}}$is decision-making delay, and $L_{\text{act}}$is actuation delay. Real-time performance requires $L_{\text{total}} \leq L_{\text{threshold}}$.

- **GPU Utilization**

GPU utilization quantifies the effectiveness of GPU resource usage during task execution:

$$U_{\text{GPU}} = \frac{T_{\text{active}}}{T_{\text{total}}} \times 100\%$$

where T_{active}denotes the time during which the GPU actively executes kernels, and $T_{\text{total}}$is the total observation time. Higher utilization indicates better exploitation of GPU parallelism.

- **3D Perception Accuracy**

3D perception accuracy evaluates the correctness of object detection and environmental reconstruction. It is measured using mean Average Precision (mAP):

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^{C} AP_c$$

where $C$is the number of object classes and $AP_c$is the average precision for class c. Higher mAP values indicate more accurate perception.

- **Task Success Rate**

Task success rate reflects the reliability of autonomous operations:

$$\text{TSR} = \frac{N_{\text{success}}}{N_{\text{total}}} \times 100\%$$

where $N_{\text{success}}$is the number of successfully completed tasks and $N_{\text{total}}$is the total number of assigned tasks. A higher task success rate indicates improved autonomous decision-making and coordination.

- **Performance Gain**

Performance improvement over baseline systems is computed as:

$$\text{Gain}(\%) = \frac{M_{\text{proposed}} - M_{\text{baseline}}}{M_{\text{baseline}}} \times 100$$

where Mrepresents a performance metric such as energy consumption, latency, or accuracy.

4 RESULTS AND DISCUSSION

4.1 Energy Consumption Analysis

The graph compares total system energy consumption between the baseline system and the proposed Cloud–IoT–HPC architecture. The proposed architecture demonstrates lower energy consumption because it uses intelligent task offloading and GPU acceleration while both systems use more energy when their robot count increases. The graph in figure 2 compares energy consumption of the baseline and proposed architecture as the number of robots increases from 5 to 30. The baseline system shows a sharp energy increase which begins at 120 J with five robots and reaches 820 J with thirty robots. The proposed system shows a gradual energy increase which starts at 95 J and ends at 520 J throughout the entire testing range. The energy gap widens with fleet size: about 25 J at 5 robots, nearly 170 J at 20 robots, and roughly 300 J at 30 robots. The proposed architecture shows increasing energy savings through its design which continues to work as the system expands.

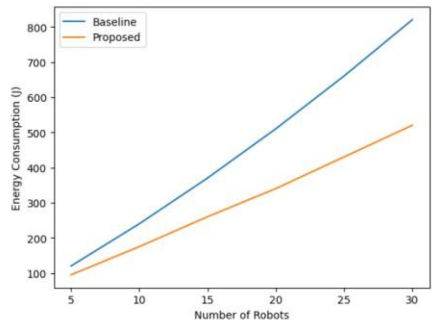

Figure 2: Comparison of Energy consumption

4.2 End-to-End Latency Performance

The latency comparison shows that the proposed architecture substantially reduces processing delay. The reduction occurs because latency-critical tasks are executed at the edge layer instead of the cloud. The graph in figure 3 presents end-to-end latency variation with increasing task load from 50 to 300 tasks for both baseline and proposed systems. The baseline system experiences a latency increase from approximately 420 milliseconds at 50 tasks to almost 1450 milliseconds at 300 tasks which shows a steep growth trend.

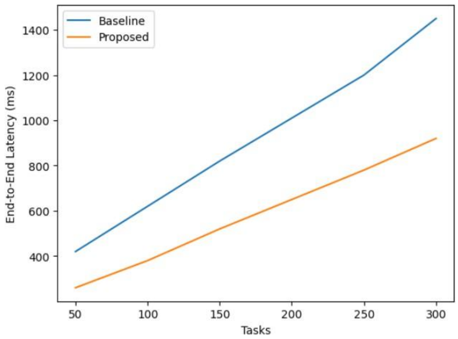

Figure 3: End-to-end latency comparison under varying task loads

The proposed Cloud–IoT–HPC architecture shows significantly lower latency which starts at approximately 260 milliseconds and reaches about 920 milliseconds during the same workload testing period. The latency difference expands as task volume increases with reductions of nearly 160 ms at 50 tasks and over 500 ms at 300 tasks. This demonstrates improved scalability and real-time responsiveness under heavy workloads.

4.3 GPU Utilization Efficiency

GPU usage grows progressively as more vehicles are added to the fleet until it reaches 88% capacity during major system rollouts. The study demonstrates that the HPC cluster achieves effective parallel computation while keeping GPU devices active during processing times. The graph in figure 4 illustrates GPU utilization as the number of robots increases from 5 to 30. Utilization rises steadily from approximately 35% at 5 robots to about 60% near 15 robots, then reaches around 72% at 20 robots and finally approaches 88% at 30 robots. The growth pattern shows that the system distributes parallel workloads effectively among its GPU processing units. The system begins to show active usage through resource optimization which results in minimal downtime after it reaches its first stage of fleet expansion. The architecture demonstrates its capability to utilize GPU acceleration while maintaining effective performance with increasing numbers of deployed robotic units.

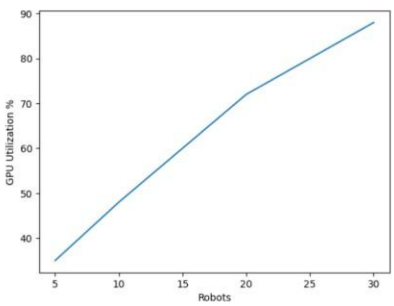

Figure 4: GPU utilization versus number of robots.

4.4 3D Perception Accuracy

The multi-sensor fusion framework enhances environmental comprehension through its ability to handle multiple robots that work together for collaborative mapping. The system achieves precise object detection and dependable scene reconstruction, which results in a mean Average Precision increase from 0.72 to 0.92. The graph in figure 5 demonstrates how mean Average Precision (mAP) accuracy changes when the robot count increases from 5 robots to 30 robots. The accuracy starts at approximately 0.72 for 5 robots and improves to about 0.78 at 10 robots. The accuracy increases to approximately 0.86 accuracy at 20 robots, and it reaches about 0.92 accuracy at 30 robots. The upward trend demonstrates that robots working together with their sensors produce better results for both object detection and environmental comprehension. The reliability of perception increases while uncertainty diminishes through the integration of sensor data from additional robots. The multi-sensor 3D vision framework gains advantages from distributed data fusion which operates together with cooperative mapping.

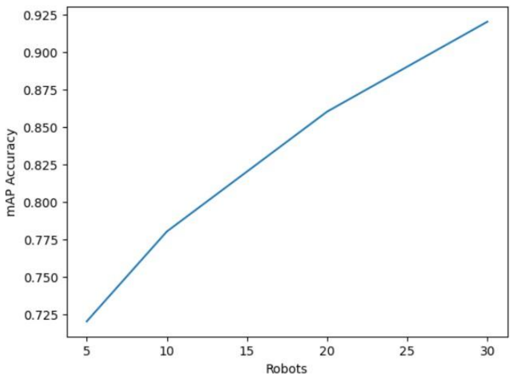

Figure 5: mAP accuracy versus number of robots

4.5 Task Success Rate

The process of autonomous coordination gets improved when people share their understanding of global information. The success rate increases from 78% to 96%, which shows that reinforcement learning decision making methods achieve successful results. The graph in figure 6 presents the Task Success Rate (TSR) percentage as the number of robots increases from 5 to 30. The TSR begins at approximately 78% for 5 robots and rises to around 84% at 10 robots.

The value increases to approximately 88% when there are 15 robots and it reaches about 91% when there are 20 robots. The success rate reaches about 94% with 25 robots and it reaches its highest point of about 96% with 30 robots. The robotic fleet exhibits improved coordination and decision-making efficiency together with enhanced collaborative intelligence as shown by the consistent upward trend. The results demonstrate that reinforcement learning–based policy optimization enhances operational reliability as fleet size scales.

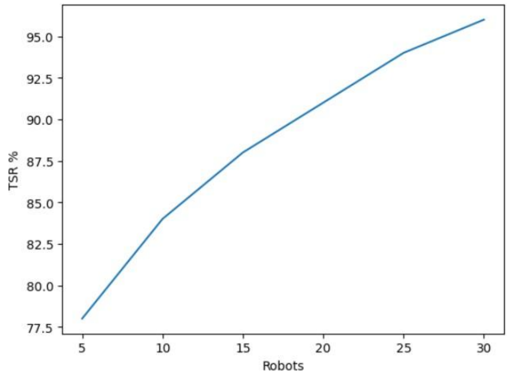

Figure 6: Task success rate versus fleet size

4.6 Energy Performance Gain

Energy savings increase for larger fleets because their operational work distribution reaches better performance. Distributed intelligence offers maximum advantages through larger robotic fleet deployments. The figure 7 shows the energy increase percentage which the proposed architecture achieves when robot count increases from 5 to 30. The energy gain starts at approximately 21% for 5 robots and rises to around 27% at 10 robots. The system reaches a 30% improvement at 15 robots which increases to 33-34% at 20 robots. The gain at 25 robots reaches 35% and it reaches its maximum between 36-37% at 30 robots. The energy-aware scheduling strategy shows better efficiency results when operational work increases for the fleet. Larger robotic deployments benefit more from optimized task distribution and GPU-accelerated processing.

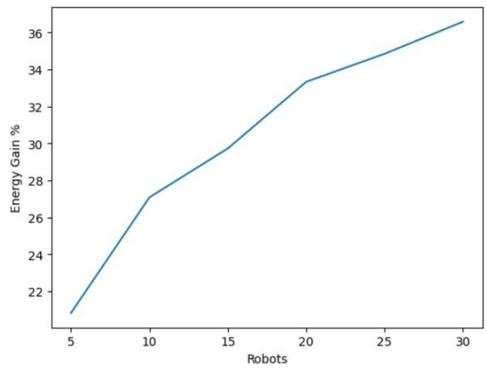

Figure 7: Energy efficiency gain versus fleet size.

4.7 Latency Performance Gain

The continuous ability of edge computing to manage real-time operations shows through stable latency enhancements across different types of workloads. The figure 8 shows latency gain percentage of the proposed system across task loads ranging from 50 to 300 tasks. The latency gain starts at 38.1% which rises to 38.7% when 100 tasks are introduced. The gain decreases with increasing workload until it reaches 36.6% at 150 tasks and 35.7% at 200 tasks which then drops to approximately 35.0% at 250 tasks. The gain at 300 tasks shows a small recovery which reaches almost 36.5%. The architectural design shows permanent substantial latency improvements although system performance decreases with increased workload because of greater communication and processing requirements.

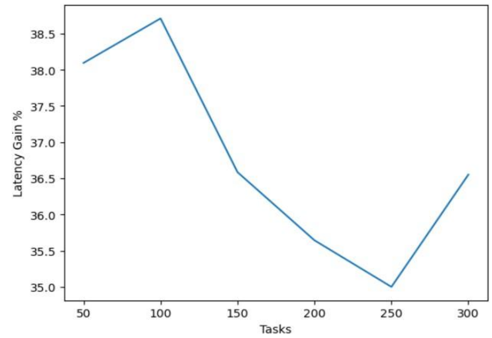

Figure 8: Latency improvement versus task load.

4.8 Edge–Cloud Workload Distribution

The initial processing work handles about 70% of its tasks through edge computing. As the fleet size increases, the cloud system begins to handle additional global optimization duties. The system achieves its two goals through a method that maintains both system responsiveness and system capacity to grow. The figure 9 shows how the edge and cloud systems divide their processing duties when the robot count rises from 5 to 30. At 5 robots, approximately 70% of the workload is processed at the edge, while 30% is handled by the cloud. The edge computing system experiences diminishing workloads which reach 60% at 15 robots and 55% at 20 robots. The two systems distribute their tasks equally between 25 robots because each system handles approximately 50% of the total workload. The edge system manages 45% of its tasks while the cloud system handles 55% of its tasks at 30 robots. The process of adaptive task offloading enables the system to achieve two objectives which include maintaining scalability and central coordination efficiency.

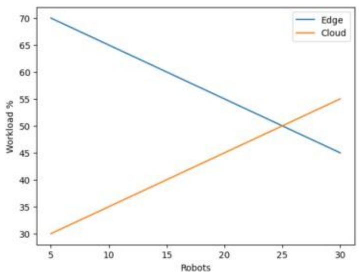

Figure 9: Edge–cloud workload distribution versus fleet size.

4.9 3D Reconstruction Error

The reconstruction error shows a reduction from 0.18 to 0.08 which demonstrates that the mapping accuracy improvements occurred because of the shared robotic systems. The graph in figure 10 shows the variation of 3D reconstruction error with increasing number of robots from 5 to 30. The error starts at approximately 0.18 for 5 robots and decreases to around 0.15 at 10 robots. The system starts with 0.18 error for 5 robots and proceeds to 0.15 error at 10 robots which continues to 20 robots with 0.11 error and 25 robots with 0.10 error before reaching its lowest point of 0.08 error at 30 robots. The continuous decline indicates improved environmental mapping accuracy as more robots contribute sensory information. The system achieves spatial consistency through collaborative multi-sensor fusion which reduces uncertainty while showing that cooperative perception boosts reconstruction accuracy for big robotic systems.

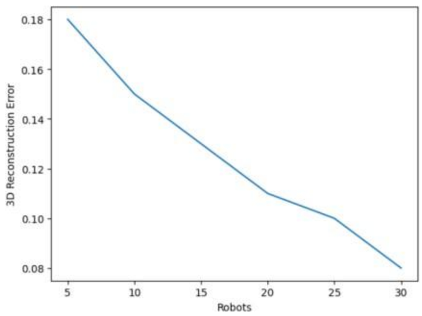

Figure 9: 3D reconstruction error versus number of robots

5 CONCLUSION

This paper introduced a High-Performance Computing (HPC) architecture, developed using Cloud-IoT enabled, GPU-accelerated energy-efficient robotic fleet management systems in smart manufacturing factories. The suggested multi-layer structure, which included the layers of robotic fleet, edge computing, cloud HPC, and IoT management, was successful in overcoming the problem of latency, scalability, and high-computational requirements. Experimental findings showed that there were great improvements in performance than the baseline system. The overall energy consumption was reduced by some 820 to 520 J at 30 robots and the energy gain was found to be almost 3637. When compared to the heavy workloads 1450 ms, the end-to-end latency decreased to 920 ms indicating approximately 3538 percent improvement. The use of GPUs reached as high as 88 percent that confirmed a high degree of resource utilization and parallel processing on the large-scale deployments.

Besides, the incorporation of the GPU-enhanced multi-sensor 3D vision and decision-making by reinforcement learning performed significantly in improving the perception and the reliability of the operations. Mean Average Precision (mAP) rose by 0.72 to 0.92 and the error of 3D reconstruction also dropped by 0.18 to 0.08 with an increase in fleet size. The success rate of the task increased to 96% out of 78% which showed the collaborative intelligence and adaptive coordination was enhanced. The responsiveness and scalability were guaranteed by the adaptive edge-cloud workload distribution at small (70% edge) and large scale (55% cloud) data. The proposed architecture, in general, offers a highly scalable, energy efficient and autonomous solution to a next-generation smart manufacturing system

REFERENCES

Bhadra, Prasenjit, Shilpi Chakraborty, and Subhajit Saha. "Cognitive iot meets robotic process automation: The unique convergence revolutionizing digital transformation in the industry 4.0 era." In Confluence of artificial intelligence and robotic process automation, pp. 355-388. Singapore: Springer Nature Singapore, 2023.

Hoe, Min, and Jamal Dargham. "High Performance Computing (HPC) Applications in Industry 4.0 (I4. 0) for the betterment of humanity." In 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), pp. 1-6. IEEE, 2020.

Tyagi, Amit Kumar, Pooja Bhatt, N. Chidambaram, and Shabnam Kumari. "Artificial intelligence empowered smart manufacturing for modern society: a review." Artificial Intelligence-Enabled Digital Twin for Smart Manufacturing (2024): 55-83.

Othman, Uqba, and Erfu Yang. "Human–robot collaborations in smart manufacturing environments: review

and outlook." Sensors 23, no. 12 (2023): 5663.
Xu, Xiaobin, Lei Zhang, Jian Yang, Chenfei Cao, Wen Wang, Yingying Ran, Zhiying Tan, and Minzhou Luo. "A review of multi-sensor fusion slam systems based on 3D LIDAR." Remote Sensing 14, no. 12 (2022): 2835.
Ekechi, Chijioke C. "Deep Learning for Real-Time Fault Detection in Wireless Robotic Systems." (2025).
Ajayi, Rhoda. "Integrating IoT and cloud computing for continuous process optimization in real-time systems." Int J Res Publ Rev 6, no. 1 (2025): 2540-2558.
Pacella, Massimo, Antonio Papa, Gabriele Papadia, and Emiliano Fedeli. "A scalable framework for sensor data ingestion and real-time processing in cloud manufacturing." Algorithms 18, no. 1 (2025): 22.
Khalid, Ayesha. "Performance Evaluation of Edge Computing for Latency-Sensitive Applications." Global Journal of Multidisciplinary and Applied Sciences 3, no. 2 (2025): 101-108.
Gathu, Simon. "High-performance computing and big data: Emerging trends in advanced computing systems for data-intensive applications." Journal of Advanced Computing Systems 4, no. 8 (2024): 22-35.
Vaithianathan, Muthukumaran. "The future of heterogeneous computing: Integrating cpus, gpus, and fpgas for high-performance applications." International Journal of Emerging Trends in Computer Science and Information Technology 6, no. 1 (2025): 12-23.
Rojek, Izabela, Piotr Prokopowicz, Maciej Piechowiak, Piotr Kotlarz, Nataša Náprstková, and Dariusz Mikołajewski. "The Impact of Data Analytics Based on Internet of Things, Edge Computing, and Artificial Intelligence on Energy Efficiency in Smart Environment." Applied Sciences 16, no. 1 (2025): 225.
Rema, Catarina, Pedro Costa, Manuel Silva, and Eduardo J. Solteiro Pires. "Task scheduling with mobile robots—a systematic literature review." Robotics 14, no. 6 (2025): 75.
Mohammed, Rahimoddin. "Artificial intelligence-driven robotics for autonomous vehicle navigation and safety." NEXG AI Review of America 3, no. 1 (2022): 21-47.
Vermesan, Ovidiu, Arne Bröring, Elias Tragos, Martin Serrano, Davide Bacciu, Stefano Chessa, Claudio Gallicchio et al. "Internet of robotic things–converging sensing/actuating, hyperconnectivity, artificial intelligence and IoT platforms." In Cognitive hyperconnected digital transformation, pp. 97-155. River Publishers, 2022.
Romanelli, Fabrizio. "Multi-Sensor Fusion for Autonomous Resilient Perception."
Kirimtat, Ayca, and Ondrej Krejcar. "GPU-based parallel processing techniques for enhanced brain magnetic resonance imaging analysis: a review of recent advances." Sensors 24, no. 5 (2024): 1591.
Ajayi, Rhoda. "Integrating IoT and cloud computing for continuous process optimization in real-time systems." Int J Res Publ Rev 6, no. 1 (2025): 2540-2558.
Alex, Bazigu, and Mwebaze Johnson. "A framework for IoT-enabled smart manufacturing for energy and resource optimization." arXiv preprint arXiv:2502.03040 (2025).
Soret, Beatriz, Lam D. Nguyen, Jan Seeger, Arne Bröring, Chaouki Ben Issaid, Sumudu Samarakoon, Anis El Gabli, Vivek Kulkarni, Mehdi Bennis, and Petar Popovski. "Learning, computing, and trustworthiness in intelligent IoT environments: Performance-energy tradeoffs." IEEE Transactions on Green Communications and Networking 6, no. 1 (2021): 629-644.
Salhaoui, Marouane. "Smart IoT monitoring and real-time control based on autonomous robots, visual recognition and cloud/edge computing services." (2021).
Ramareddy, Sathish Kaniganahalli. "A CLOUD-BASED AI FRAMEWORK FOR REAL-TIME FINANCIAL DATA VISUALIZATION AND DECISION SUPPORT." International Journal of Applied Mathematics 38, no. 11s (2025): 1206-1225.
Lilhore, Umesh Kumar, Sarita Simaiya, Yogesh Kumar Sharma, Anjani Kumar Rai, S. M. Padmaja, Khan Vajid Nabilal, Vimal Kumar, Roobaea Alroobaea, and Hamed Alsufyani. "Cloud-edge hybrid deep learning framework for scalable IoT resource optimization." Journal of Cloud Computing 14, no. 1 (2025): 5.
Liu, Liu, Shaoshan Liu, Zhe Zhang, Bo Yu, Jie Tang, and Yuan Xie. "PIRT: A runtime framework to enable energy-efficient real-time robotic applications on heterogeneous architectures." arXiv preprint arXiv:1802.08359 (2018).
Chen, Kaiyuan Eric, Yafei Liang, Nikhil Jha, Jeffrey Ichnowski, Michael Danielczuk, Joseph Gonzalez, John Kubiatowicz, and Ken Goldberg. "Fogros: An adaptive framework for automating fog robotics deployment." In 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE), pp. 2035-2042. IEEE, 2021.

Sánchez-Ribes, Víctor, Antonio Maciá-Lillo, Higinio Mora, and Antonio Jimeno-Morenilla. "Efficient GPU Cloud architectures for outsourcing high-performance processing to the Cloud." The International Journal of Advanced Manufacturing Technology 133, no. 1 (2024): 949-958.

Ahmad, Syed Zubair, and Farhan Qamar. "A hybrid AI based framework for enhancing security in satellite based IoT networks using high performance computing architecture." Scientific Reports 14, no. 1 (2024): 30695.

Sahoo, Soumya. "Sensor Fusion and Virtual Sensor Design for Enhanced Multi-Sensor Data Accuracy in Autonomous Systems." International Journal on Smart & Sustainable Intelligent Computing 1, no. 2 (2024): 21-39.

Lu, Lu, Yongtang Yuan, Yongkang Xie, Shangqin Yuan, Jingwen Song, Han Luo, Yamin Li, Jihong Zhu, and Weihong Zhang. "Autonomous intelligent additive manufacturing of continuous fiber-reinforced composites: data-enhanced knowledgebase and multi-sensor fusion." Virtual and Physical Prototyping 19, no. 1 (2024): e2412192.

Yousif, Ibrahim, Liam Burns, Fadi El Kalach, and Ramy Harik. "Leveraging computer vision towards high-efficiency autonomous industrial facilities." Journal of intelligent manufacturing 36, no. 5 (2025): 2983-3008.